LaSalle Universidad

Una obra de los Hermanos de La Salle

+300 años de experiencia en educación

Innovación y Software

VOL 4 N° 1 Marzo - Agosto 2023 ISSN N° 2708-0935

Revista de la Facultad de Ingeniería de ULASALLE

Universidad La Salle, Arequipa, Perú
facin.innosoft@ulasalle.edu.pe
https://revistas.ulasalle.edu.pe/innosoft

ARK: ark:/42411/s11
PURL: 42411/s11

ENFOQUE Y ALCANCE

La Revista Innovación y Software de la Universidad La Salle es una publicación científica arbitrada, con periodicidad semestral, especializada en Ingeniería de Software, Ciencias de la Computación y Tecnologías de la Información dando cobertura a las siguientes temáticas:

- Ingeniería de software
- Gestión de software
- Calidad de software
- Técnicas de programación
- Programación paralela y distribuida
- Tecnologías de bases de datos
- Inteligencia artificial
- Procesamiento de imágenes
- Reconocimiento de patrones
- Redes y seguridad informática
- Tecnologías de la información y las comunicaciones
- Desarrollo de aplicaciones informáticas

COMITÉ EDITORIAL

EDITORIAL

ARTÍCULOS ORIGINALES

La Revista Innovación y Software de la Facultad de Ingeniería, en la Universidad La Salle, se complace en presentar este primer número de su cuarto volumen que tiene como objetivo el promover investigaciones, los cambios y usos de nuevos elementos tecnológicos y su interrelación con la Ingeniería de Software y la Ciencia de la Computación.

Si bien uno podría pensar que los robots y las máquinas autónomos son cosa del futuro, la realidad es que hay muchos ejemplos de inteligencia artificial que ya funcionan a la sombra de la actividad humana. Y, en muchos casos, afectan a muchos aspectos de nuestra vida diaria. Hablar de inteligencia artificial (IA) es tan fácil como hablar de máquinas inteligentes. Es decir, las máquinas están programadas para realizar ciertas tareas automáticamente sin que los humanos supervisen su trabajo. De esta forma, la inteligencia artificial aparece como una rama de la informática, la disciplina encargada de programar máquinas inteligentes.

La IA puede ser una oportunidad para mejorar nuestra sociedad si comunicamos claramente las consecuencias de aplicar esta tecnología y logramos que las legislaturas promulguen leyes de redistribución de la riqueza que complazcan a todos (impuesto a los robots, salario mínimo universal, etc.). La educación de una sociedad en el uso de la tecnología es clave para desarrollar ciudadanos con conciencia crítica que puedan determinar la dirección en la que se usa esa tecnología. Por ejemplo, la automatización de tareas agrícolas, manufactureras o de transporte eliminará puestos de trabajo y creará otros nuevos. El balance en este proceso obviamente estará en la dirección de reducir el tiempo de trabajo para las necesidades básicas.

Comité Editorial

An observation toward Computer aided processes in Garments production. Comparison and analysis of CAD/CAM Software in Bangladesh

Una observación a los procesos asistidos por ordenador en la producción de prendas de vestir. Comparación y análisis del software CAD/CAM en Bangladesh en Bangladesh

Md. Shazzat Hossain
University of Dhaka, Dhaka, Bangladesh.

Md. Abdus Samad
University of Dhaka, Dhaka, Bangladesh.

Md. Hasan Ali
University of Dhaka, Dhaka, Bangladesh.

Prahlad Sutradhar
University of Dhaka, Dhaka, Bangladesh.

Jubayer Islam
University of Dhaka, Dhaka, Bangladesh.

Md. Hazrat Ali
Bangladesh University of Textiles . Dhaka, Bangladesh.

Md Al-Amin
Bangladesh University of Textiles . Dhaka, Bangladesh

Md. Zahir Uddin Babar
Green University of Bangladesh. Dhaka, Bangladesh.

Md. Rezaul Karim
University of Rajshahi. Rajshahi, Bangladesh

Mohammad Ullah
Khulna University of Engineering and Technology. Khulna, Bangladesh

ARK: ark:/42411/s11/a76

PURL: 42411/s11/a76

RECIBIDO 15/09/2022 • ACEPTADO 25/10/2022 • PUBLICADO 30/03/2023

RESUMEN

Este trabajo de investigación indaga sobre los diferentes atributos de los procesos asistidos por ordenador en la producción de prendas de vestir. Esta perspectiva de investigación fue realizada por nuestro valiente equipo de 2021 a 2022. Revela información adecuada sobre la intención de la industria de la confección y los criterios para la elección de software CAD/CAM. Por el bien de esta investigación, hemos visitado más de 600 industrias para recopilar datos en bruto; Cada industria trató de asistir a esta investigación de una región diferente de Bangladesh de buena gana. Después de recoger todos los datos en bruto de la industria de la confección. Los datos de la industria de prendas de vestir fueron coordinados por el programa Excel. Consecuentemente, los datos fueron analizados e implementados estadísticamente para identificar los atributos de la Industria de la Confección para la satisfacción con el software CAD/CAM. Este proceso también detecta muchos retos y define y aconseja una solución adecuada a los problemas a los que se enfrenta la industria de la confección en la situación actual. Este trabajo de investigación muestra información adecuada sobre los criterios y la demanda de la industria de la confección a la hora de adquirir software asistido por ordenador

Palabras claves: Software CAD y CAM textil, diseño asistido por ordenador, fabricación asistida por ordenador, proceso asistido por ordenador, software CAD y CAM de prendas de vestir.

ABSTRACT

This research paper inquires about different attributes of Computer-aided processes in garments production. This perspective Research was done by our courageous team from 2021 to 2022. It reveals adequate information on the Garments industry's Intension and criteria for choosing CAD/CAM software. For the sake of this Research, we visited more than 600 industries to gather raw data; Every Industry tried to attend this Research from a different region of Bangladesh willingly. After collecting all raw data from the garments industry. The data of the Garments industry was coordinated by Excel program. Consequently, the data was analyzed and implemented statistically to identify the Garments Industry attribute for satisfaction with CAD/CAM software. This process also detects many challenges and defines and advises a proper solution to the problems that the Garments industry is facing in the current situation. This research paper demonstrates adequate information about the Garment's criteria and demand in purchasing garments Computer-aided software.

Keywords: *Textile CAD and CAM software, Computer aided design, computer aided manufacturing, computer aided process, Garments CAD and CAM software.*

Introducción

Nowadays, the garments industry faces numerous problems that must be addressed through constant stamina and Targeted Research. Insufficiency is the constant companion of Bangladesh that surprisingly influenced Industry behavior [1]. The clothing market is dependent on foreign buyers because of targeted profit. Nevertheless, in this Modern period, buyer expectation is higher than usual. So every industry focuses on buyer demand and satisfaction. Buying behaviors of a Buyer and uses of the Computer-aided software of industry rely on different aspects of retention, perception, financial state, and the other Circumstance. Owing to the issue, every vendor needs to realize what the industry is willing to buy and use and what affects Buy er satisfaction [5]. With the modern revolution, industry and buyer behavior have alternated dramatically in the past few years. Every Buyer focuses on their money and how they spend more than ever before, Rather than wasting energy and time. The Buyer wants products that satisfy their Criteria, and the industry relies on some affordable attributes [6]. Foreign buyers are not a complicated attitude but are harder to define [3].Computer-aided processes play a very far-reaching role in garment production. The CAD and CAM software is Garments products capable of enhancing garment production. It also played a vital role in Buyer satisfaction and provoked buyer loyalty, which impacted the specific company's reputation. The majority, Companies are aware of increasing the production flow of an industry [13]. However, it depends on many affordable attributes [1]. Now a day, the demand has increased dramatically in the marker [10]. Hence, this Research focuses on the present study to explore the garments industry behavior in using CAD/CAM software and Buyer satisfaction. Regarding performing the research analysis, The garments industry and buyer satisfaction were treated as the root point. Eventually, the Different Respondent sentiment was considered about the Attribute of garments Cad and Cam software. [11] . This Research detects the preferable Attribute for Using the behavior of Garments Computer aided process [2].

Procedure

The goal of this inquiry was to collect the data with specific Research on numerous Bangladeshi Garments industry on the different aspects such as quality, time, performance, price, mistake, production size, prices, etc. The Buyer expectation was implemented using targeted Research [12]. The Cad and cam users experimented with specific designing with buyer requirements. They enabled the Buyer to write their honest opinions and expectation of demand about the Cad and cam software [2]. The Goal of this inquiry was to collect the data with a specific research of numerous Bangladeshi Garments industry on the different aspects such as quality, time, performance, price, mistake, production size, prices, etc. The Buyer expectation was implemented by using targeted research [12]. The Cad and cam user were experimented with specific designing with buyer requirements. This enabled the Buyer to write their open opinions and expectation of demand about the Cad and cam software [2].

Obstacle Analysis

The analysis has been performed to detect Buyer demand and the Attribute of Cad and cam software. Simultaneously it determines the industry aspects of using specific software and its root causes [13]. The general Research was conducted through Cad and Cam engineers in the targeted industry, and this study tried to focus on the Buyer's needs. This Research was initially carried out to collect valuable data with Targeted criteria that create a sense of overall behavior for Industry and Buyers. Overall, using software's behavior positively impacts both Attribute and gratification, ultimately leading to customer Satisfaction [4].

Evolve Questions

The provided questions and experiments are created to adequately collect information to construct a view of a user and Buyer's intentions regarding disjunction variables such as time, price, activity, design quality, preference, and reason for using specific software and Buyer expectation [5]. The user can answer with "a tick mark" and show their experiments in garments CAD and CAM lab. This study expounds on the psychology of different users regarding Garments CAD and CAM software and Buyer demands [8].

Data assortment and Inspect

There were 600 users and 600 buyers attending this research experiment; most were aged 18 to 55. Most respondents are Bangladeshi users and foreign buyers [14]. Most of the experiments and interview locations were chosen at the industry lab, where engineers from different industries use their CAD and CAM software [9].

Final Data Scraping and Analysis

Final Data scraping and analysis was a very challenging part of this Research. Because we have focused on 2 sites-

1] User experience and satisfaction
2] Buyer criteria and demands

Data analysis is regulated through the User intention and faculty to use specific software in fewer worthwhile groups. In the same way, it analyzed the Buyer's demands. The statistical data analysis succeeds using the process of an open coding system. Consequently, this method of data analysis permits us to illustrate the actual data [6]

Magnitude and measurement

For data collecting, we tried to make a sample size of 600 Industry Users for respondents we took to ensure adequate ratios where most respondents are directly CAD and CAM software experts. On the other hand, we have collected information from more than 600 buyers about why they require and Satisfy with specific software [4]. Eventually, the authors can collect Data and information from respondents from 5 different Industrial areas such as Dhaka, Tangail, Gazipur, Narayanganj, and Narshingdi [2].

Data manipulation and Integration

Data integration is a vital element of quality research, where this process considers several analyses with some data points [15]. It controlled ideal and correct decisions with decreasing misleading data. This process can evaluate and recognize data where the specific statistics come from respondents. That is why this combination is implemented for quality results [6]

Primary analysis & Discussions

1st and 2nd Priority level To Using CAD/CAM software For a Software user:

In this analytics, two specific charts represent the priority level of the different users with some attributes. However, the line and bar charts represent exact data and the same value differently. The Research Determine that the majority of the Software user give first preference to production and second priority to the quality of the Software performance and production. According to the linear priority level, this is clear that the highest preference is production size with minimum time. So every user of an industry chooses the software with Bulk production to save valuable time. The following Line chart shows the highest priority for software production, where first and second priority accounted for 71.50% and 69.08%, respectively. The bar diagram clearly expressed that the price, speed, and mistake are the maximum preferences for second priority. The following line diagram shows the volume of each priority in percentage.

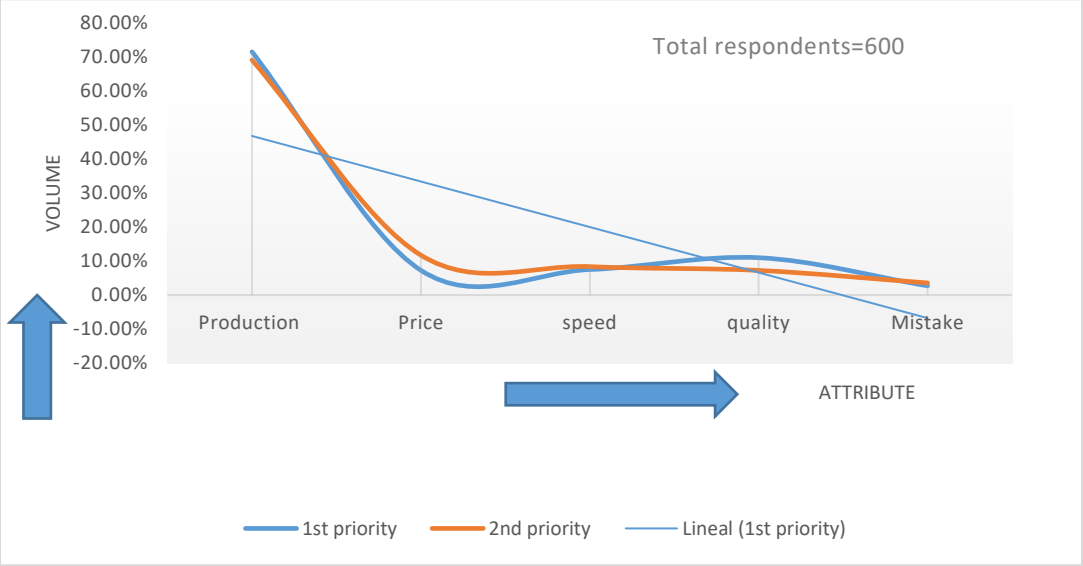


Figure 1: 1st and 2nd priority of software user with Line chart

This bar chart is illustrated the priority level of software user considering with some attribute. However the following bar diagram created with almost same data and value.

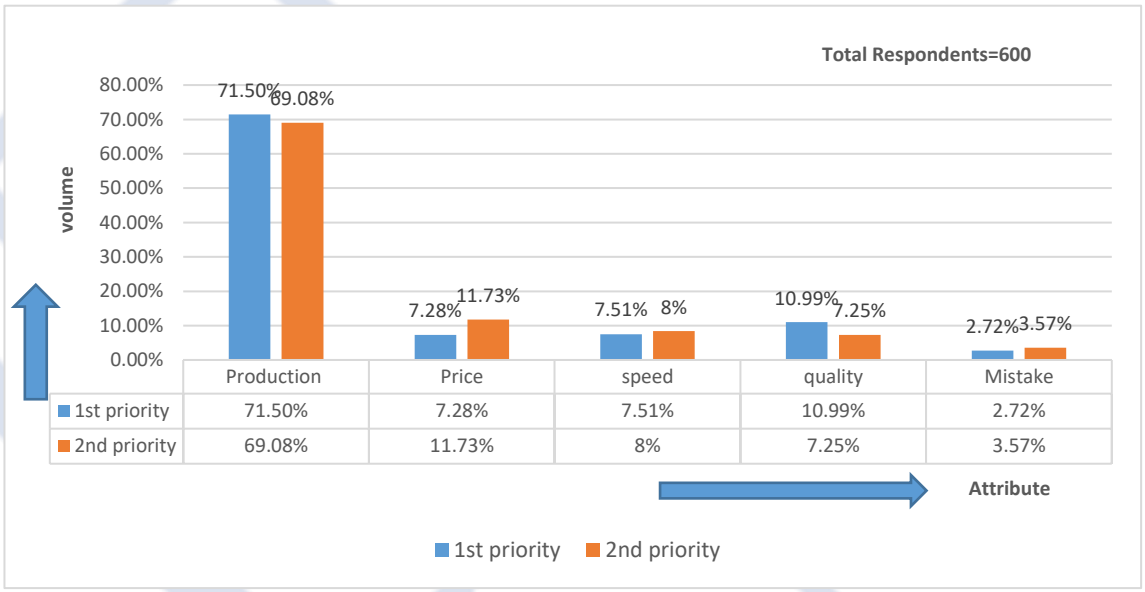


Figure 2: 1st and 2nd priority of software user with Bar chart

Here, in the upper following chart
Production = the amount of production per hour or per unit time.
Price = how much needed to buy a software/ the amount of money.

Speed= the software flexibility and how much it faster.
Quality= the quality of the design and final production
Mistake= the possibility of mistake when a software user working on a design.

Quality Selection of CAD and CAM software

For quality selection, Firstly, we have to choose five attributes, such as time to make a design, the software's brand, the software's price, the possibility of Mistakes when making the design, and the complexity of the software. A user is keen on software quality, which most users consider a maximum priority as the software brand and time for making a design. The priority accounted for Brand, Time, price, mistake, and complexity, respectively, 37%, 35%, 15%, 7%, and 6%, and it was followed by just 25%, 30%, 20%, 10%, and 15% for second priority. So, the user is Possibility of the mistake and Complexity of software in case of quality selection. The possibility of mistakes and complexity of software is almost identical for all types of software. People always try to focus on the brand and time to make a design. The volume of priority for quality is shown individually in the following bar chart and pie chart.

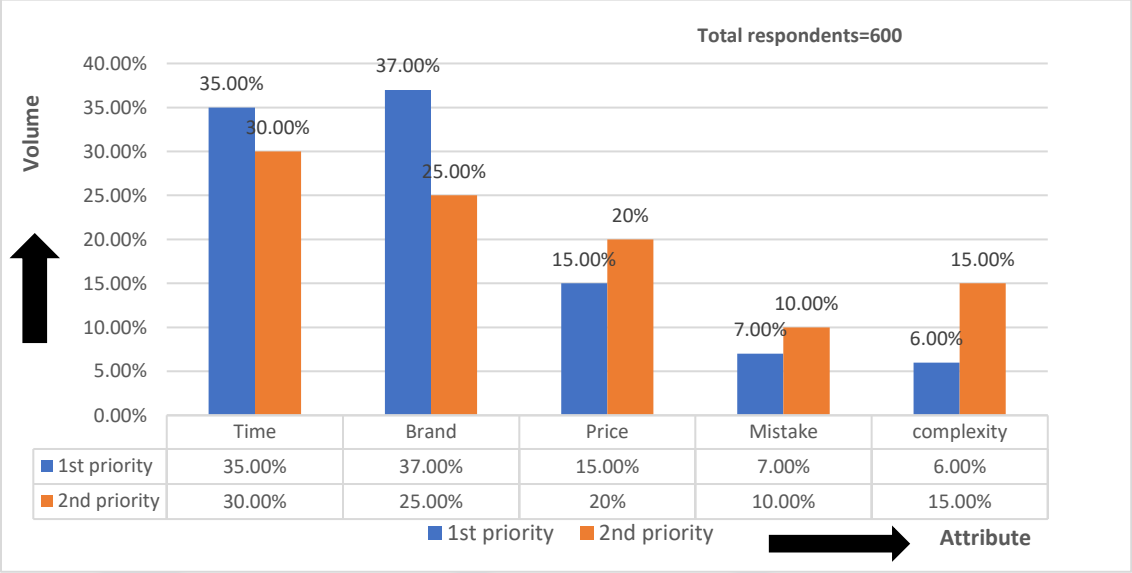


Figure 3: Bar chart for quality selection of CAD and CAM software

Buyer priority to select software with their satisfaction:

Most of the Buyers intended to Use branded and well-reputed company software rather than the price and speed of the software. However, the highest proportion of buyers chooses to use a software brand and the production quality, accounting for 60.59% and 39.41% for priority, followed by just 41.27% and 58.73%, respectively, for second priority. Among them, Buyers avoid the price of the software and how much faster it is. Buyers are confident that another software attribute will be better if the brand is well-reputed. That is why every Buyer avoided the price and faster properties of the software. So a buyer chooses the best-branded software like Lectra or Optitex software. However, Buyer focuses on Lectra software in the first place. The software selection priority percentage is shown in the following bar chart.

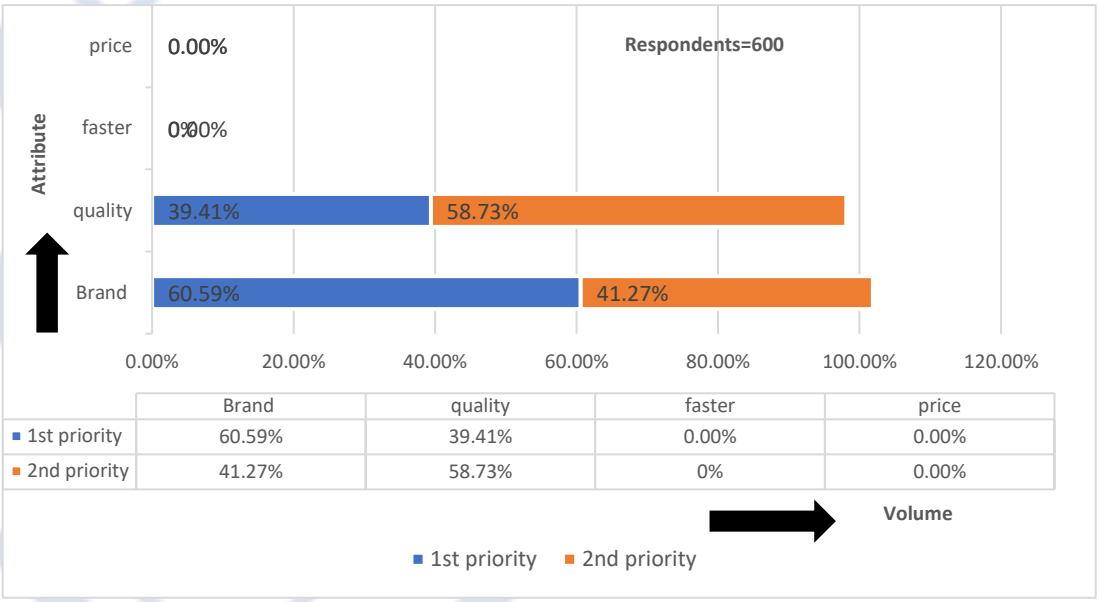
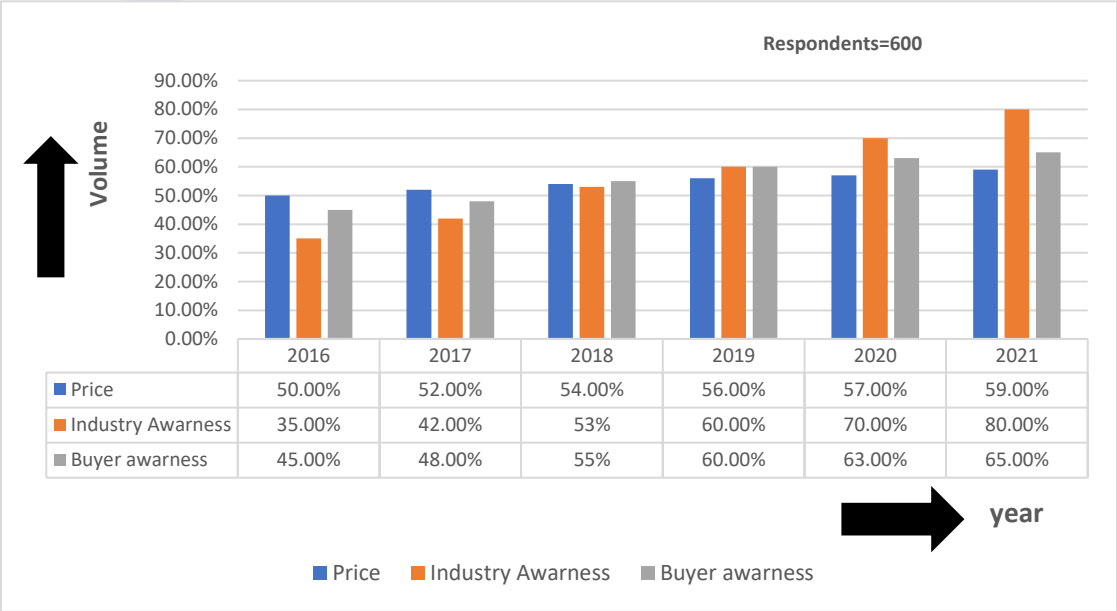


Figure 4: Buyer priority for software selection with bar chart

How garments industry and buyer awareness increases over time:

This Bar chart and line chart are created with the same data and properties. However, both charts clearly expressed the industry and buyer awareness about using CAD/CAM software. There is little comparison of software price with their awareness over the period. For the last 6 years, the software price has gradually increased, and awareness has changed with the price. According to the chart, the software prices increased in 2016, 2017, 2018, 2019, 2020, and 2021 respectively, 50%, 52%, 54%, 56%, 57%, and 59%. From the line chart, it is found that buyer awareness was maximum in 2016 when the company used the software. However, industry awareness was lower than buyer awareness. However, gradually, industry awareness surpassed buyer awareness in

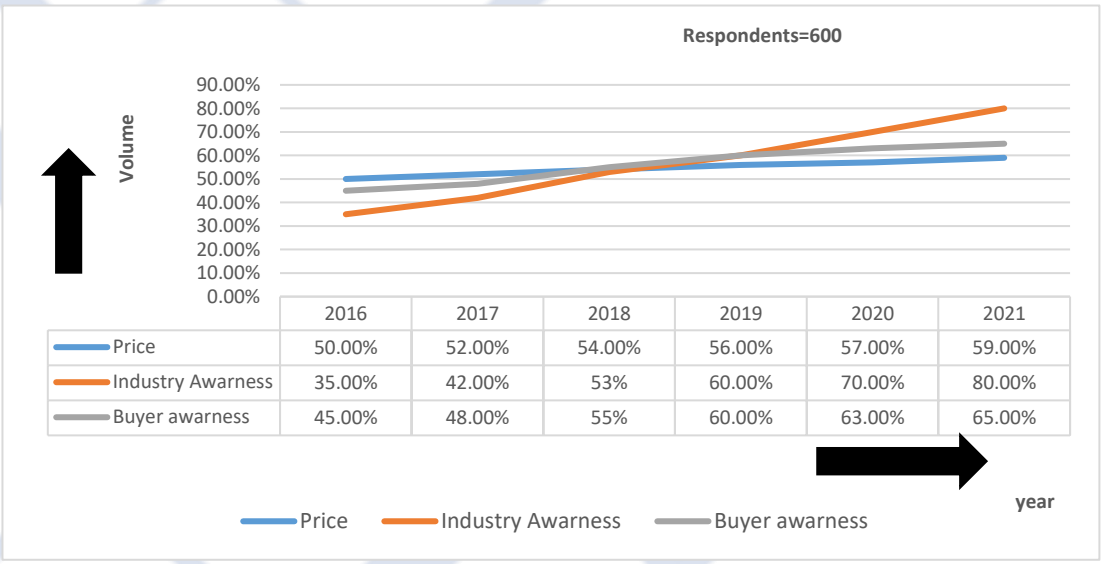
2021. The following bar sketch show the industry and Buyer awareness over time. However the Bar chart



	2016	2017	2018	2019	2020	2021
Price	50.00%	52.00%	54.00%	56.00%	57.00%	59.00%
Industry Awarness	35.00%	42.00%	53%	60.00%	70.00%	80.00%
Buyer awarness	45.00%	48.00%	55%	60.00%	63.00%	65.00%

Figure 5: Industry and buyer awareness overtime with bar chart

The Following line sketch is clearly illustrated the buyer awareness, industry awareness and price of the software. However this line sketch created with same data and same value. But this chart more clear explanation of awareness changing over time.

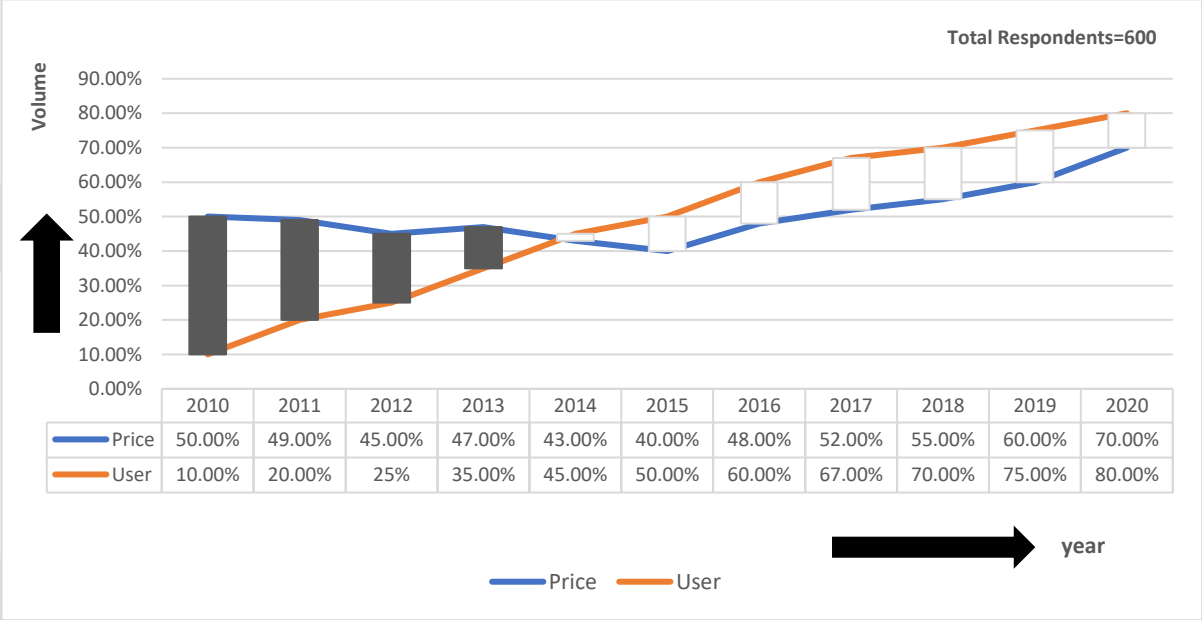


	2016	2017	2018	2019	2020	2021
Price	50.00%	52.00%	54.00%	56.00%	57.00%	59.00%
Industry Awarness	35.00%	42.00%	53%	60.00%	70.00%	80.00%
Buyer awarness	45.00%	48.00%	55%	60.00%	63.00%	65.00%

Figure 6: Industry and buyer awareness overtime with line chart

How industry user increased with software price:

The following line chart represents how the price of the software and software user changed over time. This analysis collects data from the last 11 years. The price of the CAD and CAM software increased and decreased over time from 2010 to 2020 which is accounted for 50% for 2010, 49% for 2011, 45% for 2012, 47% for 2013, 43% for 2014, 40% for 2015, 48% for 2016, 52% for 2017, 55% for 2018, 60% for 2019 and 60% for 2020. Where CAD/CAM software user increased 10% , 20% , 25% , 35% , 50%, 60% , 67% , 70% , 75% and 80% following by the year. The result of user with software price is shown in the following figure.



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Price	50.00%	49.00%	45.00%	47.00%	43.00%	40.00%	48.00%	52.00%	55.00%	60.00%	70.00%
User	10.00%	20.00%	25%	35.00%	45.00%	50.00%	60.00%	67.00%	70.00%	75.00%	80.00%

Figure 7: Industry user increasing trend with Software price

Most used software in Bangladesh with industry satisfaction:

This analysis primarily focuses on the final result of the whole Research because this Research represents which software is usually used by the industry for their satisfaction and why this software meets their industry criteria.

The following chart explains the software name, number of the software user, type of factory, and price of the software in Bangladeshi currency, taka. However, more than 355 industries

intended to use Lectra software. However, the price of the software is high, but most big factories like EPZ and others prefer to use Lectra Modaris software. On the other hand, 150 industries like to use Optitex software. Most of the factories are medium class. So they never want to spend a big budget on buying Lectra software. The small factory chooses to use Gerber software for the low price. They choose Gerber for low prices and a small amount of production. Approximately 600 industries were attended in this Research.

The following table shows the number of software user with different attribute.

Name of the software	Number of user	User percentage%	Type of factory	Price of software[Taka]
lectra	355	59%	Big factory	10,00000-15,00000
optitex	150	25%	Medium factory	300000 - 500000
Gerber	95	16%	Small Factory	250000-300000

Note: Number of respondents=600 industry

Table 1: Number of software user industry with a table

An actual Garments order sheet for analysis:

For the research purpose, we have the carry out a real test with a Buyer order sheet. Most of the respondents used their preferred software with specific measurements. The buyer KIABI company sends a requirement for a half shirt. The specific measurement below is in figure 1; on the other hand, figure 2 represents the half shirt's specific design style, where more than 600 respondents were directly involved with making this half shirt design.

We carefully observe the software performance when the software user is trying to make a design. Our observation includes some basic questions and parameters. Such as-

1. How much time is required to make this specific design?
II. The possibility of mistakes was carefully observed.
III. The production ability was calculated with this software.
IV. Quality of the final design
V. Quality of the final product with specific measurement.
VI. Buyer satisfaction
VII. Complexity of the design with the specific software, etc.

Where Lectra Modaris software performed terrifically, so every big, gigantic factory like EPZ, Square Fashions ltd, Denim, Ha-meem group, Beximco, DBL Group, Opex Sinha Group, Fakir Group, Epyllion Group, Standard Group, Asian Apparels Limited, Viyellatex Limited and the AJI group is using Lectra Modaris software.

On the other hand, Optitex is also better for medium-quantity production. The performance of using Gerber software is fewer users than in small factories. But for small production, this

software suits. During our research we have tested more than 600 user with different measurement.

The following figure 1 represent the measurement chart for the half shirt of KIABI company.

KIABI	TECHNICAL FILE : **ARKW22CDOM6** (ZH331)		GPM : TANT VALERIE	Version 1 - Page 4 / 6
	Season : HIVER 22		Designer : TRAN HUYEN TRAN	14-02-2022 09:49
Description : P6 SPE DOM TOM TH TROP 6 CHEMISE MC COL MAO IMPRIME ET UNI BCI COTTON				
Technical description :				
Market : KIDSBOY	Department : GRKIDS	Class : KIBSHIRTSS	Category : ALL	Typology : FVI	Type : SHIRT

Measurement chart >> SHIRT short sleeves

Measure		Ctrl	Tol +	Tol -	3A	4A	5A	6A	8A	10A	12A	Comment
AA - 1/2 Chest round-	CM	X	1.0	1.0	33.0	34.0	35.0	36.0	38.0	41.0	44.0	
WW - 1/2 Waist round	CM	X	1.0	1.0	32.0	33.0	34.0	35.0	37.0	40.0	43.0	
HWF - Front waist height from shoulder	CM		0.0	0.0	24.0	25.0	26.0	27.0	31.5	34.5	37.5	
VV - 1/2 Bottom round	CM		1.0	1.0	33.0	34.0	35.0	36.0	38.0	41.0	44.0	
CF - Front breadth	CM		0.5	0.5	24.5	25.5	26.0	27.0	28.5	31.0	34.0	Take at middle of armhole
CB - Back breadth	CM		0.5	0.5	25.0	26.0	26.5	28.0	29.5	32.0	35.0	Take at middle of armhole + back pleat closed
HSF - Front from shoulder lenght	CM	X	1.0	1.0	35.5	40.5	42.5	44.0	48.0	52.0	57.0	*
HSB - Back from shoulder length	CM	X	1.0	1.0	37.0	42.0	44.0	45.5	49.5	53.5	58.5	*
EB - Shoulder incline degree	CM		0.0	0.0	22.0	22.0	22.0	22.0	22.0	22.0	22.0	
EA - Shoulder length	CM		0.5	0.5	7.5	8.0	8.5	9.0	10.0	11.0	12.0	At shoulder fold
EH - Armhole height	CM		0.5	0.5	13.5	14.0	14.5	15.0	16.0	17.0	18.0	
SHW - 1/2 Upper sleeve width	CM	X	0.5	0.5	11.0	11.5	12.0	12.5	13.5	14.5	15.5	
SWB - 1/2 Bottom sleeve (short)	CM	X	0.5	0.5	10.0	10.5	11.0	11.5	12.5	13.5	14.5	
SLS - Sleeve length (short)	CM	X	1.0	1.0	12.0	12.5	13.0	13.5	14.0	15.0	16.0	

Figure 8: KIABI company measurement chart for half shirt

KIABI	TECHNICAL FILE : **ARKW22CDOM6** (ZH331)		GPM : TANT VALERIE	Version 1 - Page 2 / 6
	Season : HIVER 22		Designer : TRAN HUYEN TRAN	14-02-2022 09:49
Description : P6 SPE DOM TOM TH TROP 6 CHEMISE MC COL MAO IMPRIME ET UNI BCI COTTON				
Technical description :				
Market : KIDSBOY	Department : GRKIDS	Class : KIBSHIRTSS	Category : ALL	Typology : FVI	Type : SHIRT

Figure 9: CAD/CAM design for half shirt

Automatic template design system:

An automatic template design system is a process where a user is capable of making a CAD/CAM design with a previously created template. This Template provides the facility to make a garment design easily. This automatic template design system allows users to re-edit the design with their criteria [16]. A designer can re-edit the measurement also. So the time to make a design is less in this automatic system. The purpose of discussing these properties is to make it clear which software consists of the highest facilities.

Where Lectra has a rich resource of automatic templates because Lectra is a software created by a well-reputed company, they are trying to develop the software within a period. They updated the software gradually and added a lot of automatic template designs.
Optitex software recently added some automatic template design systems. Nevertheless, this is less effective compared to Lectra software.
On the other hand, Gerber still has backdated software compared to Lectra and Gerber. So this software has no resources for an automatic template design system.

Final analysis and discussion:

Now the final analysis is done from previous preliminary Research. The final analysis represents the actual parameter and Attribute, which is why a company chooses a specific software. The following table is more complex than the whole Research. Where it represents the different attributes, this specifically represents why companies choose software for their industry [12] . The positive Attribute of Lectra software is the quality of design, the possibility of mistake, and the production rate, which accounted for 98%, 8%, and 99%, respectively. Where Optitex is accounts for 80%, 15%, and 70%, respectively. The Gerber also accounted for 75%, 20%, and 50%, respectively. The possibility of the mistake of creating a design is less in Lectra, which plays an essential role in choosing Lectra.
However, Gerber and Optitex are faster, have minimum complexity, and fewer prices. Where Gerber accounted for 90%, 70%, and 25%. On the other hand, Optitex accounted for 80%
However, Optitex and Gerber have many more limitations than Lectra. The following bar chart represents the different Attributes with different software with the following data table.

Software	speed	quality	mistake	complexity	price	production
lectra	75% ↓	98% ↑	8% ↑	80% ↓	100% ↓	99% ↑
optitex	80%	80%	15%	75%	35%	70%
gerber	90%	75%	20%	70%	25%	50%

Table 2: different attribute for a software

Here,
Speed = the working speed of the software
Quality =quality of the garments design which done by CAD/CAM software
Mistake= the possibility of mistake when a user using a software to making design
Complexity= how difficult the software to learn and how difficult to Operate the software
Price= price of the specific software to buy.
Production = production size or amount of production per hour.

= IT INDICATE POSITIVE IMPACT = IT INDICATE NEGATIVE IMPACT

The following bar chart illustrated in accordance with data and value of table 2.which clearly explains the software popularity with user priority level.

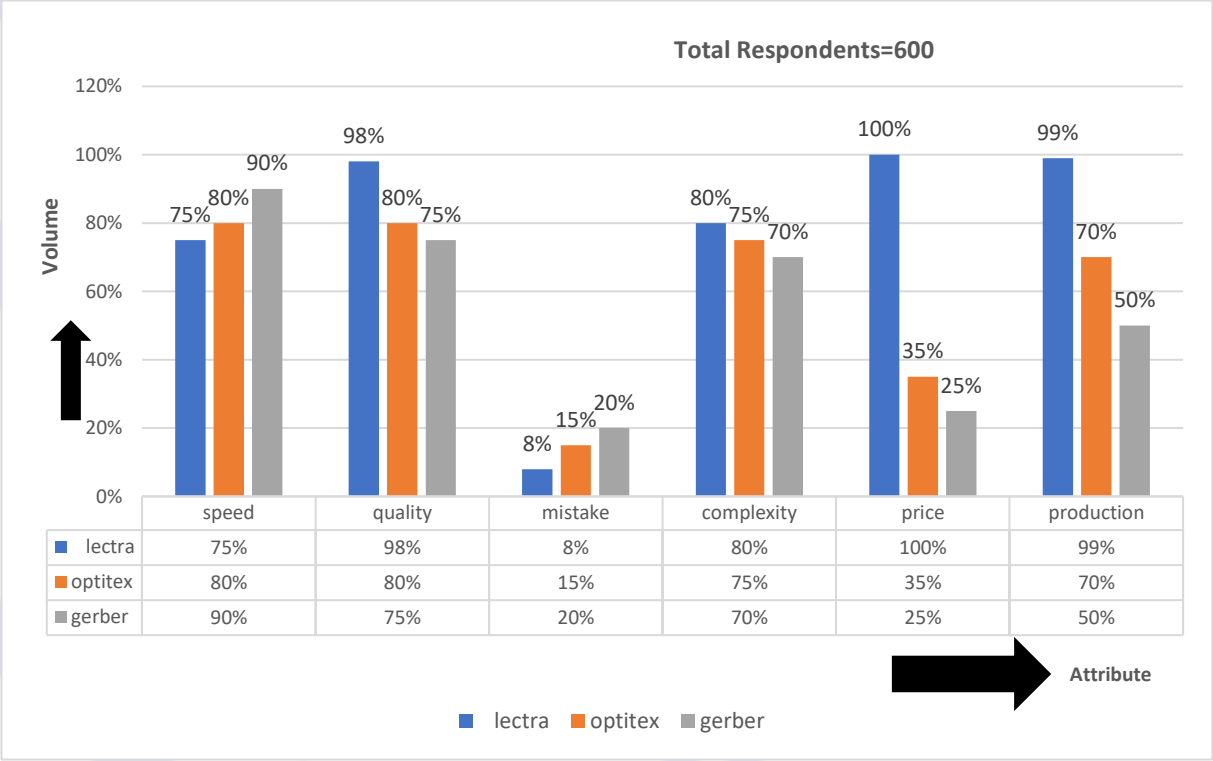


Figure 10: Different software attribute with a bar chart

Conclusion:

From the Research that was performed successfully, it can conclude the Bangladeshi industry's attitude toward CAD/CAM software and Identify Attributes that affect User behavior for particular Garments Software [4]. Overall, customers' highest priority was Production, speed, and Brand. The speed table represents that the faster software is Gerber than optics and Lectra.

However, why an industry chooses Lectra for their industry? If we look up, the quality of the design of Lectra is maximum. The possibility of mistakes in Lectra is less compared to other software.
On the other hand, the production rate is also maximum in Lectra software. However, the speed, complexity, and price are maximum, negatively impacting Lectra software. Nevertheless, the design quality, the possibility of the mistake, and production accounted for 98%, 8%, and 99%. The price of the Lectra software is maximum, accounting for 10,00000 to 15,00000 taka. However, the price of the software is the too high price. However, the other Attribute plays a positive impact. That is why the user has ignored the price of the software. Lectra software also has complex and slower than other software.
So the production and design quality rate hit the maximum, positively impacting Lectra Modaris software. The possibility of a mistake is 8% means Lectra Modaris software has a minimum possibility of mistakes. The other significant factor is that Lectra software is the leading template system, which helps a user make automatic software, the most extensive facility for making a design quickly. That is why Lectra choose a maximum number of big industries in Bangladesh. Owing to compatibility, the production of the software plays a vital role in influencing the user's tendency to buy CAD/CAM software.

For the medium industry, they focus on some attributes for their company satisfaction. So every medium-class factory intended to buy Optitex software. Optitex is software that is faster than Lectra software. Complexity is also lower than the Lectra software. The price of the software is low compared to Lectra software. That is the way the 3 attribute of the speed of the software, complexity, and price, which is accounted for 80%, 75%, and 35%, plays an essential role in choosing Optitex software. For medium-class factories, the production size is less. So they try to ignore the production size.

For a small factory, they always preferred to buy Gerber software for their satisfaction. Gerber is enough to meet their criteria, also. Where the small factory ignored 3 attributes such as quality of design, the possibility of mistakes, and production size; on the other hand, they focused on the software's speed, complexity, and price. Where speed, complexity, and price are accounted for 90%, 70%, and 25%. So it represents that the speed and price of the software are minimum. Moreover, the complexity of the software is less. So it is pretty easy to learn and use. To recapitulate, 59% of users prefer to buy Lectra Software, 25% prefer Optitex software, and 16% prefer to use Gerber software. There are 3 types of Gerber software available. Such as gamine, bock, Emma, Etc. Every Buyer wants faster and quality software that involves the proper time delivery. Buyer also focuses on attributes such as quality and brand, accounting for39.41% and 60.59% for priority, followed by 58.73% and 41.27% for second priority. However, the Buyer has no choice about the price of the software. Because of that, there is no influence of the price upon a buyer. [6]. To encapsulate considering all the properties user is intended to buy 59% Lectra, 25 % Optitex, and 16% Gerber.

ACKNOWLEDGEMENTS

This valuable thesis is the culmination of years of work that was associated with and approved by numerous people. It is fascinating that we are now all respectfully thanking each contributor for their contributions. I also want to thank infra polytechnic institute and university of global village for helping us to complete this project.

Referencias

[1]. Ali, J., Kapoor, S., & Moorthy, J. (2010, January 1). Buying behaviour of consumers for food products in an emerging economy. British Food Journal, 112(2), 109-124. doi:10.1108/00070701011018806

[2]. Au, C., & Ma, Y.-S. (2010, August 1). Garment pattern definition, development and application with associative feature approach. Computers in Industry, 61(6), 524-531. doi:https://doi.org/10.1016/j.compind.2010.03.002

[3]. Dabke, P., Cox, A., & Johnson, D. (1998). NetBuilder: an environment for integrating tools and people. Computer-Aided Design, 30(6), 465-472. doi:10.1016/S0010-4485(97)00098-5

[4]. Fayoomy, A., & Tahan, A. (2014, January 1) . Basic garment pattern design and the development for the standard figure. World Journal of Engineering, 11(2), 171-180. doi:10.1260/1708-5284.11.2.171

[5]. Gulati , S. (2017, June 29) . IMPACT OF PEER PRESSURE ON BUYING BEHAVIOUR. International Journal of Research, 5(6), 280-291. doi:10.5281/zenodo.820988

[6]. Prof.Lakshminarayana, K., & Dr.Sreenivas, D. (2018, March). A Study of Consumer Buying Behavior towards Branded Apparels in Selected cities of Karnataka. International Journal of Advanced in Management, Technology and Engineering Sciences, 8(3), 1129-1145.

[7]. Stjepanovic, Z. (1995, January 1) . Computer-aided processes in garment production: features of CAD/CAM hardware. International Journal of Clothing Science and Technology, 7(2/3), 81-88. doi:10.1108/09556229510087236.

[8]. Vinaja, R. (2012, October 1) . Research in Systems Analysis and Design: Models and Methods. Journal of Global Information Technology Management, 15(4), 89-90. doi:10.1080/1097198X.2012.10845626

[9]. Zhujun, W., Jianping, W., Xianyi, Z., Xuyuan, T., Yingmei, X., & Pascal, B. (2021, June 11). Construction of Garment Pattern Design Knowledge Base Using Sensory Analysis, Ontology and Support Vector Regression Modeling. International Journal of Computational Intelligence Systems, 14(1), 1687-1699. doi:10.2991/ijcis.d.210608.002

[10]. Md. Shazzat Hossain, Md. Hasan Ali, Md. Abdus Samad (2021), Attitude of Customers to Buy Face Masks Cloth in Bangladesh - An Observation toward Customers Psychology of Face Masks Fabric, International Journal of Textile Science 2020, 9(2): 28-34 ,DOI: 10.5923/j.textile.20200902.02

[11]. Sabeur Bettaieb ., Frédéric Noël ., Serge Tichkiewitch. (January 2004), "Interface between CAD/CAM Software and an Integrative Engineering Design Environment", Methods and Tools for Co-operative and Integrated Design, pp.315-326, DOI: 10.1007/978-94-017-2256-8_27 .

[12]. Sarita Chaudhary; Pardeep Kumar; Prashant Johri (January 2020) , "Maximizing performance of apparel manufacturing industry through CAD adoption" , International Journal of Engineering Business Management ,12(4):184797902097552 , DOI:10.1177/1847979020975528 ,LicenseCC BY 4.0

[13]. Siming Guo; Cynthia L. Istook (December 2019) ,"Evaluation of CAD Technology for Mass Customization" International Textile and Apparel Association Annual Conference , DOI:10.31274/itaa.8362 .

[14]. Abu Sadat Muhammad Sayem; William Richard Kennon (July 2010) ," 3D CAD systems for the clothing industry" , International Journal of Fashion Design Technology and Education Technology and Education(2):45-53 , DOI:10.1080/17543261003689888

[15]. Naznin Kamrun Nahar ; Summiya Sultana (January 2022) , "Process & Effective Methods of Pattern Making For the RMG (Readymade-Garment) Sector" , IOSR Journal of Research & Method in Education (IOSRJRME) Volume 7(, Issue 3 Ver. II (May - June 2017)):46-48 .

[16]. Jun Zhang ; Noriaki Innami; Kyoungok Kim; Masayuki Takatera (November 2015) , "Upper garment 3D modeling for pattern making" , International Journal of Clothing Science and Technology 27(6):852-869 , DOI:10.1108/IJCST-01-2015-0003

Sistema de reconocimiento facial para el control de accesos mediante Inteligencia Artificial

Facial recognition system for access control through Artificial Intelligence

Jean E. Manuel Reyes Campos
Universidad Nacional de Trujillo. La Libertad, Perú.
jereyesc@unitru.edu.pe
https://orcid.org/0000-0001-8635-4560

Christian Castañeda Rodríguez
Universidad Nacional de Trujillo. La Libertad, Perú.
ccastaneda@unitru.edu.pe
https://orcid.org/0000-0003-1409-7870

Luis Daniel Alva Luján
Universidad Nacional de Trujillo. La Libertad, Perú.
ldalval@unitru.edu.pe
https://orcid.org/0000-0003-1587-6366

Alberto Mendoza de los Santos
Universidad Nacional de Trujillo. La Libertad, Perú.
amendozad@unitru.edu.pe
https://orcid.org/000-0002-0469-915X

ARK: ark:/42411/s11/a78
PURL: 42411/s11/a78

RECIBIDO 22/09/2022 • ACEPTADO 05/11/2022 • PUBLICADO 30/03/2023

RESUMEN

El presente artículo tiene como objetivo principal el desarrollo de un sistema que permita el reconocimiento facial de una persona para el control de accesos mediante Inteligencia Artificial. Para el desarrollo del sistema se tuvo como algoritmo Redes Neuronales Convolucionales, el cual es un modelo de reconocimiento. Así mismo se utilizó el lenguaje de programación Python y las librerías siguientes como Numpy, Os, OpenCV e Imutils para su implementación. Los resultados obtenidos según el acierto y utilizando un dataset de 450 imágenes por individuo son de un 88% aproximadamente en cuanto la predicción por persona, concluyendo que el sistema de reconocimiento es eficaz y tiene mayor eficiencia incrementando el tamaño de datasets generados por individuos.

Palabras claves: Control de acceso, Inteligencia Artificial, Redes Neuronales Convolucionales.

ABSTRACT

The main objective of this article is the development of a system that allows the facial recognition of a person for access control through Artificial Intelligence. For the development of the system, the Convolutional Neural Networks algorithm was used, which is a recognition model. Likewise, the Python programming language and the following libraries such as Numpy, Os, OpenCV and Imutils were used for its implementation. The results obtained according to the hit and using a dataset of 4500 images are approximately 88% in terms of the prediction per person, concluding that the recognition system is effective and has greater efficiency by increasing the size of datasets generated by individuals.

Keywords: *Access Control, Artificial Intelligence, Convolutional Neural Networks.*

INTRODUCCIÓN

Actualmente la relación entre computadora y humano va reduciendo brechas, lo cual genera consigo una gran necesidad en implementación de seguridad informática. La seguridad informática es considerada un tema de gestión alineado a estándares y buenas prácticas [13], por lo cual requiere de un sistema de autenticación para restringir el acceso de los usuarios a cierta información almacenada en computadores.

Dentro de los diferentes tipos de autenticación se encuentra la autenticación biométrica que incluye la detección de una señal biométrica, la extracción de diversas características contenidas en la señal biométrica, y el uso de clasificadores para manejar las características extraídas [14]. La autenticación biométrica es aplicable a distintas características genéticas de los individuos, tales como: iris de los ojos [15], huellas dactilares [16] y el rostro [1].

Los diferentes tipos de software cuentan con una autenticación débil que puede ser vulnerada si se consiguen las credenciales necesarias, es decir que cualquier individuo puede acceder al sistema si cuenta con el usuario y contraseña correctos para ser admitido, debido a que la autenticación se realiza bajo un solo factor[7].Por otra parte se pudo identificar información relevante que debe ser tomada en cuenta para el desarrollo de un sistema de autenticación biométrico facial, elementos como el nivel de iluminación, nivel de brillo del ambiente y perspectiva de la imagen afectan a la recopilación de características de las imágenes, sobre todo si se trabaja bajo un concepto de reconocimiento facial automatizado 2D[1].

En este artículo se tratará de abordar un sistema que permita el reconocimiento facial de una persona para el control de accesos mediante Inteligencia Artificial captados por una videocámara, cubriendo las brechas o limitaciones como son el nivel de iluminación, nivel de brillo del ambiente [1], mediante un trabajo en 3D, además que vamos a realizar múltiples capturas de imagen para incrementar el reconocimiento facial [10] y reducir la repercusión del ambiente.

Materiales y métodos o Metodología computacional

Estado del arte:

Un sistema de reconocimiento facial integrado de bajo costo para el control de acceso a puertas mediante Deep Learning

El presente artículo [2] presenta cómo se utiliza un procesador Neural Compute Stick 2 junto con una placa Raspberry Pi 3 B+ para controlar una cerradura electromagnética usando técnicas de aprendizaje profundo para generar un sistema de control de acceso con capacidades de reconocimiento facial integradas desarrollado en el lenguaje de programación Python.

Automatizado basado en redes neuronales convolucionales Sistema de Asistencia mediante Reconocimiento Facial Dominio

El presente artículo [7] tiene como finalidad lograr el reconocimiento de rostros en imágenes de video, videos pregrabados o a través de una cámara en vivo para la implementación de un sistema de asistencia automático desarrollado con lenguaje de programación Python que marcará la asistencia solo a los alumnos que hayan asistido a más del 60% de las clases, dejando al resto con inasistencia. Los métodos que implementados son: Análisis de Componentes, análisis de discriminante lineal, patrón binario local y el clasificador de gabo; los cuales reducen la cantidad de cálculos, reducen la complejidad y aumentan la precisión en el reconocimiento facial.

Fundamentación Teórica

Reconocimiento Facial

Identificación de rostros con bajos niveles de sesgo mediante la aplicación de la Inteligencia artificial.[9]
El reconocimiento biométrico facial es definido como una tecnología de inteligencia artificial que implementa comparaciones automáticas de diversos rasgos faciales.[10]

Inteligencia Artificial

La inteligencia artificial (IA) es un conjunto de algoritmos (reglas que definen con precisión un conjunto de operaciones) que permiten realizar cálculos para percibir, razonar y actuar. La IA es

usada para llevar a cabo la realización de múltiples tareas, pero puede usarse para brindar mejoras a la inteligencia humana.[8]

Machine learning

Machine learning es definido como aprendizaje automático, y su uso está enfocado en el análisis masivo de datos [19]. Dentro de sus algoritmos más usados tenemos: SVM, BOSQUE ALEATORIO, Árbol de decisiones KNN y Adaboost clasificadores [20].

Deep Learning

Definido como aprendizaje profundo, ofrece una estrategia de optimización global. Dentro de sus usos tenemos: Procesamiento de información, reducción de ruido en imágenes [21], procesamiento del lenguaje natural [22], máquina de traducción [24] e ingeniería de software [25].
Además, el aprendizaje profundo es una rama derivada del aprendizaje automático (Machine learning) [23].

Redes Neuronales

Es un algoritmo inspirado en el cerebro humano diseñado para reconocer patrones en conjuntos de datos numéricos. Los datos del mundo real, por ejemplo, imagen, audio de texto, video, etc. necesita ser transformado en vectores numéricos para usar redes neuronales. Una red neuronal se compone de diferentes capas y una capa se compone de múltiples nodos.
Según el tipo de patrón que la red neuronal está tratando de aprender, a cada dato de entrada que ingresa a un nodo se le asigna cierto peso. Estos pesos determinan la importancia de los datos de entrada para producir el resultado final. Se calcula la suma ponderada de los datos de entrada y, dependiendo de algunos sesgos de umbral, se determina la salida para el nodo. La asignación de entrada a salida se realiza mediante alguna función de activación. [17]

Redes Neuronales convolucionales

Son un tipo de red neuronal que se utiliza principalmente en el campo de clasificación de imágenes, particularmente en el reconocimiento facial. Las redes neuronales convolucionales toman una imagen de entrada y modifican los pesos de la red en función de la imagen de entrada para que pueda diferenciarla de otras imágenes. Esto permite que la red aprenda e identifique las características importantes por sí misma. Por la tanto, se minimiza la necesidad de supervisión humana, reducen la necesidad de procesamiento requerido para entrenar el modelo.[7]

Seguridad de la información

Es la disciplina que, con base en políticas y normas internas y externas de la empresa, se encarga de proteger la integridad y privacidad de la información almacenada en un sistema informático, contra cualquier tipo de amenaza, minimizando los riesgos tanto físicos como físicos. lógica, a la que está expuesto.[18]

Control de Accesos

Es implementado como un método de seguridad para delimitar un conjunto de usuarios autorizados para acceder a una determinada información [2].

Herramientas y elementos

Visual Studio Code Visual Studio Code es definido como una plataforma de código abierto, un editor de código de multiplataforma que pertenece a Microsoft y proporciona todos los componentes necesarios de un IDE como: IntelliSense, depuración, control de versiones, creación de plantillas y API de extensiones que brindan muchas facilidades a los desarrolladores.[11]

Python Python es un lenguaje de programación ampliamente utilizado en las aplicaciones web, el desarrollo de software, la ciencia de datos y el machine learning (ML). Los desarrolladores utilizan Python porque es eficiente y fácil de aprender, además de que se puede ejecutar en muchas plataformas diferentes.[6]

Las librerías utilizadas en este lenguaje de programación para este proyecto fueron las siguientes:

Numpy NumPy es un módulo de Python. El nombre es un acrónimo de Python Numérico. Es una librería que consiste en objetos de matrices multidimensionales y una colección de rutinas para procesar esas matrices. Es un módulo de extensión para Python, escrito en su mayor parte en C. Esto asegura que las funciones y funcionalidades matemáticas y numéricas recompiladas de NumPy garantizan una gran velocidad de ejecución.[3]

Os Este módulo provee una manera versátil de usar funcionalidades dependientes del sistema operativo, algunas como leer, escribir, manipular archivos y demás.[4]

OpenCV Es una biblioteca de código abierto que incluye varios cientos de algoritmos de visión artificial.[5]

Imutils Es un paquete de OpenCV que evita la pérdida de fotogramas en el procesamiento de imágenes mediante el uso de múltiples subprocesos que llevan a cabo la lectura y procesamiento de las imágenes en forma simultánea.[12]

Uso del Sistema de detección

Creación de los Datasets

```python
import cv2, os, imutils

personCode = 'Jean'
personPath = './Data/' + personCode

if not os.path.exists(personPath):
    os.makedirs(personPath)
    print('Directory created: ', personPath)

cap = cv2.VideoCapture(0, cv2.CAP_DSHOW)
faceClassif = cv2.CascadeClassifier(
    cv2.data.haarcascades + 'haarcascade_frontalface_default.xml')
count = 0

while True:
    ret, frame = cap.read()
    if ret == False: break

    frame = imutils.resize(frame, width = 720)
    gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
    auxFrame = frame.copy()
    faces = faceClassif.detectMultiScale(gray, 1.3, 5)

    for (x,y,h,w) in faces:
        cv2.rectangle(frame, (x,y),(x+w, y+h), (0,255,0), 2)
        rostro = auxFrame[y:y+h, x:x+w]
        rostro = cv2.resize(rostro, (720, 720),
            interpolation=cv2.INTER_CUBIC)
        cv2.imwrite(
            personPath + '/'+personCode+'_{}.jpg'.format(count),
            rostro)
        count+=1

    cv2.imshow('Camera', frame)
    if (cv2.waitKey(1) == 27) or count==450: break

cap.release()
cv2.destroyAllWindows()
```

Figura 1. Script para la creación de un dataset personalizado por persona

Se usó un dataset creado por un script escrito en python que podemos apreciar en la Figura 1, el cual se encarga de primero trabajar con el frame que es capturado por la cámara para luego

evaluarlo por "haarcascade" y detectar la presencia de un rostro, de existir la presencia de algún rostro, es extraído, redimensionado a 720x720 y almacenado como un archivo de formato JPG en una carpeta con el nombre de la persona que estamos registrando, el cual definimos en la línea 3 como *personCode*, se mantendrá en un bucle hasta alcanzar la cantidad de 450 imágenes registradas para posteriormente finalizar el proceso.

Entrenamiento

Se ejecuta un script en python el cual se encarga de usar el algoritmo que será usado para el reconocimiento facial es el conocido como "Local Binary Pattern Histogram", el cual nos ayuda con el reconocimiento de una persona Figura 2.

```python
import cv2, numpy as np, os

dataPath = './Data/'

peopleList = os.listdir(dataPath)

print("Lista de usuarios registrados", peopleList)

labels = []
facesData = []
label = 0

for code in peopleList:
    personPath = dataPath+code
    print("Leyendo Imagenes de :", code)

    for fileName in os.listdir(personPath):
        labels.append(label)

        facesData.append(
            cv2.imread(personPath+'/'+fileName, 0))

    label+=1

cv2.destroyAllWindows()

faceRecognizer = cv2.face.LBPHFaceRecognizer_create()
#Entrenamiento
print("Entrenando")
faceRecognizer.train(facesData, np.array(labels))
print("Modelo Entrenado")
#GuardarModelo
faceRecognizer.write('Modelo.xml')
print("Modelo guardado")
```

Figura 2. Script para el entrenamiento de la red con los datos de los individuos registrados

Lo que haremos es brindarle la data (imágenes generadas anteriormente) junto a sus respectivos labels que no son más que valores asignados a cada persona registrada, que van desde el 0 hasta donde sea necesario, todo esto con el fin de que el algoritmo pueda ser entrenado y finalmente nos genere un "Modelo" en formato xml que contenga todos los valores necesarios para que sea capaz de reconocer a las personas que protagonizan las imágenes entregadas anteriormente.

El modelo generado contiene los valores de los pesos y sesgos necesarios en la red para que ésta sea capaz de reconocer a las personas que han sido registradas en el entrenamiento, no es necesario que comprendamos cada valor que está incluido en este modelo, simplemente con comprender que en su conjunto está personalizado para el reconocimiento facial de aquellos individuos que registramos con su respectivo dataset. Figura 3.

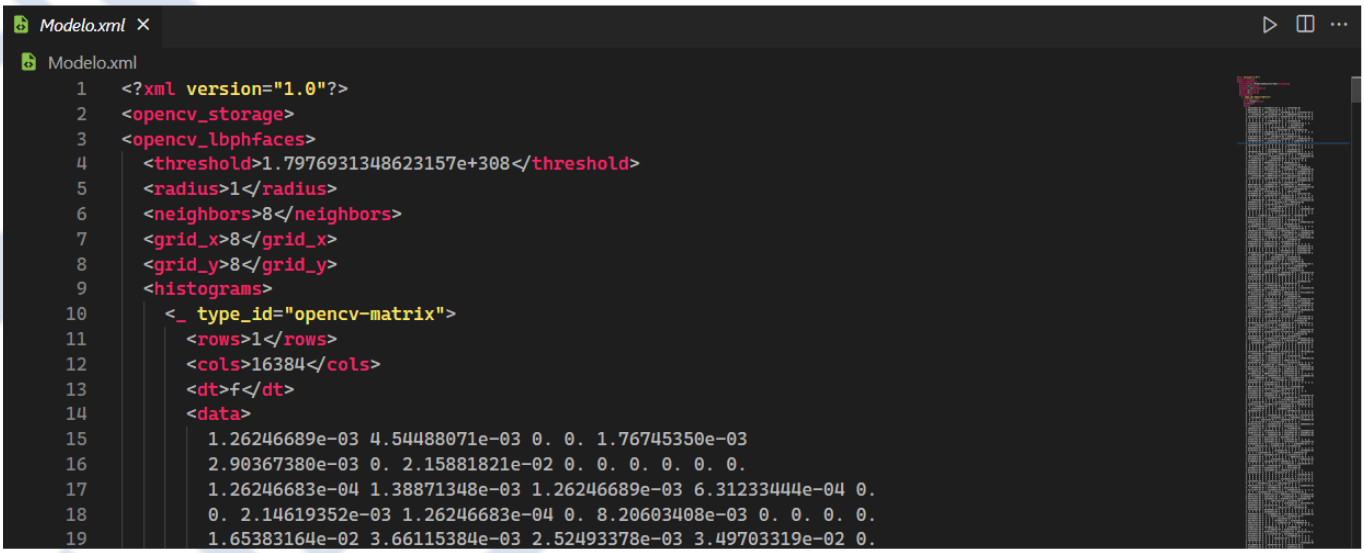


Figura 3. Modelo que contiene los valores numéricos para que la red sea capaz de reconocer a los individuos registrados

Predicción

Una vez entrenado, el modelo está listo para realizar predicciones, para ello desarrollamos un script de python el cual se encarga de primero trabajar con el frame que es capturado por la cámara para luego evaluarlo por "haarcascade" y detectar la presencia de un rostro, de existir la presencia de algún rostro, es extraído, redimensionado a 720x720, y este último será el frame que ingresará al modelo para poder predecir y que este nos devuelva su predicción. Figura 4.

```python
import cv2
import os

dataPath = './Data/'
Nombres = os.listdir(dataPath)

faceRecognizer = cv2.face.LBPHFaceRecognizer_create()
faceRecognizer.read('Modelo.xml')

cap = cv2.VideoCapture(0, cv2.CAP_DSHOW)

faceClassif = cv2.CascadeClassifier(cv2.data.haarcascades + 'haarcascade_frontalface_default.xml')

#colors BGR
blue, green, red, black = (255, 0, 0), (0,255,0), (0,0,255), (0,0,0)

verificado, tiempo, maxTime, auth, color, acumulado = 0, 0, 200, "", black, 0

while True:
    ret, frame = cap.read()
    if ret == False: break
    gray = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
    auxFrame = gray.copy()
    faces = faceClassif.detectMultiScale(gray, 1.3, 5)

    for (x,y,h,w) in faces:
        face = auxFrame[y:y+h, x:x+w]
        face = cv2.resize(face, (720,720), interpolation=cv2.INTER_CUBIC)
        result = faceRecognizer.predict(face)

        if(tiempo<=maxTime):
            if(verificado/maxTime>=0.8):
                auth, color="Reconocido", green
            else:
                auth="No reconocido"

            if result[1]<=20:
                cv2.putText(frame, '{}'.format(result[1]), (x,y-5), 1, 1.3, blue, 1, cv2.LINE_AA)
                cv2.putText(frame, '{}'.format(Nombres[result[0]]), (x,y-25), 2, 1.1, blue, 1, cv2.LINE_AA)
                cv2.rectangle(frame, (x,y), (x+w,y+h), blue, 2)
                acumulado+=result[1]
                verificado+=1
            else:
                cv2.putText(frame, '{}'.format(result[1]), (x,y-5), 1, 1.3, red, 1, cv2.LINE_AA)
                cv2.putText(frame, 'Intruso!', (x,y-20), 2, 0.8, red, 1, cv2.LINE_AA)
                cv2.rectangle(frame, (x,y), (x+w,y+h), red, 2)
        else:
            cv2.putText(frame, '{}'.format(result[1]), (x,y-5), 1, 1.3, color, 1, cv2.LINE_AA)
            cv2.putText(frame, auth, (x,y-25), 2, 1.1, color, 1, cv2.LINE_AA)
            cv2.rectangle(frame, (x,y), (x+w,y+h), color, 2)
        tiempo+=1

    cv2.imshow('Reconociendo', frame)
    k = cv2.waitKey(1)
    if k==27 or tiempo==maxTime+60: break
print("=============================================")
print("Margen de error promedio de predcción")
print("=============================================")
print(acumulado/verificado)
cap.release()
cv2.destroyAllWindows()
```

Figura 4. Script para la predicción usando el modelo entrenado

La predicción del modelo nos devuelve 2 valores, primero el número del label con el cual tiene mejor coincidencia y en segundo lugar, la distancia o que tan alejado se encuentra el frame actual de los registrado para el entrenamiento que pertenecen a dicho label, el nivel de confiabilidad puede ser personalizado dependiendo del nivel de confianza que se desea en el momento de la autenticación, en la línea 37 se estipula la lejanía máxima a la que se puede encontrar la imagen de entrada de la predicción para ser aceptado como reconocido, mientras que en la línea 17, el valor de maxTime indica la cantidad de frames máximos que se evaluarán para la predicción, por su parte en la línea 32 se establece que el mínimo de frames en los que debe ser reconocido el individuo para considerarlo como tal, reconocido, 0.8(80%), luego de que se cumpla el maxTime, se pinta en pantalla el resultado, sea positivo o negativo durante unos segundos para luego finalizar su ejecución.

En la Figura 5 podemos visualizar la apariencia de la ejecución del script de reconocimiento tanto cuando reconoce cuando al individuo, como cuando no.

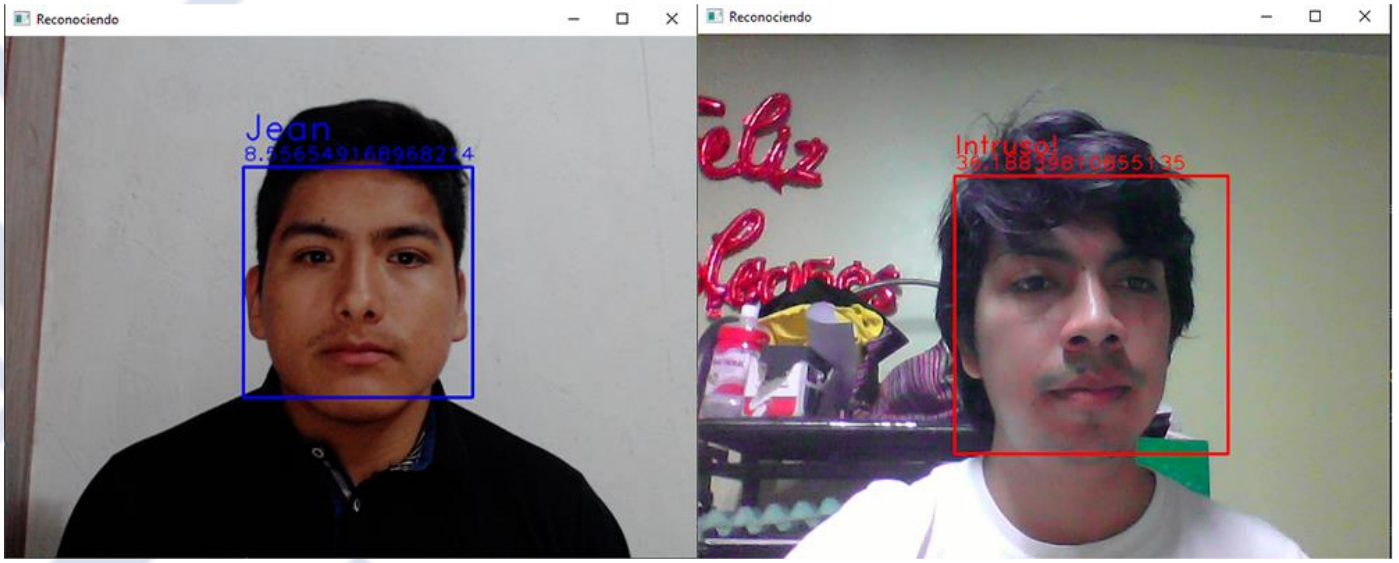


Figura 5. Ejecución del script de reconocimiento

En la Figura 6 podemos observar cómo se muestran los resultados finales de autenticación, tanto positiva como negativa.

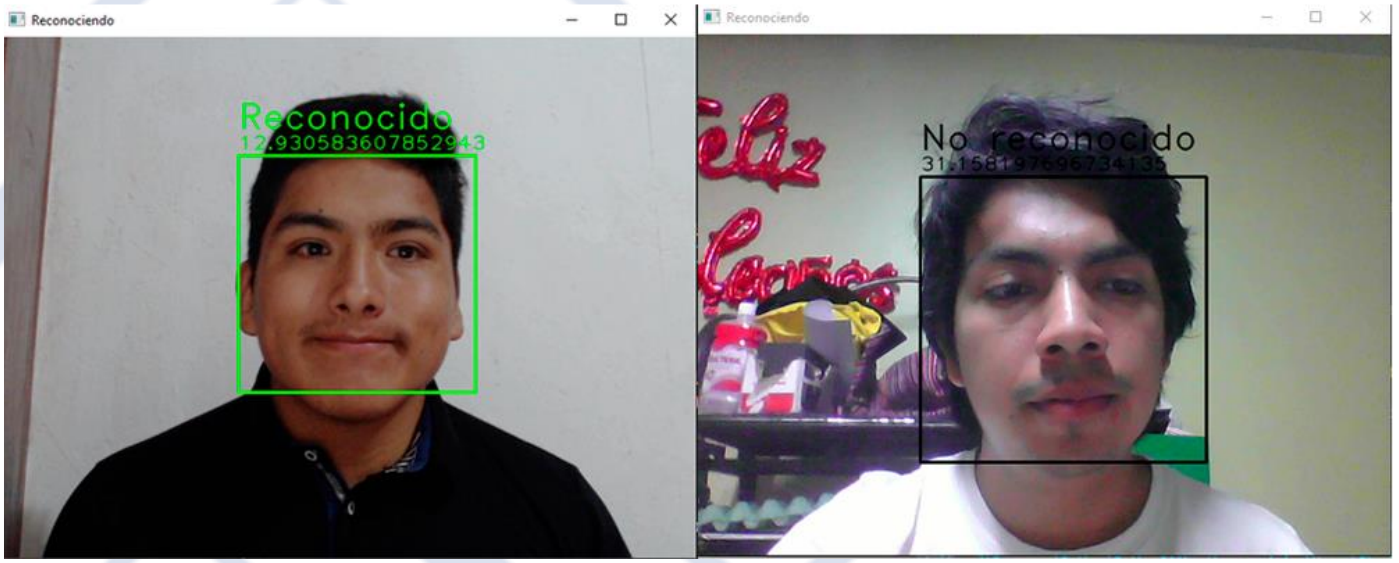


Figura 6. Proyección de resultados positivos y negativos de autenticación

Resultados y discusión

Los parámetros que usa el sistema pueden ser perfectamente ajustados de acuerdo a lo que uno necesite, en este caso se usaron 450 imágenes por persona registrada, pero a mayor este número, mayor será la precisión del sistema.
Respecto a la precisión, al realizar pruebas se obtiene como segundo valor devuelto, que tan alejado está el frame de la predicción hecha por el sistema, al sacar una media de estos valores devueltos, se obtuvo como resultado el valor de 12.070609534085545, esto se traduce en un porcentaje de acierto de un 88% aproximadamente en cuanto a las predicciones por persona se refiere, Figura 8.

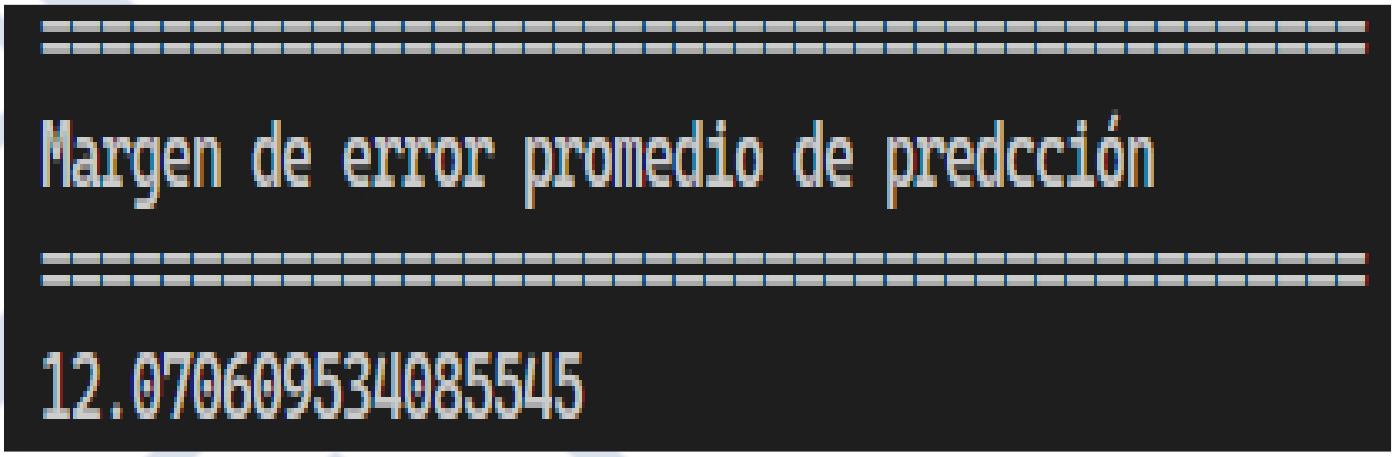


Figura 8. Margen de error promedio de predicción

Mientras que el porcentaje de aciertos totales dentro de una prueba unitaria fue considerado como un valor mínimo a alcanzar para validar la autenticación de la persona, en este caso como se indicó previamente, es del 80%, al igual el tamaño unitario de los datasets, se puede ajustar manualmente.

Conclusiones

El sistema de reconocimiento se puede ajustar para una mayor eficiencia siempre y cuando se incremente el tamaño de los datasets generados por individuo para un posterior más robusto entrenamiento.

El lenguaje de programación Python brinda muchas facilidades para la programación de este tipo, refiriéndonos a Visión Artificial como a Inteligencia artificial, con la gran cantidad de librerías que ofrecen métodos para una mejor experiencia en estos campos.

Algunos rasgos faciales pueden ser compartidos por una o más personas, para ello es mejor ampliar el tamaño de los datasets y así poder ser más específicos al entrenar a la red neuronal. Los datasets de imágenes solo son usados para crear el modelo, una vez esto esté hecho ya no es necesario mantenerlos, puede deshacerse de ellos o almacenarlos, de aquí en adelante para las predicciones sólo es necesario usar los valores que están dentro del modelo generado.

Referencias

[1] "Facial Expression Recognition Using Machine Learning Techniques", *International Journal of Advance Engineering and Research Development*, vol. 1, n.º 06, junio de 2020. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.21090/ijaerd.010633

[2] R. F. Rahmat, E. N. Zai, I. Fawwaz y I. Aulia, "Facial Recognition-Based Automatic Door Access System Using Extreme Learning Machine", *IOP Conference Series: Materials Science and Engineering*, vol. 851, p. 012065, mayo de 2020. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1088/1757-899x/851/1/012065

[3] "Librería NumPy - Aprende IA". Aprende IA. https://aprendeia.com/libreria-de-python-numpy-machine-learning/ (accedido el 17 de noviembre de 2022).

[4] "os - Interfaces miscelíneas del sistema operativo — documentación de Python - 3.10.8". 3.11.0 Documentation. https://docs.python.org/es/3.10/library/os.html (accedido el 17 de noviembre de 2022).

[5] "OpenCV: OpenCV modules". OpenCV documentation index. https://docs.opencv.org/4.x/ (accedido el 17 de noviembre de 2022).

[6] "¿Qué es Python? | Guía de Python para principiantes de la nube | AWS". Amazon Web Services, Inc. https://aws.amazon.com/es/what-is/python/ (accedido el 17 de noviembre de 2022).

[7] Anuja Jadhav, Yash Joshi y Vishakha Kalambe, "Face Based Attendance System Using Convolutional Neural Network", *International Journal of Advanced Research in Science, Communication and Technology*, pp. 51–54, febrero de 2022. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.48175/ijarsct-2506

[8] O. Niel y P. Bastard, "Artificial Intelligence in Nephrology: Core Concepts, Clinical Applications, and Perspectives", *American Journal of Kidney Diseases*, vol. 74, n.º 6, pp. 803–810, diciembre de 2019. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1053/j.ajkd.2019.05.020

[9] T. Walsh, "The troubling future for facial recognition software", *Communications of the ACM*, vol. 65, n.º 3, pp. 35–36, marzo de 2022. Accedido el 15 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1145/3474096

[10] M. Smith y S. Miller, "The ethical application of biometric facial recognition technology", *AI & SOCIETY*, abril de 2021. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1007/s00146-021-01199-9

[11] S. Latifi, Ed., *17th International Conference on Information Technology–New Generations (ITNG 2020)*. Cham: Springer International Publishing, 2020. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1007/978-3-030-43020-7

[12] S. K. Shammi, S. Sultana, M. S. Islam y A. Chakrabarty, "Low Latency Image Processing of Transportation System Using Parallel Processing co-incident Multithreading (PPcM)", en *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Kitakyushu, Japan, 25–29 de junio de 2018. IEEE, 2018. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1109/iciev.2018.8640957

[13] M. Nicho, S. Khan y M. S. M. K. Rahman, "Managing Information Security Risk Using Integrated Governance Risk and Compliance", en 2017 International Conference on Computer and Applications (ICCA), Doha, United Arab Emirates, 6–7 de septiembre de 2017. IEEE, 2017. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1109/comapp.2017.8079741

[14] S. Rasnayaka, S. Saha y T. Sim, "Making the most of what you have! Profiling biometric authentication on mobile devices", en *2019 International Conference on Biometrics (ICB)*, Crete, Greece, 4–7 de junio de 2019. IEEE, 2019. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1109/icb45273.2019.8987402

[15] I. Sluganovic, M. Roeschlin, K. B. Rasmussen y I. Martinovic, "Analysis of Reflexive Eye Movements for Fast Replay-Resistant Biometric Authentication", *ACM Transactions on Privacy and Security*, vol. 22, n.º 1, pp. 1–30, enero de 2019. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1145/3281745

[16] L. Monastyrskii, V. Lozynskii, Y. Boyko y B. Sokolovskii, "Fingerprint recognition in inexpensive biometric system", *Electronics and Information Technologies*, vol. 9, 2018. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.30970/eli.9.120

Implementación de controles de acceso para un sistema web de gestión de prácticas profesionales

Implementation of access controls for a professional practice management system

Edisson Alejandro Galvez Mori
Universidad Nacional de Trujillo. La Libertad, Perú.
egalvez@unitru.edu.pe
https://orcid.org/0000-0003-3280-7661

Emmanuel Robert Torres Correa
Universidad Nacional de Trujillo. La Libertad, Perú.
rtorresc@unitru.edu.pe
https://orcid.org/0000-0002-7729-9026

Jorge Valdivia Valderrama
Universidad Nacional de Trujillo. La Libertad, Perú.
jvaldiviav@unitru.edu.pe
https://orcid.org/0000-0003-3280-7661

Alberto Mendoza de los Santos
Universidad Nacional de Trujillo. La Libertad, Perú.
amendozad@unitru.edu.pe
https://orcid.org/0000-0002-0469-915X

ARK: ark:/42411/s11/a80
PURL: 42411/s11/a80

RECIBIDO 12/10/2022 • ACEPTADO 03/12/2022 • PUBLICADO 30/03/2023

RESUMEN

El presente artículo presenta un sistema de información web que facilitará la gestión de las prácticas profesionales en la Universidad Nacional de Trujillo. Para ello, se aplicó una metodología ágil SCRUM debido a que es adaptable a los constantes cambios en el proceso de desarrollo del software y los controles de acceso que proporciona la normativa ISO 27001:2013 con el objetivo de garantizar la integridad, autenticidad y disponibilidad de los activos de información que posee el sistema. Los resultados obtenidos apoyan la implementación de la autenticación de doble factor y la firma digital como controles esenciales para la gestión de la información de acuerdo con la confiabilidad calculada. Se concluye que los controles de acceso proporcionan seguridad en la información manejada en el sistema de acuerdo a las pruebas de confianza.

Palabras claves: Sistema web de información, prácticas profesionales, ISO/IEC-27001, SCRUM, control de acceso.

ABSTRACT

This article presents a web information system that will facilitate the management of professional practices at the National University of Trujillo. For this, an agile SCRUM methodology was applied because it is adaptable to the constant changes in the software development process and the access controls provided by the ISO 27001:2013 standard with the aim of guaranteeing the integrity, authenticity and availability of data. the information assets that the system possesses. The results obtained support the implementation of two-factor authentication and digital signature as essential controls for information management according to the calculated reliability. It is concluded that the access controls provide security in the information handled in the system according to the trust tests.

Keywords: *Web information system, professional practices, ISO/IEC-27001, SCRUM, access control.*

INTRODUCCIÓN

Las prácticas preprofesionales corresponden a la primera aproximación al mercado laboral y a la vida profesional que tienen los estudiantes de la Facultad de Ingeniería, cuyo propósito es, según la Ley 28518: Ley sobre modalidades formativas laborales, permitir a la persona en formación durante su condición de estudiante aplicar sus conocimientos teóricos y prácticos mediante el desempeño de tareas programadas de capacitación y formación profesional en una situación real de trabajo [4]. En la Universidad Nacional de Trujillo se requiere de un sistema que realice la gestión de las prácticas profesionales de las distintas facultados. Además, se necesita que el sistema actúe como una vía de comunicación entre los estudiantes y las empresas que requieran de la fuerza laboral de los estudiantes a través de las convocatorias de prácticas preprofesionales. Esto porque la universidad, como institución responsable de administrar la educación superior a la sociedad emergente del conocimiento, está obligada a reinventarse para evolucionar al ritmo de los tiempos [2]. En lo que respecta la seguridad de la información se aprecia una vital importancia en los procesos que comprenden la cadena de valor de los negocios y empresas, y es por esto que los riesgos en los SI están cada vez más presentes, afectando a la integridad y protección de los mismos, resultando en un impacto negativo [1]. Algunas investigaciones sugieren que, además de las implementaciones de SGSI a los sistemas de información siguiendo el estándar ISO-27001, se debe considerar una mejora continua de los servicios ofrecidos, es decir, dar el respectivo seguimiento para garantizar de esa forma una comunicación eficaz que contribuyan a los objetivos de seguridad planeados [2, 3].

Por lo tanto, este artículo busca elaborar un prototipo con los respectivos controles de acceso de un sistema de gestión de prácticas profesionales, específicamente autenticaciones desde doble factor y firma digital para los certificados necesarios en los procesos de evaluación. Esto permitirá

la integridad, confidencialidad y disponible, además de ayudar en la agilidad de todos los procesos implicados. También se busca analizar la confiabilidad de dichos controles de acceso para analizar la factibilidad de su respectiva implementación. Esta investigación se justifica en la necesidad de implementar el sistema mencionado, puesto que no existe uno actualmente y para las demandas actuales, son necesarias de implementar para evitar las vulneraciones de la información que se puedan dar en los sistemas ya existentes.

Métodos y Metodología computacional

Descripción del proyecto

La función principal del sistema de prácticas profesionales es optimizar el proceso de gestión de prácticas profesionales para las diferentes escuelas profesionales de la universidad. Facilitando el registro, seguimiento y evaluación de las prácticas profesionales realizadas por los estudiantes. Además de reducir tiempo en realizar los trámites, disminuir los costos de impresión de documentos, ahorrar espacio y almacenar documentos en la base datos sin riesgo a que se dañen [13]. En la Figura 1 se visualiza el diagrama general del sistema que representa la forma en que interactúan los usuarios con el sistema.

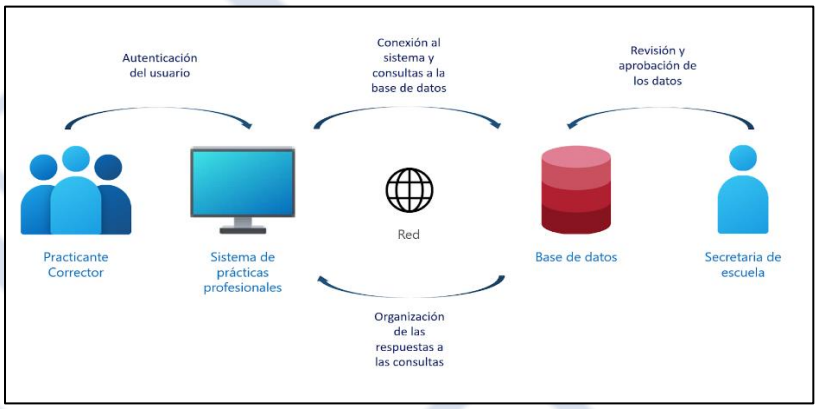


Figura 1. Diagrama general del sistema.

Modelo del proyecto

En la Figura 2 se presenta el modelo planteado para el sistema de prácticas profesionales. Dicho modelo se basa en 3 módulos: gestión de usuarios, relacionado a la administración de cuentas de los usuarios y sus roles; gestión de actividades, relacionado al control de las tareas que deben

cumplir los usuarios en el sistema como la entrega de los documentos en determinados plazos de tiempo; y la gestión de documentos, relacionado a la verificación y almacenamiento de la documentación generada por las actividades.

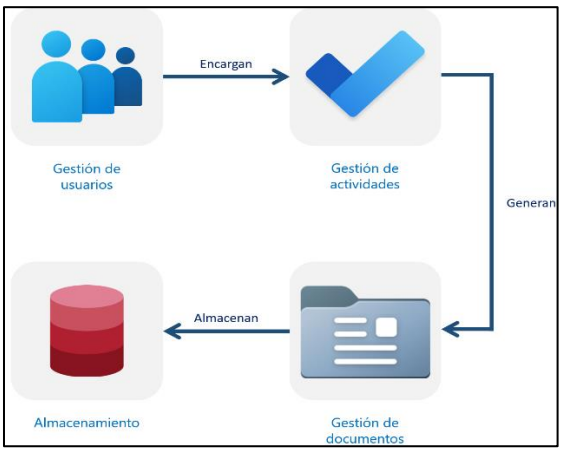


Figura 2. Modelo planteado para el sistema.

Diseño de base de datos

Para el diseño de la base de datos, las tablas se organizaron en dos categorías: tablas de cuentas de usuarios y tablas de prácticas preprofesionales. En la Figura 1, se muestra el diseño lógico de la base de datos con los campos y relaciones.

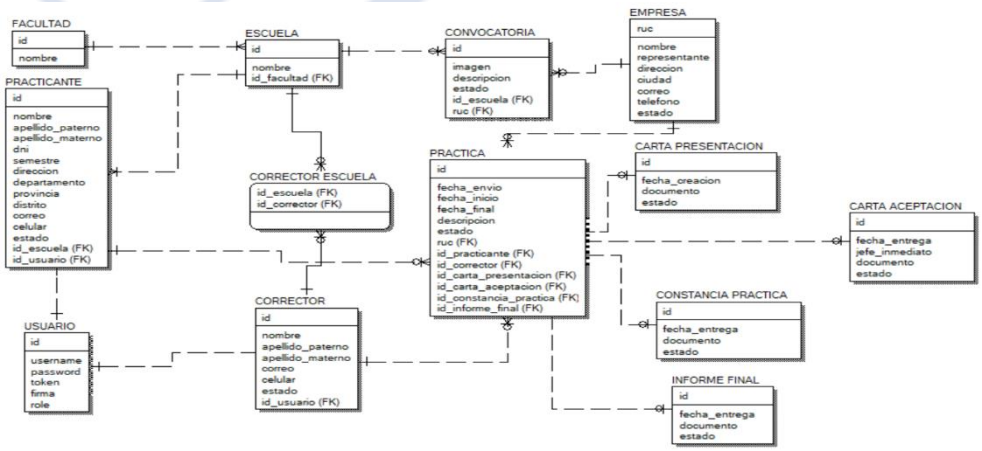


Figura 3. Diseño lógico de base de datos para el sistema.

Arquitectura de solución

El software a utilizar para el desarrollo de la aplicación web será gratuito y de código abierto con el fin de disminuir costos. El presente proyecto se realizó bajo el framework de aplicaciones web Django escrito en el lenguaje de programación Python y con el uso del motor de base de datos MySQL. Se empleó la metodología de desarrollo de software SCRUM y se realizaron 6 Sprint.

Desarrollo Sprint 1 - Funcionalidades base

El propósito del Sprint 1 es comenzar con las funcionalidades básicas de la aplicación que son importantes para el primer caudal dentro de lo planificado. Se revisan temas como la seguridad de la cuenta del usuario y la actualización de datos. En el sistema existen dos tipos de usuarios: el practicante (el estudiante) y el corrector (el docente). En la figura 4, se muestra la interfaz de inicio de la aplicación web donde se ingresan las credenciales del usuario. En la figura 5, se muestra la vista de los datos de la cuenta del usuario con la posibilidad de actualizar algunos datos.

Figura 4. Diseño de interfaz entregado en Sprint 1 en interfaz de inicio de sesión.

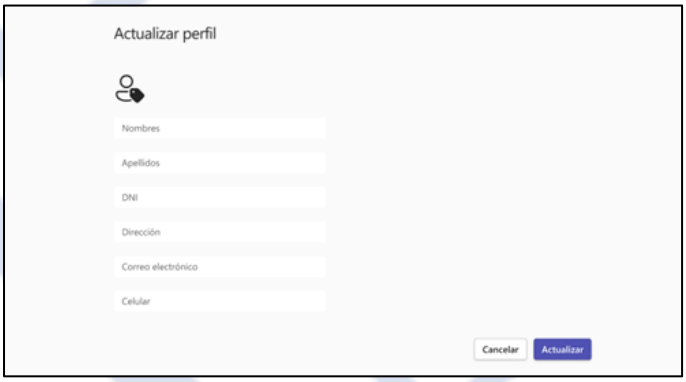


Figura 5. Diseño de interfaz entregado en Sprint 1 en interfaz de datos de cuenta de usuario.

Desarrollo Sprint 2 - Funcionalidades base 2

Con Sprint 2 se pretende implementar funcionalidades que complementen el proceso de preinscripción a la práctica preprofesional. En la figura 6, se muestra la interfaz de inscripción donde se mostrará la lista detallada de las inscripciones hechas por el practicante. En la figura 7, se muestra el formulario para el registro de una carta de presentación donde se ingresarán datos de la organización donde el practicante ofrecerá sus servicios, una vez registrado hay un tiempo límite de 7 días para que el practicante registre la carta de aceptación por parte de la empresa, en caso contrario la carta de presentación es anulada. En la figura 8, se muestra el formulario para el registro de la carta de aceptación por parte de la empresa, detallando la fecha de inicio y fin de la práctica, así como los datos del jefe encargado de monitorear las actividades del practicante.

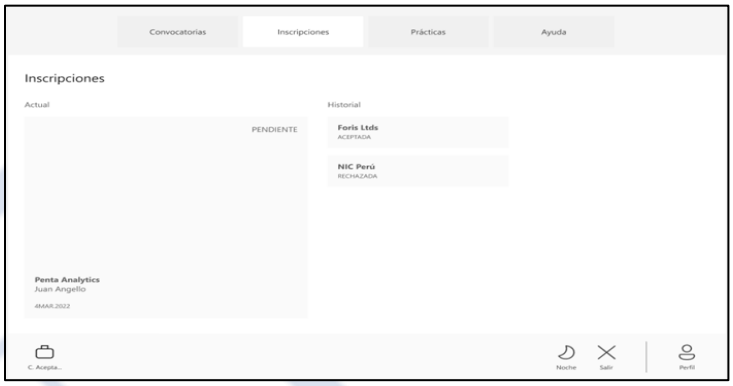


Figura 6. Diseño de interfaz entregado en Sprint 2 en interfaz de inscripción para el practicante.

Figura 7. Diseño de interfaz entregado en Sprint 2 en interfaz de nueva carta de presentación.

Figura 8. Diseño de interfaz entregado en Sprint 2 en interfaz de registro de carta de aceptación.

Desarrollo Sprint 3 - Funcionalidades base 3

Con Sprint 3 se pretende implementar funcionalidades que complementen el proceso de seguimiento de prácticas preprofesionales. En la figura 9, se muestra la interfaz de práctica donde se mostrará la lista detallada de las prácticas pendientes o terminadas del practicante. En la figura 10, se muestra el formulario para subir entregas tales como la constancia de prácticas (documento que detalla las actividades que realizó el practicante en la organización) y el informe final (documento que evidenciará el progreso del practicante durante la práctica).

Figura 9. Diseño de interfaz entregado en Sprint 3 en interfaz de práctica para el practicante.

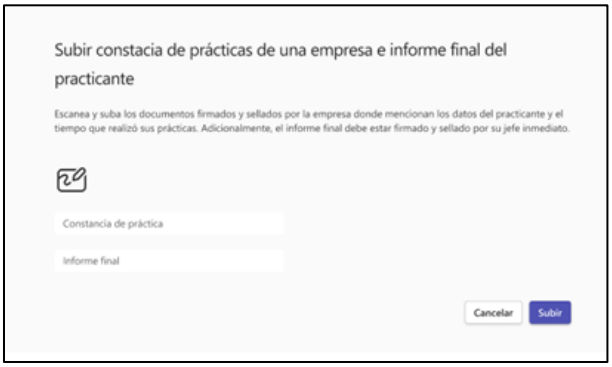


Figura 10. Diseño de interfaz entregado en Sprint 3 en interfaz de entregas de constancia de prácticas e informe final.

Desarrollo Sprint 4 - Funcionalidades base 4

Con Sprint 4 se pretende implementar funcionalidades que complementen el proceso de publicación de convocatorias de prácticas preprofesionales. En la figura 11, se muestra la interfaz de convocatorias donde se visualizará los anuncios por parte de las empresas solicitando practicantes.

Figura 11. Diseño de interfaz entregado en Sprint 4 en interfaz de convocatoria.

Desarrollo Sprint 5 – Autenticación de doble factor (2AF)

Con Sprint 5 se plantea implementar autenticación de doble factor que proporcione seguridad al proceso de inicio de sesión al sistema. Después que el usuario ingrese sus credenciales, el sistema emitirá un correo electrónico con un código aleatorio al buzón del usuario, el usuario deberá

introducir el código al sistema para completar la autenticación en un tiempo determinado, en caso contrario el sistema volverá a pedir credenciales al usuario. En la figura 12 se muestra el formulario de ingreso del código de verificación enviado previamente por el correo electrónico del usuario. En la figura 13 se muestra un ejemplo del correo que contiene el código que envía el sistema a los usuarios.

Figura 12. Diseño de interfaz entregado en Sprint 5 en interfaz de ingreso de código de verificación por 2AF.

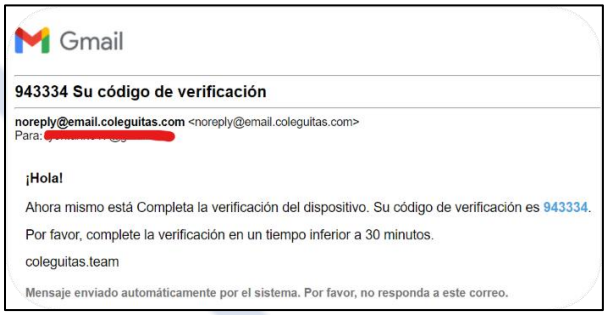


Figura 13: Formato entregado en Sprint 5 en formato de correo electrónico generado por el sistema que contiene el código de verificación de autenticidad.

Desarrollo Sprint 6 – Firma digital

Con Sprint 6 se plantea implementar firma digital para que asocie la identidad del usuario a cada uno de los documentos, además de asegurar la integridad de los documentos generados o registrados. Para ello se registrará las firmas de los usuarios, en la Figura 14 se muestra el formulario de registro, y se guardará en la base de datos, también se añadirán campos donde se realizará la firma en los formularios que impliquen entrega de documentos. El sistema pedirá previamente la contraseña del usuario para realizar la firma, tal como se muestra en la Figura 15, y comprobará que la firma sea similar a la firma registrada en la cuenta del usuario, en caso contrario se anulará el formulario.

Figura 14. Diseño de interfaz entregado en Sprint 6 en interfaz de registro de firma digital a la cuenta.

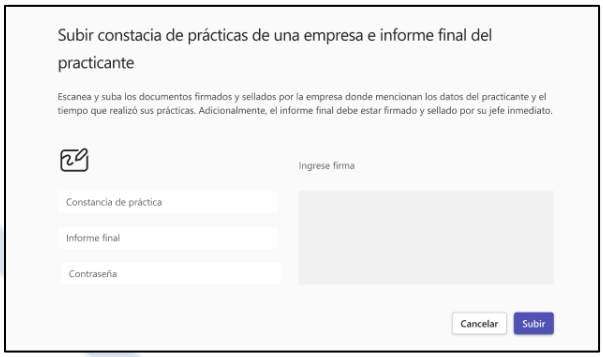


Figura 15. Diseño de interfaz entregado en Sprint 6 en interfaz de formulario de entrega de documentos implementando firma digital.

Fundamentación teórica

Prácticas profesionales

Corresponden a la primera aproximación al mercado laboral y a la vida profesional que tienen los estudiantes universitarios, cuyo propósito es, según la Ley 28518: Ley sobre modalidades formativas laborales, permitir aplicar sus conocimientos teóricos y prácticos mediante el desempeño de tareas programadas de capacitación y formación profesional en una situación real de trabajo [4].

Sistema de información

Conjunto de elementos organizados y relacionados que permite, según las políticas de la organización donde opera, almacenar, analizar y utilizar datos a favor para las actividades que

se realizan dentro de la empresa. Esto ayuda a la toma de decisiones dentro de la organización con el fin de mejorar sus procesos y satisfacer las necesidades de los clientes. Actualmente, son alojadas en servidores web para obtener alcance a grandes masas de usuarios mediante internet [5].

Control de acceso

El control de acceso es el proceso de decidir si el usuario tiene permiso para ejecutar algo o no. También llamado autorización, se refiere a la gestión del acceso a los recursos protegidos y al proceso de determinar si un usuario está autorizado a acceder a un recurso particular [7].

Metodología de desarrollo de software SCRUM

Metodología no ágil y flexible para el desarrollo de software, cuyo principal objetivo es maximizar el retorno de la inversión para la organización. Se basa en construir primero la funcionalidad de mayor valor para los usuarios y en los principios de inspección continua, adaptación, autogestión e innovación [6].

Sistema de gestión de seguridad de información

Un SGSI es un enfoque sistemático para establecer, implementar, operar, monitorear, revisar, mantener y mejorar la seguridad de la información de una organización para lograr sus objetivos. Consta de políticas, procedimientos, directrices y recursos y actividades asociados, gestionados colectivamente por una organización, con el fin de proteger sus activos de información [7].

Norma ISO 270001:2013

La ISO 27001:2013 es la norma internacional que proporciona un marco de trabajo para los sistemas de gestión de seguridad de la información (SGSI) con el fin de proporcionar confidencialidad, integridad y disponibilidad continuada de la información, así como cumplimiento legal [12]. Es esencial para resguardar los activos de información de los empleados y clientes, y la reputación de la organización.

Resultados y discusión

En este trabajo de investigación se trata de explicar la implementación de los controles de acceso definidos por la normativa ISO 27001:2013. Aquellos controles mencionados son los pilares de la seguridad de los sistemas de información, garantizando que solo los usuarios autorizados accedan a la información además de saber si realmente son quien dicen ser [8].

Un ejemplo clásico es el uso de contraseñas para realizar la autenticación de usuario, sin embargo, con el crecimiento de los casos de ataques cibernéticos y de la necesidad de preservar la integridad de la información se ha optado por usar la autenticación de doble factor con el fin de proteger el sistema de información.

La autenticación de doble factor (2AF) consiste en una capa adicional de seguridad para las cuentas de usuario con la finalidad de garantizar que el usuario sea el único quien pueda acceder con su cuenta al sistema, aunque alguien más conozca su contraseña [9]. Es decir, se añade un paso adicional para iniciar sesión en el sistema que puede ser por una llave de seguridad física, por una aplicación (como Google Authenticator o Microsoft Authenticator) o por un código de verificación de un solo uso [14]. En el sistema de prácticas profesionales se aplica 2AF por código de verificación, implementado en Sprint 5, que se envía al correo electrónico registrado en la cuenta del usuario a autenticar, tal como se presenta en la Figura 16.

Figura 16: Pasos de la autenticación de doble factor por código de verificación.

Además de la autenticación de doble factor, al tratarse de un sistema que se centra básicamente en la gestión de los documentos generados por el proceso de prácticas profesionales plantea la necesidad de poder adjuntar la identidad del autor con el documento que genera con el fin de proteger la integridad de la documentación [10]. Para cubrir aquella necesidad se implementa la firma digital, lo cual nos permite demostrar que el autor del documento existe, que no puede negar que envió el documento y que el documento no ha sido modificado desde el envío. En la Figura 17 se muestran los pasos para realizar una firma digital en el sistema.

Figura 17: Pasos de la firma digital.

Estas dos herramientas disponibles en el sistema están basadas en las medidas de control de accesos que la norma ISO 27001:2013 detalla, con el objetivo de controlar y monitorizar los accesos a los medios de información según las políticas definidas por la organización [11]. Y, para comprobar el funcionamiento de las herramientas mencionadas, se ha realizado pruebas para verificar cuántas veces puede ejecutarse sin presentar errores, el procedimiento se muestra a continuación. Primero se define la Ecuación 1 para evaluar el programa.

$$Confiabilidad = 100\% * \frac{N° \ de \ casos \ fallidos}{N° \ de \ casos \ totales} \quad (1)$$

También se determinó una escala, la cual se muestra:

0% -10%, **Seguro**
11% -20%, **Regular**
21% - 100%, **Inseguro**

Se realizaron 200 casos de uso para cada apartado. Los casos fallidos son aquellas situaciones donde se ejecuta el programa y se lanzan errores de codificación o el programa no responde como debería. Con la fórmula de confiabilidad (Ecuación 1) se obtienen los siguientes resultados:

Evaluación de efectividad: Autenticación de dos factores

$$Confiabilidad = 100\% * \frac{19}{200} = 9{,}5\%$$

Evaluación de efectividad: Autenticación de dos factores

$$Confiabilidad = 100\% * \frac{32}{200} = 16\%$$

Conclusiones

Concluimos que los controles de acceso proporcionan seguridad en la información manejada en el sistema de acuerdo a las pruebas de confianza. De esta manera, es posible afirmar que los controles de acceso en los sistemas académicos representan un activo importante en los planes de seguridad de los sistemas académicos, y de acuerdo a ISO 27001, son importantes en la elaboración a futuro de un Sistema de Gestión de Seguridad de la Información (SGSI). También se concluye que la implementación hipotética de dicho servicio junto a los controles de acceso para el proceso de gestión de prácticas profesionales ayudaría enormemente a la agilización de los procesos, además de la seguridad de los datos que son utilizados dentro del sistema. Se recomienda realizar la mejora continua de estos controles de acceso para el beneficio de los sistemas de información implementados a futuro en la Universidad Nacional de Trujillo.

Referencias

[1] Dronov, V.Y. and Dronova, G.A. Principles of Information Security Management System - Journal of Physics: Conference Series. Disponible: https://iopscience.iop.org/article/10.1088/1742-6596/2182/1/012092 (Accedido: November 25, 2022).

[2] Molina, D. 2002, Material de Apoyo Instruccional. Curso Orientación Educativa. Barinas: Unellez.

[3] Tatiara, R., Fajar, A.N., Siregar, B. & Gunawan, W. 2018, "Analysis of factors that inhibiting implementation of Information Security Management System (ISMS) based on ISO 27001", Journal of Physics: Conference Series. (Accedido: November 26, 2022).

[4] Congreso de la República. Ley No 28518. Ley sobre Modalidades Formativas Laborales. (2018, 12 diciembre). SITEAL. https://siteal.iiep.unesco.org/sites/default/files/sit_accion_files/pe_6114.pdf

[5] RODRIGUEZ PEROJO, Keilyn & RONDA LEON, Rodrigo. (2006). El web como sistema de información. ACIMED [online]. 2006, vol.14, n.1. ISSN 1024-9435.

[6] Xavier Albaladejo. Qué es SCRUM. (2021, 20 septiembre). Proyectos Ágiles. https://proyectosagiles.org/que-es-scrum/

[7] Sistema de gestión de seguridad de la información. (2022, 25 diciembre). Orientación - Presidencia del Consejo de Ministros - Gobierno del Perú. https://www.gob.pe/14086-sistema-de-gestion-de-seguridad-de-la-informacion

[8] ISOTools Chile. (2017, 14 diciembre). Cómo la autenticación de dos factores permite cumplir con los controles de acceso ISO 27001. https://www.isotools.cl/como-la-autenticacion-de-dos-factores-permite-cumplir-con-los-controles-de-acceso-iso-27001/

[9] Apple. (2022, 11 noviembre). Autenticación de doble factor para el ID de Apple. Apple Support. https://support.apple.com/es-es/HT204915

[10] Soto, L. (2022, 22 noviembre). ¿Qué es una firma digital? https://blog.signaturit.com/es/que-es-una-firma-digital

[11] Norma ISO 27001. (2013). ISO 27002 punto por punto A9 Control de acceso - Caso Práctico. (s. f.-b). ISO 27001. https://normaiso27001.es/a9-control-de-acceso/

[12] What is ISO 27001 and How To Get an ISO 27001 Certification. (s. f.). NQA Certification Body. https://www.nqa.com/es-pe/certification/standards/iso-27001

[13] Margarita Labastida Roldán, José Luis Ruiz Islas & Fernando Saldaña Ramírez. (2019). SSGPP: Sistema de Semi-Automatización y Gestión de Prácticas profesionales. Iztatl Computación, 15, 48. https://ingenieria.uatx.mx/docs/RevistaIztatlComputacionNo15.pdf

[14] Nica Latto. (2020). ¿Qué es la autenticación de doble factor (2FA)? ¿Por qué la necesita? Blog AVG Signal. https://www.avg.com/es/signal/what-is-two-factor-authentication

[15] Junta de Andalucía. (2012). Control de Acceso y Autenticación. Marco de Desarrollo de la Junta de Andalucía. https://www.juntadeandalucia.es/servicios/madeja/contenido/subsistemas/desarrollo/control-acceso-y-autenticacion

Identificador de sentimientos de comentarios de hoteles utilizando BERT

Hotel feedback sentiment identifier using BERT

Walther Medina Pauca
Universidad La Salle. Arequipa, Perú
wmedinap@ulasalle.edu.pe

Camila Huamani Tito
Universidad La Salle. Arequipa, Perú
chuamanit@ulasalle.edu.pe

ARK: ark:/42411/s11/a63
PURL: 42411/s11/a63

RECIBIDO 22/10/2022 • ACEPTADO 10/12/2022 • PUBLICADO 30/03/2023

RESUMEN

La forma de escribir del ser humano fue cambiando con el tiempo siendo reducidas/abreviadas por las nuevas generaciones. El proyecto investigará estas formas de escribir de la personas a través de comentarios de hoteles, para poder identificar y realizar su clasificación de acuerdo si este es un comentario formal o informal; a la vez se tratará de identificar cada uno de estos si cuenta con información positiva o negativa. Todos los procesos para identificar textos serán usados con el Procesamiento de lenguaje natural (NLP), así lograremos identificar diferentes oraciones de acuerdo al contexto que se encontrará en el comentario de la base de datos, la cual será sacada de TripAdvisor.

Palabras claves: Clasificación, comentarios, dataset, exactitud, promedio, procesamiento de lenguaje natural, NLP.

ABSTRACT

The way of writing of the human being was changing over time being reduced/abbreviated by the new generations. The project will investigate these forms of writing of people through hotel comments, in order to identify and classify them according to whether it is a formal or informal comment; at the same time we will try to identify whether each of these has positive or negative information. All the processes to identify texts will be used with Natural Language Processing (NLP), so we will be able to identify different sentences according to the context that will be found in the comment database, which will be taken from TripAdvisor.

Keywords: *Accuracy, averaging, classification, comments, dataset, natural language processing, NLP.*

INTRODUCCIÓN

Cuando se requiere buscar un hotel y ver si es adecuado con las características que se requieren, las personas pocas veces van a los comentarios, donde les brindan una idea externa a la experiencia o la crítica que dan por su estancia en un particular hotel, con esto en mente el proyecto consiste en clasificar los comentarios de acuerdo a los ejemplos vistos de otros sitios de reserva de hoteles, entres esto destaca el tamaño del comentario, la forma de describir el entorno y la experiencia en dicho establecimiento, usando algoritmos de clasificación de textos y obteniendo archivos CSV para los comentarios de los hoteles.

Para la prueba del buen funcionamiento del sistema a implementar es adquirir datos de varias páginas de reserva de hoteles que cuenten con comentarios tanto positivos y negativos para sacar la forma de escritura de algunos clientes o críticos para así realizar las debidas comparaciones, de acuerdo a la base de datos lingüística que tendrá el proyecto. Por eso se buscará en kaggle.com, que tiene más de 515,000 comentarios de clientes en inglés, además el documento nos mostrará diferentes atributos, pero sólo tendremos en cuenta: Negative_Review, Positive_Review y Reviewer_Score. Dichos datos serán obtenidos por el CSV que dará la página. Además, su obtención del archivo CSV se usará un árbol para ir buscando las palabras reservadas Como solo 3 atributos son elegidos su almacenamiento irá a una base de datos para que la máquina aprenda los patrones para futuras consultas.

Un buen machine learning, además de almacenar datos para futuras llamadas es identificar escritura informal o formal, está usará un árbol de recorrido y búsqueda usando otro método de machine learning para hacer las debidas comparaciones y mandar en el caso de ser formal o informal, la cantidad exacta que contiene el comentario entre las dos formas de escritura. Lo mismo sucedería con la identificación de ser comentarios críticos o positivos, pasará por un método de Machine Learning realizando las debidas consultas. Para la realización de clasificar los comentarios de acuerdo a experiencias positivas como negativas el análisis de sentimiento es el adecuado para realizar dicha acción.

El proyecto será realizado en una página web, mostrando los comentarios de diferentes hoteles de manera "filtrada" como pantalla principal, donde los comentarios buenos y descriptivos de las características del hotel serán los primeros.

Además, como son textos y no simplemente palabras soltadas al azar es un poco más complicado identificar el contexto de estas; no es algo raro de encontrar ya que se considera como el lenguaje

neutral siendo ambiguo, pero usando el Part of Speech Tagging, se logrará identificar oraciones para clasificarlas gramaticalmente.

Materiales y métodos o Metodología computacional

Usando NLP y cado uno de las propuestas que se fueron estableciendo, cuentan con diferentes librerías o la misma librería, en todos los proyectos presentados siempre estará la librería NLTK (Natural Language Toolkit), el cual es la que tiene más información y más material para su desarrollo, el problema que al ser una librería completa es pesada al momento de ser cargada y ejecutada para las aplicaciones. A pesar que el NLTK sea completo, el proyecto usará la librería BeautifulSoup, transformers, torch sklearn, ya que éste simplifica el NLTK de una manera intuitiva sacando lo que es el Part-of-speech tagging, usa el método de Naive Bayes, Decision Tree (métodos más certeros al momento de identificar palabras, pero a la vez los más lentos), también esta librería nos ofrece funciones para una clasificación rápida de las palabras negativas como positivas.

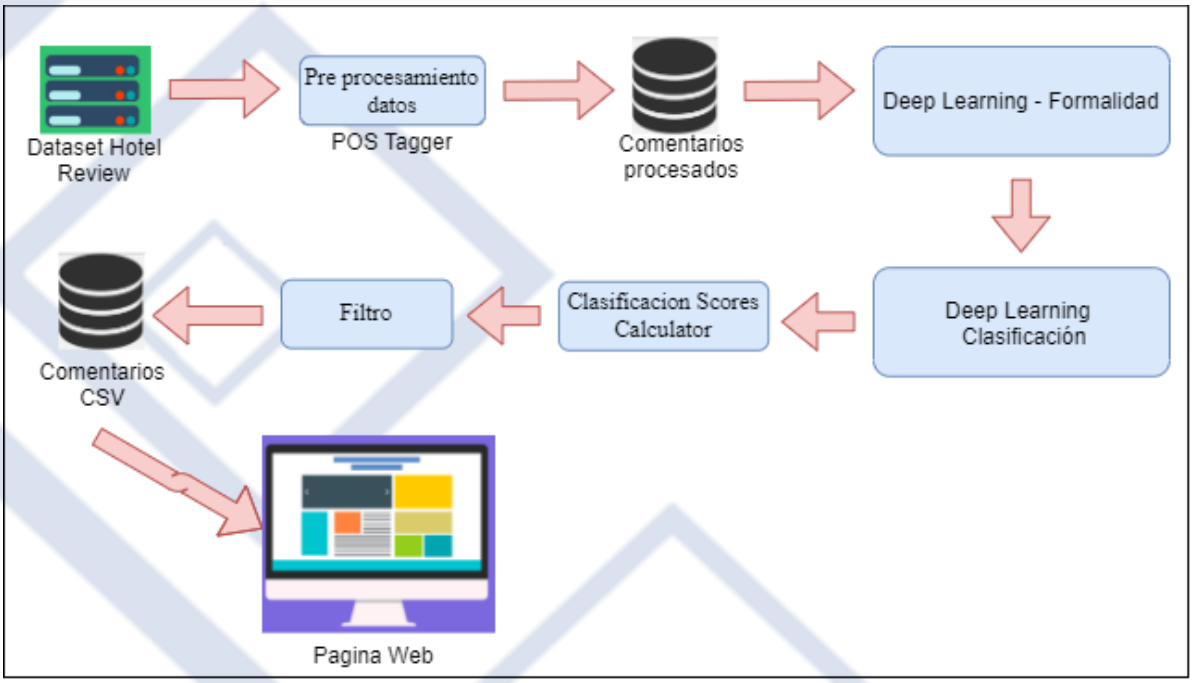


Fig 1. Pipeline de clasificación de comentarios

Para la clasificación de los comentarios/reseñas seguiremos los pasos descritos en la Fig 2.

- Dataset Hotel Review: Estamos obteniendo la información de un dataset que nos servirá de modelo para el software. Este dataset podrá ser cambiado por una base de datos real con reseñas reales.

- Pre-procesamiento datos: Haciendo uso de Part of Speech Taggin para lograr identificar los comentarios reales de comentarios con palabras aleatorias.
- Comentarios procesados: Corresponde al almacenamiento de los comentarios previamente clasificados como reales para utilizarlos posteriormente.
- Deep Learning - Formalidad: Corresponde al identificador de comentarios formales y los comentarios informales que contengan jergas.
- Deep Learning - Clasificación: Corresponde al identificador de comentarios críticos constructivos y de los demás.
- Clasificación Scores Calculator: Permite agregar una clasificación a comentarios que se han considerado como neutros dependiendo de la puntuación que le brindaron en la reseña.
- Filtro: Este filtro está dado por nosotros, pudiendo ser variable permitiendo así obtener los datos que realmente necesitamos.
- Comentarios CSV: Los comentarios ya luego de ser filtrados se almacenarán en formato CSV. Para su creación, utilizamos diferentes librerías que nos permitieron realizar las diferentes operaciones sobre el dataset y este muestre los resultados mejor esperados.

Bibliotecas utilizadas:

Transformers.- Proporciona APIs para descargar y entrenar fácilmente modelos pre entrenados de última generación.
Torch.- Es un framework de redes neuronales, para diseñar y entrenar redes neuronales.
BeautifulSoup.- biblioteca de Python para extraer datos de archivos HTML y XML.
Sklearn.- Proporciona algoritmos para machine learning.

```
from transformers import AutoTokenizer, AutoModelForSequenceClassification
import torch
import requests
import numpy as np
import pandas as pd
from bs4 import BeautifulSoup
import re
from sklearn.metrics import *
```

Entrenando el algoritmo con el diccionario de datos de BERT MULTILINGUAL.

```
tokenizer = AutoTokenizer.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')
model = AutoModelForSequenceClassification.from_pretrained('nlptown/bert-base-multilingual-uncased-sentiment')
```

Se generan tokens para hacer la prueba con la oración ingresada manualmente.

```
tokens = tokenizer.encode('meh, it was okay', return_tensors='pt')
result = model(tokens)
##print(result.logits)
```

Se carga el dataset para comenzar con la limpieza del mismo.

```
df = pd.read_csv('dataset2.csv')
print( df.head())
```

Reduciendo los datos mostrados:
Filtramos el dataframe.

```
dfreview = df[['Review']]
```

Verificamos datos duplicados

```
print('Cantidad de datos duplicados: ', dfreview.duplicated().sum())
```

Verificamos la existencia de datos NULL

```
print('Cantidad de datos nulos: ', df.isnull().sum())
```

Hacemos un conteo de los datos luego de la limpieza.

```
print('Cantidad de datos en Review: ', df['Review'].value_counts())
```

Encapsulamos el proceso de sentimiento en una función, esto hace que sea más fácil procesar múltiples cadenas.

```
def sentiment_score(review):
    tokens = tokenizer.encode(review, return_tensors='pt')
    result = model(tokens)
    return (int(torch.argmax(result.logits))+1)
```

Se utiliza la función para poder generar la revisión en el marco de datos.

```
df['Sentiment'] = df['Review'].apply(lambda x: sentiment_score(x[:512]))
```

Se convierte el valor de las columnas en listas, para poder ser procesadas.

```
sentiment_column = df.loc[:,'Sentiment']
sentiment_nums = sentiment_column.values
sentiment_nums = sentiment_nums.tolist()
rating_column = df.loc[:,'Rating']
rating_nums = rating_column.values
rating_nums = rating_nums.tolist()
```

Convirtiendo los valores de ambas listas a clases A, B y C.

```
filter_actual = []
```

```
for element in rating_nums:
    if element < 3:
        filter_actual.append('A')
    elif element == 3:
        filter_actual.append('B')
    else:
        filter_actual.append('C')
filter_pred = []
for element in sentiment_nums:
    if element < 3:
        filter_pred.append('A')
    elif element == 3:
        filter_pred.append('B')
    else:
        filter_pred.append('C')
```

Se construye una serie con pandas, para que se pueda construir la matriz de confusión.

```
filter_actual = pd.Series(filter_actual, name='actual')
filter_pred = pd.Series(filter_pred, name='pred')
```

Se halla la matriz de confusión, un informe de clasificación y el valor exactitud.

```
cm = confusion_matrix(filter_actual, filter_pred)
matrix = classification_report(filter_actual, filter_pred)
accuracy = accuracy_score(filter_actual, filter_pred)
```

Macro average metrics:

Macro average precision implementado para casos en los que se tengan más de dos clases.

```
print('La presicion Score de macro es: ', precision_score(filter_actual,filter_pred, average='macro'))
```

Macro average recall

```
print('El macro promedio recall es: ', recall_score(filter_actual,filter_pred, average='macro'))
```

Informe de clasificación para la precisión, el recall, la f1-score y la accuracy.

```
matrix = classification_report(filter_actual, filter_pred)
```

En esta sección se explica cómo se hizo la investigación. Se describe el diseño de la misma y se explica cómo se llevó a la práctica, justificando la elección de métodos y técnicas de forma tal que un lector pueda repetir el estudio.

Subepígrafes en caso de utilizarse

Los párrafos se escribirán en Times New Roman a 11 puntos y con espaciado 1,5 y una línea en blanco como separador.

Resultados y discusión

Luego de la limpieza y el procesamiento del dataset, podemos presentar los siguientes resultados. Conforme a la primera obtención de resultados se presentan las siguientes listas.

Valores de clasificación del dataset:

[4, 2, 3, 5, 5, 5, 5, 4, 5, 5, 2, 4, 4, 3, 4, 1, 2, 5, 5, 3, 5, 5, 4, 5, 2, 3, 4, 3, 4, 4, 4, 4, 1, 2, 4, 4, 4, 5, 4, 4, 1, 4, 2, 4, 2, 2, 3, 3, 5, 5, 5, 4, 5, 5, 3, 5, 3, 5, 5, 5, 4, 5, 5, 5, 4, 1, 5, 3, 3, 1, 4, 2, 5, 5, 4, 5, 1, 1, 4, 2, 2, 5, 5, 2, 4, 4, 5, 4, 1, 4, 4, 5]

Valores obtenidos luego del procesamiento:

[4, 2, 3, 5, 1, 5, 4, 5, 5, 5, 2, 4, 4, 3, 5, 1, 2, 4, 4, 3, 4, 5, 5, 5, 4, 4, 5, 3, 4, 4, 4, 4, 2, 3, 4, 4, 4, 5, 4, 3, 1, 5, 2, 5, 2, 1, 2, 3, 5, 4, 5, 5, 5, 5, 2, 4, 3, 5, 5, 5, 3, 5, 5, 5, 5, 1, 5, 2, 3, 3, 5, 2, 5, 5, 4, 4, 1, 1, 4, 1, 1, 5, 4, 2, 4, 4, 4, 4, 1, 2, 3, 5]

En el cual podemos observar que los valores ordenados presentan cierta similitud, ya que al momento de poder comparar con la clasificación dada del dataset, resaltamos casos en los que los valores son exactamente iguales.

2. Luego de filtrar los valores para crear 3 clases se obtiene el siguiente resultado:

Valores de clasificación del dataset:

['C', 'A', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'C', 'B', 'C', 'A', 'A', 'C', 'C', 'B', 'C', 'C', 'C', 'C', 'A', 'B', 'C', 'B', 'C', 'C', 'C', 'C', 'A', 'A', 'C', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'A', 'C', 'A', 'A', 'B', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'B', 'C', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'B', 'B', 'A', 'C', 'A', 'C', 'C', 'C', 'C', 'A', 'A', 'C', 'A', 'A', 'C', 'C', 'A', 'C', 'C', 'C', 'C', 'A', 'C', 'C', 'C']

Valores obtenidos luego del procesamiento:

['C', 'A', 'B', 'C', 'A', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'C', 'B', 'C', 'A', 'A', 'C', 'C', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'C', 'B', 'C', 'C', 'C', 'C', 'C', 'A', 'B', 'C', 'C', 'C', 'C', 'C', 'B', 'A', 'C', 'A', 'C', 'A', 'A', 'A', 'B', 'C', 'C', 'C', 'C', 'C', 'C', 'A', 'C', 'B', 'C', 'C', 'C', 'B', 'C', 'C', 'C', 'C', 'A', 'C', 'A', 'B', 'B', 'C', 'A', 'C', 'C', 'C', 'C', 'A', 'A', 'C', 'A', 'A', 'C', 'C', 'A', 'C', 'C', 'C', 'C', 'A', 'A', 'B', 'C']

Como se puede observar, se presentan más datos resaltados por el rango que se dio al generar las clases.

Confusion matrix:
[[17 2 1]
[3 7 1]
[2 3 56]]

Accuracy value: 0.8695652173913043

La precisión Score de macro es: 0.7738592824799722

El macro average recall es: 0.8014654744162941

Informe de clasificación:

	precision	recall	f1-score	support
A	0.77	0.85	0.81	20
B	0.58	0.64	0.61	11
C	0.97	0.92	0.94	61
accuracy			0.87	92
macro avg	0.77	0.80	0.79	92
weighted avg	0.88	0.87	0.87	92

Conclusiones

Terminando de realizar los estudios necesarios. llegamos a la conclusión de que una de las mejores librerías para trabajar NLP es BERT, ya que gracias a su sofisticada tecnología nos proporciona herramientas para poder acercarnos de forma muy clara a las expresiones reales de las personas.

Al momento de comparar los diferentes resultados obtenidos, vemos que la diferencia entre uno y otro es muy corta, sabiendo que se utilizó una base de datos medianamente corta, el promedio de valores desiguales que se pudieron presentar en este, eran de un nivel muy bajo.

Pudiendo observar los valores obtenidos tanto Accuracy y f1-score podemos reconocer que el algoritmo empleado para procesar los sentimientos en comentarios sobre un hotel, es confiable. ya que presenta una exactitud de 0.87.

Referencias

[1] S. Vidya (2018). "Cross Domain Sentiment Classification Using Natural Language Processing". Assistant Professor, Department of Computer Science and Engineering, Kalasalingam Institute of Technology, KrishnaSnkoil, Tamil Nadu, India. Recuperado de: https://www.researchgate.net/profile/Vidya-

Soundarapandian/publication/339901166_Cross_Domain_Sentiment_Classification_Using_Natural_Language_Processing/links/5e6b74c2a6fdccf321d98d41/Cross-Domain-Sentiment-Classification-Using-Natural-Language-Processing.pdf

[2] Ishaq, A., Umer, M., Mushtaq, M., Medaglia, C., Siddiqui, H., Mehmood, A. and Choi, G., 2020. Extensive hotel reviews classification using long short term memory. Journal of Ambient Intelligence and Humanized Computing, 12(10), pp.9375-9385.

[3] Ghabayen, A. and Ahmed, B., 2019. Polarity Analysis of Customer Reviews Based on Part-of-Speech Subcategory. Journal of Intelligent Systems, 29(1), pp.1535-1544.

[4] Shin, S., Du, Q. and Xiang, Z., 2018. What's Vs. How's in Online Hotel Reviews: Comparing Information Value of Content and Writing Style with Machine Learning. Information and Communication Technologies in Tourism 2019, pp.321-332.

[5]Colab.research.google.com. 2022. Google Colaboratory. [online] Available at: <https://colab.research.google.com/github/bentrevett/pytorch-sentiment-analysis/blob/master/1%20-%20Simple%20Sentiment%20Analysis.ipynb?authuser=1#scrollTo=RMDoLMlxaUql> [Accessed 10 July 2022].

[6] Medium. 2022. Churning the Confusion out of the Confusion Matrix. [online] Available at: <https://blog.clairvoyantsoft.com/churning-the-confusion-out-of-the-confusion-matrix-b74fb806e66> [Accessed 10 July 2022].

[7] E. Cambria and B. White, "Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]," in IEEE Computational Intelligence Magazine, vol. 9, no. 2, pp. 48-57, May 2014, doi: 10.1109/MCI.2014.2307227.

[8] SALAMI, Salami. IMPLEMENTING NEURO LINGUISTIC PROGRAMMING (NLP) IN CHANGING STUDENTS' BEHAVIOR: RESEARCH DONE AT ISLAMIC UNIVERSITIES IN ACEH. Jurnal Ilmiah Peuradeun, [S.l.], v. 3, n. 2, p. 235-256, may 2015. ISSN 2443-2067. Available at: <http://www.journal.scadindependent.org/index.php/jipeuradeun/article/view/65>. Date accessed: 10 july 2022.

[9] D. Yu, K. Yao and Y. Zhang, "The Computational Network Toolkit [Best of the Web]," in IEEE Signal Processing Magazine, vol. 32, no. 6, pp. 123-126, Nov. 2015, doi: 10.1109/MSP.2015.2462371.

[10] E. H. Houssein, R. E. Mohamed and A. A. Ali, "Machine Learning Techniques for Biomedical Natural Language Processing: A Comprehensive Review," in IEEE Access, vol. 9, pp. 140628-140653, 2021, doi: 10.1109/ACCESS.2021.3119621.

[11] A. Elnagar, S. M. Yagi, A. B. Nassif, I. Shahin and S. A. Salloum, "Systematic Literature Review of Dialectal Arabic: Identification and Detection," in IEEE Access, vol. 9, pp. 31010-31042, 2021, doi: 10.1109/ACCESS.2021.3059504.

[12] D. Mahendran, C. Luo and B. T. Mcinnes, "Review: Privacy-Preservation in the Context of Natural Language Processing," in IEEE Access, vol. 9, pp. 147600-147612, 2021, doi: 10.1109/ACCESS.2021.3124163.

[13] X. Feng and Y. Zeng, "Neural Collaborative Embedding From Reviews for Recommendation," in IEEE Access, vol. 7, pp. 103263-103274, 2019, doi: 10.1109/ACCESS.2019.2931357.

[14] X. Feng and Y. Zeng, "Multi-Level Fine-Grained Interactions for Collaborative Filtering," in IEEE Access, vol. 7, pp. 143169-143184, 2019, doi: 10.1109/ACCESS.2019.2941236.

[15] S. Salloum, T. Gaber, S. Vadera and K. Shaalan, "A Systematic Literature Review on Phishing Email Detection Using Natural Language Processing Techniques," in IEEE Access, vol. 10, pp. 65703-65727, 2022, doi: 10.1109/ACCESS.2022.3183083.

El modelo COBIT 5 para Auditoría Informática de los Sistemas de Información Académica de la Universidad Nacional Jorge Basadre Grohmann

The cobit 5 model for Computer Audit of the academic information systems of the National University Jorge Basadre Grohmann

Rene Aquino Arcata
Universidad Nacional Jorge Basadre Grohmann, Tacna, Perú.
rene.aquino@unjbg.edu.pe
https://orcid.org/0000-0002-5041-7344

Ronald Cuevas Machaca
Universidad Nacional Jorge Basadre Grohmann, Tacna, Perú.
ronald.cuevas@unjbg.edu.pe
https://orcid.org/0000-0002-3887-6396

Gustavo Adolfo Villarroel Laura
Universidad Nacional Jorge Basadre Grohmann, Tacna, Perú.
gustavo.villarroel@unjbg.edu.pe
https://orcid.org/0000-0002-6047-5679

ARK: ark:/42411/s11/a56
PURL: 42411/s11/a56

RECIBIDO 30/10/2022 • ACEPTADO 19/12/2022 • PUBLICADO 30/03/2023

RESUMEN

La Universidad Nacional Jorge Basadre Grohmann en su proyecto "MEJORAMIENTO DEL SERVICIO INFORMÁTICO DE LA PLATAFORMA WEB DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN DE TACNA", con el fin de mejorar la calidad e integración de la información, tiene como fin la implantación de una plataforma educativa. El presente trabajo de investigación titulado "El modelo COBIT 5 para Auditoría Informática de los Sistemas de Información Académica de la Universidad Nacional Jorge Basadre Grohmann", tiene como objetivo general proponer los lineamientos y controles necesarios, mediante la utilización del COBIT 5, para el posterior mantenimiento del sistema de información, a fin de lograr la excelencia operativa a través de una aplicación fiable y uso eficiente de la tecnología..

Palabras claves: Sistema de auditoría, Sistema de Información, Cobit 5, Requerimientos.

ABSTRACT

The Jorge Basadre Grohmann National University in its project "IMPROVEMENT OF THE COMPUTER SERVICE OF THE WEB PLATFORM OF THE JORGE BASADRE GROHMANN NATIONAL UNIVERSITY OF TACNA". in order to improve the quality and integration of the information, its purpose is the implementation of the educational platform. The present research work entitled "The COBIT 5 model for Computer Audit of Academic Information Systems of the Jorge Basadre Grohmann National University", has as a general objective to carry out the necessary guidelines and controls in the use of COBIT 5, for its subsequent maintenance. of the information system in order to achieve operational excellence through a reliable, efficient application of technology.

Keywords: *Systems audit, Information Systems, Cobit 5, Requirements.*

INTRODUCCIÓN

Los sistemas de información son base primordial para el desarrollo empresarial, los sistemas de información transcurrido los años han sufrido transformaciones los cuales han proporcionado beneficios, y las formas de auditar han sido modificadas en función de las necesidades que se van presentando una a continuación de otra.

En tal sentido, el Cobit es un aporte a través de un modelo que permite revisar cuidadosamente el trabajo realizado por los sistemas informáticos en relación a las necesidades empresariales. La tecnología de la información está avanzando cada vez más y se ha generalizado en las empresas y en entornos sociales, públicos y de negocios.

CobiT, es un marco de referencia y un juego de herramientas de soporte que permiten a la gerencia cerrar la brecha con respecto a los requerimientos de control, temas técnicos y riesgos de negocio, y comunicar ese nivel de control a los participantes. CobiT, permite el desarrollo de políticas claras y de buenas prácticas para el control de TI por parte de las empresas. CobiT constantemente se actualiza y armoniza con otros estándares; por lo tanto, CobiT se ha convertido en el integrador de las mejores prácticas de TI y el marco de referencia general para el gobierno de TI que ayuda a comprender y administrar los riesgos y beneficios asociados con TI. Con lo cual, la estructura de procesos de CobiT y su enfoque de alto nivel orientado al negocio brindan una visión completa de TI y de las decisiones a tomar.

El objetivo de COBIT es brindar buenas prácticas a través de un marco de trabajo de dominios y procesos, y presentar las actividades de una manera manejable y lógica. Estas prácticas están enfocadas más al control que a la ejecución.

Asimismo COBIT 5 ayuda a las empresas de todos los tamaños aportando beneficios tales como: Mantener la información de alta calidad para apoyar las decisiones de negocios, Alcanzar los objetivos estratégicos y obtener los beneficios de negocio a través del uso efectivo e innovador de las TI, Lograr la excelencia operativa a través de una aplicación fiable, eficiente de la tecnología, Mantener los riesgos relacionados con TI a un nivel aceptable, Optimizar los servicios y el coste de las TI y la tecnología, Apoyar el cumplimiento de las leyes, reglamentos, acuerdos contractuales y las políticas.

El objetivo de esta investigación es estructurar los lineamientos para la aplicación del modelo cobit 5, a los sistemas de información académica de la universidad nacional Jorge Basadre Grohmann los que componen los siguientes módulos. (Aplicativo informático web de gestión del sistema de admisión, Sistema de escalafón, Aplicativo informático web de gestión del sistema de bienestar universitario, Sistema de tutoría, Aplicativo informático de gestión y seguimiento de egresados, Aplicativo informático web de gestión de la bolsa de trabajo y capacitación, Aplicativo informático web de sistema de matrícula, Aplicativo informático web de gestión docente, Sistema de intranet, Aplicativo informático web de gestión de registro académico, Aplicativo informático web de gestión de aprendizaje virtual, Aplicativo informático web de gestión institucional con base a indicadores, Aplicativo informático web de gestión de proyectos de investigación - ogin, Aplicativo informático web de gestión del comedor universitario, Aplicativo informático web de gestión de cooperación nacional e intercambio académico, Website facultad, Website de escuela de pregrado,website de instituto de informática y telecomunicaciones-itel, Aplicativo informático web de gestión de equipo informático, Aplicativo informático web de gestión de biblioteca), para el control de los sistemas de información.

Problemática

Por todo lo anteriormente expuesto nos planteamos la siguiente pregunta de investigación:

La universidad nacional Jorge Basadre Grohmann en la ejecución del Proyecto “MEJORAMIENTO DEL SERVICIO INFORMÁTICO DE LA PLATAFORMA WEB DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN DE TACNA”, en la fase de entrega del sistema de información requerido se ha evidenciado que actualmente no cuenta con los lineamientos y controles necesarios para su posterior mantenimiento a fin de lograr la excelencia operativa a través de una aplicación fiable y uso eficiente de la tecnología.

Marco teórico y Referencial

Definición de Auditoría.

Según Manual Latinoamericano de Auditoría Profesional en el sector Público define a la Auditoría como: Es el examen objetivo, sistemático y profesional de las operaciones ejecutadas con la finalidad de evaluarlas, verificarlas y emitir un informe que contenga comentarios, conclusiones y recomendaciones. (Instituto Latinoamericano de Ciencias Fiscalizadores - ILACIF, 1981, pág. 44)

Para Auditing Concepts Committee es Un proceso sistemático para obtener y evaluar evidencia de una manera objetiva respecto de las afirmaciones concernientes a actos económicos y eventos para determinar el grado de correspondencia entre estas afirmaciones y criterios establecidos y comunicar los resultados a los usuarios interesados. (Osorio Sanchez, 1999, pág. 22)

Definición de Auditoría de Sistemas.

La Auditoría de Sistema tiene varias definiciones, para el caso se tomaron en cuenta las siguientes definiciones:

Es un conjunto de procedimientos y técnicas para evaluar y controlar total o parcialmente un sistema informático, con el fin de proteger sus activos y recursos, verificar si sus actividades se desarrollan eficientemente y de acuerdo con la normativa informática y general existente en cada empresa y para conseguir la eficacia exigida en el marco de la organización correspondiente. (González Gallego , 2004)

Es la revisión técnica, especializada y exhaustiva que se realiza a los sistemas computacionales, software e información utilizados en una empresa, sean individuales, compartidos y/o de redes, así como a sus instalaciones, telecomunicaciones, mobiliario, equipos periféricos y demás componentes. (Muñoz Razo, 2002, pág. 19)

Es una función que ha sido desarrollada para asegurar la salvaguarda de los activos de los sistemas de computadoras, mantener la integridad de los datos y lograr los objetivos de la organización en forma eficaz y eficiente. (Weber, 2016)
Es el proceso de recoger, agrupar y evaluar evidencias para determinar si un sistema informatizado salvaguarda los activos, mantiene la integridad de los dalos, lleva a cabo

eficazmente los fines de la organización y utiliza eficientemente los recursos. (Piattini, 2001, pág. 28)

Auditoria Informática.

La Auditoría Informática es un proceso llevado a cabo por profesionales especialmente capacitados para el efecto, y que consiste en recoger, agrupar y evaluar evidencias para determinar si un Sistema de Información salvaguarda el activo empresarial, mantiene la integridad de los datos ya que esta lleva a cabo eficazmente los fines de la organización, utiliza eficientemente los recursos, cumple con las leyes y regulaciones establecidas

Sistemas de Información.

Un sistema de información es un conjunto de componentes que interactúan entre sí con un fin común. En informática, los sistemas de información ayudan a administrar, recolectar, recuperar, procesar, almacenar y distribuir información relevante para los procesos fundamentales y las particularidades de cada organización.

La importancia de un sistema de información radica en la eficiencia en la correlación de una gran cantidad de datos ingresados a través de procesos diseñados para cada área con el objetivo de producir información válida para la posterior toma de decisiones.

Sistema de Información Académica.

El SIA es el Sistema de Información Académica que proporciona una plataforma informática de trabajo para la interacción de usuarios y equipo computacional que facilita la captura, almacenamiento, procesamiento, acceso y salida de información confiable y actualizada sobre programas, proyectos y actividades académicas.

COBIT

COBIT (Control Objectives for Information and Related Technologies / Objetivos de Control Para la Información y Tecnologías Relacionadas) creado por ISACA (Information Systems Audit and

Control Association /Asociación de Auditoría y Control de Sistemas de Información) una asociación internacional que apoya y patrocina el desarrollo de metodologías y certificaciones para la realización de actividades auditoría y control en sistemas de información.

Es un enfoque que permite la auditoría y evaluación de los servicios informáticos de una organización, para apreciar el rendimiento y robustez en términos de seguridad y conformidad. Constituye un repositorio completo que permite controlar el conjunto de las operaciones relacionadas con la información. Ayuda a los dirigentes a entender y gestionar los riesgos relativos a la informática. (Baud, 2015, pág. 24)

Permite auditar y evaluar los servicios informáticos de una empresa para valorar el rendimiento y la robustez en términos de seguridad y conformidad. Forma toda una disciplina completa que permite controlar el conjunto de operaciones relacionadas con la información. Ayuda a los responsables a entender y gestionar los riesgos relativos a la informática. (Baud, 2015, pág. 28)

Como complemento a lo anterior, COBIT enfatiza el cumplimiento regulatorio de los lineamientos y buenas prácticas de control de las tecnologías, ayuda a las organizaciones a incrementar su valor a través de las TI, y permite su alineamiento con los objetivos del negocio.

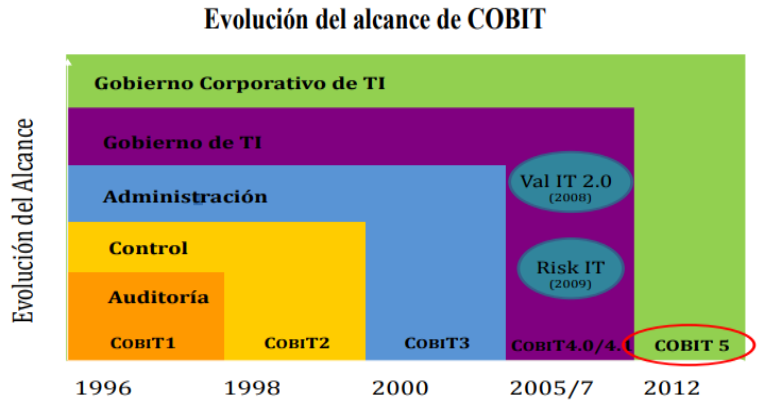


Figura 1 Evolución Del Alcance De COBIT
En: (Information Systems Audit and Control Association - ISACA, 2012)

COBIT 5

COBIT 5 es el resultado de la mejora estratégica de ISACA, el cual Provee de un marco de trabajo integral que ayuda a las empresas a alcanzar sus objetivos para el gobierno y la gestión de las TI corporativas. Dicho de una manera sencilla, ayuda a las empresas a crear el valor óptimo desde IT manteniendo el equilibrio entre la generación de beneficios y la optimización de los niveles de riesgo y el uso de recursos. (Isaca.org, 2016)

COBIT 5 permite a las TI ser gobernadas y gestionadas de un modo holístico para toda la empresa, abarcando al negocio completo de principio a fin y las áreas funcionales de responsabilidad de TI, considerando los intereses relacionados con TI de las partes interesadas internas y externas. COBIT 5 es genérico y útil para empresas de todos los tamaños, tanto comerciales, como sin ánimo de lucro o del sector público. (Information Systems Audit and Control Association - ISACA, 2012).

Materiales y métodos o Metodología computacional

Materiales

Encuestas

Se utilizarán encuestas, dirigidas al responsable de la administración de sistemas y el equipo de ingeniería del área de informática y sistemas de la Universidad Jorge Basadre Grohmann. Estas encuestas tienen el objetivo de conocer los siguientes aspectos:

- Frecuencia de prácticas de auditoría
- Nivel de definición de objetivos
- Nivel de consideración de las partes interesadas en la toma de decisiones
- Factores que influyen en el desarrollo empresarial
- Metas de la organización

Objetivos de Control

En base a los CUATRO (04) dominios que posee el modelo COBIT 5: Planeación y Organización, Adquisición e Implementación, Entrega de Servicios y Soporte, y Monitoreo se definirá su aplicación para el caso propuesto en relación a los 34 procesos del referido modelo y su impacto sobre los Criterios de Información y Recursos de TI.

Métodos

Considerando la problemática identificada, se ha definido los siguientes objetivos de control relacionados a los dominios del modelo COBIT 5

Dominios del COBIT 5.

Como se ha señalado, el modelo COBIT 5 presenta CUATRO (04) dominios, de los cuales, para nuestro caso de estudio, vamos a aplicar el segundo que corresponde al de Adquisición e Implementación (AI) y específicamente a DOS (02) procesos:

AI1: Identificar soluciones de automatización
AI2: Adquirir y Mantener Software de Aplicación

DOMINIO: Adquisición e Implementación (AI)

AI1 Identificar Soluciones de automatización:

El objetivo es asegurar el mejor enfoque para cumplir con los requerimientos del usuario, mediante un análisis claro de las oportunidades alternativas comparadas contra los requerimientos de los usuarios.

AI2 Adquirir y Mantener Software de Aplicación:

El objetivo es proporcionar funciones automatizadas que soporten efectivamente la organización mediante declaraciones específicas sobre requerimientos funcionales y operacionales, y una implementación estructurada con entregables claros.

AI3 Adquirir y Mantener Arquitectura de TI:

El objetivo es proporcionar las plataformas apropiadas para soportar aplicaciones de negocios mediante la realización de una evaluación del desempeño del hardware y software, la provisión de mantenimiento preventivo de hardware y la instalación, seguridad y control del software del sistema.

AI4 Desarrollar y Mantener Procedimientos relacionados con TI:

El objetivo es asegurar el uso apropiado de las aplicaciones y de las soluciones tecnológicas establecidas, mediante la realización de un enfoque estructurado del desarrollo de manuales de procedimientos de operaciones para usuarios, requerimientos de servicio y material de entrenamiento.

AI5 Instalar y Acreditar Sistemas:

El objetivo es verificar y confirmar que la solución sea adecuada para el propósito deseado mediante la realización de una migración de instalación, conversión y plan de aceptaciones adecuadamente formalizadas.

AI6 Administrar Cambios:

El objetivo es minimizar la probabilidad de interrupciones, alteraciones no autorizadas y errores, mediante un sistema de administración que permita el análisis, implementación y seguimiento de todos los cambios requeridos y llevados a cabo a la infraestructura de TI actual.

Sin embargo, también se tiene una identificación más holística del caso de estudio, aun cuando su tratamiento en específico va a corresponder al dominio Adquisición e Implementación (AI)

DOMINIO	PROCESO		Criterios de Información							Recursos de TI				
			Efectividad	Eficiencia	Confidencialidad	Integridad	Disponibilidad	Cumplimiento	Confiabilidad	Recursos	Sistemas de Aplicación	Tecnología	Instalaciones	Datos
Planeación y Organización	PO05	Administrar las inversiones (en TI)	X	X					X	X	X	X		

	P07	Administrar los recursos humano	X	X						X			
	P08	Asegurar el cumplimiento de requerimientos externo	X					X	X	X	X		X
	PO10	Administrar proyectos	X	X						X	X	X	X	
	PO11	Administrar calidad	X	X		X			X	X	X	X	X	
Adquisición e Implementación	AI1	Identificar soluciones de automatización	X	X							X	X	X	
	AI2	Adquirir y Mantener Software de Aplicación	X	X		X		X	X		X			
	AI3	Adquirir y Mantener la Arquitectura Tecnológica	X	X		X						X		
	AI4	Desarrollar y mantener procedimientos	X	X		X		X	X	X	X	X	X	
	AI5	Instalar y acreditar sistemas de información	X			X	X			X	X	X	X	X
	AI6	Administrar cambios	X	X		X	X		X	X	X	X	X	X
	DS5	Garantizar la seguridad de sistemas			X	X	X	X	X	X	X	X	X	X

Entrega de servicios y soporte	DS9	Administrar la configuración	X				X		X		X	X	X	
	DS13	Administrar la operación	X	X		X	X			X	X		X	X
Monitoreo	M1	Monitorear el proceso	X	X	X	X	X	X	X	X	X	X	X	X
	M2	Evaluar lo adecuado del control interno	X	X	X	X	X	X	X	X	X	X	X	X
	M3	Obtener aseguramiento independiente	X	X	X	X	X	X	X	X	X	X	X	X
	M4	Proporcionar auditoria independiente	X	X	X	X	X	X	X	X	X	X	X	X

Procedimiento

De acuerdo a lo expuesto anteriormente, como parte específica del problema identificado, se plantea la aplicación de los siguientes dominios: Adquisición e implementación.

Se elaboró un modelo de evaluación considerando el dominio de Adquisición e Implementación ya que la Universidad Nacional Jorge Basadre Grohmann en su proyecto "MEJORAMIENTO DEL SERVICIO INFORMÁTICO DE LA PLATAFORMA WEB DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN DE TACNA", ha adquirido el software y adecuado a los procedimientos de la universidad. Este modelo de evaluación está basado en COBIT 5, se diseñó una tabla la cual se orienta al departamento de sistemas como indica el COBIT 5.

Este modelo de evaluación consta de una tabla con los campos de preguntas, respuesta y calificación. Se acordó con los integrantes del grupo de investigación determinar que la calificación será de 1(uno) si posee la documentación pertinente y 0 (cero) si no posee. Dentro de este formato no se consideran los puntos intermedios.

A continuación se visualizan las tablas a emplear para dicha evaluación.

Adquisición e Implementación (AI)

PROCESO: AI1: Identificar soluciones automatizadas	**RESPUESTA**	**CALIFICACIÓN**	**DOCUMENTACIÓN**
ACTIVIDADES DEL PROCESO: Definir los requerimientos funcionales y técnicos del negocio			
1.-¿Se identifican los requerimientos funcionales del negocio que cubran el alcance completo de los programas de inversión en TI?			
2.- ¿Se identifican los requerimientos técnicos del negocio que cubran el alcance completo de los programas de inversión en TI?			
3.- ¿Se priorizan los requerimientos funcionales del negocio que cubran el alcance completo de los programas de inversión TI?			
4.- ¿Se priorizan los requerimientos técnicos del negocio que cubran el alcance completo de los programas de inversión en TI?			
5.-¿Se especifican los requerimientos técnicos del negocio que cubran el alcance completo de los programas de inversión en TI?			

PROCESO: AI1: Identificar soluciones automatizadas	RESPUESTA	CALIFICACIÓN	DOCUMENTACIÓN
6.-¿Se especifican los requerimientos técnicos del negocio que cubran el alcance completo de los programas de inversión en TI?			
7.-¿Se acuerdan de los requerimientos funcionales del negocio que cubran el alcance completo de los programas de inversión en TI?			
8.-¿Se acuerdan de los requerimientos técnicos del negocio que cubran el alcance completo de los programas de inversión en TI?			

Fuente: Manual COBIT 5 y Tesis Diagnóstico para la implantación de COBIT en una Empresa de Producción de Martha Elizabeth de la Torre Morales.
Elaborado: Por Autores de la Investigación

PROCESO: AI2: Adquirir y mantener software de aplicación	RESPUESTA	CALIFICACIÓN	DOCUMENTACIÓN
ACTIVIDADES DEL PROCESO: Especificar los controles de aplicación dentro del diseño			
1.-¿Se aborda la seguridad de las aplicaciones?			

PROCESO: AI2: Adquirir y mantener software de aplicación	RESPUESTA	CALIFICACIÓN	DOCUMENTACIÓN
2.-¿Se abordan los requerimientos de disponibilidad en respuesta a los riesgos identificados?			
3.-¿Se abordan los requerimientos en línea con la clasificación de datos?			
4.-¿Se aborda la arquitectura de la información?			
5.-¿Se aborda la arquitectura de seguridad de la información?			
6.-¿Se aborda la tolerancia a riesgos de la organización?			

Fuente: Manual COBIT 5 y Tesis Diagnóstico para la implantación de COBIT en una Empresa de Producción de Martha Elizabeth de la Torre Morales.
Elaborado: Por Autores de la Investigación

PROCESO: AI2: Adquirir y mantener software de aplicación	**RESPUESTA**	**CALIFICACIÓN**	**DOCUMENTACIÓN**
ACTIVIDADES DEL PROCESO: Desarrollar las metodologías y procesos formales para administrar el proceso de desarrollo de la aplicación			
1.-¿Se garantiza que la funcionalidad de automatización se desarrolla de acuerdo con las especificaciones de diseño?			
2.- ¿Se garantiza que la funcionalidad de automatización se desarrolla de acuerdo los estándares de desarrollo y documentación?			
3.- ¿Se garantiza que la funcionalidad de automatización se desarrolla de acuerdo a los requerimientos de calidad y estándares de aprobación?			
4.- ¿Se asegura que todos los aspectos legales se identifican y direccionan para el software aplicativo desarrollado por terceros?			

PROCESO: AI2: Adquirir y mantener software de aplicación	RESPUESTA	CALIFICACIÓN	DOCUMENTACIÓN
5.- ¿Se asegura que todos los aspectos legales se identifican y direccionan para el software aplicativo desarrollado por terceros?			

Fuente: Manual COBIT 5 y Tesis Diagnóstico para la implantación de COBIT en una Empresa de Producción de Martha Elizabeth de la Torre Morales.
Elaborado: Por Autores de la Investigación

PROCESO: AI2: Adquirir y mantener software de aplicación	RESPUESTA	CALIFICACIÓN	DOCUMENTACIÓN
ACTIVIDADES DEL PROCESO: Crear un plan de aseguramiento de la calidad del software para el proyecto			
1.-¿Se desarrolla un plan de aseguramiento de calidad de software?			
2.- ¿Se implementa los recursos de un plan de aseguramiento de calidad del software?			

PROCESO: AI2: Adquirir y mantener software de aplicación	RESPUESTA	CALIFICACIÓN	DOCUMENTACIÓN
3.- ¿Se ejecuta un plan de aseguramiento de calidad del software?			

Fuente: Manual COBIT 5 y Tesis Diagnóstico para la implantación de COBIT en una Empresa de Producción de Martha Elizabeth de la Torre Morales.
Elaborado: Por Autores de la Investigación

PROCESO: AI2: Adquirir y mantener software de aplicación	RESPUESTA	CALIFICACIÓN	DOCUMENTACIÓN
ACTIVIDADES DEL PROCESO: Desarrollar un plan para el mantenimiento de aplicaciones de software			
1.-¿Se desarrolla una estrategia para el mantenimiento de aplicaciones de software?			
2.-¿Se desarrolla una plan para el mantenimiento de aplicaciones de software?			

Fuente: Manual COBIT 5 y Tesis Diagnóstico para la implantación de COBIT en una Empresa de Producción de Martha Elizabeth de la Torre Morales.
Elaborado: Por Autores de la Investigación

Resultados y discusión

Los resultados que se obtengan luego de la aplicación de las encuestas, que representan la verificación del cumplimiento del modelo COBIT 5 planteado para el problema identificado, evidenciarán de forma objetiva las condiciones actuales en que se está implementando el software adquirido como parte del proyecto "MEJORAMIENTO DEL SERVICIO INFORMÁTICO DE LA PLATAFORMA WEB DE LA UNIVERSIDAD NACIONAL JORGE BASADRE GROHMANN DE TACNA".

El resultado ideal es que se cumpla con la mayor cantidad de requisitos propuestos por el modelo COBIT 5; sin embargo, conociendo parte de la realidad ya expuesta en el presente artículo, se trata de evidenciar las falencias que se tiene a fin de recomendar los aspectos a subsanar previo al despliegue de la solución adquirida como parte del proyecto.

En ese sentido, si del total de calificaciones adquiridas por cada uno de los dominios y procesos relacionados, más del 30% están en el espacio de no cumplimiento, se deberá de recomendar a la alta dirección su subsanación, antes del despliegue de la solución, caso contrario, se pondría en alto riesgo el funcionamiento integral de la misma y en consecuencia la actividad operativa, estratégica de la organización, que para el caso es el área académica de la Universidad Jorge Basadre Grohmann de Tacna.

Conclusiones

La aplicación del modelo COBIT 5, es aplicable antes y después de iniciado el proceso de adquisición de una solución para la atención de una necesidad y mejora de las actividades de una organización.

El marco de trabajo COBIT 5, es una guía a seguir para evaluar el desempeño en la gestión de las tecnologías de la información, tanto en organizaciones públicas como privadas. Para el caso, en la ciudad de Tacna y el ámbito educativo de nivel superior, sería una primera experiencia en su aplicación.

Los criterios definidos en el modelo COBIT 5, deben conciliarse con las regulaciones y exigencias de cumplimiento que regulan las adquisiciones en el sector público del estado peruano, debiendo adecuarse a etapas previas y posteriores según sea el caso.

Referencias

[1] Jose Luis Aro Maquera, "Auditoría Informática Del Sistema De Administración Tributaria De La Municipalidad Distrital De Pilcuyo" (2021), [Online] http://repositorio.unap.edu.pe/bitstream/handle/UNAP/15399/Aro_Maquera_Jose_Luis.pdf?sequence=3&isAllowed=y.

[2] Medina Christian Javier, "Aplicación De Cobit 5.0 En El Diseño De Un Gobierno Y Gestión De Ti Para El Centro De Educación Contínua" (2015), [Online] https://drive.google.com/file/d/1JmJyNYONvAzvbS_TqZcMhvDnTQiYNyaT/view?usp=sharing

[3] /Joffre V. León-Acurio, COBIT como modelo para auditorías y control de los sistemas de información, (2018). [Online]. https://drive.google.com/file/d/1VofgFPQT6elSvJcp7LO7lZpY-Y0X_S1n/view?usp=sharing

Modelo de Autentificación de Doble Factor

Two-Factor Authentication Model

Anderson Jhanyx Reyes Riveros
Universidad Nacional de Trujillo. La Libertad, Perú.
ajreyesr@unitru.edu.pe
https://orcid.org/0000-0002-7324-5055

Jhon Erick Salinas Meza
Universidad Nacional de Trujillo. La Libertad, Perú.
jsalinas@unitru.edu.pe
https://orcid.org/0000-0003-1715-2716

Alberto Mendoza de los Santos
Universidad Nacional de Trujillo. La Libertad, Perú.
amendozad@unitru.edu.pe
https://orcid.org/0000-0002-0469-915X

ARK: ark:/42411/s11/a81
PURL: 42411/s11/a81

RECIBIDO 15/11/2022 • ACEPTADO 27/12/2022 • PUBLICADO 30/03/2023

RESUMEN

El presente artículo tiene como objetivo principal el desarrollo de un modelo que permita la autentificación de un usuario para el control de accesos mediante el modelo de Autentificación de doble factor. Para el desarrollo de dicho modelo presentamos un esquema seguro de autentificación de dos factores(TFA) basado en la posesión por el usuario de una contraseña y un dispositivo con capacidad criptográfica. La seguridad de este modelo es de extremo a extremo en el sentido de que el que quiera acceder de una manera fraudulenta se le va a complicar y asi garantizar la seguridad del usuario de dicho sistema, se tuvo como algoritmo Redes criptográficas, el cual es un modelo de doble autentificación. Así mismo se utilizó el lenguaje de programación cakephp 4.0, además de utilizar el programa visual studio code para poder realizar los algoritmos requeridos para que funciones el modelo de doble autentificación.

Palabras claves: Control de acceso, Autenticación de dos factores, Criptografía.

ABSTRACT

The main objective of this paper is the development of a model that allows the authentication of a user for access control using the Two-Factor Authentication model. For the development of such

a model we present a secure two-factor authentication (TFA) scheme based on the user's possession of a password and a cryptographically capable device. The security of this model is end-to-end in the sense that whoever wants to access in a fraudulent way is going to find it difficult and thus guarantee the security of the user of the system, the algorithm used was Cryptographic Networks, which is a double authentication model. Also the programming language cakephp 4.0 was used, in addition to using the visual studio code program to perform the algorithms required for the double authentication model to work.

Keywords: *Access Control, Two-factor Authentication, Convolutional Cryptography.*

INTRODUCCIÓN

Actualmente las contraseñas son el mecanismo dominante de autentificación digital y por ende van a proteger una gran cantidad de información importante. Sin embargo, las contraseñas son vulnerables ataques en línea y fuera de línea. Un adversario de la red puede probar las contraseñas adivinadas en interacciones en línea con el servidor, mientras que un atacante que compromete los datos de autenticación almacenados por el servidor (es decir, una base de datos de contraseñas) puede organizar un ataque de diccionario fuera de línea comparando la información de autenticación de cada usuario con un diccionario de posibles contraseñas. Los ataques de diccionario fuera de línea son una amenaza importante, experimentada rutinariamente por las ventas comerciales, y conducen al compromiso de miles de millones de cuentas de usuario [13], por lo cual requiere de un sistema de autenticación para restringir el acceso de los usuarios a cierta información almacenada en computadores.

Dentro de los diferentes tipos de autenticación se encuentra la autenticación donde los usuarios autorizados pueden acceder a los datos almacenados en la nube, según el esquema del mecanismo de autenticación para sistemas de banca por Internet en la nube con autenticación multifactorial. Los usuarios se autentican utilizando una combinación de factores como su nombre de usuario, contraseña, número aleatorio y huella dactilar biométrica. -La huella biométrica del usuario se utiliza para cifrar el número aleatorio [6]. Sin embargo, en este proceso, el número arbitrario cifrado se envía al número de teléfono registrado a través de un entorno abierto y vulnerable, lo que da lugar a diversos ataques. Además, para validar las muestras biométricas de huellas dactilares se necesita una potencia de computación adicional. Una base conceptual para la técnica 2FA (verificación de dos factores) mezcla elementos de verificación de contraseñas (basada en el conocimiento) y biométrica (dinámica de pulsación de teclas) [7]. Una solución de autenticación multifactorial (MFA) basada en el polinomio de Lagrange invertido, como expansión de la función de compartición de secretos de Shamir, aborda las situaciones de verificación de la identidad incluso si algunas de las partes están desalineadas o ausentes. También ayuda en la calificación de los elementos que faltan sin revelar información sensible al validador; por lo tanto, cuando un usuario pierde o sigue olvidando sus claves 2F, una asignación apropiada está lista

para ayudar con la autenticación mediante el envío de una información privada al usuario [17]. Para completar los pasos de la MFA, la solución propuesta se concibe explícitamente, por lo que su gestión para 2FA y SFA no es aclamada. Añadir un factor de tiempo aleatorio no es capaz de proporcionar un nivel útil de protección de datos biométricos, ya que el espía podría ser capaz de recuperar instantáneamente el secreto del factor.

En este artículo se tratará de abordar un modelo de autenticación de contraseña de dos factores (TFA), en la que el usuario U se autentica ante el servidor S "demostrando la posesión" de un dispositivo personal auxiliar D (por ejemplo, un smartphone o un token USB) además de conocer su contraseña, constituye una defensa común contra los ataques de contraseña en línea, así como una segunda línea de defensa en caso de fuga de contraseñas. Un esquema TFA que utiliza un dispositivo que no está directamente conectado al terminal cliente C de U suele funcionar de la siguiente manera: D muestra un breve PIN secreto de un solo uso, recibido de S (por ejemplo, mediante un mensaje SMS) o calculado por D basándose en una clave compartida con S, y el usuario teclea manualmente el PIN en el cliente C además de su contraseña. Ejemplos de sistemas basados en PINs de un solo uso incluyen PINs basados en SMS, TOTP [10], HOTP [14], Google Authenticator [4], FIDO U2F [2], y esquemas en la literatura como [3].

Materiales y métodos o Metodología computacional

Estado del arte:
Propuesta de esquema seguro de autenticación multifactor en la nube. Nuestro sistema propuesto tiene una estructura modular que permite analizar cada riesgo y su respuesta por separado. -is facilita la gestión del sistema en la nube y permite a los administradores y usuarios integrar soluciones especializadas para combatir los riesgos. El sistema en nube tiene dos tipos de entidades: el servidor en nube y el usuario en nube. El procedimiento de autenticación propuesto consta de las dos fases siguientes: fase de registro y fase de inicio de sesión

Fase de registro.

Esta fase consta además de tres pasos. Los usuarios de la nube se registran en el servidor de la nube para utilizar los servicios prestados por el servidor de la nube con los tres pasos principales siguientes:
Primer paso. En el primer paso, el usuario de la nube se registra con su ID de usuario, ID de correo electrónico y número de móvil; el servidor valida y verifica toda la información proporcionada y envía el correo electrónico y el número de móvil OTP para validar al cliente.
(1) Cliente: el cliente envía su nombre de usuario, correo electrónico y número de móvil al servidor en la nube.

(2) Servidor: el servidor almacena la información recibida y envía la OTP por correo electrónico y SMS al cliente.
(3) Cliente: el cliente almacena y recibe la OTP por correo electrónico y SMS del servidor.

Segundo paso. En el segundo paso, el cliente introduce las OTP válidas y recibe la clave del servidor para seguir comunicándose de forma segura con el servidor y otra información útil como la contraseña.
(1) Cliente: el cliente introduce la OTP por correo electrónico y la OTP por SMS en el servidor en la nube.
(2) Servidor: el servidor verifica la OTP enviada y genera el par de claves EC.
(3) Servidor: el servidor envía la clave pública EC generada al cliente para una comunicación segura.
(4) Cliente: el cliente recibe la clave pública EC del servidor.

Tercer paso. En el tercer paso, el cliente elige la contraseña, el tipo de servicio y la duración del servicio y envía toda la información de forma segura utilizando la clave compartida del servidor con las siguientes medidas de seguridad avanzadas:
(1) El cliente genera y guarda la contraseña segura salada utilizando PBDKF2, como se muestra en la Figura 1. Enccs_pk {H(UserID) ||H(SecureSaltedPassword) || nonceX}
(2) Servidor: el servidor recibe la contraseña segura y los datos de suscripción, servicio y duración.
A continuación, el servidor almacena toda la información relativa a la identificación del usuario en la base de datos del servidor de forma segura y genera un certificado en la nube que contiene la identificación del usuario, la suscripción y la duración, que se enviará al usuario en formato cifrado. Servidor: el servidor cifra el certificado de suscripción utilizando su clave privada.

Inicio de sesión y autenticación tiene las siguientes dos capas o autenticación de dos factores para verificar la identidad del usuario

Autenticación de primer factor en la nube. Durante la primera fase, el usuario solicita el primer factor de autenticación. -e solicitud es recibida y procesada por la plataforma en nube. -La plataforma en nube consta de un servidor de base de datos y un servidor web. -El servidor de la base de datos autentica las credenciales electrónicas comparándolas con los registros ya registrados y, tras la autenticación, se envía una confirmación al servidor en la nube para comunicar al usuario que la autenticación se ha realizado correctamente. Aquí, el usuario de la nube proporciona el ID de usuario y la contraseña al servidor de la nube. EncMsg1 Enccs_pk {H(UserID) ||H (Password)}

(i) El servidor en nube recibe el mensaje cifrado (EncMsg1) y primero lo descifra y luego verifica la firma digital del usuario.

(ii) (ii) Si se verifica el paso anterior, el servidor en nube envía la OTP al correo electrónico registrado (OTP1) y al móvil (OTP2) y espera al segundo paso de autenticación.

Autenticación de segundo factor en la nube. Una vez verificado con éxito el 1FA, se pide al usuario que verifique el segundo factor a través de una aplicación autenticadora. Al recibir la solicitud, el servidor en la nube reverifica el primer factor y envía la confirmación o el rechazo a la nube para que la procese. Tras la reverificación satisfactoria del 1FA, la nube envía la solicitud para verificar el segundo factor y envía una solicitud OTP al usuario para verificar el dispositivo. Tras una autenticación correcta, se concede al usuario acceso a la nube. El usuario de la nube pasa por un procedimiento de autenticación multifactor en el que, en el primer proceso, envía de forma segura las credenciales tradicionales, como el ID de usuario y una palabra clave con sal fuerte, al servidor de la nube para que se verifiquen[7]; si se verifican, el servidor pide al usuario que envíe otro factor de autenticación, que es el certificado ya proporcionado por el servidor de la nube como su certificado de identidad y la OTP en el correo electrónico y el móvil. -De este modo, se envían los tres elementos siguientes:

(1) certificado en la nube;
(2) OTP en el correo electrónico;
(3) OTP en el móvil. (Figura 4) EncMsg2 Enccs_pk { H(CloudCertificate) ||H (OTP2) || H (OTP2) || nonceY}
(i) Si se verifican los tres factores anteriores, el usuario estará autenticado y podrá utilizar los servicios en la nube.

Fundamentación Teórica

Autentificación de Doble Factor

La autenticación de dos factores proporciona una capa secundaria de seguridad que hace que sea más difícil para los piratas informáticos acceder a los dispositivos y las cuentas en línea de una persona para robar información personal. Con la autenticación de dos factores habilitada, incluso si el pirata informático conoce la contraseña de su víctima, la autenticación seguirá fallando y evitará el acceso no autorizado. Además, también proporciona a las organizaciones un nivel adicional de control de acceso a sistemas sensibles y datos y cuentas en línea, protegiendo esos datos de ser comprometidos por piratas informáticos armados con contraseñas de usuario robadas.[24]

Inteligencia Artificial

La inteligencia artificial (IA) es un conjunto de algoritmos (reglas que definen con precisión un conjunto de operaciones) que permiten realizar cálculos para percibir, razonar y actuar. La IA es usada para llevar a cabo la realización de múltiples tareas, pero puede usarse para brindar mejoras a la inteligencia humana.[8]

Machine learning

Machine learning es definido como aprendizaje automático, y su uso está enfocado en el análisis masivo de datos [19]. Dentro de sus algoritmos más usados tenemos: SVM, BOSQUE ALEATORIO, Árbol de decisiones KNN y Adaboost clasificadores [20].

Deep Learning

Definido como aprendizaje profundo, ofrece una estrategia de optimización global. Dentro de sus usos tenemos: Procesamiento de información, reducción de ruido en imágenes [22], procesamiento del lenguaje natural [21]
Además, el aprendizaje profundo es una rama derivada del aprendizaje automático (Machine learning) [23].

Seguridad de la información

Es la disciplina que, con base en políticas y normas internas y externas de la empresa, se encarga de proteger la integridad y privacidad de la información almacenada en un sistema informático, contra cualquier tipo de amenaza, minimizando los riesgos tanto físicos como físicos. lógica, a la que está expuesto.[18]

Control de Accesos

Es implementado como un método de seguridad para delimitar un conjunto de usuarios autorizados para acceder a una determinada información [20].

Herramientas y elementos

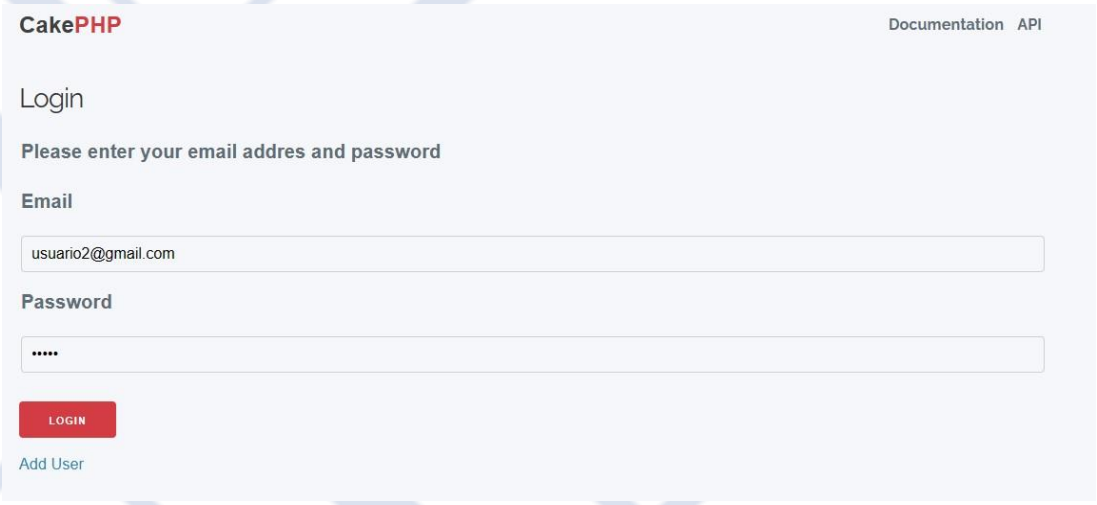
Visual Studio Code Visual Studio Code es definido como una plataforma de código abierto, un editor de código de multiplataforma que pertenece a Microsoft y proporciona todos los componentes necesarios de un IDE como: IntelliSense, depuración, control de versiones, creación de plantillas y API de extensiones que brindan muchas facilidades a los desarrolladores.[11]

Cakephp 4.0 Proporciona una estructura organizativa básica que cubre los nombres de las clases, archivos, tablas de base de datos y otras convenciones más. Aunque lleva algo de tiempo aprender las convenciones, siguiéndolas CakePHP evitará que tengas que hacer configuraciones innecesarias y hará que la estructura de la aplicación sea uniforme y que el trabajo con varios proyectos sea sencillo. El capítulo de convenciones muestra las que son utilizadas en CakePHP.[12]

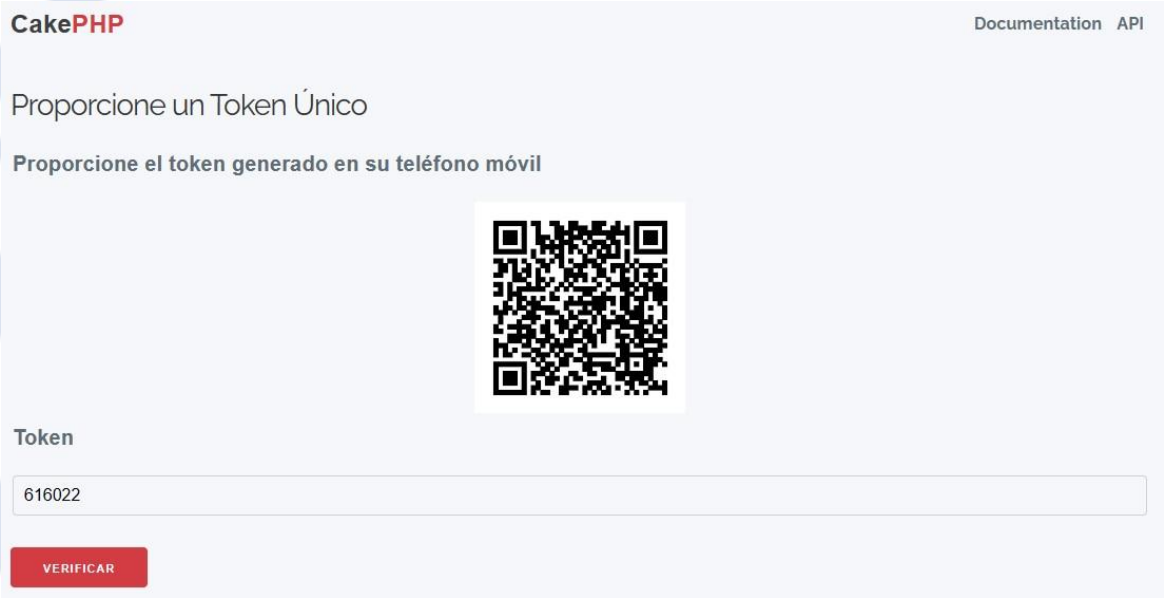
Encuesta nos ayudara a poder sacar resultados con los usuarios que se comprometieron a testear este modelo y poder analizar si es eficiente su aplicación en varios sistemas o en la nube, además si es rentable, para las empresas corporativas.

Uso del modelo de doble factor

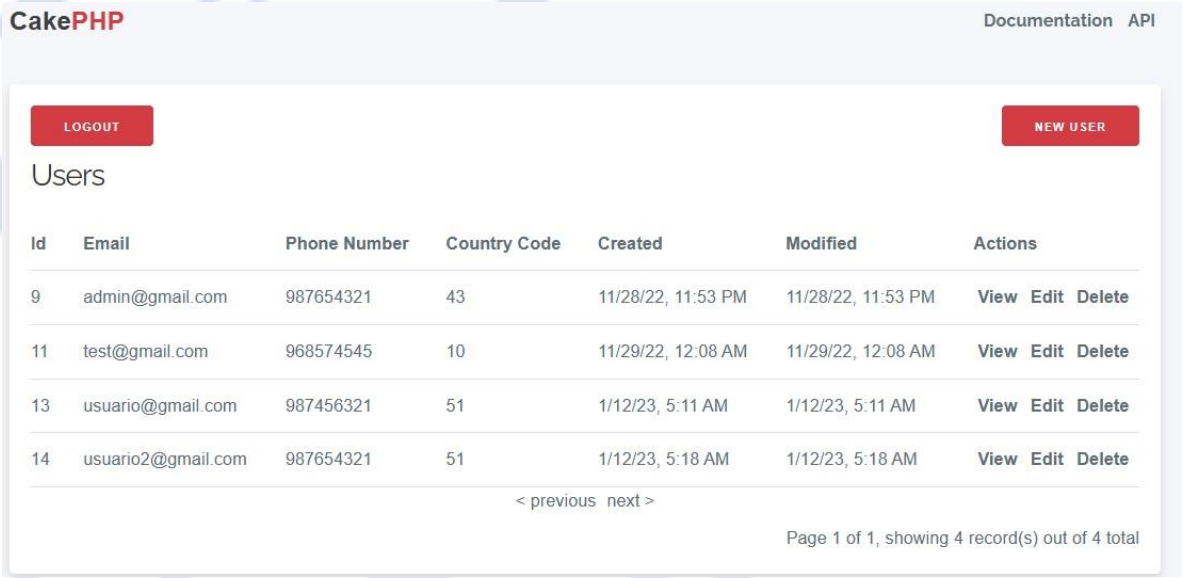
Ingresamos nuestro usuario para poder acceder al sistema



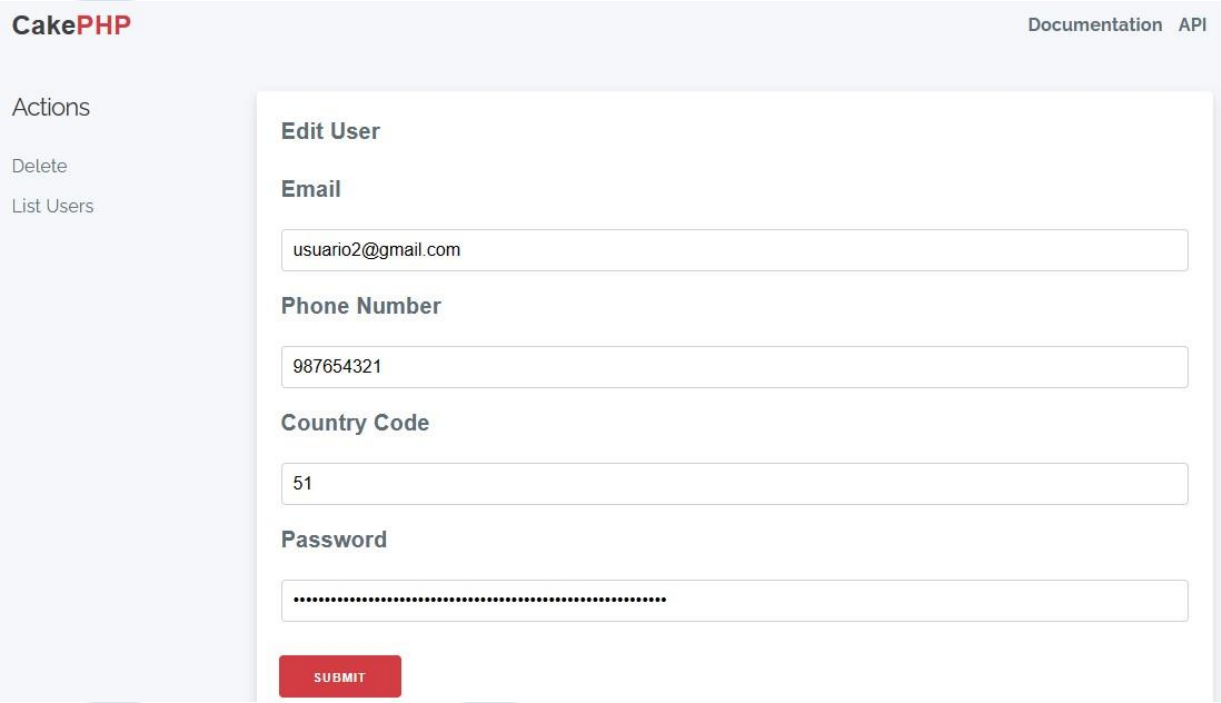
Cuando digitamos correctamente los datos del usuario el siguiente paso es digitar el token o código enviado un numero móvil para poder validar o también podrías escanear el qr con nuestro móvil vinculado, es decir tratar de usar el modelo de doble autentificación.



Aquí vemos los usuarios vinculados a dicho sistema, la cuales podrán usar el método de doble autentificación.



Además, el usuario podrá editar sus datos y en especial su número móvil, la cual es usada para usar el modelo de doble autenticación, esto es algo fundamental, mas cuando uno es propenso a perder el móvil o que será robado.



Resultados y discusión

¿Le ha resultado útil la autenticación doble factor? (protección doble).

Entre los encuestados respondieron sobre la pregunta de le ha resultado útil la autenticación doble factor, en un 77-1% afirman que les pareció útil y en un 29-9%, que no están útil. En concreto, el 51.6% de la muestra señalaba la seguridad como su razón principal. Esto puede deberse a que la mayor parte de los encuestados no están muy familiarizados con la doble autenticación.

Utiliza un sistema de seguridad en computadora de escritorio (PC).

En la utilización de un sistema de seguridad en computadora de escritorio (PC), los encuestados respondieron en un 48.6% que, si utilizan, en un 11.8% que no utilizan y en un 39-3%, que no tienen PC. Según Arellano y Peralta (2014), un 42,7% de las empresas que utiliza Internet no

cuenta ni utiliza instalaciones o procedimientos internos de seguridad, coincidiendo aproximadamente con los porcentajes encontrados en el trabajo.

Utiliza un sistema de seguridad en Smartphone (Android, IOs).

En la utilización de un sistema de seguridad en smartphone (Android, IOs), respondieron los encuestados en un 54.9% que, si utilizan, en un 27-2% que no utilizan y en un 17-6%, que no tienen Smartphone. En relación a la pregunta no se encontró investigaciones realizadas.

Sabe lo que es el Phishing.

La respuesta de los encuestados sobre si sabe lo que es el Phishing, solo en un 19-1% conocen y en un 80.9% que no conocen. En un 46% de los encuestados afirmó haber recibido un mensaje fraudulento que afirmaba provenir de servicios de correo electrónico como Yahoo!, Microsoft y Gmail. Las siguen las redes sociales con un 45%, los bancos 44% y tiendas en línea 37%.

Sabe la diferencia entre hacker y cracker.

Los encuestados respondieron sobre la pregunta, si sabe la diferencia entre hacker y cracker, en un 59-4% no conocen y en un 40.6% que conoce.

Con qué frecuencia realiza usted copias de seguridad.

La respuesta de los encuestados sobre con qué frecuencia realiza usted copias de seguridad, en un 33-5% hace copias Anualmente, en un 15-3% hace copias Diariamente, en un 26.3% hace copias Mensualmente, en un 14.2% hace copias Semanalmente y en un 10.4% hace copias Trimestralmente.

Sabe lo que es una Dirección IP

La respuesta de los encuestados sobre si sabe lo que es una dirección IP, en un 37-7% no conocen y en un 62.3% que si conocen.

Sabe lo que es una dirección MAC.

Los encuestados respondieron sobre la pregunta, si sabe lo que es una dirección MAC, en un 77-4% no sabe y en un 22.6% que si saben lo que es una dirección MAC.

Sabe lo que significa el protocolo https.

La respuesta de los encuestados sobre si Sabe lo que significa el protocolo https, mencionan que en un 52.5% no sabe y en un 47-5% que sabe lo que significa el protocolo https.

¿Sabe lo que es una VPN?

Entre los que saben lo que es una VPN, están en un 39-7% y entre los que no saben están en un 60.3%.

Cómo considera la seguridad para hacer pagos a través de Internet.

Entre los encuestados que respondieron con respecto a cómo considera usted la seguridad para hacer pagos a través de Internet, se observa que, el 3-8% indica que es muy seguro, el 19-9% menciona que es nada seguro, el 47.1% indica que es poco seguro y el 28.9% indica que es seguro.

Utiliza autentificador de seguridad

Entre los encuestados que respondieron con respecto a utiliza autentificador de seguridad, se observa que, el 3-5% indica que es sí, el 93-95% menciona que no, ya que o desconocen su uso o no tienen tiempo que perder.

Conclusiones

Entre los encuestados respondieron sobre la pregunta de le ha resultado útil la autenticación doble factor, en un 77-1 % afirman que les ha resultado util; En la utilización de un sistema de seguridad en computadora de escritorio (PC), los encuestados respondieron en un 48.6% que si utilizan. En la utilización de un sistema de seguridad en smartphone (Android, IOs), en un 54.9% que si utilizan. La respuesta de los encuestados sobre si sabe lo que es el Phishing, solo en un 19-1% conocen. Los encuestados respondieron sobre la pregunta, si sabe la diferencia entre hacker y cracker, en un 40.6% que conoce. La respuesta de los encuestados sobre con qué frecuencia realiza usted copias de seguridad, en un 33-5% hace copias Anualmente, los demás en menor porcentaje. La respuesta de los encuestados sobre si sabe lo que es una dirección IP, en un 62.3% que conocen.

La respuesta de los encuestados sobre si Sabe lo que significa el protocolo https, mencionan que en un 47-5% que sabe lo que significa el protocolo https. Entre los encuestados que respondieron con respecto a cómo considera usted la seguridad para hacer pagos a través de Internet, el 47-1 % indica que es poco seguro, entre los de más porcentaje. El Género Masculino obtuvo 8.0%

de conocimiento de seguridad informática. Los promedios de Ocupación con referente a la seguridad informática. El conocimiento de seguridad informática en la ocupación el que más porcentaje obtuvo es la ocupación Profesional independiente con 46%. El Género Masculino alcanzó 2 % que utiliza autentificador de seguridad. Y de estos el 70-75% son población laboral activa.

Referencias

[1] "Facial Expression Recognition Using Machine Learning Techniques", *International Journal of Advance Engineering and Research Development*, vol. 1, n.º 06, junio de 2020. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.21090/ijaerd.010633

[2] FIDO Universal 2nd Factor. https://www.yubico.com/

[3] Shirvanian, M., Jarecki, S., Saxena, N., Nathan, N.: Two-factor authentication resilient to server compromise using mix-bandwidth devices. In: Network & Distributed System Security Symposium (2014)

[4] Google Authenticator Android app. https://goo.gl/Q4LU7k

[5] "OpenCV: OpenCV modules". OpenCV documentation index. https://docs.opencv.org/4.x/ (accedido el 17 de noviembre de 2022).

[6] S. Nagaraju and L. Parthiban, "Trusted framework for onlinebanking in public cloud using multi-factor authentication and privacy protection gateway," Journal of Cloud Computing, vol. 4, no. 1, pp. 1–23, 2015.

[7] M. Olalere, M. Taufik Abdullah, R. Mahmod, and A. Abdullah, "Bring your own device: security challenges and A theoretical framework for two-factor Authentication," Inaternational Journal of Computer Networks and Communications Security, vol. 4, no. 1, pp. 21–32, 2016, http://www.ijcncs.org.

[8] O. Niel y P. Bastard, "Artificial Intelligence in Nephrology: Core Concepts, Clinical Applications, and Perspectives", *American Journal of Kidney Diseases*, vol. 74, n.º 6, pp. 803–810, diciembre de 2019. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1053/j.ajkd.2019.05.020

[9] T. Walsh, "The troubling future for facial recognition software", *Communications of the ACM*, vol. 65, n.º 3, pp. 35–36, marzo de 2022. Accedido el 15 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1145/3474096

[10] TOTP: Time-Based One-Time Password Algorithm. https://goo.gl/9Ba5hv

[11] S. Latifi, Ed., *17th International Conference on Information Technology–New Generations (ITNG 2020)*. Cham: Springer International Publishing, 2020. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1007/978-3-030-43020-7

[12] S. K. Shammi, S. Sultana, M. S. Islam y A. Chakrabarty, "Low Latency Image Processing of Transportation System Using Parallel Processing co-incident Multithreading (PPcM)", en *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Kitakyushu, Japan, 25–29 de junio de 2018. IEEE, 2018. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1109/iciev.2018.8640957

[13] Kaur, S., Kaur, G., & Shabaz, M. (2022). A Secure Two-Factor Authentication Framework in Cloud Computing. *Security and Communication Networks*, *2022*. https://doi.org/10.1155/2022/7540891

[14] RFC 4226 HOTP: An HMAC-based One-Time Password Algorithm (2005). https://goo.gl/wxHBvT

[15] I. Sluganovic, M. Roeschlin, K. B. Rasmussen y I. Martinovic, "Analysis of Reflexive Eye Movements for Fast Replay-Resistant Biometric Authentication", *ACM Transactions on Privacy and Security*, vol. 22, n.º 1, pp. 1–30, enero de 2019. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1145/3281745

[16] L. Monastyrskii, V. Lozynskii, Y. Boyko y B. Sokolovskii, "Fingerprint recognition in inexpensive biometric system", *Electronics and Information Technologies*, vol. 9, 2018. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.30970/eli.9.120

[17] A. Ometov, S. Bezzateev, N. Makitalo, S. Andreev, ¨T. Mikkonen, and Y. Koucheryavy, "Multi-factor authentication: a survey," Cryptography, vol. 2, no. 1, pp. 1–31, 2018

[18] H. AYDIN, "The Importance of Cyber Security in Management Information Systems (MIS)", *Bilgisayar Bilimleri ve Teknolojileri Dergisi*, octubre de 2022. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.54047/bibted.1138252

[19] M. Meroni, F. Waldner, L. Seguini, H. Kerdiles y F. Rembold, "Yield forecasting with machine learning and small data: What gains for grains?", *Agricultural and Forest Meteorology*, vol.

308-309, p. 108555, octubre de 2021. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1016/j.agrformet.2021.108555

[20] M. S. Bouzakraoui, A. Sadiq y A. Y. Alaoui, "Customer Satisfaction Recognition Based on Facial Expression and Machine Learning Techniques", *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, n.º 4, p. 594, agosto de 2020. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.25046/aj050470

[21] A. Bazaga, N. Gunwant y G. Micklem, "Translating synthetic natural language to database queries with a polyglot deep learning framework", *Scientific Reports*, vol. 11, n.º 1, septiembre de 2021. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1038/s41598-021-98019-3

[22] B. Liu *et al.*, "Unsupervised Deep Learning for Random Noise Attenuation of Seismic Data", *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2021. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1109/lgrs.2021.3057631

[23] J. Han, E. Shihab, Z. Wan, S. Deng y X. Xia, "What do Programmers Discuss about Deep Learning Frameworks", *Empirical Software Engineering*, vol. 25, n.º 4, pp. 2694–2747, abril de 2020. Accedido el 17 de noviembre de 2022. [En línea]. Disponible: https://doi.org/10.1007/s10664-020-09819-6

[24] Kaur, S., Kaur, G., & Shabaz, M. (2022). A Secure Two-Factor Authentication Framework in Cloud Computing. *Security and Communication Networks*, *2022*. https://doi.org/10.1155/2022/7540891

Política informática y la gestión de la seguridad de la información en base a la norma ISO 27001

IT policy and information security management based on ISO 27001

Roy Guiller Ramos Mamami
Universidad Nacional Jorge Basadre Grohmann, Lima, Perú.
roy.ramos@unjbg.edu.pe
https://orcid.org/0000-0001-7208-4977

Rogelio Cahuaya Ancco
Universidad Nacional Jorge Basadre Grohmann, Lima, Perú.
rogelio.cahuaya@unjbg.edu.pe
https://orcid.org/0000-0003-1350-688X

Roberto René Llanqui Argollo
Universidad Nacional Jorge Basadre Grohmann, Lima, Perú.
roberto.llanqui@unjbg.edu.pe
https://orcid.org/0000-0003-1871-5066

ARK: ark:/42411/s11/a57
PURL: 42411/s11/a57

RECIBIDO 15/11/2022 • ACEPTADO 27/12/2022 • PUBLICADO 30/03/2023

RESUMEN

La tecnología en los últimos tiempos ha ido cambiando a pasos agigantados y es natural aclimatarse a estos cambios tecnológicos y más aún por la pandemia covid-19 en el mundo, provocando el cambio drástico en las estrategias de las empresas. El artículo que se presenta a continuación pretende describir las causas y efectos de la norma ISO 27000 enfocándose a la norma ISO 27001. En este sentido para asegurar una adecuada política de seguridad de la información que, ante la falta de una legislación nacional sobre el tema, debe basarse en los estándares internacionales, el derecho comparado y autonomía de la voluntad. La metodología empleada para explicar las diversas áreas de impacto es la seguida por la norma ISO 27001 en el dominio que hace referencia al cumplimiento y que comprende: La norma ISO 27001 auxilia a las empresas en el cumplimiento de los requisitos legales, los cuales, tienen la finalidad de eludir la vulneración de la legislación o el incumplimiento de toda obligación legal de las entidades de cualquier requisito de seguridad.

Palabras claves: TIC's (Política Informática), cumplimiento, protección de datos, seguridad, propiedad intelectual.

ABSTRACT

Technology in recent times has been changing by leaps and bounds and it is natural to acclimatize to these technological changes and even more so due to the covid-19 pandemic in the world, causing a drastic change in company strategies. The article presented below aims to describe the causes and effects of the ISO 27000 standard, focusing on the ISO 27001 standard.

Keywords: *TIC's (Computer Policy), compliance, data protection, security, intellectual property.*

INTRODUCCIÓN

Antes de la pandemia COVID-19 las herramientas digitales en las empresas eran muy pocas, a diferencia de otras organizaciones con mayor proporción que, si tienen el uso de tecnologías adecuadas norma ISO, pero cabe señalar que la gran mayoría de empresas informales y formales se limitan a tan solo a contar con un correo electrónico y tener una línea de internet básica. Las organizaciones enfrentan un desafío que es la transformación digital, para hacer frente a este contexto así mismo es necesario el uso de las herramientas tecnológicas de conectividad, conocimientos y de gestión. Pero este reto de adopción tecnológica trae altos costos de hardware, software.

La política de información es relevante e importante en estrategias y tomas de decisiones, para una organización, abarcando el área de las tecnologías de la información con el fin de asegurar de la información y datos. Identificando las reglas y procedimientos que deberán cumplir los usuarios al utilizar los recursos informáticos en una organización o empresa, su cumplimiento es fundamental para lograr los objetivos trazados.

El impacto de las Tecnologías de la Información y las Comunicaciones TIC no es ajeno al Derecho, por el contrario, cada día los avances de la tecnología imponen mayores retos a los operadores jurídicos, a los cuales hay que responder desde la legislación nacional si ésta existe, la legislación internacional, el derecho comparado, la autonomía de la voluntad privada, las mejores prácticas existentes en la industria y las normas que permitan dar un tratamiento uniforme a problemáticas que experimentan las organizaciones, cualquiera que sea la latitud en que estén ubicadas.

El desarrollo en nuestro país de normas jurídicas que respondan a los problemas que surgen del fenómeno de las TIC´s es mínimo. La Ley 527 de 1999 constituye uno de los pocos desarrollos importantes en este sentido. Esta situación genera un grado importante de inseguridad e incertidumbre no sólo para las organizaciones, sino para también los ciudadanos, en su condición de usuarios, consumidores y titulares de datos personales.

La información se ha convertido no sólo en un activo valioso, sino también estratégico en las organizaciones. La información puede ser protegida de muchas maneras. Desde el Derecho pudiera pensarse que se logra contar con un adecuado nivel de protección, con la encriptación, teniendo en cuenta que la mayor de las veces la compresión del tema tecnológico es poca; sin embargo, la encriptación es un mecanismo para otorgar a la información atributos de confidencialidad, integridad, autenticidad, y dependiendo del mecanismo de encriptación, podría reputarse el no repudio. En la protección de la información intervienen diferentes disciplinas, desde la informática, la gerencial, la logística, la matemática hasta la jurídica, entre muchas otras.

Así pues, el punto de partida de este estudio será acudir a los conceptos de Información y Seguridad, para lo cual se tendrá en cuenta las definiciones otorgadas por el Diccionario de la Real Academia de la Lengua Española, ello con el fin de partir de conceptos básicos.

El objetivo de este artículo es analizar parte legal como en la gestión de la protección de la información, de esta manera se puede concluir el valor que tiene la información como resultado de un conocimiento especializado en un área determinada y su seguridad, que a su vez requiere de ciertos mecanismos para garantizar su buen funcionamiento, en aras de protegerlo y asegurar su permanencia frente a los actos violentos que se pueden perpetrar contra la información. Anteriormente la seguridad de la información estaba entendida como la aplicación de un conjunto de medidas de orden físico y lógico a los sistemas de información, para evitar la pérdida de la misma, siendo ésta una tarea de responsabilidad exclusiva de los departamentos de informática de las organizaciones.

Análisis Teórico

Importancia de la seguridad de la información

La seguridad informática ha hecho tránsito de un esquema caracterizado por la implantación de herramientas de software, que neutralicen el acceso ilegal y los ataques a los sistemas de información, hacia un modelo de gestión de la seguridad de la información en el que prima lo dinámico sobre lo estacional.

Para lograr niveles adecuados de seguridad se requiere el concurso e iteración de las disciplinas que tengan un impacto en el logro de este cometido, teniendo siempre presente que un sistema de gestión no garantiza la desaparición de los riesgos que se ciernen con mayor intensidad sobre la información.

Entonces, el problema es determinar cómo desde una disciplina como el Derecho se contribuye a la gestión de la seguridad de la información. Los enfoques de intervención jurídica podrían ser muchos; de hecho, no existe limitación alguna, para que una organización adopte las medidas

que considere pertinentes con el fin de neutralizar un riesgo. La ISO 27 001 es una herramienta de gestión estratégica que conduce a lograr la protección de la información, bien en un contexto en el cual la empresa pretenda alcanzar una certificación, o bien que sólo pretenda incorporar buenas prácticas de seguridad de la información, no sólo en sus procesos internos, sino también en sus procesos externos.

Métodos y Metodología computacional

La norma consagra un conjunto significativo de dominios que pretenden establecer un ciclo de seguridad lo más completo posible, advirtiendo que no todos ellos tienen impacto jurídico. Desde ya es importante mencionar que el enfoque que se propone se alimenta tanto de normatividad nacional como internacional, así como de otras fuentes del Derecho, en razón de la escasa legislación que existe.

La norma ISO 27 001:2013, contempla 14 dominios:

1. Políticas de seguridad de la información.
2. Organización de la seguridad de la información.
3. Seguridad de los recursos humanos.
4. Gestión de activos.
5. Controles de acceso.
6. Criptografía – Cifrado y gestión de claves.
7. Seguridad física y ambiental.
8. Seguridad operacional.
9. Seguridad de las comunicaciones.
10. Adquisición, desarrollo y mantenimiento del sistema.
11. Gestión de incidentes de seguridad de la información.
12. Cumplimiento.

Cumplimiento

El último capítulo del anexo A de la norma ISO 27001 está dedicado a controles que nos permitan garantizar el cumplimiento con las políticas, normas y legislación aplicable enfocándose principalmente en lo que se refiere a seguridad de la información.

Objetivo

- Cumplimiento de los requisitos y contractuales.
- Evitar los incumplimientos de las obligaciones legales, estatutarias, reglamentarias o contractuales relacionadas con la seguridad de la información y con los requisitos de seguridad.
- Revisiones de seguridad de la información.
- Garantizar que la seguridad de la información es implementada y operada de acuerdo con las políticas y procedimientos organizacionales.

Identificación de la legislación aplicable y de los requisitos contractuales

El control establece la necesidad de identificar de forma documentada todos los requisitos legales y contractuales que afecten a la organización, además de mantenerlos actualizados. Las leyes y reglamentos que afectan a una actividad son algo que evidentemente cambia con el tiempo y pueden ser distintas en cuanto a:

- Leyes o reglamentos sectoriales que afectan a una actividad.
- Leyes del parlamento europeo o leyes locales.
- Requisitos legales aplicables a tipos de información por su clasificación.

En la actualidad dado el crecimiento de los incidentes relacionados con la seguridad y la magnitud de su impacto han hecho que los gobiernos de todo el mundo sean conscientes de la necesidad de proteger a las personas y las empresas contra la gestión inadecuada de la información confidencial.

Derechos de propiedad intelectual

Este control nos requiere que se establezcan procedimientos que garanticen el uso del software de acuerdo a los términos previstos en la Ley de Propiedad.

Para ello se establecen los siguientes puntos a tener en cuenta para cumplir con este control:

- Se dispone de una política de uso legal de productos de software.
- Se asegura la no violación de derechos de copia.
- Dónde se compra los productos de software.
- Se mantiene la política de licencias del software comprado.
- Controlar el número máximo de usuarios por licencia.

- Revisar periódicamente que se estén utilizando solamente productos software con licencia.
- Cumplir con los derechos de copia de material audiovisual, libres e informes.
- Comunicar al personal la política de uso legar de software aclarando que cosas están permitidas y cuáles no.
- Advertir al personal sobre las consecuencias de la violación de las políticas de uso legal de software estableciendo las medidas disciplinarias oportunas.
- Identificar (listado de activos) los activos de información que estén afectados por derechos de propiedad intelectual.
- Mantener la documentación que justifique o acredite la propiedad de las licencias (discos, manuales, etc.).

Protección de los registros

Este control nos pide mantener un análisis de los requisitos contractuales legales en cuanto a las obligaciones de debido control sobre la protección de los registros en cuanto a evitar su pérdida, falsificación o acceso no autorizado.

Se clasifican los registros de información y aplica los controles necesarios según los requisitos legales:

- Registros contables.
- Bases de datos.
- Bases de datos de transacciones.
- Registros de auditoría (propios del sistema de Gestión de la Seguridad de la Información).
- Procedimientos operativos.
- Registros documentales en papel, microfichas, archivos electrónicos.
- Archivos cifrados (contraseñas, firmas digitales).

Para cumplir con la protección de los registros se nos requiere revisar el cumplimiento con:

- La definición y publicación de las directrices sobre la retención, almacenamiento, tratamiento y eliminación de os registros y la información.
- Mantenimiento de un calendario de retenciones donde se identifique los registros y los períodos de tiempo que deberían retenerse.
- Mantenga un inventario de los registros de información clave o critica.

Protección de los datos y privacidad de la información personal

Este control nos requiere el establecimiento de controles para el cumplimiento de la legislación en materia de cumplimiento con la legislación vigente en materia de protección de datos personales.

Regulación de los controles criptográficos

En caso de utilizar mecanismos de cifrado deben tenerse en cuentas las normativas sobre uso de controles criptográficos vigentes. Deberemos tener en cuenta leyes como:

- La Ley General de Telecomunicaciones.
- Ley de Firma Electrónica.

Limitaciones en el uso de medios criptográficos

En muchos países existen limitaciones en el uso de medios criptográficos por lo que tendremos que tener en cuenta estas limitaciones o restricciones en las importaciones o exportaciones de Hardware y/o Software para funciones criptográficas

También deberemos considerar si existen métodos obligatorios de cifrado de información y cumplir los requisitos legales del cifrado de información establecidos por los reglamentos de cada país.

Códigos de conducta

Determinadas organizaciones y organismos pueden elaborar códigos de conducta sobre el tratamiento de datos personales. En este caso si nos adherimos a dichos códigos deberemos cumplir con los requisitos en cuanto al uso y obligatoriedad de los sistemas de cifrado

Evaluación de riesgos

La evaluación de riesgos, también obligatoria en el reglamento RGPD sobre el tratamiento de datos personales puede determinar la necesidad de utilizar cifrado de datos como resultado de un control necesario para mitigar o evitar un riesgo para la seguridad de la información.

Cifrado legal

En el caso que nos veamos obligados a cifrar datos según lo expuesto hasta ahora deberemos dedicar los recursos necesarios para cumplir con estos requisitos. Sin embargo, cualquier sistema de cifrado no es suficiente ya que los archivos PDF o archivos comprimidos ZIP con clave no son considerados válidos para garantizar que la información no sea inteligible ni manipulada por terceros.

Revisión independiente de la seguridad de la información

Las revisiones deben ser llevadas a cabo por personal independiente al personal que es auditado. Aunque pueden ser llevadas a cabo por personal interno siempre de áreas o departamentos independientes al auditado, conviene que de forma regular se realicen auditorias de cumplimiento de la seguridad de la información por personal externo. Las auditorías realizadas por personal externo siempre podrán aportar beneficios como:

- Garantizar la independencia de las revisiones o auditorias.
- Aportar un punto de vista imparcial.
- Aportar la experiencia de profesionales de la seguridad de la información que conocen otras organizaciones y pueden aportar mejoras.

Cumplimiento de la política y las normas de seguridad

Este control nos pone como requisito la necesidad de que los responsables de cada área deben revisar que los procedimientos de la organización sean aplicados de acuerdo a los requisitos definidos.

Para ello los responsables deberían:

- Determinar la forma de revisar cómo se cumplen los requisitos de seguridad de la información definidos en las políticas, normas y en otras regulaciones aplicables.
- Tener en cuenta la implementación de sistemas de medición automática y herramientas de informes.

Cuando se identifican incumplimientos se deberá:

- Identificar las causas.
- Evaluar la necesidad de tomar medidas.
- Implementar las acciones correctivas apropiadas.
- Revisar la eficacia de las acciones correctivas.
- Identificar las deficiencias y debilidades del sistema.

Revisión del cumplimiento técnico

Para la evaluación de los sistemas de información debe revisarse periódicamente si están configurados correctamente de acuerdo a las reglas y políticas definidas:

- Identificar fallos en las actualizaciones de los sistemas.
- Establecer medidas correctivas antes de que estos fallos puedan suponer una amenaza real para el sistema.

Resultados y discusión

Cada cierto tiempo, la Seguridad de la Información basada en la ISO 27001 tiene que examinarse. Estas revisiones se realizan a través de diferentes políticas de seguridad y tiene que auditarse que las plataformas técnicas y los sistemas de información satisfagan la totalidad de normas de implementación de seguridad y controles de seguridad documentados que sean aplicables.

El control de que se está llevando a cabo el cumplimiento técnico únicamente tiene que estar revisado por los sujetos cualificados y además, estas personas tienen que estar autorizadas por la entidad, aunque puede pasar que lo realice otra persona bajo la supervisión del responsable. Se considera un elemento de reconocida importancia, por lo que hablamos del examen que tienen que pasar las entidades de sus sistemas operativos con el fin de asegurar que los softwares y los hardware han sido implantados perfectamente.

Conclusiones

La norma ISO 27001 auxilia a las empresas en el cumplimiento de los requisitos legales, los cuales, tienen la finalidad de eludir la vulneración de la legislación o el incumplimiento de toda obligación legal de las entidades de cualquier requisito de seguridad.

La política informática, como administrador de riesgos, se encarga de dotar de seguridad los diferentes activos de información de una organización; desde esa perspectiva se requiere una gestión jurídica permanente de los riesgos, amenazas y vulnerabilidades, como medio para adoptar las medidas y controles que disminuyan los mismos.

Las Tecnologías de la Información reclaman de la política informática respuestas innovadoras y globales respecto de los retos que le son intrínsecos; por tanto, los operadores jurídicos deben estar capacitados y entrenados para apoyar a la sociedad en la solución de las problemáticas propias de la política informática.

Referencias

[1] Marlon A. Di Luca. (2019). *Modelo para la gestión de la seguridad de la información y los riesgos asociados a su uso*. Modelo para la gestión de la seguridad de la información y los riesgos asociados a su uso (redalyc.org)

[2] Aplicación de ISO 27001 y su influencia en la seguridad de la información de una empresa privada peruana. http://dx.doi.org/10.20511/pyr2020.v8n3.786. [Accessed: Set, 2020].

[3] Diana L. Carvajal, A. Cardona, F. J. Valencia. (2020). *Una propuesta de gestión de la seguridad de la información aplicado a una entidad pública colombiana*. Una propuesta de gestión de la seguridad de la información aplicado a una entidad pública colombiana | Entre Ciencia e Ingeniería (ucp.edu.co)

[4] Erick G., Harold N., Jorge L. Díaz, J. Patiño. (2021). *Desarrollo de un sistema de gestión para la seguridad de la información basado en metodología de identificación y análisis de riesgo en bibliotecas universitarias*. Development of an information security management system based on analysis methodology and risk identification in university libraries (scielo.cl)

[5] Navira G. Angulo Murillo, María F. Zambrano Vera, G. García Murillo, F. Bolaños-Burgos. (2018). *Propuesta metodológica de seguridad de información para proveedores de servicios de internet en ecuador.* Microsoft Word - 165-176 (core.ac.uk)

[6] David A. Aguirre Mollehuanca. (2014). *Diseño de un Sistema de Gestión de Seguridad de la Información para Servicios Postales del Perú S.A*. http://hdl.handle.net/20.500.12404/5677

[7] Juan D. Aguirre Cardona, C. Aristizábal Betancourt. (2013). *Diseño del Sistema de Gestión de Seguridad de la Información para el grupo empresarial La Ofrenda*. https://hdl.handle.net/11059/4117

[8] Alexander, A. G., & Buitrago, L. J. (2007). *Diseño de un sistema de gestión de seguridad de información: Óptica ISO 27001: 2005*.

[9] Flores, L. C. A. (2013). *Diseño de un sistema de gestión de seguridad de información para un instituto educativo (Doctoral dissertation, Pontificia Universidad Católica del Perú, Facultad de Ciencias e Ingeniería. Mención: Ingeniería Informática).* http://hdl.handle.net/20.500.12404/4721

[10] Ampuero Chang, C. E. (2011). *Diseño de un sistema de gestión de seguridad de información para una compañía de seguros.* *http://hdl.handle.net/20.500.12404/933*

[11] Aráoz Severiche, I. (2020). *Implementación ISO/IEC 27001:2013: Un enfoque práctico (Spanish Edition) Edición Kindle.*

[12] Calder, A. (2017). *ISO27001/ISO27002: A Pocket Guide.*

[13] García, F. Albarrán, S. (2015). *Guía para Implantar un Sistema de Gestión de Seguridad de Información: Basada en la Norma ISO/IEC 27001 (Spanish Edition).*

[14] Fernández, C. Piattini, M. (2012). *Modelo para el gobierno de las TIC basado en las normas ISO*.

[15] Muñoz, C. (2002). *Auditoria en sistemas computacionales. Pearson Educación*.

Predicción de la clasificación ESRB para videojuegos según su contenido usando árboles de decisión

Predicting ESRB ratings for video games by content using decision trees

Rodrigo S. Huamán Maqque
Universidad Nacional de San Agustín.
Arequipa, Perú.
rhuamanma@unsa.edu.pe

Graciela Condori Anahua
Universidad Nacional de San Agustín.
Arequipa, Perú.
gcondoria@unsa.edu.pe

Fátima Gigi Rojas Carhuas
Universidad Nacional de San Agustín.
Arequipa, Perú.
frojasca@unsa.edu.pe

Rodolfo Robert Quispe Huacho
Universidad Nacional de San Agustín.
Arequipa, Perú.
rquispehuach@unsa.edu.pe

ARK: ark:/42411/s11/a83
PURL: 42411/s11/a83

RECIBIDO 18/12/2022 • ACEPTADO 03/02/2022 • PUBLICADO 30/03/2023

RESUMEN

Diversos estudios han comprobado que los niveles de violencia en los videojuegos pueden influir negativamente en el desarrollo de los niños, especialmente en la adolescencia y es por ello que se debe tener cuidado en que la clasificación sea la adecuada según el contenido presente. Para el análisis se utilizó la clasificación ESRB, que contiene 7 diferentes categorías, junto con la implementación de un modelo de árboles de decisión, que es una técnica de minería de datos capaz de representar gráficamente la relación entre las variables. Los resultados arrojaron que el nivel de precisión para un árbol de nivel 6 no supera el mínimo requerido.

Palabras claves: Árbol de decisión, ESRB, Minería de datos.

ABSTRACT

Various studies have proven that the levels of violence in video games can negatively influence the development of children, especially in adolescence and that is why care must be taken that the classification is appropriate according to the content present. For the analysis, the ESRB

classification was used, which contains 7 different categories, together with the implementation of a decision tree model, which is a data mining technique capable of graphically representing the relationship between the variables. The results showed that the precision level for a level 6 tree does not exceed the minimum required.

Keywords: *Decision tree, Data mining, ESRB.*

INTRODUCCIÓN

Los videojuegos se han convertido en un mercado crecimiento año a año, llegando a facturar más que la industria cinematográfica y de música a nivel global; desde su creación en 1962 por Steve Rusell, en el Massachusetts Institute of Technology inventando el Spacewar [1], ha ido mejorando en gráficos y mecánicas nuevas en busca de nuevos jugadores, pero también a todo su largo recorrido fue acompañado de polémicas por el tipo de contenido sin control que ofrecían.

Los videojuegos se definen [2] como programas informáticos que mantienen a los usuarios interactuando a través de imágenes que se muestran en dispositivos con pantallas de varios tamaños, a través de un incentivo implícito para ganar. Actualmente se están desarrollando videojuegos que pueden ser controlados por voz o movimiento, siendo un avance a como antes solo se podía controlar con los dedos u otros instrumentos adicionales (como guitarras, rifles y pistolas).

Los sistemas de clasificación previo a 1994 no existían, sin embargo, después del lanzamiento de algunos videojuegos en los años 90 como Funk, Buchman y principalmente por la polémica de Mortal kombat, causada por los elementos violentos que siempre tuvo la saga, se crea en Estados Unidos la Entertainment Software Rating Board (ESRB) con el fin de garantizar contenido adecuado a partir de diferentes parámetros [3].

Categorías de clasificación y su sugerencia de nivel de edad [4]:

- EC Early Childhood: Contenido apto para niños de tres a diez años, no contiene material inapropiado.
- E Everyone: Contenido apto para personas de seis años en adelante, contiene una mínima violencia, travesuras y lenguaje obsceno.
- E10+ Everyone 10+: Contenido apto para mayores de diez años. Puede contener mayor cantidad de violencia de caricatura o temas sugestivos.
- T Teen: Contenido apto para personas de trece años en adelante. Puede contener violencia, humor grosero y temas sugestivos con un mínimo de sangre.
- M Mature: Contenido apto para personas de diecisiete años en adelante. Puede contener temas sexuales y lenguaje o violencia más intensa con derramamiento de sangre.

- AO Adults Only: Contenido apto sólo para adultos de 18 años a más. Puede incluir contenido sexual gráfico o escenas prolongadas de violencia.
- RP: La clasificación se encuentra en espera.

Además de esta metodología de clasificación de videojuegos ERP existen otras relevantes como son:

- USK es la organización de clasificación de software utilizada en Alemania que viene a ser un tablero de clasificación bastante estricto, en parte debido a la paranoia del país entorno a los videojuegos con violencia, ya que existen distintos videojuegos que han sido prohibidos o han requerido que los desarrolladores editen o quiten ciertas funciones.
- SMECCV es una clasificación mexicana orientada a las especificaciones gráficas para las advertencias, descriptores de contenidos y elementos interactivos que deberán implementar objetos, respecto de los videojuegos distribuidos, comercializados o arrendados en el territorio nacional. Con el objetivo de crear parámetros específicos que apoyen a desarrollar una adecuada categorización tanto el contenido de los mismos, así como de las personas a la que va dirigido.
- CERO Computer Entertainment Rating Organization es una organización que clasifica la edad recomendada de los juegos en Japón, y fue fundada en 2002 como una división de "CESA".
- GSRR es el sistema de clasificación de contenido de videojuegos utilizado en Taiwán, Hong Kong y sureste de Asia. La ley realizada el 6 de julio de 2006 y se enmendó el 29 mayo 2012.

Los padres tienden a preocuparse por el contenido del juego en especial si son violentos, ya que podrían afectar la conducta del niño. Según Russell N. Laczniak, Les Carlson, Doug Walker, y E. Deanne Brocato los ESRB pueden ayudar a que los niños jueguen juegos menos violentos y asimismo reducir la participación de los niños en comportamientos negativos en la escuela [5], [6].

En una investigación sobre consecuencias de los videojuegos en niños y adolescentes en aspectos de la vida social, muestra como posibles efectos psicológicos y fisiológicos pueden sufrir los menores, donde ahonda como el entregar un juego con contenido de acuerdo a su edad le puede evitar comportamiento agresivos, problemas de atencion, gasto energético, respuestas hormonales y entre otros el efecto negativo de la exposición prolongada a los videojuegos, tal como la llamada «epilepsia fotosensible» [7], [8].

Los árboles de decisión han sido utilizados en el aprendizaje, sirve para hallar una respuesta o curso de acción, dadas k variables o entradas, donde cada nodo suyo hace una pregunta sobre una variable; también se puede haber más de k preguntas antes de emitir una respuesta, por

otro lado, las hojas o nodos finales "dan la respuesta" o decisión a tomar. El método usa una muestra de datos llamada generalmente conjunto de entrenamiento para formar el árbol, por lo cual resulta una buena opción para poder clasificar [9].

Este trabajo tiene como finalidad utilizar los árboles de decisión y la inteligencia que estos proveen para predecir la clasificación de los videojuegos según el contenido que presenten, tales como: sangre, lenguaje explicito, violencia, etc; en las diferentes categorías que tiene la ESRB, con el propósito de brindar ayuda al control de efectos negativos en menores de edad.

Marco Teórico

Modelos de árbol

Son modelos precisos, estables y más sencillos de interpretar principalmente porque construyen unas reglas de decisión, las cuales se pueden representar usando un diagrama similar a un árbol. A diferencia de los modelos lineales conocidos, pueden representar relaciones no lineales para resolver problemas más diversos. En estos modelos, destacan los árboles de decisión y los random forest. Al ser más precisos y elaborados, obviamente ganamos en capacidad predictiva, pero perdemos en rendimiento [10].

Árbol de decisión

Es una forma gráfica y analítica de representar todos los eventos o sucesos que pueden surgir a partir de una decisión asumida en cierto momento. Nos ayudan a tomar la decisión más "acertada", desde un punto de vista probabilístico, ante un abanico de posibles decisiones tomando en cuenta cierta precisión. Estos árboles permiten examinar los resultados y determinar visualmente cómo fluye el modelo. Los resultados visuales ayudan a encontrar subgrupos específicos y relaciones que tal vez no se podría hallar con estadísticos más tradicionales [11].

Herramientas

- *DataSet:* Un conjunto de datos corresponde a los contenidos de una única tabla de base de datos, donde cada columna del dataset representa una variable única, y cada fila está reemplazando a un miembro en concreto del conjunto de datos en la tabla. En un dataset tenemos todos los valores que puede tener cada una de las variables, como por ejemplo la sangre que se pueda visualizar y las agresiones que tengan dentro del juego. Cada uno de estas variables se conoce con el nombre de dato, también puede incluir datos para uno o más miembros en función de su número de filas [14].

- *Google Colab:* Viene a ser una herramienta que nos da la posibilidad de ejecutar y compilar scripts del lenguaje Python a mediante los servidores de Google. Permitiéndonos ejecutar porciones de código similar a un cuaderno de Jupyter Notebook para linux. Orientado para implementar machine learning, ya que al ser una máquina virtual no se limita a los recursos de hardware. Teniendo en cuenta que se puede usar tanto el GPU y la TPU del mismo servidor de Google para poder potenciar el procesamiento del proyecto [15].
- *Python:* Es un lenguaje 100% gratuito. Tratándose de un lenguaje open source siendo disponible para todas las plataformas. Ya que podemos instalarlo y ejecutarlo en diferentes sistemas operativos como Windows, Linux o MacOS [15].

Métodos y Metodología computacional

Descripción de la Aplicación

Se usó para el desarrollo de este clasificador de ESBR usando un árbol de decisión, la cual comprende las siguientes fases:

- La fuente de información fue un Dataset.csv, esto fue sacado de la página Kaggle.com, necesaria para poder construir un árbol de decisión utilizando las librerías sklearn, y se hizo uso de las funciones que nos brinda.
- Análisis de Datos: con el dataset que se obtuvo, se vio la cantidad de información que tiene, y las variables. las clasificamos en números para su uso correcto con la librería o lenguaje que se trabajaría.
- Modelado: Se escoge una técnica adecuada para poder desarrollar la clasificación.
- Evaluación: Tuvimos que corroborar efectivamente que el modelo escogido se ajuste a lo que estamos buscando, en este caso poder clasificar según ESBR y que pueda ser modificado en el futuro.

Trabajos relacionados

A continuación se muestran una serie de artículos que se han analizado para poder desarrollar nuestro trabajo, se procederá a hablar brevemente sobre los puntos que nos sirvieron de apoyo.

Algoritmos de aprendizaje automático para análisis y Predicción de datos

Sandoval [10] expone sobre la técnica de Machine Learning, rama de la inteligencia artificial, como elemento fundamental de la ciencia de datos, así como tipos de aprendizaje, diferentes

métodos que son utilizados para la predicción de los mismos y la fase de desarrollo hasta su presentación.
Se obtuvo una fundamentación teórica para el modelado del proyecto, además brinda información sobre cómo cual método de predicción podría ser más factible.

Cómo aplicar árboles de decisión en SPSS

Complemento al trabajo anterior se tiene el realizado por Berlanga, Rubio y Vilà [11] donde explica como funciona un árbol de decisión como parte de la minería de datos, usos del mismo y terminología relacionada; se presenta una información más específica en contraste con el anterior.

Predicción del consumo de cocaína en adolescentes mediante árboles de decisión

Gervilla y Palmer [12] presentan un estudio sobre cómo implementar técnicas de Data Mining, especialmente árboles de decisión, que permitan analizar el valor predictivo de la impulsividad y la búsqueda de nuevas sensaciones sobre consumo de cocaína y/u otras sustancias, principalmente en la etapa de la adolescencia, partiendo de la suposición que los rasgos anteriormente mostrados evidencian una marcada relación con las conductas de experimentación y la adicción a ciertas sustancias.

Predicción, análisis y pronóstico de covid-19 utilizando un modelo de árbol de decisión

Otro ejemplo desarrollado es la tesis de Castellanos y Haro [13], donde elaboran un modelo de aprendizaje automático basado en IA, utilizando un utilizando algoritmo de aprendizaje supervisado, árbol de decisión. Además, la creación y estructuración de una base de datos que les permitió predecir, analizar y pronosticar el grado de afectación en los pacientes contagiados de Covid-19.

Resultados y discusión

Teniendo en cuenta las fases del Machine Learning, se tiene lo siguiente:

Análisis de los datos

El dataset contiene contienen el nombre de 1895 juegos con 34 características de contenido de calificación ESRB con el nombre y la consola como características para cada juego.

Un solo punto de datos se representa como un valor binario 0-1 para la consola y un vector binario para las características del contenido de ESRB.

Variables de entrada

Característica	*Tipo*	*Descripción*	*Llave*
title	string	Nombre del juego.	- - -
console	int	La consola en la que se comercializo el videojuego.	0 = no 1 = si
Alcohol_Reference	int	Referencia a y/o imágenes de bebidas alcohólicas.	0 = no 1 = si
Animated_Blood	int	Representaciones de sangre descoloridas y/o poco realistas.	0 = no 1 = si
Blood	int	Representaciones de sangre.	0 = no 1 = si
BloodandGore	int	Representaciones de mutilación o sangre en distintas partes del cuerpo.	0 = no 1 = si
Cartoon_Violence	int	Acciones violentas que involucran situaciones escenográficas y personajes de dibujos animados. Puede incluir violencia en la que un personaje sale ileso después de que se haya infligido la acción.	0 = no 1 = si

Crude_Humor	int	Representaciones o diálogos que involucren travesuras vulgares, incluido el humor de "baño".	0 = no 1 = si
DrugRe_ference	int	Referencia a y/o imágenes de drogas ilegales.	0 = no 1 = si
Fantasy_Violence	int	Acciones violentas de naturaleza fantástica, que involucran personajes humanos o no humanos en situaciones fácilmente distinguibles de la vida real.	0 = no 1 = si
Intense_Violence	int	Representaciones gráficas y realistas de conflictos físicos. Puede incluir sangre, gore, armas y representaciones extremas y/o realistas de heridas y muertes humanas.	0 = no 1 = si
Language	int	Uso moderado de blasfemias.	0 = no 1 = si
Lyrics	int	Referencias a la blasfemia, la sexualidad, la violencia, el consumo de alcohol o drogas en la música.	0 = no 1 = si
Mature_Humor	int	Representaciones o diálogos que involucren humor "adulto", incluidas las referencias sexuales.	0 = no 1 = si
Mild_Blood	int	Un poco de sangre.	0 = no 1 = si

MildCartoonViolence	int	Algunas acciones violentas que involucran dibujos animados.	0 = no 1 = si
MildFantasyViolence	int	Algunas acciones violentas de naturaleza fantasiosa.	0 = no 1 = si
Mild_Language	int	Uso leve a moderado de blasfemias.	0 = no 1 = si
Mild_Lyrics	int	Leve referencias a blasfemias, sexualidad, violencia, consumo de alcohol o drogas en la música.	0 = no 1 = si
Mild_Violence	int	Algunas escenas que implican un conflicto agresivo.	0 = no 1 = si
No_Descriptors	int	Sin descriptores de contenido.	0 = no 1 = si
Nudity	int	Representaciones gráficas o prolongadas de desnudez.	0 = no 1 = si
Partial_Nudity	int	Representaciones breves y/o leves de desnudez.	0 = no 1 = si
Sexual_Content	int	Representaciones no explícitas de comportamiento sexual, que posiblemente incluyan desnudez parcial.	0 = no

			1 = si
Sexual_Themes	int	Referencias al sexo o la sexualidad.	0 = no 1 = si
Strong_Language	int	Uso explícito y/o frecuente de malas palabras.	0 = no 1 = si
Strong*Sexual*Content	int	Representaciones explícitas y/o frecuentes de comportamiento sexual, posiblemente incluyendo desnudez.	0 = no 1 = si
Suggestive_Themes	int	Referencias o materiales provocativos.	0 = no 1 = si
Use*of*Alcohol	int	The consumption of alcoholic beverages.	0 = no 1 = si
Use*of*Drugs*and*Alcohol	int	El consumo de bebidas alcohólicas y estupefacientes.	0 = no 1 = si
Violence	int	Escenas que implican un conflicto agresivo. Puede contener desmembramiento sin derramamiento de sangre.	0 = no 1 = si

Variable de salida

ESRB_rating	*Description*
RP	Calificación Pendiente
EC	NIñez temprana
E	Todo público
E 10+	Todo público 10+
T	Adolescente
M	Maduro
A	Adulto

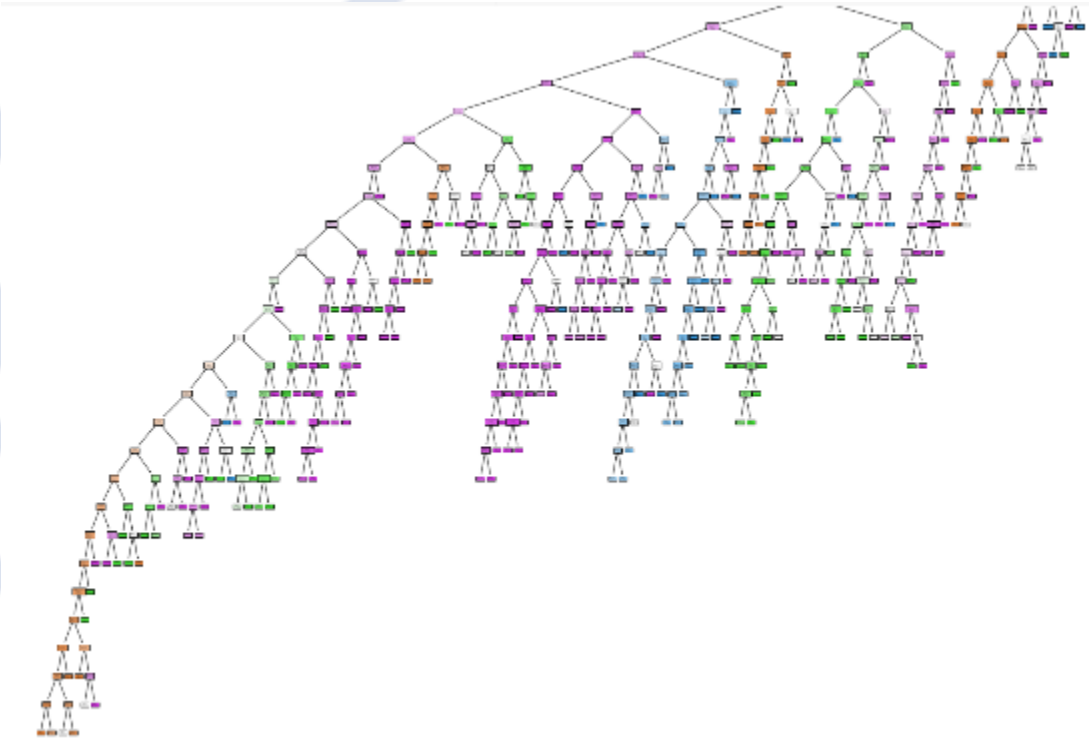
Implementación

Fase de entrenamiento

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, train_size=0.75, random_state=0)
X_train.info()
```

```
[5]  arbol = DecisionTreeClassifier();
     arbol_cat = arbol.fit(X_train, Y_train)
```

```
from matplotlib import pyplot as plt
from sklearn import tree
fig = plt.figure(figsize=(25,20))
tree.plot_tree(arbol_cat, feature_names=list(X.columns.values),class_names=list(Y.values), filled=True)
plt.show()
```



En esta fase se tiene una cantidad enorme de datos, de la cual separamos una parte para entrenar al algoritmo y darle toda esta información para que encuentre los patrones necesarios y después pueda hacer predicciones.

Fase de prueba

El resto de los datos que quedan, se van a utilizar para hacer las pruebas. Así podemos desarrollar algunas preguntas al algoritmo y evaluar si las respuestas están bien o mal, y saber si está aprendiendo o no. Si vemos que no llegan a coincidir los datos, se tiene que añadir más datos o cambiar el método que se esta utilizando. Pero si se observa que hay entre un 70% a 90% de respuestas correctas, podemos decir que hay un buen grado de aprendizaje y poder utilizar este algoritmo.

Análisis de resultados

Como se observa en la Figura la precisión después de la fase de entrenamiento es de 0.71 en 6 niveles y completo es de 0.86. A continuación, se muestra una comparación entre los árboles arrojados.

```
[17] from sklearn.metrics import confusion_matrix
     mdc = confusion_matrix(Y_test, Y_pred)
     mdc
```

```
array([[ 87,   0,   0,  18],
       [  9,  50,   0,  37],
       [  1,   4,  53,  42],
       [  6,  16,   2, 149]])
```

```
import numpy as np
pg = np.sum(mdc.diagonal())/np.sum(mdc)
pg
```

```
0.7151898734177216
```

```
[ ] from sklearn.metrics import confusion_matrix
    mdc = confusion_matrix(Y_test, Y_pred)
    mdc
```

```
array([[102,   2,   0,   1],
       [  4,  78,   0,  14],
       [  0,   0,  87,  13],
       [  2,  24,  15, 132]])
```

```
[ ] import numpy as np
    pg = np.sum(mdc.diagonal())/np.sum(mdc)
    pg
```

```
0.8417721518987342
```

Conclusiones

Los resultados obtenidos con el modelo de clasificación por árboles de decisión, indican que este es capaz de generar modelos consistentes con la realidad observada y el respaldo teórico, basándose únicamente en los datos obtenidos por el DataSet, con una precisión del 0,71 que depende de la cantidad de datos de entrenamiento que se le proporciona.

Como trabajos futuros se plantea utilizar otros clasificadores para comparar con estos resultados, además de ver que tan efectivo sería nuestro modelo, y para aumentar el porcentaje de evaluación se podría utilizar tanto un dataset que contenga más pruebas como otras técnicas de

machine learning o aplicar tareas descriptivas de minería de datos como asociación y agrupación, con el fin de encontrar relaciones y similitudes.

Referencias

[1] S. Belli and C. López, "A brief history of videogames," Athenea Digit. Rev. pensam. investig. soc., no. 14, p. 159, 2008.

[2] H. Y. S. P. C. H. P. L. Videojuegos: Conceptos, "Historia y su potencial como herramientas para la educación."

[3] A. B. Fernández Revelles, "Sistemas de calificación de video juegos, revisión narrativa," ESPHA 2018, p. 49839, 2018.

[4] "Guía de clasificaciones," ESRB Ratings. [Online]. Available: https://www.esrb.org/ratings-guide/es/. [Accessed: 23-Jun-2022].

[5] Laczniak, R. N., Carlson, L., Walker, D., & Brocato, E. D. "Parental restrictive mediation and children's violent video game play: the effectiveness of the Entertainment Software Rating Board (ESRB) rating system," Journal of Public Policy & Marketing, vol. 36, no 1, p. 70-78, 2017. [Abstract]. Available: SageJournals, https://journals.sagepub.com/. [Accessed June 22, 2022].

[6] Stroud, N. J., & Chernin, A. "Video games and the ESRB: An evaluation of parental beliefs about the rating system," Journal of Children and Media, vol. 2, no 1, p. 1-18, 2008. [Abstract]. Available: TaylorFrancisOnline, https://www.tandfonline.com/. [Accessed June 22, 2022].

[7] H. Y. S. P. C. H. P. L. "Educación, VIDEOJUEGOS: CONCEPTEl efecto de los videojuegos en variables sociales, psicológicas y fisiológicas en niños y adolescentes".

[8] Gonzálvez, M. T., Espada, J. P., & Tejeiro, R. (2016). El uso problemático de videojuegos está relacionado con problemas emocionales en adolescentes. Adicciones, 29(3), 180. https://doi.org/10.20882/adicciones.745

[9] García, A., & Martínez, G. L. (s/f). CLASIFICACIÓN SUPERVISADA INDUCCIÓN DE ARBOLES DE DECISIÓN, ALGORITMO k-d. Ipn.mx. Recuperado el 29 de junio de 2022, de https://www.cic.ipn.mx/aguzman/papers/109%20Algoritmo%20KD.%20Clasificacion%20supervisada.pdf.

[10] L. J. Sandoval Serrano, "Algoritmos de aprendizaje automático para análisis y predicción de datos," Revista Tecnológica, no. 11, 2018.

[11] R. V. Baños and M. Torrado, "CÓMO APLICAR ÁRBOLES DE DECISIÓN EN SPSS. Cómo aplicar las arbres de decisión en SPSS. Applying SPSS decision trees." .

[12] E. G. García and A. L. P. Pol, "Predicción del consumo de cocaína en adolescentes mediante árboles de decisión," Revista de investigación en educación, vol. 6, no. 1, pp. 7–13, 2009.

[13] K. S. Daza Rosado and M. A. García Reyes, "Predicción, análisis y pronóstico de COVID-19 utilizando un modelo de Machine Learning basado en el análisis forecasting sobre series temporales," Universidad de Guayaquil. Facultad de Ciencias Matemáticas y Físicas. Carrera de Ingeniería en Sistemas Computacionales., 2021.

[15] D. Blanco Álvarez, Aprendizaje y evaluación de un dataset para predecir áreas de interés en una secuencia de vídeo. 2021.

[14] Q. Wu, "geemap: A Python package for interactive mapping with Google Earth Engine," Journal of Open Source Software, vol. 5, no. 51, 2020.

Anexos

- https://colab.research.google.com/drive/1E1K12bAVVTXY77fkcidt6Rb8Gc49N6c5?usp=sharing#scrollTo=pNKE5qVR7sGs
- https://www.kaggle.com/datasets/imohtn/video-games-rating-by-esrb

Predicción del éxito del telemarketing bancario mediante el uso de árboles de decisión

Predicting the success of banking telemarketing through the use of decision trees

Rony Tito Ventura Ramos
Universidad Nacional de San Agustín.
Arequipa, Perú.
rventurar@unsa.edu.pe
https://orcid.org/0000-0002-2170-9217

Andrew Pold Jacobo Castillo
Universidad Nacional de San Agustín.
Arequipa, Perú.
ajacoboc@unsa.edu.pe
https://orcid.org/0000-0003-0949-2139

Jesus Begazo Ticona
Universidad Nacional de San Agustín.
Arequipa, Perú.
jbegazoti@unsa.edu.pe
https://orcid.org/0000-0002-8259-4601

Brian Jhosep Gomez Velasco
Universidad Nacional de San Agustín.
Arequipa, Perú.
bgomezv@unsa.edu.pe
https://orcid.org/0000-0003-0623-4353

ARK: ark:/42411/s11/a84

PURL: 42411/s11/a84

RECIBIDO 28/12/2022 • ACEPTADO 10/02/2022 • PUBLICADO 30/03/2023

RESUMEN

El telemercadeo es una técnica interactiva de mercadeo directo en la que un agente de telemercadeo solicita clientes potenciales a través del teléfono para realizar una venta de mercadería o servicio. Uno de los grandes problemas del telemarketing es especificar la lista de clientes que presentan una mayor probabilidad de comprar el producto que se ofrece. En este artículo proponemos un sistema de apoyo en la toma de decisiones personalizado que puede predecir automáticamente la decisión del público objetivo luego de realizar una llamada de telemarketing, con el fin de aumentar la efectividad de las campañas publicitarias directas y en consecuencia reducir el costo y tiempo de la campaña. El método de inteligencia artificial utilizado en este trabajo es el árbol de decisión evaluado con las métricas de precisión, exactitud y exhaustividad. Luego de aplicar el método de inteligencia artificial obtenemos una exactitud, precisión y exhaustividad mayor al 80%. Las conclusiones a los que el equipo llegó son que para mejorar el modelo de árbol de decisión es importante realizar un análisis previo de los datos mediante técnicas estadísticas o diagramas, para obtener referencia de los datos y aplicar técnicas de balanceo para obtener el mejor modelo posible.

Palabras claves: Telemarketing, Árboles de decisión, Inteligencia artificial.

ABSTRACT

Telemarketing is an interactive direct marketing technique in which a telemarketing agent solicits potential customers over the phone to make a sale of merchandise or a service. One of the great problems of telemarketing is to specify the list of clients that presents a greater probability of buying the product that is offered. In this article, we propose a personalized decision support system that can automatically predict the decision of the target audience after making a telemarketing call, in order to increase the effectiveness of direct advertising campaigns and consequently reduce the cost and cost. campaign time. The artificial intelligence method used in this work is the decision tree evaluated with the metrics of precision, accuracy and completeness. After applying the artificial intelligence method we obtain an accuracy, precision and completeness greater than 80%. The conclusions reached by the team are that in order to improve the decision tree model it is important to carry out a prior analysis of the data using statistical techniques or diagrams, to obtain a reference to the data and apply balancing techniques to obtain the best possible model.

Keywords: *Telemarketing, Decision trees, Artificial Intelligence.*

INTRODUCCIÓN

Las campañas publicitarias constituyen una estrategia tradicional que las organizaciones utilizan para aumentar el número de clientes, es decir es el proceso mediante el cual se gestiona responsablemente las necesidades del cliente con el fin de entregar e intercambiar ofertas de valor con el cliente, socios o público en general [1]. Actualmente el marketing está cada vez más relacionado a los datos, automatización e inteligencia, los avances tecnológicos han producido cambios significativos en el marketing de modo que puede trabajar con la inteligencia artificial y generar oportunidades para la empresa [2].

El telemercadeo es una técnica interactiva de mercadeo directo en la que un agente de telemercadeo solicita clientes potenciales a través del teléfono para realizar una venta de mercadería o servicio [3]. Esta tecnología permite repensar el marketing enfocándose en maximizar el valor de por vida del cliente a través de la evaluación de la información disponible y métricas del cliente [4].

Existen 2 tipos de metodologías por las cuales las organizaciones promueven sus productos: a) centradas en la población general y b) campañas directas. La primera solo genera menos del 1% de reacciones positivas en toda la población, sin embargo, las campañas directas que se concentran en un grupo pequeño de personas que se cree que tienen una mayor probabilidad de

sentirse interesados en el producto, son mucho más productivas para la empresa financiera [5, 7]. Uno de los grandes problemas del telemarketing es especificar la lista de clientes que presentan una mayor probabilidad de comprar el producto que se ofrece [6].

Para ello, los sistemas de soporte de decisiones que utilizan modelos predictivos pueden proporcionar una toma de decisiones mejor informada [6]. Incluso en la actualidad, muchos bancos han optado por la minería de datos para predecir los datos del cliente para clasificar a los clientes antes de ofrecer servicios especiales, estas técnicas se centran en hacer coincidir los atributos del cliente y otras características a diferentes resultados [5]. Sin embargo, un factor importante que afecta el rendimiento de la predicción es el número de características de entrada. Especialmente, la información del cliente del banco es que normalmente tiene muchas características, por lo que hace que disminuya el rendimiento de la predicción [3].

Otra de los problemas presentamos es manejar la distribución de conjuntos de datos desequilibrados de manera confiable; los enfoques comúnmente usados imponen una sobrecarga de procesamiento o conducen a la pérdida de información. Las Redes Neuronales Artificiales (RNN) como se indica en [8][9] en algunos casos son muy competitivas, ya que pueden manejar cualquier tipo de distribución de datos desequilibrada, pero si se crean modelos sensibles al costo tenemos un fuerte candidato para dar una solución eficiente.

En este artículo proponemos un sistema de apoyo en la toma de decisiones personalizado que puede predecir automáticamente la decisión del público objetivo luego de realizar una llamada de telemarketing, con el fin de aumentar la efectividad de las campañas publicitarias directas y en consecuencia reducir el costo y tiempo de la campaña, para ello nos enfocamos en una rango limitado de características de entrada, principalmente indicadores socioeconómicos como: edad, ocupación, educación, moroso, etc. Los registros son analizados y se les aplica una técnica de Inteligencia artificial para obtener el sistema de apoyo.

El artículo está organizado de la siguiente manera en la sección de metodología se explica la aplicación de una técnica de inteligencia artificial y el procedimiento que se realizó para obtener el sistema soporte de decisiones en la siguiente sección presentamos la discusión de los resultados y finalmente las conclusiones.

Trabajos relacionados

El trabajo realizado por Albrecht, Tobias, Rausch aborda los métodos y prácticas de implementación de inteligencia artificial en las llamadas que ingresan a un call center, este trabajo investiga las capacidades de los modelos de machine learning para pronosticar las llamadas que ingresan diariamente en un call center y medir su precisión y practicidad para ello los investigadores hacen uso de dos datasets de dos empresas retail, apoyo al cliente y reclamos del

cliente, como primer paso realizan el análisis preliminar, luego diseñan la investigación, las variables predicativas, finalmente obtienen el resultado con diferentes algoritmos, random forest, gradient boosting machine, K-nearest neighbor y vector regression, los resultados que los investigadores obtuvieron muestran que el algoritmo de random forest es el modelo que tuvo mejor rendimiento en las predicciones [10].

El artículo sirve de perspectiva para el desarrollo de esta investigación, sin embargo, difiere en el algoritmo usado y el enfoque de la investigación, ya que el artículo presenta un punto de partida para el pronóstico de llamadas de un call center mediante los algoritmos de inteligencia artificial.

El trabajo realizado por Zeinulla, Bekbayeva y Yazici [11] tiene como objetivo evaluar varios modelos de clasificación para la predicción de resultados de campañas de telemarketing bancario en cuanto a la probabilidad de suscripción del cliente al depósito. Para este propósito, la eficacia de los algoritmos ha sido evaluados mediante el análisis de la curva de características del operador de receptor (ROC) y el perfil de precisión acumulada (CAP), la precisión del algoritmo y la varianza de las predicciones.

Según los resultados de la investigación, los mejores modelos para la predicción de la efectividad del telemarketing bancario son Random Forest con una precisión de 0.90 por ROC y Redes Neuronales Artificiales Profundas (DANN) con una precisión de 0.86. El artículo sirve de perspectiva para el desarrollo de esta investigación, sin embargo, difiere del algoritmo usado y la metodología de análisis.

Marco Teórico

Árbol de decisión

Un árbol de decisiones es una herramienta de apoyo con una estructura similar a un árbol que modela los resultados probables, el costo de los recursos, las utilidades y las posibles consecuencias. Los árboles de decisión proporcionan una forma de presentar algoritmos con declaraciones de control condicional, estos árboles incluyen ramas que representan pasos de toma de decisiones que pueden conducir a un resultado favorable.

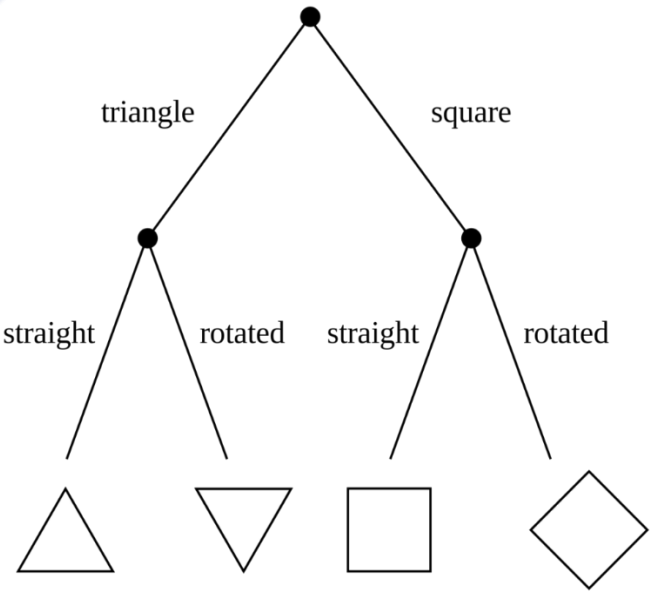


Figura 1. Árbol de decisión simple

La estructura del diagrama de flujo incluye nodos internos que representan pruebas o atributos en cada etapa y cada rama representa un resultado para los atributos, mientras que el camino desde la hoja hasta la raíz representa reglas para la clasificación. Los árboles de decisión son una de las mejores formas de algoritmos de aprendizaje basados en varios métodos de aprendizaje. Impulsan los modelos predictivos con precisión, facilidad de interpretación y estabilidad. Las herramientas también son efectivas para ajustar relaciones no lineales, ya que pueden resolver desafíos de ajuste de datos, como la regresión y las clasificaciones.

Python

El lenguaje de programación python es un lenguaje de alto nivel, las propiedades de este lenguaje es que es un lenguaje de alto nivel orientado a objetos, su semántica fue desarrollado por Guido van Rossum, Este lenguaje fue diseñado para ser fácil y entendible de programar. Python es un lenguaje de programación fácil de entender por los programadores principiantes por la sintaxis sencilla de entender y la gran cobertura de sus métodos. Python es utilizado para el desarrollo web, el desarrollo de programas de software, las matemáticas y la creación de secuencias de comandos del sistema y lenguaje de secuencias de comandos debido a sus estructuras de datos integradas de alto nivel. escritura dinámica y enlace dinámico. Las librerías creadas para Python facilitan los programas modulares y la reutilización del código. Python es un lenguaje comunitario de código abierto, por lo que numerosos programadores independientes crean continuamente bibliotecas y funcionalidades para él.

Algoritmo C4.5

El algoritmo C4.5 es una versión mejorada del algoritmo ID3 desarrollado por Ross Quinlan, este algoritmo utiliza el ratio de ganancia y la función de bondad para dividir el conjunto de datos, en cambio el algoritmo ID3 usa la ganancia de información. La función de ganancia de información prefiere las características con más categorías, los árboles de decisión que nos da C4.5 son usados mayormente para la clasificación. La ventaja de usar este algoritmo para el proyecto es que acepta variables categóricas y variables numéricas del tipo continuo y discretas [20].

Telemarketing bancario

El telemercadeo es una técnica interactiva de mercadeo directo en la que un agente de telemercadeo solicita clientes potenciales a través del teléfono para realizar una venta de mercadería o servicio [3]. Esta tecnología permite repensar el marketing enfocándose en maximizar el valor de por vida del cliente a través de la evaluación de la información disponible y métricas del cliente [4].
Existen 2 tipos de metodologías por las cuales las organizaciones promueven sus productos: a) centradas en la población general y b) campañas directas. La primera solo genera menos del 1% de reacciones positivas en toda la población, sin embargo, las campañas directas que se concentran en un grupo pequeño de personas que se cree que tienen una mayor probabilidad de sentirse interesados en el producto, son mucho más productivas para la empresa financiera [5, 7]. Uno de los grandes problemas del telemarketing es especificar la lista de clientes que presentan una mayor probabilidad de comprar el producto que se ofrece [6].

Sistema de soporte a la decisión

Un sistema de soporte a la decisión (DSS) es una herramienta que se encuentra dentro del paraguas de Inteligencia de negocios, enfocada al análisis de datos de una organización para ayudar en la toma de decisiones de una organización, un DSS usa las tecnologías de la información para apoyar la toma de decisiones, existen muchos subcampos en los cuales se puede aplicar, desde pequeños sistemas personales hasta grandes sistemas para organizaciones, estos sistemas usan técnicas de inteligencia artificial (IA) para apoyar la toma de decisiones, por ende es la columna vertebral detrás de estos sistemas [4].

Los DSS están compuestos por una interfaz de usuario que recibe y responde las consultas y un motor de inferencias, el cual es capaz de usar todo el conocimiento acumulado, base de conocimiento, y dado unos parámetros de entrado es capaz de ejecutar una serie de reglas lógicas y responder preguntas de alto nivel de acuerdo a la consulta que se realizó. El administrador

tiene acceso al resultado del análisis consolidado y toma decisiones en beneficio de la organización [13].

Herramientas Utilizadas

Sklearn

También conocido como scikit-learn es una librería hecha en Python que nos proporciona un amplio repertorio de algoritmos de aprendizaje automático además que son de última generación, dicha librería prioriza llevar el aprendizaje automático a los no especialistas, teniendo en cuenta que se realiza por medio de un lenguaje de alto nivel como lo es Python. Esta librería proporciona distintas ventajas como la facilidad de uso que tiene, rendimiento, documentación, coherencia de la API. Además que su dependencia llega a ser mínima, se distribuye bajo licencia BSD. Es bastante utilizado en sectores comerciales y también en los entornos académicos[18].

Google Colab

Colaboratory o "Colab", es un producto de Google Research. Permite a cualquier usuario escribir y ejecutar código de Python en el navegador. Este software es a menudo utilizado para tareas de aprendizaje automático, análisis de datos y educación. La razón de esto es que Colab nos permite cambiar los ajustes del entorno para realizar ejecuciones más potentes que usualmente serían poco eficientes ejecutarlas en nuestro propio ordenador [12].

Pandas

Pandas es un paquete de Python que provee estructuras de datos rápidas, flexibles y expresivas, diseñado para hacer tareas con datos etiquetados fácil e intuitivos, pretende ser un fundamental bloque de alto nivel para realizar prácticas reales en análisis de datos en Python. Adicionalmente tiene como propósito fundamental llegar a ser la herramienta open source más poderosa y flexible para el análisis y manipulación de datos en cualquier lenguaje de programación [15].
Pandas fue desarrollado a inicios del 2008 y fue publicado como open source a finales del siguiente año, algunas características fundamentales de esta herramienta son las siguientes, permite el fácil manejo de datos faltantes, las estructuras de datos que provee son fácilmente mutables en sus dimensiones, permite convertir datos, además de mapeos y mezclas entre otras funciones útiles [15, 16].

Resultados y discusión

Análisis del problemas

Los datos están relacionados con campañas de marketing directo de una entidad bancaria portuguesa. Las campañas de marketing se basaron en llamadas telefónicas. El objetivo de la clasificación es predecir si el cliente suscribirá (si/no) un depósito a plazo (variable).

Tenemos 9 variables categóricas de entrada que son: job, marital, education, default, month, housing, loan, contact, poutcome. Además 7 variables numéricas de entrada que son: age, balance, day, duration, campaign, pdays, previous.
Finalmente, una variable categórica de salida que es: y (yes, no)

Variable	Descripción
Age	Edad del objetivo, numérico
Job	Tipo de trabajo, categórico
Marital	Estado marital, categórico.
Education	Grado educativo, categórico.
Default	Tiene crédito en mora, binario.
Balance	Saldo medio anual en años, numérico.
Housing	Tiene crédito hipotecario, binario.
Loan	Tiene crédito personal, binario.
Contact	Tipo de comunicación, categórico.
Day	Último día de contacto del mes, numérico.
Duration	Tiempo de duración del último contacto en segundos, numérico.
Campaign	Número de contactos realizados durante la campaña actual, numérico.
Pdays	Número de días desde la última campaña, numérico.
Previous	Número de contactos realizados antes de esta campaña, numérico.
Poutcome	Resultado de la campaña anterior, categórico.
Month	Último mes de contacto en el año, categórico.
Salida	El cliente se suscribe o no (yes, no)

Tabla 1: Descripción de variables.

Análisis de los datos

Con la ayuda de Matplot se graficó todas las variables con el objetivo de localizar datos no relevantes (en el caso de las variables categóricas) y datos outliers (en el caso de las variables numéricas). Es importante aplicar un método de tratamiento sobre estas para que el modelo mejore, especialmente si se usa árboles de decisión. En nuestro caso
elegimos eliminarlas.

A. Data Cleaning

En las variables categóricas hacemos uso de diagramas de barras para representar los datos con el fin de
detectar anomalías en los datos.

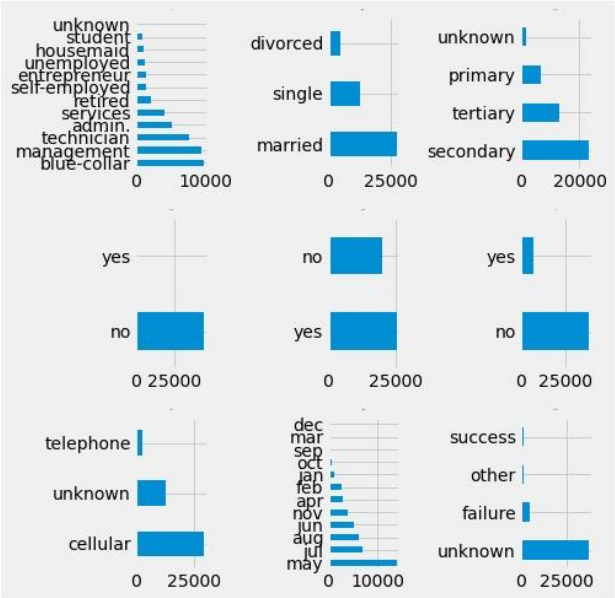


Figura 2: Diagrama de Barras para las variables categóricas

Primero eliminamos las variables irrelevantes que no aportan mucho a nuestro modelo que son: marital, month, day, contact, previous y default. En las variables que si consideramos vemos que hay valores
"unknown", estas se eliminan. En job, el valor "student" lo consideramos irrelevante al igual que el valor "other" de poutcome. Esto nos deja con 40716 filas.

Para encontrar outliers en nuestras variables numéricas nos hemos apoyado de diagramas de Caja y Bigote que nos permiten visualizar si estas tienen datos atípicos. Se ha utilizado la librería seaborn con la función plotbox().

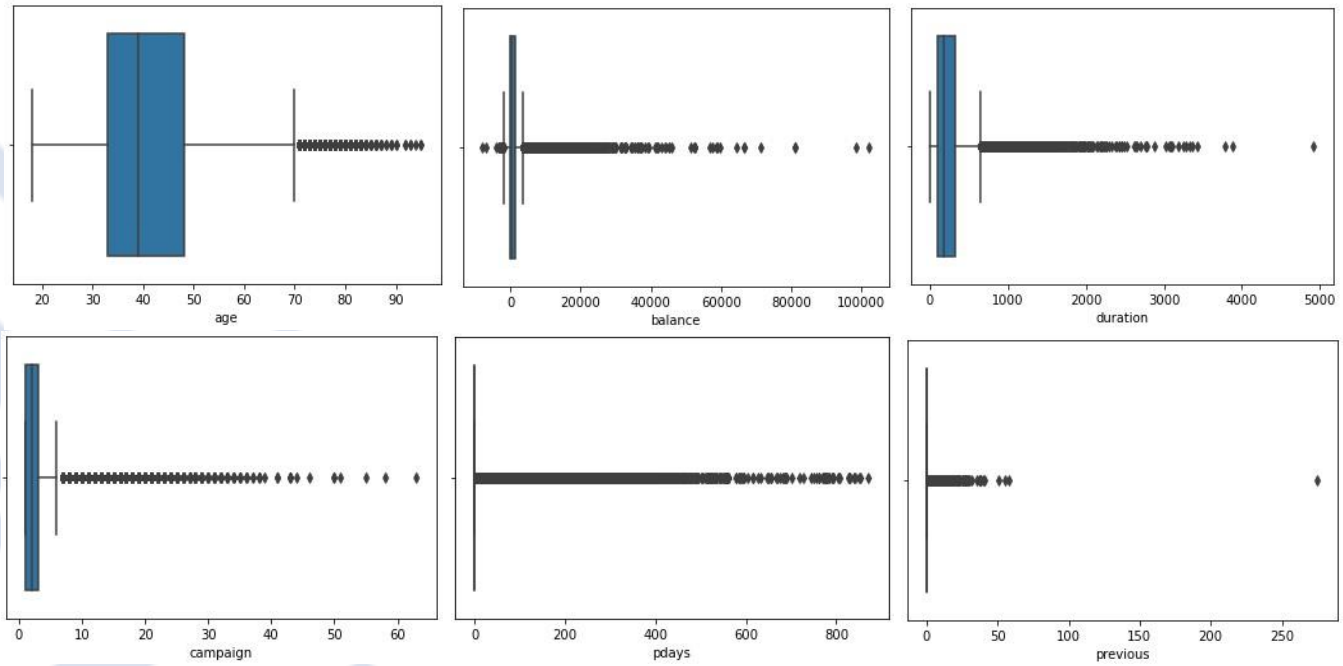


Figura 3: Diagramas de Caja y Bigote para las variables numéricas

Para eliminar este tipo de valores se ha empleado el puntaje estándar o también conocido como z-score, este valor estadístico no dará una idea de qué tan lejos de la media está un punto de datos. Para cumplir con nuestro objetivo hemos empleado la librería scipy con el módulo stats y su función zscore() que nos deja con
un total de 39178 filas.

B. Transformación de los datos

El algoritmo C4.5 utilizado para el entrenamiento de nuestro árbol de decisión es una adaptación del ID3 que permite tanto variables numéricas (continuas y discretas) como categóricas. Sin embargo, sólo admite entradas numéricas lo cual lleva a realizar una transformación en los datos especialmente en las variables categóricas las cuales debemos asignar una representación numérica por cada clase diferente. Con la ayuda de pandas y sus funciones map() podemos hacer el reemplazo de estas mismas y, además, eliminar valores vacios como NaN o Null.

Variable	Valores
job	"admin.": 0, "unknown": 1, "unemployed": 2, "management": 3, "housemaid": 4, "entrepreneur": 5, "blue-collar": 6, "self-employed": 7, "retired": 8, "technician": 9, "services": 10
education	"secondary": 0, "primary": 1, "tertiary": 2
housing	"yes": 1, "no": 0
loan	"yes": 1, "no": 0
poutcome	"unknown": 0, "failure": 1, "success": 2
y	"yes": 1, "no": 0

Tabla 2: Transformación de las variables categóricas

C. Balanceo de datos

Algo que caracteriza a los datos usados para este trabajo es su gran desbalance con respecto a la variable de salida. En la literatura consultada se presentan diferentes métodos para resolver este tipo de problema. En una fase inicial se probó eliminando las filas que contengan "no" en la variable objetivo, sin embargo, al realizar este proceso provocó que las demás variables de entrada se desbalanceen dando como resultado un árbol con bastante precisión en su modelo, pero mal entrenado. La solución más óptima encontrada fue aumentar las salidas de la clase "yes" para que de esta manera no tengamos tanta diferencia en el total de los resultados. Para ello se ha utilizado Smote que es una herramienta que se encarga de sintetizar nuevos ejemplos para la clase minoritaria. Al finalizar este proceso obtenemos 33280 filas de clase mayoritaria y 26624 de la clase minoritaria.

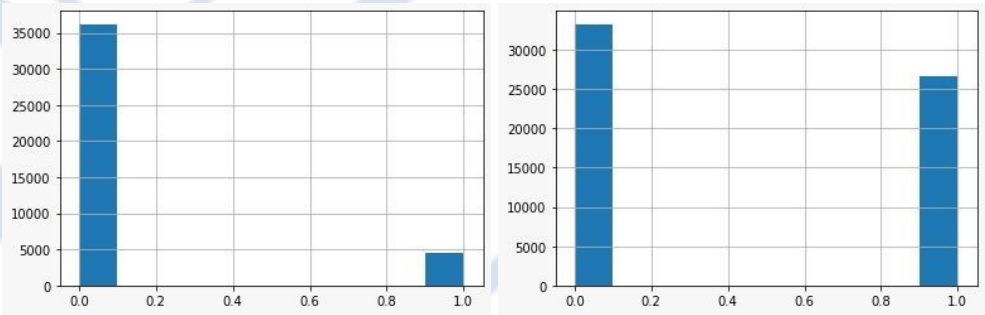


Figura 4: Balanceo de la clase minoritaria en la variable de salida

Entrenamiento

En esta sección describimos la manera de cómo se prepara el algoritmo, para el árbol de decisión, se separa los datos para el entrenamiento y otros para la prueba, en el caso propuesto se selecciona el 80% de los datos para el entrenamiento y el 20% para las pruebas. También se

utiliza el criterio de "entropy" para el algoritmo, luego se determina el nivel máximo del árbol en este caso es 6, luego se realiza el entrenamiento con los datos de prueba y los parámetros establecidos.

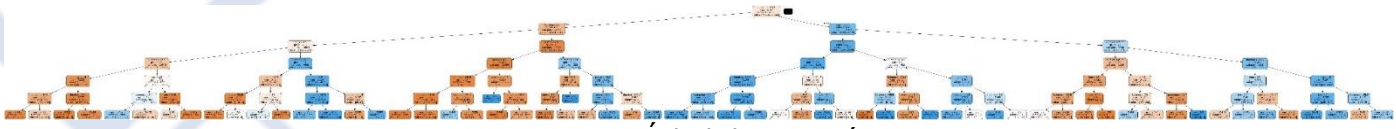


Figura 5: Árbol de Decisión.

Métricas

A. Precisión

La precisión es la capacidad del modelo de no clasificar como positiva una nuestra que es negativa, se calcula mediante la razón tp / (tp + fp), donde tp es el número de verdaderos positivos y fp el número de falsos positivos. En nuestro caso de estudio usamos la librería sklearn para obtener la precisión del modelo y obtenemos 0.81 lo cual indica que el modelo es aceptable pero no óptimo.

```
precision = precision_score(y_test, y_pred)
print('Precisión del modelo:')
print(precision)

Precisión del modelo:
0.8139244458320325
```

Figura 6: Precisión del modelo.

B. Exactitud

La exactitud mide el porcentaje de casos que el modelo ha acertado, este modelo no funciona bien cuando los datos están desbalanceados, es mucho mejor el uso de métricas como precisión, exhaustividad , F1.

```
accuracy_score = accuracy_score(y_test, y_pred)
print('Exactitud del modelo:')
print(accuracy_score)

Exactitud del modelo:
0.8473192616018239
```

Figura 7: Exactitud del modelo

C. Exhaustividad

La exhaustividad es intuitivamente la capacidad que tiene el modelo para encontrar todas las muestras positivas, se calcula mediante, tp / (tp + fn), donde tp son los verdaderos positivos y fn los falsos negativos.

```
recall_score = recall_score(y_test, y_pred)
print('Exhaustividad del modelo:')
print(recall_score)

Exhaustividad del modelo:
0.8550344375204986
```

Figura 8: Exhaustividad del modelo

D. Validación

Los valores obtenidos en las métricas de precisión, exactitud y exhaustividad son aceptables, se encuentran por encima de ochenta puntos, sin embargo no son valores óptimos, para comprobar que el modelo es capaz
de predecir correctamente se tomó un valor aleatorio dentro del conjunto de datos y observar el resultado.

```
------------------Valor X----------------------
age            35
job             4
education       3
balance      1047
housing         0
loan            0
duration      528
campaign        2
pdays          -1
Name: 50431, dtype: int64
------------------Valor Y----------------------
1
**************** Resultado ***********************
[1]
```

Figura 9: Validación del modelo

Conclusión

Dentro de la industria del marketing es importante optimizar el público objetivo para el telemercadeo, ya que escoger el público correcto reduce los costos y aumenta la probabilidad de ganancias. En este estudio nosotros presentamos la implementación de un árbol de decisión como modelo de solución, para saber si un cliente acepta o no un crédito bancario. El conjunto de datos obtenidos muestra un desbalance excesivo, por ello se aplicó técnicas para el balanceo como,

smote, eliminación de valores desconocidos y variables irrelevantes con el objetivo de mejorar el modelo de
árbol de decisión.

Adicionalmente se realizó la transformación de los datos mapeando el dominio de los datos, además de asignar valores a los datos desconocidos. Para comprobar si el modelo es útil se aplicó las siguientes métricas: precisión, exactitud y exhaustividad, obteniendo como resultado valores aceptables que están por encima del 80%, es importante realizar un análisis previo de los datos, mediante técnicas estadísticas o gráficas como los diagramas de barras, caja y bigote según el tipo de dato y su dominio.

Trabajos Futuros

En trabajos futuros, planteamos realizar estudios sobre el mismo conjunto de datos aplicando otras técnicas de inteligencia artificial como redes neuronales, regresión logística. Luego de aplicar las técnicas anteriormente mencionadas, realizar estudios comparativos para obtener la técnica y el objeto que mejor se ajuste a los datos.

Referencias

[1]. ESAN, "El Marketing y sus definiciones | Conexión ESAN," 2016. https://www.esan.edu.pe/conexion-esan/el-marketing-y-sus-definiciones (accessed Jun. 22, 2022).

[2]. S. Chintalapati and S. K. Pandey, "Artificial intelligence in marketing: A systematic literature review," *Int. J. Mark. Res.*, vol. 64, no. 1, pp. 38–68, 2022, doi: 10.1177/14707853211018428.

[3]. Vajiramedhin, C., & Suebsing, A. (2014). Feature selection with data balancing for prediction of bank telemarketing. *Applied Mathematical Sciences*, *8*(114), 5667-5672.

[4]. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22-31.

[5]. Asare-Frempong, J., & Jayabalan, M. (2017, September). Predicting customer response to bank direct telemarketing campaign. In *2017 International Conference on Engineering Technology and*

[6]. Moro, S., Cortez, P., & Rita, P. (2018). A divide-and-conquer strategy using feature relevance and expert knowledge for enhancing a data mining approach to bank telemarketing. *Expert Systems*, *35*(3), e12253.

[7]. Lau, Kn., Chow, H. & Liu, C. A database approach to cross selling in the banking industry: Practices, strategies and challenges. J Database Mark Cust Strategy Manag 11, 216–234 (2004).

[8]. Ghatasheh, N., Faris, H., AlTaharwa, I., Harb, Y., & Harb, A. (2020). Business analytics in telemarketing: cost-sensitive analysis of bank campaigns using artificial neural networks. *Applied Sciences*, *10*(7), 2581.

[9]. Kim, K. H., Lee, C. S., Jo, S. M., & Cho, S. B. (2015, November). Predicting the success of bank telemarketing using deep convolutional neural network. In *2015 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)* (pp. 314-317). IEEE.

[10]. Albrecht, Tobias; Rausch, Theresa Maria; Derra, Nicholas Daniel (2021). Call me maybe: Methods and practical implementation of artificial intelligence in call center arrivalsâ□□ forecasting. Journal of Business Research, 123(), 267–278, doi:10.1016/j.jbusres.2020.09.033

[11]. Zeinulla, K. Bekbayeva and A. Yazici, "Comparative study of the classification models for prediction of bank telemarketing," 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), 2018, pp. 1-5, doi: 10.1109/ICAICT.2018.8747086.

[12]. "Google Colab". Google Research.https://research.google.com/colaboratory/intl/es/faq.html#:~:text=Colaboratory,%20o%20"Colab"%20para,an álisis%20de%20datos%20y%20educación. (accedido el 17 de agosto de 2022).

[13]. Stalidis, D. Karapistolis, and A. Vafeiadis, "Marketing Decision Support Using Artificial Intelligence and Knowledge Modeling: Application to Tourist Destination Management," *Procedia - Soc. Behav. Sci.*, vol. 175, pp. 106–113, 2015, doi: 10.1016/j.sbspro.2015.01.1180.

[14]. *pandas - Python Data Analysis Library*. (n.d.). Retrieved August 14, 2022, from https://pandas.pydata.org/

[15]. *pandas · PyPI*. (n.d.). Retrieved August 14, 2022, from https://pypi.org/project/pandas/

[16]. MARTÍNEZ, Guillermo Roberto Solarte; MEJÍA, José A. Soto. Árboles de decisiones en el diagnóstico de enfermedades cardiovasculares. Scientia et technica, 2011, vol. 16, no 49, p. 104-109.

[17]. GARCÍA PICHARDO, Víctor Hugo, et al. Algoritmo ID3 en la detección de ataques en aplicaciones Web.

[18]. PEDREGOSA, Fabian, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 2011, vol. 12, p. 2825-2830.

[19]. Larranaga, P., Inza, I., & Moujahid, A. Tema 10: árboles de clasificación. 2020 from http://www. sc. ehu.es/ccwbayes/docencia/mmcc/docs/t10arboles. pdf.

Creación de un Árbol de Decisión para la Predicción de Tonos a Partir de un Data Set

Analysis of an input dataset to perform a tonal analysis system

Víctor Manuel Vilca Rojas
Universidad Nacional de San Agustín.
Arequipa, Perú.
vvilcaro@unsa.edu.pe
https://orcid.org/0000-0002-6193-8057

Aldair Bryan Salcedo Chávez
Universidad Nacional de San Agustín.
Arequipa, Perú.
asalcedoch@unsa.edu.pe
https://orcid.org/0000-0001-7692-4064

Jairo Miguel Castillo Rojas
Universidad Nacional de San Agustín.
Arequipa, Perú.
jcastillo@unsa.edu.pe
https://orcid.org/0000-0001-5952-7323

Valery Byrne Macias
Universidad Nacional de San Agustín.
Arequipa, Perú.
vbyrne@unsa.edu.pe
https://orcid.org/0000-0003-2819-7127

ARK: ark:/42411/s11/a85

PURL: 42411/s11/a85

RECIBIDO 07/01/2022 • ACEPTADO 16/02/2022 • PUBLICADO 30/03/2023

RESUMEN

El análisis musical es un proceso que se ha llevado a cabo desde hace años donde diferentes expertos han buscado estudiar variadas piezas musicales. Este proceso inicia con el aprendizaje de detección de tonos, notas y acordes, donde los estudiantes tienen que entrenar el oído para poder llevarlo a cabo. Bajo este contexto, en el siguiente trabajo se ha realizado un árbol de decisión en base a un dataset de coros de Bach con el fin de predecir acordes a partir de tonos. Se dividió el dataset en 80% para crear el árbol y 20% para pruebas, después se realizó la transformación de datos para realizar un análisis de los mismos, con esto finalmente se creó un árbol de decisión con una profundidad de 15 y una exactitud del 75.52%, finalmente se realizaron las pruebas y encontramos buenos resultados de la exactitud del árbol.

Palabras claves: Inteligencia artificial, árbol de decisión, análisis musical, detección de tonos, música.

ABSTRACT

Musical analysis is a process that has been carried out for years where different experts have sought to study various musical pieces. This process begins with the learning of tone, note and chord detection, where students have to train their ears to be able to carry it out. In this context, in the following work a decision tree has been made based on a dataset of Bach choirs in order to predict chords from tones. The dataset was divided into 80% to create the tree and 20% for testing, then the data transformation was performed to perform an analysis of the data, with this a decision tree was finally created with a depth of 15 and an accuracy of 75.52%, the tests were finally carried out and we found good results for the accuracy of the tree.

Keywords: *Artificial intelligence, decision tree, music analysis, pitch detection, music.*

INTRODUCCIÓN

Las ondas de sonido se propagan a través de varios medios y permiten la comunicación o el entretenimiento para nosotros, los humanos. La música que escuchamos o creamos se puede percibir en aspectos como el ritmo, la melodía, la armonía, el timbre o el estado de ánimo. Todos estos elementos de la música pueden ser de interés para los usuarios de sistemas de recuperación de información musical. Dado que vastos repositorios de música están disponibles para todos en el uso diario (tanto en colecciones privadas como en Internet), es deseable y se vuelve necesario explorar las colecciones de música por contenido. Por lo tanto, la recuperación de información musical puede ser potencialmente de interés para todos los usuarios de computadoras e Internet [1].

Dado un flujo musical, la tarea del análisis de la armonía musical consiste en asociar una etiqueta a cada punto de tiempo. Tales etiquetas revelan la armonía subyacente al indicar una nota fundamental (raíz) y un modo, usando nombres de acordes como 'Do menor'. La tarea de análisis musical puede representarse naturalmente como un problema de aprendizaje secuencial supervisado. De hecho, al considerar sólo las clases de tonos actualmente resonantes, difícilmente se producirían análisis razonables. Las evidencias experimentales sobre la cognición humana revelan que para eliminar la ambigüedad de los casos poco claros, los compositores y los oyentes también se refieren a las transiciones de acordes: en estos casos, el contexto juega un papel fundamental y las claves contextuales pueden ser útiles para el análisis [6].

El dominio musical siempre ha ejercido una fuerte fascinación sobre investigadores de diversos campos. En los últimos años se ha invertido un gran esfuerzo de investigación para analizar la música, bajo una presión académica e industrial. Las técnicas de búsqueda y análisis de música inteligente son cruciales para diseñar sistemas para varios propósitos, como la identificación de música, para decidir sobre la similitud de la música, para la clasificación de música basada en

algún conjunto de descriptores, para la generación algorítmica de listas de reproducción y para el resumen de música. De hecho, los avances tecnológicos recientes mejoraron significativamente la forma en que los entornos automáticos componen música, la interpretan expresivamente, acompañan a músicos humanos y la forma en que se vende y compra música a través de tiendas web [5].

El análisis musical es un paso necesario para componer, interpretar y en última instancia, comprender la música, tanto para los seres humanos como para los entornos artificiales. Dentro del área más amplia del análisis musical, destacamos la tarea del reconocimiento de acordes. Este es un problema desafiante para los estudiantes de música, que dedican una cantidad considerable de tiempo a aprender la armonía tonal, así como para los sistemas automáticos. Es un problema interesante y un paso necesario para realizar un análisis estructural de alto nivel que considere los principales elementos estructurales de la música en sus interconexiones mutuas. En la música tonal occidental, en cada momento del flujo musical (o vertical) se puede determinar qué acorde está sonando: el reconocimiento de acordes normalmente consiste en indicar la nota fundamental (o raíz) y el modo del acorde (Figura 1) [5].

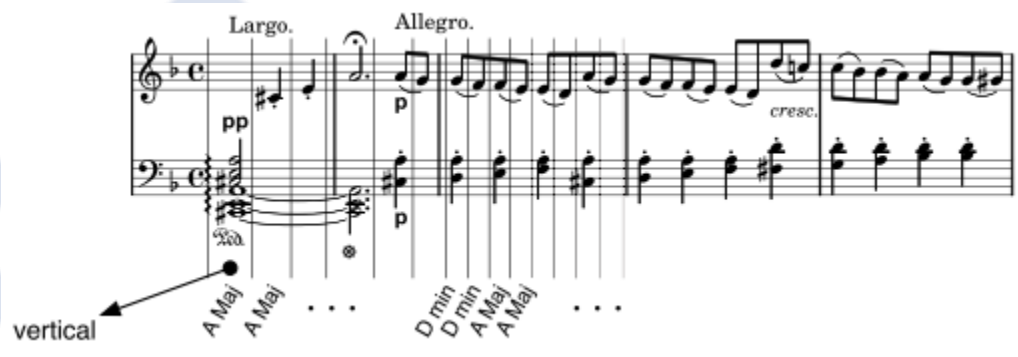


Figura 1 :El problema de reconocimiento de acordes consiste en indicar para cada vertical qué acorde está sonando en ese momento. Extracto de la Sonata para piano Opus 31 n.2 de Beethoven, 1er movimiento.

En síntesis, como dice Bent [7], el análisis musical consiste en exponer y describir sintéticamente la estructura musical y la manera en que se relacionan esos elementos más simples en la estructura general. El análisis puede ir desde una parte de una pieza musical hasta una colección de obras en un periodo de tiempo exacto. Pero el proceso seguirá siendo el mismo utilizando tradicionalmente el oído, papel y un lápiz [8]. Sin embargo, si nos proyectamos a un ámbito más actual, donde frente a nosotros no tenemos una partitura sino una colección de CD´s o hasta un celular con una lista de reproducción de Spotify, las técnicas anteriormente utilizadas para analizar este contenido resultan difíciles de usar, por no decir inservibles.

En estas situaciones del mundo moderno es que surgen diferentes métodos acompañados de algoritmos, los cuales participan en nuevos sistemas de análisis [9].

En este párrafo se expondrá los métodos a utilizar en este artículo: como primer punto se empleará la técnica de árbol de decisión, ya que es una técnica de forma gráfica y analítica de representar todos los eventos o sucesos que pueden surgir a partir de una decisión asumida en cierto momento. Esto nos permitirá tomar la decisión más acertada, ante una variedad de posibles decisiones [10]. Cómo segunda técnica a emplear tenemos clustering, esta técnica es utilizada como un proceso para encontrar una estructura significativa, procesos subyacentes explicativos, características generativas y agrupaciones inherentes a un conjunto de ejemplos; lo que la hace idónea para la presente investigación [11]. Entre otras técnicas de minería de datos.

Entonces, ya habiendo comprendido las bases teóricas de esta investigación, explicaremos la finalidad de este artículo, el cual es: "Desarrollar el análisis del dataset de entrada para realizar un sistema de análisis tonal" el cual nos dará como resultado, un acorde después de recibir los siguientes datos de ingreso: archivos de Bach Central [WebLink]; donde se reconocerá las notas musicales con un SI/NO dependiendo en qué lugar se presente el tono.

Herramientas

Árbol de decisiones: Se utilizó árboles de decisiones ya que es una manera de representación de todos los eventos que pueden surgir por una decisión asumida en cierto momento la cual puede ser gráfica y analítica que nos permite organizar el trabajo de cálculos correspondientes para así un despliegue visual del problema. Además de ello nos ayuda a tomar la decisión más idónea, desde el punto de vista probabilístico de un sin fin de posibles decisiones [17].

Correlación de Pearson: La correlación de Pearson mide la existencia (dada por un valor p) y la fuerza (dada por el coeficiente r entre -1 y +1) de una relación lineal entre dos variables continuas. Solo debe usarse cuando se satisfacen sus supuestos subyacentes. Si el resultado es significativo, concluimos que existe una correlación. Para mejor entendimiento un valor absoluto de r de 0,1 se clasifica como pequeño, un valor absoluto de 0,3 se clasifica como medio y de 0,5 se clasifica como grande [18].

Métodos y Metodología computacional

Google Colab: Es un entorno colaborativo de Google que permite trabajar con Notebooks sin alguna configuración requerida que se ejecuta en la nube y te proporciona acceso gratuito a GPU

lo cual te permite escribir y ejecutar código de Python en tu navegador y que a su vez permite almacenar dichos cuadernos y trabajar con datos que tengas almacenados en el Drive y compartirlos con tu equipo de trabajo [12].

Python: Es un lenguaje de programación de alto nivel, además es un lenguaje interpretado, por lo cual no requiere ser compilado. Por este motivo el programador puede utilizar el lenguaje de forma directa en el programa o aplicación que realice. Una de las características más resaltantes es que cuenta con una amplia bibliotecas de librerías que permiten obtener diversos recursos de código abierto aplicables para la inteligencia artificial [13].

NumPy: Es una biblioteca idónea para la manipulación de muchos datos. El problema yace en que Python usa listas y no arreglos por ello NumPy nos provee estructuras muy eficientes para manipular muchos datos [14].

Pandas: Es un buen toolkit para hacer análisis de datos. Tiene herramientas para tener tablas y otras estructuras de datos. Además de ello nos permite cargar con gran facilidad archivos csv [15].

Seaborn: Es una librería que usa Matplotlib por debajo para trazar gráficos. Lo cual será usado para visualizar distribuciones aleatorias [16].

Trabajos relacionados

Si vemos en un entorno específico, dentro de la enseñanza de la música, las TICs (Tecnologías de la Información y la Comunicación) aportaron al proceso de enseñanza-aprendizaje. Por un lado, ayudaron a los maestros que tenían alrededor 15 alumnos a personalizar más sus sesiones y por otro lado, el estudiante podía solventar sus dudas en cualquier momento al interactuar con este sistema. Estas TICs como herramientas comenzaron a complementar las sesiones en diferentes conservatorios, que cabe recalcar que solo se llevaban una vez a la semana, y de tal modo el descenso del nivel de conocimiento de los estudiantes de años actuales, a comparación de años anteriores, comenzó a menguar [9].

Siguiendo esta investigación, Illescas logró desarrollar un software que realizaba análisis musical, el cual orientado a la pedagogía lo probó en el Conservatorio Superior de Música de Murcia. Para realizar una buena interpretación de una pieza, se necesita realizar un buen análisis de la partitura, por lo que este software demostró su ayuda tanto como para alumnos y como para maestros [9].

Resultados y discusión

Inicialmente se importan las librerías necesarias para ejecutar nuestro árbol de decisiones.

```
# Imports necesarios
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as plt
%matplotlib inline
plt.rcParams['figure.figsize'] = (16, 9)
plt.style.use('ggplot')
from sklearn import tree
from sklearn.metrics import accuracy_score
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from IPython.display import Image as PImage
from subprocess import check_call
from PIL import Image, ImageDraw, ImageFont
from sklearn.model_selection import train_test_split
```

Cargamos los valores de entrada.

```
[ ] bach_choral = pd.read_csv(r"bach_choral_set_dataset.csv")

[ ] from google.colab import drive
    drive.mount('/content/drive')

    Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

[ ] bach_choral.shape

    (4532, 17)

[ ] bach_choral.head()
```

	choral_ID	event_number	pitch_1	pitch_2	pitch_3	pitch_4	pitch_5	pitch_6	pitch_7	pitch_8	pitch_9	pitch_10	pitch_11	pitch_12	bass	meter	chord_label
0	000106b_	1	YES	NO	NO	NO	NO	YES	NO	NO	NO	YES	NO	NO	F	3	F_M
1	000106b_	2	YES	NO	NO	NO	YES	NO	NO	YES	NO	NO	NO	NO	E	5	C_M
2	000106b_	3	YES	NO	NO	NO	YES	NO	NO	YES	NO	NO	NO	NO	E	2	C_M
3	000106b_	4	YES	NO	NO	NO	NO	YES	NO	NO	NO	YES	NO	NO	F	3	F_M
4	000106b_	5	YES	NO	NO	NO	NO	YES	NO	NO	NO	YES	NO	NO	F	2	F_M

Para ello se usó Data Transformation para cambiar los valores “NO” y “YES”
Los valores respectivos son 0 = no, 1 = sí

```
bach_choral.groupby('chord_label').size()
```

```
chord_label
A#d       3
A#d7      2
A_M     235
A_M4     11
A_M6      2
         ...
G_M6      2
G_M7     48
G_m     153
G_m6      3
G_m7     18
Length: 95, dtype: int64
```

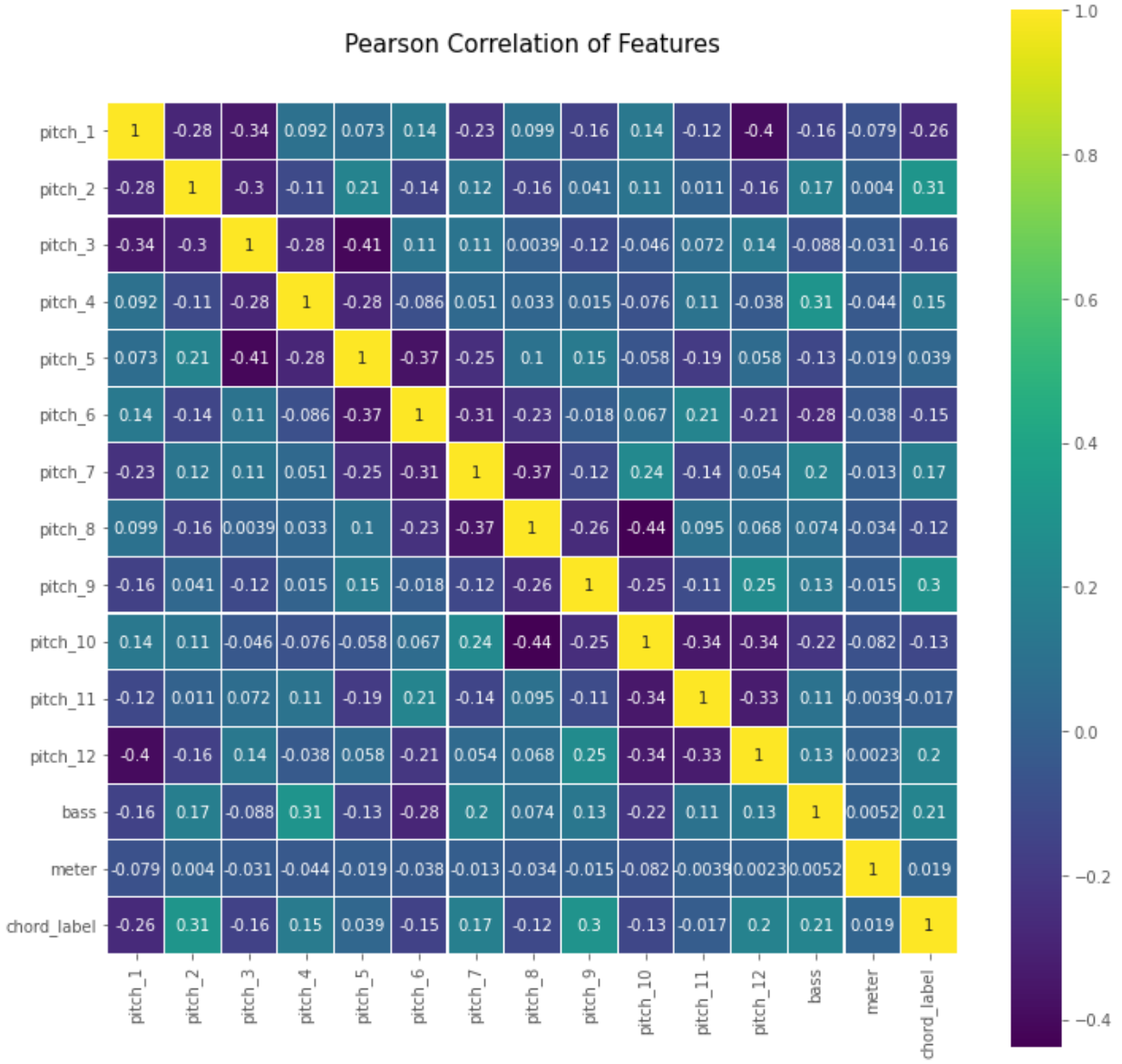
```
categorical_cols = ['bass','chord_label']
for column in categorical_cols:
    bach_choral[column] = pd.factorize(bach_choral[column])[0]

for i in range(1,13):
    pitchChange = "pitch_" + str(i)
    bach_choral[pitchChange] = bach_choral[pitchChange].replace('YES', 1)
    bach_choral[pitchChange] = bach_choral[pitchChange].replace('NO', 0)

drop_elements = ['choral_ID','event_number']
bach_choral_encoded = bach_choral.drop(drop_elements, axis = 1)

bach_choral_encoded
```

	pitch_1	pitch_2	pitch_3	pitch_4	pitch_5	pitch_6	pitch_7	pitch_8	pitch_9	pitch_10	pitch_11	pitch_12	bass	meter	chord_label
0	1	0	0	0	0	1	0	0	0	1	0	0	0	3	0
1	1	0	0	0	1	0	0	1	0	0	0	0	1	5	1
2	1	0	0	0	1	0	0	1	0	0	0	0	1	2	1
3	1	0	0	0	0	1	0	0	0	1	0	0	0	3	0
4	1	0	0	0	0	1	0	0	0	1	0	0	0	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4527	0	1	0	0	1	0	0	1	0	1	0	0	9	3	11
4528	0	0	1	0	0	0	1	0	0	1	0	0	2	4	11
4529	0	0	1	0	0	0	1	0	0	1	0	0	2	3	11
4530	0	0	1	0	1	0	0	1	0	1	0	0	1	2	11
4531	0	0	1	0	0	0	1	0	0	1	0	0	8	5	11

4532 rows × 15 columns

Luego de ello se realizó un análisis de nuestros datos de entrada categóricos el cual mostrará una correlación de Pearson en su mayoría pequeña(o débil) entre los datos ya que la mayoría ésta inferior al 0.3.

```
colormap = plt.cm.viridis
plt.figure(figsize=(12,12))
plt.title('Pearson Correlation of Features', y=1.05, size=15)
sb.heatmap(bach_choral_encoded.astype(float).corr(),linewidths=0.1,vmax=1.0, square=True, cmap=colormap, linecolor='white', annot=True)
```



Para continuar se realiza la creación del Árbol Decisión

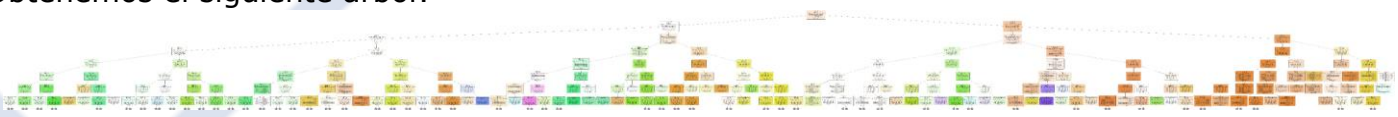
```
# Crear arrays de entrenamiento y las etiquetas que indican si llegó a top o no
y_train = bach_choral_encoded['chord_label']
x_train = bach_choral_encoded.drop(['chord_label'], axis=1).values
```

```
# Crear Arbol de decision con profundidad = 4
decision_tree = tree.DecisionTreeClassifier(criterion='entropy',
                                            min_samples_split=20,
                                            min_samples_leaf=5,
                                            max_depth = 15,
                                            class_weight={1:3.5})
decision_tree.fit(x_train, y_train)
```

```
# exportar el modelo a archivo .dot
with open(r"tree1.dot", 'w') as f:
    f = tree.export_graphviz(decision_tree,
                             out_file=f,
                             max_depth = 7,
                             impurity = True,
                             feature_names = list(bach_choral_encoded.drop(['chord_label'], axis=1)),
                             class_names = ['F_M',
'C_M',
'D_m',
'BbM',
'C_M7',
'D_m7',
'G_M',
'A_m',
'C_M4',
'G_m',
'G_M7',
'D_M',
'F#d',
'AbM',
'C#d7',
'D_M7',
'A_M',
'EbM',
'F_M7',
```

```
'Bbd',
'Dbm7',
'Abm',
'DbM7',
'Dbm',
'F#m6',
'G#m',
'B_d',
'C_M6',
'D#m',
'D#M',
'BbM7',
'F_d7',
'C#d6',
'G_d',
'G#M',
'C#M4',
'D#d6',
'D#d7'],
                                rounded = True,
                                filled= True )

# Convertir el archivo .dot a png para poder visualizarlo
check_call(['dot','-Tpng',r'tree1.dot','-o',r'tree1.png'])
PImage("tree1.png")
```

Obtenemos el siguiente árbol:



Se realiza el mapeo de atributos y las predicciones del árbol de decisión.

```
#predecir

x_test = pd.DataFrame(columns=['meter','pitch_1_encoded', 'pitch_2_encoded', 'pitch_3_encoded','pitch_4_encoded','pitch_5_encoded','pitch_6_encoded', 'pitch_7_encoded','pitch_8_encoded', 'pitch_9_encoded', 'pitch_10_encoded', 'pitch_11_encoded', 'pitch_12_encoded', 'bass_encoded', 'chord_label_encoded'])
x_test.loc[0] = (0,0,1,0,0,0,1,1,0,1,0,0,9,2,10)

y_pred = decision_tree.predict(x_test.drop(['chord_label_encoded'], axis = 1))
print("Prediccion: " + str(y_pred))
y_proba = decision_tree.predict_proba(x_test.drop(['chord_label_encoded'], axis = 1))
print("Probabilidad de Acierto: " + str(round(y_proba[0][y_pred[0]]* 100, 2))+"%")
```

```
Prediccion: [11]
Probabilidad de Acierto: 81.82%
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:444: UserWarning: X has feature names, but DecisionTreeClassifier was fitted without feature names
  f"X has feature names, but {self.__class__.__name__} was fitted without"
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:444: UserWarning: X has feature names, but DecisionTreeClassifier was fitted without feature names
  f"X has feature names, but {self.__class__.__name__} was fitted without"
```

```
#predecir test data

x_test = pd.read_csv(r"bach_choral_set_test.csv")
categorical_cols = ['bass','chord_label']
for column in categorical_cols:
    x_test[column] = pd.factorize(x_test[column])[0]

for i in range(1,13):
    pitchChange = "pitch_" + str(i)
    x_test[pitchChange] = x_test[pitchChange].replace('YES', 1)
    x_test[pitchChange] = x_test[pitchChange].replace('NO', 0)

drop_elements = ['choral_ID','event_number']
x_test_encoded = x_test.drop(drop_elements, axis = 1)

x_test_encoded
```

	pitch_1	pitch_2	pitch_3	pitch_4	pitch_5	pitch_6	pitch_7	pitch_8	pitch_9	pitch_10	pitch_11	pitch_12	bass	meter	chord_label
0	0	0	1	0	0	0	1	1	0	1	0	0	0	2	0
1	0	0	1	0	0	0	1	0	0	0	0	1	1	3	1
2	0	0	0	0	1	0	0	1	0	0	0	1	2	4	2
3	0	0	0	0	1	0	0	1	0	0	0	1	2	3	2
4	0	0	1	0	1	0	0	1	0	0	0	1	3	2	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1128	0	0	1	0	0	0	0	1	0	0	1	0	6	4	44
1129	0	0	1	0	0	0	0	1	0	1	0	0	6	3	44
1130	1	0	0	0	1	0	0	1	0	0	0	0	8	5	7
1131	1	0	0	0	1	0	0	1	0	0	1	0	8	3	7
1132	0	0	0	0	0	1	0	0	0	1	0	0	7	4	40

1133 rows x 15 columns

Imprimimos las predicciones

```
y_pred = decision_tree.predict(x_test_encoded.drop(['chord_label'], axis = 1))
print("Prediccion: " + str(y_pred))
y_proba = decision_tree.predict_proba(x_test_encoded.drop(['chord_label'], axis = 1))
print("Probabilidad de Acierto: " + str(round(y_proba[0][y_pred][0]* 100, 2))+"%")
```

Conclusiones

Se puede concluir que nuestro modelo de clasificación genera modelos acertados con un 75.52% de exactitud. Esto nos indica que nuestro árbol de decisión es capaz de predecir correctamente la nota musical en el 75.52% de los casos, esto a partir de los datos de entrada que consideramos en este trabajo así como las características que indicamos para la creación del árbol.

También encontramos que nuestro dataset, al tratarse de notas musicales de distintas canciones, y ademas, teniendo en cuenta que los eventos del dataset están numerados consecutivamente, indicándonos el orden en que estos hacen su aparición en la canción, es que podemos entender porque la correlación entre las variables de entrada es tan baja, una mejor opción sería realizar otros análisis de correlación.

Además, en la búsqueda de un mayor grado de exactitud del árbol de decisión, encontramos que la profundidad más adecuada para este era de 15, un mínimo de hojas para dividir el nodo de 20 y un mínimo de eventos para considerar hoja de 5, además, al aumentar el peso de la clase pitch_1 del dataset de 1 a 3.5 con una diferencia de 2.5 respecto a las demás clases, encontramos que el árbol de decisión incrementó su exactitud en un porcentaje mayor al 30%.

Como trabajo futuro. Teniendo ya la base del conocimiento de nuestro proyecto se puede realizar un identificador de acordes para cualquier artista, el cual con un dataset de sus pistas podemos deducir cuales son los acordes más tocados en sus pistas.

Referencias

[1]. Raś, Z. W., & Wieczorkowska, A. A. (Eds.). (2010). Advances in Music Information Retrieval. Studies in Computational Intelligence. doi:10.1007/978-3-642-11674-2

[2]. R. Esposito and D. P. Radicioni, "CarpeDiem: Optimizing the viterbi algorithm and applications to supervised sequential learning," *J. Mach. Learn. Res.*, vol. 10, pp. 1851–1880, 2009.

[3]. M. Rohrmeier and I. Cross, "Statistical Properties of Tonal Harmony in Bach ' s Chorales Statistical Properties of Tonal Harmony in Bach ' s Chorales," no. January 2008, 2014.

[4]. D. P. Radicioni and R. Esposito, "Learning tonal harmony from Bach chorales," *Procs. 7th Int. Conf. Cogn. Model.*, no. August, 2006.

[5]. D. P. Radicioni and R. Esposito, "BREVE: An HMPerceptron-based chord recognition system," *Stud. Comput. Intell.*, vol. 274, no. December, pp. 143–164, 2010, doi: 10.1007/978-3-642-11674-2_7.

[6]. D. P. Radicioni and R. Esposito, "UCI Machine Learning Repository: Bach Choral Harmony Data Set", Archive.ics.uci.edu, 2014. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Bach+Choral+Harmony#. [Accessed: 20- Jun- 2022].

[7]. J. Bent. "Music analysis in the nineteenth century" 1994.

[8]. Federico Sammartino. "Ceros y unos en la musicología. Software y análisis musical". http://resonancias.uc.cl/images/PDFs_n_37/Sammartino.pdf.

[9]. P. R. Illescas Casanova. "Análisis tonal asistido por ordenador". Dialnet. https://dialnet.unirioja.es/servlet/tesis?codigo=60733 .

[10]. V. Berlanga-Silvente, M. J. Rubio-Hurtado, and R. Vilà-Baños, "Cómo aplicar árboles de decisión en SPSS," REIRE. Rev. d'Innovació i Recer. en Educ., vol. 6, no. 1, pp. 65–79, 2013, doi: 10.1344/reire2013.6.1615.

[11]. D. Tamm, "Road Map," Dtsch. Arztebl. Int., vol. 115, no. 35–36, p. A1554, 2018, doi: 10.4324/9781003191056-1.

[12]."Te damos la bienvenida a Colab." https://colab.research.google.com/?utm_source=scs-index.

[13]."Python 3.10.6 documentation," [Online]. Available: https://docs.python.org/3/.

[14]."NumPy." https://numpy.org/.

[15]."Pandas." https://pandas.pydata.org/.

[16]."Seaborn." https://www.w3schools.com/python/numpy/numpy_random_seaborn.asp.

[17]. M. C. Ruiz Abellón, "Introducción a los árboles de decisión," pp. 1–7, 2014, [Online]. Available: http://www.dmae.upct.es/~mcruiz/Telem06/Teoria/arbol_decision.pdf.

[18]. P. Samuels, "וסריפ מאתמ Pearson Correlation ? מאתמ והמ," no. April 2014, pp. 1–5, 2015, [Online]. Available: https://www.researchgate.net/publication/274635640.

Pruebas de Software para Microservicios

Software Testing for Microservices

Cesar Adolfo Laura Mamani
Universidad La Salle. Arequipa, Perú.

clauram@ulasalle.edu.pe

ARK: ark:/42411/s11/a86

PURL: 42411/s11/a86

RECIBIDO 21/01/2023 • ACEPTADO 27/02/2023 • PUBLICADO 30/03/2023

RESUMEN

Los microservicios han surgido como un estilo arquitectónico que ofrece muchas ventajas, pero también plantea desafíos. Uno de estos desafíos gira alrededor de las pruebas, puesto que una aplicación puede tener cientos o miles de servicios que funcionan juntos, y cada uno de ellos requiere ser probado a medida que evolucionan. Para superar este desafío, la automatización adquiere un papel clave, y junto con ella, el uso de herramientas de pruebas eficientes y eficaces..

Palabras claves: Microservicios, pruebas de software, seguridad, rendimiento.

ABSTRACT

Microservices have emerged as an architectural style that offers many advantages, but also poses challenges. One of these challenges revolves around testing, as an application may have hundreds or thousands of services running together, each requiring testing as they evolve. To overcome this challenge, automation takes on a key role, and along with it, the use of efficient and effective testing tools.

Keywords: *microservices, software testing, security, performance.*

INTRODUCCIÓN

Actualmente existe un gran cambio y desafío para el desarrollo de las aplicaciones de software, ya que atravesamos una época de transiciones a lo digital, ya que desde el 2020 la pandemia provocó apresuradamente un gran cambio en el manejo de la información. Entonces al tener una gran demanda de usuarios por consumo de X aplicación, se requiere una mayor eficiencia y

mantenibilidad de los servicios de las aplicaciones, por ello la transición de los servicios monolíticos a microservicios fue el mayor acierto para afrontar esta demanda.

Se pretende tratar todo lo referente a microservicios, ya que es el estándar actual en el desarrollo de software. Pero se necesita mantener una delgada línea entre la calidad y rapidez de entrega para estos servicios, entonces una manera confiable para respaldar que los microservicios desarrollados hagan lo que debían hacer, para esto es necesario realizar diferentes pruebas y estrategias para garantizar la calidad de los microservicios.

Para ello se necesitan herramientas eficientes y eficaces que acompañen el proceso (Heinrich et al., 2017) y faciliten la automatización de las pruebas, ayudando a la vez a enfrentar los retos asociados a las pruebas en el contexto de microservicios (Heinrich et al., 2017)(Stefano Munari, Sebastiano Valle, 2018). De ahí que el estudio de herramientas que soporten la automatización de pruebas para microservicios sea un terreno fértil de investigación.

Métodos y Metodología computacional

El objetivo es entender la importancia de las pruebas de software aplicadas a microservicios, así de esta manera encontrar la mejor manera para aumentar la calidad de los microservicios.

Arquitecturas de servicios

Las dos arquitecturas principales usadas para descomponer un sistema en servicios son: la arquitectura orientada a servicios (SOA, por sus siglas en inglés: service-oriented architecture) y la arquitectura de microservicios. A continuación se describe cada una de ellas.

Arquitectura SOA

Los servicios SOA consisten en un diseño de descomposición de servicios integrados en un proyecto por mecanismos de enrutamiento inteligente, que proporciona una gobernanza global (o administración centralizada) (Cerny et al., 2017). De ahí que es frecuente la connotación de "arquitectura orquestación" o monolítica. Por ello, los procesos de los servicios se encuentran vinculados a un único contexto general. Además, SOA se encarga de encapsular la funcionalidad de negocio en una única interfaz (servicios SOA como web service o restfull), por medio de la cual el productor proporciona la funcionalidad y el consumidor la puede solicitar. El Enterprise

Service Bus es el medio de comunicación entre productores y consumidores, permitiendo tener comunicación punto a punto (Quenum & Aknine, 2018).
SOA continúa siendo una arquitectura utilizada por las organizaciones. Sin embargo, SOA no se adapta bien a las necesidades de las metodologías ágiles (Chen, 2018), ni de movimientos como DevOps, Integración y Entrega Continua. Estas corrientes imponen la necesidad de implementar piezas pequeñas de software y colocarlas lo más rápido posible en los ambientes de producción. Por su parte, los proyectos creados con la arquitectura SOA adquieren grandes dimensiones y para colocar cambios en producción se requiere enviar todo el proyecto, por lo que no se alinean a las corrientes ágiles.

Arquitectura de microservicios

A inicios del año 2000 comienza la conceptualización de la arquitectura de los microservicios, de la mano del surgimiento de metodologías ágiles. Empresas como Amazon expresan la necesidad de una arquitectura con mayor capacidad de escalabilidad, en la cual sus componentes tengan mayor aislamiento e independencia (Carneiro & Schmelmer, 2016). En el año 2011 aparece por primera vez el término "microservicios" (Ueda et al., 2016)(Carneiro & Schmelmer, 2016). Los microservicios surgen como una alternativa de arquitectura para el diseño e implementación de sistemas distribuidos, dando como resultado sistemas con bajo acoplamiento de componentes que exhiben propiedades como flexibilidad, escalabilidad, adaptabilidad, y tolerancia a fallas, entre otros (Heinrich et al., 2017).

La definición con mayor aceptación por parte de la comunidad de software es la de Martín Fowler, pionero en la arquitectura de microservicios. Fowler define los microservicios como un enfoque para el desarrollo de aplicaciones compuesto por un conjunto de servicios pequeños, donde cada servicio se ejecuta en su propio proceso y se comunica a través de mecanismos ligeros, a menudo usando APIs Http (Fowler, M., Lewis, 2018)(Lewis, James; Fowler, 2014)[12]. Esto permite dividir sistemas complejos en múltiples componentes operacionales pequeños e independientes (Liu et al., 2016). El nivel de independencia de la arquitectura de microservicios permite que los servicios puedan ser implementados con diferentes lenguajes de programación y delega la gestión de los datos a cada servicio.

Relación entre SOA y microservicios

Los microservicios están relacionados con la arquitectura SOA al punto de que existe un debate sobre si los microservicios son una arquitectura nueva o son más bien una subcategoría (caso especial) de la arquitectura SOA (Quenum & Aknine, 2018)(Vera-Rivera, 2018). Ambas

arquitecturas están estrechamente relacionadas, puesto que los microservicios están basados en SOA, razón por la que a menudo se considera a los microservicios como una versión de SOA para sistemas distribuidos (Vera- Rivera, 2018) o simplemente una versión extendida de SOA (Quenum & Aknine, 2018).
No obstante, sí existen elementos que las diferencian. En el caso de los servicios SOA, los componentes se encapsulan en una única interfaz. Además, en la arquitectura SOA es necesaria una orquestación, que es facilitada por el Enterprise Service Bus. Este tiene la responsabilidad de ser el punto de integración para las comunicaciones con los servicios (Quenum & Aknine, 2018). La arquitectura de microservicios, por el contrario, no necesita del Enterprise Service Bus, dada la capacidad de independencia de los servicios que la componen. Los microservicios se diferencian por el nivel de granularidad de los servicios y su eliminación de la gobernanza central (de Camargo et al., 2016). Quizás la principal diferencia entre las arquitecturas SOA y microservicios es el propósito con el que fueron creadas (Heinrich et al., 2017): los microservicios surgen como arquitectura emergente ante la necesidad de aplicaciones de servicios con mayor capacidad de escalabilidad e independencia, mientras que la arquitectura SOA tiene características de centralización de los servicios controlado por el bus de servicios empresariales (Cerny et al., 2017) (Zúñiga-Prieto, Insfran, Abrahão, & Cano-Genoves, 2017)[13].

Pruebas end-to-end para microservicios

Las pruebas end-to-end surgieron en la última década como una herramienta valiosa para diagnosticar problemas de corrección y rendimiento en sistemas distribuidos (Las-Casas, Mace, Guedes, & Fonseca, 2018)[14].

Este tipo de pruebas se consideran pruebas funcionales de caja negra (Sotomayor et al., 2019), puesto que emulan el funcionamiento real del sistema y verifican su comportamiento (García, Gallego, Gortazar, & López, 2017). En el contexto de microservicios, las pruebas E2E son relevantes por ser la mejor manera de evaluar si el sistema como un todo funciona apropiadamente (Lei, Liao, Jiang, Yang, & Li, 2019).

Un aspecto clave de las pruebas E2E radica en las métricas y el "trazado" que pueden generar, los cuales permiten a los desarrolladores tomar decisiones sobre el rendimiento del sistema o la identificación de fallas. El "trazado" de las pruebas se refiere a la representación visual del flujo (Shahin, Babar, Zahedi, & Zhu, 2017), donde se muestra información como el orden de las información como el orden de las invocaciones, los eventos ejecutados, y la relación entre componentes y errores (Las-Casas et al., 2018).

Otras métricas importantes son las cargas de trabajo, y el uso de recursos y tiempo (Shahin et al., 2017). Estas métricas ayudan a comprender cómo funciona el sistema, a detectar anomalías en tiempo de ejecución, y a analizar por qué está fallando (Las-Casas et al., 2018).

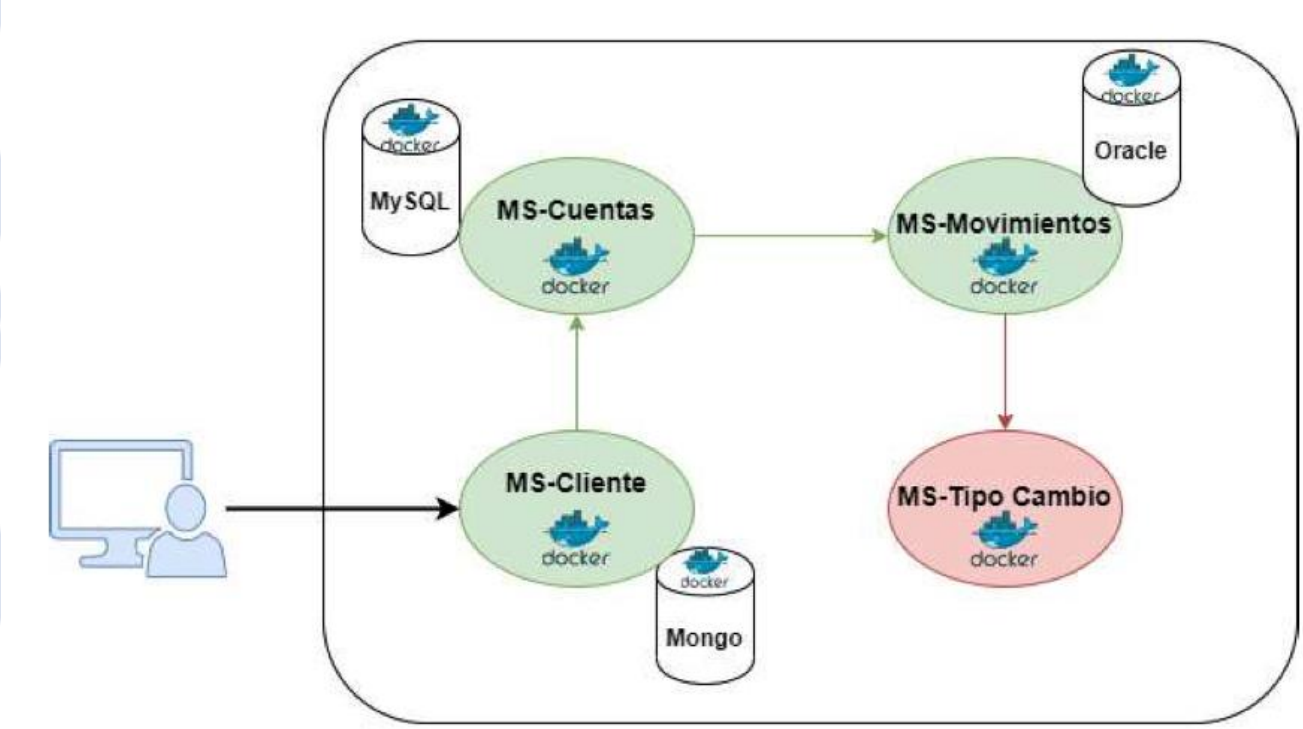


Figura 1: Ejemplo de arquitectura de microservicios para consultar el estado de cuenta

Prueba de la unidad

Una práctica que a menudo se pasa por alto cuando se prueban microservicios es la prueba unitaria. ¿Qué son las pruebas unitarias? Estas pruebas verifican que los métodos y clases que los desarrolladores escriben funcionan como se esperaba. Si bien las pruebas unitarias son una tarea muy técnica para los desarrolladores, un conjunto sólido de pruebas unitarias proporciona una red de seguridad crítica para detectar consecuencias no deseadas cuando los desarrolladores cambian el código. Y paga dividendos al alertar a los desarrolladores exactamente en qué parte del código han roto la funcionalidad existente.

Esta es una práctica valiosa para escribir software de alta calidad. Sin embargo, las pruebas unitarias por sí solas no son suficientes. Como analogía, el hecho de que todas las partes de un motor estén mecanizadas con una especificación perfecta no significa que el motor funcionará y funcionará como se espera.

Prueba de componentes

Esta prueba de microservicios no se concentra en cómo el desarrollador escribió el código de microservicios, sino que se enfoca en ejecutar el microservicio como una caja negra y probar el tráfico que se mueve a través de la interfaz. Desde la perspectiva de un único microservicio, ahora está probando el motor para asegurarse de que cumple con sus requisitos.
En la mayoría de los casos, está probando un servicio REST. Por lo que desea pruebas automatizadas que actúen como clientes del servicio, enviando varias solicitudes positivas y negativas al servicio y verificando las respuestas que devuelve el servicio.
Un desafío es que puede ser complejo y difícil probar microservicios de forma aislada porque a menudo llaman a muchos otros microservicios para responder a la solicitud de su cliente de prueba. Para probar un microservicio, es posible que necesite docenas más implementadas y disponibles para que su microservicio hable y lo pruebe correctamente.

Pruebas de integración

Al usar la virtualización de servicios para simplificar y estabilizar las pruebas del microservicio como un componente individual, también desea probar que el microservicio funciona con los otros microservicios REALES involucrados. Los desarrolladores a menudo hacen esto en una etapa de "QA" o "integración", donde muchos de los sistemas requeridos en el ecosistema general se implementan e integran juntos. Con esta práctica de prueba, está comenzando a ensamblar el automóvil para asegurarse de que todas las piezas encajen y funcionen juntas, pero aún no lo está probando en la carretera.

Pruebas de extremo a extremo

También se llama prueba del sistema. En algún momento, una gran red de microservicios tiene puntos de entrada donde interactúan los usuarios finales de la aplicación. Por ejemplo, una aplicación de Netflix en su Apple TV habla con microservicios dentro del centro de datos de Netflix. Pero representan solo una pequeña parte de su funcionalidad principal, que son componentes pequeños e individuales responsables de cosas específicas como un servicio de recomendaciones, un servicio de transmisión de video, un servicio de detalles de cuenta, etc. Así que esta también es una oportunidad para probar microservicios.
A menudo es dolorosamente lento y de alto mantenimiento probar estas interacciones automáticamente, ya sea web o móvil. Para las rutas críticas y los recorridos del usuario, es imprescindible, pero poder representar estas transacciones completas o de un extremo a otro desde la perspectiva del usuario final como una secuencia de llamadas de microservicio tiene muchos beneficios.

Básicamente, elimina la interfaz de usuario y simula todas las llamadas a la API que la interfaz de usuario hace a su arquitectura de microservicios para que pueda verificar que todos los microservicios funcionan juntos correctamente para el contexto de requisitos comerciales / de usuario final más amplio. Ahora está conduciendo el automóvil en la carretera, poniéndose en el lugar de sus clientes y asegurándose de que el automóvil cumpla sus promesas.

Pruebas de seguridad

Los piratas informáticos pueden explotar áreas bajo el paraguas de los microservicios. Por lo tanto, los desarrolladores deben probar los microservicios a fondo para que estén protegidos contra las vulnerabilidades de seguridad. Parasoft Jtest tiene tecnología de análisis de código estático que escanea el código fuente subyacente e identifica las debilidades de codificación segura para que los desarrolladores puedan corregirlas. Parasoft también permite a los evaluadores reutilizar los casos de prueba funcionales que han escrito en SOAtest para realizar pruebas de penetración del microservicio.

Pruebas de carga y rendimiento

Para asegurarse de que el microservicio pueda mantener los SLA (acuerdos de nivel de servicio), los desarrolladores deben comprender cómo funcionan los SLA bajo carga y también determinar los puntos de ruptura.

Resultados y discusión

Discusión

Cinco consejos para probar microservicios

1. Considere cada servicio como un módulo de software. Realice pruebas unitarias en un servicio como lo haría con cualquier código nuevo. En la arquitectura de microservicios, cada servicio se considera una caja negra. Por lo tanto, pruebe cada uno de manera similar.
2. Determine los vínculos fundamentales en la arquitectura y pruébalos. Por ejemplo, si existe un vínculo sólido entre el servicio de inicio de sesión, la interfaz que muestra los detalles del usuario y la base de datos para obtener los detalles, pruebe estos vínculos.
3. No se limite a probar escenarios de caminos felices. Los microservicios pueden fallar y es importante simular escenarios de fallas para generar resiliencia en su sistema.

4. Haz tu mejor esfuerzo para probar en todas las etapas. La experiencia ha demostrado que los probadores que utilizan una combinación diversa de prácticas de prueba, comenzando en el desarrollo y progresando a través de alcances de prueba más grandes, no solo aumentan la posibilidad de que los errores se revelen, sino que lo hacen de manera eficiente. Esto es particularmente cierto en entornos virtuales complicados donde existen pequeñas diferencias entre varias bibliotecas y donde la arquitectura de hardware subyacente puede producir resultados imprevistos e indeseables a pesar de la capa de visualización.
5. Utilice "pruebas canarias" en código nuevo y pruebe en usuarios reales. Asegúrese de que todo el código esté bien instrumentado. Y también utilice toda la supervisión que ofrece su proveedor de plataforma. Esto cumple con las pruebas de "cambio a la izquierda" con las pruebas de "cambio a la derecha" porque también está probando "en la naturaleza".

Tres estrategias para maximizar el ROI de las pruebas de microservicios

1. Aumente la calidad de la cobertura de prueba de API funcional con IA para garantizar que los servicios implementados cumplan con los requisitos.
2. Automatice flujos de trabajo complejos basados en eventos para acelerar las pruebas.
3. Mejorar el entorno de prueba para mejorar la confiabilidad y estabilidad de las pruebas.

Selección de herramientas de pruebas E2E para microservicios

Esta primera fase, correspondiente al primer objetivo específico de la investigación, requirió en primera instancia identificar las herramientas de pruebas E2E para microservicios que se iban a evaluar mediante el estándar IEEE-14102-2010. En segundo lugar, se definieron los criterios mediante los cuales se evaluarían las herramientas. En tercer lugar, se aplicaron los procesos definidos en el estándar, con el fin de evaluar y posteriormente seleccionar las mejores herramientas. A continuación se describen estas tres actividades.

Aplicación del estándar IEEE 14102-2010

Para evaluar y seleccionar herramientas El estándar IEEE 14102-2010 se divide en cuatro procesos: preparación, estructuración, evaluación y selección. A continuación se detallan las actividades realizadas para cada proceso.

1. Preparación: establecimos el objetivo de la evaluación y definimos los criterios de selección.
2. Estructuración: definimos la lista de candidatos, y establecimos los pesos para cada criterio de selección, según su importancia relativa. Además, definimos las

características a evaluar por cada criterio. Por ejemplo: el criterio de métrica tiene un peso de 30 sobre 100; si la herramienta genera métricas de trazados y tiempos obtiene un valor de 30, si genera únicamente una de las dos, se le asigna 15, y si no genera ninguna obtiene 0.

3. Evaluación: evaluamos la lista de candidatos, valorando las herramientas según los criterios y pesos establecidos.
4. Selección: seleccionamos las dos herramientas con mejor puntaje, ya que son las herramientas que cuentan con las características más apropiadas para realizar pruebas end- to-end para microservicios. Las herramientas seleccionadas obtuvieron una calificación similar.

Conclusiones

Los fundamentos de las pruebas de microservicios no son nuevos en comparación con los servicios web tradicionales o las pruebas SOA, pero la importancia de hacerlo solo se ha vuelto más crítica en los sistemas modernos. Al repasar por los diferentes conceptos para obtener una mayor calidad en los microservicios, conocimos lo suficiente como para tener en cuenta muchos métodos para realizar eficazmente pruebas de software para microservicios, por ende, aumentar su calidad para enfrentar a desafíos actuales.

Referencias

[1]. Aderaldo, C. M., Mendonça, N. C., Pahl, C., & Jamshidi, P. (2017). Benchmark Requirements for Microservices Architecture Research. 2017 IEEE/ACM 1st International Workshop on Establishing the Community-Wide Infrastructure for Architecture-Based Software Engineering (ECASE), 8–13. https://doi.org/10.1109/ECASE.2017.4

[2]. Estructuración: definimos la lista de candidatos, y establecimos los pesos para cada criterio de selección, según su importancia relativa. Además, definimos las características a evaluar por cada criterio. Por ejemplo: el criterio de métrica tiene un peso de 30 sobre 100; si la herramienta genera métricas de trazados y tiempos obtiene un valor de 30, si genera únicamente una de las dos, se le asigna 15, y si no genera ninguna obtiene 0.

[3]. Antichi, G., & Rétvári, G. (2020). Full-Stack SDN: The Next Big Challenge? Proceedings of the Symposium on SDN Research, 48–54. https://doi.org/10.1145/3373360.3380834

[4]. Arcuri, A. (2019). RESTful API Automated Test Case Generation with EvoMaster. ACM Trans. Softw. Eng. Methodol., 28(1), 3:1--3:37. https://doi.org/10.1145/3293455

[5]. \item BA, K., & Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. 2.

[6]. Gil, D. G., & D\'\iaz-Heredero, R. A. (2018). A Microservices Experience in the Banking Industry. Proceedings of the 12th European Conference on Software Architecture: Companion Proceedings. https://doi.org/10.1145/3241403.3241418

[7]. Gu, G., Hu, H., Keller, E., Lin, Z., & Porter, D. E. (2017). Building a Security OS With Software Defined Infrastructure. Proceedings of the 8th Asia-Pacific Workshop on Systems. https://doi.org/10.1145/3124680.3124720

[8]. Harsh, P., Ribera Laszkowski, J. F., Edmonds, A., Quang Thanh, T., Pauls, M., Vlaskovski, R., ... Gallego Carrillo, M. (2019). Cloud Enablers For Testing Large-Scale Distributed Applications. Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing Companion, 35– 42. https://doi.org/10.1145/3368235.3368838

[9]. Hasselbring, W., & Steinacker, G. (2017). Microservice architectures for scalability, agility and reliability in e-commerce. Proceedings - 2017 IEEE International Conference on Software Architecture Workshops, ICSAW 2017: Side Track Proceedings, 243–246. https://doi.org/10.1109/ICSAW.2017.11

[10]. Heinrich, R., van Hoorn, A., Knoche, H., Li, F., Lwakatare, L. E., Pahl, C., ... Wettinger, J. (2017). Performance Engineering for Microservices: Research Challenges and Directions. Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion, 223–226. https://doi.org/10.1145/3053600.3053653

[11]. Heorhiadi, V., Rajagopalan, S., Jamjoom, H., Reiter, M. K., & Sekar, V. (2016). Gremlin: Systematic Resilience Testing of Microservices. Proceedings - International Conference on Distributed

[12]. MARTÍNEZ HERNÁNDEZ, Cristian Fernando. Evaluación de una herramienta de pruebas end-to-end para microservicios implementados en Java y Node. js.

[13]. Fowler, M., Lewis, J. (2018). Microservices. 1–15.

[14]. Zúñiga-Prieto, M., Insfran, E., Abrahão, S., & Cano-Genoves, C. (2017). Automation of the incremental integration of microservices architectures. In Lecture Notes in Information Systems and Organisation. https://doi.org/10.1007/978-3-319-52593-8_4

Uso de las redes neuronales para determinar la calificación de una aplicación publicada en Google Play Store

Use of neural networks to determine the rating of an application published in the Google Play Store

Rudy Roberto Tito Durand
Universidad Nacional de San Agustín.
Arequipa, Perú.
rtitod@unsa.edu.pe

Marcelo A. Guevara Gutierrez
Universidad Nacional de San Agustín.
Arequipa, Perú.
mguevarag@unsa.edu.pe

Jeampier Anderson Moran Fuño
Universidad Nacional de San Agustín.
Arequipa, Perú.
jmoran@unsa.edu.pe

Edsel Yael Alvan Ventura
Universidad Nacional de San Agustín.
Arequipa, Perú.
ealvan@unsa.edu.pe

ARK: ark:/42411/s11/a87
PURL: 42411/s11/a87

RECIBIDO 05/02/2023 • ACEPTADO 13/03/2023 • PUBLICADO 30/03/2023

RESUMEN

La inteligencia artificial es la combinación de algoritmos escritos en forma de código computacional con el fin de que se ejecuten en una computadora para emular comportamientos similares a la inteligencia humana. En este trabajo, se buscará el uso de las redes neuronales, las cuales son parte de la inteligencia artificial y, nos permiten solucionar problemas de predicción. Se modificará un dataset basado en las descargas de aplicaciones de Google Play Store y sus características y se procesará la información para obtener información de salida.

Palabras claves: Redes neuronales, numpy, flujo tensorial, valores perdidos, valores atípicos.

ABSTRACT

Artificial intelligence is the combination of algorithms written in the form of computer code in order to be executed on a computer to emulate behaviors similar to human intelligence. In this work, the use of neural networks will be sought, which are part of artificial intelligence and allow

us to solve prediction problems. A dataset based on Google Play Store app downloads and their features will be modified and the information processed to obtain output information.

Keywords: N*eural networks, numpy,tensorflow, missing values, outliers.*

INTRODUCCIÓN

La inteligencia artificial es la combinación de algoritmos escritos como código computacional con el fin de que las computadoras en donde se ejecuten estos códigos ejecuten comportamientos similares a la inteligencia humana. Desde su invención, ha tenido muchas aplicaciones. Las aplicaciones que se le pueden dar a la Inteligencia Artificial han estado incrementando en casi todos los dominios de la vida humana desde los sistemas de recomendación para compras, gestión de inventarios y aspectos logísticos hasta la solución de problemas de salud [1].

Existe un gran número de problemas dentro de la ingeniería en los cuales se puede aplicar Inteligencia Artificial pero debido a la naturaleza de los datos con los que se trabaja no se puede tratar de las formas tradicionales. Esto tardó ya que recién en los años 70 entendieron el hecho de que era imposible el manejo de los millones de combinaciones posibles necesarias para poder evaluar situaciones reales, esto debido al infinito número de perturbaciones que pueden existir dentro de una sola situación [2]. Las redes neuronales son sistemas de procesamiento de información, que a su vez son reconocidas como un paradigma matemático de computación. Son ampliamente utilizadas en diversos ambientes teóricos y prácticos, son un conjunto de unidades llamadas neuronas artificiales conectadas [3].

Como un problema relacionado con los aspectos logísticos está relacionado con el desarrollo de software para móviles y si una empresa que inicia puede saber si una aplicación va a ser exitosa o no podría condicionar el desarrollo, evitando pérdidas económicas o buscando mayores ganancias al redefinir el proyecto. Estudios realizados por diferentes organizaciones demuestran que la tasa de éxito de los proyectos está alrededor del 32%, y de proyectos fallidos alrededor del 24% [4].

A través de este proyecto se busca predecir a partir de factores vinculados al software desarrollado , que calificación se puede obtener al ser puesto al público, a través de Google Play, que es la tienda de aplicaciones creada por Google donde puedes encontrar juegos, películas, música, libros y más, la cual está disponible para cualquier dispositivo móvil que cuente con sistema operativo Android [5], aunque hay estudios que aclaran que la calificación de una aplicación no necesariamente indica que será la más popular [6].

Para lograr tal fin, se usará una de las herramientas de la inteligencia artificial la cual son las redes neuronales, porque son un método de resolver problemas, por ejemplo, el mapeo

autoorganizado suelen ser utilizado como herramienta para la predicción de tendencias y como clasificador de conjuntos de datos. [7]. Las cuales consisten en unidades de procesamiento interconectadas de manera densa, llamadas neuronas; una de sus características más importantes es la capacidad del aprendizaje adaptativo mediante datos dados, además pueden ser combinadas con otras herramientas como la lógica difusa, los algoritmos genéticos o los sistemas expertos [8].

Se buscará implantar un modelo estándar que a su vez, requerirá de sistemas de información apoyados en datos históricos, para lo cual estas redes neuronales leerán la información de un archivo CSV, el cual será tratado en una fase preliminar para eliminar o completar información errónea obtenida en la fase de recolección.

Marco Teórico

Logística

Para Ferrel, Hirt, Adriaenséns, Flores y Ramos, la logística es "una función operativa importante que comprende todas las actividades necesarias para la obtención y administración de materias primas y componentes, así como el manejo de los productos terminados, su empaque y su distribución a los clientes"[9]

Redes neuronales

Las redes neuronales simulan la estructura y el comportamiento del cerebro, usan lo que se conoce como procesos de aprendizaje para dar solución a los problemas para los que fueron programadas; son un conjunto de algoritmos matemáticos que encuentran las relaciones no lineales entre conjuntos de datos; suelen ser utilizadas como herramientas para la predicción de tendencias y como clasificadoras de conjuntos de datos.[18]

Se denominan neuronales porque están basadas en el funcionamiento de una neurona biológica cuando procesa información.[10]

Data Cleaning

Se conoce como al proceso que se encarga de corregir los errores en los datos, se convierte por tanto en un mecanismo necesario para que las estadísticas, los informes y en última instancia las decisiones que se tomen por los directivos sean confiables, pues en la medida en que esté garantizada la calidad de los datos, así mismo habrá seguridad y fiabilidad en las acciones posteriores que se produzcan a partir de su análisis.[11]

Data Set

Un Dataset no es más que un conjunto de datos tabulados en cualquier sistema de almacenamiento de datos estructurado. El término hace referencia a una única base de datos de origen, la cual se puede relacionar con otras, cada columna del Dataset representa una variable y cada fila corresponde a cualquier dato que estemos tratando[12]

Red Neuronal Secuencial

Una red neuronal sequential es un tipo de modelo de red neuronal que se conforma por capas de neuronas, cada capa se agrega una después de la otra. La metodología usada durante la construcción del modelo es paso a paso y trabajando en una sola capa en un momento determinado.

Herramientas

Colab Research

Es una herramienta para escribir y ejecutar código Python en la nube.de Google. También es posible incluir texto enriquecido, “links” e imágenes. En caso de necesitar altas prestaciones de cómputo, el entorno permite configurar algunas propiedades del equipo sobre el que se ejecuta el código.[13]

Python

Python es un lenguaje de programación de alto nivel que se utiliza para desarrollar aplicaciones de todo tipo. A diferencia de otros lenguajes como Java o .NET, se trata de un lenguaje interpretado, es decir, que no es necesario compilarlo para ejecutar las aplicaciones escritas en

Python, sino que se ejecutan directamente por el ordenador utilizando un programa denominado interpretador, por lo que no es necesario "traducirlo" a lenguaje máquina.[15]

TensorFlow

Tensor Flow es una biblioteca de software de código abierto para computación numérica, que utiliza gráficos de flujo de datos. Los nodos en las gráficas representan operaciones matemáticas, mientras que los bordes de las gráficas representan las matrices de datos multidimensionales (tensores) comunicadas entre ellos.
Tensor Flow es una gran plataforma para construir y entrenar redes neuronales, que permiten detectar y descifrar patrones y correlaciones, análogos al aprendizaje y razonamiento usados por los humanos.[16]

Keras

Keras es una biblioteca que funciona a nivel de modelo: proporciona bloques modulares sobre los que se pueden desarrollar modelos complejos de aprendizaje profundo. A diferencia de los frameworks, este software de código abierto no se utiliza para operaciones sencillas de bajo nivel, sino que utiliza las bibliotecas de los frameworks de aprendizaje automático vinculadas, que en cierto modo actúan como un motor de backend para Keras.[17]

Métodos y Metodología computacional

Se usó para el desarrollo de este detector de sitios web fraudulentos usamos el modelo CRISP-DM, la cual comprende de seis fases: análisis del problema, análisis de Datos, preparación de los Datos, modelado, evaluación y explotación [10].

- La fuente de información fue un Dataset.csv, esto fue sacado de internet con datos reales, para poder construir un árbol de decisión utilizando las librerías *sklearn*, y se hizo uso de las funciones que nos brinda.
- Análisis del problema: Identificamos qué tipos de clasificación tendrán las páginas que se van a probar, viendo que pueden ser tres, y también identificamos la información necesaria para poder hacer esta clasificación.
- Análisis de Datos: con el *dataset* que obtuvimos, vimos que cantidad de información tiene, y las variables lingüísticas las clasificamos en números para su uso correcto con la librería que se trabajaría.
- Preparación de Datos: Se hizo una limpieza del *dataset*, con las funciones que nos brinda la libre, como eliminación de campos vacíos, datos irrelevantes.
- Modelado: Seleccionamos la técnica adecuada para poder hacer la clasificación.

- Evaluación: Tuvimos que corroborar efectivamente que el modelo escogido se ajuste a lo que estamos buscando, en este caso poder clasificar los diferentes sitios web.

Resultados y Discusión

Análisis del Problema

Los datos de las aplicaciones de Play Store tienen un enorme potencial para impulsar a las empresas para la creación de aplicaciones exitosas.
Por lo cual pretendemos realizar una red neuronal que permita la predicción de ratings usando como datos de entrenamiento una fuente de datos provenientes de un dataset recolectado de las descargas de la Google Play Store, este consiste en datos extraídos de la web de 10,000 aplicaciones de Play Store.

Análisis de los Datos

Fase de data cleaning:

El dataset posee 13 variables (columnas) las cuales son: App, Category, Rating, Reviews, Size, Installs,Type, Price, Content Rating, Genres, Last Updated, Current Ver y Android Ver. Se cuenta con 10841 filas. Luego de la descarga del dataset se procedió al análisis para eliminar los valores missing.

Fase de Transformación:

En la fase de transformación, en la sección Genre se detectó que existían datos múltiples valores en la variable, pues esta variable proveía mas de un valor en sus celdas, por lo que se decidió dividir la cadena en dos tokens usando el símbolo ";" y quedándonos con el primer token obtenido.

En el caso de la variable "Content Rating","Category", se decidió conservar cada uno de sus datos para la codificación. Mientras que para las variables "Last Updated", se procedió a la transformación de sus valores restando la fecha indicada en la fila y el año de publicación aproximado del dataset. La variable "Android Ver", se transformó en un número de tipo float, debido a que el orden de precedencia de las versiones esta determinado por la posición que ocupa. Con respecto a la variable "Size", se quitó y transformó cada valor por el sufijo que tuvo al final. Por ejemplo si el valor es "233k", entonces esto indicaría que debemos multiplicar 233*0.001, mientras que si se tenía "233M", se conservaría igual, en síntesis, se transformó y se

puso en una misma unidad de medida en la variable "Size". Con respecto a la variable "Installs" se transformó a números enteros y se les redimensiona a un valor general que es el valor actual entre 100,000.
Cabe destacar que todos los valores "Varies with Device" en las filas de "Android Ver", "Size", entre otras, fueron eliminadas pues no representan una gran cantidad de datos y se decidió eliminarlas debido a la falta de datos que este valor provee.

Fase de eliminación de Outliers:

Para la fase de eliminación de outliers se procedió a identificar los cuartiles de los datos que fueron convertidos en la anterior fase. El método usado para la eliminación de outliers que se usó fue la eliminación por el valor máximo y mínimo dados por el Método Caja y Bigotes.
En el cual se determina que el valor máximo y minimo estan dados en la siguiente formula:
máximo = q3+(1.5*iqr)
mínimo = q1+(1.5*iqr)
Donde "q1" es el cuartil que divide a los datos en 25% y 75%. Mientras que "q3", representa el tercer cuartil representa que los valores por debajo de ese valor son el 75% de los datos. Con respecto a "iqr", es la diferencia entre el tercer cuartil y el primero. Los valores de máximo y mínimo son los valores que nos ayudarán a identificar los outliers de nuestra gráfica de Caja y Bigotes.
Se procedió a reemplazar(en algunos casos eliminar) a los valores outliers con la media del total de valores, de esta manera no afectamos al balance de los datos y por lo tanto se conservó los demás atributos de esas filas para el entrenamiento de nuestro modelo Secuencial de red Neuronal.

Fase de creación de red neuronal:

- En esta fase se identifican cada uno de los valores como "Categórico" o "Continuo", por lo tanto las variables "Content Rating","Category" se consideran categóricos mientras que las variables "Android Ver", "Installs","Reviews" son considerados valores continuos, pues los valores pueden incrementar o disminuir, pero no se mantienen en un rango fijo de valores a los cuales es posible clasificarlos.

Lueg
o de este análisis, se procedió a categorizar todos los valores como una variable en específico, es decir todos los valores categóricos han sido puestos como características propias de la Aplicación. De este modo si hay las características categóricas A,B y C, se le asigna este tipo de codificación a sus valores:

Nombre de la Aplicación	A	B	C

App A	1	0	0

La tabla anterior representa que la aplicación "App A" tiene la característica A, y no tiene las características B y C. Esto beneficia al modelo Secuencial, pues los valores altos o secuenciales 1,2,3,4,5 pueden afectar a los pesos de cada
neurona. Por lo tanto este modelo de codificación tendrá más precisión a la hora de entrenar nuestro modelo de red neuronal.

Preparación de los datos

Los valores missings se trataron asi:

- A continuación se muestran la tabla de variables que contienen valores missing y sus respectivos porcentajes:

	Total	%
Rating	1474	13.597
Current Ver	8	0.074
Android Ver	3	0.028
Type	1	0.009
Content Rating	1	0.009

- Luego de eliminar los valores missing en la variable objetivo "Rating", es muestra los valores missing restantes:

DATAFRAME REMOVIDO

	Total	%
Current Ver	4	0.043
Android Ver	3	0.032
Content Rating	1	0.011

- Se eliminan los otros missing Values
- Se observa las filas anteriores con respecto las nuevas filas sin valores missing:

```
Original: (10841, 13)
Clean dataframe: (9360, 13)
```

Se procede a un análisis de las columnas a utilizar

- 'App' : No se utilizara porque es el nombre
- 'Category': Si se utilizara
- 'Rating': Es nuestro OUTPUT
- 'Reviews': Si se utilizara
- 'Size': Si se utilizara
- 'Installs': Si se utiliza
- 'Type': No se utilizara porque la mayoría son valores Free
- 'Price': No se utilizara porque la mayoría son valores 0
- 'Content Rating': Si se utilizara
- 'Genres': No se utilizara porque era redundante respecto a Category.
- 'Last Updated': Se utilizara
- 'Current Ver': No se utilizará porque es un valor subjetivo dependiendo del creador de la App.
- 'Android Ver': Se utilizara

En este sentido las únicas variables a usar son: Reviews, Category, Rating,Size, Installs, Content Rating, Last Updated, Current Ver, Android Ver.

Se arreglan los datos de algunas columnas

- Al ya haber realizado la obtención de los datos, se deben analizar cuáles requieren de una transformación para que sean útiles.

.

- Para arreglar los datos del campo Size debemos convertir todos sus datos a la misma base, en este caso serían kilobytes, para ello se asigna un valor para poder convertir los datos a esta base, y esta acción se repite en todas los registros de la tabla eliminando el carácter y transformando el dato en un float y reemplazandolo en la tabla..

```
print("DF-SHAPE:",df.shape)
M = 1
k = 0.001
varies_devices=0
for index, row in df.iterrows():
  if(row["Size"][-1] == "M"):
    r = row["Size"][:-1]
    s = float(r)
    s = s*M
    row["Size"] = s*M
  elif(row["Size"][-1] == "k"):
    r = row["Size"][:-1]
    s = float(r)
    s = s*k
    row["Size"] = s*k
```

- Ahora también existe un valor no numérico "varies in device" que básicamente indica que el tamaño depende del dispositivo en el cual se instale la aplicación evaluada, debido a que no debemos indagar el peso de la aplicación en otra fuente, debemos descartar estos y para eso tenemos un else dentro del bucle que contabiliza cuantas veces se repite este valor en el campo Size de la tabla

```
    else:
      varies_devices+=1
      pass#sacar o rescatar los "varies with device"
    df.at[index,"Size"] = s
      #df.set_value(index,'Size',s)
```

- Después de haber evaluado cada fila de la tabla se nos muestra la siguiente información y los datos ya registrados como floats de 64 bits

```
#plt.show()
print("Varies with device: ", varies_devices)
df["Size"].describe()
```

```
DF-SHAPE: (9360, 13)
Varies with device:  1637
count    9360.000000
mean       23.143466
std        23.245147
min         0.008500
25%         5.500000
50%        15.000000
75%        33.000000
max       100.000000
Name: Size, dtype: float64
```

- Como ya se ha evaluado toda la tabla ya tenemos una vista más limpia de los datos de el campo Size con datos numéricos y ya retirados los valores "varies in device"

```
df["Size"].head(10)
```

```
0    19.0
1    14.0
2     8.7
3    25.0
4     2.8
5     5.6
6    19.0
7    29.0
8    33.0
9     3.1
Name: Size, dtype: float64
```

- Para arreglar los datos del campo Installs, lo que se requiere es eliminar el signo "+", además que en algunos casos se están trabajando con números demasiado

grandes, por lo que convendría reducirlos para no ocupar tanto espacio por el tamaño de los números

```
array(['10,000+', '500,000+', '5,000,000+', '50,000,000+', '100,000+',
       '50,000+', '1,000,000+', '10,000,000+', '5,000+', '100,000,000+',
       '1,000,000,000+', '1,000+', '500,000,000+', '100+', '500+', '10+',
       '5+', '50+', '1+'], dtype=object)
```

- Para lo cual se ha tomado como base el valor de 100.000 para poder reducir el tamaño de los valores originales. logrando así reducir espacio, y al retirar el símbolo "+" se establecen como floats de 64 bits

```
for i, row in df.iterrows():
  tmp = row["Installs"].replace("+","")
  tmp = tmp.replace(",","")
  df.at[i,"Installs"] = str(float(tmp)/100000)

df["Installs"] = df["Installs"].astype(float)
```

- Como ya se ha evaluado toda la tabla ya tenemos una vista más limpia de los datos de el campo Installs con datos numéricos más manejables

```
0      0.1
1      5.0
2     50.0
3    500.0
4      1.0
5      0.5
6      0.5
7     10.0
8     10.0
9      0.1
10    10.0
11    10.0
Name: Installs, dtype: float64
```

- Ahora en el caso del campo Reviews, lo único que se hace es establecer el tipo del campo como floats ya que los datos originales de por sí son manejables y no requieren mayor conversion

```
df_test = df.copy()
df_test["Reviews"] = df_test["Reviews"].astype(float)
```

- Aquí se ve ya convertido a float los datos iniciales

Reviews
159.0
967.0
87510.0
215644.0
967.0

- Para el caso de Genres el obstáculo que tenemos es que algunos registros poseen más de un géneros y no podemos clasificar las fusiones de géneros para realizar análisis, afortunadamente cuando es más de un género en el registro están separados por un ";" por lo que este será el límite para separar dichas fusiones obteniendo todos los valores diferentes dentro de la tabla.

```
df_test["Genres"].unique()
```

```
array(['Art & Design', 'Art & Design;Pretend Play',
       'Art & Design;Creativity', 'Auto & Vehicles', 'Beauty',
       'Books & Reference', 'Business', 'Comics', 'Comics;Creativity',
       'Communication', 'Dating', 'Education;Education', 'Education',
       'Education;Creativity', 'Education;Music & Video',
       'Education;Action & Adventure', 'Education;Pretend Play',
       'Education;Brain Games', 'Entertainment',
       'Entertainment;Music & Video', 'Entertainment;Brain Games',
       'Entertainment;Creativity', 'Events', 'Finance', 'Food & Drink',
       'Health & Fitness', 'House & Home', 'Libraries & Demo',
       'Lifestyle', 'Lifestyle;Pretend Play',
       'Adventure;Action & Adventure', 'Arcade', 'Casual', 'Card',
       'Casual;Pretend Play', 'Action', 'Strategy', 'Puzzle', 'Sports',
       'Music', 'Word', 'Racing', 'Casual;Creativity',
       'Casual;Action & Adventure', 'Simulation', 'Adventure', 'Board',
       'Trivia', 'Role Playing', 'Simulation;Education',
       'Action;Action & Adventure', 'Casual;Brain Games',
       'Simulation;Action & Adventure', 'Educational;Creativity',
       'Puzzle;Brain Games', 'Educational;Education', 'Card;Brain Games',
       'Educational;Brain Games', 'Educational;Pretend Play',
       'Entertainment;Education', 'Casual;Education',
```

- Luego de hallar todos los valores diferentes pasamos a separar las fusiones en elementos individuales y así obtenemos los géneros únicos

```
[ ]  df_test["Genres"]=df_test["Genres"].str.split(';').str[0]
```

```
[ ]  df_test["Genres"].unique()
```

```
array(['Art & Design', 'Auto & Vehicles', 'Beauty', 'Books & Reference',
       'Business', 'Comics', 'Communication', 'Dating', 'Education',
       'Entertainment', 'Events', 'Finance', 'Food & Drink',
       'Health & Fitness', 'House & Home', 'Libraries & Demo',
       'Lifestyle', 'Adventure', 'Arcade', 'Casual', 'Card', 'Action',
       'Strategy', 'Puzzle', 'Sports', 'Music', 'Word', 'Racing',
       'Simulation', 'Board', 'Trivia', 'Role Playing', 'Educational',
       'Music & Audio', 'Video Players & Editors', 'Medical', 'Social',
       'Shopping', 'Photography', 'Travel & Local', 'Tools',
       'Personalization', 'Productivity', 'Parenting', 'Weather',
       'News & Magazines', 'Maps & Navigation', 'Casino'], dtype=object)
```

- Ya habiendo obtenido los géneros individuales ya se puede realizar un conteo de registros coincidentes con cada género.

```
df_test.groupby("Genres")["App"].nunique()
```

Genres	
Action	304
Adventure	78
Arcade	186
Art & Design	61
Auto & Vehicles	73
Beauty	42
Board	57
Books & Reference	171
Business	263
Card	46
Casino	37
Casual	217
Comics	54
Communication	258
Dating	134
Education	498
Educational	93
Entertainment	502
Events	45
Finance	302
Food & Drink	94
Health & Fitness	246
House & Home	62
Libraries & Demo	63
Lifestyle	302
Maps & Navigation	118
Medical	291

-

- Para el campo Last Update ya que se esta trabajando con fechas debemos convertir la cadena de caracteres de toda la tabla a un valor de tipo Date, y como el formato en el que se presentan en estas cadenas es en casi toda la tabla es solo separar los valores y registrarlos con el tipo Date

```
print("CURRENT DAY : ",current_date)
date_pub = current_date - timedelta(days=365*4)#FUE HACE 4 AÑOS

# date_now = datetime.today().date()#BUSCAR LA FECHA DE PUBLICACION DEL DATASET
date = date_pub
NO = 0
import math
for i, row in df_test.iterrows():
  tmp = row["Last Updated"]
  try:
    format = datetime.strptime(tmp,"%B %d, %Y").date()
    days = (date_pub-format).days
    df_test.at[i,"Last Updated"] = days
  except:
    df_test.at[i,"Last Updated"] = "NO"
    print("No")
    NO +=1
condition = df_test.loc[df_test["Last Updated"] == "NO"]
df_test.drop(condition.index, inplace=True)
print("Total Exceptions: ",condition.shape,"->", NO)
print("Total Success: ", df_test.shape)
```

- Ya para trabajar es más sencillo trabajar con datos numéricos que con fechas cambiaremos el significado del campo, pasando de registrar la fecha de la última actualización, se contarán los días desde la última actualización de cada aplicación de la tabla hasta el dia de hoy y pasa de ser de tipo Date a tipo float

```
[ ]  #CONVIRTIENDO A FLOAT-> PARA MAS ADELANTE USARLO EN RED NEURONAL
     df_test["Last Updated"] = df_test["Last Updated"].astype(float)
```

```
[ ]  df = df_test
     df["Last Updated"].head(30)
     df.rename(columns = {'Last Updated':'Days passed',}, inplace = True)
```

- Por último para arreglar los datos de Android ver, hay caracteres que obstaculizan el análisis, la cadena “and up”, “nan” y “NO”, lo primero se debe retirar dichos caracteres

```
for i, row in df_test.iterrows():
  val = row["Android Ver"]
  if(val == "Varies with device" or val=="nan"):
    tmp = "NO"
    total_out+=1
  else:
    tmp = val.replace("and up","")
    total_succ += 1
  df_test.at[i,"Android Ver"] = tmp
print("#Success: ", total_succ)
print("#Fails: ", total_out)
print("#Total", df_test.shape)
```

```
#Success:  8041
#Fails:  1319
#Total (9360, 13)
```

- Luego tenemos algunos rangos por ejemplo "4.0 - 5.0" en el cual se nos muestran 2 valores, en este caso se toma el primer valor osea el menor, así obtenemos todos los valores diferentes y realizamos un conteo

```
for i, row in df_test.iterrows():
  val = row["Android Ver"]
  if val != "NO":
    val = val[0:3]
    df_test.at[i,"Android Ver"] = val
```

```
[ ] df_test.groupby("Android Ver")["App"].nunique()
```

```
Android Ver
1.0       2
1.5      15
1.6      87
2.0      34
2.1     112
2.2     203
2.3     780
3.0     201
```

- Ya por ultimo convertimos estos valores de las versiones a floats

```
arr= np.array(df_test["Android Ver"].unique())
arr = arr[arr != "NO"]
another = np.array([])
for item in arr:
  another = np.append(another, float(item))
arr.astype(float)
print(arr,another)
mean = np.mean(another)
print("mean: ", mean)
```

```
['4.0' '4.2' '4.4' '2.3' '3.0' '4.1' '2.2' '5.0' '6.0' '1.6' '1.5' '2.1'
 '7.0' '4.3' '2.0' '3.2' '5.1' '7.1' '8.0' '3.1' '1.0'] [4.  4.2 4.4 2.3 3.  4.1 2.2 5.  6.  1.6 1.5 2.1 7.  4.3 2.  3.2 5.1 7.1
 8.  3.1 1. ]
```

```
df_test["Android Ver"].astype(float)#convirtiendo a float
```

```
0       4.000
1       4.000
2       4.000
3       4.200
4       4.400
        ...
9355    4.100
9356    4.100
9357    4.100
9358    3.867
9359    3.867
Name: Android Ver, Length: 9360, dtype: float64
```

1. **Los outliers se trataron asi:**

En el paso anterior se arreglaron los datos con el fin de obtener coherencia en el resultado de la red neuronal. Se tiene el dataset con 13 columnas y 9360 filas

	App	Category	Rating	Reviews	Size	Installs	Type	Price	Content Rating	Genres	Days passed	Current Ver	Android Ver
0	Photo Editor & Candy Camera & Grid & ScrapBook	ART_AND_DESIGN	4.1	159.0	19.0	0.100	Free	0	Everyone	Art & Design	223.0	1.0.0	4.000
1	Coloring book moana	ART_AND_DESIGN	3.9	967.0	14.0	5.000	Free	0	Everyone	Art & Design	215.0	2.0.0	4.000
2	U Launcher Lite – FREE Live Cool Themes, Hide ...	ART_AND_DESIGN	4.7	87510.0	8.7	50.000	Free	0	Everyone	Art & Design	17.0	1.2.4	4.000
3	Sketch - Draw & Paint	ART_AND_DESIGN	4.5	215644.0	25.0	500.000	Free	0	Teen	Art & Design	71.0	Varies with device	4.200
4	Pixel Draw - Number Art Coloring Book	ART_AND_DESIGN	4.3	967.0	2.8	1.000	Free	0	Everyone	Art & Design	59.0	1.1	4.400
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9355	FR Calculator	FAMILY	4.0	7.0	2.6	0.005	Free	0	Everyone	Education	426.0	1.0.0	4.100
9356	Sya9a Maroc - FR	FAMILY	4.5	38.0	53.0	0.050	Free	0	Everyone	Education	389.0	1.48	4.100
9357	Fr. Mike Schmitz Audio Teachings	FAMILY	5.0	4.0	3.6	0.001	Free	0	Everyone	Education	43.0	1.0	4.100
9358	The SCP Foundation DB fr nn5n	BOOKS_AND_REFERENCE	4.5	114.0	3.6	0.010	Free	0	Mature 17+	Books & Reference	1307.0	Varies with device	3.867
9359	iHoroscope - 2018 Daily Horoscope & Astrology	LIFESTYLE	4.5	398307.0	19.0	100.000	Free	0	Everyone	Lifestyle	24.0	Varies with device	3.867

9360 rows × 13 columns

Se puede observar los tipos de datos, resumidos a object(string) y float64

```
[32] df.dtypes

App                object
Category           object
Rating            float64
Reviews           float64
Size              float64
Installs          float64
Type               object
Price              object
Content Rating     object
Genres             object
Days passed       float64
Current Ver        object
Android Ver       float64
dtype: object
```

Se realizó una función que detecta outliers en las columnas del dataset (una observación anormal y extrema en una muestra estadística). Se calcula el valor iqr (Los Outliers son los puntos que caen a más de 1.5 veces el IQR, a partir de la caja) y se hallan límites superiores e inferiores, con los cuales se creará una estructura con los outliers y sus valores:

```
[33] def detect_outlier(df,cols):
       outlier_dict = {}
       for col in cols:
         print(col)
         q1 = np.quantile(df[col],0.25)
         q3 = np.quantile(df[col],0.75)
         iqr = q3 - q1
         upper = q3+(1.5*iqr)
         lower = q1-(1.5*iqr)
         outlier = df.loc[ ((df[col]>=upper) | (df[col]<=lower) )][[col]]
         outlier_dict[col] = [q1,q3,iqr,lower,upper,outlier.values]
       return outlier_dict
```

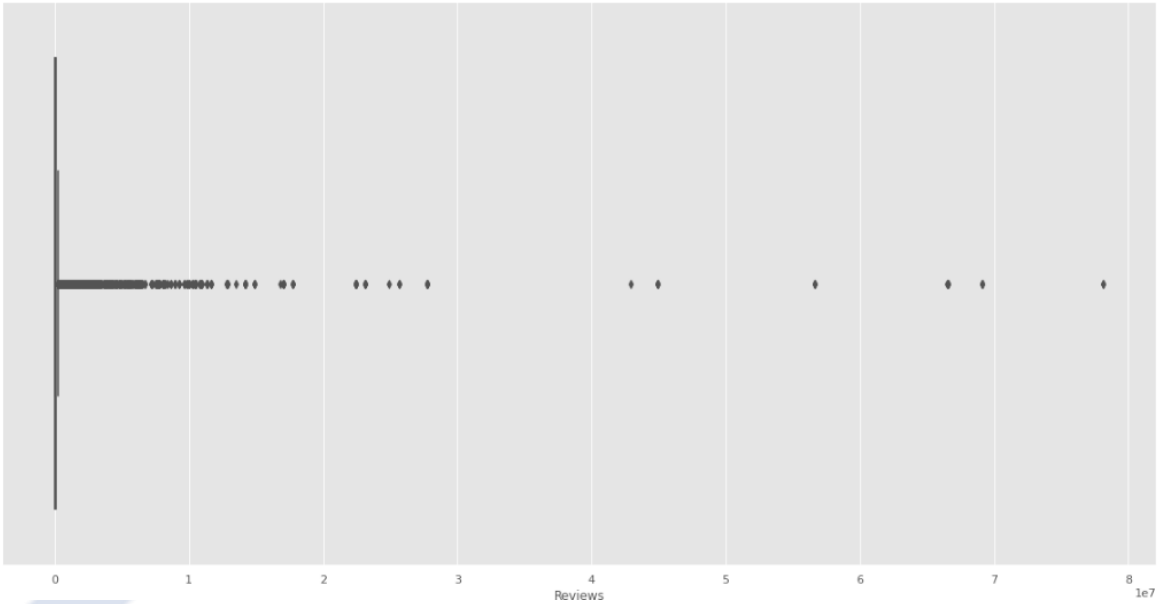
Función para remover outliers

```
def showData(column,df_outlier):
  lower = df_outlier[column]["lower"]
  upper = df_outlier[column]["upper"]
  q1 = df_outlier[column]["q1"]
  q3 = df_outlier[column]["q3"]
  iqr =  df_outlier[column]["inter_q"]
  print("q1=",q1,"\nq3=",q3,"\niqr=",iqr,"\nupper=",upper,"\nlower=",lower)
  return (lower,upper)

def remove_outliers(df,column,df_outlier,method="capping"):
  lower,upper = showData(column,df_outlier)
  new_df = df.copy()
  outliers = new_df.loc[(df[column] >= upper) | (df[column] <= lower)]
  print("METHOD= ",method)
  if(method.lower() == "capping"):
    #show boxplot before capping
    # sns.boxplot(new_df[column])
    #do capping
    new_df.loc[(new_df[column] > upper), column] = df[column].mean()#upper#PONER LA MEDIA/MEDIANA
    new_df.loc[(new_df[column] < lower), column] = df[column].mean()#lower#PONER LA MEDIA/MEDIANA
    #show after capping
    # sns.boxplot(new_df[column])
  elif(method.lower() == "remove"):
    #show boxplot before has removed
    sns.boxplot(new_df[column])
    #Removiing...
    new_df = new_df.loc[(df[column] < upper) & (df[column] > lower)]
    #showing after removing outliers
    sns.boxplot(new_df[column])

  print("#rows of Original DF: ",len(df))
  print("#rows of New DF: ",len(new_df))
  print("#Outliers: ", len(outliers), len(df)-len(new_df))
  return (new_df,outliers)
```

Al procesar la información con la función detect_outliers, se obtuvieron los siguientes valores

	Reviews	Size	Installs	Days passed	Android Ver
q1	186.75	5.5	0.1	25.0	3.867
q3	81627.5	33.0	50.0	313.0	4.1
inter_q	81440.75	27.5	49.9	288.0	0.233
lower	-121974.375	-35.75	-74.75	-407.0	3.5175
upper	203788.625	74.25	124.85	745.0	4.4495
outlier	[[215644.0], [224399.0], [295221.0], [271920.0...	[[77.0], [77.0], [84.0], [97.0], [76.0], [76.0...	[[500.0], [1000.0], [1000.0], [10000.0], [500....	[[1453.0], [1128.0], [1402.0], [857.0], [1072....	[[2.3], [3.0], [2.3], [2.3], [3.0], [2.3], [2....

En base a esto, se corrigen los outliers para cada una de las variables de la tabla anterior
Para reviews se grafica primero para ver la magnitud del problema

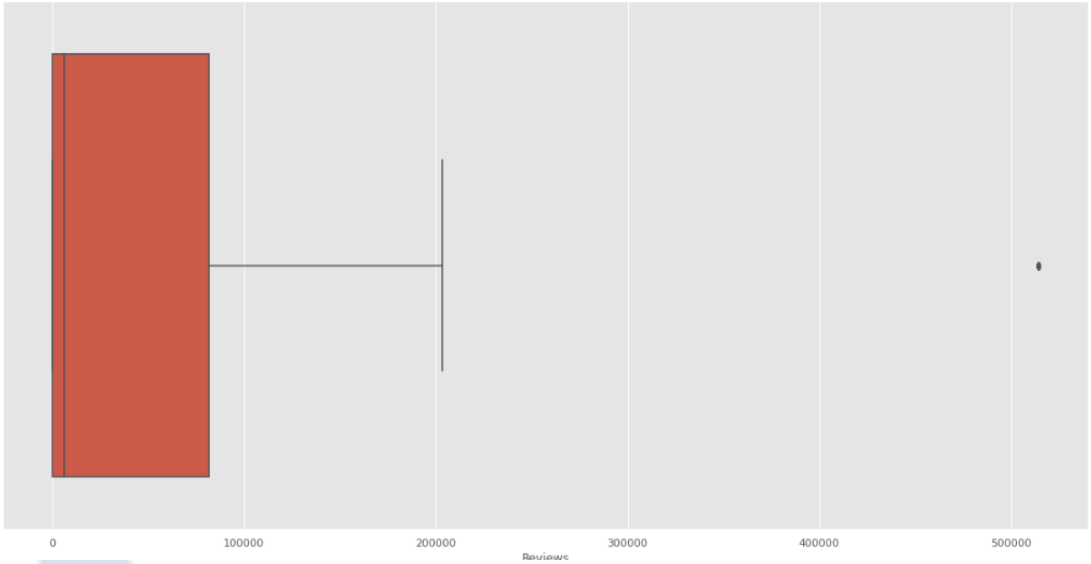


Se realiza un llamado a la función *remove_outliers*() para remover los outliers. Se ha de aclarar que se usará el método capping para tal fin:

```
df_test, outliers = remove_outliers(df_test,"Reviews",df_outlier,method="Capping")
```

```
q1= 186.75
q3= 81627.5
iqr= 81440.75
upper= 203788.625
lower= -121974.375
METHOD=  Capping
#rows of Original DF:  9360
#rows of New DF:  9360
#Outliers:  1634 0
```

Se visualiza el gráfico de cajas y bigotes:



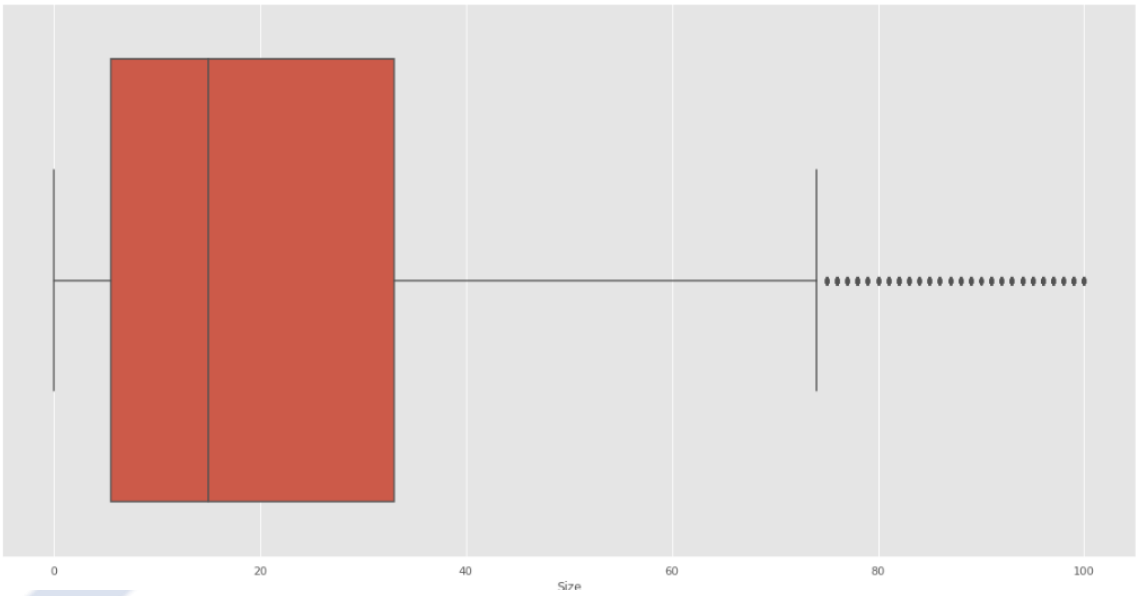
Se visualizan los resultados:

```
df_test.loc[df_test["Reviews"]>500000,"Reviews"] = df_test["Reviews"].mean()
```

```
df_test["Reviews"].describe()
```

```
count      9360.000000
mean      36645.841596
std       49036.140636
min           1.000000
25%         186.750000
50%        5955.000000
75%       81627.500000
max      203130.000000
Name: Reviews, dtype: float64
```

Para sizes tenemos lo siguiente:
Se visualiza el gráfico de cajas y bigotes donde se puede ver los puntos separados de la caja del diagrama:



Un análisis de los datos arroja lo siguiente:

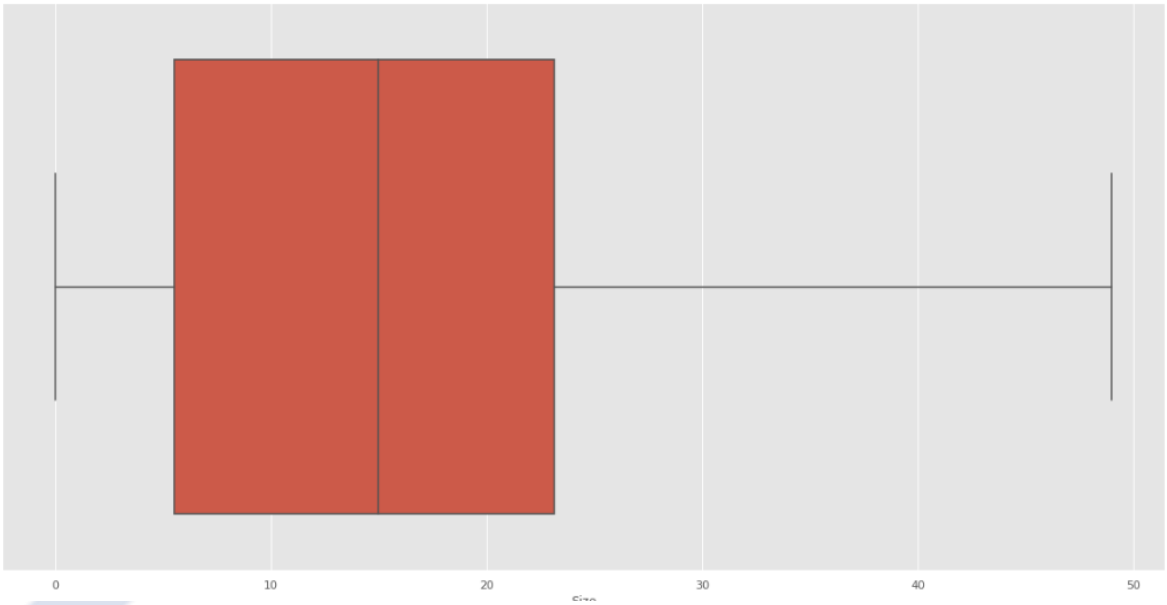
	Size	Android Ver
q1	5.5	3.867
q3	33.0	4.1
inter_q	27.5	0.233
lower	-35.75	3.5175
upper	74.25	4.4495
outlier	[[77.0], [77.0], [84.0], [97.0], [76.0], [76.0...	[[2.3], [3.0], [2.3], [2.3], [3.0], [2.3], [2....

Se hace el llamado para corregir ese problema usando capping:

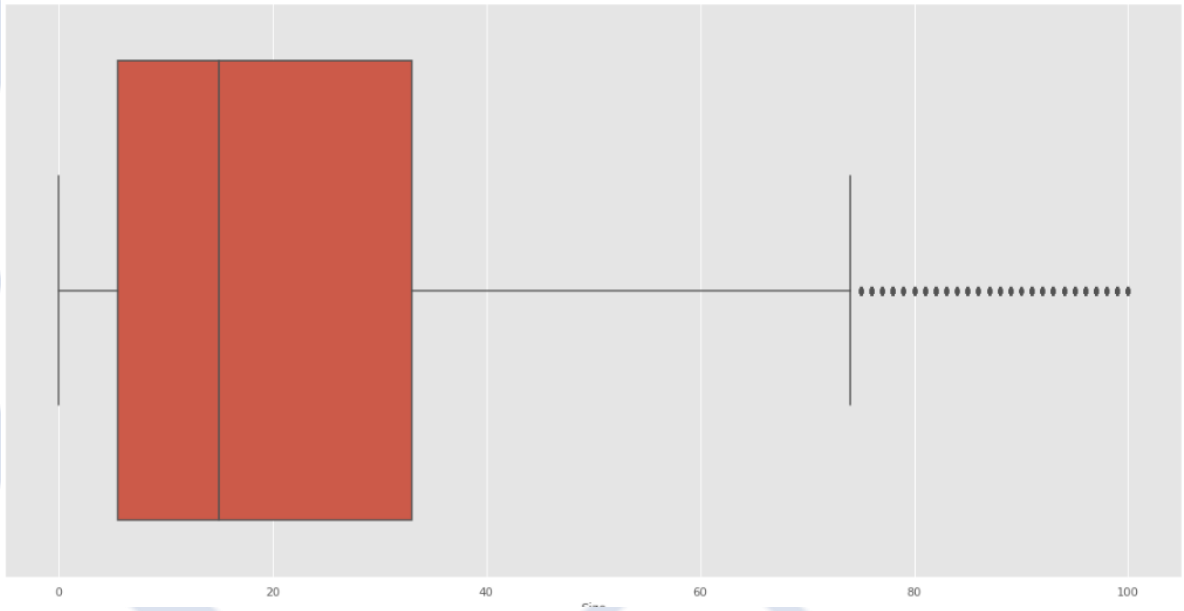
```
df_test, outliers = remove_outliers(df_test,"Size",df_outlier,method="Capping")
```

```
q1= 5.5
q3= 33.0
iqr= 27.5
upper= 74.25
lower= -35.75
METHOD=  Capping
#rows of Original DF:  9360
#rows of New DF:  9360
#Outliers:  490 0
```

Se resolvió el problema de outliers:



Para Installs tenemos que tiene puntos distanciados considerablemente de la caja:

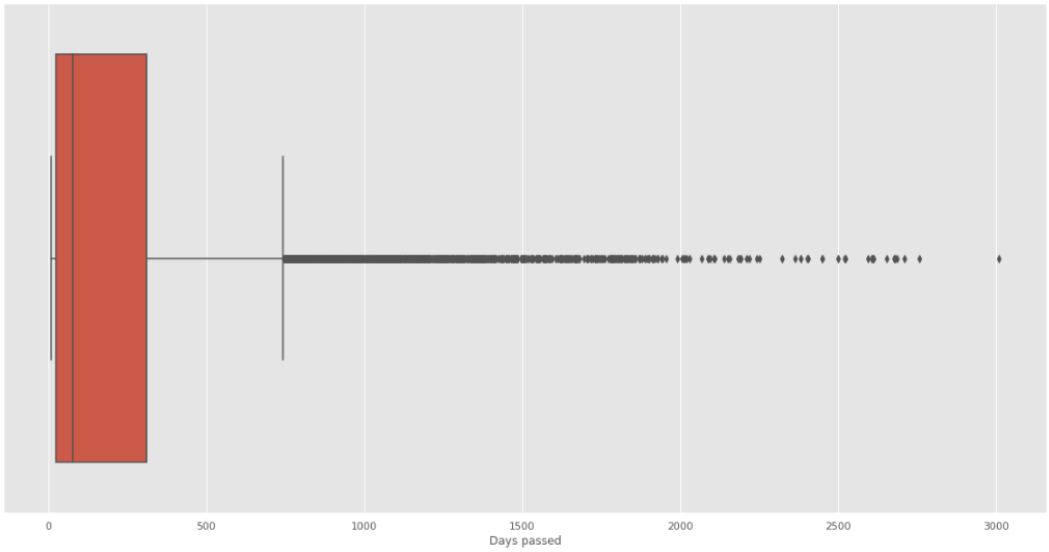


Análisis de límites de los datos y visualización de los outliers

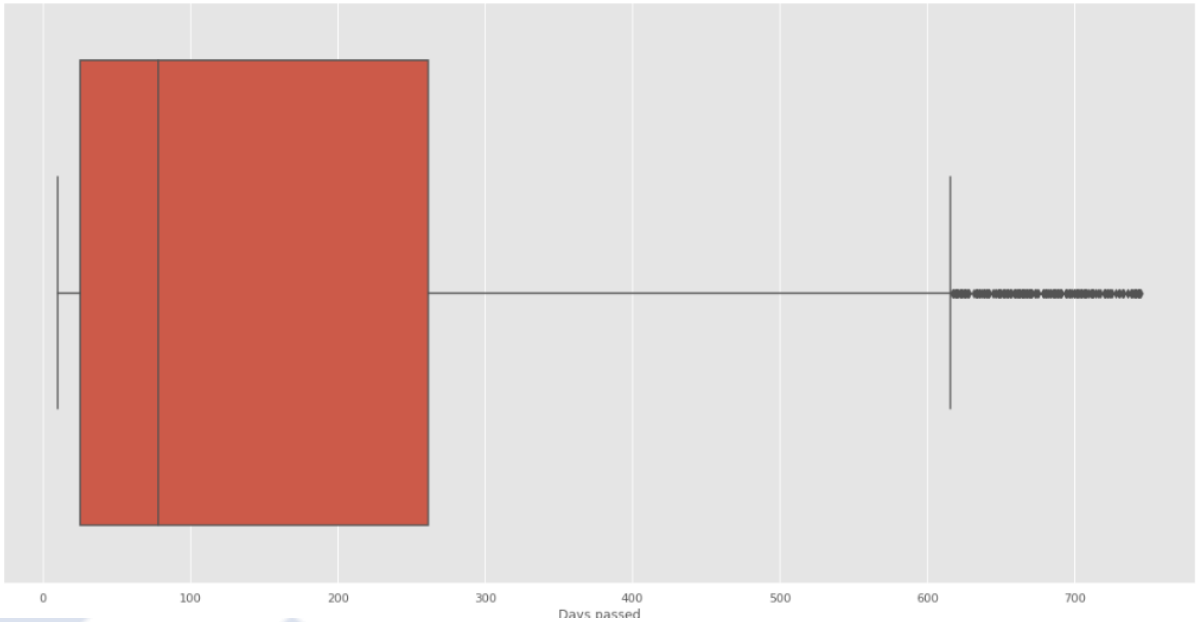
	Installs
q1	0.1
q3	50.0
inter_q	49.9
lower	-74.75
upper	124.85
outlier	[[500.0], [1000.0], [1000.0], [10000.0], [500....

Luego de la corrección se puede ver el siguiente gráfico en función a los datos corregidos :

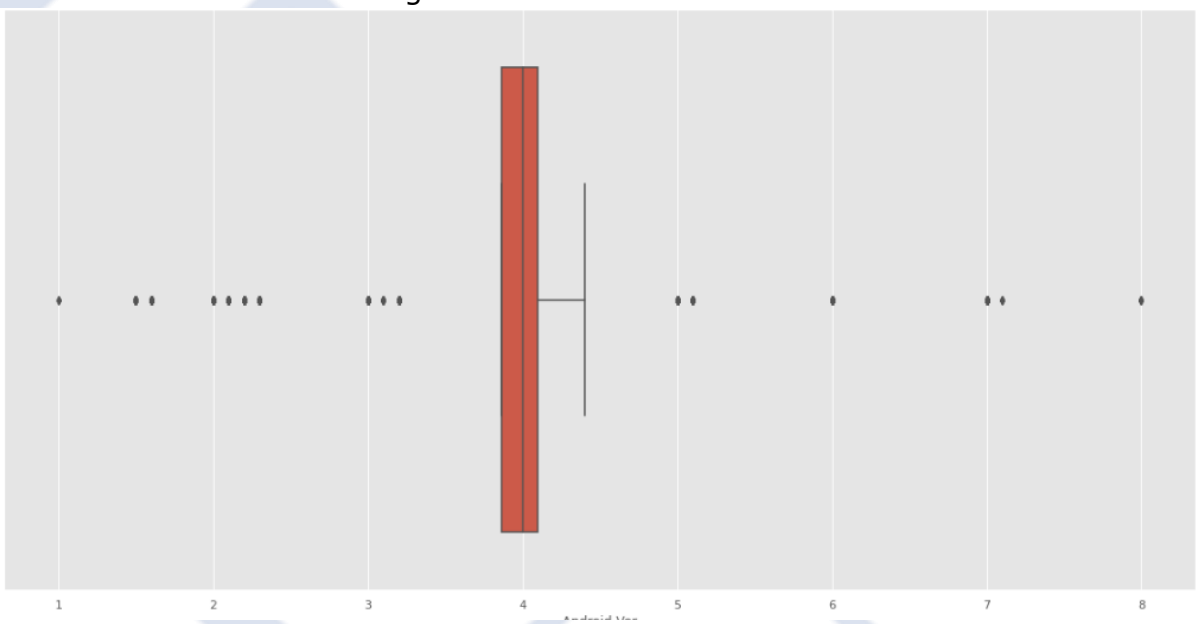
Para Days Passed tenemos lo siguiente:



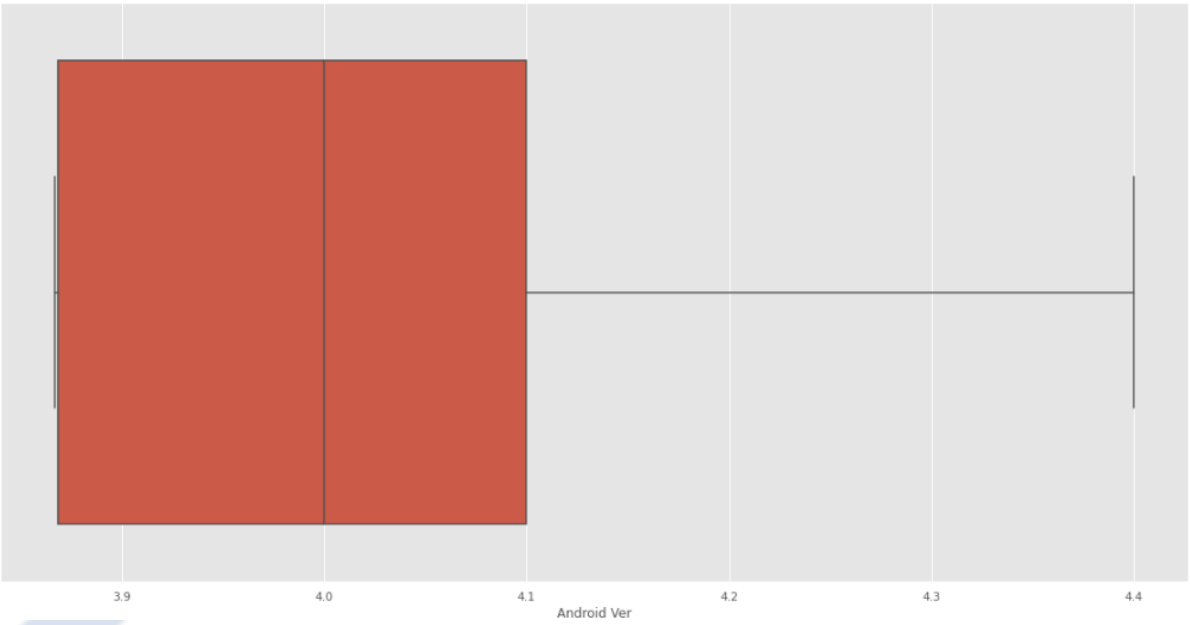
Se visualiza la corrección de los datos:



Para Android Version tenemos lo siguiente:



Finalmente se ve la corrección de los datos:



Redes Neuronales

- Se seguirán unos ciertos pasos

PASOS

- CODIFICAR Y/O CATEGORIZAR
- CONTRUIR EL MODELO
- COMPILAR
- ENTRENAR
- PREDECIR

- Ahora se removerán las columnas que no se utilizaran

REMOVE COLUMNS

```
[ ]  dataset.drop(columns=["Type","Price","Genres","Current Ver","App"],inplace=True)
```

- Luego se ordenara la tabla

	Category	Reviews	Size	Installs	Content Rating	Days passed	Android Ver	Rating

- Observamos la columna “Category” esta columna.

```
df.groupby("Category").nunique()
```

AUTO_AND_VEHICLES	18	69	50	11	3	49	7
BEAUTY	14	41	30	10	3	37	5
BOOKS_AND_REFERENCE	20	139	90	14	4	143	6
BUSINESS	31	191	112	15	2	177	7
COMICS	19	55	44	10	5	39	7
COMMUNICATION	25	188	120	14	3	161	7
DATING	27	138	68	12	3	64	7
EDUCATION	13	111	57	9	4	79	6
ENTERTAINMENT	16	80	41	8	4	53	7
EVENTS	14	40	34	11	3	42	6

- Se procede a codificar “Category” y los valores que poseía esta columna

	Reviews	Size	Installs	Content Rating	Days passed	Android Ver	Rating	ART_AND_DESIGN	AUTO_AND_VEHICLES	BEAUTY	...	SPORTS	TRAVEL_AND_LOCAL	TOOLS	PERSONALIZATION	PRODUCTIVITY	PARENTING	WEATHER	VIDEO_PLAYERS	NEWS_AND_MAGAZIN
0	159.000000	19.0	0.100000	Everyone	222.0	4.000	4.1	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
1	967.000000	14.0	5.000000	Everyone	214.0	4.000	3.9	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
2	87510.000000	8.7	50.000000	Everyone	16.0	4.000	4.7	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
3	107649.319055	25.0	35.367881	Teen	70.0	4.200	4.5	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
4	967.000000	2.8	1.000000	Everyone	58.0	4.400	4.3	1.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

- Se procede a codificar “Rating y los valores que poseía esta columna”

```
cols_content = list(dataset["Content Rating"].unique())
cols_content
```

```
['Everyone',
 'Teen',
 'Everyone 10+',
 'Mature 17+',
 'Adults only 18+',
 'Unrated']
```

```
data_test = dataset.copy()
```

```
col_droped = data_test.pop("Content Rating")
```

```
for col in cols_content:
  data_test[col] = (col_droped == col)*1.0
```

```
data_test
```

	Reviews	Size	Installs	Days passed	Android Ver	Rating	ART_AND_DESIGN	AUTO_AND_VEHICLES	BEAUTY	BOOKS_AND_REFERENCE	...	WEATHER	VIDEO_PLAYERS	NEWS_AND_MAGAZINES	MAPS_AND_NAVIGATION	Everyone	Teen	Everyone 10+	Mature 17+	A

- Se procede a tener dos variables para entrenar y los datos para el testing

```
train_dataset = dataset.sample(frac=0.8,random_state=0)
test_dataset = dataset.drop(train_dataset.index)
```

Ahora se procede a normalizar

- Primero se obtienen las estadísticas

```
train_stats = train_dataset.describe()
train_stats.pop("Rating")
train_stats = train_stats.transpose()
train_stats
```

	count	mean	std	min	25%	50%	75%	max
Reviews	7488.0	36646.543287	49051.651664	1.00000	190.750000	5998.5	80909.750000	203130.0
Size	7488.0	23.383751	23.385373	0.00850	5.600000	15.0	33.000000	100.0
Installs	7488.0	12.258876	15.685476	0.00001	0.100000	5.0	22.654223	50.0
Days passed	7488.0	259.702858	396.855706	9.00000	24.000000	76.0	311.000000	3010.0
Android Ver	7488.0	4.023885	0.163303	3.86700	3.868779	4.0	4.100000	4.4
ART_AND_DESIGN	7488.0	0.006944	0.083049	0.00000	0.000000	0.0	0.000000	1.0

- Se obtienen las etiquetas del train y test (Se quita Rating porque es nuestro output)

```
train_labels = train_dataset.pop('Rating')
#En este instante, se borro Rating en train_dataset
test_labels = test_dataset.pop('Rating')
#En este instante, se borro Rating en test_dataset
```

- Se normalizan los datos (Con los datos de traint_stats)

```
def norm(x):
  return (x - train_stats['mean']) / train_stats['std']

normed_train_data = norm(train_dataset)
normed_test_data = norm(test_dataset)
```

- Resultado de la tabla *normed_train_data*

```
normed_train_data
```

	Reviews	Size	Installs	Days passed	Android Ver	ART_AND_DESIGN	AUTO_AND_VEHICLES	BEAUTY	BOOKS_AND_REFERENCE	BUSINESS	...	PARENTING	WEATHER	VIDEO_PLAYERS	NEWS_AND_MAGAZINES	MAPS_AND_NAVIGATION	Everyone	Teen
1710	1.447510	0.539493	1.473274	-0.593926	1.690812	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	-1.938248	2.752509
5355	2.495664	2.634820	-0.144011	-0.178158	-0.146259	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	0.515861	-0.363256
4540	-0.740007	-0.935788	-0.775168	3.349573	-0.949799	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	0.515861	-0.363256
5071	-0.538015	0.667779	-0.144011	-0.026969	1.690812	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	0.515861	-0.363256
6394	-0.747020	-0.897302	-0.781537	0.298086	0.466098	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	0.515861	-0.363256
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5732	-0.746082	-0.880198	-0.775168	0.101541	0.466098	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	0.515861	-0.363256
8872	0.263670	-0.828884	0.662737	-0.467935	0.466098	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	0.515861	-0.363256
505	-0.729732	0.154637	-0.717790	-0.609045	0.466098	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	-1.938248	-0.363256
7388	-0.746938	-0.713427	-0.780906	1.419400	0.466098	-0.083619	-0.089111	-0.059024	-0.136	-0.182348	...	-0.070464	-0.090621	-0.131342	-0.158708	-0.116922	0.515861	-0.363256

- Se construye el modelo y ejecuta

```
def build_model():
  model = tf.keras.Sequential([
    tf.keras.layers.Dense(32, activation='relu', input_shape=[len(train_dataset.keys())]),
    tf.keras.layers.Dense(64, activation='relu'),
    tf.keras.layers.Dense(128, activation='relu'),
    tf.keras.layers.Dense(64, activation='relu'),
    tf.keras.layers.Dense(32, activation='relu'),
    tf.keras.layers.Dense(1)
  ])
  #OPTIMIZERS
  optimizer = tf.keras.optimizers.RMSprop(0.001)
  #COMPILE
  model.compile(loss='mse',
                optimizer=optimizer,
                metrics=['mae', 'mse'])
  return model
```

- Se entrena al modelo

```
history = model.fit(
  normed_train_data, train_labels,
  epochs=EPOCHS, validation_split = 0.2, verbose=0,
  callbacks=[PrintDot()])
```

- Se realiza un ejemplo de predicción

```
example_batch = normed_train_data[:10]
example_result = model.predict(example_batch)
example_result
```

```
array([[4.396088 ],
       [3.8018947],
       [4.0594616],
       [4.571604 ],
       [3.97579  ],
       [3.6644804],
       [2.948675 ],
       [4.6022177],
       [4.6018677],
       [4.2051067]], dtype=float32)
```

- Se imprime la historia de los 1000 epoch iterados

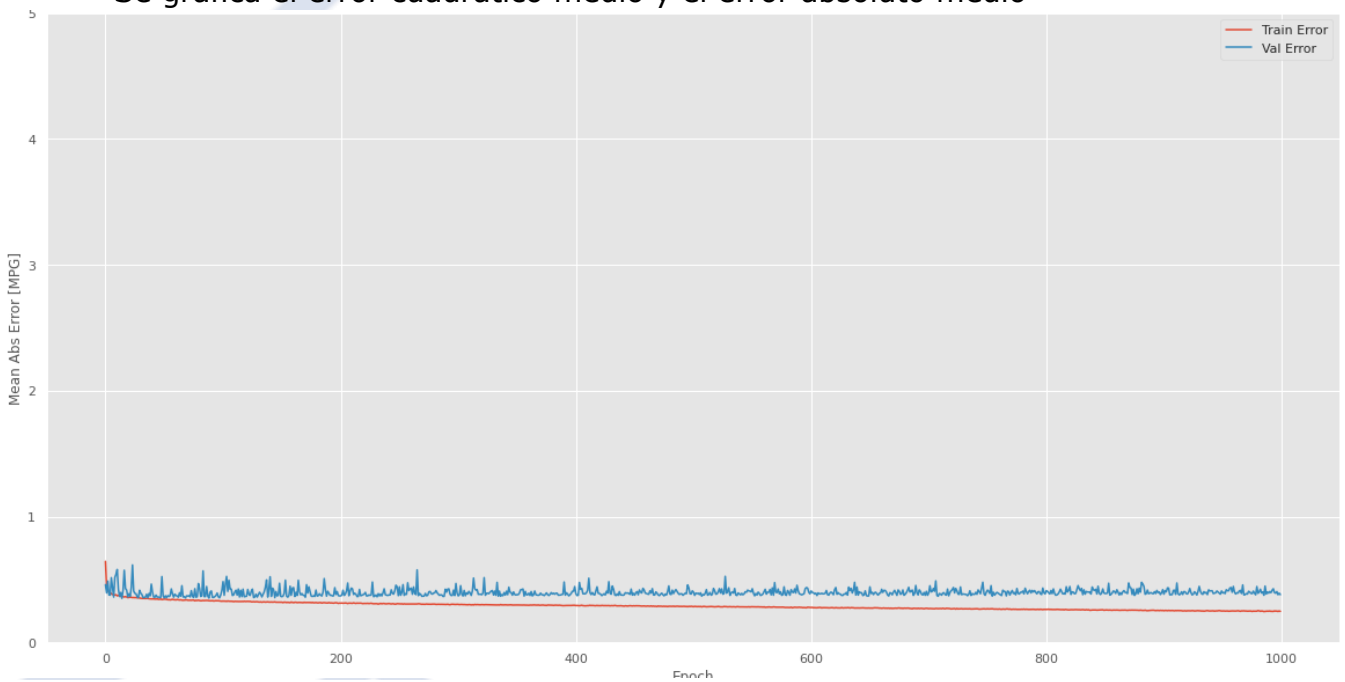
```
hist = pd.DataFrame(history.history)
hist['epoch'] = history.epoch
hist.tail()
```

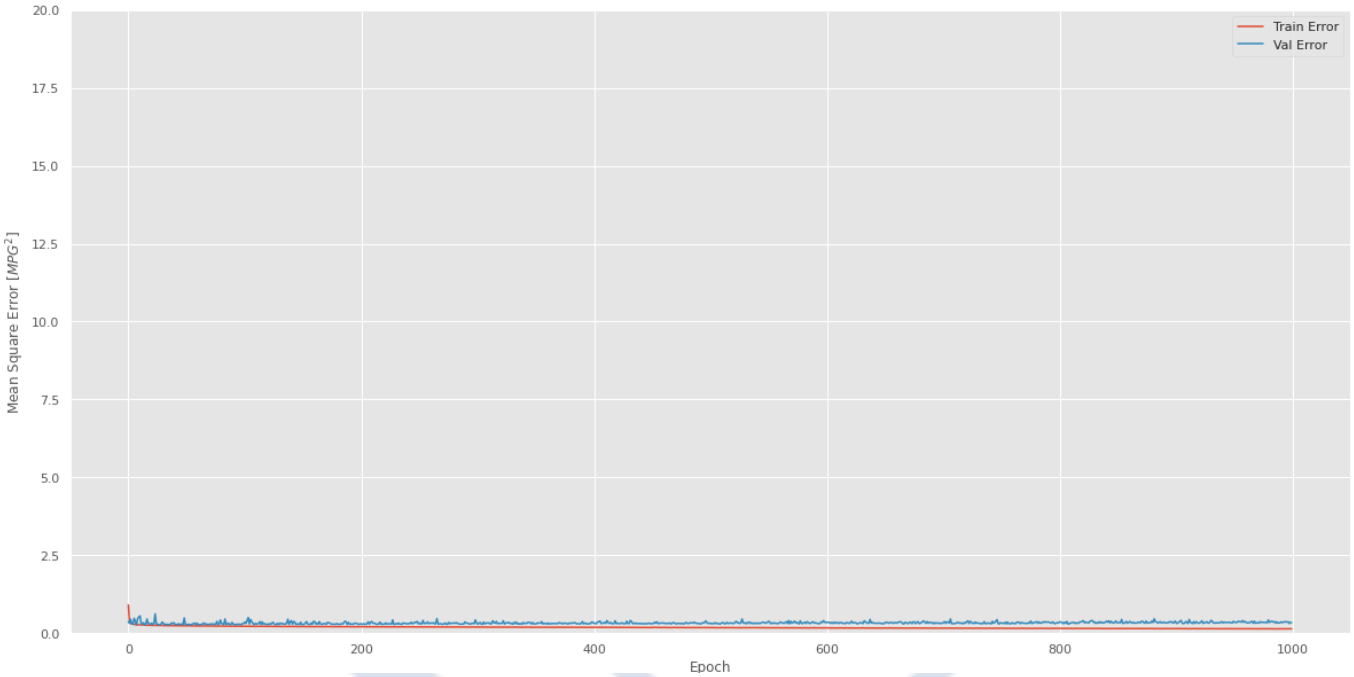
	loss	mae	mse	val_loss	val_mae	val_mse	epoch
995	0.141928	0.248724	0.141928	0.352853	0.391610	0.352853	995
996	0.143987	0.249435	0.143987	0.370625	0.404294	0.370625	996
997	0.142637	0.246242	0.142637	0.311485	0.379217	0.311485	997
998	0.143063	0.250076	0.143063	0.347423	0.391196	0.347423	998
999	0.138982	0.247082	0.138982	0.340982	0.379805	0.340982	999

- Se crea una función para ver el error de entrenamiento
-

```
def plot_history(history):
  hist = pd.DataFrame(history.history)
  hist['epoch'] = history.epoch

  plt.figure()
  plt.xlabel('Epoch')
  plt.ylabel('Mean Abs Error [Rating]')
  plt.plot(hist['epoch'], hist['mae'],
           label='Train Error')
  plt.plot(hist['epoch'], hist['val_mae'],
           label = 'Val Error')
  plt.ylim([0,5])
  plt.legend()

  plt.figure()
  plt.xlabel('Epoch')
  plt.ylabel('Mean Square Error [$Rating^2$]')
  plt.plot(hist['epoch'], hist['mse'],
           label='Train Error')
  plt.plot(hist['epoch'], hist['val_mse'],
           label = 'Val Error')
  plt.ylim([0,20])
  plt.legend()
  plt.show()


plot_history(history)
```

- Se grafica el error cuadrático medio y el error absoluto medio





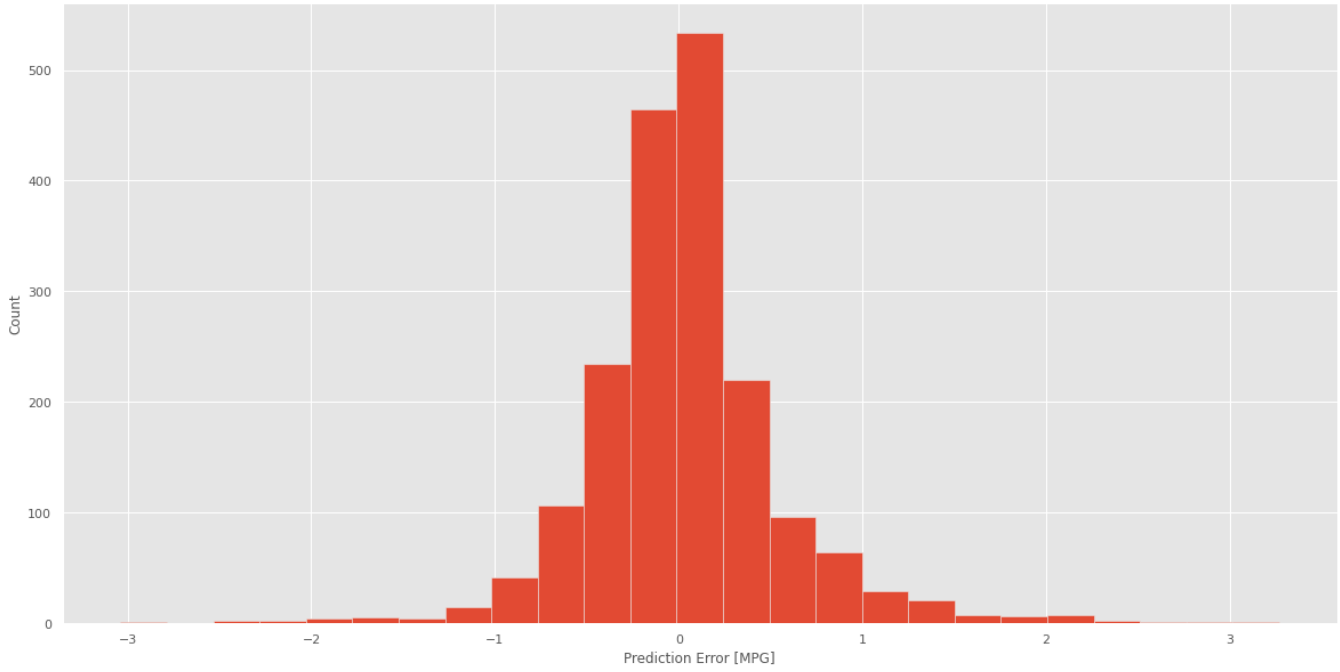
- Se analizan las predicciones

```
loss, mae, mse = model.evaluate(normed_test_data, test_labels, verbose=2)

print("Testing set Mean Abs Error: {:5.2f} MPG".format(mae))
```

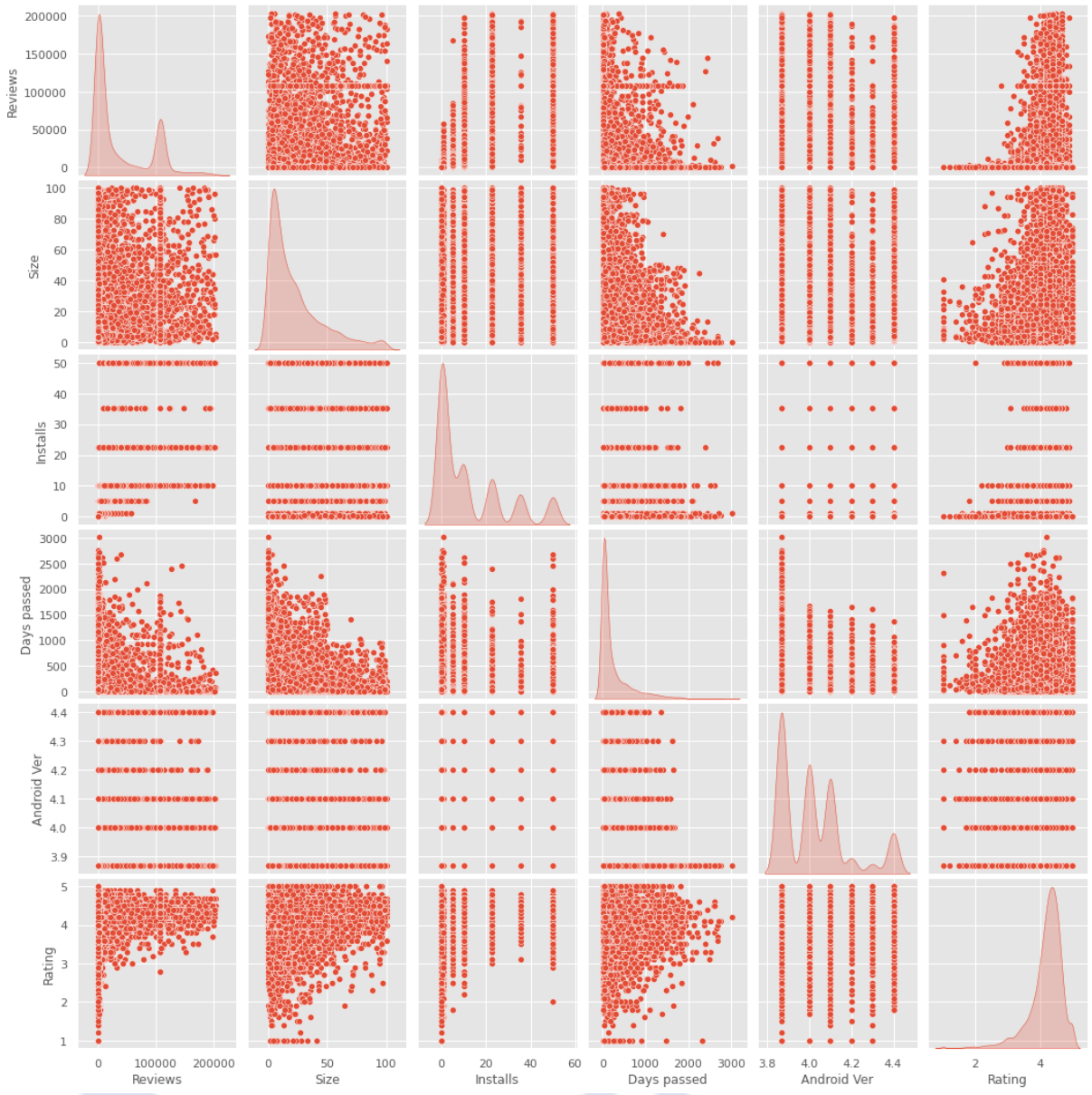
```
59/59 - 0s - loss: 0.2757 - mae: 0.3544 - mse: 0.2757 - 87ms/epoch - 1ms/step
Testing set Mean Abs Error:  0.35 MPG
```

- Gráfico en el error medio absoluto

```
error = test_predictions - test_labels
plt.hist(error, bins = 25)
plt.xlabel("Prediction Error [MPG]")
_ = plt.ylabel("Count")
```



- Se usa *pairplot* que genera una gráfica de pares de variables. nos permite ver tanto la distribución de variables individuales como las relaciones entre dos variables



Evaluación

Para la evaluación de nuestros datos, se usaron dos variantes de un mismo modelo, en el cual el modelo versión 1, se implementó sin un modelo de estrategia determinado para corregir a la red neuronal con respecto al error cuadrático medio.

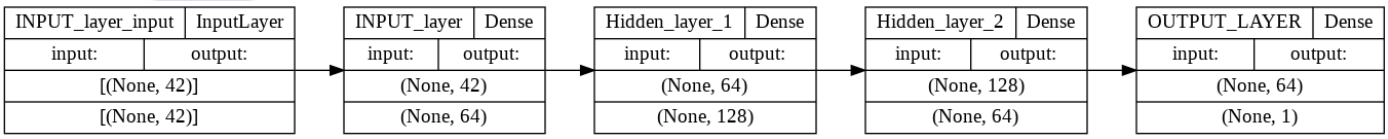
Mientras que en el modelo versión 2, si se usó una estrategia de corrección llamada "EarlyStopping", este tipo de técnica logra corregir a la red neuronal según el error absoluto medio, el cual deberá bajar en cada iteración o época de entrenamiento, pero si el valor aumenta

con respecto al anterior entonces se debe orientar a la red neuronal para que consiga un Error Absoluto Medio(considerado en el parámetro *"monitor"*) menor en cada iteración, para esto se debe especificar el parámetro "*patience*", el cual indicará hasta cuantas épocas es permitido el valor de incremento y cada cuanto corregir este valor.

Modelo Genérico:

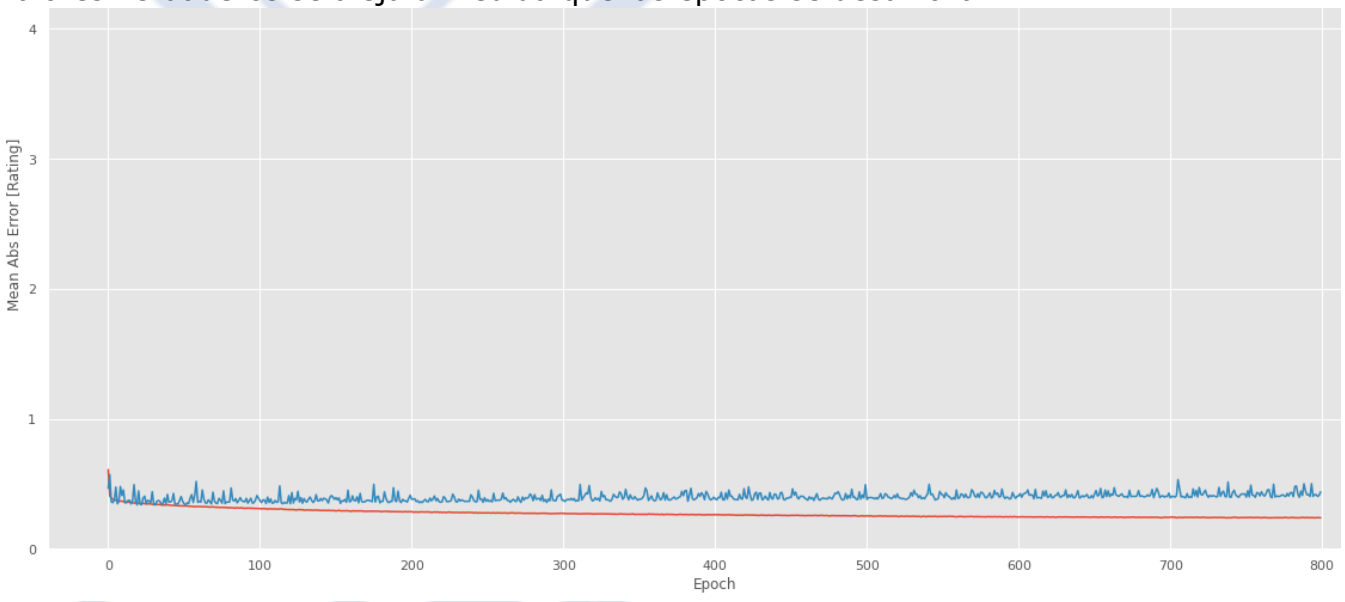
El modelo genérico para las dos versiones de siguientes está determinado por esta arquitectura:

1. La primera capa, está constituida por el Input Layer, donde entran los datos de las variables de cada columna.
2. La segunda, tercera y cuarta columna son las denominadas "Hidden Layers", las cuales iniciarán con pesos aleatorios en sus conexiones y a medida se entrene el modelo, se modificarán las neuronas conforme a los patrones de los datos.
3. Y finalmente, la última capa es llamada "Output Layer", donde se entregará el valor final de la variable Rating predecida.



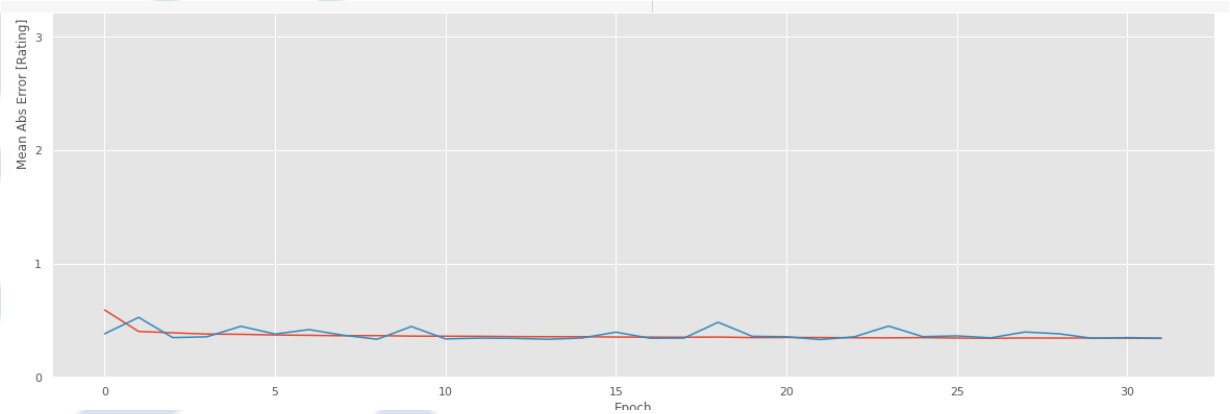
Modelo 1:

A continuación, se muestra el Mean Absolute Error con respecto a cada época, se puede ver claramente que la diferencia entre el Mean Absolute Error entre los valores predecidos y los valores verdaderos se aleja a medida que las épocas se desarrollan.



Modelo versión 2:

En este modelo, se puede ver a simple vista que los valores convergen y se acercan más y más, debido a la estrategia "EarlyStopping" que corregirá a los valores predecidos conforme al lote de testeo modificando así a las redes neuronales cada vez que las diferencias de Mean Absolute Error son mayores con una corrección de cada 10 épocas.



Métricas del Modelo 1 y Modelo 2:

Luego fueron desarrolladas las métricas con respecto a los dos modelos, como se presenta en la siguiente tabla de comparaciones:

Modelo/Métricas	Mean Absolute Error	Mean Square Error
Modelo versión 1	0.4582	0.4286
Modelo versión 2	0.3581	0.2870

Implementación

En base al análisis realizado se realizó la implementación de la red neuronal usando la herramienta Collaboratory de Google Research, el cual nos permite ejecutar código desde notebooks(páginas donde el código puede ser dividido , con el fin de poder ser ejecutados en pasos) . Se creó un directorio ubicado en el Google Drive personal de cada desarrollador (/MyDrive/INTELIGENCIA_ARTIFICIAL) donde se colocó el archivo original que contiene al

dataset, este directorio también sirvió para poder recibir la respuesta de la aplicación, la cual contenía el dataset modificado en su fase de limpieza.

En el momento de la ejecución el servicio de google nos suministra todas las librerías, el almacenamiento, el poder de procesamiento y la RAM, por lo cual simplemente se ejecutan las instrucciones y se obtienen los resultados a través del navegador Web, sin consumir muchos recursos de nuestros sistemas.

Conclusiones

Las conclusiones al respecto de los modelos construidos es que tienen una grán relación entre las variables usadas para la predicción, así el modelo cumple en general con el objetivo de este trabajo, y puede ser una herramienta para la predicción de rating de una App, dada sus variables de input.
De tal manera el modelo hace posible analizar la relación del "Rating" con respecto a otras variables, y ayuda a predecir cómo se comporta la predicción frente a ciertas variables cuando las personas necesiten analizar cómo se comportaría su app en diversas situaciones estimadas o supuestas.

Como trabajo futuro, nos proponemos mejorar y probar el dataset con varios optimizadores, diferentes modelos, y probar con varias "Hidden Layers". Con el fin de mejorar las predicciones.

Referencias

[1] NewsEuropeanParliament, "What is artificial intelligence and how is it used? | News | European Parliament," News European Parliament, 2021.://www.europarl.europa.eu/news/en/headlines/society/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used (accessed Jun. 29, 2022).

[2] R. P. Diez, Introducción a la inteligencia artificial: Sistemas expertos, redes neuronales artificiales y computación evolutiva. Oviedo: Universidad de Oviedo, Servicio de Publicaciones, 2001.

[3] S. E. T. Sánchez, M. O. Rodríguez, A. E. Jiménez, and H. J. P. Soberanes, "Implementación de Algoritmos de Inteligencia Artificial para el Entrenamiento de Redes Neuronales de Segunda Generación," Jóvenes En La Cienc., vol. 2, no. 1, pp. 6–10, 2016, Accessed: Jun. 29, 2022. [Online]. Available: ://www.jovenesenlaciencia.ugto.mx/index.php/jovenesenlaciencia/article/view/715.

[4] M. Gallego Gallego and J. Hernández Cáceres, "Identificación de factores que permitan potencializar el éxito de proyectos de desarrollo de software," Sci. Tech., vol. 20, no. 1, p. 70, Mar. 2015, doi: 10.22517/23447214.9241. (accessed Jun. 29, 2022).

[5] GCFGLOBAL, "¿Cómo usar Android?: Qué es y cómo usar Google Play Store," 2019. https://edu.gcfglobal.org/es/como-usar-android/que-es-y-como-usar-google-play-store/1/ (accessed Jun. 29, 2022).

[6] E. Noei and K. Lyons, "A Survey of Utilizing User-Reviews Posted on Google Play Store," 2019, doi: 10.1145/1122445.1122456.

[7] "(PDF) Capítulo 1: Generalidades de las redes neuronales artificiales." https://www.researchgate.net/publication/327703478_Capitulo_1_Generalidades_de_las_redes_neuronales_artificiales (accessed Jun. 29, 2022).

[8] E. Varela and E. Campbells, "Redes Neuronales Artificiales: Una Revisión del Estado del Arte, Aplicaciones y Tendencias Futuras," Investig. y Desarro. en TIC, vol. 2, no. 1, pp. 18–27, 2011, Accessed: Jun. 22, 2022. [Online]. Available: http://revistas.unisimon.edu.co/index.php/identic/article/view/2455.

[9] F. O.C, *Introducción a los Negocios en un Mundo Cambiante*, 4ª ed. México D.F.. México: McGraw-Hill Interamericana, 2004.

[10]"López Porrero - 2011 - Limpieza de datos reemplazo de valores ausentes y estandarización," Accessed: Aug. 19, 2022. [Online]. Available: https://1library.co/title/limpieza-datos-reemplazo-valores-ausentes-estandarizacion.

[11]"¿Qué son los Datasets? [4 sitios donde encontrarlos]". KeepCoding Tech School. https://keepcoding.io/blog/que-son-datasets/ (accedido el 19 de julio de 2022).

[12]"Google colaboratory". Google Colab. https://colab.research.google.com/?hl=es (accedido el 25 de julio de 2022).

[13]"Mapa autoorganizado". Los diccionarios y las enciclopedias sobre el Académico. https://es-academic.com/dic.nsf/eswiki/683112 (accedido el 5 de agosto de 2022).

[14] J. M. Uriarte, "Google Drive: qué es, cómo funciona y características," *2020*, 2020. https://www.caracteristicas.co/google-drive/#ixzz7c8jGC900 (accessed Aug. 19, 2022).

[15] Santander Universidades, "¿Qué es Python? | Blog Becas Santander," 2021. https://www.becas-santander.com/es/blog/python-que-es.html (accessed Aug. 19, 2022).

[16] "Todo lo que necesitas saber sobre TensorFlow, la plataforma para Inteligencia Artificial de Google – Puentes Digitales," 2021. https://puentesdigitales.com/2018/02/14/todo-lo-que-necesitas-saber-sobre-tensorflow-la-plataforma-para-inteligencia-artificial-de-google/ (accessed Aug. 19, 2022).

[17] Ionos, "¿Qué es Keras? Introducción a la biblioteca de redes neuronales - IONOS," 2020. https://www.ionos.mx/digitalguide/online-marketing/marketing-para-motores-de-busqueda/que-es-keras/ (accessed Aug. 19, 2022).

[18] F. Pérez and H. Fernández, "Las redes neuronales y la evaluación del Riesgo de Crédito," Rev. Ing. Univ. Medellín, pp. 77–91, 2007, Accessed: Aug. 19, 2022. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1692-33242007000100007

Clasificación de texto con NLP en tweets relacionados con desastres naturales

NLP text classification in tweets related to natural disasters

Patrik Renee Quenta Nina
Universidad Nacional de San Agustín.
Arequipa, Perú.
pquenta@unsa.edu.pe
https://orcid.org/0000-0002-6184-1378

Frank Berly Quispe Cahuana
Universidad Nacional de San Agustín.
Arequipa, Perú.
quispecah@unsa.edu.pe
https://orcid.org/0000-0001-5584-1593

ARK: ark:/42411/s11/a88
PURL: 42411/s11/a88

RECIBIDO 25/02/2023 • ACEPTADO 23/03/2023 • PUBLICADO 30/03/2023

RESUMEN

Actualmente existe una gran cantidad de información circula a través de las redes sociales, esta no siempre tiende a ser verídica y tratándose de desastres naturales su falsedad podría llegar a tener bastantes consecuencias como histeria colectiva en la población. Para evitar esto se propuso un análisis eficiente para la comprobación de tweets con información falsa utilizando algoritmos de procesamiento de lenguaje natural.

Palabras claves: Desastres naturales,NLP , sentimientos, Twitter.

ABSTRACT

Currently there is a large amount of information circulating through social networks, this does not always tend to be true and in the case of natural disasters its falsity could have quite consequences such as mass hysteria in the population. To avoid this, an ef icient analysis was proposed to check tweets with false information using natural language processing algorithms.

Keywords: *Feelings, natural disasters, NLP, Twitter.*

INTRODUCCIÓN

Las plataformas de redes sociales como Facebook y Twitter se han convertido en herramientas de comunicación predominantes en la sociedad moderna. Estas plataformas proporcionan un

mecanismo para recopilar datos dinámicos sobre el comportamiento y el sentimiento humanos [1]. Estudios recientes consideran el uso de las redes sociales durante los desastres naturales, estudiando principalmente el estado de ánimo de la población o las diversas reacciones del público durante un incidente específico [1]. Sin embargo es posible captar datos incorrectos debido a que las personas generan percepciones erróneas sobre los peligros y que su conciencia situacional general se vea equivocada también cuando se exponen a información falsa o engañosa [1].

Debido al creciente uso de las redes sociales, la vulnerabilidad de las personas a los rumores falsos está determinada cada vez más por nuevas formas de verificación colectiva de los hechos y de dar sentido a los riesgos y desastres [2]. La existencia de diversas fuentes oficiales y no oficiales hacen que información errónea llegue a las personas durante una crisis. Lo cual puede hacer que juzgar la relevancia y la credibilidad de la información recibida sea una tarea difícil, por lo tanto, no pueden tomar las medidas de protección adecuadas [2].

Ya se han visto trabajos relacionados como AI-SocialDisaster, este es un sistema de apoyo a la toma de decisiones para identificar y analizar desastres naturales como terremotos, inundaciones e incendios forestales utilizando fuentes de redes sociales [3]. Con este software, los estrategas y planificadores de desastres pueden comprender las características de un desastre en un área en particular, además de obtener datos como análisis de sentimientos [3].

Debido a la disponibilidad omnipresente de datos en tiempo real, muchas agencias de rescate monitorean estos datos regularmente para identificar desastres, reducir riesgos y salvar vidas [4]. Sin embargo, es imposible que los humanos verifiquen manualmente la gran cantidad de datos e identifiquen desastres en tiempo real [4]. Con este propósito, se han propuesto muchas investigaciones para presentar palabras en representaciones comprensibles por máquina y aplicar métodos de aprendizaje automático en las representaciones de palabras para identificar el sentimiento de un texto [4].

Un claro ejemplo de la propagación de información falsa es “El caso del tifón Mangkhut en China”. En este caso se realizaron simulaciones de los escenarios reales, de aislamiento y de empotramiento. En esta simulación se probó la relevancia de la Primera Ley de Tobler en las redes sociales [5]. De tres escenarios, un escenario real, un escenario de aislamiento y un escenario de incrustación, se probó que la estrategia de incrustación controlaba mejor la transmisión de información falsa [5] de esta manera se plantearon sugerencias prácticas a la hora de filtrar información falsa en un desastre natural. Basándonos en estos casos podemos afirmar que enseñar a las computadoras cómo entender y hablar lenguajes naturales ofrece una gran cantidad de beneficios [6].

El subcampo de la IA que hace que las computadoras parezcan inteligentes para comprender y generar lenguajes como los humanos se llama procesamiento de lenguaje natural [6]. NLP se centra en la traducción de lenguaje natural, recuperación de información, extracción de

información, resumen de texto, respuesta a preguntas,modelado de temas, y el reciente sobre minería de opiniones [7]. Para resolver el problema de la información falsa generada a través de las redes sociales, nos apoyaremos en las incrustaciones y modelos pre-entrenados, debido a que estos proporcionan una visión de las estrategias fundamentales a la hora de recopilar información [8].

Teniendo en cuenta que Twitter se ha convertido en un importante canal de comunicación en tiempos de emergencia y, la ubicuidad de los teléfonos inteligentes que permite a las personas anunciar una emergencia que están observando en tiempo real. El objetivo del presente trabajo es desarrollar un modelo de deep learning capaz de clasificar si un tweet sobre un desastre natural es real o falso, haciendo uso de algoritmos de procesamiento de lenguaje natural (NLP).

Fundamentación Teórica

Redes neuronales artificiales

Una red neuronal artificial consiste en una red de unidades simples de procesamiento de información, llamadas neuronas. El poder de las redes neuronales para modelar relaciones complejas no es el resultado de modelos matemáticos complejos, sino que surge de las interacciones entre un gran conjunto de neuronas simples. Es normal pensar en las neuronas como un trabajo organizado en capas. [Deep learning]

Procesamiento de Lenguaje Natural (NLP)

El procesamiento del lenguaje natural (NLP) es una técnica de inteligencia artificial que se ocupa de la comprensión del lenguaje humano. El cual implica técnicas de programación para crear un modelo que pueda comprender el lenguaje, clasificar el contenido e incluso generar y crear nuevas composiciones de lenguaje humano. [ML for coders]

Clasificación de texto (Análisis de sentimientos)

El análisis de sentimientos, también conocido como minería de opiniones, se ocupa de inspeccionar estos sentimientos dirigidos hacia cualquier entidad. Liu (2010) utilizó el término objeto para representar la entidad objetivo mencionada en el texto. Un objeto está constituido por componentes y algún conjunto de atributos. Por ejemplo, se considera la siguiente oración "la pantalla de la computadora portátil está dañada y la duración de la batería es terrible". El objeto aquí es una computadora portátil que tiene pantalla y batería como componentes. La

calidad de visualización es un atributo de la pantalla y la duración de la batería es el atributo de la batería. Este texto se puede clasificar en una opinión positiva, negativa o neutral. []

Materiales

- **Google Drive**

Google Drive proporciona una solución de almacenamiento basada en la nube para archivos de Google Workspace y otros datos de usuario.[]

- **Google Colaboratory**

Permite escribir y ejecutar Python en el navegador, no se requiere configuración, acceso a las GPU sin cargo y es fácil de compartir

- **Python**
- **Matplotlib**
- **Numpy**
- **Pandas**
- **nltk (Natural Language Toolkit)**
- **Pytorch**
- **TensorFlor**

Metodología

Dataset

El dataset se obtuvo de kaggle el cual contiene los siguientes datos, cada muestra en los datos de entrenamiento y los datos de prueba tiene la siguiente información:

- El "text" de un tweet
- Una "keyword" de ese tweet (¡aunque esto puede estar en blanco!)

• La “location” desde la que se envió el tweet (también puede estar en blanco):

Análisis exploratorio de datos

Consiste en analizar e investigar conjuntos de datos y resumir sus principales características, a menudo empleando métodos de visualización de datos. el cual ayuda a determinar la mejor manera de manipular las fuentes de datos para obtener las respuestas que se necesitan.

Limpieza de datos (Data cleaning)

Es el proceso de detectar, corregir o eliminar registros corruptos o imprecisos de un conjunto de registros, donde pueden ser tablas o bases de datos con información incorrecta, incompleta, mal formateada o duplicada.

Modelado

Para la solución al problema se propuso usar una Red Neuronal Recurrente, debido a que su arquitectura lo permite para el manejo de datos secuenciales como lo es el texto.

Conclusiones

El análisis de los datos que nos brindan las redes sociales es importante debido a que con estos se podría evitar casos de histeria colectiva como se han mencionado anteriormente.
Para evitar la gran acumulación de datos innecesarios que toda red social tiene se destacó la parte de la limpieza de estos datos, de esta manera se podría hacer un mejor y más efectivo análisis con las metodologías de procesamiento del lenguaje natural.

Referencias

[1] S. K. Theja Bhavaraju, C. Beyney y C. Nicholson, "Quantitative analysis of social media sensitivity to natural disasters", International Journal of Disaster Risk Reduction, vol. 39, p. 101251, octubre de 2019. [En línea]. Disponible: https://doi.org/10.1016/j.ijdrr.2019.101251

[2] S. Hansson et al., "Communication-related vulnerability to disasters: A heuristic framework", International Journal of Disaster Risk Reduction, vol. 51, p. 101931, diciembre de 2020. [En línea]. Disponible: https://doi.org/10.1016/j.ijdrr.2020.101931

[3] F. K. Sufi, "AI-SocialDisaster: An AI-based software for identifying and analyzing natural disasters from social media", Software Impacts, p. 100319, mayo de 2022. [En línea]. Disponible: https://doi.org/10.1016/j.simpa.2022.100319

[4] S. Deb y A. K. Chanda, "Comparative analysis of contextual and context-free embeddings in disaster prediction from Twitter data", Machine Learning With Applications, vol. 7, p. 100253, marzo de 2022. [En línea]. Disponible: https://doi.org/10.1016/j.mlwa.2022.100253

[5] Y. Lian, Y. Liu y X. Dong, "Strategies for controlling false online information during natural disasters: The case of Typhoon Mangkhut in China", Technology in Society, vol. 62, p. 101265, agosto de 2020. [En línea]. Disponible: https://doi.org/10.1016/j.techsoc.2020.101265

[6] Raina, V., Krishnamurthy, S., "Natural Language Processing". In: Building an Effective Data Science Practice.Apress, Berkeley, CA, diciembre de 2021 Disponible. https://doi.org/10.1007/978-1-4842-7419-4_6

[7] K. R. Chowdhary, Fundamentals of Artificial Intelligence. New Delhi: Springer India, 2020.[En línea]. Disponible: https://doi.org/10.1007/978-81-322-3972-7

[8] J. K. Tripathy et al., "Comprehensive analysis of embeddings and pre-training in NLP", Computer Science Review, vol. 42, p. 100433, noviembre de 2021. [En línea]. Disponible: https://doi.org/10.1016/j.cosrev.2021.100433

[9] Kelleher, J. D. (2019). Deep Learning. MIT Press.

[10] Yadav, A. y Vishwakarma, D. K. (2019). Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), 4335–4385. https://doi.org/10.1007/s10462-019-09794-5

Predicción de la presión de burbujeo utilizando aprendizaje automático

Prediction of bubble pressure using machine learning

Oscar G. Gil M.
Universidad del Zulia (L.U.Z).
Maracaibo, Venezuela.
profesoroscargil@gmail.com
https://orcid.org/0000-0002-4108-2431

ARK: ark:/42411/s11/a82
PURL: 42411/s11/a82

RECIBIDO 10/03/2023 • ACEPTADO 29/03/2023 • PUBLICADO 30/03/2023

RESUMEN

En el presente estudio se utilizó la colección de algoritmos de aprendizaje automático del programa Weka para predecir la presión de burbujeo de 36 muestras de petróleo, determinando la precisión de sus resultados con el método de prueba validación cruzada de 10 pliegues. Posteriormente, para efectos de comparación, se calcularon las presiones de burbujeo con la correlación generada en el trabajo del cual se tomaron las muestras y sus resultados fueron más precisos que los obtenidos por los algoritmos en 4 de las 7 métricas de rendimiento utilizadas. En virtud de esta situación, y considerando que la correlación fue evaluada con los mismos datos con los que fue generada, se cambió el método de prueba a validación con los datos de entrenamiento y se volvieron a predecir las presiones de burbujeo. En igualdad de condiciones, el aprendizaje automático obtuvo mayor precisión que la correlación en todas las métricas de rendimiento.

Palabras claves: Algoritmos, Aprendizaje automático, Método de prueba, Presión de burbujeo, Weka.

ABSTRACT

In the present study, the collection of machine learning algorithms of the Weka program was used to predict the bubble pressure of 36 oil samples, determining the accuracy of their results with the 10-fold cross-validation test method. Subsequently, for comparison purposes, the bubble pressures were calculated with the correlation generated in the work from which the samples were taken and their results were more precise than those obtained by the algorithms in 4 of the

7 performance metrics used. Due to this situation, and considering that the correlation was evaluated with the same data with which it was generated, the test method was changed to validation with the training data and the bubble pressures were predicted again. Other things being equal, machine learning was more accurate than correlation on all performance metrics.

Keywords: *Algorithms, Machine learning, Test method, Bubble pressure, Weka.*

INTRODUCCIÓN

Las propiedades físicas de los fluidos son de gran importancia en los estudios de ingeniería de petróleo debido a que son necesarias para calcular los hidrocarburos inicialmente en sitio, para simular el comportamiento de los yacimientos y de los pozos, así como también para realizar el diseño de las facilidades de superficie.

De todas estas propiedades, la presión de burbujeo (Pb) es probablemente la más importante porque determina la existencia o no de una fase gaseosa que cambia las características del flujo en el yacimiento, en los pozos y en las facilidades de superficie. Adicionalmente, la presión de burbujeo aparece como una discontinuidad, ya que las tendencias con presión de otras propiedades como la relación gas petróleo en solución (Rs) cambian en ese punto [1].

Para determinar la presión de burbujeo y el resto de propiedades se realizan en un laboratorio un conjunto de pruebas comúnmente denominadas análisis PVT, ya que en ellas se analizan las relaciones entre presión, volumen y temperatura de una muestra de los fluidos del yacimiento. Sin embargo, cuando no se tiene esta información experimental (no se puede tomar una muestra que sea representativa, no están garantizados los costos asociados, etc.) se debe recurrir a correlaciones empíricas, ecuaciones de estado o modelos de aprendizaje automático [2].

El uso de modelos de aprendizaje automático en la industria petrolera se ha incrementado substancialmente a raíz de que la cantidad de datos que tienen que manejar, procesar y analizar es cada vez mayor. En este sentido, se deben destacar los esfuerzos dedicados desde finales de los años 1990 para utilizar el aprendizaje automático en la predicción de las propiedades obtenidas de los análisis PVT [3].

Elsharkawy [4] y Gharbi y col. [5, 6] estuvieron entre los primeros que utilizaron modelos de aprendizaje automático para predecir la presión de burbujeo de los fluidos de un yacimiento. Desde entonces, se han presentado muchos estudios que buscan reemplazar a los métodos tradicionales con técnicas de inteligencia artificial/aprendizaje automático debido a su exactitud, confiabilidad, rápida velocidad de respuesta y robusta capacidad de generalización [7] – [10].

En el presente estudio se utilizó la colección de algoritmos de aprendizaje automático del programa Weka [11] para predecir la presión de burbujeo de 36 muestras de petróleo y se determinó la precisión de sus resultados con los métodos de prueba validación cruzada de 10 pliegues y validación con los datos de entrenamiento. Posteriormente, para efectos de comparación, se calcularon las presiones de burbujeo con la correlación generada en el trabajo del cual se tomó la información de las muestras [12]. Las variaciones que ocasionaron los cambios de método de prueba en los resultados del estudio determinaron la conveniencia de su extensión para incorporar muestras de petróleo de diferentes regiones y composiciones, porque cuando los algoritmos se prueben con datos de diferentes bases químicas se podrá evaluar la capacidad que tuvieron de comprender y aprender patrones durante los entrenamientos [13].

Métodos y Metodología computacional

Construcción de la base de datos.

La base de datos se construyó con información de 36 muestras de petróleo provenientes de 12 yacimientos localizados costa afuera de los Emiratos Árabes Unidos (U.A.E.). Esta selección se produjo en virtud de que la presión de burbujeo es función tanto de la composición del petróleo como de la presión y temperatura del yacimiento, por lo que estos factores pueden ser aproximados utilizando las gravedades específicas del gas y del petróleo (gg y go, adimensionales), la relación gas petróleo en solución a presiones mayores o iguales a la de burbujeo (Rsb, PCN/BN) y la temperatura del yacimiento (T, °F) [12]. En la **Tabla 1** se presentan algunos parámetros estadísticos de estos datos y en la **Ecuación 1** se describe su relación:

Tabla 1. Parámetros estadísticos de la base de datos.

Parámetro Estadístico	Presión de Burbujeo (Lpcm)	Gravedad Específica del Gas	Gravedad Específica del Petróleo	Relación Gas-Petróleo en Solución (PCN/BN)	Temperatura (°F)
Máximo	4822	1,116	0,9254	3588	306
Mínimo	541	0,746	0,6731	128	190
Promedio	2190,444	0,932	0,831	810,806	241,444
Desv. Estándar	1154,035	0,089	0,037	799,244	24,976
Coef. de Variación	0,527	0,095	0,045	0,986	0,103

$$Pb = f(\gamma g, \gamma o, Rsb, T) \quad (1)$$

Modelos de aprendizaje automático.

Los modelos de aprendizaje automático desarrollados en este estudio se definen como modelos de regresión, ya que cuando tratan de predecir un caso desconocido producen un resultado numérico (en este caso presión de burbujeo) dentro de un conjunto infinito de posibles resultados [14].

Para desarrollar estos modelos se utilizó la colección de algoritmos del programa Weka (Waikato Environment for Knowledge Analysis), el cual es un software de código abierto emitido bajo la licencia pública general GNU y creado en la universidad de Waikato en Nueva Zelanda. Este contiene herramientas para la preparación, clasificación, regresión, agrupación, minería de reglas de asociación y visualización de datos [11].

Adicionalmente se debe destacar que este programa se considera un punto de referencia en la historia de las investigaciones de minería de datos y aprendizaje automático porque es el único que ha tenido una adopción tan generalizada y se ha mantenido vigente por un período de tiempo tan extenso [15].

Con la base de datos establecida se construyó el archivo Datos.ARFF (Attribute-Relation File Format) utilizado por el programa y sobre el cual se realizaron todas las corridas y sensibilidades. En la **Figura 1** se presenta un ejemplo de las características de uno de los modelos y en el **Anexo 1** se encuentra el archivo.

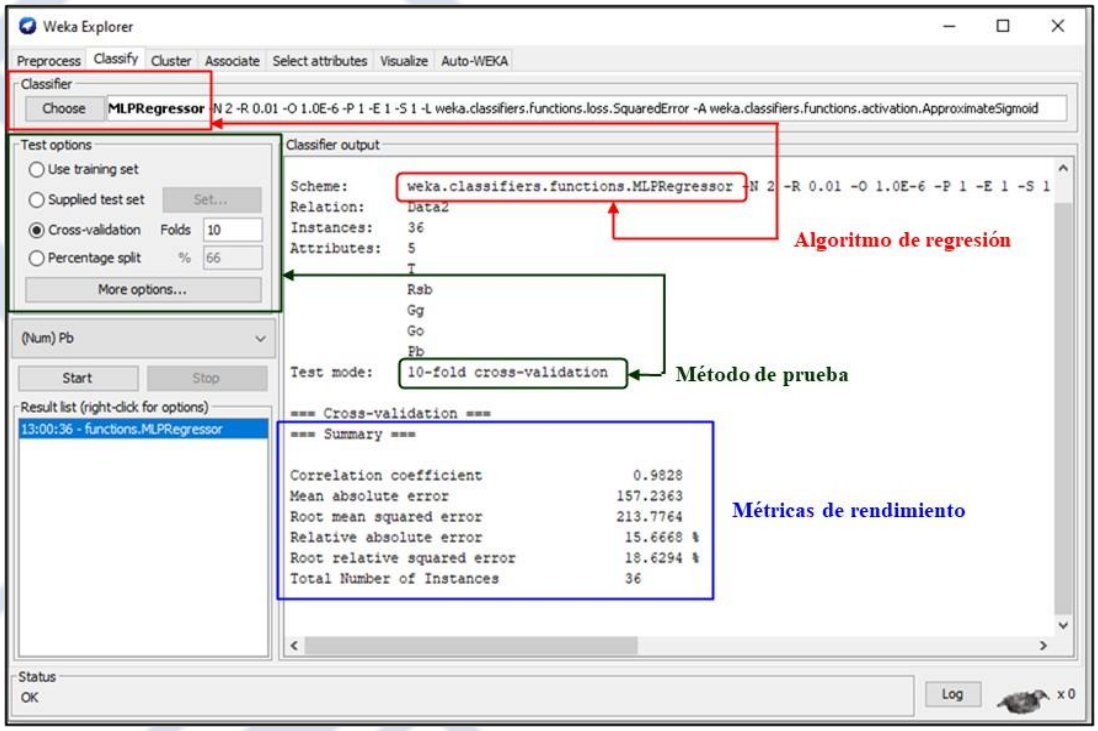


Figura 1. Características de un modelo de regresión de aprendizaje automático con el programa Weka.

Métricas de rendimiento.

El programa Weka utiliza 5 métricas de rendimiento para evaluar la precisión de sus modelos de aprendizaje automático de regresión:

1. Coeficiente de correlación (Correlation coefficient, r^2).
2. Error absoluto medio (Mean absolute error, MAE).
3. Raíz del error cuadrático medio (Root mean squared error, RMSE).
4. Error absoluto relativo (Relative absolute error, RAE).
5. Error absoluto relativo (Relative absolute error, RAE).

En este estudio, además de estas 5 métricas, y en virtud de su uso frecuente en la literatura, también se calcularon los errores porcentuales absolutos (%Eai) de cada predicción y posteriormente, el error porcentual absoluto medio (mean absolute percentage error, MAPE) y la desviación estándar (s). En el **Anexo 2** se encuentran las ecuaciones utilizadas para los cálculos.

El hecho de utilizar 7 métricas de rendimiento (5 del programa Weka y 2 de la literatura) se debe a que cada una de ellas condensa un gran número de datos en un solo valor, por lo que en realidad ninguna es inherentemente mejor que otra, sino que solamente provee una proyección y enfatiza un aspecto de las características del error del modelo. Por consiguiente, y considerando que diferentes tipos de modelos tienen diferentes distribuciones del error, es evidente que se necesitan diferentes métricas (o incluso una combinación de ellas) para poder evaluar la precisión de los resultados de un modelo de [16, 17, 18].

Resultados y discusión

Aprendizaje automático utilizando validación cruzada de 10 pliegues.

Para determinar la precisión que tuvieron los algoritmos de aprendizaje automático del programa Weka para predecir las 36 presiones de burbujeo de la base de datos se realizó una validación cruzada de 10 pliegues. Los resultados indicaron que los 7 algoritmos de mejor rendimiento fueron: MLPRegressor, AdditiveRegression, MultilayerPerceptron, RBFRegressor, Kstar, RandomizableFilteredClassifier y M5Rules. Después de esta primera selección, se realizaron sensibilidades en los parámetros internos de estos algoritmos para definir las configuraciones con

las que se obtuvieron los mejores resultados. En la **Tabla 2** se presentan las métricas de rendimiento:

Tabla 2. Métricas de rendimiento de los 7 algoritmos de aprendizaje automático de mayor precisión.

	Validación cruzada 10 pliegues						
Parámetro Estadístico	MLPRegressor	AdditiveReg.	MultilayPerc.	RBFRegres.	Kstar	RandomF.	M5Rules
Coeficiente de correlación, r^2	0,9911	0,9864	0,9857	0,9660	0,9527	0,9515	0,9403
Error absoluto medio, MAE	114,6275	146,2712	156,0691	207,7583	252,0619	257,5596	305,2491
Raíz cuadrada del error cuadrático medio, RMSE	156,0982	190,3949	200,2859	295,3871	361,0441	352,6713	407,6970
Error relativo absoluto, RAE	11,5198	14,5742	15,5505	20,7007	25,1151	25,6628	30,4146
Raíz del error cuadrado relativo, RRSE	13,7182	16,5918	17,4538	25,7413	31,4629	30,7333	35,5285
Error absoluto porcentual medio, MAPE	6,2865	8,9878	9,2214	12,8721	15,3572	13,9560	17,3722
Desviación estándar, σ	8,9052	13,4847	13,4929	21,1952	28,1750	21,5082	25,3005

Posteriormente, estas métricas fueron normalizadas y representadas gráficamente de forma adimensional para evaluar sus tendencias de forma comparativa (**Figura 2**).

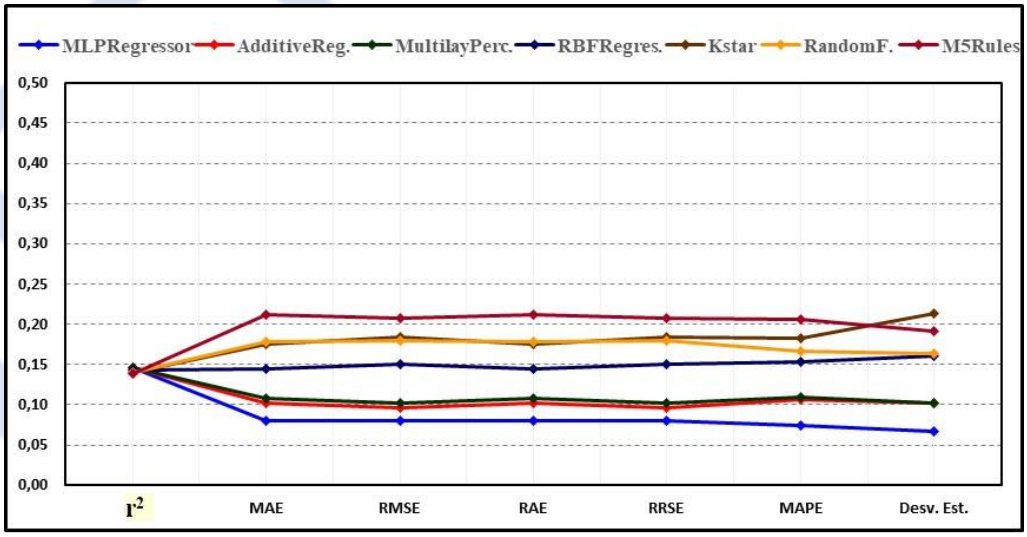


Figura 2. Métricas de rendimiento adimensionales de los 7 algoritmos de aprendizaje automático de mayor precisión.

En la información presentada se observa que con la validación cruzada el algoritmo de mayor precisión fue MLPRegressor, sin embargo, en virtud de que AdditiveRegression y MultilayerPerceptron también obtuvieron buenos resultados y sus tendencias tuvieron un comportamiento similar, se seleccionaron estos 3 para continuar el estudio y se descartó el resto.

Correlación generada utilizando análisis de regresión.

En el trabajo del cual se tomó la información de las 36 muestras de petróleo se realizó un análisis de regresión para generar la correlación que mejor ajustaba las propiedades bajo consideración. En este estudio, para efectos de comparación, se utilizó esa correlación para calcular las presiones de burbujeo y sus resultados fueron más precisos que los obtenidos por los algoritmos de aprendizaje automático en 4 de las 7 métricas de rendimiento: MAE, RAE, MAPE y s (MLPRegressor lo fue en r^2, RMSE y RRSE).

Esta elevada precisión que tuvo la correlación se debe tratar con cuidado, porque se consiguió en una prueba con los mismos datos con los que fue generada, los cuales como ya tienen la tendencia regional de los Emiratos Árabes Unidos, "ya conocen" los resultados correctos, pero si la prueba se realiza con muestras de petróleo de diferentes regiones o composiciones (las cuales tienen diferentes bases químicas), se pueden obtener errores significativos [19].

Por el contrario, cuando el aprendizaje automático realiza la validación cruzada de 10 pliegues se está desarrollando un modelo mucho más generalizado, menos sesgado y que no tiene tanta dependencia de los datos utilizados porque se construye 10 veces tomando cada vez 1/10 de los datos para la prueba y el resto para la construcción. Después que el proceso se ha repetido las 10 veces se calcula un promedio con la precisión de cada uno de los modelos [20].

En la **Tabla 3** y **Figura 3** se presenta una comparación de las métricas de rendimiento de los algoritmos de aprendizaje automático y de la correlación:

Tabla 3. Comparación de las métricas de rendimiento de los algoritmos de aprendizaje automático y de la correlación.

	Validación cruzada 10 pliegues			
Parámetro Estadístico	MLPRegressor	AdditiveReg.	MultilayPerc.	Correlación
Coeficiente de correlación, r^2	0,9911	0,9864	0,9857	0,9846
Error absoluto medio, MAE	114,6275	146,2712	156,0691	114,4554
Raíz cuadrada del error cuadrático medio, RMSE	156,0982	190,3949	200,2859	203,7676
Error relativo absoluto, RAE	11,5198	14,5742	15,5505	11,5025
Raíz del error cuadrado relativo, RRSE	13,7182	16,5918	17,4538	17,9074
Error absoluto porcentual medio, MAPE	6,2865	8,9878	9,2214	4,5215
Desviación estándar, σ	8,9052	13,4847	13,4929	6,2040

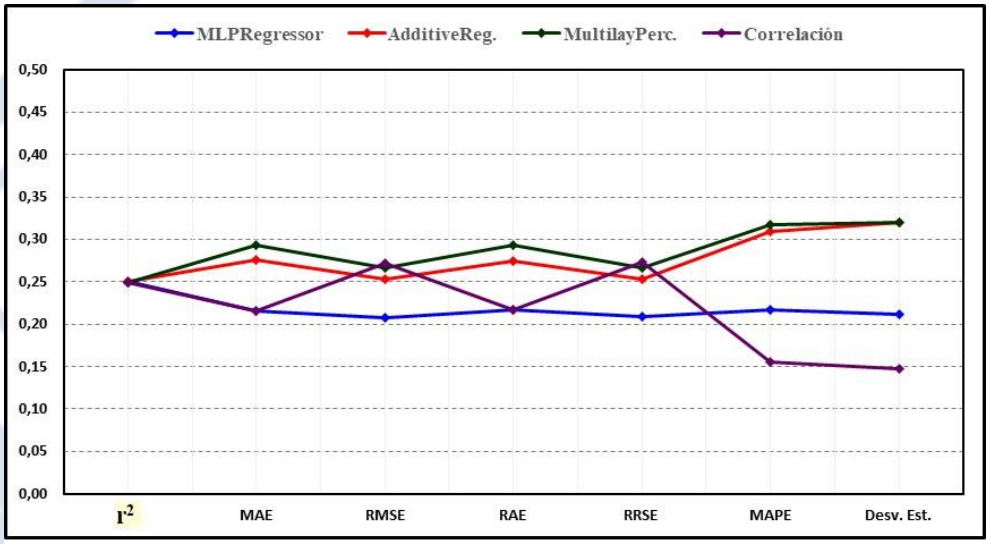


Figura 3. Comparación de las métricas de rendimiento adimensionales de los algoritmos de aprendizaje automático y de la correlación.

Aprendizaje automático utilizando validación con los datos de entrenamiento.

Para comparar en igualdad de condiciones los algoritmos de aprendizaje automático y la correlación se cambió el método de prueba a validación con los datos de entrenamiento y se volvieron a predecir las presiones de burbujeo. De igual manera, también se volvieron a realizar sensibilidades en los parámetros internos de los algoritmos y estas demostraron que existen diferencias entre las configuraciones con las que se obtuvieron los mejores resultados utilizando validación cruzada de 10 pliegues y las que lo hacen cuando se valida con los datos de entrenamiento. En la **Tabla 4** y **Figura 4** se presenta la comparación de las métricas de rendimiento:

Tabla 4. Comparación de las métricas de rendimiento de los algoritmos de aprendizaje automático y de la correlación.

	Validación con los datos de entrenamiento			
Parámetro Estadístico	MLPRegressor	AdditiveReg.	MultilayPerc.	Correlación
Coeficiente de correlación, r^2	0,9950	0,99997	0,9961	0,9846
Error absoluto medio, MAE	91,5330	5,5375	99,4378	114,4554
Raíz cuadrada del error cuadrático medio, RMSE	114,0809	8,2665	125,2687	203,7676
Error relativo absoluto, RAE	9,1989	0,5565	9,9933	11,5025
Raíz del error cuadrado relativo, RRSE	10,0256	0,7265	11,0088	17,9074
Error absoluto porcentual medio, MAPE	5,4272	0,4119	6,0424	4,5215
Desviación estándar, σ	7,9282	0,7362	8,4451	6,2040

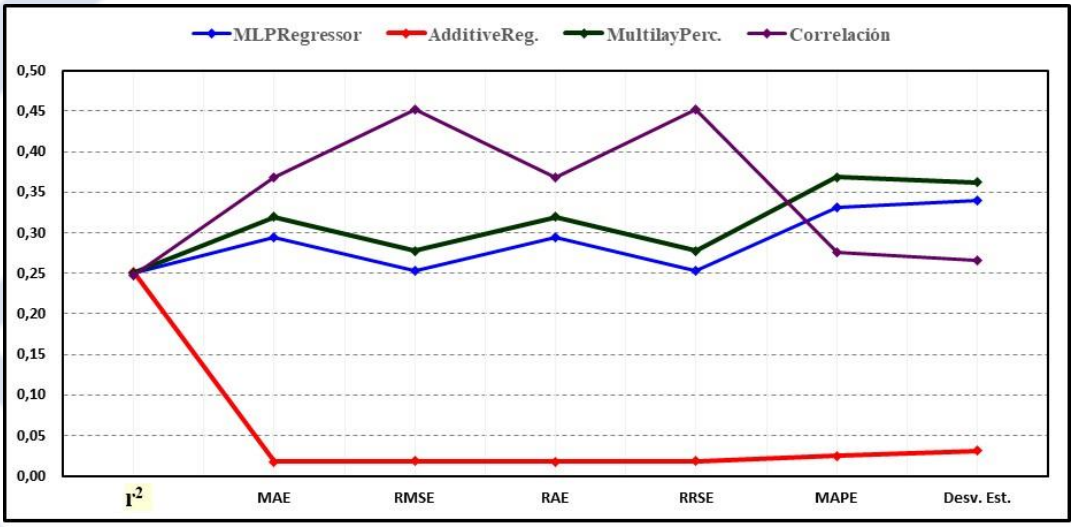


Figura 4. Comparación de las métricas de rendimiento adimensionales de los algoritmos de aprendizaje automático y de la correlación.

En la información presentada se observa que el algoritmo AdditiveRegression obtuvo los mejores resultados en todas las métricas de rendimiento, mientras que MLPRegressor y MultilayerPerceptron fueron más precisos en r^2, MAE, RMSE, RAE y RRSE que la correlación (esta fue superior en MAPE y s).

Capacidad de comprender y aprender patrones de los algoritmos de aprendizaje automático.

Los análisis numéricos y gráficos de las métricas de rendimiento de los resultados determinaron que con la validación cruzada el algoritmo MLPRegressor obtuvo los mejores resultados en 3 métricas de rendimiento (la correlación lo hizo en 4), y que posteriormente, cuando se cambió el método de prueba a validación con los datos de entrenamiento fue el algoritmo AdditiveRegression el que obtuvo mejores resultados en todas las 7 métricas de rendimiento (MLPRegressor y MultilayerPerceptron fueron superiores a la correlación en 5).
Las variaciones que ocasionaron los cambios de método de prueba en los resultados del estudio determinan la conveniencia de su extensión para incorporar muestras de petróleo de diferentes regiones y composiciones, porque cuando los algoritmos se prueben con datos de diferentes bases químicas se podrá evaluar la capacidad que tuvieron de comprender y aprender patrones durante los entrenamientos.

Conclusiones

Los algoritmos de aprendizaje automático del programa Weka que obtuvieron mayor precisión para predecir la presión de burbujeo de las 36 muestras de petróleo utilizando validación cruzada de 10 pliegues fueron MPLRegressor, AdditiveRegression y MultilayerPerceptron.

Cuando se utilizó validación cruzada de 10 pliegues el algoritmo MLPRegressor obtuvo resultados más precisos en 3 métricas de rendimiento y la correlación generada con análisis de regresión lo hizo en 4.

Cuando se utilizó validación con los datos de entrenamiento el algoritmo AdditiveRegression obtuvo resultados más precisos en todas las métricas de rendimiento.
Las sensibilidades realizadas en los parámetros internos de los algoritmos de aprendizaje automático demostraron que existen diferencias entre las configuraciones con las que se obtuvieron los mejores resultados utilizando validación cruzada de 10 pliegues y las que lo hacen cuando se valida con los datos de entrenamiento.

Los resultados del estudio determinaron la conveniencia de su extensión para incorporar muestras de petróleo de diferentes regiones y composiciones, porque cuando los algoritmos se prueben con datos de diferentes bases químicas se podrá evaluar la capacidad que tuvieron de comprender y aprender patrones durante los entrenamientos.

Referencias

[1] X. Yang, B. Dindoruk and L. Lu, "A comparative analysis of bubble point pressure prediction using advanced machine learning algorithms and classical correlations", Journal of Petroleum Science and Engineering, vol. 185, 106598, 2020. Available: https://doi.org/10.1016/j.petrol.2019.106598.

[2] M. M. Almashan, Z. Arsalan, Y. Narusue and H. Morikawa, "Estimating Pressure-Volume-Temperature Properties of Crude Oil Systems Using Boosted Decision Tree Regression", Journal of the Japan Petroleum Institute, vol. 65, nº. 6, pp. 221-232, 2022. Available: https://doi.org/10.1627/jpi.65.221.

[3] K. Ghorayeb, A. Mawlod, A. Maarouf, Q. Sami, N. El Droubi, R. Merrill, O. El Jundi and H. Mustapha, "Chain-based machine learning for full PVT data prediction", Journal of Petroleum Science and Engineering, vol. 208, Part D, 109658, 2022. Available: https://doi.org/10.1016/j.petrol.2021.109658.

[4] A. M. Elsharkawy, "Modeling the Properties of Crude Oil and Gas Systems Using RBF Network." Paper presented at the SPE Asia Pacific Oil and Gas Conference and Exhibition, Perth, Australia, October 1998. Available: https://doi.org/10.2118/49961-MS.

[5] R. B. Gharbi and A. M. Elsharkawy, "Neural network model for estimating the PVT properties of Middle East crude oils". Paper presented at the Middle East Oil Show and Conference, Bahrain, March 1997. Available: https://doi.org/10.2118/37695-MS.

[6] R. B. Gharbi, A. M. Elsharkawy and M. Karkoub, "Universal Neural-Network-Based Model for Estimating the PVT Properties of Crude Oil Systems". Energy Fuels, vol. 13, pp. 454–458, 1999. Available: https://doi.org/10.1021/ef980143v.

[7] S. Alatefi and A. M. Almeshal. "A New Model for Estimation of Bubble Point Pressure Using a Bayesian Optimized Least Square Gradient Boosting Ensemble". Energies, vol. 14, nº. 9, pp. 2653, 2021. Available: https://doi.org/10.3390/en14092653.

[8] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist and H. Oza. "Application of machine learning and artificial intelligence in oil and gas industry". Petroleum Research, vol. 6, n°. 4, pp. 379-391, 2021. Available: https://doi.org/10.1016/j.ptlrs.2021.05.009.

[9] M. Ahmadi, M. Pournik and S. Shadizadeh. "Toward connectionist model for predicting bubble point pressure of crude oils: Application of artificial intelligence". Petroleum, vol. 1, n°. 4, pp. 307-317, 2015. Available: https://doi.org/10.1016/j.petlm.2015.08.003.

[10] F. Alakbari, M. Mohyaldinn, M. Ayoub, A. Muhsan and I. Hussein. "A reservoir bubble point pressure prediction model using the Adaptive Neuro-Fuzzy Inference System (ANFIS) technique with trend analysis. PLoS ONE vol. 17, n°. 8, e0272790, 2022. Available: https://doi.org/10.1371/journal.pone.0272790.

[11] Weka 3-Data Mining with Open Source Machine Learning Software in Java. Available online: https://www.cs.waikato.ac.nz/ml/weka/.

[12] R. A. Al-Mehaideb, "Improved PVT Correlations for UAE Offshore Crudes", Journal of The Japan Petroleum Institute, vol. 40, nº. 3, pp. 232-235, 1997. Available: https://doi.org/10.1627/jpi1958.40.232.

[13] S.M. Macary and M.H. El-Batanoney "Derivation of PVT Correlations for the Gulf of Suez Crude Oils", Journal of The Japan Petroleum Institute, vol. 36, nº. 6, pp. 472-478, 1993. Available: https://doi.org/10.1627/jpi1958.36.472.

[14] IArtificial.net: ¿Clasificación o Regresión? [Online]. Available: https://www.iartificial.net/clasificacion-o-regresion/.

[15] M. Senthamilselvi and P.S.S. Akilashri, "A Comparative Study on Weka, orange Tool for Mushroom data Set", International Journal of Computer Sciences and Engineering, vol. 06, nº. 11, pp. 231-236, 2018. Available: www.ijcseonline.org.

[16] T. Chai and R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literatura", Geoscientific Model Development, vol. 7, nº. 3, pp. 1247–1250, 2014. Available: https://doi.org/10.5194/gmd-7-1247-2014, 2014.

[17] T. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not", Geoscientific Model Development, vol. 15, nº. 14, pp. 5481–5487, 2022. Available: https://doi.org/10.5194/gmd-15-5481-2022.

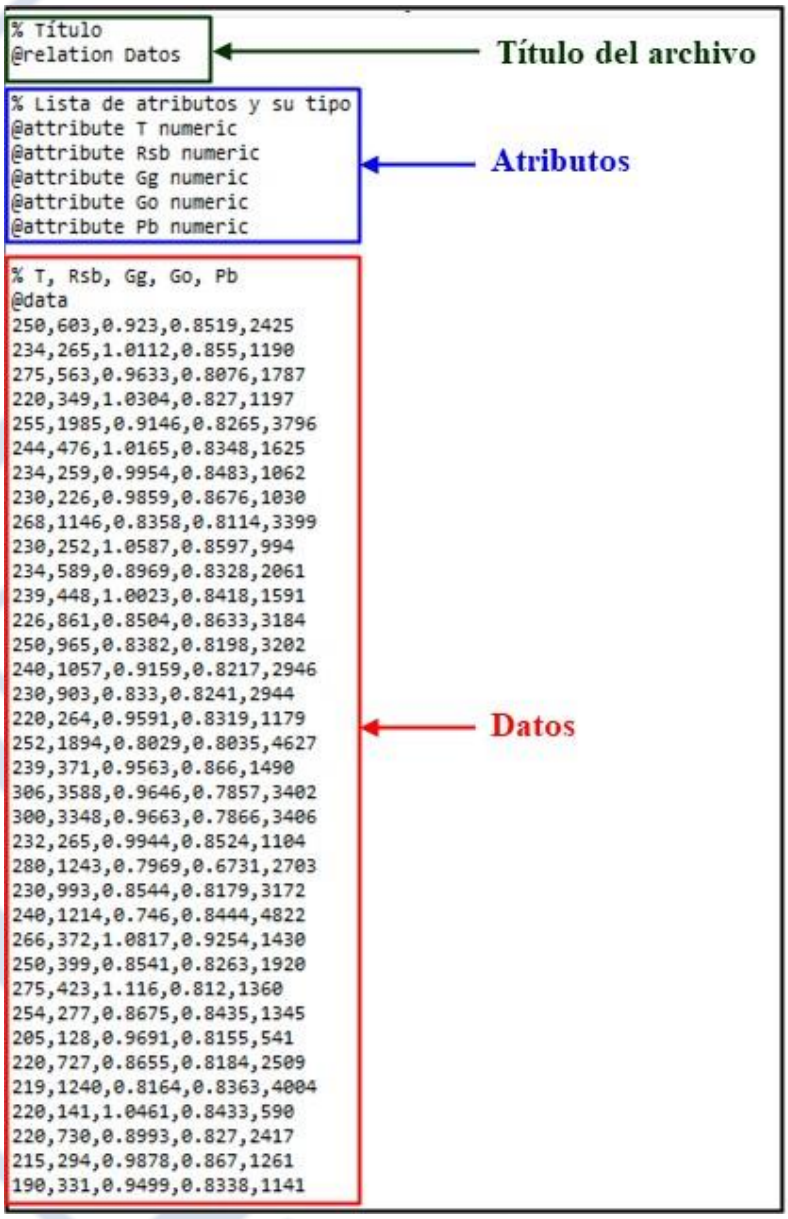
[18] S. Taipe y G. Ampuño, "Modelo del proceso de producción de energía en centrales de generación térmica considerando el perfil de funcionamiento", Ciencia Latina Revista Científica Multidisciplinar, vol. 6, nº. 4, pp. 5541-5560, 2022. Available: https://doi.org/10.37811/cl_rcm.v6i4.3032.

[19] B. Moradi, E. Malekzadeh, M. Amani, F. Boukadi, and R. Kharrat. "Bubble Point Pressure Empirical Correlation." Paper presented at the Trinidad and Tobago Energy Resources Conference, Port of Spain, Trinidad, June 2010. Available: https://doi.org/10.2118/132756-MS.

[20] Data science: Validación cruzada K-Fold [Online]. Available: https://datascience.eu/es/aprendizaje-automatico/validacion-cruzada-de-k-fold/.

Anexos

Anexo 1. Archivo Datos.ARFF utilizado por el programa Weka.



Anexo 2. Ecuaciones utilizadas para calcular las 7 métricas de rendimiento y los errores porcentuales absolutos.

Nomenclaturas:
Xi = Valores reales de la presión de burbujeo.
Ẍ = Promedio aritmético de los valores reales de la presión de burbujeo.
Yi = Valores calculados de la presión de burbujeo.
Ÿ = Promedio aritmético de los valores calculados de la presión de burbujeo.
n = Número de muestras.

A.2.1. Coeficiente de correlación (Correlation coefficient, r^2).

$$r^2 = \frac{\sum_{i=1}^{n}(Xi - \ddot{X}) * (Yi - \ddot{Y})}{\sqrt{\sum_{i=1}^{n}(Xi - \ddot{X})^2 \sum_{i=1}^{n}(Yi - \ddot{Y})^2}}$$

A.2.2. Error absoluto medio (Mean absolute error, MAE).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Xi - Yi|$$

A.2.3. Raíz del error cuadrático medio (Root mean squared error, RMSE).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Xi - Yi)^2}$$

A.2.4. Error absoluto relativo (Relative absolute error, RAE).

$$RAE = \frac{\sum_{i=1}^{n}|Xi - Yi|}{\sum_{i=1}^{n}|Xi - \ddot{X}|}$$

A.2.5. Raíz del error cuadrado relativo (Root relative squared error, RRSE).

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n}(Xi - Yi)^2}{\sum_{i=1}^{n}(Xi - \ddot{X})^2}}$$

Errores porcentuales absolutos (%Eai).

$$\%Eai = \left|\frac{Xi - Yi}{Xi}\right| * 100$$

A.2.6. Error porcentual absoluto medio (mean absolute percentage error, MAPE).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|Xi - Yi|}{Xi} * 100$$

A.2.7. Desviación estándar (s).

$$\sigma = \sqrt{\frac{1}{n-1}\left[\left(\frac{Xi - Yi}{Xi}\right) * 100\right]^2}$$

Seguridad de la información en el comercio electrónico basado en ISO 27001 : Una revisión sistemática

Information security in e-commerce based on ISO 27001: A systematic review

Gerson De La Cruz Rodríguez
Universidad Nacional de Trujillo. La Libertad, Perú.
gdelacruz@unitru.edu.pe
https://orcid.org/0000-0002-9276-3376

Ronny A. Méndez Fernández
Universidad Nacional de Trujillo. La Libertad, Perú.
rmendezf@unitru.edu.pe
https://orcid.org/0000-0003-0867-7326

Alberto Carlos Mendoza De Los Santos
Universidad Nacional de Trujillo. La Libertad, Perú.
amendozad@unitru.edu.pe
https://orcid.org/0000-0002-0469-915X

ARK: ark:/42411/s11/a79
PURL: 42411/s11/a79

RECIBIDO 30/11/2022 • ACEPTADO 10/01/2023 • PUBLICADO 30/03/2023

RESUMEN

En los últimos años, con la popularización tan acelerada del eCommerce (comercio electrónico), que facilita mucho la vida de las personas que, solo dando un clic, tiene la posibilidad de adquirir innumerables productos prescindiendo de la infraestructura física del mundo real. Este crecimiento va de la mano con la seguridad de la información por el valor de esta por lo tanto se vio necesario analizar las evidencias aportadas desde la investigación para conocer el estado actual de la gestión de la seguridad de la información en el ámbito del eCommerce. Se ha llevado a cabo una revisión sistemática siguiendo las directrices PRISMA de los artículos publicados encontrados en Scopus, incluyendo un total de 6 artículos. Los resultados señalan consistentemente que los sistemas de eCommerce son vulnerables en gran manera, y para esto se requiere de una mejora en la gestión de la seguridad de la información y una gestión de riesgos de seguridad consciente de las amenazas que van en aumento, para así ofrecer un buen servicio de ciberseguridad. Actualmente se encuentran en el mercado muchos gestores que ayudan a tener segura la información de las empresas, los cuales abarcan las necesidades de los sistemas y sus vulnerabilidades en conjunto, correspondientes a la gestión de la seguridad de la información relacionada con el eCommerce, pero la norma ISO 27001 abarca en gran manera

muchas áreas de la seguridad de la información en una empresa, la cual brinda una mayor protección y confianza de los datos de sus clientes.

Palabras claves: Ciberseguridad, comercio electrónico, gestión de la seguridad de la información, ISO 27001, seguridad de la información..

ABSTRACT

In recent years, with the rapid popularization of eCommerce (electronic commerce), which greatly facilitates the lives of people who, with just one click, have the possibility of acquiring innumerable products regardless of the physical infrastructure of the real world. This growth goes hand in hand with the security of information due to its value, therefore it was necessary to analyze the evidence provided from the investigation to know the current state of information security management in the field of eCommerce. A systematic review has been carried out following the PRISMA guidelines of the published articles found in Scopus, including a total of 6 articles. The results consistently indicate that eCommerce systems are highly vulnerable, and this requires an improvement in information security management and security risk management aware of the threats that are increasing, in order to offer a good cybersecurity service. Currently there are many managers on the market that help to keep company information secure, which cover the needs of the systems and their vulnerabilities as a whole, corresponding to the management of information security related to eCommerce, but the ISO 27001 standard largely covers many areas of information security in a company, which provides greater protection and confidence in customer data.

Keywords: *Cybersecurity, e-commerce, information security management, information security, ISO 27001.*

INTRODUCCIÓN

En los últimos años, con la popularización tan acelerada del comercio electrónico que facilita mucho la vida de las personas que solo dando un clic tienen la posibilidad de adquirir innumerables productos prescindiendo de la infraestructura física del mundo real, ha llegado a tomar un papel muy importante la seguridad de la información de los clientes que realizan este tipo de transacciones.

Definiendo de forma técnica, según Torre y Codner, "electrónico hace referencia a la infraestructura mundial de la información, compuesta por la conjunción del hardware, el software, las redes informáticas y las telecomunicaciones, que permiten la transmisión, el procesamiento, el almacenamiento y la recuperación de datos en formato digital. En conjunto, estas tecnologías

han dado origen a Internet, una gran red de carácter abierto y multifuncional cuyo acceso es cada vez más económico y amigable para gran parte de la población mundial" [1].

Y ya adentrándonos en el comercio electrónico para Turban, Volonino y Wood, "el comercio electrónico describe el proceso de compra, venta, transferencia, servicio o intercambio de productos y servicios o información mediante una red de computadores, incluyendo Internet" [2]. Podemos ahora decir que el comercio electrónico es una serie de operaciones comerciales y financieras realizadas mediante el procesamiento y la transmisión de información. Dicha información puede ser el objeto principal o un elemento relacionado con una transacción, lo que incluye compartir información sobre un negocio entre proveedores, consumidores, agencias gubernamentales y otras organizaciones a través de cualquier medio electrónico como correo principal o un elemento relacionado con una transacción, lo que incluye compartir información sobre un negocio entre proveedores, consumidores, agencias gubernamentales y otras organizaciones a través de cualquier medio electrónico como correo electrónico, sitio web, etc., para realizar y ejecutar transacciones en actividades comerciales, administrativas y de consumo. Para [3] "las tiendas virtuales son páginas web cuyo objetivo es la venta de productos o servicios", ofreciendo al cliente un nuevo espacio para realizar transacciones en sus compras, de una forma más rápida desde cualquier lugar y a cualquier hora con acceso a todos los productos y el pago sea de forma virtual y rápida con un envió según la estructura física de lo adquirido.
El comercio electrónico se define como "transacciones comerciales habilitadas de manera digital entre organizaciones e individuos". En donde estas transacciones sean mediadas a través de la tecnología digital, es decir, internet y la web, y en la cual se encuentren involucrados el intercambio de valores, como el dinero, entre los límites organizacionales o individuales a cambio de productos y servicios [4].

"El comercio electrónico entendido en sentido estricto cubre, principalmente, dos tipos de actividades: el pedido electrónico de bienes materiales que se entregan a través de canales tradicionales como el correo o los servicios de mensajería (comercio electrónico indirecto, que depende de factores externos, como la eficacia del sistema de transporte); y el pedido, el pago y la entrega en línea de bienes y servicios intangibles, como programas informáticos, revistas electrónicas, servicios recreativos y de información (comercio electrónico directo, que aprovecha todo el potencial de los mercados electrónicos mundiales)" [5].

De manera más doctrinaria, según [6] "el comercio electrónico constituye un fenómeno jurídico y se concibe como la oferta y la contratación electrónica de productos y servicios a través de dos o más ordenadores o terminales informáticos conectados a través de una línea de comunicación dentro del entorno de red abierta que constituye Internet. Representa un fenómeno en plena expansión con votos de crecimiento extraordinario en número de conexiones, clientes y operaciones".

En cuanto a la Seguridad de la Información, es uno de los conceptos más importantes en el comercio electrónico, y es muy importante tener en claro lo que esto significa. Muchos son los riesgos que han tenido las tiendas virtuales desde que comenzaron hasta la actualidad, se debe saber que ningún sistema es seguro al 100%, pero se debe cubrir las brechas existentes para que la información no llegue a personas no autorizadas.

"La información, es como el aparato circulatorio para las organizaciones y requiere que se proteja ante cualquier amenaza que pueda poner en peligro las empresas, tanto públicas como privadas, pues en otro caso podría dañarse la salud empresarial. La realidad nos muestra, que las organizaciones empresariales se enfrentan en la actualidad con un alto número de riesgos e inseguridades procedentes de una amplia variedad de fuentes" [7].

"Un Sistema de Gestión de Seguridad de la Información (SGSI) es, tal como su nombre lo indica, un elemento para administración relacionado con la seguridad de la información, aspecto fundamental de cualquier empresa. Un SGSI implica crear un plan de diseño, implementación, y mantenimiento de una serie de procesos que permitan gestionar de manera eficiente la información, para asegurar la integridad, confidencialidad y disponibilidad de la información" [8]. La norma ISO 27001 especifica los requisitos para una correcta implementación de un sistema de gestión de la seguridad de la información. Esta norma fue publicada por primera vez en el año 2005 y brinda las pautas para establecer, implementar, mantener y mejorar continuamente dicho sistema dentro del contexto de la organización [9].

"En los actuales momentos la norma ISO 27001:2007, presenta un compendio que proporciona una base común para la elaboración de reglas, un método de gestión eficaz de la seguridad y permite establecer informes de confianza en las transacciones y las relaciones entre empresas" [10].

La seguridad de la información son todas las medidas preventivas y reactivas de las personas, organizaciones y sistemas técnicos que permiten guardar y proteger la información con el fin de mantener su confidencialidad, autenticidad e integridad.

Podríamos decir también que la seguridad de información se enfoca en la data o aquellos activos que tienen las empresas y se ven representados de diferentes formas como correo, papel, etc. Si se le llega a dar un mal uso a esta información, se puede comprometer de forma negativa a la empresa; en conclusión, es el conjunto de medidas que le permite a la empresa asegurar la confidencialidad, integridad y disponibilidad de su información, lo que conocemos como la triada de la información.

"La seguridad de la información es conjunto de técnicas y medidas para controlar todos los datos que se manejan dentro de una institución y asegurar que no salgan de ese sistema establecido por la empresa" [11].

Podemos ahora indicar que la seguridad de la información es la protección de la confidencialidad, integridad y disponibilidad de la información (ver Figura 1.1); es decir, cuidar de la triada de la información, lo que esto quiere decir es que la información sea accesible solo a las personas autorizadas, sea exacta sin modificaciones no deseadas y que esté disponible a los usuarios cuando lo requieran.

Figura 1. Triada de la Información

Métodos y Metodología computacional

Tipo de Estudio

En este trabajo se ha llevado a cabo una revisión sistemática de la literatura científica publicada en materia de seguridad de la información en aplicaciones de compraventa electrónica. Para su elaboración, se han seguido las directrices de la metodología PRISMA para la correcta elaboración de revisiones sistemáticas.

La pregunta seleccionada que conducirá el proceso metodológico fue la siguiente: ¿Cuál es el estado actual de la seguridad de la información en el comercio electrónico basado en ISO 27001?

Fundamentación de la Metodología

La revisión sistemática es la evaluación ordenada y explícita de la literatura a partir de una pregunta clara de investigación, junto a un análisis crítico de acuerdo a diferentes herramientas y un resumen cualitativo de la evidencia [12].

Teniendo en cuenta esta definición podemos ver la importancia de sintetizar información para poder destacar lo más relevante ante la gran diversidad de documentación existente que tratan el mismo punto desde diferentes contextos pero que tienen el mismo objetivo al final de todo.
A continuación, se detalla el proceso de realización en sus distintas etapas.

Búsqueda inicial

Las primeras búsquedas se realizaron en la primera semana de octubre del 2022 combinando los términos 'seguridad de la información', 'compraventa electrónica', 'ISO 27001' en las bases de datos Scopus, Scielo, PubMed, Cochrane y Eric. Posteriormente, se amplió con una combinación, usando los operadores booleanos AND y OR según convino, de los términos 'e-commerce', 'ecommerce', 'electronic commerce', 'ISO 27001', 'ISO/EIC', 'comercio electrónico', 'compraventa electrónica', 'tienda virtual', 'security information'. Todas estas búsquedas nos dieron como resultado una cantidad considerable de resultados, dentro de los cuales algunos se repetían, sin embargo, esto nos permitió tener una vista global de la amplitud del tema. Debido a la poca cantidad de artículos encontrados en "Cochrane y Eric" y que estos además tenían poca relación y/o relevancia para con la revisión, se optó por retirarlos de la búsqueda sistemática.

Búsqueda sistemática

Se realizó una nueva búsqueda sistemática en la segunda semana de octubre de 2022, en Scopus, Scielo y PubMed.
Acortando la cantidad de resultados a las publicaciones realizadas en un marco temporal no mayor a 5 años, es decir, con fecha de publicación desde el año 2018 (incluyéndolo) hasta la actualidad.
La fórmula de búsqueda que se usó en el buscador **Scopus** fue:
KEY (e-commerce, OR ecommerce, OR information OR security OR management, OR information OR security, OR cybersecurity, AND iso AND 27001.) AND (LIMIT-TO (PUBYEAR, 2022) OR LIMIT-TO (PUBYEAR, 2021) OR LIMIT-TO (PUBYEAR, 2020) OR LIMIT-TO (PUBYEAR, 2019) OR LIMIT-TO (PUBYEAR, 2018)) AND (LIMIT-TO (LANGUAGE, "English") OR LIMIT-TO (LANGUAGE, "Spanish"))

En los buscadores **Pubmed** se usó la siguiente fórmula de búsqueda:
("information security management" OR "information security" OR security OR cybersecurity OR cyber-security OR e-commerce OR ecommerce OR "electronic commerce" OR e-business OR "online shopping" AND ISO 27001) Filters: in the last 5 years, English, Spanish.
Mientras que en **Scielo** fue:
("seguridad" OR "seguridad de la información" OR "information security management" OR "information security" OR security OR cybersecurity OR cyber-security OR "comercio electronico" OR "venta en linea" OR e-commerce OR ecommerce OR "electronic commerce" OR e-business OR "online shopping" OR ISO 27001).

Concretamente, se obtuvieron 80 resultados en Scopus, 10 en Scielo y 8 en PubMed. Antes de proceder a depurar los artículos, se definieron los criterios de inclusión y exclusión.

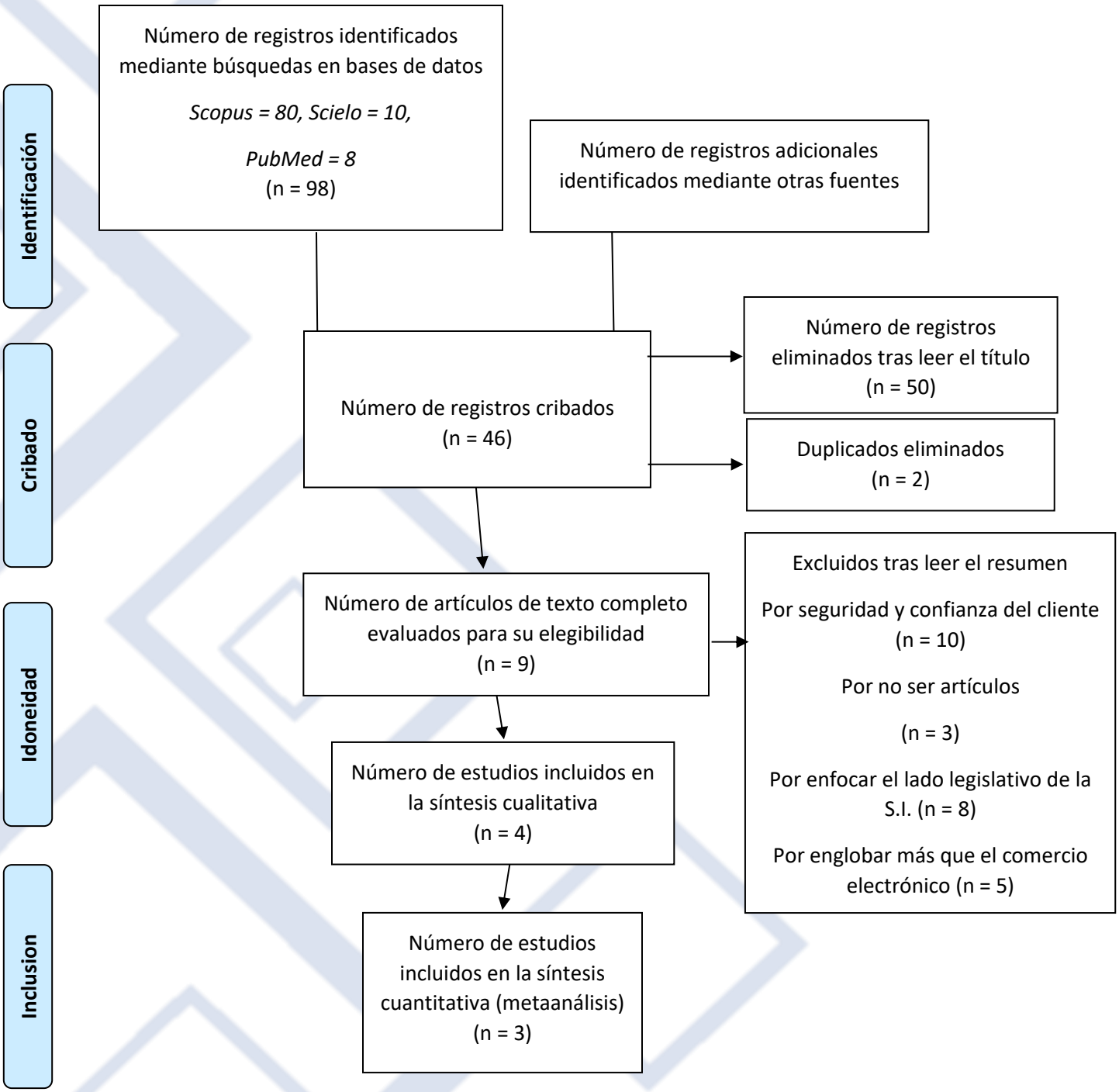


Figura 2. Diagrama de flujo PRISMA en cuatro niveles

Criterios de inclusión

* Para el desarrollo de la presente revisión sistemática, se incluyeron artículos originales publicados entre los años 2018 y 2022.
* Se seleccionan artículos redactados en inglés y español.
* Los artículos deben describir un enfoque de la gestión de la seguridad de la información en el comercio electrónico desde la perspectiva del sistema y la seguridad del mismo basado en la norma ISO 27001.
* Los artículos deben tartar comercio electrónico local como internacional.

Criterios de exclusión

* Se excluyeron aquellos artículos que se enfocan en la seguridad personal y confianza del cliente para participar en el comercio electrónico.
* Los documentos que no eran artículos.
* Aquellos que hablan del lado legislativo de la seguridad de la información.
* Los que englobaban contextos irrelevantes para la revisión (salud, finanzas, educación, etc.)
* Los que usaban tecnologías poco conocidas y/o sospechosas.

Según estos criterios, y sólo con la lectura del título, se consideraron adecuados 46 artículos (tras eliminar 2 articulo duplicado). Se procedió a leer el resumen y, a partir de esta lectura, se descartaron 30, principalmente por centrarse en la perspectiva del consumidor y la confianza del mismo para su participación en el comercio electrónico (n = 10), por tratarse de artículos de revisión(n = 3), por darle un enfoque más legislativo o político (n = 8), por tratar temas más generales que solo el comercio electrónico (n = 5) y por usar tecnologías poco mencionadas y/o sospechosas, lo que dificultaría la interpretación y síntesis de los resultados (n = 4).
Finalmente, 9 cumplieron los criterios de inclusión y se seleccionaron para llevar a cabo la revisión sistemática, pero 2 de ellos eran inaccesibles, nos indicaba error al tratar de ubicarlos mediante el link que nos proporcionaba la base de datos (Scopus), por lo tanto, se redujeron a 7 los artículos que formarán parte de la revisión sistemática. Todos ellos trataban el tema de la seguridad de la información en el comercio electrónico, así como algunos proponen soluciones o sistemas de mejora para el mismo y dentro de un contexto de compraventa entre empresas, consumidores o empresa y consumidor (ver Figura 2).

Resultados y discusión

En primera instancia se procedió a identificar los datos de los artículos seleccionados para poder ver el desenvolvimiento y abordamiento del tema en el mundo como se muestran en las revistas donde fueron publicados, todo esto expresado en la Tabla 1.

Tabla 1. Artículos elegidos para la revisión sistemática

N.º	Año	Autor(es)	País	Base de datos	Título
1	2019	Ehikioya, Sylvanus A. Olukunle, Adepele A.	Nigeria, Canadá	Scopus	A formal model of distributed security for electronic commerce transactions systems
2	2020	Akinyede, Raphael Olufemi Adegbenro, Sulaiman Omolade Omilodi, Babatola Moses	Nigeria	Scopus	A security model for preventing e-commerce related crimes
3	2019	Khan, Shazia W	India	Scopus	Cyber Security Issues and Challenges in E-Commerce
4	2020	Dushyant, Kaushik Ankur, Gupta Swati, Gupta	India	Google Scholar	E-Commerce Security Challenges A Review
5	2022	Alfadli, Ibrahim	Arabia Saudita	Scopus	Integrated e-commerce security model for websites

6	2020	Affia, Abasi Amefon O. Matulevičius, Raimundas Nolte, Alexander	Estonia, USA	Scopus	Security risk management in E-commerce systems: A threat-driven approach
7	2020	Christopher A. Kanter-Ramirez, Josue A. Lopez-Leyva, Lucia Beltran-Rocha & Dominica Ferková	México, Austria	Scopus	Marco para el diseño óptimo de un sistema de información para diagnosticar el nivel de seguridad empresarial y gestionar el riesgo de la información basado en ISO/IEC-27001

En la Figura 3 se muestra un gráfico donde se representan la cantidad de autores, distribuidos por nacionalidad, que participaron en la redacción de los documentos seleccionados para la revisión, se considera como 0.5 de valor a aquellos autores que participaron en conjunto, pero su compañero era de otra nacionalidad.

Esto nos permite ver en qué partes del globo se encuentran los investigadores que abordan el tema de estudio de nuestra revisión sistemática.

Ahora mostraremos el porcentaje de artículos por año (ver Tabla 2), donde se observa que, en el año 2018 y 2021 se tiene un porcentaje de 0%, esto quiere decir que en estos años no se publicaron artículos relevantes que hayan pasado nuestros filtros de selección, en el año 2019 se encuentran 2 publicaciones, en el año 2020 se realizaron más de la mitad de las publicaciones que fueron seleccionadas para la realización de esta revisión sistemática y finalmente en el año 2022 se obtuvo un porcentaje de 14,29% siendo el año con menos publicaciones pero existentes, es decir, más que 0.

El proceso de compra es muy complicado, ya que se debe mantener toda la información protegida desde que los clientes hacen su pedido, pasando por el pago y envío, por lo cual se debe tener en cuenta ciertas técnicas que aseguren que todo saldrá bien, y sin filtración de datos importantes que puedan perjudicar a los clientes. A continuación, se muestra un diagrama de la interacción segura entre el cliente y la página en donde está comprando:

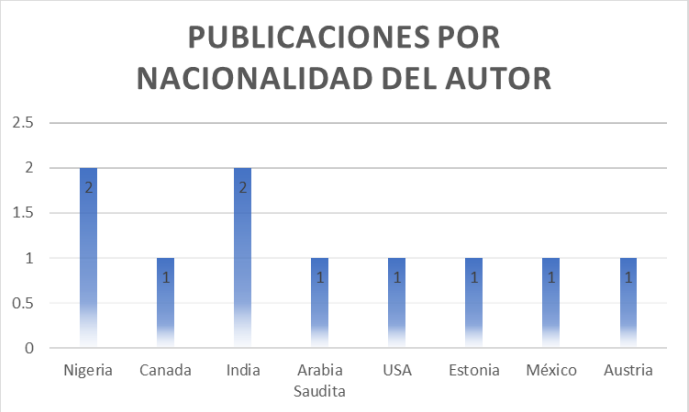


Figura 3. Publicaciones por nacionalidad del autor

Tabla 2. Artículos publicados por año

Año	Número de publicaciones	Porcentaje
2018	0	0%
2019	2	28,57%
2020	4	57,14%
2021	0	0%
2022	1	14,29%

La Figura 4 nos muestra que para que una transacción sea segura, se debe trabajar con un apartado para la seguridad de la información que se verá afectada en dicho proceso, y dividiéndolo en varias partes para abarcar cada punto posible de filtrado de datos y asegurándolos pasando por ciertos muros de protección y verificando que el usuario sea el que dice ser y todo se realice con integridad.

En la Figura 5, según [13], se muestra las principales empresas a nivel mundial que han optado por tener sus e-commerce y cuanto han llegado a costar a través de los años, estas empresas son de las pocas que han llegado a invertir mas en su seguridad, debido a que tienen millones de clientes y los riesgos a los que están expuestos son muy grandes, ya que los ciberdelincuentes suelen fijarse en empresas grandes para robar la información.

En la Figura 6, podemos apreciar una estimación de las ventas de estas empresas en corto plazo, mostrando el aumento que tendrá el mercado electrónico en un futuro no tan lejano y su gran

impacto a nivel mundial, debido a que estas deben contar con una gran protección de toda la información de sus clientes, tomando en cuenta la triada de la información.

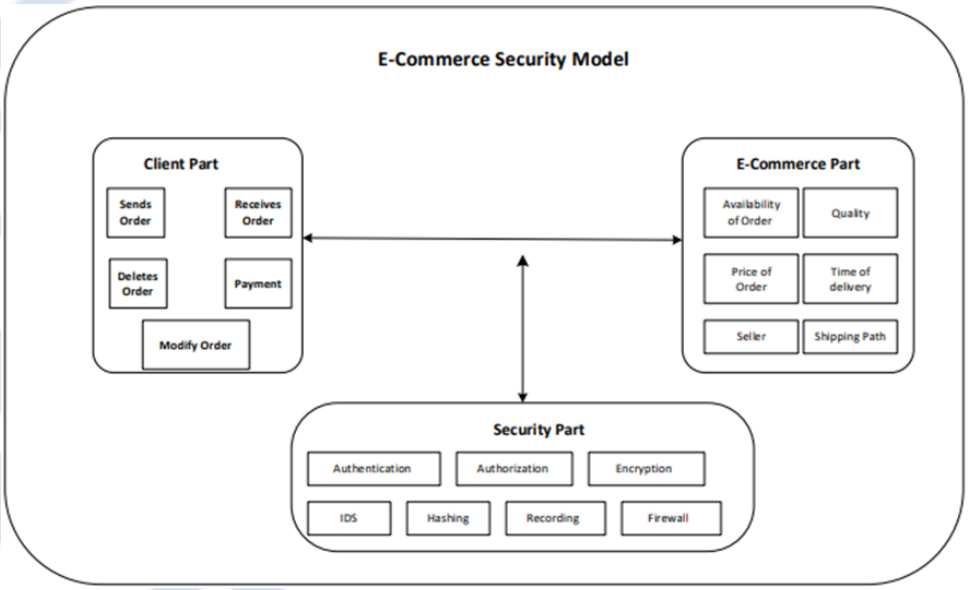


Figura 4. E-commerce security model for websites

Figura 5. Top e-commerce companies by market value

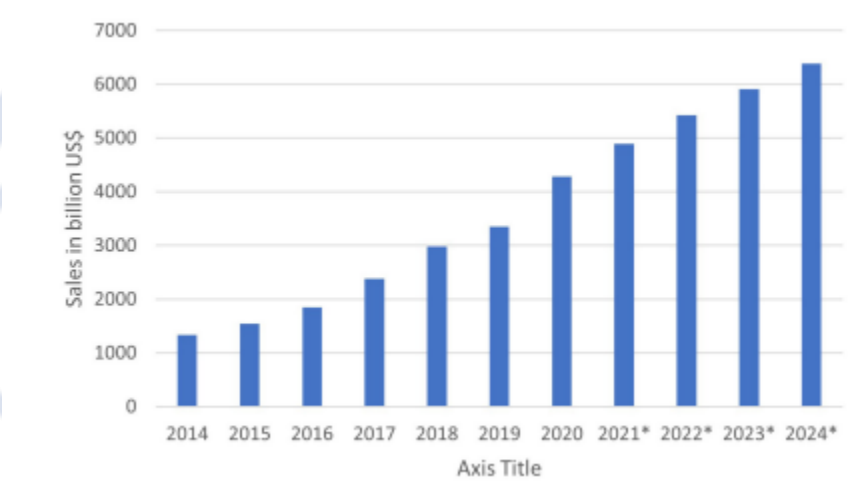


Figura 6. Top e-commerce companies by market value

En los últimos años, las empresas han optado por invertir más dinero en la seguridad de su información, debido al constante aumento de ataques cibernéticos que, por la alta competencia en el crecimiento de la tecnología y nuevas maneras de vulnerar sistemas de información, en este caso un comercio electrónico, han implementado estándares que les facilitan tener un control más eficiente, lo cual brinda mayor seguridad y confianza a sus clientes.

MAPA DE CALOR

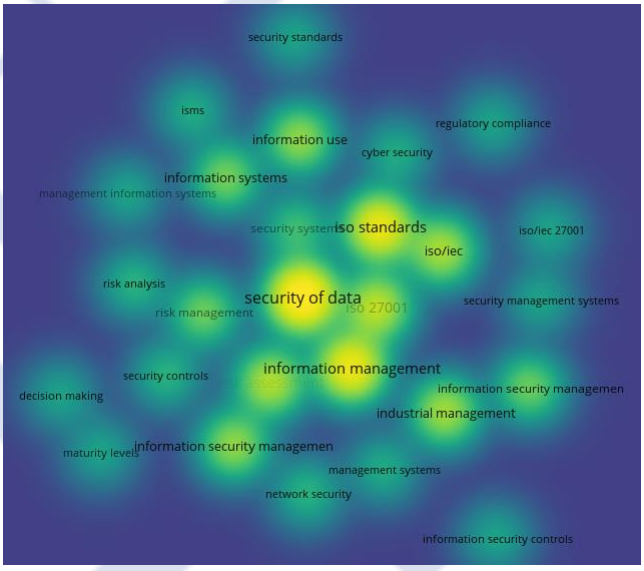


Figura 7. Palabras clave

En las búsquedas realizadas de revisiones sistemáticas previas y con la misma línea que nuestro tema de revisión, y gracias a la herramienta VosViewer se puede observar en la Figura 7 que las palabras más utilizadas en las búsquedas en los últimos 5 años son referentes a seguridad de datos y estándares ISO, entre las cuales resalta ISO 27001 ya que es una de las más aplicadas al momento de implementar un sistema de gestión de seguridad de la información, la cual permite a las organizaciones tener un estándar internacional y tener mayor confiabilidad en el comercio electrónico.

Concepto Comercio electrónico

El comercio electrónico ha transformado la industria del comercio tal como la conocemos, introduciendo mejores compras, envíos y servicios al cliente. Estos servicios comerciales generan y utilizan información confidencial, como compras de clientes, información financiera y personal, que son de gran valor para los atacantes [14].

El comercio electrónico (comercio electrónico) es la compra y venta de mercancías y empresas, o la transmisión de activos o información, a través de una red electrónica, esencialmente Internet. Estos intercambios comerciales ocurren como b a b (empresa a empresa), b a c (empresa a consumidor), c a c (consumidor a consumidor) o c a b (consumidor a empresa). es el intercambio de artículos o servicios que utilizan redes informáticas como Internet o comunidades informales en línea [15].

Los sistemas de comercio electrónico permiten a los clientes realizar compras en línea. Un proceso de pedido típico en los sistemas de comercio electrónico de empresa a cliente permite a los clientes buscar y encontrar artículos para comprar, negociar el precio de los artículos, agregar artículos a un carrito de compras, pagar artículos (es decir, comprar artículos) y pagar artículos. comprado; el sistema también permite a los comerciantes de comercio electrónico actualizar su inventario, verificar los métodos de pago de los clientes y planificar la logística para enviar artículos al cliente [16].

Concepto Gestión de la seguridad de la información

Proteger la información que se maneja en los sistemas de comercio electrónico exige una gestión de seguridad de la información y una gestión riesgos de seguridad consciente de las amenazas de seguridad en evolución [14].

La gestión de la seguridad de la información se puede describir como la implementación de la colección de protocolos o regulaciones que llevan a cabo todas las transacciones de información. Estos criterios de seguridad deben estar en condiciones de proteger la seguridad de la información de diferentes empresas contra una serie de peligros innegables [17].

A pesar de la cantidad de modelos de seguridad para prevenir delitos relacionados con el comercio electrónico que se han propuesto en el pasado, no muchos de ellos son practicables o implementables. La autenticación de usuario proporciona la garantía de que los detalles del cliente están protegidos y se mantiene la privacidad. El sitio web del comerciante brinda una gran seguridad, asegurando así a los clientes que la transacción se lleva a cabo sin un ápice de duda o temor a la inseguridad y también se mantiene la integridad, la privacidad y la confidencialidad [18].

La gestión de la seguridad de la información en el comercio electrónico es una fracción de la estructura de seguridad de la información. Básicamente se considera el uso de componentes que inciden en el comercio electrónico. Incluye la protección informática, la seguridad de datos y otros ámbitos más amplios relacionados con la estructura de seguridad de la información. La seguridad del comercio electrónico también se conoce como la protección de los activos de comercio electrónico contra los piratas informáticos [19].

En la literatura revisada se puede apreciar que los autores trabajan con las dimensiones de la seguridad de la información, que incluyen a la triada de la información, que son la confidencialidad, integridad, disponibilidad además de a la autenticación y el no repudio, mismas que son mencionadas, como los ejes de la gestión de la seguridad de la información para su comprensión y gestión de sus riesgos; además de haber utilizado la norma ISO 27001 la cual ayudo a los autores a implementar un sistema de gestión de seguridad de la información para las empresas en las cuales se enfocó el trabajo.

Concepto de la norma ISO 27001

La norma fue diseñada y publicada conjuntamente por la Organización Internacional de Normalización (ISO) y la Comisión Electrotécnica Internacional (IEC) en 2005. Fue planteada como una evolución de BS 7799. La norma 27001 especifica con detalle los requisitos para establecer, implementar, mantener y mejorar continuamente un sistema de gestión de la seguridad de la información (SGSI) dentro del contexto interno y externo de la organización. Y algunos de los requisitos que dicta la norma son genéricos y están destinados a ser aplicables a todas las organizaciones, independientemente de su tipo, tamaño o naturaleza [20].

ISO/IEC 27001 es el único estándar de toda la familia de normas ISO 27000 que se utiliza para brindar certificaciones a las organizaciones. Las demás normas de esta familia se usan para brindar un apoyo robusto y profundo para que la organización pueda construir e implementar un Sistema de Gestión de la Seguridad de la Información de manera correcta y duradera [21].

Discusión

La presente revisión sistemática muestra la realidad actual de la gestión de la seguridad de la información dentro de los sistemas de comercio electrónico, a pesar de la cantidad de modelos de seguridad para prevenir delitos relacionados con el comercio electrónico que se han propuesto en el pasado, no muchos de ellos son practicables o implementables, teniendo en cuenta los múltiples aspectos en los cuales la información puede verse afectada, cumpliendo así las buenas prácticas de la norma ISO 27001 para implementarla correctamente en las empresas de compra y venta en línea.

Una gran ventaja fue el tiempo de las publicaciones para la revisión, siendo el comercio electrónico un área relativamente nueva, el margen fueron 5 años, pero en años de comercio electrónico fue como abarcar todo lo existente hasta ahora. Además de claro empezar a enlazar nuevos términos a este campo para futuras investigaciones, como la utilización de Block Chain para el desarrollo de nuevos sistemas de gestión en el ámbito de la seguridad de la información.

En los artículos encontrados en nuestra investigación tenemos que con las tecnologías actuales se puede complementar de muy buena manera la implementación de medidas de seguridad tal y como dictan las normas ISO 27000. Según nuestra bibliografía la más utilizada viene a ser la ISO 27001 ya que es la única que tiene una certificación que garantiza que la organización es segura. Los clientes, aunque no sepan exactamente lo que dicta la norma, pueden sentirse en confianza al comprar en empresas de comercio electrónico que tengan una certificación en seguridad de la información.

Conclusiones

En los últimos años, las empresas que han conseguido un espacio en la web han crecido considerablemente y se han hecho muy conocidas alrededor del mundo, lo cual ha incrementado el número de sus clientes, y esto ha dado pase a que personas que no tienen que ver con las empresas o clientes y quieren robar información de ellos, para fines de enriquecerse ilegalmente. Este es un problema que aqueja a todas las empresas, ya que ninguna es 100% segura y aunque las brechas que tienen no son siempre las mismas, cabe resaltar que todo el tiempo estarán expuestas a ciberdelincuentes.

La presente investigación indicó la gran importancia de contar con un sistema gestión de la seguridad de la información y que por esto se debe invertir mucho más tiempo y dinero para conseguir resguardar la información del peligro al que suele estar expuesto, siendo los principales problemas que enfrentan tanto los proveedores como los consumidores las transacciones, la privacidad, la seguridad del sistema en el que se desarrollan estas y sus vulnerabilidades. Se puede además apreciar que uno de los aspectos que más aprovechan los ciberdelincuentes es el

robo de sesión y muchos métodos más de ataques, lo cual perjudica a muchas empresas, ya sean pequeñas o grandes, y muchas de ellas no cuentan con políticas y controles bien definidos para un correcto manejo de la información. Una manera de proteger los comercios electrónicos seria implementar ISO 27001 que nos brinda una serie de pasos y requerimientos que deben tener para poder tener un alto nivel de seguridad, ya sea de la empresa y sus clientes. Esto lo hace mediante la creación e implementación de un SGSI para que la organización pueda tener el control y la confianza de que sus activos más valiosos como pueden ser: servidores, información, recursos humanos, etc., estén resguardados ante cualquier riesgo de amenaza latente.

Esto ayudará a los investigadores y académicos que trabajan actualmente en este campo a tener una idea de las tendencias y el estado actual para futuras investigaciones sobre el comercio electrónico, para que puedan tener una vista global de cómo existe la necesidad de tener un sistema de gestión de seguridad de la información en las empresas.

Refrencias

[1] G. S. Torre y D. G. Codner, Fundamentos de Comercio, Buenos Aires: Universidad Virtual de Quilmes, 2013.

[2] E. Turban, L. Volonino y G. R. Wood, Information Technology for Management, Nueva York: Wiley, 2010.

[3] S. Carrasco Fernández, Venta online, Madrid: Ediciones Paraninfo, 2014.

[4] K. Laudon y C. Traver, E Commerce: Business, Technology, Society, Nueva York: Pearson Prentice Hall, 2009.

[5] A. Martínez Nadal, Comercio electrónico, firma digital y autoridades de certificación, Madrid: Aranzadi, 2001.

[6] R. Mateu De Ros, El consentimiento y el proceso de contratación electrónica, Pamplona: Aranzadi, 2000.

[7] C. M. Fernández, «La norma ISO 27001 del Sistema de Gestión de la Seguridad de la Información,» *Calidad,* pp. 40-44, 2012.

[8] F. Pacheco, «Welivesecurity,» 10 Septiembre 2010. [En línea]. Available: https://www.welivesecurity.com/la-es/2010/09/10/la-importancia-de-un-sgsi/.

[9] M. Podrecca, G. Culot, G. Nassimbeni y M. Sartor, «Information security and value creation: The performance implications of ISO/IEC 27001,» *Computers in Industry,* vol. 142, pp. 2-10, 2022.

[10] D. Freitas, «Análisis y evaluación del riesgo de la información: caso de estudio Universidad Simón Bolivar,» *Enlace,* vol. 6, nº 1, p. 13, 2009.

[11] A. Pérez, «OBS Business School,» 09 Octubre 2017. [En línea]. Available: https://www.obsbusiness.school/blog/seguridad-de-la-informacion-un-conocimiento-imprescindible.

[12] H. A. García-Perdomo, «Conceptos fundamentales de las revisiones sistemáticas/metaanálisis,» *Urología Colombiana,* pp. 28-34, 2015.

[13] L. Xiang, F. A. Sayed, K. A. Muhammad, K. Jingying, I. Muhammad, U.-H. Jabbar y A. Shujaat, «Cyber security threats: A never-ending challenge for e-commerce,» *Frontiers in Psychology,* vol. 13, nº 2, p. 7, 2022.

[14] A. Nolte, A. Abasi-amefon y M. Raimundas, «Security Risk Management in E-commerce Systems: A Threat-driven Approach,» *Modern Computing,* vol. 8, nº 2, p. 28, 2020.

[15] S. Khan, «Cyber Security Issues and Chanllenges in E-Commerce,» *SSRN,* vol. 10, nº 5, p. 8, 2019.

[16] E. Sylvanus y O. Adepele, «A Formal Model of Distributed Security for Electronic Commerce Transactions Systems,» *International Journal of Networked and Distributed Computing,* vol. 7, nº 2, p. 17, 2019.

[17] A. Ibrahim, «Integrated e-commerce security model for websites,» *International Journal of Advanced and Applied Sciences,* vol. 9, nº 4, p. 8, 2022.

[18] A. O. Raphael, A. O. Sulaiman y O. M. Babatola, «A SECURITY MODEL FOR PREVENTING E-COMMERCE RELATED CRIMES,» *Applied Computer Science,* vol. 16, nº 3, p. 12, 2020.

[19] K. Shweta y G. Charu, «Ensure Hierarchal Identity Based Data Security in Cloud Environment,» *International Journal of Cloud Applications and Computing,* vol. 9, nº 4, p. 16, 2019.

[20] G. Culot, «The ISO/IEC 27001 information security management standard: literature review and theory-based research agenda,» *The TQM Journal,* 2021.

[21] X. Zhu y Y. Zhu, «Extension of ISO/IEC27001 to Mobile Devices Security Management,» *Communications in Computer and Information Science,* 2019.


LaSalle
Universidad - PERÚ