VOL 2 Nº 2 Septiembre - Febrero 2021 ISSN Nº 2708-0935

Innovación y Software

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

ARK: ark:/42411/s6
PURL: 42411/s6

Vol. 2 N°. 2 2021 Septiembre - Febrero

ISSN N°: 2708-0935

DOI: 10.48168/innosoft.s6

ARK: ark:/42411/s6

PURL: 42411/s6

Depósito Legal: 2023-08884

Periodicidad: Semestral

Publicado: 30/09/2021

Editado por:

Universidad La Salle

RUC: 20456344004

Av. Alfonso Ugarte N° 517, Cercado, Arequipa

COMITÉ EDITORIAL

EDITORIAL

Las investigaciones y el desarrollo tecnológico en la Ingeniería de Software y las Ciencias de la Computación

p. 4 - 5

ARTÍCULOS CORTOS

Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica

Autores: Kevin Rivera Vergaray

p. 6 – 13

ARTÍCULOS ORIGINALES

Predicción de mortalidad a causa del Covid 19 en Perú utilizando redes neuronales artificiales

Autores: Cesar Mayta Avalos, Jesús Cristian Valdivia Mamani, Fernando Rosales Castilla, Milca Gines Colana

p. 14 – 26

Propuesta de un plan de seguridad de la información para incrementar la fiabilidad de datos en una financiera

Autores: Wilmer Aufredy Apaza Chávez

p. 27 - 43

Aplicación de regresión logística para la predicción de demanda por especialidad médica en consulta externa hospitalaria

Autores: Rene Aquino Arcata, Ronald Cuevas Machaca, Luis Godoy Montoya, Heber Rodríguez Puma

p. 44 - 59

Predicción de hipertensión arterial a través de un sistema de regresión logística

Autores: Cynthia Mayumi Tesillo Gómez, Yuri Alexander Escobar Arcaya, Edwin Daniel León Gutiérrez

p. 60 - 74

Sistema para proponer la nota final de los estudiantes mediante Redes Neuronales

Autores: Kleber Ernesto Baldarrago Salas, Erika Cayllahua Chicaña, Fanny Lorena Lorenzo Quilla, Maria Quijia Álvarez

p. 75 - 91

Selección de una red social para apoyar la docencia universitaria empleando computación con palabras

Autores: Dargel Veloz Morales, Laritza González Marrero

p. 92 - 105

La Revista Innovación y Software de la Facultad de Ingeniería, en la Universidad La Salle, se complace en presentar este segundo número de su segundo 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.

En el futuro, debemos prestar atención a la importancia de la capacitación y la educación en diferentes campos, en lugar de descuidar el apoyo a la ingeniería de software y la informática que respalda las encuestas de modelos.

En la actual pandemia mundial, los factores clave como el talento y la tecnología respaldados por actividades de investigación destinadas a encontrar soluciones adecuadas a estos problemas son cada vez más evidentes. Despierta la sabiduría de las personas y les permite encontrar soluciones viables a los problemas que respaldan el uso de modelos, tecnologías y estándares de métodos de informática e ingeniería de software.

Se puede observar que estas tecnologías han traído cambios revolucionarios al mundo en diversos campos del desarrollo como la sociedad, la economía, la política y la educación, por lo que la importancia de la educación en este campo aumenta día a día. Este número se centra en el avance de la investigación y el desarrollo tecnológico de investigadores, académicos y estudiantes en diferentes campos; propone el avance de la resolución de problemas prácticos en base a diferentes principios.

El primer artículo se desarrolla un modelo e inteligencia artificial que permite detectar estudiantes universitarios con un alto riesgo de deserción académica. Se comparan varias técnicas de inteligencia artificial como la regresión logística, los árboles de decisión, las redes neuronales y el agrupamiento por medio del algoritmo el vecino más cercano. La aplicabilidad del modelo presentado permite a las universidades tomar acciones y decisiones centradas en disminuir en alto grado de deserción que tienen los estudiantes.

En el segundo artículo se desarrolla un modelo de redes neuronales artificiales con el fin de predecir la cantidad de fallecidos durante la pandemia del Covid 19. Este modelo de inteligencia artificial utiliza los datos públicos del Ministerio de Salud del Perú para predecir la cantidad de fallecidos que van a presentarse en los próximos meses. La efectividad de dicho modelo se puede apreciar con un error cuadrático medio bastante bajo así como una exactitud del modelo alta superando el 90%.

El tercer artículo muestra una propuesta de plan de seguridad de la información para mejorar la fiabilidad de los datos en las entidades financieras. Se aplican conceptos analizados en las normas ISO 27001 y 27002. Se tienen en cuenta la disponibilidad de la información, así como la confidencialidad y la integridad de la misma. Estos factores son de vital importancia en las

entidades financieras donde se manejan datos sensibles y delicados por lo que deben de preservarse de la mejor manera.

El cuarto artículo se enfoca la predicción mediante un modelo de regresión logística para determinar la cantidad de pacientes que solicitarán consultas externas hospitalarias. Durante el Covid 19 la gestión hospitalaria se ha convertido en una tarea complicada porque han aumentado la cantidad de pacientes que necesitan ser atendidos. La predicción de la cantidad de pacientes que van a asistir a un servicio de consulta externa puede ayudar a planificar mejor los recursos del hospital optimizando así la atención hospitalaria.

En el artículo quinto se muestran cómo con un sistema de regresión logística puede ser usado para predecir si una persona tiene probabilidades de desarrollar hipertensión arterial. Se utiliza un conjunto de datos de pacientes hipertensos del Perú donde se exponen un conjunto de factores que pueden influir en el desarrollo de esta enfermedad. Con un 87% de exactitud y un total de 5615 registro de pacientes el sistema trabaja con más de diez variables para cumplir con la predicción, haciendo uso de la técnica de inteligencia artificial de regresión logística.

El sexto artículo propone una herramienta basada en inteligencia artificial de apoyo a los docentes. Dicha herramienta es capaz de proponer la nota final de una asignatura en función de un conjunto de variables analizadas. Utiliza una red neuronal con alto grado de precisión que le permite a los docentes entender que variables influyen en el desempeño académico de sus estudiantes.

Finalmente, el séptimos artículo propone un marco metodológico para seleccionar una red social a ser usada en la docencia universitaria. Este artículo muestra como se puede utilizar la metodología de computación con palabras para tomar una decisión. Los problemas de toma de decisiones que tienen varias alternativas, con varios criterios y múltiples expertos suelen convertirse en tareas engorrosas cuando se hacen manualmente. Existen software que pueden ser usados para esta tarea y en el presente artículo se ilustra como se usa uno de ellos.

Comité Editorial

Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica

Predictive model for the early detection of students with high risk of academic dropout

Kevin Rivera Vergaray
Universidad Nacional Mayor de San Marcos
kevin.rivera1@unmsm.edu.pe
https://orcid.org/0000-0001-5393-4382

ARK: ark:/42411/s6/a40
PURL: 42411/s6/a40

RECIBIDO 05/04/2021 • ACEPTADO 08/06/2021 • PUBLICADO 30/09/2021

RESUMEN

Se comparan los resultados de 4 modelos predictivos, de regresión logística, árboles de decisión, KNN y una red neuronal para predecir la deserción académica de estudiantes en la Universidad Nacional Intercultural de la Amazonía, aplicado a un *dataset* extraído de la base de datos del sistema de gestión académica de la universidad, que contiene datos socioeconómicos y de rendimiento académico los cuales fueron procesados y formateados utilizando técnicas de *one hot encoding* para así poder aplicar los modelos predictivos ya mencionados. Para el procesamiento y formateo de datos se utilizó consultas Transac Sql y la aplicación de los modelos predictivos se hizo a través del Software Knime y utilizando Python a través de Google Colab. Los resultados obtenidos al aplicar 4 modelos predictivos son muy buenos ya que todos superaron el 80% de *accuracy*, lo cual garantiza que puedan ser puestos en producción para el beneficio de la universidad y así pueda tomar mejores decisiones a la hora de abordar la deserción académica. Se concluye que aplicar un modelo predictivo en las universidades para la detección temprana de estudiantes con alto riesgo de deserción académica es viable y muy beneficioso para que las universidades a través de sus gestores académicos puedan aplicar estrategias más focalizadas para reducir sus índices de deserción académica.

Palabras claves: Deserción académica, Modelo predictivo, Dataset.

ABSTRACT

The results of 4 predictive models, logistic regression, decision trees, KNN and a neural network are compared to predict the academic dropout of students at the National Intercultural University of the Amazon, applied to a dataset extracted from the system's database. of academic management of the university, which contains socioeconomic and academic performance data which were processed and formatted using one hot encoding techniques in order to apply the predictive models already mentioned. For data processing and formatting, Transac Sql queries were used and the application of predictive models was done through Knime Software and using Python through Google Colab. The results obtained by applying 4 predictive models are very good since they all exceeded 80% of Accuracy, which guarantees that they can be put into production for the benefit of the university and thus can make better decisions when addressing academic dropout. It is concluded that applying a predictive model in universities for the early detection of students with high risk of academic dropout is viable and very beneficial so that universities, through their academic managers, can apply more focused strategies to reduce their academic dropout rates.

Keywords: *Academic dropout, Dataset, Predictive model.*

INTRODUCCIÓN

La deserción estudiantil universitaria no es un problema nuevo ni exclusivo del Perú. Este fenómeno se da en todo el mundo, es un viejo problema que tiene muchas variables y el cual no es preocupación exclusiva del mundo académico. La deserción estudiantil universitaria trae como consecuencia el aumento del número de alumnos con educación superior incompleta que se incorporan al mundo laboral y se convierten en sub empleados sin obtener los ingresos deseados; lo cual, perjudica al mismo estudiante, a sus familiares, al país y a la universidad pues esta ve afectado su presupuesto [1].

No importa el tipo de universidad o casa de estudio. Hay factores muy comunes que disminuyen las tasas de retención estudiantil en la educación superior. Pueden ser problemas individuales o una mezcla de factores. Por eso, las facultades deben trabajarlas de manera adecuada para reducir la deserción [2].

Actualmente el 40% de estudiantes de la UNIA son indígenas y el otro 60% de los estudiantes son mestizos (en su gran mayoría viven en el casco urbano) No se cuenta con información procesada de las deserciones por semestre académico. Se percibe la mayor cantidad de deserciones en las carreras de ingeniería, las principales causas, la dificultad de la carrera y la situación económica del estudiante. Es así que se pretende aplicar conceptos de *machine learning* y análisis de datos con la finalidad de predecir la deserción académica y así tomar las previsiones

necesarias para reducirla. En el presente artículo se utilizó la base de datos del Sistema de Gestión Académica de la UNIA.

Materiales y métodos o Metodología computacional

En el presente trabajo de probaron con 4 modelos predictivos de *machine learning* para poder predecir el riesgo de deserción académica.

Deberá entenderse por deserción estudiantil o deserción académica, el abandono definitivo de las aulas de clase por diferentes razones y la no continuidad en la formación académica, que la sociedad quiere y desea en y para cada persona que inicia sus estudios universitarios [3].

- **Regresión logística:** La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables "*dummy*", es decir variables simuladas. El propósito del análisis consiste en: predecir la probabilidad de que a alguien le ocurra cierto "evento": por ejemplo, estar desempleado =1 o no estarlo = 0, ser pobre = 1 o no pobre = 0, recibirse de sociólogo =1 o no recibirse = 0). Determinar que variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso que cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos [4].

- **Arboles de decisión:** Los árboles de decisión proveen de una herramienta de clasificación muy potente. Su uso en el manejo de datos la hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos sus resultados por cualquier usuario. El árbol en sí mismo, al ser obtenidos, determinan una regla de decisión. Esta técnica permite:

1. **Segmentación**: establecer que grupos son importantes para clasificar un cierto ítem.
2. **Clasificación**: asignar ítems a uno de los grupos en que está particionada una población.
3. **Predicción**: establecer reglas para hacer predicciones de ciertos eventos.

4. **Reducción de la dimensión de los datos**: Identificar qué datos son los importantes para hacer modelos de un fenómeno.

5. **Identificación-interrelación**: identificar que variables y relaciones son importantes para ciertos grupos identificados a partir de analizar los datos.

6. **Recodificación**: discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante [5].

- **KNN:** *K-Nearest-Neighbor* es un algoritmo basado en instancia de tipo supervisado de *Machine Learning*. Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos). Al ser un método sencillo, es ideal para introducirse en el mundo del Aprendizaje Automático. Sirve esencialmente para clasificar valores buscando los puntos de datos "más similares" (por cercanía) aprendidos en la etapa de entrenamiento y haciendo conjeturas de nuevos puntos basado en esa clasificación [6].

- **Red Neuronal:** Las redes neuronales Artificiales (RNAs) son modelos computacionales como un intento de conseguir formalizaciones matemáticas acerca de la estructura del cerebro. Las RNAs imitan la estructura *hadware* del sistema nervioso, centrándose Enel funcionamiento el cerebro humano, basado en el aprendizaje a través de la experiencia, con la consiguiente extracción de conocimiento a partir de la misma [7].

Para aplicar los modelos mencionados se utilizó un *dataset*, extraído a través de consultas SQL de la base de datos del sistema académico de la Universidad Nacional Intercultural de la Amazonía, dicho *dataset* este compuesto con datos registrados desde el 2005, en total se generaron 17 variables junto con el target "desercion". La tabla de datos contiene 18 columnas y 5803 filas, con un peso de 680.2 kb.

Tabla 1: Contenido de la data académica procesada

N°	Variable	Descripción
1	codigo	Código que identifica al estudiante.
2	sexo	Sexo del estudiante, 1 es masculino y 0 femenino.
3	mestizo	Si el estudiante es mestizo 1 y si no lo es 0.
4	indigena	Si el estudiante es indigena 1 y si no lo es 0.
5	pobre	Si el estudiante es pobre 1 y si no lo es 0.
6	pobre_extremo	Si el estudiante es pobre extremo 1 y si no lo es 0.
7	no_pobre	Si el estudiante es No pobre 1 y si no lo es 0.
11	educacion	Si el estudiante es de la facultad de educación 1 y si no lo es 0.
12	ingenieria	Si el estudiante es de la facultad de ingeniería 1 y si no lo es 1.
13	matriculas	Número de matriculas que tiene el estudiante durante su estadía en la universidad.
14	matriculas_aprobadas	Número de matriculas aprobadas que tiene el estudiante durante su estadía en la universidad.
15	matriculas_desaprobadas	Número de matriculas desaprobadas que tiene el estudiante durante su estadía en la universidad.
16	egresado	Si el estudiante es egresado 1 y si aun no lo es 0.
17	ponderado	Promedio ponderado acumulado del estudiante
18	semestres	Numero de semestres que se ha ,atriculado el estudiante.
19	desercion	1 si el estudiante deserto 0 si el estudiante no deserto.

Fuente: Elaboración propia.

Para aplicar los modelos se utilizó el software KNIME y a través de Python usando Google Colab.

Resultados y discusión

Luego de aplicar y evaluar los modelos aplicados al conjunto de datos extraídos de la base de datos del sistema académico, los mejores resultados se obtuvieron con el KNN y el árbol de decisión. En el modelo KNN se obtuvo un *accuracy* de 88,844% y un error de 11,156%, y en el modelo de árbol de decisión se obtuvo un *accuracy* de 88,4% y un Error de 11,6%.

A continuación, se detallan los resultados obtenidos por cada modelo aplicado.

a) **Regresión logística:** Al aplicar regresión logística se obtuvo un *accuracy* de 84,57% y un error de 15,43%.

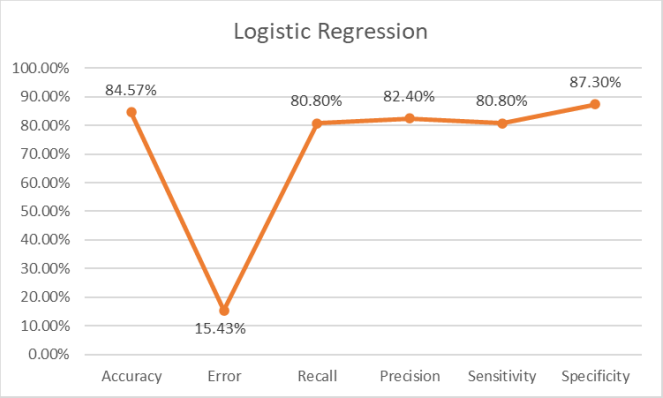

Figura 1: Resultados obtenidos aplicando regresión logística.

b) **Árbol de decisión:** Al aplicar arboles de decisión se obtuvo un *accuracy* de 88,4% y un Error de 11,6%.

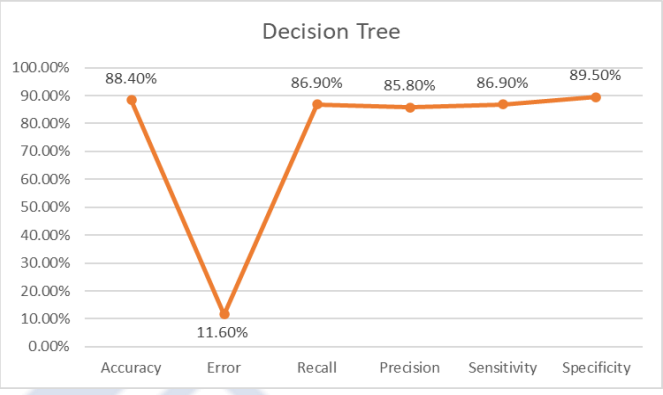

Figura 2: Resultados obtenidos aplicando arboles de decisión.

c) KNN: Al aplicar KNN se obtuvo un *accuracy* de 88,4% y un Error de 11,6%.

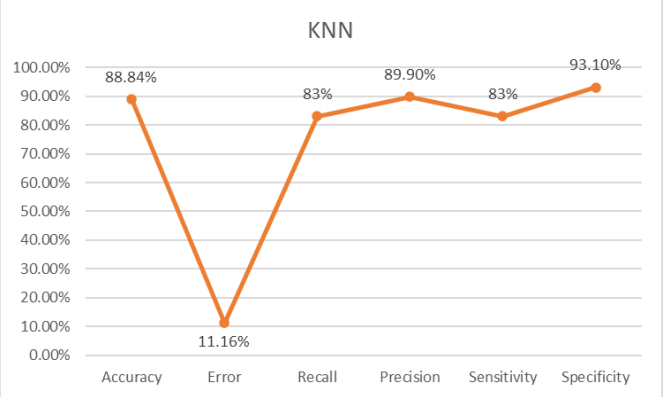

Figura 3: Resultados obtenidos aplicando KNN.

d) Red neuronal: Al aplicar la Red Neuronal se obtuvo un *accuracy* de 82,33% y un Error de 17,67%.

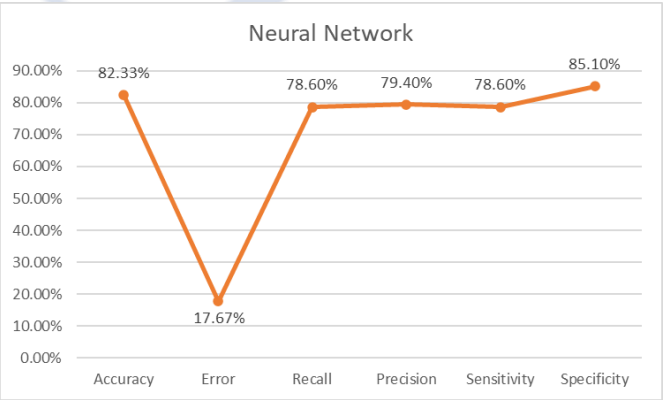

Figura 4: Resultados obtenidos aplicando una red neuronal.

Se puede ver que en los 4 modelos predictivos aplicados se obtuvieron un *accuracy* mayor al 80%, lo cual en gran parte es por el trabajo previo que se realizó para preparar la data utilizada,

en el cual se aplicaron técnicas de *one hot encoding* entre otros. Finalmente presentamos una comparación de los resultados obtenidos con los 4 modelos predictivos utilizados. Tabla 2: Cuadro comparativo de los resultados obtenidos.

Tabla 2: Cuadro comparativo de los resultados obtenidos.

Medida	Logistic Regression	Decision Tree	KNN	Neural Network
Accuracy	84.57%	88.40%	88.84%	82.33%
Error	15.43%	11.60%	11.16%	17.67%
Recall	80.80%	86.90%	83%	78.60%
Precision	82.40%	85.80%	89.90%	79.40%
Sensitivity	80.80%	86.90%	83%	78.60%
Specificity	87.30%	89.50%	93.10%	85.10%

Fuente: Elaboración propia.

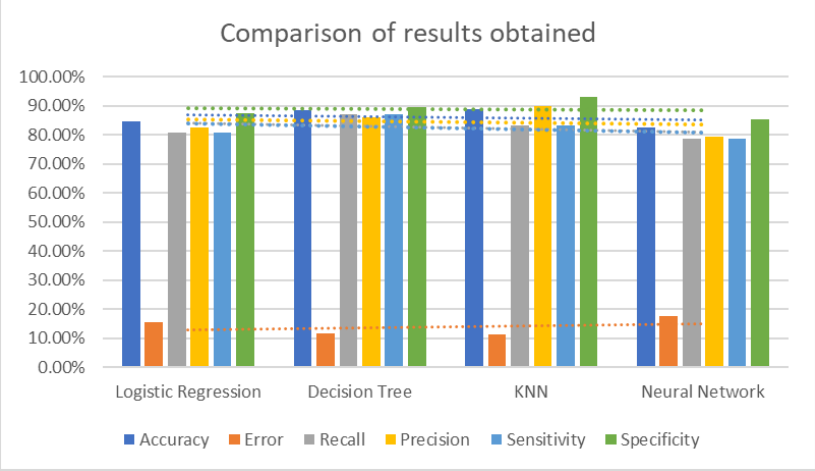

Figura 5: Comparación de resultados obtenidos.

CONCLUSIONES

Se determina que el modelo de KNN y Arboles de decisión son los que describen un mayor ajuste de los datos analizados, estos modelos presentan un *accuracy* bastante aceptables de 88.844 % y 88.4%.

Los modelos predictivos aplicados a datos académicos, como en nuestro caso pueden ayudar en la generación de mejores estrategias para los problemas que aquejan a las universidades, sobre todo públicas, como es la deserción académica.

El modelo debe analizar más datos referidos a la enseñanza no presencial implementada en las universidades a causa de la pandemia por el COVID-19, lo que posiblemente influya en los resultados obtenidos con la data utilizada.

AGRADECIMIENTOS

Mi agradecimiento al curso de Tópicos Avanzados en Ingeniería de Software de la Maestría en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software en UNMSM. Mi mayor agradecimiento a mi familia.

REFERENCIAS

[1] T. Viale, «Enfoque UPC,» 10 enero 2020. [En línea]. Available: https://enfoque.upc.edu.pe/mas-temas/educacion/desercion-estudiantil-universitaria-accionamos-o-reaccionamos/.

[2] uPlanner, «uPlanner,» 27 Marzo 2017. [En línea]. Available: https://uplanner.com/es/blog/8-causas-de-desercion-estudiantil-en-la-educacion-superior/.

[3] Gabriel, Jaime, Páramo, Arturo y Correa, Deserción Estudiantil Universitaria. Conceptualización, Revista Universidad Eafit, 1999.

[4] H. Chitarroni, La regresión logística, Buenos Aires: Instituto de Investigación en Ciencias Sociales, 2002.

[5] C. N. Bouza y A. Santiago, MODELACIÓN MATEMÁTICA DE FENÓMENOS DEL MEDIO AMBIENTE Y LA SALUD, Mexico: Universidad Autónoma de Guerrero, 2012.

[6] aprendemachinelearning, «Clasificar con K-Nearest-Neighbor ejemplo en Python,» July 2018. [En línea]. Available: https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/.

[7] R. F. López y J. M. F. Fernández, Las Redes Neuronales Artificiales - Fundamentos teoricos y aplicaciones prácticas., España: lorena bello, 2008.

Predicción de mortalidad a causa del Covid 19 en Perú utilizando redes neuronales artificiales

Prediction of mortality due to Covid 19 in Peru using artificial neural networks

Cesar Mayta Avalos
Universidad Jorge Basadre Grohman
cesar.mayta@unjbg.edu.pe
https://orcid.org/0000-0002-5722-1854

Jesús Cristian Valdivia Mamani
Universidad Jorge Basadre Grohman
jesus.valdivia@unjbg.edu.pe
https://orcid.org/0000-0001-9566-6988

Fernando Rosales Castilla
Universidad Jorge Basadre Grohman
fernando.rosales@unjbg.edu.pe
https://orcid.org/0000-0003-0668-2885

Milca Gines Colana
Universidad Jorge Basadre Grohman
milca.gines@unjbg.edu.pe
https://orcid.org/0000-0002-3596-2803

ARK: ark:/42411/s6/a43
PURL: 42411/s6/a43

RECIBIDO 18/04/2021 • ACEPTADO 30/05/2021 • PUBLICADO 30/09/2021

RESUMEN

Con el desarrollo de la pandemia en Perú, la cantidad de fallecidos ha ido en aumento y lamentablemente no se han tomado las medidas adecuadas, esto por no tener una herramienta que nos permita saber la cantidad de fallecidos posibles en un tiempo determinado. El objetivo del presente artículo es proponer una herramienta capaz de predecir la cantidad de fallecidos por COVID-19 en función del tiempo. La metodología utilizada fue redes neuronales artificiales utilizando series temporales con información obtenida del Ministerio de Salud del estado peruano a través de su portal de datos abiertos. Los resultados alcanzados tuvieron un error cuadrático medio de 0.0037 y pérdida de 0.0480. Los resultados obtenidos a lo largo del artículo confirman la validez de esta herramienta y la efectividad en la predicción de la cantidad de fallecidos a causa del COVID 19.

Palabras claves: Inteligencia Artificial, Series Temporales, COVID 19, Predicción, Pronóstico.

ABSTRACT

With the development of the pandemic in Peru, the number of deaths has been increasing and unfortunately the appropriate measures have not been taken, this because we do not have a tool that allows us to know the number of possible deaths in a given time. The objective of this article is to propose a tool capable of predicting the number of deaths from COVID-19 as a function of time. The methodology used was artificial neural networks using time series with information obtained from the Ministry of Health of the Peruvian state through its open data portal. The results achieved had a mean square error of 0.0037 and a loss of 0.0480. The results obtained throughout the article confirm the validity of this tool and its effectiveness in predicting the number of deaths from COVID 19.

Keywords: *Artificial Intelligence, Time Series, COVID 19, Prediction, Forecast.*

INTRODUCCIÓN

Hoy en día, la prioridad de todas las naciones, es lograr el máximo nivel de salud, tanto en el aspecto físico, mental y social. A nivel mundial se toman decisiones a favor del pueblo, con el fin de lograr este objetivo. Durante muchos años hemos sido afectados por diversos tipos de enfermedades, de los cuales hemos salido airosos a pesar de las pérdidas. Esto conlleva a preguntarnos: ¿Estaremos preparados para una pandemia de grandes magnitudes? ¿Nuestro sistema de salud llegaría a colapsar?

En diciembre de 2019, se detectó por primera vez un brote causado por el nuevo coronavirus humano del síndrome respiratorio agudo grave de tipo 2 (SARS-CoV-2), en la ciudad de Wuhan, provincia de Hubei, en China [1]. Desde entonces se ha propagado por todo el mundo produciendo una severa crisis económica, social y de salud, nunca antes vista. Debido a su rápida extensión, la Organización Mundial de la Salud declaró la pandemia por la enfermedad de coronavirus del 2019 (COVID-19) el 11 de marzo del 2020. El COVID-19 hizo su aparición en Perú un poco más tarde en comparación con otros países. Al primer caso confirmado, se tomaron medidas como la instauración de un estado de emergencia sanitaria, inmovilización total obligatoria, medidas de higiene, cierre de fronteras y de aeropuertos. A pesar de todas las medidas tomadas por el estado peruano, se tuvo la tasa de mortalidad más alta a nivel mundial durante la primera ola de contagios.

Con el desarrollo de la pandemia en Perú, la cantidad de fallecidos ha ido en aumento, según el registro de fallecimientos del Ministerio de Salud (MINSA) [2], las últimas cifras han sido alarmantes alcanzando un total de fallecidos de 193230 hasta el 03 de julio del 2021. Como estado, nuestro sistema de salud falló, nuestras entidades hospitalarias no se dieron abasto, incluso instituciones funerarias colapsaron, nuestro sistema educativo no está preparado para adversidades, nuestro sistema económico es ineficaz. ¿Cuántas personas más tendrán que morir para tener un buen sistema, para así prever las muertes de más peruanos?

Lamentablemente en nuestro país se tiene el pensamiento arcaico de "ver para creer". Es por eso que teniendo cifras adecuadas, que nos permitan saber la cantidad de fallecidos posibles en un tiempo determinado, se podrían plantear nuevas estrategias. Para así evitar la aglomeración de pacientes y la desesperación de un pueblo que no solo se queja por la enfermedad, sino también por la falta de dinero, la crisis que produjo esta pandemia. Actualmente no se tiene un modelo predictivo que permita conocer el número de fallecidos diarios por COVID 19 en el Perú, al no tener este modelo no se podría estimar si la tasa de mortalidad incrementará en el tiempo.

La inteligencia artificial es una nueva forma de resolver problemas dentro de los cuales se incluyen los sistemas expertos, el manejo y control de robots y los procesadores, que intenta integrar el conocimiento en tales sistemas, en otras palabras, un sistema inteligente capaz de escribir su propio programa. Un sistema experto definido como una estructura de programación capaz de almacenar y utilizar un conocimiento sobre un área determinada que se traduce en su capacidad de aprendizaje.

El objetivo de la investigación es proponer una herramienta capaz de predecir la cantidad de fallecidos por COVID-19 en función del tiempo. Para el cumplimiento de dicho objetivo, el presente artículo consta de cuatro secciones, aparte de la introducción y se estructura de la siguiente manera: La sección Materiales y métodos o Metodología computacional, describe el desarrollo de las actividades realizadas a lo largo del trabajo de investigación, partiendo de conceptos previos y definición de las herramientas utilizadas, la sección de Análisis de los resultados, muestra la descripción del *dataset* utilizado, su respectivo tratamiento, entrenamiento de la herramienta y predicción; Finalmente en la sección de conclusiones, se presenta en base al objetivo planteado.

MATERIALES Y MÉTODOS O METODOLOGÍA COMPUTACIONAL

Conceptos

Para dar solución al problema identificado es necesario conocer el concepto de Inteligencia Artificial, Series Temporales, Estacionalidad y las Principales librerías utilizadas en Python.

Inteligencia Artificial

Una de las definiciones que se puede considerar más ajustada a la realidad es la reflejada en la Encyclopedia Of Artificial Intelligence:

"La Inteligencia Artificial es un campo de la ciencia y la ingeniería que se ocupa de la comprensión, desde el punto de vista informático, de lo que se denomina comúnmente comportamiento inteligente. También se ocupa de la creación de artefactos que exhiben este comportamiento".

Otros autores prefieren otras definiciones como:

"La Inteligencia Artificial es el estudio de las ideas que permiten ser inteligentes a los ordenadores" (H.Winston).

"Parte de la informática que estudia procesos simbólicos, razonamiento no algorítmico y representaciones simbólicas de conocimiento" [3].

Series Temporales

Como lo describe [4], una serie temporal es un conjunto de observaciones de una variable tomadas al largo de intervalos regulares de tiempo, como el número de automóviles vendidos por un fabricante cada mes durante los últimos diez años. Las series temporales aparecen prácticamente en todos los campos de actividad. El interés de su análisis estadístico radica en el estudio del comportamiento de la serie, lo que permite explicar sus variaciones y, sobre todo, en la posibilidad de predecir valores futuros.

Estacionalidad

La estacionalidad se define por una fluctuación cíclica o periódica de la serie temporal que se repite de forma regular.

Si no fuera por la estacionalidad, el análisis de las series temporales se convertiría en una tarea muy simple.[5]

Un ejemplo claro de variación estacional es la que se produce a lo largo del año, particularmente en climas templados y fríos, en cuanto la temperatura, las precipitaciones inundaciones y sequías periódicas, la duración del día y la noche y la ecología, y que dan lugar a las estaciones del año.

Herramientas

En esta investigación propone como aporte una herramienta para determinar la predicción del crecimiento de los fallecimientos debido al Covid 19 en el mes de Julio, usando para ello Series temporales mediante herramientas de *machine learning*. Se utilizó como lenguaje de programación Python, la cual utiliza programación multiparadigma ya que soporta parcialmente la orientación a objetos, programación imperativa y en menor medida la programación funcional.

Principales librerías de Python para *Machine Learning*

- **Pandas**: Librería más utilizada para el tratamiento de datos en Python, una de sus grandes virtudes que tiene esta librería es la carga de datos como los archivos de texto plano como CSV "Comma Separated Valúes"[6].

- **Numpy**: Librería que por excelencia tiene su virtud en el procesamiento de arrays. Debido a que contiene una gran colección de funciones que permite realizar cálculos matemáticos complejos sobre *arrays* multidimensionales. [7]
- **Matplotlib**: Esta librería es importante en las tareas de visualización y entre sus cualidades destacan que es open *source* y trabaja a bajo nivel.
- **Seaborn**: Es una librería gráfica basada en matplotlib, especializada en la visualización de datos estadísticos. Se caracteriza por ofrecer un interfaz de alto nivel para crear gráficos estadísticos visualmente atractivos e informativos.
- **Skelearn**: Es una librería que cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad.
- **Keras**: Es una interfaz de alto nivel para manipular las redes neuronales, con keras podemos ver fácilmente si nuestras ideas darán buenos resultados inmediatos.[8]

Google Colaboraty

Como se indica. [9], utilizaremos la herramienta de Google Colaboratory también llamada "Colab" la cual nos permitirá ejecutar y programar en Python desde nuestros navegadores mediante una Jupyter Notebook.

ANÁLISIS Y RESULTADOS

Diseño y Población

El presente estudio fue descriptivo y estuvo basado en un análisis de series de tiempo correspondiente al período entre el 3 de marzo del 2020 al 3 de julio del 2021 en Perú. Se utilizó un *dataset* de los fallecidos por COVID 19, extraídos del portal del Gobierno del Estado Peruano en la sección de datos abiertos proporcionados por el Ministerio de Salud.

Tratamiento del *Dataset*

El *dataset* consta de 193,230 registros pertenecientes al periodo del 3 de marzo del 2020 al 3 de julio del 2021, a continuación, se muestra el diccionario de datos:

Tabla 1. Diccionario de Datos.

Variable	**Descripción**
FECHA_CORTE	Fecha de corte de información
UUID	ID de la persona fallecida por covid-19
FECHA_FALLECIMIENTO	Fecha de fallecimiento que ocurre por covid-19
EDAD_DECLARADA	Edad de la persona fallecida por covid-19
SEXO	Sexo de la persona fallecida por covid-19
CLASIFICACION_DEF	Criterios utilizados para la confirmacion de la defunción por covid-19
UBIGEO	Código de Ubicación Geografica que denotan "DDppdd" (Departamento, provincia,distrito), fuente INEI
DEPARTAMENTO	Departamento donde reside la persona fallecida por covid-19
PROVINCIA	Provincia donde reside la persona fallecida por covid-19
DISTRITO	Distrito donde reside la persona fallecida por covid-19

Fuente: MINSA.

Se ejecutaron las acciones de retirar aquellas filas con datos anómalos y anormales con la finalidad de no afectar considerablemente a los resultados, la estrategia desarrollada para dicho tratamiento consistió en eliminarlas del *dataset*.

Seguidamente y disponiendo de un *dataset* depurado se procedió a eliminar las columnas FECHA_CORTE, UUID, CLASIFICACION_DEF, UBIGEO, DEPARTAMENTO, PROVINCIA y DISTRITO, la justificación de lo señalado se basa en el aspecto que la serie de tiempo será aplicada sobre la columna FECHA_FALLECIMIENTO.

Se realiza una conversión de la columna FECHA_FALLECIMIENTO que es de tipo INT a formato DATETIME, el cual nos permitirá realizar el agrupamiento de cantidad de fallecidos por fecha.

A continuación, se procede a crear un índice con la correspondiente columna de FECHA_FALLECIMIENTO, es importante señalar que una de las características más poderosas y recomendadas de las series temporales es la indexación basada en el tiempo, el uso de fechas y horas para organizar y acceder de forma intuitiva a los datos.

A partir del índice creado iniciamos las acciones de obtener la cantidad de filas agrupadas por fecha con las que se va a procesar siendo un total de: 479 filas.

La reducción de resolución es volver a muestrear con *resample* un conjunto de datos de series de tiempo a un marco de tiempo más amplio, en el caso aplicado se está pasando de días a meses originando un número reducido de filas.

Estacionalidad

En el siguiente modelo se ajusta la media de los fallecidos por COVID 19 del año 2020 con el año 2021, observando que las líneas de color azul y naranja no coinciden, no son exactas, pero sí muestran una misma tendencia. Esto muestra un comportamiento de un periodo de repetición por lo cual esta serie temporal es considerada Estacionaria.

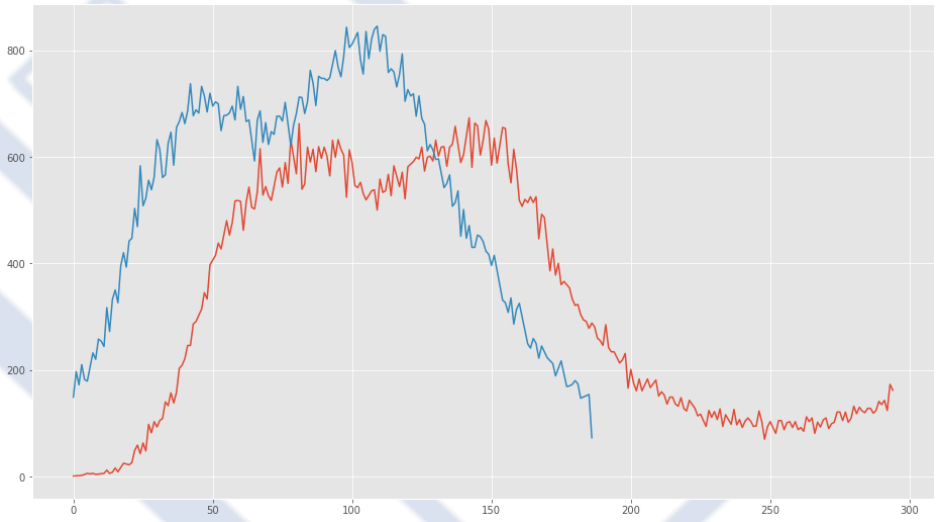

Figura 1. Comportamiento de la Serie temporal de los fallecidos por COVID 19 del año 2020 y 2021.

PREPROCESAMIENTO

Debido a que la cantidad de variables de entrada es una sola columna, se necesita hacer un preprocesamiento de nuestro *dataset* para tener más variables de entrada. Para este fin estamos considerando el hacer una conversión de series a aprendizaje supervisado considerando 7 días.

El resultado de esta conversión es el siguiente:

	var1(t-7)	var1(t-6)	var1(t-5)	var1(t-4)	var1(t-3)	var1(t-2)	var1(t-1)	var1(t)
7	-1.000000	-1.000000	-1.000000	-0.997633	-0.992899	-0.988166	-0.990533	-0.988166
8	-1.000000	-1.000000	-0.997633	-0.992899	-0.988166	-0.990533	-0.988166	-0.992899
9	-1.000000	-0.997633	-0.992899	-0.988166	-0.990533	-0.988166	-0.992899	-0.990533
10	-0.997633	-0.992899	-0.988166	-0.990533	-0.988166	-0.992899	-0.990533	-0.990533
11	-0.992899	-0.988166	-0.990533	-0.988166	-0.992899	-0.990533	-0.990533	-0.988166

Figura 2. Resultado del preprocesamiento.

Entrenamiento

Se aplicó el modelo de series temporales utilizando un 60% para entrenamiento y 40% para pruebas. Este procesamiento se hizo sobre un total 479 registros de agrupamiento de fallecimientos por fecha.

Con esto hemos procesado para entrenamiento 287 registros y para la validación 192 días teniendo como resultado del entrenamiento los siguientes valores

0s 2ms/step - loss: 0.0553 - mse: 0.0049 - val_loss: 0.0480 - val_mse: 0.0037

El siguiente gráfico nos muestra en los puntos verdes el valor esperado del entrenamiento y los puntos rojos el valor resultante de la red neuronal que es valor de validación de prueba y como se puede apreciar sin muy cercanos.

Respecto a la pérdida (*loss*) se puede apreciar el siguiente gráfico que muestra en azul la pérdida del valor esperado y en rojo la pérdida del valor hallado en la prueba de nuestro modelo lo que indica que también son similares.

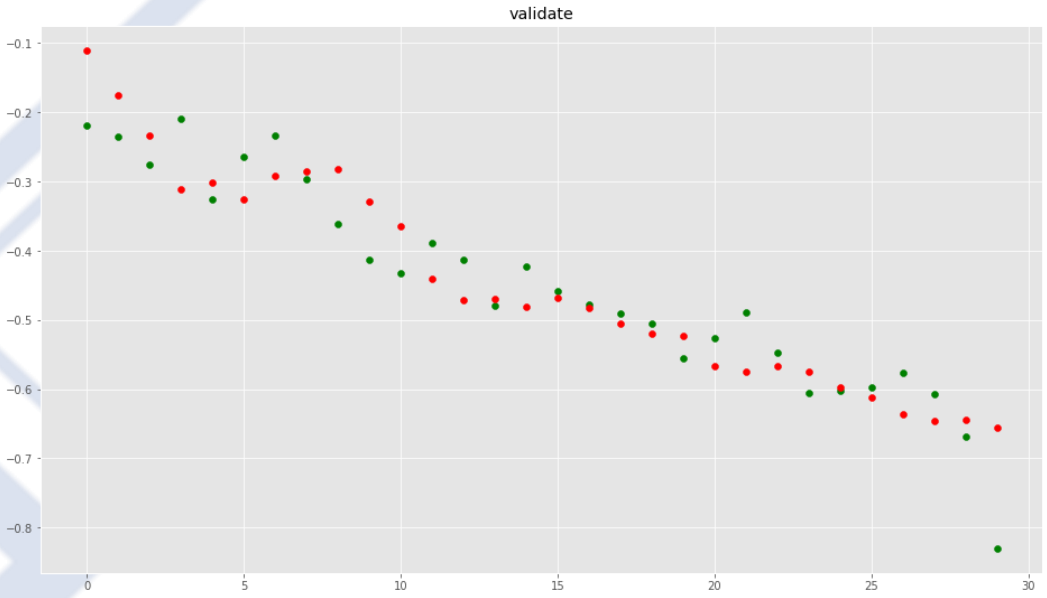

Figura 3. Tendencia de valores esperados vs valores resultantes de la red neuronal.

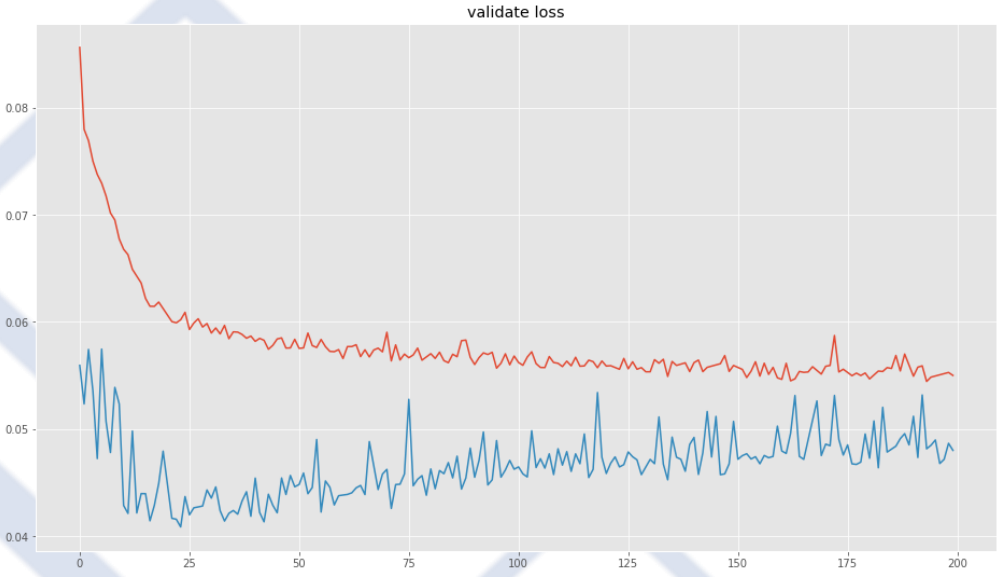

Figura 4. Tendencia de valor de Loss (pérdida) de datos esperados y resultantes de la red neuronal.

En este gráfico tenemos la variación del valor de error cuadrático medio que es la suma de las diferencias entre el valor esperado vs el valor de validación elevado al cuadrado el cual nos da al final un valor de precisión de 0.037.

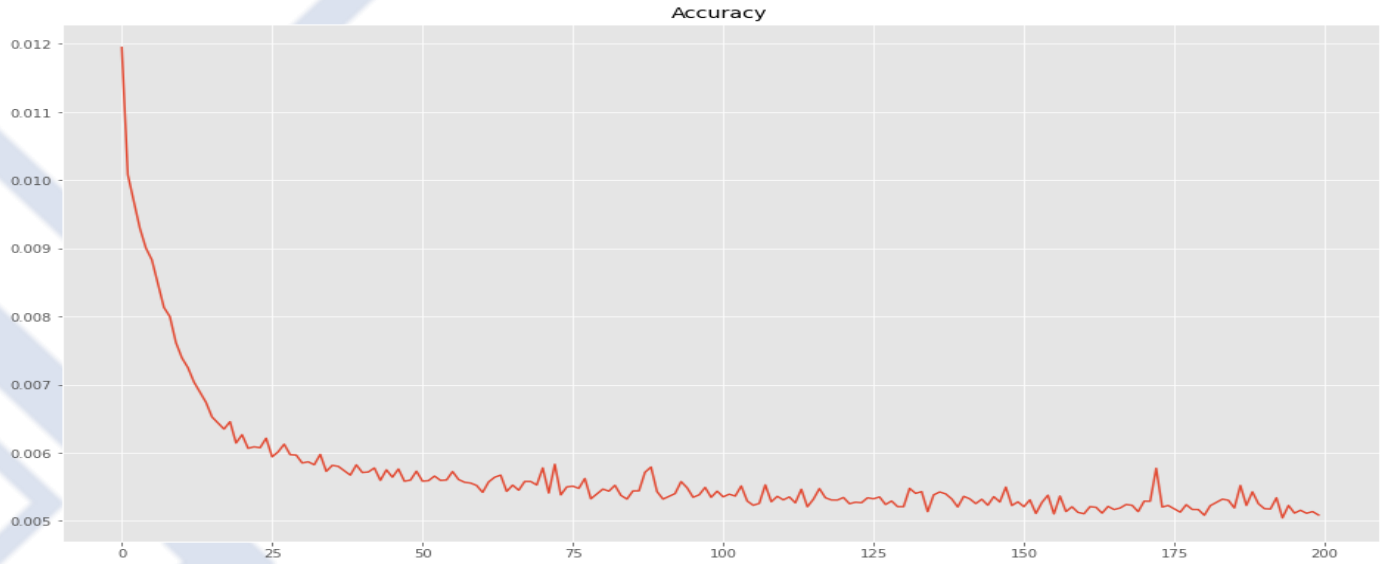

Figura 5. Comportamiento del error cuadrático medio.

Finalmente, para poder visualizar en la Figura 6 la diferencia entre los valores que nuestro modelo predijo en el entrenamiento vs el valor real de fallecidos tenemos el siguiente cuadro que nos muestra que mantiene la misma tendencia.

Predicción

Como nuestro modelo ya se encuentra entrenado y validado, procedemos a realizar la predicción; ya que, nuestro *dataset* solo nos muestra la cantidad de fallecidos hasta el 03 de Julio del 2021, se realizó la predicción de la cantidad de fallecidos para el periodo restante hasta fines del mes de Julio del 2021. Para la predicción se utilizará los últimos 30 días de nuestro *dataset* para predecir el mes de Julio.

En la Figura 7 se muestra la distribución de elementos por día, mostrando una tendencia a la baja en la cantidad de fallecidos, Se puede evidenciar que para fines del mes de Julio se llegará a los 110 fallecidos, la mitad de la cantidad de fallecidos a inicios del mismo mes que es 220; asimismo, observamos dos picos durante el mes, al tercer día del mes en el que hay un aumento de 208 a 211 fallecidos y el segundo pico lo vemos al sexto día con un aumento de 180 a 182 fallecidos.

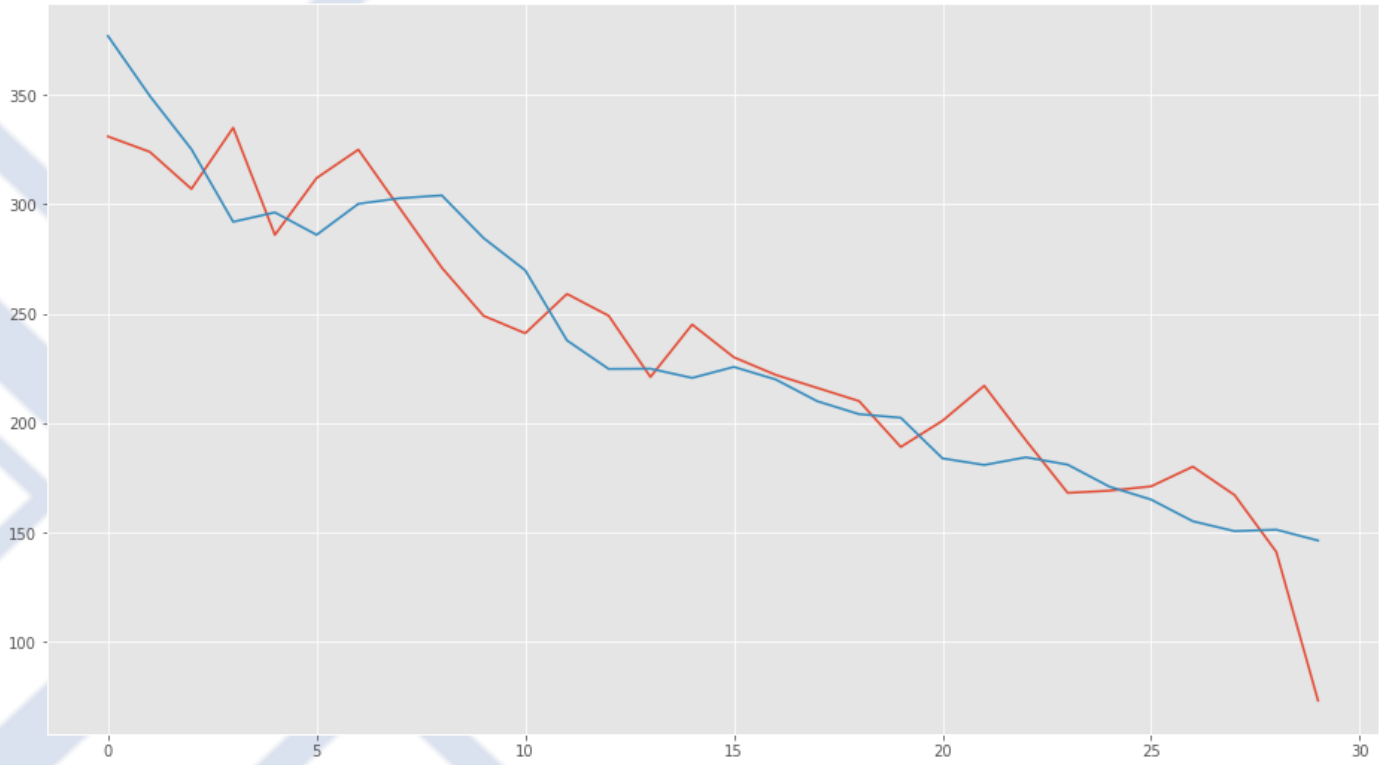

Figura 6. Tendencia de cantidad de fallecidos de la predicción del entrenamiento vs el valor real.

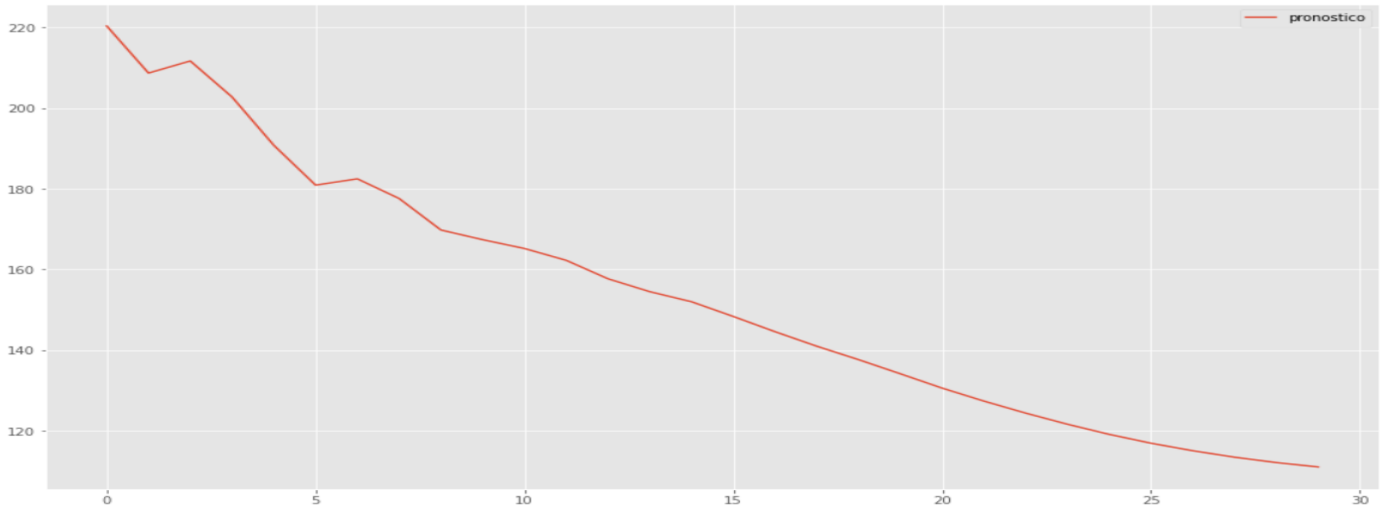

Figura 7. Pronóstico para el mes de Julio 2021.

CONCLUSIONES

Con los datos disponibles hasta la fecha se presenta una herramienta para la predicción de la cantidad de fallecidos a causa del COVID 19 basada en redes neuronales artificiales, que complementa a la información proporcionada por el ministerio de salud a través del portal de datos abiertos.

Los resultados obtenidos a lo largo del artículo confirman la validez de esta herramienta y la efectividad en la predicción de la cantidad de fallecidos a causa del COVID 19, lo cual se evidencia en la sección de análisis y resultados.

Los modelos identificados a lo largo del artículo presentan un horizonte de estimación que depende de la estacionalidad, en aras de obtener una buena predicción. Para el desarrollo de la presente herramienta se utilizó con la cantidad de fallecidos por Covid 19 durante la pandemia, por lo que dicha estacionalidad es temporal debido a que al término de la pandemia volvería a sus valores normales.

En otras palabras, se requiere construir un modelo que simule de forma más precisa utilizando un *dataset* más amplio que no considere el tiempo de pandemia para que se pueda aplicar para futuros años post pandemia.

REFERENCIAS

[1] Organización Mundial de la Salud, "coronavirus COVID 19" July, 2021. [Online]. Available: https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019?gclid=CjwKCAjw55-HBhAHEiwARMCszrbbBSFmekHH9cphVjelvC85L8pGGpKMcOMiNDkbJPAMYeUrpSEXaRoCT7MQAvD_BwE. [Accessed Jul. 09, 2021].

[2] Ministerio de salud, "datos abiertos," July, 2021. [Online]. Available: https://www.datosabiertos.gob.pe/dataset/fallecidos-por-covid-19-ministerio-de-salud-minsa/resource/4b7636f3-5f0c-4404-8526. [Accessed Jul. 09, 2021].

[3] R. Pino, A. Gómez, N.de Abajo, "Introducción a la inteligencia artificial: sistemas expertos, redes neuronales artificiales y computación evolutiva," Universidad de Oviedo, pp. 01, 2001.

[4] C. Guisande, A. Vaamonde, A. Barreiro, "Tratamiento de datos con R, Statistica y SPSS," Ediciones Diaz de santos, pp. 585, 2013.

[5] J. Arnau, "Diseños de Series Temporales: Técnicas de Análisis," Edicions Universitat Barcelona, pp. 92, 2001.

[6] W. McKinney, "Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython," O'Reilly Media, pp. 04, 2012.

[7] F. Nelli, "Python Data Analytics: With Pandas, NumPy, and Matplotlib," Apress, pp. 47, 2018.

[8] J. Torres, "DEEP LEARNING Introducción práctica con Keras," CC BY-NC-SA, pp. 97, 2018.

[9] B. Auffarth, "Artificial Intelligence with Python Cookbook: Proven recipes for applying AI algorithms and deep learning techniques using TensorFlow 2.x and PyTorch 1.6," Packt Publishing Ltd, pp. 10, 2020.

Propuesta de un plan de seguridad de la información para incrementar la fiabilidad de datos en una financiera

Proposal of an information security plan to increase the reliability of data in a financial company

Wilmer Aufredy Apaza Chávez
Universidad Nacional de Trujillo
wapazac@unitru.edu.pe

ARK: ark:/42411/s6/a39
PURL: 42411/s6/a39

RECIBIDO 12/05/2021 • ACEPTADO 24/06/2021 • PUBLICADO 30/09/2021

RESUMEN

La entidad financiera tiene como función principal ofrecer sus servicios de colocación de tarjetas, préstamos, etc., hacia los clientes que soliciten en sus diferentes establecimientos. Ante ello se identificó que en el banco existen actividades que están generando mal manejo de la información por parte del personal hacia los clientes lo cual está ocasionando reclamos de los mismos por inconsistencia de los datos que trae como consecuencia la desafiliación de sus servicios. Por ese motivo se desarrolló una propuesta de un plan de seguridad de la información en los procesos y áreas del banco Ripley, teniendo como objetivo el incremento de la fiabilidad de sus datos, logrando los tres principios para un SGSI como son disponibilidad, integridad y confidencialidad. Para lograr dicho objetivo se seleccionó las normas ISO/IEC 27001 y 27002 para aplicar los controles de la propuesta del plan de seguridad de la información en el banco Ripley, quedando claramente establecidos los responsables y la información que se maneja en cada una de los procesos y áreas. Como resultado se realizó el alcance del plan, así como definir las políticas, análisis de gestión de riesgos, se dio prioridad al manejo de la información por áreas, además se analizó los activos del banco donde se garantice la fiabilidad de los datos, luego se definió el plan aplicando los controles de la ISO/IEC 27002. Se concluyó en definir los indicadores para evaluar la propuesta del plan de seguridad de la información para incrementar la fiabilidad de sus datos.

Palabras claves: Datos, Gestión, Información, ISO/IEC 27002, Plan ,SGSI.

ABSTRACT

The main function of financial entity is to offer its services for the placement of cards, loans, etc., to customers who request in their different establishments. Given this, it was identified that in the bank there are activities that are generating mishandling of information by the staff towards the clients, which is causing complaints from them due to the inconsistency of the data that results in the disaffiliation of their services.

For this reason, a proposal for an information security plan was developed in the processes and areas of Ripley Bank, with the objective of increasing the reliability of its data, achieving the three principles for an ISMS such as availability, integrity and confidentiality. To achieve this objective, the ISO / IEC 27001 and 27002 standards were selected to apply the controls of the proposal for the information security plan in the Ripley bank, with the responsible parties and the information handled in each of the processes being clearly established. and areas. As a result, the scope of the plan was carried out, as well as defining the policies, risk management analysis, priority was given to the management of the information by areas, and the assets of the bank were analyzed where the reliability of the data is guaranteed, then defined the plan applying the controls of ISO / IEC 27002

It was concluded in defining the indicators to evaluate the proposal of the information security plan to increase the reliability of its data.

Keywords: *Data, Management, Information, ISO / IEC 27002, Plan, SGSI.*

INTRODUCIÓN

En el Perú, en la actualidad en las organizaciones financieras es incuestionable que la gran mayoría de los procesos del negocio son soportados, automatizados y gestionados por sistemas informáticos, así como los sistemas de información apoyan la actividad gerencial y la toma de decisiones; incluso muchas veces, es la propia información y el acceso a la misma, el producto o servicio que se intercambia como el principal objeto del negocio. La seguridad de la información ya no puede ser considerada como el resultado de un accionar defensivo y reactivo para preservar los activos del negocio, ya que muchas veces, es un activo del mismo, una condición para operar y/o competir con el sector financiero, un generador de valor. Requiere un accionar proactivo y su incorporación como elemento estratégico. A modo de ejemplo, un banco que gestione adecuadamente la seguridad de la información, por un lado, da cumplimiento a sus obligaciones y regulaciones y a su vez genera confianza entre sus clientes.

la información es el principal activo de toda organización según los más modernos paradigmas de la administración empresarial pudiendo hacer su aparición de muchas formas, impresa o escrita en el papel, almacenada electrónicamente, transmitida por correo, ilustrada en películas

o hablada en conversaciones. En el ambiente de negocios financieros de hoy, esa información está constantemente bajo la amenaza de muchas fuentes, que pueden ser internas o externas, accidentales o maliciosas para con el banco. Con el incremento del uso de nueva tecnología para almacenar, transmitir y recobrar información se han abierto canales para un mayor número y variedad de amenazas. Es por ello que se requiere establecer, un planeamiento de seguridad de la información dentro de cualquier tipo de organización. Es necesario asegurar la confidencialidad, integridad y disponibilidad de la información vital para el banco y de sus clientes [1].

Una estrategia de gestión de información es esencial para sobrevivir en el mercado financiero actual, un ejemplo es ver un escenario donde los activos de información de la organización están rodeados por un complejo ambiente de objetos y amenazas que van desde simples virus de una computadora hasta robos de la propiedad intelectual del negocio [2].

Por lo tanto, cada vez hay más conciencia y consenso en la importancia de la seguridad de la información en las empresas y organizaciones cualquiera sea el sector de la economía o rol en la sociedad que desempeñen, en lo particular en las empresas medianas y grandes. Sin embargo, existen diversas industrias y estructuras empresariales que hacen que algunos temas deban ser analizados y estudiados con una estrategia diferente, ya sea por criticidad de la información que manejan, su dimensión o estructura empresarial [3].

A modo de ejemplo, pequeñas empresas, con una infraestructura limitada y sistemas informáticos de gestión que no requieren almacenamiento y procesamiento de información confidencial o critica ni están sujetos a estrictas normas regulatorias, normalmente van a enfrentar riesgos menores, deben considerar aspectos diferentes a la de una gran corporación o grupo empresarial financiero, y también una dimensión del problema diferente, tanto en la problemática como en la del capacidad de gestión de la solución. Por lo tanto, su estrategia y decisiones responderán a estas diferencias estructurales.

Por otra parte, organizaciones más grandes como pueden ser: empresas del sector financiero, salud, operadores de telefonía, gubernamentales, etc., deben afrontar la seguridad de la información de forma metodológica y planificada y con planes concretos, con un enfoque de continuidad del negocio y mejora continua. Además de parámetros y dimensiones diferentes en su relación costo-beneficio, existen motivos legales, regulaciones y contratos que requieren de la protección de información personal y sensible además de la crítica y estrategia del negocio.

De acuerdo a algunas encuestas internacionales, el mayor riesgo a la seguridad de la información está dado por el factor humano, específicamente errores, conductas inapropiadas y/o negligencias generadas internamente. también existen referencias donde se asegura que la inversión en la gestión de la seguridad (de T.I.) es más efectiva que la inversión tecnológica para mejorar los niveles de seguridad [4][5].

El desafío es entonces lograr el desarrollo del plan de seguridad de la información que conduzca en una solución eficaz y eficiente, desde el punto de vista técnico y económico que provea los niveles de seguridad requeridos y brinde la confianza necesaria en la entidad financiera, los socios del negocio y los clientes y bajo este enfoque es necesario considerar lo siguiente: las necesidades de los procesos del negocio con respecto a la información, aplicaciones. El uso eficaz y eficiente de los recursos tecnológicos como soporte de estos procesos de negocio. Un enfoque predictivo, estratégico y económicamente raciona en la evaluación y tratamiento de riesgos. La confiabilidad de las soluciones con particular atención en la continuidad del negocio y de los procesos estratégicos (críticos y de mayor valor)

Para alinear todo lo dicho anteriormente en cuanto a la seguridad de la información se utilizará la norma ISO/IEC 27001, controles de la ISO 27002, esta norma ha sido preparada para proporcionar los requisitos para establecer, implementar, mantener y mejorar de manera continua un sistema de gestión de seguridad de la información. Esta debe conservar la confidencialidad, integridad y disponibilidad de la información al aplicar un proceso de gestión de riesgo y la entrega confianza a las partes interesadas cuyos riesgos son gestionados de manera adecuada.

La Información

Es un activo del negocio, tiene una función importante en la organización y por consecuencia debe estar protegido adecuadamente (ISO/IEC, 2014). La información puede ser clasificada de diversas formas, pero por la manera de comunicarse tenemos: Hablada en reuniones, impresa o escrita en papel, almacenada electrónicamente, transmitida por correo convencional o electrónicamente, Exhibido por videos corporativos. Los activos de la información son todo lo que tiene valor para la organización como: Software, servicios, intangibles como la reputación e imagen, personas y sus habilidades, certificaciones, Computadora, servidor etc.

Figura 1: Activos de la Información. Fuente Bendermacher.

Seguridad de la Información

La Seguridad de la información es mucho más que establecer firewalls, aplicar parches para corregir nuevas vulnerabilidades de un sistema de software o guardar copias de seguridad.

"Seguridad de información es determinar que requiere ser protegido y por qué, de que debe ser protegido y cómo protegerlo", es cualquier medida que impida la ejecución de operaciones no autorizadas sobre un sistema o red informática, cuyos efectos pueden conllevar daños sobre la información, comprometer su confidencialidad, autenticidad o integridad y disponibilidad, disminuyendo el rendimiento de los equipos o bloquear el acceso a usuarios autorizados al sistema.

Dependiendo del tipo de información manejada y de los procesos realizados por una organización, esta podrá destinar más o menos recursos a garantizar la confidencialidad, la integridad o autenticidad y la disponibilidad de sus activos de información. Para toda organización, es fundamental contar con un SGSI. Muchas empresas creen tener sistemas eficientes y eficaces para proteger y asegurar la información, tienen controles e inclusive software para aplicar los controles, pero los aplican solo cuando hay incidente de seguridad; de manera que actúan de forma reactiva, sin tener un enfoque claro y bien estructurado para un SGSI. Es por ello que las organizaciones requieren de un sistema que permita asegurar la información de manera proactiva, no obstante, hay organizaciones que han diseñado verdaderos SGSI, pero al momento de la puesta en marcha resultan caducos o incluso no aplican los controles.

La seguridad de la información involucra la tecnología, las personas y la estructura organizacional (procesos), las normativas, lo cual hace necesario un amplio conocimiento sobre la gestión de estos recursos. Sin embargo, esta gestión puede servir parcialmente, poco o nada si existen fallas de hardware, de software, fallas humanas, desastres naturales, ataques terroristas, entre otros, sin que la organización haya estado preparada para estos eventos

Es importante que, en todo este proceso, saber de qué proteger, de quien proteger y cómo proteger, esta es la palabra clave para poder direccionar el diseño y el mejoramiento continuo del SGSI. Dada la competencia de la globalización y las nuevas formas de comercio internacional, las empresas, sin importar su tamaño, su actividad o ubicación, deben estar preparadas para asegurar su información.

Calidad de la Información

Se caracteriza por la preservación de los siguientes aspectos:

- Confidencialidad: Se asegura que la información sea accesible solo para aquellos que estén autorizados
- Integridad: Salvaguardando la exactitud de la información en su procesamiento, así como su modificación autorizada
- Disponibilidad: Asegurando que los usuarios autorizados tengan acceso a la información y a los activos asociados cuando sea necesario

La sensibilización de los directivos y responsables de la organización, que deben ser conscientes de la necesidad de destinar recursos a esta función.

Los conocimientos, las capacidades e implicación de los responsables del sistema informático: dominio de la tecnología utilizada en el sistema y conocimiento sobre posibles amenazas y los tipos de ataque. La mentalización, formación de responsabilidades de todos los involucrados en el sistema. Instalación, configuración y mantenimiento correcto de los equipos. Soporte de los fabricantes de hardware y software que permitan realizar actualizaciones y mejoras para cubrir fallos y problemas relacionados con la seguridad. Considerar que hay amenazas internas como externas en la seguridad de la información.

Seguridad Informática

Entre los conceptos que más resalta la seguridad informática es la física que cubre todo lo referido a los equipos informáticos, computadores, servidores y equipamiento de la red. La seguridad lógica se refiere a las distintas aplicaciones que se ejecutan en cada uno de estos equipos. Los desastres naturales (Incendios, inundaciones, terremotos, etc.), los tenemos en cuenta a la hora de ubicar los emplazamientos del centro de proceso de datos donde alojamos los principales servidores de la empresa; pero, aunque tengamos el mejor sistema de extinción de incendios o la sala esté perfectamente sellada, siempre deberíamos tener un segundo CPD (centro de procesamiento de datos) para que la actividad no pare. Robos nuestros equipos, y sobre todo la información que contienen, resultan valiosos para otros individuos u organizaciones. Debemos proteger el acceso a la sala del CPD mediante múltiples medidas de seguridad: vigilantes, tarjetas de acceso, identificación mediante usuario y contraseña, etc. Fallas de suministro Los ordenadores utilizan corriente eléctrica para funcionar y necesitan redes externas para comunicarse con otras empresas y con los clientes. Estos servicios los contrataremos con determinados suministradores, pero debemos estar preparados para las ocasiones en que no puedan proporcionarlo: unas baterías o un grupo electrógeno por si falla la corriente eléctrica, una segunda conexión a Internet como línea de *backup*, incluso podemos optar por una solución inalámbrica para estar protegidos ante un corte del servicio. Las amenazas lógicas: Virus troyanos o malwares, en general. Como ocurre con el spam en el correo electrónico, el malware es software

no deseado y que debemos eliminar, perdida de datos en general. Como ocurre con el spam en el correo electrónico, el malware es software no deseado y que debemos eliminar, ataque a las aplicaciones de los servidores en general. Como ocurre con el spam en el correo electrónico, el malware es software no deseado y que debemos eliminar.

Tabla 1: Consecuencia de la seguridad de la Información. Fuente Elaboración propia.

Imagen	Volumen del Negocio	Productividad y Presentación del Servicio
Pérdida de imagen respecto al cliente Pérdida de imagen respecto a los proveedores Pérdida de imagen a otras partes Ventaja de los competidores, etc.	Pérdida de ingresos/facturación Posibles indemnizaciones a terceros Posibles sanciones Pérdida de oportunidades del negocio Pérdida de contratos/caída de acciones	Disminución rendimiento laboral Interrupción de procesos productivos Retraso de entregas Cese de transacciones Enfado de empleados

Sistema de Gestión de Seguridad de la Información

El manejo de la seguridad de información es como un proceso, requiere de conocimientos, habilidades y capacidades de las áreas técnicas, legal, humana y organizacional. Un sistema en la que se puedan integrar todos los factores con todos los requerimientos y consideraciones organizacionales, recibe el nombre de sistema de gestión de seguridad de la información, conocido por las siglas como SGSI. La parte del sistema general de gestión, que comprende la política, la estructura organizativa, los procedimientos, los procesos y los recursos necesarios para implantar la gestión de la seguridad de la información en una organización se denomina SGSI. Para implementar un sistema de gestión de seguridad de información en una organización, debe considerar lo siguiente: Formalizar la gestión de seguridad de información, Analizar y gestionar los riesgos. Establecer los procesos de gestión de seguridad en base a la metodología, Certificar la gestión de la seguridad. Para esto debe tenerse en cuenta el marco legal, los estándares, las metodologías, los requerimientos, entre otros aspectos fundamentales.

Por su trascendencia, es importante mencionar algunos de los enfoques más relevantes al tema de la gestión para la continuidad de las operaciones: Plan para recuperación ante desastres

(Disaster recovery planning RDP), se enfoca en la recuperación de los servicios de TI y los recursos, dados un evento que ocasionara una interrupción mayor en su funcionamiento. Plan para reanudación del negocio *(Business resumption planning BRP),* se centraliza en la reanudación de los procesos de los negocios afectados por una falla en las aplicaciones de TI. Se enfoca en la utilización de procedimientos relacionados con el área de trabajo.Plan para la continuidad de las operaciones *(Continuity or operation planning COOP)*, busca la recuperación de las funciones estratégicas de una organización que se desempeñan en sus instalaciones corporativas. Plan de contingencia *(contingency planning CP)*, se enfoca en las recuperaciones de los servicios y recursos de TI, después de un desastre de dimensiones mayores o de una interrupción menor. Especifica procedimientos lineamientos para la recuperación, tanto en áreas de la empresa como las alternas. Plan de respuesta ante emergencias *(Emergency response planning),* su objetivo es salvaguardar a los empleados, el público, el ambiente y los activos de la empresa.

Últimamente se busca de inmediato llevar la situación de crisis a un estado de control. Todos los enfoques tienen como denominador común, el cual, es un limitado alcance. Cada una de estas ópticas de planeación se centra en proyección de aspectos específicos de la organización, ignorando otras áreas críticas. Para atender esta limitación, se requiere un enfoque de planeación integrado, que permita proteger todas las áreas críticas de la organización. Plan de continuidad del negocio PNC (o sus siglas en inglés *BCP-Businnes continuity plan*), integra el alcance y los objetivos de todos estos enfoques [6].

Figura 2: Proyectos que constituyen un SGSI. Fuente Gómez.

La norma ISO/27001

La norma ISO 27001, es un estándar desarrollado como modelo para el establecimiento, implantación, operación, monitorización, revisión, mantenimiento y mejora de un SGSI para cualquier tipo de organización. Permite diseñar e implementar un SGSI, se encuentra influenciado por las necesidades, objetivos, requisitos de seguridad, los procesos, los empleados. Los sistemas de soporte y la estructura de la organización. Su origen es británico se creó en el año 2005, pero tiene su versión más reciente que data del año 2013, la Organización Internacional para la Normalización (ISO) la oficializo como norma.

El ISO 27001:2013 es el único estándar certificable, aceptado internacionalmente de manera global para la gestión de la seguridad de información; aplica a todo tipo de organizaciones, tanto por su tamaño como su actividad. La norma ISO 27001, actúa bajo el enfoque de procesos. La aplicación de unos sistemas de procesos, dentro de la organización, junto con la identificación y las intersecciones de estos procesos, así como su gestión, puede denominarse como enfoque basado en procesos. El enfoque basado en procesos para la gestión de la seguridad de información presentada en esta norma, enfatiza a los usuarios, la importancia de: La comprensión de los requisitos de la seguridad de una organización y la necesidad de establecer políticas y objetivos para la seguridad de información, Implementar y operar controles para dirigir los riesgos de la seguridad de información de una organización con el contexto de los riesgos globales del negocio de la organización, Realizar seguimiento y revisar el desempeño y la eficiencia del SGSI, Mejora continua con base en mediciones objetivos. Es una forma sistemática de administrar la información sostenible de una institución, para que permanezca segura. Abarca a las personas, los procesos y las tecnologías de información. La forma total de la seguridad de la información, y a la integración de las diferentes iniciativas de seguridad necesitan ser administradas para cada elemento sea completamente efectivo. Aquí en donde entra el SGSI, que permite coordinar esfuerzos de seguridad con mayor efectividad.

Órganos gubernamentales regulatorias para entidades financieras

La Superintendencia de Banca, Seguros y AFP es el organismo encargado de la regulación y supervisión de los Sistemas Financiero, de Seguros y del Sistema Privado de Pensiones, así como de prevenir y detectar el lavado de activos y financiamiento del terrorismo. Su objetivo primordial es preservar los intereses de los depositantes, de los asegurados y de los afiliados al SPP.

La SBS es una institución de derecho público cuya autonomía funcional está reconocida por la Constitución Política del Perú. Sus objetivos, funciones y atribuciones están establecidos en la Ley General del Sistema Financiero y del Sistema de Seguros y Orgánica de la Superintendencia de Banca, Seguros y AFP (Ley 26702)[7].

La oficina Nacional de Gobierno Electrónico e Informática, la Presidencia del Consejo de Ministros – PCM a través de la ONGEI, se encarga de normar, coordinar, integrar y promover el desarrollo de la actividad informática en la Administración Pública (DS Nº 066-2003-PCM, DS Nº 067-2003-PCM). Impulsa y fomenta el uso de las TICs para la modernización y descentralización del estado. Actúa como ente rector del Sistema Nacional de Informática, dirige y supervisa la política nacional de informática y gobierno electrónico. (Oficina Nacional de Gobierno Electrónico e Informática, 2018).

MATERIALES Y MÉTODOS

Población y muestra

La población (N) está compuesta por los directivos de las diferentes áreas del negocio como operaciones y gerencia de sistemas de la sucursal del Banco Ripley -Trujillo, Es de tipo poblacional ya que lo conforman los usuarios de los departamentos del banco de operaciones y gerencia de sistemas. De ello, se estima la muestra (n), la cual al ser menor o igual que 30, se calcula directamente como sigue:

N=29 personas

Instrumentación

A continuación, se procede a citar las técnicas e instrumentos que se utilizarán para la recolección de datos en el desarrollo de la presente investigación.

Para definir el alcance a nivel estratégico de seguridad de la información en el banco se procederá a identificar el organigrama y su plan estratégico de la entidad bancaria en el aspecto de seguridad de la información

Para diseñar las políticas y controles del diseño del plan de seguridad de la información basado en la norma ISO/IEC 27002 se procederá a recolectar la información de los controles de acuerdo a la ISO/IEC 27002 y se elaborará un plan de seguridad de la información que contenga: cuadro de niveles de control, en sus diferentes capítulos como: política de seguridad, aspectos organizativos de la seguridad de la información, gestión de activos, seguridad ligada a los recursos humanos, seguridad física y del entorno, gestión de comunicaciones y operaciones, control de acceso, adquisición desarrollo y mantenimiento de los sistemas de información, gestión de incidentes de la seguridad de la información y mejoras, gestión de la continuidad del negocio y cumplimiento arquitectura del programa de indicadores, de la metodología ISO/IEC 27002 , la técnica de Entrevistas a los directivos y como instrumento a las guías de entrevistas.

Para analizar la matriz de gestión de riesgos de acuerdo a la planificación. Se procederá con recolectar la información de los activos y se procederá a evaluar en cuanto a la gestión de riesgo, para ello tomamos en cuenta los activos como son personas, procesos y tecnología

Para el análisis de los datos recolectados, se utilizará el Método Deductivo, pues se va de lo general a lo específico. Este comienza dando paso a los datos en cierta forma válidos, para llegar a una deducción a partir de un razonamiento de forma lógica o suposiciones; o sea se refiere a un proceso donde existen determinadas reglas y procesos donde gracias a su asistencia, se llegan a conclusiones finales partiendo de ciertos enunciados o premisas. Ver figura 3

La validación y confiabilidad fue presentado por los docentes Ing. Nelson Ángeles Quiñones con CIP: 185097, Ing. Edward Vega Gavidia CIP:130533 y el Ing. David Agreda Gamboa CIP:86691, todos colegiados para que puedan dar sus observaciones y luego de ella procedieron a firmar las constancias de validez.

Figura 3: Plan de seguridad de la información. Fuente Elaboración propia.

RESULTADOS Y DISCUSIÓN

Realizamos el análisis FODA a nivel de TI y estratégico, como el de los procesos que maneja el banco, luego de ello elaboramos el documento final que conforma el alcance del plan de seguridad de la información.

MP1: Tarjetas de crédito: Administrar las tarjetas de crédito

MP2. Captaciones: Administrar las captaciones de nuevos clientes

MP3: Atención al cliente: Administrar los servicios de atención y contacto directo con el cliente

MP4. Soporte financiero: Administrar presupuesto y contabilidad

MP5: Soporte Operativo Administrativo: Todas las transacciones financieras del banco deben estar respaldadas y su gestión debe ser continua y de soporte a los demás macroprocesos.

Tabla 2: Matriz RAM de micropocesos y áreas del banco.

	Operaciones	Comercial	Finanzas	Riesgos	Cobranzas	CRM	RRHH
MP1	RAM	RAM		RA	A	RA	
MP2		RAM		RA			A
MP3	RA	RAM		RA	RA	RA	A
MP4	A	A	RAM	A	A	A	A
MP5	RAM		RAM				RA
MP6	RA	RA	RA	RA	RA	RA	RAM
MP7	RA	RA	A	RAM	A	RA	
MP8	RAM		RA	RA	RA		
MP9		A	A	A	RAM	A	
MP10	A	RA			A	RAM	

Nota: M. Ejecuta el Proceso | R: Recibe información | A: Brinda información

MP6.Recursos Humanos: Selección de nuevo personal y administración de personal, control de tiempos y compensaciones

MP7.Gestion de Riesgos: Riesgo de crédito, de mercado, de operación de liquidez

MP8. Control de cumplimiento: Soporte al directorio, que involucra la auditoría interna y el cumplimiento normativo

MP9.Cobranzas: Recuperar temprana o tardíamente las cuotas atrasadas de compromisos de pago de los clientes

MP10: Fidelización: Administración de clientes, perfil de uso y fidelización

Identificamos que macroprocesos son estratégicos, tácticos y operativos, de acuerdo a la participación de cada proceso en los objetivos de negocio.

Para empezar con el diseño del plan de Seguridad de Información para el banco Ripley tendremos en cuenta los siguientes puntos a desarrollar: Ver figura 4.

Se diseñará las políticas de seguridad de la información con el propósito de proteger la información del banco, estas servirán como una guía para una futura implementación de medidas de seguridad que ayuden a cumplir con la integridad, confidencialidad y la disponibilidad de los datos dentro de los sistemas de aplicación, redes, instalaciones de cómputo y procedimientos manuales. Este documento de políticas de seguridad se realizó tomando como base la siguiente documentación. Normas internas del banco referidas a la seguridad de la información. Requerimientos de la superintendencia de banca y seguros (SBS) sobre riesgo de la tecnología de la información. Estándar de la seguridad de la información ISO/IEC 27002.

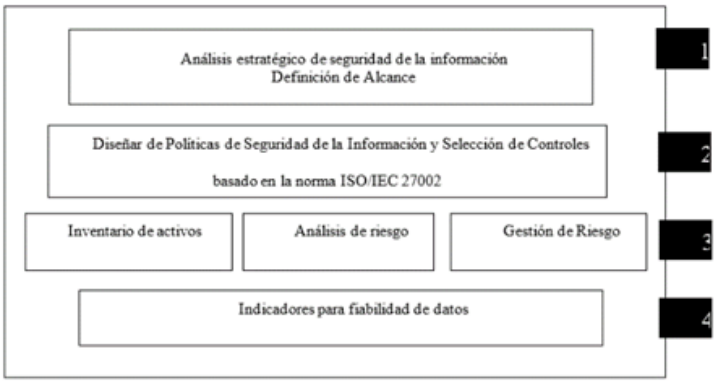

Figura 4: Esquema de diseño del plan SGSI ISO-27002. Fuente Elaboración propia.

El inventario de activos es la recopilación de todos aquellos elementos indispensables para que la administración electrónica pueda prestarse con todas las garantías, de manera que los ciudadanos tengas confianza en ella. Metodología para analizar, evaluar y gestionar los riesgos.

Figura 5: Pasos para aplicar la metodología en la gestión de riesgos. Fuente Elaboración Propia.

El nivel de riesgo vendrá dado por el valor más alto para cada activo. Tanto el nivel de vulnerabilidad como el nivel (o probabilidad) de amenaza se valoran de 0 a 3 (no aplicado, bajo, medio y alto).

Nivel de riesgo = Nivel de amenaza x Nivel de vulnerabilidad x Nivel de impacto.

Luego de aplicar los controles a los activos, procedemos a establecer los indicadores los cuales se usarán en la implementación para medir el impacto de la planificación y por consiguiente una implementación futura, cabe señalar que el objetivo de esta investigación es estipular los indicadores en este planeamiento de la seguridad de la información en la fiabilidad de datos, lo que quiere decir que se enfoca en los tres pilares de un SGSI que son:

Tabla 3: Cálculo del riesgo para aplicaciones comerciales. Fuente elaboración propia.

Amenaza	Impacto (valor del activo)	Nivel de amenaza	Vulnerabilidad	Nivel de Riesgo
Fuego	7	1	0	0
Robo	7	1	1	7
Error de mantenimiento	7	1	3	21
Fallo de software	7	3	2	42
Fallo de comunicaciones	7	2	1	14
Errores de usuario	7	2	2	28

Confidencialidad: el cual previene el acceso no autorizado a la información, de manera intencional o no.

Integridad: Evita modificaciones de la información por parte de personal no autorizado.

Disponibilidad: Proporciona acceso seguro a la información en el momento en que se precisa.

Luego de "Definir los indicadores para evaluar el plan de seguridad de la información", podemos apreciar que solo se contemplan 10 indicadores que ayudarán en la implementación de controles de un SGSI, de acuerdo a estos parámetros se debe realizar una evaluación en la cual se determinará el impacto que generará la planificación en el banco.

Elección de los indicadores a través de la evaluación de expertos en la aplicación del método V de Aiken. En qué medida los indicadores de la prueba son una muestra representativa del constructo. Medida para cuantificar el acuerdo de los jueces (expertos) A continuación,

presentamos los resultados al aplicar el método de Aiken para la elección y fiabilidad de los datos de acuerdo a la evaluación de los expertos.

De la figura 4, como podemos apreciar los coeficientes de Aiken (A_k) para ambas representatividades y luego de la evaluación de los jueces, los puntos de cortes aceptables del método de Aiken nos dicen que estamos en la métrica permitida la cual da validez de nuestros indicadores para aplicarse y garantizar el incremento de la fiabilidad de los datos. $0.6<=A_k<=1.0$, el intervalo de resultado obtenido: $A_k=0.96$ y $A_k=0.98$.

Tabla 4: Análisis del método V de Aiken mediante expertos para la elección de indicadores. Fuente elaboración propia.

Indicadores	Fiabilidad de datos A_k por indicador	Elección de indicadores A_k por indicador
ID 1	0,92	0,92
ID 2	1,00	1,00
ID 3	0,92	1,00
ID 4	1,00	1,00
ID 5	0,92	1,00
ID 6	1,00	1,00
ID 7	1,00	0,92
ID 8	0,92	1,00
ID 9	1,00	1,00
ID 10	0,92	1,00
Ak Total	0,96	0,98

Se realizó un análisis del método V de Aiken para determinar la elección de indicadores mediante expertos, el cual nos dio como resultado un coeficiente que está en el margen de corte como aceptable en la escala de Aiken y el cual garantiza la valides de los indicadores para lograr el incremento de la fiabilidad de los datos.

CONCLUSIONES

Se logró realizar el documento oficial del alcance para la aplicación del plan de seguridad de la información basado en la norma ISO/IEC 27001 y sus controles contemplados en la ISO/IEC 27002.

Se logró desarrollar el documento de las políticas de seguridad de la información, el cual representa una herramienta muy importante para la aplicación de los controles en los procesos que se determinó en el alcance.

Se logró evaluar la matriz de gestión de riesgos de acuerdo a los activos (personal, procesos y tecnología) al cual se les determinó un valor de acuerdo a la vulnerabilidad que ocasionaría este

en caso de alguna eventualidad, y de acuerdo a ello se determinó los controles a aplicar en los procesos donde se manipulan estos activos.

Se logró definir los indicadores para evaluar el plan de la seguridad de la información, donde se aprecian los indicadores, garantizan la confiabilidad de los datos y está lista para ser aplicadas en una futura implementación.

AGRADECIMIENTOS

A la universidad Nacional de Trujillo por ser mi alma mater y a todo el personal docente y administrativo que labora en la Escuela de Postgrado Sección de Ingeniería, por sus enseñanzas y apoyo durante el tiempo que curse la maestría.

REFERENCIAS

[1] Moreira, M. (2015). Auditoría del sistema informático del ministerio de transporte y obras públicas. España: Escuela politécnica nacional.

[2] Mega, G. (2014). Metodología de Implementación de un SGSI en un grupo Empresarial Jerarquico. Montevideo-Uruguay: Universidad de la República.

[3] Pressman, R. (2013). Ingeniería del Software. McGraw Hill, 10-15.

[4] Security, F. O. (5 de Mayo de 2017). Federal Office for Information Security – Germany. Obtenido de "The IT Security Situation in Germany in 2017: http://www.bsi.bund.de/english/publications/securitysituation/Lagebericht_2017_englisch.pdf

[5] Grundschutz, G. (8 de Junio de 2017). Federal Office for Information Security – Germany. Obtenido de http://www.bsi.bund.de/

[6] Guerra, A., & Mantilla, R. (2009). Diseño de un Sistema de Gestión de Seguridad de la Información para Cooperativas de Ahorro y Credito en Base a la norma ISO 27001.

[7] SBS. (15 de Enero de 2017). Superintendecia de Banca , Seguros y AFP. Obtenido de https://www.sbs.gob.pe/

[8] Aguilar, M., & Villena, A. (2015). Sistema de Gestión de Seguridad de la Información en una Institución Financiera. TESIS PUCP, 6-9.

[9] BCRP. (15 de Enero de 2017). Banco de Reserva del Perú. Obtenido de http://www.bcrp.gob.pe/sitios-de-interes/entidades-financieras.html.

[10] Bendermacher, J. (s.f.). Auditoría interna y auditoría externa. Obtenido de https://global.theiia.org/translations/PublicDocuments/GPI-Distinctive-Roles-in-Organizational-Governance-Spanish.pdf

[11] Betarte, G. (2014). Information Security – Security conscious. Uruguay. Obtenido de https://www.fing.edu.uy/inco/pedeciba/bibliote/cpap/tesis-pallas.pdf

[12] Munoz, G. Wastewater treatment for the U.C. Davis Arboretum. Recuperado de: http://lda.ucdavis.edu/people/2013/GMunoz.pdf Tomado el 04/01/2015.

[13] Córdova, L., & Muñoz, R. (2014). Planeamiento Estratégico de Tecnología de Información de Banco Ripley Perú. Tesis - UPC, 20-30.

[14] Franco, D., & Guerrero, C. (2013). Sistemas de Controles de la Seguridad Informática basado en ISO/IEC 27002. 5-8.

[15] Gomez, G. (2016). Interpretación de la Norma ISO/IEC 27001:2013. Informe de investiagacion, Universidad ESAN, Lima.

[16] Indacochea, A. (2012). Una Propuesta para mejorar las prácticas de Gobierno Corporativo en el Perú. CENTRUM,Pontificia Universidad Catolica del Perú, 12-15.

[17] Mantilla, A. (junio de 2009). Diseño de un sistema de seguridad de la información para cooperativas de ahorro y crédito en base a la norma ISO 27001. Obtenido de http://bibdigital.epn.edu.ec/bitstream/15000/8108/1/CD-2254.pdf

[18] Muñoz, R. (2017). Planeamiento Estratégico de Tecnología de la Información de Banco Ripley Perú. Lima: Repositorio UPC.

[19] NCh-ISO. (2013). Tecnología de Información-Técnicas de Seguridad -Sistema de Gestión de Seguridad de la Información. Norma Chilena - 27001, 1-2.

[20] SBS. (15 de Enero de 2017). Superintendecia de Banca, Seguros y AFP. Obtenido de https://www.sbs.gob.pe/

Aplicación de regresión logística para la predicción de demanda por especialidad médica en consulta externa hospitalaria

Application of logistic regression for the prediction of demand by medical specialty in hospital outpatient consultation

Rene Aquino Arcata
Universidad Jorge Basadre Grohman
reneaquino@gmail.com
https://orcid.org/0000-0002-5041-7344

Ronald Cuevas Machaca
Hospital Regional de Moquegua
rzcuevasm@gmail.com
https://orcid.org/0000-0002-3887-6396

Luis Godoy Montoya
Universidad Jorge Basadre Grohman
luis.godoy.montoya@gmail.com
https://orcid.org/0000-0001-8860-8843

Heber Rodríguez Puma
Universidad Jorge Basadre Grohman
wisapuma@gmail.com
https://orcid.org/0000-0003-1779-5738

ARK: ark:/42411/s6/a45
PURL: 42411/s6/a45

RECIBIDO 25/05/2021 • ACEPTADO 02/07/2021 • PUBLICADO 30/09/2021

RESUMEN

En este trabajo se realizó el análisis de la información producto de la atención de pacientes en el servicio de consulta externa. Se han revisado trabajos que guardan relación con las metodologías posibles de utilizar, antes de la elección de una en particular. Posteriormente, se ha justificado y aplicado la metodología de regresión logística para evaluar, clasificar y pronosticar los resultados esperados conforme al objetivo trazado. En el Hospital Regional de Moquegua, desde el inicio de la emergencia sanitaria por el Covid-19, se suspendió la atención en el servicio de consulta externa, vale decir desde Marzo del 2020 a Junio 2021 no se tiene información de cuánto hubiese sido la demanda por especialidad en dicho servicio. El objetivo del trabajo es predecir, en base a variables de edad y sexo, la cantidad de pacientes de sexo femenino que solicitarán una cita para las especialidades de consulta externa, en un período de tiempo. Para la resolución del objetivo planteado, se aplicó el modelo de regresión logística de scikit-learn que, en un inicio ha permitido clasificar y determinar el grupo de importancia en base al cual está orientado nuestro objetivo, tomando como variables independientes y relevantes: el sexo y la edad. Los resultados iniciales obtenidos del procedimiento del modelo no mostraron correspondencia real a la predicción esperada. Las conclusiones determinan que el modelo propuesto requiere la inclusión de otras variables de entrada.

Palabras claves: atenciones médicas, covid-19, predicción, regresión logística.

ABSTRACT

In this work, the analysis of the information produced by the care of patients in the outpatient service was carried out. Studies have been reviewed that are related to the possible methodologies to be used, before choosing one in particular. At the Regional Hospital of Moquegua, since the beginning of the health emergency due to Covid-19, care in the outpatient service was suspended, that is, from March 2020 to June 2021 there is no information on how much the demand would have been by specialty in said service. The objective of the work is to predict, based on age and sex variables, the number of female patients who will request an appointment for outpatient specialties, in a period of time. To solve the problem, the logistic regression technique was used, which initially allowed us to classify and determine the importance group on the basis of which our objective is oriented, taking sex and age as relevant variables. The results obtained from the initial procedure of the model did not show real correspondence to the expected prediction. The conclusions determine that the proposed model requires the inclusion of other input variables.

Keywords: *medical care, covid-19, prediction, logistic regression*.

INTRODUCCIÓN

En los hospitales del país, el servicio de consulta externa, es el segundo nivel de atención después de los puestos y centros de salud. La diferencia en la prestación de servicios está dada por la atención especializada, por profesionales de la salud que han profundizado en el estudio especializado relativo a un área específica del cuerpo humano, a técnicas quirúrgicas específicas o a un método diagnóstico determinado.

El servicio de consulta externa del Hospital Regional de Moquegua es la unidad orgánica encargada de sistematizar la atención integral de la salud y la referencia y contrarreferencia de los pacientes nuevos y/o continuadores, a los cuales el Hospital venía atendiendo de forma continua hasta la declaratoria de la emergencia sanitaria en el país por la pandemia del Covid-19. Al servicio de consulta externa concurren pacientes con diferentes características, como tipos de seguro, sexo, etnia, de forma continua o por primera vez, tanto nacionales como extranjeros. De la información publicada en el Boletín Estadístico del Hospital – 2019 [1], se registra un total de 50,659 atenciones que corresponde al servicio de consulta externa, de las cuales 19,958 fueron de sexo masculino y 30,701 de sexo femenino.

Desde el 16 de marzo de 2020, fecha que entra en vigencia la cuarentena a consecuencia de la emergencia sanitaria por el Covid-19 en el Perú, se suspendió la atención total en el servicio de consulta externa del Hospital Regional de Moquegua. Los pacientes asegurados y no asegurados

dejaron de recibir atención médica en un espacio de tiempo particular, dado que de enero a marzo abarca la estación de verano y el período vacacional para la mayoría de trabajadores y estudiantes del país.

De los pacientes atendidos en consulta externa, un porcentaje asiste por primera vez al hospital o regresan después de períodos prolongados de tiempo, que pueden ser meses o años. Otro conjunto de pacientes asiste continuamente al hospital ya sea para seguir un tratamiento prolongado o por nuevas dolencias derivadas o no de su enfermedad principal. En conjunto, todos los pacientes que se atendían por consulta externa han dejado de recibir atención por un lapso aproximado de 16 meses, lo que posiblemente habría provocado un mayor deterioro en la salud de la población. Otro hecho derivado de esta situación, luego de iniciada la reactivación económica, es la probable concurrencia de los pacientes a Instituciones Prestadoras de Servicios de Salud (IPRESS) privadas, lo que habría involucrado una afectación mayor a su economía personal o familiar, considerando los costos diferenciados con respecto a IPRESS públicas como el Hospital Regional de Moquegua. Estos hechos deben estar siendo evaluados por el máximo ente rector del sector salud en Perú, el Ministerio de Salud, así como por las demás autoridades en los distintos niveles de gobierno y el propio hospital, entonces es importante conocer cuál será la probable demanda futura en el servicio de consulta externa para las distintas especialidades ofertadas.

Con el avance y aplicación de la Inteligencia Artificial como herramienta para predecir, agrupar o clasificar grandes cantidades de datos, es la tecnología elegida para la resolución del problema planteado en el presente trabajo. De sus técnicas desarrolladas, es la regresión logística, por su aplicación al aprendizaje automático para clasificación, considerada como una red neuronal en miniatura y dado que ampliamente ha demostrado que funciona muy bien cuando hay muchísimos datos y las interrelaciones entre ellos no son muy complejas, sirve como sustento para priorizar su uso y aplicación.

De similar importancia son las Redes Neuronales, que más allá de imitar el funcionamiento de las redes neuronales de los organismos vivos. Se basan en una idea sencilla: dados unos parámetros hay una forma de combinarlos para predecir un cierto resultado. En suma, son un modelo para encontrar esa combinación de parámetros y aplicarla al mismo tiempo. En el lenguaje propio, encontrar la combinación que mejor se ajuste es "entrenar" la red neuronal. Una red ya entrenada se puede usar luego para hacer predicciones o clasificaciones, es decir, para "aplicar" la combinación.

En consecuencia, el objetivo del trabajo es predecir, en base a las variables de edad y sexo de los pacientes, cuántas mujeres, niñas o adultas, se atenderán en una especialidad de consulta externa, considerando un periodo de tiempo específico. Es necesario justificar nuestro enfoque en el grupo de interés señalado, principalmente porque representa el 60.60 % de la totalidad de

pacientes atendidos en el último año (2019) del período considerado para nuestro análisis (2015-2019), vale decir dos tercios de la población atendida en el año previo a la pandemia del Covid-19. Para este propósito, se ha se hecho un análisis de otros trabajos relacionados y su utilidad respecto a sus objetivos, éstos abarcan a áreas como la educación, enfocado en el rendimiento académico; también, al área de los negocios y el riesgo de quiebra por variables financieras; así como, el comportamiento de los afiliados a una organización profesional y el cumplimiento con sus obligaciones económicas. En conjunto, todos estos trabajos tienen una variable común, el tiempo.

Trabajos relacionados

En la investigación propuesta en [2] El objetivo de este estudio es evaluar la capacidad de la regresión lineal y de la regresión logística en la predicción del rendimiento y del éxito/fracaso académico, partiendo de variables, como la asistencia y la participación en clase, cuya relevancia ya ha sido puesta de manifiesto en anteriores trabajos de nuestro equipo (Alvarado y García Jiménez, 1997). La muestra la constituyeron 175 universitarios de primero de psicología, tomándose los datos en la asignatura de «Métodos y Diseños de Investigación en Psicología I», del área de Metodología. Las conclusiones de este estudio son que (a) el rendimiento previo es un buen predictor del rendimiento futuro y (b) la asistencia y sobre todo la participación son variables con un peso importante en la predicción del rendimiento. Sin embargo, esta investigación tiene el inconveniente que el *dataset* está constituido por un reducido conjunto de datos a comparación del *dataset* que se utiliza en el presente trabajo.

En [3] se explora un método para realizar la predicción de la tendencia de cierre del indicador S&P 500 con un horizonte de pronóstico de 1 día, adecuando el problema de interés a un problema de clasificación binaria; asignando 1 si la predicción del indicador es creciente y 0 si es decreciente. Al final, se evalúan algunas pruebas de hipótesis para establecer conclusiones sobre la estacionalidad de la serie temporal, analizando los resultados más relevantes obtenidas en las implementaciones, de los cuales destacan niveles de exactitud de 52.51% y 64.04% para los modelos LSTM y Regresión Logística respectivamente.

En la investigación [4] se puso como objeto el desarrollo de un modelo de predicción de riesgo de quiebra con base en la metodología de regresión logística, para las micro y pequeñas empresas del sector comercial del Ecuador, identificando como factores influyentes las razones financieras de liquidez, solvencia, actividad, endeudamiento y rentabilidad, así también, las variables de edad y tamaño de las empresas. Se identifica cuál de estos factores, son los que mayor impacto generan en la estabilidad.

Se concluye que el modelo propuesto en la investigación permite medir de una forma aceptable el nivel de probabilidad de riesgo de quiebra al que se expone una empresa comercial del Ecuador, logrando un 69.76% y 100% a tres y un año antes de que el fracaso ocurra.

En el trabajo [5] se presenta una comparación de indicadores de rendimiento del modelo de deserción actual de la Universidad del Bío-Bío (UBB), el cual está basado en la técnica de regresión logística y se compara con un nuevo modelo basado en árboles de decisión. El nuevo modelo se obtiene a través de metodologías de minería de datos y fue implementado a través de la herramienta SAP Predictive Analytics. Para entrenar, validar y aplicar el modelo se dispone de información real de las bases de datos de la Universidad de Bio Bio -UBB.

El modelo propuesto obtiene una exactitud del 86%, una precisión del 97% con un porcentaje de error del 14% en la predicción de la deserción estudiantil, muy superior al valor que entrega el modelo basado en regresión logística. Posteriormente, el modelo de predicción obtenido es optimizado considerando para ello otras variables logrando de este modo mejoras en los indicadores de predicción.

En el trabajo [6] se comparó una técnica de aprendizaje automático y una técnica clásica, con el objetivo de determinar qué técnica es más eficiente en la predicción de la morosidad de cuotas sociales en el Colegio de Ingenieros del Perú Consejo Departamental de Lambayeque. Las técnicas a comparar seleccionadas fueron máquina de soporte vectorial y regresión logística. Para efectuar la comparación de las técnicas se dispone de datos históricos de los colegiados cuya información se obtuvo de fuentes internas y externas a la organización. Posteriormente los datos recopilados por medio del proceso de extracción, transformación y carga (ETL) se limpió y estandarizó obteniéndose datos concisos y relevantes. Finalmente se aplicó las técnicas predictivas cuyos resultados son favorables para la máquina de soporte vectorial en comparación con la regresión logística.

Concluyendo que la técnica máquina de soporte vectorial es más eficiente para predecir la morosidad de cuotas sociales en el Colegio de Ingenieros del Perú Consejo Departamental de Lambayeque.

En la investigación [7] se estableció como objetivo el comparar el Modelo de regresión lineal Múltiple frente al Árbol de regresión, para ello se utilizó las variables Precio máximo de las acciones de Intel en función al Precio de apertura y Volumen de ventas, de acciones por día. El *dataset* está conformado por todas las acciones de la empresa Intel desde su creación y a través del tiempo; se empleó el muestreo no probabilístico por conveniencia, se consideró desde mayo del 2018 hasta octubre del 2019 siendo un total de 410 registros recopilados a partir de la revisión documentaria. Las pruebas estadísticas usadas fueron el Análisis de regresión lineal múltiple y los Árboles de regresión. Los resultados obtenidos fueron; el Modelo de Regresión lineal múltiple con la técnica de eliminación de datos atípicos queda definida por la siguiente ecuación Y=0.02856+1.003X_1+0.000000009405X_2. Alcanzando una prueba F significativa y la bondad de ajuste es bastante alta R^2=0.9979, y un Error Estándar Residual de 0.2257 dólares, El Árbol de regresión establece que la variable para explicar el Precio máximo de acciones es el Precio de apertura, eliminando la variable volumen, el Error Medio Cuadrático es de 1.4480 dólares.

Finalmente se concluye que el mejor modelo para predecir el precio máximo de acciones de Intel es el modelo de Regresión Lineal Múltiple con eliminación de puntos Outliers.

MATERIALES Y MÉTODOS O METODOLOGÍA COMPUTACIONAL

En la actualidad nos vemos inmersos en la denominada era de la información, en la que el conocimiento es un gran activo para las compañías, sobre todo cuando se trata de conocer más al público objetivo únicamente basándose en información que se genera como resultado del usos de servicios que de alguna forma son almacenados en sistemas de información como datos históricos, datos que con las herramientas y estrategias adecuadas pueden servir para identificar hábitos de consumo, preferencias y costumbres ya sea en la gran red Internet como en el uso de servicios como supermercados, tiendas de consumo, grifos, entidades financieras, etc.

Toda la información que se produce diariamente y en cada instante influye en gran medida para orientar campañas de publicidad, campañas de lanzamiento de nuevos productos, habiendo predicho de antemano la respuesta del público. Este universo de conocimiento es el insumo del cual se nutre la Ciencia de Datos mediante la técnica del *Machine Learning*, logrando con gran aproximación identificar comportamientos, tendencias, patrones a lo largo de periodos de tiempo que con mucha probabilidad se volverían a presentar en un futuro próximo. En resumen, es así como se genera la predicción sobre el conjunto de datos.

En el presente trabajo de investigación se empleó como herramienta principal el entorno web GOOGLE COLABORATORY, adicionalmente se emplearon librerías para la implementación de algoritmos de *machine learning* (para el presente trabajo, librerías que permitan implementar algoritmos de árbol de decisiones: pandas, numphy, keras, matplotlib) en el lenguaje de programación Python sobre la plataforma de cuaderno de notas (Notebook) de Jupiter (Jupiter Note Book).

Google Colaboratory. Colaboratory, también llamado "Colab", permite ejecutar y programar en Python en nuestro navegador o como se conoce actualmente en la “nube” y tiene las siguientes ventajas:

- No requiere configuración
- Da acceso gratuito a GPUs
- Permite compartir contenido fácilmente

Machine learning. Es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos.

Regresión Logística. Es una técnica de aprendizaje automático que proviene del campo de la estadística.

Es uno de los algoritmos más utilizados actualmente en aprendizaje automático. Este algoritmo tiene como principal aplicación los problemas de clasificación binaria. Dada su simplicidad, en el que se pueden interpretar fácilmente los resultados obtenidos e identificar por qué se obtiene un resultado u otro. A pesar de su simplicidad funciona realmente bien en muchas aplicaciones y se utiliza como referencia para pruebas de rendimiento respecto a otros algoritmos.

Python. En la actualidad como lenguaje de programación interpretado goza de gran preferencia por los programadores, entusiastas y todo aquel que se ve atraído por temas como inteligencia artificial, robótica, procesamiento de imágenes, etc. Este lenguaje da mucha importancia a la legibilidad de su código. Lenguaje de programación multiparadigma, soporta parcialmente la orientación a objetos, programación imperativa y, en menor grado, la programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma, pudiéndose ejecutar en sistemas operativos como Mac OS, Windows y Linux.

Pandas. Es una librería de las muchas que han sido desarrolladas por la gran legión de programadores de Python que la mantienen, está orientada al manejo especializado y análisis de estructuras de datos. Soporta archivos en formatos CSV, Excel y base de datos SQL.

Keras. Es una biblioteca de Redes Neuronales de Código Abierto escrita en Python. Es capaz de ejecutarse sobre TensorFlow, Microsoft Cognitive Toolkit o Theano, siempre de manera transparente hacia el usuario, sin que este tenga que preocuparse de nada.

Ha sido diseñada para hacer posible la experimentación en muy poco tiempo con redes de Aprendizaje Profundo.

Matplotlib. Es una librería de Python especializada en la creación de gráficos en dos dimensiones.

Permite crear y personalizar los tipos de gráficos más comunes, entre ellos:

Diagramas de barras

- Histograma
- Diagramas de sectores
- Diagramas de caja y bigotes

- Diagramas de violín
- Diagramas de dispersión o puntos
- Diagramas de líneas
- Diagramas de áreas
- Diagramas de contorno
- Mapas de color

y combinaciones de ellos.

Seaborn. Es una librería desarrollada para Python, dotando a este lenguaje de la posibilidad de generar fácilmente elegantes gráficos. Seaborn está basada en matplotlib y proporciona una interfaz de alto nivel con una carva rápida de aprendizaje. Dada su gran popularidad se encuentra instalada por defecto en la distribución Anaconda, y también puede ser importada en la plataforma Google COLAB.

Scikit-learn. Es una biblioteca desarrollada para Python orientada al aprendizaje automático de software libre. Esta librería incluye varios algoritmos de clasificación, regresión y análisis de grupos como ser: máquinas de vectores de soporte, bosques aleatorios, Gradient boosting, K-means y DBSCAN. Así también está diseñada para interoperar con las bibliotecas numéricas y científicas NumPy y SciPy.

Numpy. Es una librería de Python especializada en el cálculo numérico y el análisis de datos, sobre todo grandes volúmenes de datos.

Esta librería, incorpora una clase de objetos llamados arrays que permite representar colecciones de datos de un mismo tipo en varias dimensiones, y funciones muy eficientes para su manipulación.

SciPy. Es una biblioteca libre y de código abierto para Python. Se compone de herramientas y algoritmos matemáticos. SciPy contiene módulos para optimización, álgebra lineal, integración, interpolación, funciones especiales, FFT, procesamiento de señales y de imagen, resolución de ODEs y otras tareas para la ciencia e ingeniería.

Procedimientos

Se dispone de un Dataset en formato csv (BD_SALUD_ESPECIALIDADES.csv), correspondiente a las citas para atención en un establecimiento de salud tipo Hospital ubicado en la Región de Moquegua, citas que corresponden a las atenciones generadas durante un periodo de 05 años.

A continuación, se resumen los procedimientos seguidos en el presente trabajo:

- Primeramente, estando en un nuevo proyecto de Jupyter notebook en Google Colab, se importará el data set arriba mencionado.
- A continuación de verifica los nombres de las columnas (features) el contenido (filas) del *dataset* importado.
- Se realizan tareas de tratamiento de datos: conversión de dato del tipo object a int64 (sx_valor y edad_menor)
- Se generan totalizados mediante comandos de agrupamiento a fin de verificar la cantidad de registros vinculados con las variables de nuestro interés (sexo y especialidad).
- Se procede a presentar la descripción del *dataset*, lo cual nos da una idea de valores como total de registros, valor promedio, desviación estándar, percentil 25%, percentil 50%, percentil 75%, valor mínimo, valor máximo, para cada columna que forma parte del *dataset*.
- Se procede a agrupar el conjunto *dataset* por servicio y sexo mostrando el valor promedio de las columnas que conforman el *dataframe*.
- Se procede a agrupar el conjunto dataset por la columna numérica calculada menor_edad.
- Se genera un gráfico de barra mostrando el número de hombres y mujeres, agrupados en menor_edad y mayor de edad.
- Se genera un gráfico del tipo histograma mostrándonos la distribución de las citas por edades en el rango de 0 a 100 años. Este gráfico permite apreciar los bloques de edades en los que se concentran la mayor cantidad de datos del dataframe.

Preparación del modelo:

- Se crean los conjuntos de datos a X que corresponden a las entradas e Y que corresponden a las salidas esperadas; para lo cual se eliminan las columnas que a nuestro criterio no son significativas para el modelo de regresión lineal.
- Se crea el modelo de regresión lineal, se entrena con un conjunto de datos que corresponde al 80% del *dataframe*, y posteriormente se testea o valida con el restante 20% del conjunto de *dataframe*.
- Luego de la validación, se realizan las predicciones que se consideren necesarias.
- Finalmente, con los resultados obtenidos se procede al análisis y discusión de los mismos.

RESULTADOS Y DISCUSIÓN

El modelo se basa en el análisis de las tuplas generadas por las atenciones realizadas en el servicio de consulta externa del Hospital Regional de Moquegua, se han analizado un total de 171,797 tuplas, que abarcan desde el año 2015 al 2019. Cada tupla contiene 11 columnas con datos relacionados a la atención, de entre las cuales se definió las variables de entrada del modelo propuesto. La variable Sexo es relevante dado que permitió filtrar a los pacientes atendidos en dos grupos diferenciados y determinar la prioridad de uno de ellos en particular: el femenino. La variable Edad es importante porque asociada a la variable anterior nos mostró con mayor precisión el enfoque hacia donde debíamos orientar el modelo propuesto. La Especialidad Médica se considera en el modelo por ser la variable que contiene el propósito de concurrencia de un paciente al servicio de consulta externa.

En la figura 1, se muestra un cuadro de barras que representan los totales de pacientes atendidos registrados, para nuestro caso podemos observar la cantidad de pacientes femeninos menores de edad.

En la figura 2, visualizamos en formato de historial los tres Features o características de entrada con nombres “edad”, “menor_edad” y “sx_valor” podemos ver gráficamente entre qué valores se comprenden sus mínimos y máximos y en qué intervalos concentran la mayor densidad de registros.

En la siguiente tabla 1, complementando con información consolidada se tiene que el promedio de edad de los pacientes es de 39 años, que representa un 56% de la población en estudio y así mismo considerar que la edad mínima del paciente es de 0 años y la máxima de 99 años.

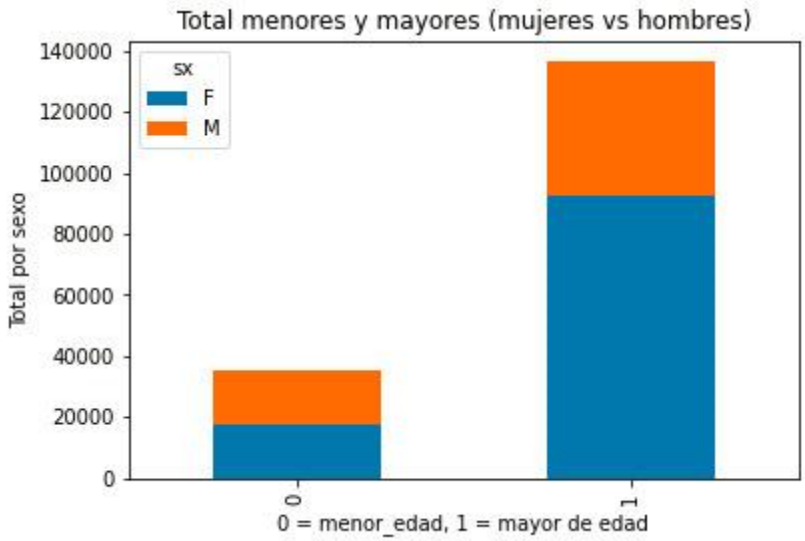

Figura 1. Cantidad de pacientes de ambos sexos y su agrupamiento por edad.

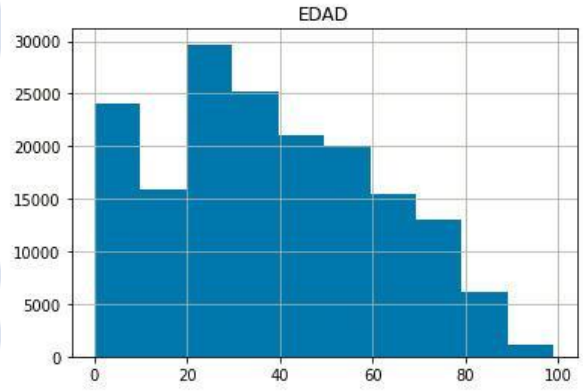

Figura 2. Agrupamiento del total de pacientes de grupos etarios por décadas.

Para un mejor procesamiento de los datos, se ha añadido al *dataset* inicial dos columnas: "menor_edad", la cual tomará como valor 0 si es menor a 18 años o 1 si es mayor de edad; así mismo, se agregó la columna "sx_valor" en la cual se transforma numéricamente los valores de sexo femenino y masculino en 0 y 1 respectivamente. Con la estructura de la data actualizada se procedió a hacer una evaluación por sexo y edad de la información, dando como resultado que un total de 17,796 son de sexo femenino y menores de edad, y 92,798 también de sexo femenino pero mayores de edad, como datos de importancia para el modelo. En un primer resultado, luego de evaluar la interrelación de las entradas de a pares, para ver cómo se concentran linealmente las salidas de especialidades médicas por colores, es decir las variables: edad, "menor_edad" que determina si es mayor o menor de edad un paciente, "sx_valor" que agrupa a los mismos en

femenino o masculino, se ha podido evidenciar su concentración en base a la clasificación por especialidad médica, representada en la variable "desc_servs".

Tabla 1. Resumen del reporte consolidado.

```
dataframe.describe()
```

	ed	EDAD	cod_servsa	menor_edad	sx_valor
count	36788.000000	36788.000000	36788.000000	36788.000000	36788.000000
mean	132.754159	38.724856	304603.051104	0.796673	0.367973
std	36.471070	23.675871	15414.122998	0.402479	0.482261
min	1.000000	0.000000	300101.000000	0.000000	0.000000
25%	120.000000	21.000000	301607.000000	1.000000	0.000000
50%	137.000000	37.000000	302401.000000	1.000000	0.000000
75%	156.000000	56.000000	303510.000000	1.000000	1.000000
max	199.000000	99.000000	410104.000000	1.000000	1.000000

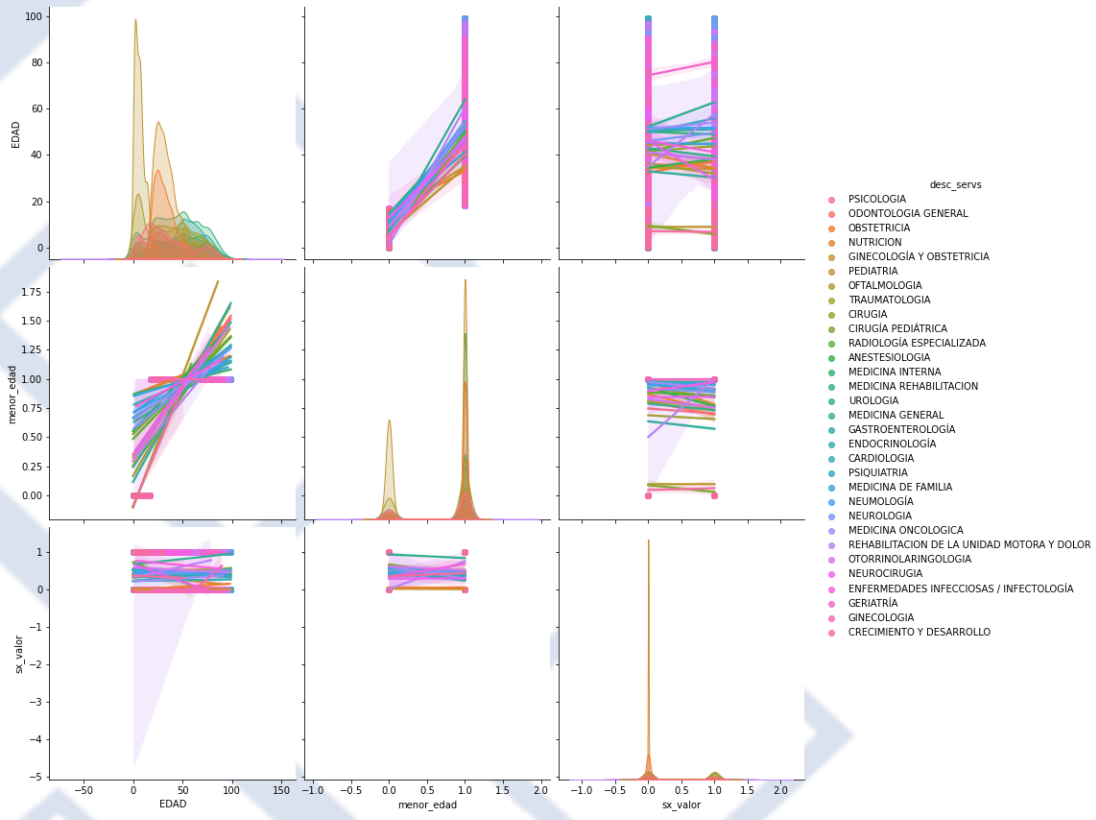

Figura 3. Clasificación por especialidad médica con la interrelación de variables de entrada por sexo y edad.

Luego de la aplicación del método, como se puede apreciar en la figura 4, los resultados del entrenamiento del modelo han obtenido el 23.67% de proximidad o confiabilidad en el modelo de la predicción inicial. Esto debido a las variables intervinientes edad, sx_valor (sexo de tipo numérico), menor_edad. Lo que indica que nuestro modelo aún es perfectible. Para una mayor precisión del modelo, se podrían considerar otras variables adicionales que tengan relación con la variable determinante o predictora.

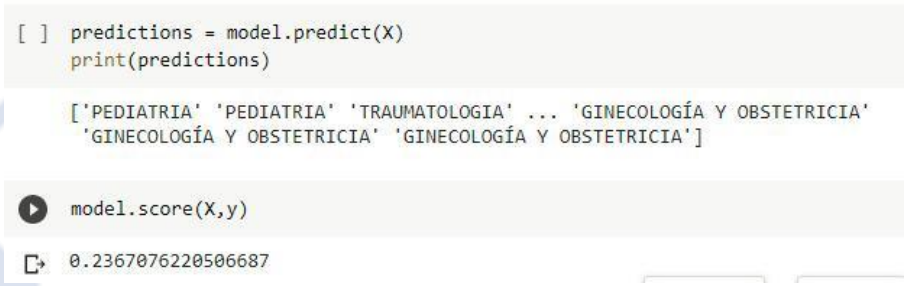

Figura 4. Resultado de la predicción inicial.

La matriz de confusión es una herramienta muy útil para valorar cómo de bueno es un modelo de clasificación basado en aprendizaje automático. En particular, sirve para mostrar de forma explícita cuándo una clase es confundida con otra, lo cual nos permite trabajar de forma separada con distintos tipos de error.

Para la evaluación del modelo, como se muestra en la figura 5, vamos a echar mano de la matriz de confusión. Para ello, dividimos el *dataset* en dos partes. Dejamos un 80% de los datos como datos de entrenamiento (train), y reservamos el 20% restante como datos de prueba (test). A continuación, entrenamos el modelo de nuevo, pero ahora sólo con los datos de entrenamiento.

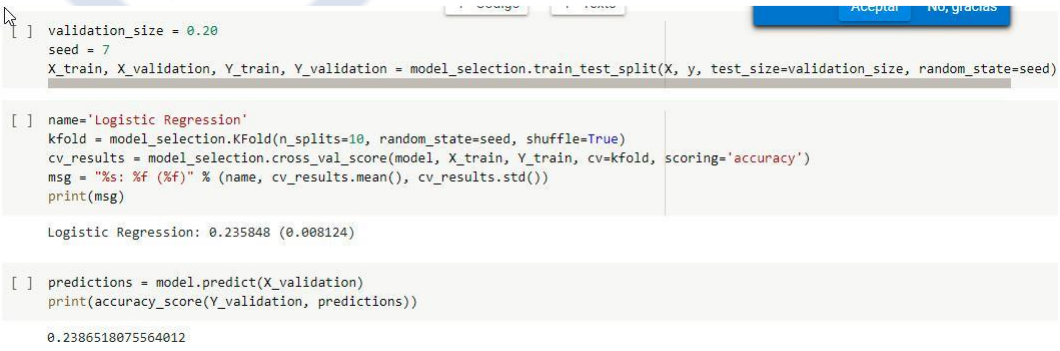

Figura 5. Resultado de la validación del modelo de regresión logística.

En consecuencia, como se observa en la tabla 2, el resultado de la regresión logística arroja un 23.58% de confiabilidad que indica que el porcentaje de precisión del modelo es muy bajo o poco confiable.

Los datos obtenidos en la matriz de confusión, como se muestran en la siguiente tabla 2, los valores están muy por debajo de la predicción acertada, esto debido a que el modelo de regresión logística tiene una validez de 23.67%, considerando las variables de entrada: sx_valor, edad y menor_edad versus las especialidades médicas.

Tabla 2. Matriz de confusión.

[[	0	0	0	0	0	0	0	0	199	0	0	68	0	0	0	0	0	0	0	0	0	0	36	0	0	0	3	19]	['PSICOLOGIA'
[	0	3	0	0	0	0	0	0	126	0	0	137	0	0	0	0	0	0	0	0	0	0	11	0	0	0	5	51]	'ODONTOLOGIA GENERAL'
[	0	3	0	0	0	0	0	0	193	0	0	165	0	0	0	0	0	0	0	0	0	0	8	0	0	0	27	30]	'OBSTETRICIA'
[	0	0	0	0	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	52	0	0	0	0	0]	'NUTRICION'
[	0	0	0	0	0	0	0	0	152	0	0	126	0	0	0	0	0	0	0	0	0	0	24	0	0	0	3	21]	'GINECOLOGÍA Y OBSTETRI
[	0	0	0	0	0	0	0	0	1	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1]	'PEDIATRIA'
[	0	2	0	0	0	0	0	0	111	0	0	120	0	0	0	0	0	0	0	0	0	0	11	0	0	0	9	28]	'OFTALMOLOGIA'
[	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3]	'TRAUMATOLOGIA'
[	0	0	0	0	0	0	0	0	822	0	0	55	0	0	0	0	0	0	0	0	0	0	18	0	0	0	0	0]	'CIRUGIA'
[	0	0	0	0	0	0	0	0	15	0	0	10	0	0	0	0	0	0	0	0	0	0	3	0	0	0	1	1]	'CIRUGÍA PEDIÁTRICA'
[	0	0	0	0	0	0	0	0	42	0	0	52	0	0	0	0	0	0	0	0	0	0	117	0	0	0	4	7]	'RADIOLOGÍA ESPECIALIZA
[	0	5	0	0	0	0	0	0	222	0	0	266	0	0	0	0	0	0	0	0	0	0	14	0	0	0	17	65]	'ANESTESIOLOGIA'
[	0	0	0	0	0	0	0	0	11	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0]	'MEDICINA INTERNA'
[	0	4	0	0	0	0	0	0	119	0	0	149	0	0	0	0	0	0	0	0	0	0	122	0	0	0	14	33]	'MEDICINA REHABILITACIC
[	0	0	0	0	0	0	0	0	16	0	0	36	0	0	0	0	0	0	0	0	0	0	2	0	0	0	2	7]	'UROLOGIA'
[	0	0	0	0	0	0	0	0	6	0	0	2	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0]	'MEDICINA GENERAL'
[	0	0	0	0	0	0	0	0	27	0	0	38	0	0	0	0	0	0	0	0	0	0	5	0	0	0	5	14]	'GASTROENTEROLOGÍA'
[	0	0	0	0	0	0	0	0	8	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0]	'ENDOCRINOLOGÍA'
[	0	0	0	0	0	0	0	0	369	0	0	57	0	0	0	0	0	0	0	0	0	0	8	0	0	0	7	0]	'CARDIOLOGIA'
[	0	0	0	0	0	0	0	0	181	0	0	85	0	0	0	0	0	0	0	0	0	0	111	0	0	0	23	22]	'PSIQUIATRIA'
[	0	0	0	0	0	0	0	0	46	0	0	101	0	0	0	0	0	0	0	0	0	0	48	0	0	0	5	24]	'MEDICINA DE FAMILIA'
[	0	0	0	0	0	0	0	0	7	0	0	12	0	0	0	0	0	0	0	0	0	0	6	0	0	0	2	3]	'NEUMOLOGÍA'
[	0	0	0	0	0	0	0	0	34	0	0	15	0	0	0	0	0	0	0	0	0	0	504	0	0	0	19	0]	'NEUROLOGIA'
[	0	3	0	0	0	0	0	0	129	0	0	104	0	0	0	0	0	0	0	0	0	0	76	0	0	0	10	37]	'MEDICINA ONCOLOGICA'
[	0	2	0	0	0	0	0	0	54	0	0	45	0	0	0	0	0	0	0	0	0	0	9	0	0	0	9	4]	'REHABILITACION DE LA U
[	0	0	0	0	0	0	0	0	61	0	0	57	0	0	0	0	0	0	0	0	0	0	17	0	0	0	4	11]	'OTORRINOLARINGOLOGIA'
[	0	3	0	0	0	0	0	0	144	0	0	193	0	0	0	0	0	0	0	0	0	0	243	0	0	0	25	38]	'NEUROCIRUGIA'
[	0	1	0	0	0	0	0	0	37	0	0	71	0	0	0	0	0	0	0	0	0	0	8	0	0	0	10	136]]	'ENFERMEDADES INFECCIOS
																													'GERIATRÍA']

En los resultados obtenidos, se observa que el modelo en su primer entrenamiento, y luego de la validación con un segundo entrenamiento no retorno valores esperados en la predicción. Se reanalizará el modelo, incorporando nuevas variables de entrada

CONCLUSIONES

Se ha logrado construir un modelo de clasificación que ayuda a predecir la cantidad de especialidades médicas, según su sexo y edad de pacientes, que podría ser requeridas en el servicio de consulta externa, para un período de tiempo en el Hospital Regional de Moquegua.

La matriz de confusión como herramienta para valorar cómo de bueno es un modelo de clasificación basado en aprendizaje automático. En particular, nos sirvió para mostrar de forma explícita cuando una clase o especialidad médica es confundida con otra, lo cual nos permite trabajar de forma separada con distintos tipos de error.

Las variables de entrada, seleccionadas para el modelo, hasta el momento, no muestran correlación respecto a la variable de salida: especialidad médica. En ese sentido, se sugiere la incorporación de otras variables de entrada que nos permitan mejorar el porcentaje de predicción del modelo propuesto.

REFERENCIAS

[1] "Estadísticas" Boletín Estadístico 2019, [online document], 2019. Disponible: Web Hospital Regional de Moquegua, http://www.hospitalmoquegua.gob.pe.

[2] Jiménez, M. V. G., Izquierdo, J. M. A., & Blanco, A. J. (2000). La predicción del rendimiento académico: regresión lineal versus regresión logística. Psicothema, 12(Su2), 248-525.

[3] Guzmán Aristizábal, S. M., & Hurtado Franco, J. C. (2021). Predicción de la tendencia del indicador S&P 500.

[4]. Erazo Garzón, J. F. (2019). Desarrollo de un modelo de predicción de riesgo de quiebra empresarial para el sector comercial del Ecuador: un enfoque de regresión logística (Doctoral dissertation, Universidad Autónoma de Nuevo León).

[5] Pérez, A., Grandón, E. E., Caniupán, M., & Vargas, G. (2013). Análisis Comparativo de Técnicas de Predicción para Determinar la Deserción Estudiantil: Regresión Logística vs Árboles de Decisión. Arquitectura, 2014, 2015.

[6] Huamán Bernilla, J. N. (2020). Comparación de máquina de soporte vectorial y regresión logística en la predicción de morosidad de cuotas sociales del colegio de ingenieros del Perú consejo departamental Lambayeque.

[7] Maydana Huanca, A. R. (2021). Elección del mejor modelo entre regresión lineal múltiple y árboles de regresión para predecir el precio máximo de las acciones de Intel en función al precio de apertura y volumen de ventas de acciones por día-2019.

Predicción de hipertensión arterial a través de un sistema de regresión logística

Prediction of arterial hypertension through a logistic regression system

Cynthia Mayumi Tesillo Gomez
Universidad Nacional Jorge Basadre Grohmann
cynthia.tesillo@unjbg.edu.pe
https://orcid.org/0000-0002-1769-9845

Yuri Alexander Escobar Arcaya
Universidad Nacional Jorge Basadre Grohmann
yuri.escobar@unjbg.edu.pe
https://orcid.org/0000-0001-5220-058X

Edwin Daniel León Gutierrez
Universidad Nacional Jorge Basadre Grohmann
edwin.leon@unjbg.edu.pe
https://orcid.org/0000-0002-2519-1785

ARK: ark:/42411/s6/a44
PURL: 42411/s6/a44

RECIBIDO 23/06/2021 • ACEPTADO 05/08/2021 • PUBLICADO 30/09/2021

RESUMEN

En el Perú y el mundo entero la hipertensión es una enfermedad que puede avanzar sin manifestar ningún síntoma o éstos ser muy leves. Se puede tener hipertensión arterial y no sentir ninguna manifestación, la hipertensión arterial es un serio problema de salud pública en países en desarrollo como el nuestro: según la Encuesta Demográfica y de Salud Familiar de 2017, aunque la prevalencia de hipertensión en personas de 15 años a más se habría reducido de 14,8 % en 2014, a 13,6 %, implica que más de 3 millones de peruanos viven con hipertensión arterial. Por ese motivo nuestro objetivo es el rápido diagnóstico de esta enfermedad silenciosa, en el presente trabajo se utilizó el sistema de regresión logística, para el cual se posee un *dataset* de 5615 registros analizados. Este artículo presenta la posibilidad de detectar una enfermedad como la hipertensión arterial basado en inteligencia artificial, ya que este mal ha ido aumentando en los últimos años. Por ese motivo el objetivo es predecir de manera rápida un posible diagnóstico de hipertensión arterial, para ello se analizó un *dataset* de 5615 registros en la aplicación web Jupyter Notebook, estableciendo 9 variables de entrada y 1 de salida, además se utilizó el sistema de regresión logística, tratamientos de datos *missing* y *outliers*, gráficas de variables, obteniendo como resultado una precisión media aceptable del 87%.

Palabras claves: Hipertensión arterial, Inteligencia Artificial, Regresión logística, Presión arterial.

ABSTRACT

In Peru and the entire world, hypertension is a disease that can progress without showing any symptoms or these being very mild. You can have high blood pressure and not feel any manifestations, arterial hypertension is a serious public health problem in developing countries like ours: According to the 2017 Demographic and Family Health Survey Survey, although the prevalence of hypertension in people aged 15 years and over would have decreased from 14.8% in 2014 to 13.6%, it implies that more than 3 million Peruvians live with high blood pressure. For this reason, our goal is the rapid diagnosis of this silent disease. In the present work, the logistic regression system was used, for which there is a dataset of 5615 analyzed records. This article presents the possibility of detecting a disease such as high blood pressure based on artificial intelligence, since this evil has been increasing in the last years. For this reason, the objective is to quickly predict a possible diagnosis of arterial hypertension, for this, a dataset of 5615 records was analyzed in the Jupyter Notebook web application, establishing 9 input variables and 1 output, in addition, the logistic regression system was used, missing data treatments and outliers, graphs of variables, obtaining as a result an acceptable average precision of 87%.

Keywords: Arterial hypertension, Artificial Intelligence, Blood Pressure, Logistic Regression.

INTRODUCIÓN

Actualmente se estima que en el mundo hay 1130 millones de personas con hipertensión, y la mayoría son de bajos recursos, en donde una de cada cinco personas hipertensas no lo tiene controlado, esta enfermedad es una de las causas principales de muerte prematura en el mundo es por ello que la Organización Mundial de Salud tiene como meta reducir la prevalencia de la hipertensión en un 25% para el año 2025 [1].

En el Perú la hipertensión arterial es una preocupación constante entre los médicos e investigadores, aún más en estas fechas de pandemia, ya que el Ministerio de Salud del Perú estima que el número personas con esta enfermedad aumentaría en un 20% lo que constituye a un problema de salud pública y conlleva a la aparición de nuevas enfermedades [2].

Por ese motivo la importancia del rápido diagnóstico de esta enfermedad silenciosa, la prevención primaria, es decir anticipar la aparición de una enfermedad que como la hipertensión arterial esencial no tiene una causa conocida, no es tarea fácil, sin embargo, hoy es evidente la existencia de factores que aumentan el riesgo de padecer la enfermedad y que deben ser conocidos por la población [3].

Se trabajó con *dataset* externo almacenado en página web de zenodo.org. del 27 de febrero de 2021, Criterios de cambio hipertensión Perú, para aplicar Predicción de hipertensión arterial a través técnica de Regresión Logística para clasificar patrones. A partir de un conjunto de datos de entrada con una variable de salida.

Para la realización del presente trabajo se tuvo que aplicar métodos de limpieza de datos a nuestro *dataset*, como la eliminación de columnas, la evaluación del procesamiento datos *missing* y la eliminación de datos *outliers*. seguidamente en este artículo se presenta la aplicación del sistema de regresión logística la cual consiste en una técnica estadística multivariante que nos permite estimar la relación existente entre una variable dependiente no métrica, en particular dicotómica y un conjunto de variables independientes métricas o no métricas [4].

Lo que se desea lograr en este artículo es predecir si una persona tiene o es propensa a sufrir de hipertensión, para así tener un diagnóstico oportuno o tratamiento adecuado. La realización del presente artículo se divide en diferentes secciones; en primer lugar, mostramos una introducción, la cual nos da una visión general del problema a atacar, luego se ve a detalle los métodos utilizados para poder realizar este trabajo conociendo a profundidad la regresión logística, debido a que esta se utiliza como base. Seguidamente pasamos a los resultados obtenidos, los cuales son satisfactorios obteniendo un 87 % de aciertos, llegando así a la parte final del trabajo expresando las conclusiones.

MATERIALES Y MÉTODOS O METODOLOGÍA COMPUTACIONAL

Herramientas

Google Colaboratory: Es un entorno gratuito de Jupyter Notebook que no requiere configuración y que se ejecuta completamente en la nube.

Google Drive: Es un servicio de alojamiento de archivos que fue introducido por la empresa estadounidense Google el 24 de abril de 2012. Es el reemplazo de Google Docs que ha cambiado su dirección URL, entre otras cosas. Es uno de los sitios de alojamiento más conocidos en el mundo.

Librerías

NumPy: proporciona una estructura de datos universal que posibilita el análisis de datos y el intercambio de datos entre distintos algoritmos. Las estructuras de datos que implementa son vectores multidimensionales y matrices con capacidad para gran cantidad de datos [5].

Pandas: es una de las librerías de python más útiles para los científicos de datos. Las estructuras de datos principales en pandas son Series para datos en una dimensión y *DataFrame* para datos en dos dimensiones. Estas son las estructuras de datos más usadas en muchos campos tales como finanzas, estadística, ciencias sociales y muchas áreas de ingeniería. Pandas destaca por lo fácil y flexible que hace la manipulación de datos y el análisis de datos [5].

scikit learn: es una de estas librerías gratuitas para Python. Cuenta con algoritmos de clasificación, regresión, *clustering* y reducción de dimensionalidad. Además, presenta la compatibilidad con otras librerías de Python como NumPy, SciPy y matplotlib [6].

sklearn.model_selection import train_test_split: nos permite dividir un dataset en dos bloques, típicamente bloques destinados al entrenamiento y validación del modelo (llamemos a estos bloques "bloque de entrenamiento " y "bloque de pruebas" para mantener la coherencia con el nombre de la función) [7].

sklearn.metrics import accuracy_score: En la clasificación de etiquetas múltiples, esta función calcula la precisión del subconjunto: el conjunto de etiquetas predichas para una muestra debe coincidir exactamente con el conjunto de etiquetas correspondiente en y true [8].

sklearn.metrics import classification_report: crea un informe de texto que muestre las principales métricas de clasificación [8].

matplotlib: es una librería de Python especializada en la creación de gráficos en dos dimensiones, como histograma, diagramas de sectores, diagramas de caja y bigotes, diagramas de violín, diagramas de dispersión o puntos, diagramas de líneas, diagramas de áreas, diagramas de contorno y mapas de color [9].

Métodos o Metodología computacional

El método elegido para realizar la investigación es el método Regresión Logística. Por lo tanto, se tratará de investigar el estudio de los principales factores de riesgo de la hipertensión arterial, cómo influyen las características, como un problema de salud.

La investigación se realiza mediante Regresión Logística con variables de entrada y salida, relacionados a la hipertensión, para poder diagnosticar esta enfermedad a partir de sus características clasificando resultados en valores discretos. Para el trabajo se ha creado un archivo hipertension.csv con datos de entrada, para el presente articulo científico se considera nueve características para el diagnóstico de la enfermedad hipertensiva, según referencia del ministerio de salud, variables como: sexo, edad, presión sistólica, presión diastólica, peso, talla, fuma, actividad física y región. como resultado de salida se considera hipertensión.

La regresión logística es el conjunto de modelos estadísticos utilizados cuando se desea conocer la relación entre

- Una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) o con más de dos categorías (regresión logística multinomial).
- Una o más variables explicativas independientes, llamadas covariables, ya sean cualitativas o cuantitativas.

Las covariables cualitativas deben ser dicotómicas, tomando valor 0 para su ausencia y 1 para su presencia. Si la covariable tuviera más de dos categorías debemos realizar una transformación de la misma en varias covariables cualitativas dicotómicas ficticias (variables *dummy*). Al hacer esta transformación cada categoría de la variable entraría en el modelo de forma individual.

Los modelos de regresión logística tienen tres finalidades:

- Cuantificar la importancia de la relación existente entre cada una de las covariables y la variable dependiente.
- Clarificar la existencia de interacción y confusión entre covariables respecto a la variable dependiente (es decir, los *odds* ratio para cada covariable).
- Clasificar individuos dentro de las categorías (presente/ausente) de la variable dependiente.

Los valores posibles de estas ecuaciones varían entre 0 y 1. Un valor cercano a 0 significa que es muy improbable que Y haya ocurrido, y un valor cercano a 1 significa que es muy probable que tuviese lugar.

Como en la regresión lineal cada variable predictora de la ecuación logística tiene su propio coeficiente. Los valores de los parámetros se estiman utilizando el método de máxima verosimilitud que selecciona los coeficientes que hacen más probable que los valores observados ocurran [10].

Para la obtención de un sistema de Predicción de hipertensión arterial, el cual para su desarrollo se utilizaron varias librerías las cuales se tienen que implementar al inicio del sistema. Las cuales se detallan a continuación.

1. **Importar las librerías las cuales nos ayudarán con el procesamiento y tratamiento de datos.**

```
import numpy as np

import pandas as pd

import seaborn as sb

import matplotlib.pyplot as plt

%matplotlib inline

#from matplotlib import cm

plt.rcParams['figure.figsize'] = (16, 9)

plt.style.use('ggplot')

from sklearn.model_selection import train_test_split

from sklearn import linear_model

from sklearn import model_selection

from sklearn.metrics import confusion_matrix

from sklearn.metrics import accuracy_score

from sklearn.metrics import classification_report
```

2. **Importamos el *dataset* almacenado en nuestra nube de Google Drive, para ello se usará la librería Pandas, así mismo se ejecutará encoding='latin-1' para evitar el Error UTF-8.**

```
dataset = pd.read_csv('/content/gdrive/My Drive/hipertension.csv',encoding='latin-1')
```

3. **Imprimimos en pantalla parte del *dataset*, para poder visualizar total las variables (columnas) y sus datos.**

```
dataset.head()
```

Tabla 1: Descripción del dataset.

	id	city	masl	sex	age_years	systolic_bp	diastolic_bp	weight_kg	height_cm	body_mass_index	diabetes_mellitus	dm_treatment
0	3574	Huancayo	3250	Female	23	119	74	58.0	163.0	22.0	No	No
1	1092	Loreto	100	Male	60	110	70	54.0	160.0	21.0	No	No
2	861	Lima	500	Female	38	120	80	65.0	163.0	25.0	No	No
3	835	Lima	500	Female	43	110	80	60.0	157.0	24.0	No	No
4	4654	Hu?nuco	1900	Female	30	95	60	50.0	152.0	22.0	No	No

4. ***Después de un análisis, se concluyó que las variables id, ciudad, masa corporal, diabetes mellitus, tratamiento, enfermedades, años de fumador, años con hipertensión, tratamiento de la hipertensión, msnm, presión arterial máxima antigua/nueva y mínima antigua/nueva no son causas fundamentales para el desarrollo de la variable de salida, o son variables las cuales tiene similitud con otras, por lo que se procedió a eliminarlas.***

dataset = dataset.drop(['id', 'city', 'masl', 'body_mass_index', 'diabetes_mellitus', 'dm_treatment', 'cv_diseases', 'cd_treatment', 'smoking_years', 'hypertension_years', 'hypertension_treatment', 'msnm', 'sist_old', 'diast_old', 'sist_new', 'diast_new'], axis=1)

dataset.head()

Tabla 2: Descripción de las variables finales.

	sex	age_years	systolic_bp	diastolic_bp	weight_kg	height_cm	smoking	physical_activity	region	hypertension_dx
0	Female	23	119	74	58.0	163.0	No	Yes	Mountain	No
1	Male	60	110	70	54.0	160.0	No	No	Jungle	No
2	Female	38	120	80	65.0	163.0	No	Yes	Coast	No
3	Female	43	110	80	60.0	157.0	Yes	No	Coast	No
4	Female	30	95	60	50.0	152.0	No	Yes	Mountain	No

DICCIONARIO DE VARIABLES.

- *sex: SEXO Condición orgánica que se distingue entre hombres y mujeres.*
- *age_years: EDAD Lapso de tiempo que transcurre desde el nacimiento hasta el momento de referencia.*
- *systolic_bp: PRESIÓN SISTÓLICA Es la presión arterial máxima*

- *diastolic_bp: PRESIÓN DIASTÓLICA Es la presión arterial mínima*
- *weight_kg: PESO_kg Cantidad de masa de la persona expresada en kilogramos*
- *height_cm: TALLA_cm Medida de la estatura de la persona expresada en centímetros*
- *smoking: FUMA Condición de la persona donde se detalla si fuma o no*
- *physical_activity: ACTIVIDAD FÍSICA Define si la persona realiza actividad física continuamente*
- *region: REGION Define el lugar donde vive la persona*
- *hypertension_dx: HIPERTENSIÓN Define si la persona es Hipertenso o no*

5. ***Al ser nuestro sistema una regresión logística, necesitamos que nuestros datos sean de tipo numérico, por lo que consultamos qué tipo de variables tenemos:***

   dataset.dtypes

Figura 1: Tipos de datos de las variables.

```
sex                 object
age_years            int64
systolic_bp          int64
diastolic_bp         int64
weight_kg          float64
height_cm          float64
smoking             object
physical_activity   object
region              object
hypertension_dx     object
dtype: object
```

En Python, el tipo de datos de texto se conoce como secuencia de caracteres (string). En Pandas se los conoce como objetos (object). Las secuencias de caracteres pueden contener números y / o caracteres.

6. ***Al ver que hay datos de tipo objetc se optó por reemplazar los valores de las columnas sex, smoking. physical_activity, region, hypertension_dx por valores tipo int o float con el método map y un diccionario entre llaves.***

   dataset['sex'] = dataset['sex'].map({'Male':0,'Female':1})

   dataset['smoking'] = dataset['smoking'].map({'No':0,'Yes':1})

dataset['physical_activity'] = dataset['physical_activity'].map({'No':0,'Yes':1})

dataset['region'] = dataset['region'].map({'Coast':1,'Mountain':2,'Jungle':3})

dataset['hypertension_dx'] = dataset['hypertension_dx'].map({'No':0,'Yes':1})

dataset.head()

Tabla 3: Variables con los mapeos correspondientes.

	sex	age_years	systolic_bp	diastolic_bp	weight_kg	height_cm	smoking	physical_activity	region	hypertension_dx
0	1	23	119	74	58.0	163.0	0.0	1	2	0.0
1	0	60	110	70	54.0	160.0	0.0	0	3	0.0
2	1	38	120	80	65.0	163.0	0.0	1	1	0.0
3	1	43	110	80	60.0	157.0	1.0	0	1	0.0
4	1	30	95	60	50.0	152.0	0.0	1	2	0.0

Se muestra el nuevo dataset con valores numéricos.

7. ***Se realizó el tratamiento de datos missing, para este caso se tomó la decisión de eliminarlos, se opta por esto ya que esos datos pueden introducir errores mayores en los resultados, por otro lado el sistema de regresión logística solo acepta números, y el dataset debe estar con valores vacíos o "NaN"***

   dataset.describe()

Tabla 4: Eliminación de outliers.

	sex	age_years	systolic_bp	diastolic_bp	weight_kg	height_cm	smoking	physical_activity	region	hypertension_dx
count	5615.000000	5615.000000	5615.000000	5615.000000	5418.000000	5414.000000	5483.000000	5615.000000	5615.000000	5495.000000
mean	0.627427	43.463224	112.792698	71.932146	64.344592	156.390469	0.095386	0.483882	1.876759	0.140855
std	0.483533	17.030004	15.844951	11.324560	11.408517	8.592941	0.293774	0.499785	0.463625	0.347904
min	0.000000	18.000000	55.000000	25.000000	30.000000	105.000000	0.000000	0.000000	1.000000	0.000000
25%	0.000000	29.000000	101.000000	65.000000	56.000000	150.000000	0.000000	0.000000	2.000000	0.000000
50%	1.000000	43.000000	112.000000	71.000000	64.000000	156.000000	0.000000	0.000000	2.000000	0.000000
75%	1.000000	55.000000	122.000000	80.000000	71.000000	163.000000	0.000000	1.000000	2.000000	0.000000
max	1.000000	97.000000	200.000000	119.000000	150.000000	185.000000	1.000000	1.000000	3.000000	1.000000

Se observa en la fila count que algunos valores difieren del total de filas del dataset 5615, lo cual se deduce la existencia de datos missing.

dataset = dataset.dropna()

dataset.describe()

Tabla 5: Eliminación de missing.

	sex	age_years	systolic_bp	diastolic_bp	weight_kg	height_cm	smoking	physical_activity	region	hypertension_dx
count	5170.000000	5170.000000	5170.000000	5170.000000	5170.000000	5170.000000	5170.000000	5170.000000	5170.000000	5170.000000
mean	0.634623	43.586654	112.962669	72.061122	64.308897	156.377950	0.097099	0.479884	1.867311	0.146422
std	0.481582	17.090609	15.848137	11.352873	11.446475	8.603421	0.296121	0.499644	0.476772	0.353563
min	0.000000	18.000000	55.000000	25.000000	30.000000	105.000000	0.000000	0.000000	1.000000	0.000000
25%	0.000000	29.000000	102.000000	65.000000	56.000000	150.000000	0.000000	0.000000	2.000000	0.000000
50%	1.000000	43.000000	112.000000	71.000000	64.000000	156.000000	0.000000	0.000000	2.000000	0.000000
75%	1.000000	56.000000	122.000000	80.000000	71.000000	163.000000	0.000000	1.000000	2.000000	0.000000
max	1.000000	97.000000	200.000000	119.000000	150.000000	185.000000	1.000000	1.000000	3.000000	1.000000

El método dropna permite, de una forma muy conveniente, filtrar los valores de una estructura de datos pandas para dejar solo aquellos no nulos.

8. ***Se realizó en tratamiento de datos outliers, para los cual se graficó las variables de entrada y salida plt.show de la librería Matplotlib***

dataset.drop(['hypertension_dx'],1).hist()

plt.show()

Luego de un análisis se determinó la no existencia de datos outliers en las variables de entrada.

dataset.drop(['sex','age_years','systolic_bp','diastolic_bp','weight_kg','height_cm','smoking','physical_activity','region'],1).hist()

plt.show()

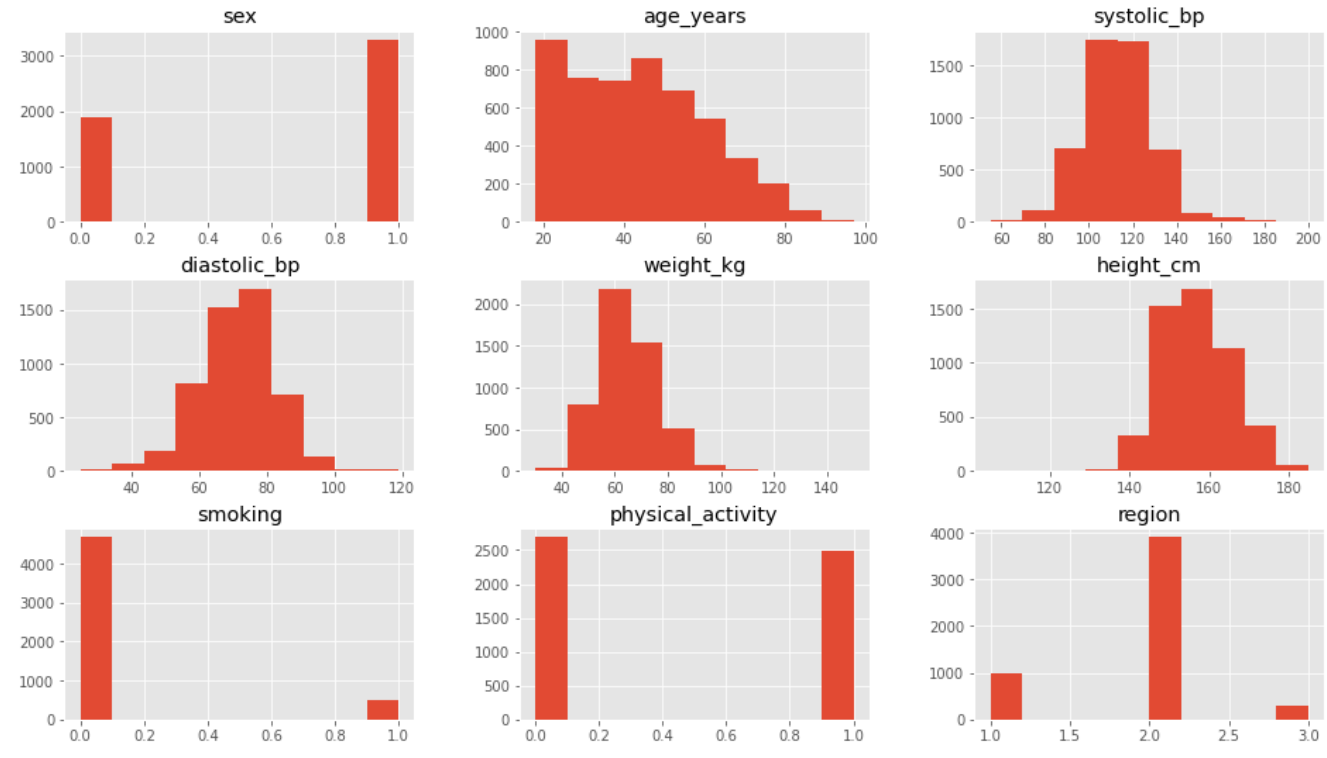

Figura 2: Histograma de las variables de entrada.

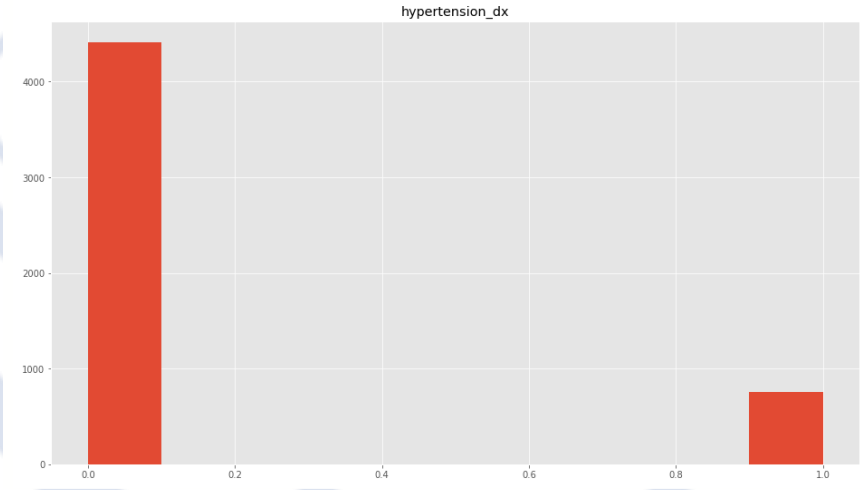

Figura 2: Histograma de la variable de salida.

Luego de un análisis se determinó la no existencia de datos outliers en la variable de salida.

9. ***Creación del modelo de Regresión Logística***

Almacenamos en una matriz las 9 variables de entrada en la variable X y la variable de salida "hypertension_dx" en la variable Y.

```
X = np.array(dataset.drop(['hypertension_dx'],1))
```

```
y = np.array(dataset['hypertension_dx'])
```

```
X.shape
```

Se creó nuestro modelo y se ajustó a nuestro conjunto de entradas X y salidas Y.

```
model = linear_model.LogisticRegression()
```

```
model.fit(X,y)
```

10. ***Una vez compilado nuestro modelo, le hacemos clasificar todo nuestro conjunto de entradas X utilizando el método "predict(X)" y revisamos algunas de sus salidas y vemos que coincide con las salidas reales de nuestro archivo csv.***

```
predictions = model.predict(X)
```

```
print(predictions)
```

11. ***Y confirmamos cuan bueno fue nuestro modelo utilizando model.score() que nos devuelve la precisión media de las predicciones, en nuestro caso del 86.94%.***

```
model.score(X,y)
```

0.8694390715667312

12. ***Validación de nuestro modelo***

Para ello se dividirá el dataset en forma aleatoria en 80% entrenamiento y 20% prueba para la validación.

```
validation_size = 0.20
```

```
seed = 7
```

X_train, X_validation, Y_train, Y_validation = model_selection.train_test_split(X, y, test_size=validation_size, random_state=seed)

RESULTADOS Y DISCUSIÓN

Y ahora hacemos las predicciones -en realidad clasificación- utilizando nuestro "cross validation set", es decir del subconjunto que habíamos apartado. En este caso vemos que los aciertos fueron del 88% que es un resultado aceptable.

predictions = model.predict(X_validation)

print(accuracy_score(Y_validation, predictions))

0.8820116054158608

Dentro de los resultados tenemos la matriz de confusión la cual tenemos que codificar de la siguiente forma:

print(confusion_matrix(Y_validation, predictions))

y esta nos da como resultado:

[[867 14]

[108 45]]

Donde muestra cuántos resultados equivocados tuvo de cada clase (los que no están en la diagonal), en nuestro caso predijo que 108 pasos negativos cuando estos eran positivos y predijo que 14 eran positivos cuando en realidad eran positivos.

También podemos ver el reporte de clasificación con nuestro conjunto de Validación. En nuestro caso vemos que se utilizaron como "soporte" 881 registros negativos y 153 positivos. La valoración que de aquí nos conviene tener en cuenta es la de F1-score, que tiene en cuenta la precisión y *recall*. El promedio de F1 es de 87% lo cual no está nada mal.

print(classification_report(Y_validation, predictions))

	precision	*recall*	*f1-score*	*support*
0.0	*0.89*	*0.98*	*0.93*	*881*

1.0 0.76 0.29 0.42 153

accuracy 0.88 1034

macro avg 0.83 0.64 0.68 1034

weighted avg 0.87 0.88 0.86 1034

Para poder comprobar se colocaron datos de dos filas, una con resultado negativo y otra con resultado positivo, obteniendo la respuesta satisfactoria.

```
#fila 590 valor respuesta 0

X_new = pd.DataFrame({'sex':[1],'age_years':[27],'systolic_bp':[71],'diastolic_bp':[37],'weight_kg':[51],'height_cm':[150],'smoking':[0],'physical_activity':[0],'region':[2]})

model.predict(X_new)


array([0.])


#fila 11 valor respuesta 1

X_new = pd.DataFrame({'sex':[0],'age_years':[58],'systolic_bp':[180],'diastolic_bp':[86],'weight_kg':[61],'height_cm':[162],'smoking':[1],'physical_activity':[1],'region':[2]})

model.predict(X_new)

array([1.])
```

CONCLUSIONES

En conclusión, se puede decir que el presente trabajo tiene un acierto del 88% en sus predicciones de hipertensión arterial a través de un sistema de regresión logística, por lo que puede predecir si una persona tiene o es propensa a sufrir de hipertensión. El uso de regresión logística, se usa normalmente cuando se quiere estimar la relación existente entre una variable dependiente, y un conjunto de variables independientes métricas o no métricas.

REFERENCIAS

[1] Organización Mundial de la Salud. https://www.who.int/es/news-room/fact-sheets/detail/hypertension

[2] Ministerio de Salud el 18 de mayo de 2021 - 2:55 p. https://www.gob.pe/institucion/minsa/noticias/493681-minsa-estima-que-pacientes-con-hipertension-arterial-aumentarian-en-20-durante-la-pandemia

[3] Dr.Raul Gamboa, Revista Peruana de Cardiología: Octubre - Diciembre 1993 https://sisbib.unmsm.edu.pe/bvrevistas/cardiologia/v19_n2/la%20hiper.htm

[4]Celia Mercedes Salcedo Poma, Capitulo 2- Modelo de Regresión Logística https://sisbib.unmsm.edu.pe/bibvirtualdata/Tesis/Basic/Salcedo_pc/enPDF/Cap2.PDF

[5] Jose Martinez Heras-Guía rápida de IArtificial.net https://www.iartificial.net/guia-rapida-iartificial-net/

[6] Universidad de Alcalá, "Scikit-learn, herramienta básica para el data science en Python" 2021. https://www.master-data-scientist.com/scikit-learn-data-science/

[7] Interactive Chaos, 2019. https://interactivechaos.com/es/python/function/sklearnmodelselectiontraintestsplit

[8] Scikit Learn, "Metrics and scoring: quantifying the quality of predictions" https://scikit-learn.org/stable/modules/model_evaluation.html

[9] Aprende con Alf, "La librería Matplotlib" https://aprendeconalf.es/docencia/python/manual/matplotlib/

[10]María Elvira Ferre Jaén.FEIR 45: Regresión logística. Jueves 04 abril 2019,23:30:47 https://gauss.inf.um.es/feir/45/

[11] Failoc-Rojas Virgilio, Dataset externo almacenado zenodo.org. Criterios de cambio hipertensión Perú, datos https://zenodo.org/record/4567767#.YOsdROhKhPZ

Sistema para proponer la nota final de los estudiantes mediante Redes Neuronales

System to propose the final grade of the students through Neural Networks

Kleber Ernesto Baldarrago Salas
Universidad Nacional de San Agustín
kbaldarrago@unsa.edu.pe

Erika Cayllahua Chicaña
Universidad Nacional de San Agustín
ecayllahua@unsa.edu.pe

Fanny Lorena Lorenzo Quilla
Universidad Nacional de San Agustín
florenzo@unsa.edu.pe

Maria Quijia Álvarez
Universidad Nacional de San Agustín
mquijia@unsa.edu.pe

ARK: ark:/42411/s6/a46

PURL: 42411/s6/a46

RECIBIDO 10/08/2021 • ACEPTADO 15/09/2021 • PUBLICADO 30/09/2021

RESUMEN

Debido al problema recurrente presentado en los alumnos en lo que se refiere a su desempeño académico, se desarrolló una aplicación de redes neuronales con el objetivo de ayudar al docente, ya que esta es capaz de dar resultados de las notas finales de los alumnos y ayudará al docente a comprender el porqué de los resultados, puesto que esta red neuronal toma en cuenta diferentes factores que conlleva al alumno a tener una nota aprobatoria o desaprobatoria. Para obtener los resultados se trabajó en el entrenamiento de la red neuronal mediante el modelo de clasificación el cual muestra en el resultado la cantidad de alumnos aprobados o desaprobados y el otro modelo de regresión el cual predice la nota de un alumno dadas las características de su encuesta inicial, ambos modelos fueron de gran ayuda para predecir el comportamiento de los datos.

Palabras claves: Clasificación, Redes neuronales, Regresión, Predicción, Compartimiento del alumno.

ABSTRACT

Due to the recurring problem presented in the students regarding their academic performance, an application of neural networks was developed with the aim of helping the teacher, since it is

capable of giving results of the final grades of the students and will help the teacher to understand the reason for the results, since this neural network takes into account different factors that lead the student to have a passing or failing grade. To obtain the results, we worked on the training of the neural network through the classification model which shows in the result the number of approved or disapproved students and the other regression model which predicts a student's grade given the characteristics of their initial survey, both models were of great help in predicting the behavior of the data.

Keywords: *Classification, Neural networks, Regression, Prediction, Student sharing.*

INTRODUCIÓN

En este trabajo se presenta el desarrollo de una aplicación con el uso de Redes Neuronales multicapa el cual busca clasificar a los alumnos de acuerdo a su nota para ver si serán aprobados o desaprobados en un curso, a su vez esta toma en cuenta los diferentes factores que influyen en la nota final del alumno los cuales fueron clasificados en variables independientes y dependientes que posteriormente serán sometidos a una limpieza de datos donde se tratarán datos duplicados y datos anómalos y estarán listos para ser normalizados y almacenados en un *dataset*.

La herramienta que se utilizó para el desarrollo de la aplicación es Anaconda en su entorno de Spyder en el lenguaje de python donde para construir la red neuronal fue necesario instalar diferentes librerías las cuales ayudarán a predecir el comportamiento de los datos y las fallas que influyen en el desempeño de los alumnos los cuales serán reflejados en las notas finales. Esta aplicación tiene como objetivo ayudar a los profesores a proveer la nota final de sus alumnos y verificar las causas y los factores que influyen a que un alumno tenga nota aprobatoria o desaprobatoria al final del curso. Así también el modelo de regresión permitirá predecir la nota final de un alumno dadas las características iniciales de este.

MATERIALES Y MÉTODOS

Descripción general de la aplicación

Un Instituto Politécnico de Informática (IPI) quiere desarrollar una aplicación que apoye a los docentes a proveer el resultado final de aprobado o no aprobado para los estudiantes. Este resultado deberá ser propuesto por la aplicación en función de los datos que los estudiantes brindaron al inicio del curso.

Así mismo, de estas encuestas se obtuvieron 1 044 encuestas recogidas en 2 IPIs de una ciudad durante el curso anterior en las asignaturas de inglés y matemáticas, así como el rendimiento de dichos estudiantes.

Selección y justificación del modelo

Debido a la capacidad de aprendizaje, las redes neuronales pueden llevar a cabo ciertas tareas basadas en un entrenamiento. Por ello, para abordar esta etapa inicial de la investigación, se pensó en un proceso de entrenamiento de la red neuronal en el cual se expondrá un conjunto de patrones de entrada y se ajustarán los pesos de forma que al final de este proceso se obtengan las salidas deseadas, que en este momento están clasificadas en dos categorías diferentes aprobado o desaprobado.

Siendo así que justificamos el uso de un modelo de clasificación para la predicción de si estará aprobado o no, y el uso de un modelo de regresión para predecir la nota como tal.

Herramientas

Anaconda

Es una distribución libre y abierta de los lenguajes Python y R, utilizada en ciencia de datos, y aprendizaje automático (*machine learning*). Este incluye procesamiento de grandes volúmenes de información, análisis predictivo y cómputos científicos. Está orientado a simplificar el despliegue y administración de los paquetes de software [1].

Spyder

Es un entorno científico escrito en Python, para Python, y diseñado por y para científicos, ingenieros y analistas de datos. El cual ofrece una combinación única de funciones avanzadas de edición, análisis, depuración y creación de perfiles de una herramienta de desarrollo integral con la exploración de datos, ejecución interactiva, inspección profunda y hermosas capacidades de visualización de un paquete científico y para aprovechar todas estas funciones es necesario instalar el entorno de anaconda [2].

Librerías y framework

Para el desarrollo de la aplicación se instalaron librerías que facilitarán el tratamiento de los datos las cuales son:

Theano

Es un compilador para matemáticas en Python. Sabe cómo tomar sus estructuras para convertirlas en código eficiente que utiliza NumPy, el cual fue diseñado específicamente para manejar los tipos de computación requeridos para grandes empresas como los algoritmos de redes neuronales usados en *deep learning* [3].

Instalación de Theano Línea de comando

pip install Theano

Tensor Flow

Es una librería Python para computación numérica la cual puede utilizarse para crear modelos de *Deep Learning* directamente o utilizando librerías de envolturas que simplifican el proceso, fue diseñado para su uso tanto en investigación y desarrollo, como por ejemplo en sistemas de producción. Puede funcionar en una sola CPUs, GPUs, así como dispositivos móviles y sistemas distribuidos a gran escala de cientos de máquinas [4].

Instalación de Tensor Flow línea de comando

#pip install tensorflow=1.0.0

Keras

Es un framework de alto nivel para el aprendizaje, escrito en Python y capaz de correr sobre los frameworks TensorFlow, CNTK, o Theano. Fue desarrollado con el objeto de facilitar un proceso de experimentación rápida [5].

Instalación de Keras línea de comando

pip install keras

Numpy

Es una extensión de Python, que le agrega mayor soporte para vectores y matrices, constituyendo una biblioteca de funciones matemáticas de alto nivel para operar con esos vectores o matrices [6].

Instalación de Numpy línea de comando:

import numpy as np

Matplotlib.pyplop

Es una interfaz basada en estados para matplotlib. Proporciona una forma de trazado similar a MATLAB.Pyplot está diseñado principalmente para gráficos interactivos y casos simples de generación de gráficos programáticos [7].

Instalación de Matplotlib.pyplop línea de comando:

import matplotlib.pyplot as plt

Pandas

Es una herramienta de manipulación y análisis de datos de código abierto rápida, potente, flexible y fácil de usar, construida sobre el lenguaje de programación Python [8].

Instalación de Pandas línea de comando:

import pandas as pd

Sklearn.metrics

Implementa funciones que evalúan el error de predicción para propósitos específicos. Estas métricas se detallan en las secciones en las métricas de clasificación, MULTILABEL métricas de clasificación, las métricas de regresión y las métricas de agrupamiento [9].

Instalación de Sklearn.metrics línea de comando:

from sklearn.metrics import mean_squared_error

Selección del lenguaje de desarrollo

La elección del lenguaje para el desarrollo de esta investigación viene dado gracias a las ventajas de la misma; en este sentido, el trabajar con el lenguaje Python puede brindar facilidades como su sencillez, la usabilidad multiplataforma que permite la obtención de paquetes para la preparación del código en distintas plataformas. Otro punto resaltante, es la flexibilidad del lenguaje al permitir trabajar con programación Orientada a Objetos o scripting.

Sin embargo, uno de los aspectos más interesantes de este lenguaje es que permite trabajar el desarrollo de RN con librerías gratuitas propias del lenguaje, las cuales permiten acelerar el proceso de desarrollo e implementación [10].

Descripción de los procesos

Tratamiento y limpieza de datos

Lo primero es tratar el *dataset* proporcionado, en éste, se corregirán registros(encuestas) erróneas. Este proceso permitirá identificar registros incompletos, incorrectos, inexactos o no pertinentes, posterior a ello y con el *dataset* limpio, se puede comenzar a trabajar en el objetivo de esta investigación.

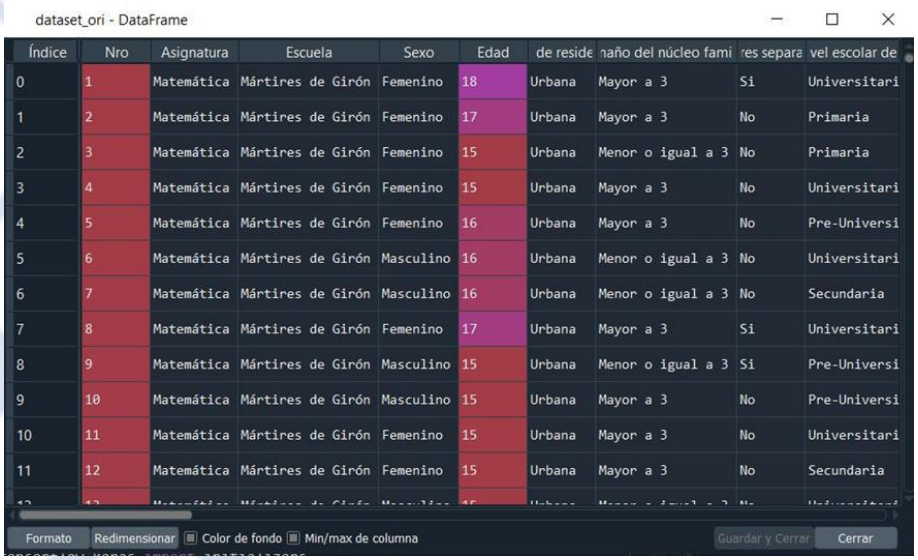

Figura 1. Dataset limpio.

Transformación de Datos

Lo primero es el reconocimiento de las variables dependientes e independientes sobre las cuales se va a trabajar y modelar la red neuronal.

Variables Independientes

Y	Nota final de curso

Dependientes

X_0	Asignatura
X_1	Escuela
X_2	Sexo del estudiante
X_3	Edad del estudiante
X_4	Tipo de residencia del estudiante (Urbana o Rural)
X_5	Tamaño del núcleo familiar del estudiante
X_6	Si el estudiante tiene padres separados
X_7	Nivel escolar de la madre

X_8	Nivel escolar del padre
X_9	Trabajo de la madre
X_10	Trabajo del padre
X_11	Razón por la que el estudiante escogió esa escuela en la secundaria
X_12	Tutor legal
X_13	Tiempo de viaje desde la casa del estudiante a la escuela
X_14	Tiempo de estudio semanal en horas
X_15	Número de asignaturas desaprobadas en la secundaria
X_16	Si la escuela le brinda al estudiante atención diferenciada
X_17	Si la familia del estudiante le presta atención diferenciada
X_18	Si el estudiante ha contratado a terceras personas para que lo ayuden en las asignaturas
X_19	Si el estudiante participa en actividades extracurriculares de la escuela
X_20	Si el estudiante ha sido atendido en la enfermería de la escuela
X_21	Si el estudiante pretende estudiar en la Universidad
X_22	Si el estudiante tiene internet en su hogar
X_23	Si el estudiante está en una relación de pareja
X_24	Calidad de la relación del estudiante con su familia
X_25	Tiempo libre después de la escuela
X_26	Si el estudiante sale con sus amigos
X_27	Cantidad de alcohol consumido entre semana
X_28	Cantidad de alcohol consumido en el fin de semana
X_29	Estado de salud
X_30	Cantidad de ausencias a la escuela durante el curso
X_31	Nota del primer trabajo de control parcial
X_32	Nota del segundo trabajo de control parcial

De esta manera, se debe identificar aquellas variables categóricas y no categóricas. Sin embargo, se ha considerado trabajar con Y_1 para la clasificación (desaprobado/aprobado) y Y_2 para la regresión (campo de notas finales).

Ahora, a las variables mostradas en el cuadro anterior se les aplicó un preprocesamiento, es decir la asignación de un equivalente numérico, teniendo en cuenta lo que representaba cada campo, para su posterior normalización, como resultado del mismo tenemos las siguientes equivalencias:

Tabla 1. Normalización de variables parte I.

X_0	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Asignatura	Escuela	Sexo	Edad	Tipo de residencia	Tamaño del núcleo familiar	¿Padres separados?	Nivel escolar de la madre
Matemática 0	Mártires de Girón 0	Femenino 0	15	Urbana 0	Menor o igual 0	Si 0	Ninguno 0
Inglés 1	Camilo Cienfuegos 1	Masculino 1	16	Rural 1	Mayor a 3 1	No 1	Primaria 1
			17				Secundaria 2
			18				Pre-Universitario 3
			19				Universitario 4
			20				
			21				
			22				

Tabla 2. Normalización de variables parte II.

X_8		X_9		X_10		X_11		X_12		X_13		X_14		X_15		X_16	
Atención diferenciada por la escuela		Atención diferenciada por la familia		Contratos a terceros para ayuda con asignaturas		¿Participa en actividades extracurriculares?		¿Atendido en la enfermería de la escuela?		¿Quiere estudiar en la Universidad?		¿Tiene internet en la casa?		¿Está en una relación de pareja?		Calidad de las relación con su familia	
Si	1	Si	1	Si	1	Si	1	Si	1	Si	1	Si	1	Si	1	Muy mala	0
No	0	No	0	No	0	No	0	No	0	No	0	No	0	No	0	Mala	1
																Regular	2
																Buena	3
																Muy Buena	4

Tabla 3. Normalización de variables parte III.

X_17		X_18		X_19		X_20		X_21		X_22		X_23		X_24
Nivel escolar del padre		Trabajo de la madre		Trabajo del padre		Razón por la que el estudiante escogió esa escuela		Tutor legal		Tiempo de viaje a la escuela		Tiempo de estudio semanal		Número de asignaturas desaprobadas en la secundaria
Ninguno	0	En casa	0	En casa	0	Asignaturas que imparte	0	Madre	0	Menos de 15 mir	0	Menos de 2 hora	0	0
Primaria	1	Salud	1	Salud	1	Otro	1	Padre	1	De 15 a 30 minu	1	De 2 a 5 horas	2	1
Secundaria	2	Otro	0	Otro	0	Distancia a su casa	0	Otro	-1	De 30 a 60 minu	2	De 5 a 10 horas	5	2
Pre-Universitario	3	Profesor	1	Profesor	1	Reputación	1			Más de 1 hora	4	Más de 10 horas	10	3
Universitario	4	Servicios	0	Servicios	0									
		En casa	0	En casa	0	Asignaturas que imparte	0	Madre	1					
		Salud	1	Salud	1	Otro	1	Padre	2					
		Otro	2	Otro	2	Distancia a su casa	2	Otro	0					
		Profesor	3	Profesor	3	Reputación	3							
		Servicios	4	Servicios	4									

Tabla 4. Normalización de variables parte IV.

X_25		X_26		X_27		X_28		X_29		X_30	X_31	X_32
Tiempo libre después de la escuela		¿Sale con amigos?		Cantidad de alcohol consumido entre semana		Cantidad de alcohol consumido en el fin de semana		Estado de salud		Cantidad de ausencias	Nota del primer trabajo de control parcial	Nota del segundo trabajo de control parcial
Muy poco	0	Muy poco	0	Muy poco	0	Muy poco	0	Mal	0	0	0	0
Poco	1	Poco	1	Poco	1	Poco	1	Muy mal	1	1	15	20
Más o menos	2	Regularmente	2	Regular	2	Regular	2	Regular	2	2	20	25
Mucho	3	Frecuentemente	3	Demasiado	3	Demasiado	3	Bueno	3	3	25	30
Demasiado	4	Muy frecuente	4	Bastante	4	Bastante	4	Muy bueno	4	4	30	35
										5	35	40
										6	40	45
										7	45	50
										8	50	55
										9	55	60
										10	60	65
										11	65	70
										12	70	75
										13	75	80
										14	80	85
										15	85	90
										16	90	95
										17	95	
										18		
										19		
										20		
										21		

Como resultado final de esta fase, se tuvo el *dataset* a entrenar.

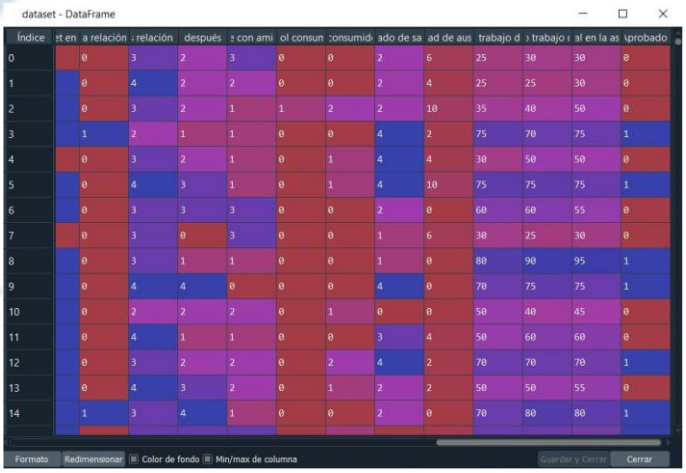

Figura 2. Dataset preprocesado - Encuestas de inicio de curso y notas.

ENTRENAMIENTO Y VALIDACIÓN

Fase de preprocesamiento de datos

```
# Parte 1 - Pre procesamiento de datos
# Importando las librerias
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

Figura 3. Implementación del importe de librerías para el preprocesamiento.

```
# Importando los datasets
dataset = pd.read_csv('data_trabajada.csv')
X = dataset.iloc[:, 1:34].values #solo aquellos campos relevantes
y = dataset.iloc[:, 35].values
```

Figura 4. Implementación de la división de variables dependientes e independientes.

Fase de entrenamiento

```
# Separando sets de datos en Entrenamiento y Prueba
from sklearn.model_selection import train_test_split
#variables
test_percent = 0.25
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = test_percent, random_state = 0)
```

Figura 5. Implementación de la división de la data de test y entrenamiento.

```
# Escalado de Caracteristicas
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
```

Figura 6. Implementación para el escalamiento de datos.

Fase de Creación de la red neuronal

Sequential es utilizada para que la red neuronal sea creada en secuencias, capa por capa, manualmente

```
# Parte 2 - Creando Red neuronal

# Importando librerias de Keras
import keras
from keras.models import Sequential
from keras.layers import Dense
from tensorflow.keras import initializers
```

Figura 7. Importe de librerías para la creación de la red neuronal.

```
# Iniciando Red Neuronal
classifier = Sequential()

i = 33 #nro de neuronas capa de entrada
o = 1    #nro de neuronas capa de salida
r = (i/o)**(1/3)
h1 = round(o*(r**2) )+1
h2 = round(o*r)+1
```

Figura 8. Implementación para el escalamiento de datos.

```
# Agregando capa Input y primera capa oculta
classifier.add(Dense(units = h1, kernel_initializer = initializers.TruncatedNormal(mean=0.0, stddev=0.05, seed=None),
                     bias_initializer=initializers.RandomNormal(), activation = 'relu', input_dim = i))

# Agregando segunda capa oculta
classifier.add(Dense(units = h2 , kernel_initializer =initializers.TruncatedNormal(mean=0.0, stddev=0.05, seed=None),
                     bias_initializer=initializers.RandomNormal(), activation = 'relu'))

# Agregando la capa output
classifier.add(Dense(units = 1, kernel_initializer = initializers.TruncatedNormal(mean=0.0, stddev=0.05, seed=None),
                     bias_initializer=initializers.RandomNormal(), activation = 'sigmoid'))

# Compilando la Red Neuronal
classifier.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['accuracy'])
```

Figura 9. Implementación del modelo (capas).

```
# Encajando Red Neuronal con set de Entrenamiento
epochs_hist = classifier.fit(X_train, y_train, batch_size = 27, epochs = 300)
# peso/bias  de h1
weights = classifier.layers[0].get_weights()[0]
biases = classifier.layers[0].get_weights()[1]
# peso/bias  de h2
weights2 = classifier.layers[1].get_weights()[0]
biases2 = classifier.layers[1].get_weights()[1]
# peso/bias  de h2
weights3 = classifier.layers[2].get_weights()[0]
biases3 = classifier.layers[2].get_weights()[1]
```

Figura 10. Implementación del encajamiento de la RN con el entrenamiento.

Fase de predicción y comprobación

```
# Parte 3 - Haciendo Predicciones y Evaluando el Modelo

# Prediccion con uso de los registros de prueba separados
y_pred = classifier.predict(X_test)
# metricas
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, y_pred)
mse
```

Figura 11. Implementación de la predicción.

```
#Midiendo data de entrenamiento
scores = classifier.evaluate(X_train, y_train, verbose=0)
print("%s: %.2f%%" % (classifier.metrics_names[1], scores[1]*100))

#Midiendo data de Prueba
scores = classifier.evaluate(X_test, y_test, verbose=0)
print("%s: %.2f%%" % (classifier.metrics_names[1], scores[1]*100))
```

Figura 12. Implementación de la medición del entrenamiento y la prueba.

```
def saveModel(model, nameModel):
    # serializa el modelo para JSON
    model_json1 = loaded_model.to_json()
    with open("model1_2.json", "w") as json_file:
    json_file.write(model_json1)
    #serializan los pesos (weights) para HDF5
    loaded_model.save_weights("model1_2.h5")
    print("Modelo guardado en el PC")
```

Figura 13. Implementación para la serialización del modelo para json.

```
def getModel():
    json_file = open('model1_2.json', 'r')
    loaded_model_json = json_file.read()
    json_file.close()
    loaded_model = model_from_json(loaded_model_json)
    # se cargan los pesos (weights) en el nuevo modelo
    loaded_model_weights = loaded_model.load_weights("model1_2.h5")
    print("Modelo cargado desde el PC")

  # se evalua el modelo cargado con los datos de los test
    loaded_model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
```

Figura 14. Implementación de la de pesos en el nuevo modelo.

```
# Gráficos
#Historial de epocs
epochs_hist.history.keys()

#Grafico-  Progreso de entrenamiento del modelo según epocs
plt.plot(epochs_hist.history['loss'])
#plt.plot(epochs_hist.history['val_loss'])
plt.title('Progreso de entrenamiento del modelo')
plt.xlabel('Epoch')
plt.ylabel('Training Loss')
plt.legend(['Training Loss'])
```

Figura 15. Implementación para mostrar los resultados gráficamente.

```
#Prediccion - gráfico de resultados de predicción
y_predict = classifier.predict(X_test) # predicción del modelo
plt.plot(y_test, y_predict, "^", color = 'g')
plt.title('Resultados de predicción del modelo')
plt.xlabel('Prediccion del Modelo')
plt.ylabel('Valores Verdaderos')
```

Figura 16. Implementación para mostrar la predicción gráficamente.

```
#Prediciendo resultados de estudiantes nuevos
##Nro,Asignatura,Escuela,Sexo,Edad,Tipo de residencia,Tamaño del núcleo familiar,¿Padres separados?,Nivel escolar de la madre,Nivel escol
#469,Inglés,Mártires de Girón,Masculino,16,Urbana,Mayor a 3,No,Pre-Universitario,Primaria,Otro,Otro,Reputación,Madre,Menos de 15 minutos,I
#469,1,0,1,16,0,1,0,3,1,0,0,1,0,0,0,0,0,0,0,0,1,1,1,0,0,4,2,1,1,1,4,0,65,65,70,1

nuevo_estudiante=loaded_model.predict(sc.transform(np.array([[1,0,1,16,0,1,0,3,1,0,0,1,0,0,0,0,0,0,0,1,1,1,0,0,4,2,1,1,1,4,0,65,65]])))
nuevo_estudiante = (nuevo_estudiante > 0.5)

#449,Inglés,Mártires de Girón,Femenino,15,Urbana,Mayor a 3,No,Universitario,Universitario,Servicios,Servicios,Asignaturas que imparte,Mad
#449,1,0,0,15,0,1,0,4,4,0,0,0,0,0,0,0,1,1,0,0,1,1,1,0,2,2,3,1,2,4,0,65,60,60,0
nuevo_estudiante2=loaded_model.predict(sc.transform(np.array([[1,0,0,15,0,1,0,4,4,0,0,0,0,0,0,0,1,1,0,0,1,1,1,0,2,2,3,1,2,4,0,65,60]])))
nuevo_estudiante2 = (nuevo_estudiante > 0.5)
```

Figura 17. Implementación para predecir resultados de estudiantes nuevos.

RESULTADOS Y DISCUSIÓN

En esta parte del documento, se presentan los resultados de la construcción de la red neuronal para el tema de investigación.

Modelo de regresión

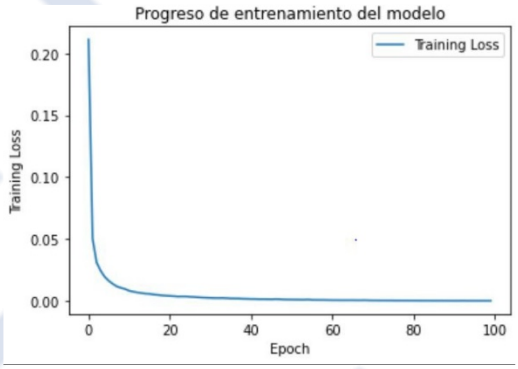

Figura 18. Progreso de entrenamiento - Modelo de regresión.

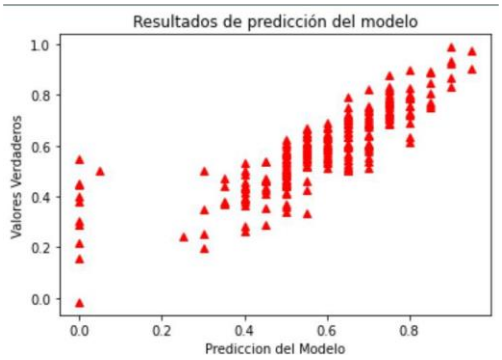

Figura 19. Resultados de predicción - Modelo de regresión.

Las métricas del modelo creado fueron las siguientes:

```
 accuracy de entrenamiento: 5.28%
accuracy de validación: 4.94%
Error absoluto medio: 5.03%
Error cuadratico medio: 0.61%
Variación: 83.30%
Error máximo: 43.79%
```

Figura 20. Métricas- Modelo de regresión.

Modelo de clasificación

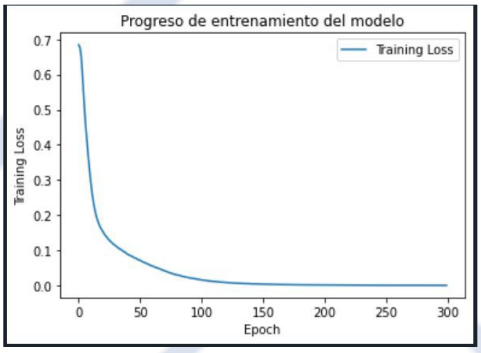

Figura 21. Progreso de entrenamiento - Modelo de clasificación.

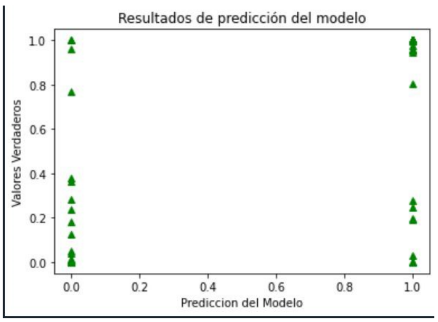

Figura 22. Resultados de predicción - Modelo de clasificación.

		predicción	
		0	1
realidad	0	TN	FP
	1	FN	TP

Figura 23. Lectura de una matriz de confusión.

La lectura de la matriz de confusión es la siguiente: True Negative [TN], True Positive [TP], False Positive [FP], False Negative [FN]. En nuestro caso, tenemos el siguiente resultado:

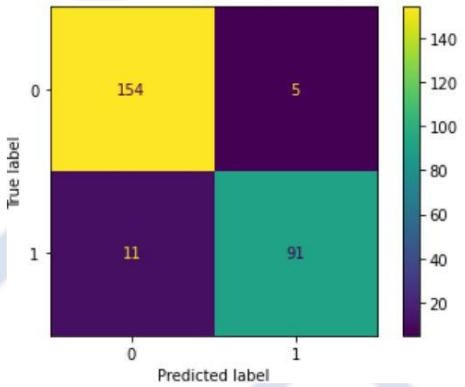

Figura 24. Gráfica de matriz de confusión - Modelo de clasificación.

Tabla 5. Resultados de métricas - Modelo de clasificación.

Nro	Métrica	Resultado
1	Error Cuadrático Medio	0.054983007933077
2	Presicion (Precisión)	0.947916666666666
3	Recall (Exhaustividad)	0.892156862745098
4	Accuracy (Exactitud)	0.938697318007662
5	F1	0.919191919191919

Comparativa

Para poder evaluar los dos métodos, adecuamos los resultados y_predictB = (y_predict>0.6) para tener el número de aprobados versus el número de desaprobados como se muestra en la figura 26. La misma la comparamos con la Figura 25, podemos ver que la predicción del modelo de clasificación con respecto al regresión es más acertada.

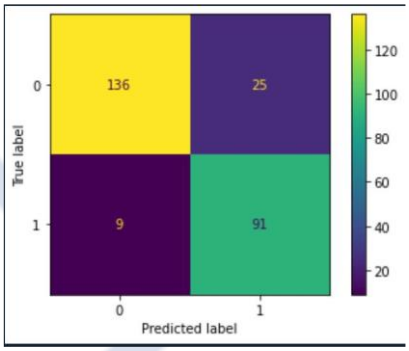

Figura 26. Gráfica de matriz de confusión - Modelo de regresión

CONCLUSIONES

En conclusión, una red neuronal de regresión nos permite predecir valores numéricos es decir la nota del alumno, sin embargo, al ser continuos los datos que calcula, hay un mayor ruido en los indicadores como la exactitud. Podemos medir la calidad de la predicción con métricas como el error absoluto medio, error cuadrático medio, varianza, entre otros. La red neuronal de clasificación nos permite predecir si el estudiante estará o no aprobado, la misma se puede medir empleando métricas tales como el cuadro de confusión, la precisión, la exactitud, entre otros. Los datos que poseemos tenemos que pre procesarlos de acuerdo al modelo que utilicemos y normalizarlos de ser necesario. entendiendo la relación de cada componente con el modelo predictivo.

REFERENCIAS

[1] Anaconda su kit de herramientas de ciencia de datos. [Online]. Available: https://www.anaconda.com/products/individual

[2] Spyder vision general. [Online]. Available: https://www.spyder-ide.org/

[3] Theano Project description. [Online]. Available: https://pypi.org/project/Theano/

[4] J. Buhigas"Todo lo que necesitas saber sobre Tensor Flow, la plataforma para Inteligencia Artificial de Google",febrero, 2018 [Online]. Available:

https://puentesdigitales.com/2018/02/14/todo-lo-que-necesitas-saber-sobre-tensorflow-la-plataforma-para-inte ligencia-artificial-de-google/#:~:text=TensorFlow%20es%20una%20biblioteca%20de,gr%C3%A1ficos%20de%20flujo%20de%20datos.&text=La%20arquitectura%20flexible%20de%20TensorFlow,m%C3%B3viles% 20con%20una%20sola%20API.

[5] Desarrollo de redes neuronales con keras. [Online]. Available:

https://unipython.com/desarrolla-primera-red-neural-python-keras-paso-paso/#:~:text=Keras%20es%20una%20librer%C3%ADa%20Python,unas%20pocas%20l%C3%ADneas%20de%20c%C3%B3digo.

[6] L.Gonzales"Introducción a la librería NumPy de Python",setiembre, 2018. [Online]. Available: https://ligdigonzalez.com/introduccion-a-numpy-python-1/#:~:text=NumPy%20es%20un%20paquete%20de, garantizan%20c%C3%A1lculos%20eficientes%20con%20matrices.

[7] Matplotlib: funciones principales. [Online]. Available: https://unipython.com/matplotlib-funciones-principales/

[8] Introducción a Pandas. [Online]. Available: https://programacion.net/articulo/introduccion_a_pandas_1632

[9] Metrics and scoring: quantifying the quality of predictions [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html

[10] Ventajas del uso de Python [Online]. Available

https://blog.enzymeadvisinggroup.com/redes-neuronales-python#:~:text=La%20principal%20caracter%C3%ADstica%20de%20Python,multiplataforma%20y%20de%20tipado%20fuerte.

Selección de una red social para apoyar la docencia universitaria empleando computación con palabras

Selection of one social network to support higher education teaching through computing with words

Dargel Veloz Morales
Universidad de las Ciencias Informáticas
dveloz@uci.cu
https://orcid.org/0000-0002-4231-5831

Laritza González Marrero
Universidad de las Ciencias Informáticas
lgmarrero@uci.cu
https://orcid.org/0000-0002-6128-8496

ARK: ark:/42411/s6/a47
PURL: 42411/s6/a47

RECIBIDO 26/08/2021 • ACEPTADO 20/09/2021 • PUBLICADO 30/09/2021

RESUMEN

Las redes sociales han impactado la sociedad, a tal punto que en muchas ocasiones se prioriza más estar al tanto de ellas que de cualquier otra aplicación en el móvil, tablet, o laptop. Por otro lado, la llegada de la enfermedad covid-19 también ha incidido en el modo de actuación de los estudiantes, profesores y el propio sistema de enseñanza. El presente trabajo tiene como objetivo seleccionar la red social más adecuada para el apoyo a la enseñanza superior, a través de la computación con palabras, utilizada para realizar el proceso de computación y razonamiento. Además, para simular los diferentes modelos, es usado el programa FLINTSTONES. Finalmente, se exponen los resultados alcanzados por cada una de las tres redes sociales analizadas: Telegram, WhatsApp y Facebook.

Palabras claves: Redes sociales, computación con palabras, CWW, 2-tupla lingüística, enseñanza superior.

ABSTRACT

Social networks have impacted society, to such an extent that in many occasions it is more important to be aware of them than any other application on the cell phone, tablet or laptop. On the other hand, the arrival of the covid-19 disease has also impacted the way students, teachers

and the education system itself act. The present work aims to select the most suitable social network for higher education support, through computing with words, used to perform the process of computation and reasoning. In addition, the FLINTSTONES program is used to simulate the different models. Finally, the results achieved for each of the three social networks analyzed are presented: Telegram, WhatsApp and Facebook.

Keywords: *Social networking, computing with words, CWW, 2-tuple linguistics, higher education.*

INTRODUCIÓN

El impacto de las TIC en la sociedad es una temática de la cual se pudiera comentar durante extensas instancias de tiempo. Su influencia es notable en la economía, la medicina, el deporte, la recreación, el desarrollo social, la educación, entre otros.

Cada vez resulta más común observar un número creciente de estudiantes de la Universidad de las Ciencias informáticas (UCI) con dispositivos móviles, haciendo diferentes usos del mismo, incluso en función de procesos docentes.

Con el desarrollo de los diferentes modelos del aprendizaje, comienzan a adoptarse con más frecuencia en los entornos de educación superior. Uno de ellos es el M-learning (mobile learning), el cual proporciona herramientas y facilita el acceso al conocimiento y procesos del aprendizaje desde cualquier dispositivo móvil, incluyendo en su concepción el empleo de las redes sociales; sin embargo, no se debe pasar por alto que tanto las redes como los dispositivos móviles, por sí mismos, no garantizan el aprendizaje.

Si las reglas o instrucciones no son claras, los dispositivos móviles se convierten en un distractor [1].

Las redes sociales son un concepto que ha tocado el mundo, modificando la forma en la que se desarrollan las relaciones, procesos de negocio, sociales, y por qué no, también personales. Existen muchas, y sus ejemplares abarcan diferentes propósitos dentro de la sociedad. Es difícil encontrar alguna persona que no esté involucrada en este contexto, utilizando al menos una red social en su vida diaria.

Algunas características fundamentales de las redes sociales, según [2], son:

- Fomentan la comunicación entre el alumnado de forma sencilla y, además, se incrementa a través de la creación de grupos de trabajo.

- Posibilitan actuaciones comunes a nivel docente, tanto en la institución educativa como a nivel de aula.
- Posibilitan el uso masivo por parte de estudiantes y docentes de forma ordenada, permitiendo una incorporación generalizada de estos recursos a nivel educativo.
- Puesto que las redes sociales son generalistas, las herramientas que incorporan son las mismas para todos los usuarios, aspecto primordial en las fases iniciales de utilización. A posteriori, estas se pueden complementar con herramientas externas más especializadas, que se pueden usar de forma complementaria.

La utilización de medios online y sitios de redes sociales ha ido adquiriendo una importancia cada vez mayor en la última década y se ha convertido en uno de los hábitos de comportamiento más extendidos entre la ciudadanía, debido a la ubicuidad y la convergencia de los dispositivos en el entorno multipantalla [3].

En este estudio se hace relevante la selección de una red social que fungirá como apoyo al proceso docente universitario a distancia. Con este propósito se decide emplear un modelo de computación con palabras (CWW), para asirnos de sus facilidades por causa de su factibilidad en procesos de selección.

Hay situaciones en las cuales no es apropiado el uso de una evaluación cuantitativa. En estos casos, la utilización de un enfoque lingüístico puede ser más conveniente, en el cual, son usados un conjunto de términos como los mostrados en la expresión [4].

La computación con palabras (CWW) es una metodología que permite realizar un proceso de computación y razonamiento, utilizando palabras pertenecientes a un lenguaje en lugar de números. Dicha metodología permite crear y enriquecer modelos de decisión, en los cuales, la información vaga e imprecisa es representada a través de variables lingüísticas [5].

Por su parte, Rente y sus coautores, la describen de este modo:

El uso de información lingüística implica la necesidad de operar con variables lingüísticas. El cálculo con palabras (CWW) es un paradigma basado en un procedimiento que emula los procesos cognitivos humanos para tomar decisiones y procesos de razonamiento en entornos de incertidumbre e imprecisión [6].

En el año 2021 la Universidad de las Ciencias Informáticas ha asumido el reto de impartir la enseñanza en pregrado de varias asignaturas completamente a distancia, a partir del uso de la plataforma Moodle (https://eva.uci.cu). A pesar de que la herramienta es idónea y que los conocimientos para su adecuada aplicación han aumentado entre estudiantes y profesores, aún

se observa una cifra creciente de estudiantes que participa activamente en procesos docentes y no docentes mediante las redes sociales.

Ante esta realidad, la Dirección de Formación de Pregrado, se propone realizar una selección de la red social más adecuada para el apoyo a la docencia universitaria.

Las redes sociales candidatas para este proceso son WhatsApp, Telegram y Facebook. Las mismas serán analizadas por tres expertos: un ingeniero en ciencias informáticas y profesor asistente, un doctor en ciencias de la educación y profesor auxiliar, y un especialista A en gestión de redes sociales (community manager).

Este análisis estará determinado por los siguientes parámetros: cantidad de funcionalidades útiles para la docencia, número de usuarios, y la calidad visual de la aplicación. Todos ellos serán valorados por los expertos utilizando los términos lingüísticos: Ninguno, Muy bajo, Bajo, Medio, Alto, Muy alto, Perfecto.

MATERIALES Y MÉTODOS

Una vez descrita la problemática, resulta de vital importancia definir cómo dar continuidad en la búsqueda de la solución pertinente; por tanto, se procede a desarrollar las etapas definidas en la metodología para la resolución de problemas de toma de decisión lingüística.

El proceso de toma de decisiones se realiza conforme al esquema que se visualiza en la figura 1.

Figura 1. Esquema de resolución de problemas de toma de decisión lingüística.

Para realizar este proceso se requiere la definición de varios conjuntos que se refieren a continuación.

Se define el conjunto de siete términos lingüísticos $S = \{S_0, S_1, S_2, S_3, S_4, S_5, S_6\}$, donde:

- S_0: Ninguno
- S_1: Muy bajo
- S_2: Bajo

- S_3: Medio
- S_4: Alto
- S_5: Muy alto
- S_6: Perfecto

Conjunto de alternativas A= {A_1; A_2; A_3}

- A_1: WhatsApp
- A_2: Telegram
- A_3: Facebook

Conjunto de atributos C = {C_1; C_2; C_3}

- C_1: Cantidad de funcionalidades útiles para la docencia
- C_2: Número de usuarios
- C_3: Calidad visual de la aplicación

Conjunto de expertos: E = {E_1; E_2; E_3}

- E_1: Ingeniero en ciencias informáticas y profesor asistente
- E_2: Doctor en ciencias de la educación y profesor auxiliar
- E_3: Especialista A en gestión de redes sociales (community manager)

La problemática se presenta como un problema de selección, actuando como decisor la Dirección de Formación de Pregrado. Este problema se clasifica de las formas siguientes.

- Número de atributos evaluados: multiatributo (3 atributos).
- Número de expertos que intervienen: multiexperto (3 expertos).
- Dominios de expresión empleado: homogéneo (siete términos lingüísticos).
- Consideración de los cambios en el tiempo: estático (una sola vez).

Luego de especificadas estas definiciones, se requiere que cada experto evalúe a las redes sociales candidatas, atendiendo a los atributos (criterios). Para la evaluación de estos criterios se deben emplear los términos lingüísticos establecidos en el conjunto S. La representación de los mismos se lleva a cabo a través del modelo 2-tuplas lingüística.

El modelo de representación lingüística de 2-tuplas permite realizar procesos de computación con palabras sin pérdida de información, basándose en el concepto de traslación simbólica [7].

La etapa de selección del operador de agregación y la agregación misma requieren de un análisis previo.

Una vez representada la información lingüística como un valor continuo, se puede operar con dicha información mediante los operadores de agregación. Para la toma de decisión, la agregación de múltiples valores es esencial. [8].

Existen un buen número de operadores de agregación, entre ellos están "la media aritmética extendida" y "la media lingüística ponderada". sus definiciones son comentadas por [9] como se muestra a continuación en las figuras 2 y 3:

$$\bar{M}_{2t}(X) = \Delta\left(\frac{1}{n}\sum_{j=1}^{n}\Delta^{-1}(s_i,\alpha)_j\right)$$

Figura 2. Formulación para el cálculo del operador media aritmética extendida para 2-tupla lingüística.

$$\bar{M}_{w2t}(X) = \Delta\left(\sum_{j=1}^{n} w_j\Delta^{-1}(s_i,\alpha)_j\right)$$

Figura 3. Formulación para el cálculo del operador media ponderada extendida para 2-tupla lingüística.

En lo sucesivo se aplican los cálculos de los operadores sobre los valores 2-tupla lingüística para realizar la agregación y, acto seguido, se da paso a la etapa de explotación. Es en esta última, donde se realiza la comparación de las inteligencias colectivas de todos los expertos para determinar cuál es la mejor opción entre las redes sociales candidatas.

Resultados y discusión

Para dar solución al marco de trabajo definido con anterioridad se da continuidad a secuencia de etapas que propone la resolución de problemas de toma de decisión lingüística.

Selección del conjunto de términos lingüísticos y su semántica.

Como se comentó en la sección precedente. se emplea un conjunto de siete términos lingüísticos $S = \{S_0, S_1, S_2, S_3, S_4, S_5, S_6\}$, y para darle solución se emplea el modelo 2-tuplas lingüística.

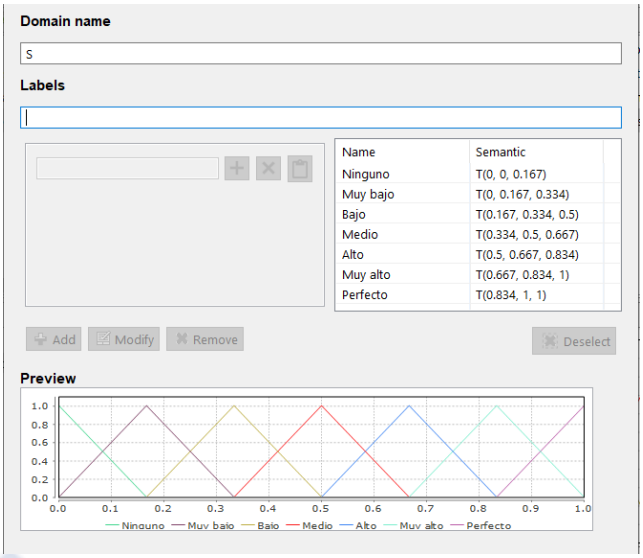

Figura 2. Conjunto de términos lingüísticos S.

Recopilación de preferencias.

Las siguientes tablas registran la recopilación de preferencias de cada experto sobre las alternativas y basándose en los criterios.

Tabla 1. Preferencias de los experto E_1.

E_1	C_1	C_2	C_3
A_1	M	MA	B
A_2	A	A	M
A_3	B	A	A

Tabla 2. Preferencias de los experto E_2

E_2	C_1	C_2	C_3
A_1	MA	A	M
A_2	MA	A	M
A_3	B	A	A

Tabla 3. Preferencias de los experto E_3

E_3	C_1	C_2	C_3
A_1	A	A	M

A_2	MA	A	M
A_3	M	A	MA

Esta información recopilada se expresa a continuación en valores 2-tupla lingüística. Al tratarse de valores originales, la traslación simbólica es 0.

Tabla 4. Transformación a 2-tupla lingüística de las preferencias de los experto E_1

E_1	**C_1**	**C_2**	**C_3**
A_1	$S_3,0$	$S_5,0$	$S_2,0$
A_2	$S_4,0$	$S_4,0$	$S_3,0$
A_3	$S_2,0$	$S_4,0$	$S_4,0$

Tabla 5. Transformación a 2-tupla lingüística de las preferencias de los experto E_2

E_2	**C_1**	**C_2**	**C_3**
A_1	$S_5,0$	$S_4,0$	$S_3,0$
A_2	$S_5,0$	$S_4,0$	$S_3,0$
A_3	$S_2,0$	$S_4,0$	$S_4,0$

Tabla 6. Transformación a 2-tupla lingüística de las preferencias de los experto E_3

E_3	**C_1**	**C_2**	**C_3**
A_1	$S_4,0$	$S_4,0$	$S_3,0$
A_2	$S_5,0$	$S_4,0$	$S_3,0$
A_3	$S_3,0$	$S_4,0$	$S_5,0$

Selección del operador de agregación.

En este punto de la investigación se decide utilizar el operador: media ponderada extendida para 2-tupla lingüística en función del cálculo de las preferencias colectivas de los criterios. Y el operador: media aritmética extendida para el de las preferencias de las alternativas.

Agregación.

En lo sucesivo se realizan los cálculos pertinentes empleando la media ponderada extendida, arrojando los siguientes resultados de las preferencias colectivas sobre cada alternativa.

Tabla 7. Preferencias colectivas de los criterios de los expertos

Inteligencia Colectiva	C_1	C_2	C_3
A_1	S_5, -0.2	S_4, -0.1	S_2, 0.4
A_2	S_6, -0.4	S_4, -0.4	S_3, -0.3
A_3	S_3, -0.2	S_4, -0.4	S_4, -0.1

Acto seguido se procede a aplicar la media aritmética extendida para obtener los siguientes valores 2-tupla de cada alternativa.

$A_1 = (S_4, -0.3)$

$A_2 = (S_4, -0.03)$

$A_3 = (S_3, 0.43)$

Explotación.

Finalmente, se realiza la comparación entre los valores 2-tupla obtenidos para cada alternativa, resultando el siguiente ranking de mayor a menor

$(S_4, -0.03) > (S_4, -0.3) > (S_3, 0.43)$

A partir de estos valores se concluye que la mejor alternativa es A2, valor que representa a la red social Telegram, como la más adecuada para el apoyo a la docencia universitaria.

A continuación, se evidencia la solución desarrollada con FLINTSTONES, teniendo en cuenta las opciones y vistas que facilita el software.

FLINTSTONES es un programa que proporciona un entorno para el análisis de los procesos de alcance de consenso mediante la simulación de diferentes modelos de CWW [10].

Seguidamente, se presentan figuras que muestran algunas fases del problema desarrollado en la herramienta FLINTSTONES.

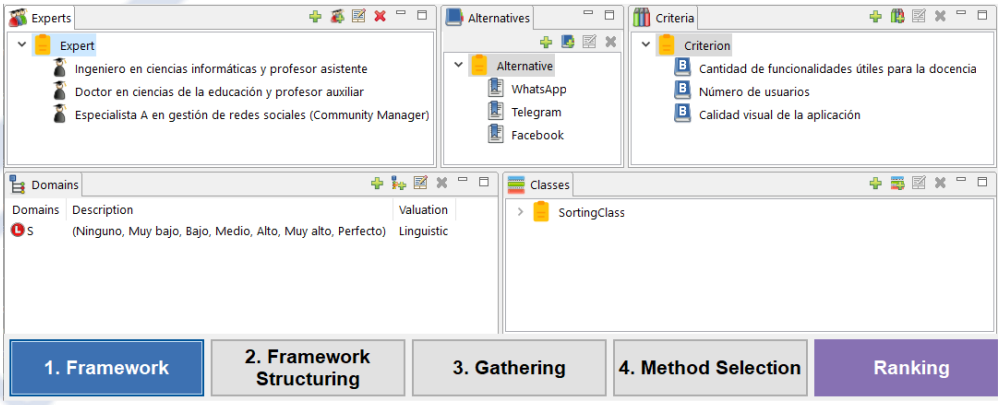

Figura 4. Framework.

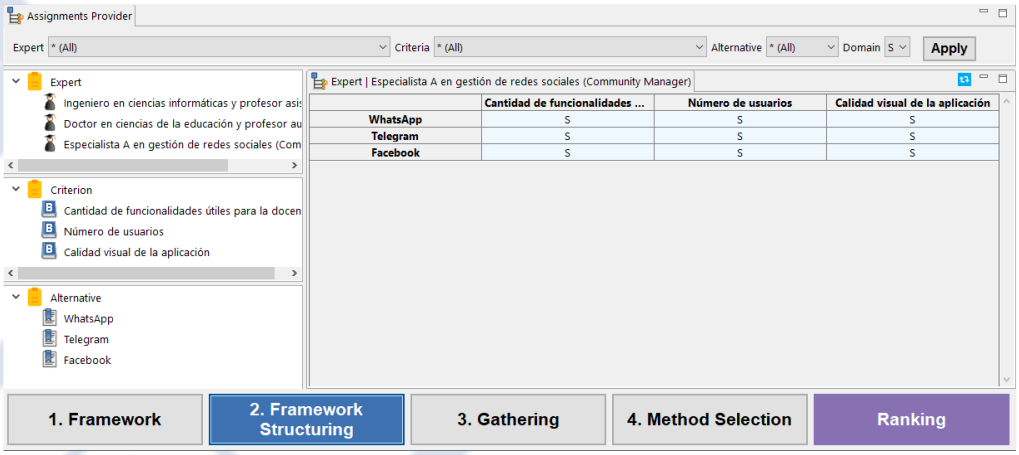

Figura 5. Framework Structuring.

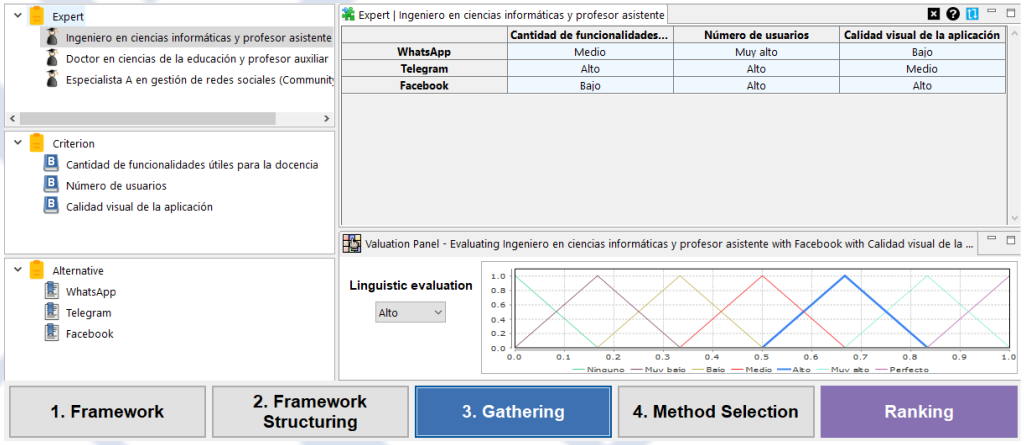

Figura 6. Gathering

Unified	Expert	Alternative	Criterion	Source domain	Evaluation
[Alto, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	Facebook	Calidad visual de la aplicación	S	Alto
[Bajo, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	Facebook	Cantidad de funcionalidades útiles para la docencia	S	Bajo
[Alto, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	Facebook	Número de usuarios	S	Alto
[Medio, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	Telegram	Calidad visual de la aplicación	S	Medio
[Muy alto, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	Telegram	Cantidad de funcionalidades útiles para la docencia	S	Muy alto
[Alto, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	Telegram	Número de usuarios	S	Alto
[Medio, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	WhatsApp	Calidad visual de la aplicación	S	Medio
[Muy alto, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	WhatsApp	Cantidad de funcionalidades útiles para la docencia	S	Muy alto
[Alto, 0.0]	Doctor en ciencias de la educación y profesor auxiliar	WhatsApp	Número de usuarios	S	Alto
[Muy alto, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	Facebook	Calidad visual de la aplicación	S	Muy alto
[Medio, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	Facebook	Cantidad de funcionalidades útiles para la docencia	S	Medio
[Alto, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	Facebook	Número de usuarios	S	Alto
[Medio, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	Telegram	Calidad visual de la aplicación	S	Medio
[Muy alto, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	Telegram	Cantidad de funcionalidades útiles para la docencia	S	Muy alto
[Alto, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	Telegram	Número de usuarios	S	Alto
[Medio, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	WhatsApp	Calidad visual de la aplicación	S	Medio
[Alto, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	WhatsApp	Cantidad de funcionalidades útiles para la docencia	S	Alto
[Alto, 0.0]	Especialista A en gestión de redes sociales (Community Manager)	WhatsApp	Número de usuarios	S	Alto
[Alto, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	Facebook	Calidad visual de la aplicación	S	Alto
[Bajo, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	Facebook	Cantidad de funcionalidades útiles para la docencia	S	Bajo
[Alto, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	Facebook	Número de usuarios	S	Alto
[Medio, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	Telegram	Calidad visual de la aplicación	S	Medio
[Alto, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	Telegram	Cantidad de funcionalidades útiles para la docencia	S	Alto
[Alto, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	Telegram	Número de usuarios	S	Alto
[Bajo, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	WhatsApp	Calidad visual de la aplicación	S	Bajo
[Medio, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	WhatsApp	Cantidad de funcionalidades útiles para la docencia	S	Medio
[Muy alto, 0.0]	Ingeniero en ciencias informáticas y profesor asistente	WhatsApp	Número de usuarios	S	Muy alto

(Ninguno, Muy bajo, Bajo, Medio, Alto, Muy alto, Perfecto)

2-tuple linguistic computational model

Figura 7. Todos los valores introducidos.

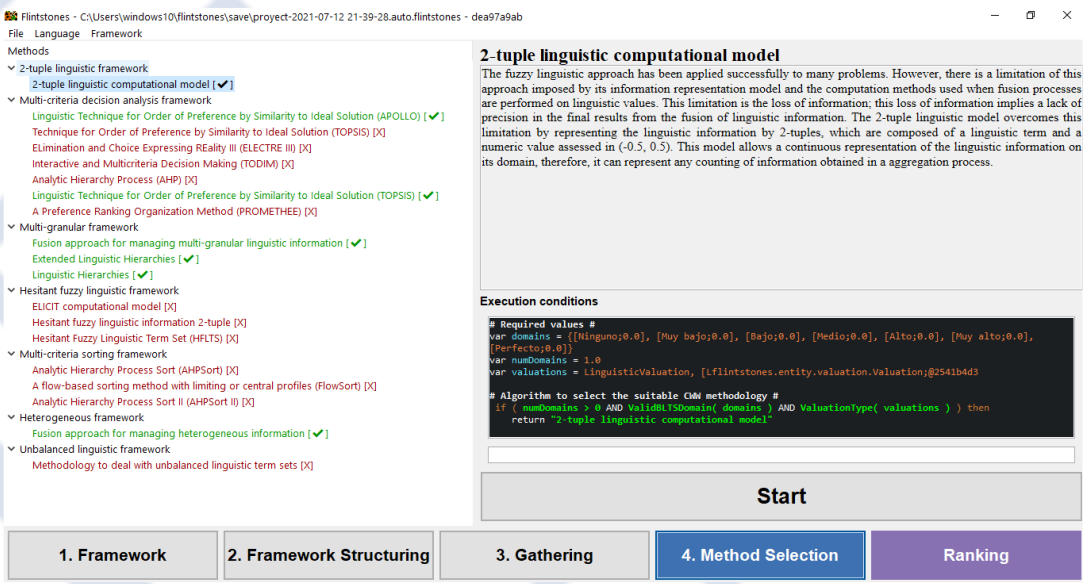

Figura 8. Method.

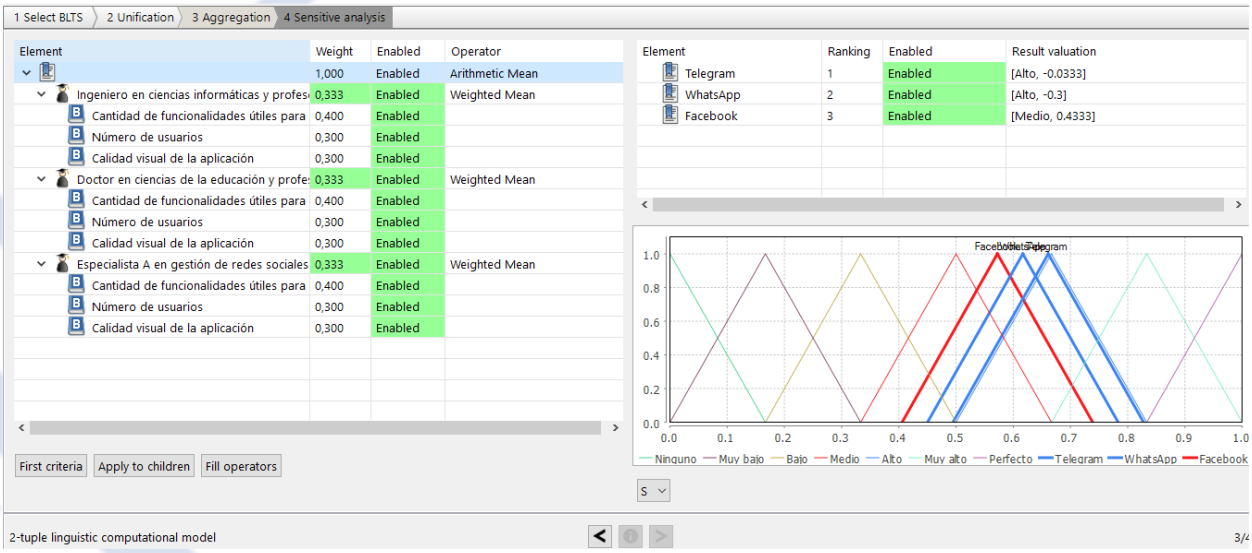

Figura 9. Ranking

CONCLUSIONES

La incorporación de las redes sociales como apoyo a la docencia universitaria a distancia constituye una fortaleza, pues reúne al profesor con estudiantes en contextos muy dinámicos. Cualquiera de las redes candidatas iniciales pudiera ejercer ese rol. Sin embargo, a través de esta investigación se arroja la más adecuada de ellas, resultando ser Telegram.

La computación con palabras, y específicamente el modelo de 2-tuplas lingüísticas, facilitaron una metodología relativamente simple para la toma de decisiones, en la transición hacia la mejor de las tres candidatas. Es un análisis dictaminado por expertos alrededor de la problemática. Teniendo en cuenta que encontrar una decisión unificada en estos contextos no siempre es sencillo; sin embargo, es sin duda una forma de cómputo asequible a los expertos, llevando sus capacidades cognitivas a un lenguaje fácil de entender y operar.

Una vez más se evidencia la superioridad de la programación con palabras en procesos semejantes a este, donde el uso de las palabras proporciona una evaluación simple y certera ante problemáticas de selección.

REFERENCIAS

[1] DÁVILA, Mario Rodrigo Mejía. M-Learning: características, ventajas y desventajas, uso. Revista Tecnológica-Educativa Docentes 2.0, 2020, vol. 8, no 1, p. 50-52. Disponible en: https://ojs.docentes20.com/index.php/revista-docentes20/article/view/80/236.

[2] HITA, María; LÓPEZ, Encarnación Rueda; PALOMINO, María. Posibilidades didácticas de las redes sociales en el desarrollo de las competencias de la educación superior: percepciones del alumnado. Pixel-Bit, Revista de Medios y Educación, 2018, no 53. Disponible en: https://www.academia.edu/download/59448268/1620190530-90867-1at3ptm.pdf.

[3] FERNÁNDEZ, Carmen Sabater; LOREA, Ion Martínez; CAMPIÓN, Raúl Santiago. La Tecnosocialidad: El papel de las TIC en las relaciones sociales. Revista Latina de Comunicación Social, 2017, no 72, p. 1592-1607. Disponible en: https://doi.org/10.4185/RLCS-2017-1236.

[4] DURAN, Diego F.; CHANCHÍ, Gabriel E.; ARCINIEGAS, Jose L. Evaluación de mapas de competencias educativas: una propuesta difusa basada en 2-tuplas. Revista Ibérica de Sistemas e Tecnologias de Informação, 2017, no 24, p. 22-38. Disponible en: https://pdfs.semanticscholar.org/88f8/00f92c24c3225021fba57d722493ccd8ab6f.pdf.

[5] QUIROZ MARTINEZ, Miguel Ángel; ARGUELLO RUIZ, Rodrigo Alexander; GOMEZ RIOS, Mónica Daniela; LEYVA VAZQUEZ, Maikel Yelandi. Evaluación de potencial del internet de las cosas en la salud mediante mapas cognitivos difusos. Conrado [online]. 2020, vol.16, n.75 [citado 2021-07-13], pp.131-136. Disponible en: http://scielo.sld.cu/pdf/rc/v16n75/1990-8644-rc-16-75-131.pdf. Epub 02-Ago-2020. ISSN 2519-7320.

[6] RENTE LABRADA, Rosa María; VALDIVIA MESA, Arianet; VEGA ALMAGUER, Manuel; GONZALEZ HIDALGO, Gilberto Enrique. Computación con palabras en la evaluación del Diseño como instrumento de la Gestión Ambiental. Rev cuba cienc informat [online]. 2021, vol.15, n.1 [citado 2021-07-13], pp.1-19. Disponible en: http://scielo.sld.cu/pdf/rcci/v15n1/2227-1899-rcci-15-01-1.pdf. Epub 31-Mar-2021. ISSN 2227-1899.

[7] ORTIZ, Bolivar Enrique Torres; ESCOBAR, Esperanza Del Pilar Araujo; ANDACHI, Jorge Washington Soxo. Análisis jurídico del abandono de causas tipificado en el Código Orgánico General de Procesos, basada en conjuntos de números de 2-tuplas. Universidad y Sociedad, 2021, vol. 13, no S1, p. 146-156. Disponible en: https://rus.ucf.edu.cu/index.php/rus/article/download/2019/2008.

[8] MORALES, Jeovani M.; MONTES, Rosana; HERRERA, Francisco. Detección del Fracaso Académico y Evaluación de la Práctica Docente mediante la Comunicación Automatizada con un Chatbot. En XVIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2018) 23-26 de octubre de 2018 Granada, España. Asociación Española para la Inteligencia Artificial (AEPIA), 2018. p. 245-250. Disponible en: https://sci2s.ugr.es/caepia18/proceedings/docs/CAEPIA2018_paper_240.pdf.

[9] HERRERA, Francisco; MARTÍNEZ, Luis. A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Transactions on fuzzy systems, 2000, vol. 8, no 6, p. 746-752. Disponible en: https://ieeexplore.ieee.org/abstract/document/890332.

[10] FLINTSTONES, Video-tutoriales. Sitio oficial del grupo investigación Sistemas Inteligentes Basados en Análisis de Decisión Difusos, 2018, [citado 2021-07-13]. Disponible en: https://sinbad2.ujaen.es/flintstones/es/video-tutorials.


LaSalle
Universidad
Perú