Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 141-155
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 19/12/2023| Aceptado: 01/03/2024 | Publicado: 30/03/2024

Identificadores persistentes:
DOI: 10.48168/innosoft.s15.a158
ARK: ark:/42411/s15/a158
PURL: 42411/s15/a158

Aplicación de técnicas de Inteligencia Artificial para la diferenciación del nivel socioeconómico

Application of Artificial Intelligence techniques for the differentiation of the socioeconomic level

Crhistian Ziegler Pacori Paucar[1] [0000-0003-4444-1273]*, **Moises Enrique Mayta Condori[2], Luis Fernando Quispe Sanomamani[3], Diego Gustavo Montana Neyra[4]**

[1] Universidad Nacional de San Agustín. Arequipa, Perú. cpacori@unsa.edu.pe
[2] Universidad Nacional de San Agustín. Arequipa, Perú. mmaytac@unsa.edu.pe
[3] Universidad Nacional de San Agustín. Arequipa, Perú. lquispesan@unsa.edu.pe
[4] Universidad Nacional de San Agustín. Arequipa, Perú. dmontanan@unsa.edu.pe

* Autor para correspondencia: cpacori@unsa.edu.pe

Resumen

En este proyecto se hace una diferenciación entre personas a travez de diferentes parametros como edad,sexo,nivel educativo entre otros,para tratar de calcular a cuanto podria asender su salario. Este problema es importante a resolver por que así una persona podría predecir su futuros ingresos a través de las decisiones que tomaría en el presente, como por ejemplo hasta qué grado de educación debe recibir y cuando ya comenzar a trabajar para obtener experiencia. Nuestro procedimiento para resolver este problema han sido dos análisis estadísticos ,el primero regresión lineal y un árbol de decisión para poder hacer una comparativa entre estos, las hemos probado usando herramientas como Colab (Python) y un dataset. Nuestra población de nuestro trabajo fue de 32000 registros (filas).Los resultados fueron que a través del árbol de decisión hubo una precisión de 0.879 y un accuracy de 0.817 .Y con respecto a la regresión logística obtuvimos una precisión de 0.80 cuando para el sueldo <=50K y 0.72 cuando el sueldo es >50K, el accuracy obtenido es de 0.7912. Dando por conclusión que entre estas dos herramientas nos quedamos con el Árbol de decisión.

Palabras clave: Inteligencia Artificial,árboles de decisión,regresión logística,dataset,nivel socioeconómico.

Abstract

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 141-155
https://revistas.ulasalle.edu.pe/innosoft

In this project, a differentiation is made between people through different parameters such as age, sex, educational level, among others, to try to calculate how much their salary could rise. This problem is important to solve because then a person could predict her future income through the decisions she would make in the present, such as how much education she should receive and when to start working to gain experience. Our procedure to solve this problem has been two statistical analyses, the first linear regression and a decision tree to be able to make a comparison between them, we have tested them using tools such as Colab (Python) and a dataset. Our population for our work was 32,000 records (rows). The results were that through the decision tree there was a precision of 0.88 and an accuracy of 0.82. And with respect to the logistic regression we obtained a precision of 0.80 when for the salary <=50K and 0.72 when the salary is >50K, the accuracy obtained is 0.7912. Concluding that between these two tools we are left with the Decision Tree.

Keywords: *Artificial Intelligence, decision trees, logistic regression, dataset, socioeconomic status.*

1. Introducción

Los ingresos económicos de una persona vendrían a ser las entradas de dinero percibidos de manera regular en un periodo y magnitud constante. Entre ellos están los salarios, pensiones, subsidios, etc. Según [1] el ingreso promedio se calcula por el ingreso nacional bruto y la población. Al dividir todos los ingresos y ganancias anuales entre la cantidad de población del país, mostrará el ingreso promedio per cápita. Se incluyen en esta cantidad todos los sueldos y salarios, pero también otros ingresos no ganados en inversiones o ganancias de capital. El ingreso promedio más alto del mundo se obtiene en las Bermudas. El presupuesto per cápita más bajo existe en Afganistán. En la comparación sobre 67 países, Perú ocupa el 49° lugar con un ingreso anual promedio de 6030 USD y un ingreso mensual promedio de 503 USD.

El Perú,[2] considerado una de las estrellas de crecimiento económico internacional en las dos últimas décadas, se ha convertido ahora en el país con mayor caída del PBI en América Latina, esperándose una contracción de 13.9% hacia finales del año 2020, según el FMI. Este resultado, y la consiguiente destrucción de millones de empleos y el aumento de la pobreza generalizada, nos ha hecho perder en pocos meses todo lo alcanzado en una década de esforzado avance económico. Según [3] datos de la Encuesta Nacional de Hogares (ENAHO), en el segundo trimestre de 2020, la población ocupada disminuyó en más de 6 millones de personas en relación a similar periodo de 2019. Los mayores incrementos en la tasa de desocupación se registraron en hombres, personas entre 25 a 44 años de edad y personas con estudios superiores no universitarios. La disminución de la población ocupada fue mayor en el área urbana (-49,0%) que rural (-6,5%), y en las actividades de construcción (-67,9%), manufactura (-58,2%), servicios (-56,6%) y comercio (-54,5%), principalmente.

Las autoridades políticas y sanitarias de un país limitaron temporalmente la cantidad de contagios o infecciones, restringiendo el funcionamiento de empresas y mercados, y obligando a las personas a permanecer en sus respectivos domicilios [4]. Obviamente, no fue posible mantener a la totalidad de la población confinada, ya que siempre algunas actividades esenciales tienen que seguir funcionando, como la producción de alimentos, el transporte de mercancías, los mercados de abastos, hospitales, farmacias, vigilancia policial, etc. Pero, aparte de este tipo de actividades, el gobierno se encontró ante un dilema. ¿Cuántas y cuáles de las restantes labores no esenciales deben de permanecer cerradas mientras dure la pandemia?

La cuarentena rígida, que obligaba a las personas a permanecer la mayor parte del tiempo en sus respectivos domicilios y obligaba también al cierre de la mayoría de empresas y actividades económicas, duró un poco más de 100 días, desde el 16 de marzo hasta los primeros días de julio [5]. Un mes antes, a inicios de junio hubo una primera apertura de la economía, que permitió la operación de algunos sectores de servicios públicos y otras operaciones de servicios técnicos privados y de distribución o reparto de mercancías y alimentos preparados a domicilio, con lo cual según cifras del MEF solo el 27.2% de la economía nacional permaneció cerrada.

La crisis de la COVID-19 [3] y la consiguiente interrupción masiva de la actividad económica afectó potencialmente a los más de 17,1 millones de trabajadores que conformaban la fuerza laboral peruana en 2019. En base a la ENAHO 2019 y a la metodología de la Organización Internacional del Trabajo (OIT) se estimó que un 40,8% del empleo de Perú se encuentra en sectores de riesgo alto y otro 8,4%, en sectores de riesgo medio-alto, lo cual se reflejó en que estos trabajadores perdieron su empleo y muchos vieron reducidas sus horas de trabajo, con recortes salariales.

El propósito de este artículo, es recolectar la información de la población vulnerable y saber a través de qué medios y cómo obtienen los recursos para mantener sus hogares y así poder brindar una ayuda más especializada y específica a ellos. Para tal efecto se utilizará la Inteligencia Artificial utilizando la Regresión Logística, junto a las librerías de aprendizaje que nos proporciona Python para realizar el análisis de los datos obtenidos.

2. Estado del Arte

Nuestra fundamentación teórica está basada en los siguientes artículos:

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 141-155
https://revistas.ulasalle.edu.pe/innosoft

INVESTIGACIÓN INTERNACIONAL N-1

Ricardo Timarán-Pereira, R. ; Caicedo-Zambrano, J. ; Hidalgo-Troya, A. ;"Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11° "; Revista de Investigación Desarrollo e Innovación: RIDI; Vol.11, Nº.1; 2019, Barcelona, España, doi: 10.19053/20278306.v9.n2.2019.9184

OBJETIVO: Se presentan los resultados obtenidos al aplicar el modelo de clasificación basado en árboles de decisión, con el fin de detectar factores asociados al desempeño académico de los estudiantes colombianos de grado undécimo de educación media,se seleccionó,la información socioeconómica, académica e institucional de estos estudiantes.Y se generaron árboles de decisión que permitieron identificar patrones asociados al buen o mal desempeño académico de los estudiantes en las pruebas para mejorar la calidad de la educación en Colombia.

MUESTRA: Las encuestas realizadas en KDNuggets en 2002,2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada.La esta metodología para proyectos de minería de datos no es la "más actual" o "la mejor", pero es muy útil para comprender esta tecnología o extraer ideas para diseñar o revisar métodos de trabajo para proyectos de similares características.CRISP-DM es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de minería de datos y contempla seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación.

CONCEPTOS CLAVES QUE SE ESTÁ ANALIZANDO:

Árboles de decisión: Diagrama en forma de árbol que muestra la probabilidad estadística o determina un curso de acción. Muestra a los analistas y, a los que toman las decisiones, qué pasos deben tomar y cómo las diferentes elecciones podrían afectar todo el proceso.

Nivel Socioeconómico: Descripción de la situación de una persona según la educación, los ingresos y el tipo de trabajo que tiene. El nivel socioeconómico por lo general se define como bajo, medio o alto.

Calidad de la educación: La calidad del sistema educativo es la cualidad que resulta de la integración de las dimensiones de pertinencia, relevancia, eficacia interna, eficacia externa, impacto, suficiencia, eficiencia y equidad.

INVESTIGACIÓN INTERNACIONAL N-2

Rodríguez Garcés, C.Sandoval Muñoz, D.; "Consumo tecnológico: Análisis de los determinantes del equipamiento doméstico mediante Árboles de Decisión" Revista Internacional de Investigación en Ciencias Sociales, Vol. 11, Nº. 1, 2015 ,Chile,2015,págs. 70-85

OBJETIVO: Se aprovecha la base de datos del Centro de aprendizaje de Chile,haciendo un análisis de tendencia a través de un árbol de decisión acerca de los niveles de penetración de tecnología doméstica y factores diferenciadores.

Los resultados muestran que, a pesar de cierta diferenciación por tipo de dispositivo y del perfil de usuario, la rápida y masiva integración de dispositivos tecnológicos se ha dado según el nivel socioeconómico y también que es el vector de mayor segmentación.

Lo observado ha revelado que los dispositivos bien integrados, como teléfonos celulares; denotan un mayor poder adquisitivo,como se hizo notar también en la tecnología del pasado, como la televisión por cable o satélite.

MUESTRA: El universo de estudio está definido por la población de 18 años y más de zonas urbanas y rurales. El muestreo es probabilístico estratificado por conglomerados múltiples entrevistando a una muestra de alrededor de 1.500 personas en cada año, con un error de muestreo del +3% y un nivel de confianza del 95%, estableciéndose un margen de respuesta efectiva promedio para todos los años cercano al 85%.

CONCEPTOS CLAVES QUE SE ESTÁ ANALIZANDO:

Árboles de decisión: Mapa de los posibles resultados de una serie de decisiones relacionadas. Permite que un individuo o una organización comparen posibles acciones entre sí según sus costos, probabilidades y beneficios. Se pueden usar para dirigir un intercambio de ideas informal o trazar un algoritmo que anticipe matemáticamente la mejor opción.

Nivel Socioeconómico: Es una medida total económica y sociológica que combina la preparación laboral de una persona, de la posición económica y social individual o familiar en relación a otras personas, basada en sus ingresos, educación y empleo.

Calidad de la educación: Implica una búsqueda de constante mejoramiento en todos sus elementos, en insumos (recursos disponibles en las escuelas), procesos de enseñanza (tiempo destinado a la enseñanza escolar, cantidad de tareas y estipulaciones curriculares) y en los productos (logros estudiantiles).

INVESTIGACIÓN INTERNACIONAL N-3

3-Blanca CUJI; Wilma GAVILANES; Rina SANCHEZ;"Modelo predictivo de deserción estudiantil basado en árboles de decisión";Revista Espacios; Vol. 38 (Nº 55) Año 2017. Pág. 17 Colombia

OBJETIVO: Muestra la construcción de un modelo predictivo de deserción estudiantil, para pronosticar la probabilidad, que un estudiante abandone su programa académico, mediante técnicas de clasificación, basadas en árboles de decisión. Se construyó un árbol con cuatro niveles de profundidad y mismo número reglas, que evalúan a los posibles desertores. Llevando a concluir que las variables nivel y notas tienen mayor influencia en la deserción.

MUESTRA: Se tomó datos, de 485 estudiantes, almacenados en hojas de cálculo y base de datos relacionales, de la DITIC. Los datos fueron transformados, a variables, según los tipos de atributos propuestos por el estadístico S. Stevens (1946), se clasificaron en tres tipos: nominal, ordinal y cuantitativo,estos fueron:Género,estado civil,etnia,edad,lugar de nacimiento,ciudad de residencia,nivel.

CONCEPTOS CLAVES QUE SE ESTÁ ANALIZANDO:

Árboles de decisión: Permite evaluar mediante una representación gráfica los posibles resultados, costos y consecuencias de una decisión compleja. Este método es muy útil para analizar datos cuantitativos y tomar una decisión basada en números

Nivel Socioeconómico: La condición socioeconómica, una medida de situación social que incluye típicamente ingresos, educación y ocupación, está ligada a una amplia gama de repercusiones de la vida, que abarcan desde capacidad cognitiva y logros académicos hasta salud física y mental.

Calidad de la educación: Está determinada por los conocimientos y competencias por las que se adquieren el reconocimiento a los derechos humanos. Para avanzar en la mejora de la calidad educativa es necesario integrar las aptitudes, la innovación educativa, la eficiencia y la igualdad.

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 141-155
https://revistas.ulasalle.edu.pe/innosoft

INVESTIGACIÓN INTERNACIONAL N-4

EMMANUEL VAZQUEZ ”SEGREGACIÓN ESCOLAR POR NIVEL SOCIOECONÓMICO: MIDIENDO EL FENÓMENO Y EXPLORANDO SUS DETERMINANTES”

OBJETIVO: Proveer una cuantificación de los niveles y la evolución de la segregación escolar por nivel socioeconómico en el mundo y contribuir a la discusión de sus determinantes.

Muestra:Este trabajo utiliza una base de datos producida por el Programa para la Evaluación Internacional de Estudiantes (PISA) como fuente de información. La primera prueba PISA se realizó en 2000 con la participación de 43 países. 7 La segunda (2003) se realizó en 41 países, la tercera (2006) en 57 países, la cuarta (2009) y quinta (2012) en 65,8 y la sexta edición (2015) en 72. De hecho, además de los países miembros de la OCDE, cada vez más países de diferentes partes del mundo se han sumado a la iniciativa, ampliando la cobertura del programa. En 2015, un total de 542 385 estudiantes en 18 602 escuelas completaron las evaluaciones, lo que representa casi 27 millones de estudiantes en todo el mundo.

CONCEPTOS CLAVES QUE SE ESTÁ ANALIZANDO:

Nivel Socioeconómico: Situación de una persona según la educación, los ingresos y el tipo de trabajo que tiene.

Calidad de la educación: Funcionamiento en los centros educativos que permite tener un control de todos los procesos llevados a cabo en los mismos, así como la correcta gestión de éstos.

3. Marco teórico

3.1 Inteligencia Artificial:

Según Winston[6] la inteligencia artificial está definida como un estudio de computación que hace posible percibir, razonar y actuar. En el campo de la ingeniería artificial es resolver problemas del mundo real utilizándose como un arsenal de ideas sobre la representación del conocimiento, el uso del conocimiento y el montaje de sistemas

3.2 Árboles de decisión:

Según Fletcher et al. [7] Los árboles de decisión son un método de aprendizaje supervisado no paramétrico utilizado para clasificación y regresión. No hacen suposiciones sobre la distribución de los datos subyacentes y están capacitados en datos etiquetados para clasificar correctamente los datos no vistos anteriormente.

3.3 Regresión Logística:

Según Dominguez et al. [8] Los modelos de regresión logística son modelos estadísticos en los que se evalúa la relación entre una variable cualitativa dependiente, dicotómica (regresión logística binaria o binomial) o variable con más de dos valores (regresión logística multinomial). Una o más variables explicativas independientes, o co-variables, ya sean cualitativas o cuantitativas.

3.4 Proceso de la IA:

Describir el proceso de transformación de datos.

a) Limpieza de datos

Según Lomet [9] la limpieza de datos, también llamada limpieza o depuración de datos, se ocupa de detectar y eliminar errores e inconsistencias de los datos para mejorar la calidad de los datos. Los problemas de calidad de los datos están presentes en recopilaciones de datos individuales, como archivos y bases de datos, por ejemplo, debido a faltas de ortografía durante el ingreso de datos, falta de información u otras colecciones, como archivos y bases de datos, por ejemplo, debido a faltas de ortografía durante la entrada de datos, falta de información u otros

b) Transformación de datos

Según Astera [10] la transformación de datos es el proceso de convertir datos de un formato a otro formato que sea más utilizable por el sistema o la aplicación de destino. Incluye múltiples actividades: puede 'transformar' sus datos filtrándose según ciertas reglas y uniendo diferentes campos para obtener una vista consolidada. Las herramientas de transformación ayudan a lograr su resultado final con facilidad.

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 141-155
https://revistas.ulasalle.edu.pe/innosoft

c) Cargar los datos

En esta etapa [11], los datos procedentes de la fase anterior (fase de transformación) son cargados en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes.

d) Eliminación de Outliers

En esta etapa[12] se eliminan los Outliers(valores atípicos) que son observaciones que se desvían tanto de otras observaciones como para despertar la sospecha de que fue generada por un mecanismo diferente.

4. Herramientas

4.1 Dataset:

Los datos usados tienen los siguientes campos:

Age: edad de la persona

Workclass : la clase de trabajo del individuo

fnlwgt: el peso del muestreo

Education: el grado de educación de la persona

Education-num: número de años de educación en total

Marital-status: estado civil del individuo.

Occupation: la ocupación/trabajo que desempeña el individuo

Relationship: el tipo de relación familiar

Race: la raza del individuo.

Sex: el género del individuo.

Capital-gain: ingresos ganados de fuentes de inversión que no son sueldos/salarios

Capital-loss: ingresos perdidos de fuentes de inversión que no son sueldos/salarios

Hours-per-week: horas de trabajo por semana

Native-country: Ciudad de nacimiento

El Dataset está compuesto por 32561 entradas sin valores nulos, de los cuales 24720 entradas son <= 50K y 7841 entradas son > 50K.

4.2 Colab:

Colab,[13] también conocido como "Colaboratory", permite programar y ejecutar Python en el navegador con las siguientes ventajas: No requiere configuración, Da acceso gratuito a GPUs, Permite compartir contenido fácilmente. Colab puede facilitar el trabajo de estudiantes, científicos de datos o investigadores de IA.

4.3 Librerías:

pandas: ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales. Es un software libre distribuido bajo la licencia BSD versión tres cláusulas

numpy: da soporte para crear vectores y matrices grandes multidimensionales, junto con una gran colección de funciones matemáticas de alto nivel para operar con ellas.

sklearn: Cuenta con varios algoritmos de clasificación , regresión y agrupamiento , que incluyen máquinas de vectores de soporte , bosques aleatorios , aumento de gradiente , k -means y DBSCAN , y está diseñado para interactuar con las bibliotecas numéricas y científicas de Python NumPy y SciPy

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 141-155
https://revistas.ulasalle.edu.pe/innosoft

5. Resultados Obtenidos:

INTERPRETACIÓN DE NUESTROS RESULTADOS

ARBOL DE DECISION

1. Metricas :

Confusión Matrix:

[[4346 614]

[560 993]]

Precision: 0.88; 0.61

Recall: 0.88; 0.63

F1 - Score: 0.88; 0.63

Accuracy: 0.82

True Positives(TP) = 4346

True Negatives(TN) = 993

False Positives(FP) = 560

False Negatives(FN) = 614

Revista Innovación y Software
Vol. 5, No. 1, Mes Marzo-Agosto, 2024
ISSN: 2708-0935
Pág. 141-155
https://revistas.ulasalle.edu.pe/innosoft

REGRESION LOGISTICA

1. Precisión : 0.80; 0.72

2. Accuracy: 0.792199815743679

3. Confusion matrix

[[7096 256]
[1774 643]]

True Positives(TP) = 7136
True Negatives(TN) = 594
False Positives(FP) = 1762
False Negatives(FN) = 277

La matriz de confusión muestra 7136 + 594 = 7730 predicciones correctas, y 1762 + 277 = 2039 predicciones incorrectas.

	precision	recall	f1-score	support
<=50K	0.80	0.97	0.87	7352
>50K	0.72	0.27	0.39	2417
accuracy			0.79	9769

Conclusiones:

1- Podemos ver que la puntuación de precisión de nuestro modelo en Árbol de Decisión es 0.87968241 y R. Logística 0.80. Entonces, podemos concluir que nuestro modelo de Árbol de Decisión está haciendo un mejor trabajo al predecir.

2- Dado que la Precisión identifica la proporción de resultados positivos predichos correctamente Para nuestro caso la precisión obtenida con el Árbol de decisión es de 0.88; esto implica que de cada 100 personas, a 88 las clasifica con una predicción correcta y al resto no.

En Regresión Logística Precisión nuestra precisión es de 0.80 quiere decir que nuestra prediccion es correcta en un 80% para personas que perciben ingresos menores e iguales a $50K y 0.68(68%) para los que ganan más de $50K.

3- A nivel de las recall y F1 el Árbol de Decisión tiene valores mayores con respecto a la Regresión Logística.

4- Dado que nuestro Recall en ambos casos son mayores al 0.8 , 87% en Árbol de Decisión y 97% R. Logística concluimos que de nuestros resultados predichos correctamente abarcan la mayoría de los resultados positivos reales.

5- Ya que ninguno de nuestros F1-score llega a 1.0 pero están por encima de 0.8 podríamos decir que nuestros valores solo son buenos.

Contribución de Autoría

Crhistian Ziegler Pacori Paucar: Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Moises Enrique Mayta Condori:** Conceptualización, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos. **Luis Fernando Quispe Sanomamani:** Conceptualización, Análisis formal, Investigación, Visualización, Metodología, Software, Validación, Redacción - borrador original, Curación de datos, Escritura, revisión y edición. **Diego Gustavo Montana Neyra:** Visualización, Software, Validación, Redacción - borrador original, Curación de datos.

Referencias

[1] "Ingresos promedio a nivel mundial." https://www.datosmundial.com/ingreso-promedio.php .

[2] J. Vega, "Departamento de economía," *Pontif. Univ. Católica del Perú*, p. 25, 2020, [Online]. Available: https://repositorio.pucp.edu.pe/index/handle/123456789/176236

[3] J. Gamero and J. Pérez, "Perú: Impacto de la COVID - 19 en el empleo y los ingresos laborales," *Organ. Int. de Trab. Panor. Labor. en tiempos la COVID- 19*, vol. I, no. I, p. 64, 2020, [Online]. Available: https://www.ilo.org/wcmsp5/groups/public/---americas/---ro-lima/documents/publication/wcms_756474.pdf

[4] "Decreto Supremo N° 051-2020-PCM" https://cdn.www.gob.pe/uploads/document/file/572157/DECRETO_SUPREMO_N%C2%BA_051-2020-PCM.pdf (accessed Jun. 27, 2022).

[5] "Decreto Supremo N° 116-2020-PCM" https://cdn.www.gob.pe/uploads/document/file/898487/DS_116-2020-PCM.pdf

[6] Patrick Henry Winston, *Artificial Intelligence*, 3rd ed., vol. 110, no. 5. Addison-Wesley Publishing Company, 1993.

[7] S. Fletcher and M. Z. Islam, "Decision tree classification with differential privacy: A survey," *ACM Comput. Surv.*, vol. 52, no. 4, 2019, doi: 10.1145/3337064.

[8] S. Domínguez-Almendros, N. Benítez-Parejo, and A. R. Gonzalez-Ramirez, "Logistic regression models," *Allergol. Immunopathol. (Madr).*, vol. 39, no. 5, pp. 295–305, 2011, doi: 10.1016/j.aller.2011.05.002.

[9] D. B. Lomet, "Bulletin of the Technical Committee on Data Engineering," *Bull. Tech. Comm. Data Eng.*, vol. 24, no. 4, pp. 1–56, 2001, [Online]. Available: papers2://publication/uuid/30073F7F-1B7C-4496-ADA4-94FF4E6EE8F7

[10] "Transformación de datos y por qué es importante para las empresas | Astera." https://www.astera.com/es/type/blog/data-transformation-tools/

[11] "ETL: Extracción, transformación y carga de datos - Evaluando Software." https://www.evaluandosoftware.com/etl-extraccion-transformacion-carga-datos/

[12] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "OPTICS-OF: Identifying local outliers," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 1704,

pp. 262–270, 1999, doi: 10.1007/978-3-540-48247-5_28.

[13] “Te damos la bienvenida a Colaboratory - Colaboratory.” https://colab.research.google.com/?hl=es#scrollTo=5fCEDCU_qrC0.