Artículo original/ Original article
ISSN 2953-6553

Implementación de un Modelo de Regresión Lineal Múltiple aplicando (PLN) para predicción de artículos científicos.

Implementation of a Multiple Linear Regression Model applying (PLN) for prediction of scientific articles.

David Crespo-Campoverde[1] 0009-0006-3127-6626, Juan Campoverde-Villalta[2] 0009-0000-5371-1402,
Wilson Sánchez-Bermeo[3] 0009-0002-5515-9986, Diana Romero-Córdova[4] 0000-0001-8456-4660
david.crespo.est@tecazuay.edu.ec, juan.campoverde.est@tecazuay.edu.ec,
wilson.sanchez.est@tecazuay.edu.ec, diana.romero@tecazuay.edu.ec
Instituto Superior Universitario Tecnológico del Azuay, Cuenca, Ecuador

DOI 10.36500/atenas.2.008

Resumen

En este artículo se explica la importancia de la regresión lineal múltiple en diferentes aplicaciones, como el procesamiento del lenguaje natural y la predicción de series de tiempo. Se destaca su relevancia en la creación de sistemas de tutoría inteligente que se adapten a las necesidades individuales de los estudiantes. Se presenta el marco teórico sobre los modelos de regresión, desde la regresión lineal simple hasta la regresión lineal múltiple, y se describen los diferentes tipos de variables que intervienen en el modelo. Además, se discuten los tipos de datos, técnicas de preparación de datos, evaluación del modelo y metodología utilizada en un trabajo de análisis de datos. Se aplicó la metodología de crisp-dm, que se divide en seis fases: recolectar datos, preparar los datos, modelar, evaluar, implementar y mantener. Se describe el proceso de recolección de datos y etiquetación de la data de IoT, la carga y visualización de la base de datos, y las técnicas utilizadas en la limpieza y transformación de variables. También se explican algunas técnicas de preparación de datos y medidas comunes utilizadas para evaluar la calidad del ajuste del modelo.

Abstract

This article explains the importance of multiple linear regression in different applications, such as natural language processing and time series forecasting. Its relevance in the creation of intelligent tutoring systems that adapt to the individual needs of students is highlighted. The theoretical framework on regression models is presented, from simple linear regression to multiple linear regression, and the different types of variables involved in the model are described. In addition, the types of data, data preparation techniques, model evaluation, and methodology used in data analysis work are discussed. The crisp-dm methodology was applied, which is divided into six phases: collect data, prepare the data, model, evaluate, implement and maintain. The process of data collection and labeling of the IoT data, the loading and visualization of the database, and the techniques used in the cleaning and transformation of variables are described. Some common data preparation techniques and measures used to assess the quality of model fit are also explained.

Palabras Claves – Regresión lineal, **Análisis de texto, Clasificación de texto, Aprendizaje Automático, Predicción de resultados, Modelos de regresión.**

Recibido: 2023-05-05, Aprobado tras revisión: 2023-11-10

Keywords– Linear regression, **Text analysis, Text classification, Machine learning, Prediction of results, Regression models**

I. INTRODUCCIÓN

La regresión lineal múltiple es una técnica de aprendizaje automático que se utiliza en una variedad de aplicaciones, como el procesamiento del lenguaje natural, la minería de datos y la predicción de series de tiempo. El conocimiento de la regresión lineal múltiple es importante para los científicos de datos y los ingenieros informáticos que trabajan en estas áreas, la regresión lineal múltiple se puede utilizar en modelos de clasificación para predecir la probabilidad de que una observación pertenezca a una categoría en particular.

Los algoritmos de regresión lineal múltiple son esenciales en el procesamiento del lenguaje natural (PLN) porque permiten modelar la relación entre múltiples variables de entrada y una variable de salida continua. En el contexto del PLN, estas variables de entrada pueden representar características lingüísticas, como el número de palabras en una oración, la frecuencia de ciertas palabras, la complejidad sintáctica y semántica, entre otras. Al aplicar la regresión lineal múltiple a datos lingüísticos, se pueden realizar análisis cuantitativos y predecir el valor de la variable de salida en función de las variables de entrada.

La capacidad de predecir el rendimiento del lenguaje puede tener importantes aplicaciones prácticas, como en el diseño de sistemas de tutoría inteligente que pueden adaptarse a las necesidades individuales de los estudiantes.

II. MARCO TEÓRICO

Perspectiva general sobre los modelos de regresión

En este apartado se explican los aspectos claves de los modelos de regresión usados más adelante. Se divide en dos secciones: Regresión Lineal Simple y Regresión Lineal Múltiple.

El valor del modelo de RLS.

Las RLS es un componente del aprendizaje supervisado, permite predecir cantidades continuas a partir de datos históricos y etiquetados. Este aprendizaje utiliza un tipo de modelo que correlaciona una relación entre ciertas características Xcontinuas con una variable objetivo Ycontinua, su respuesta an una variable dependiente identificada con Y, con la condición de que se de al menos una variable independiente X, representada a través de una recta Y= f(X), es decir, de una o más variables dependientes describidas en valores de la combinación de forma basado en la cantidad de variables independientes xi que se relacionan con el modelo mencionado anteriormente. (Roque López, 2021)

Este tipo de modelo (RLS) hace referencia en la expresión de dependencia a la lineal de su variable meta o dependiente además de la variable referente a la independiente o explicativa xy de la ecuación del error o desviación ε del modelo. Donde: y_t es la variable dependiente, explicada o endógena. X_t es una variable independiente, muy explicativa o también exógena.

Figura 1

fórmula matemática de regresión lineal simple

$$y_t = \alpha + \beta x_t + \varepsilon_t \quad \text{con } t = 1,...,T$$

Determinación del modelo de Regresión Múltiple

Para desarrollar la técnica, es fundamental y muy importante el comprender tanto la variable llamadas dependiente como las variables llamadas independientes, las cuales deben ser continuas. Como resultado, esto permitirá utilizar la técnica principalmente para relacionar una variable dependiente continua con un conjunto de variables categóricas; o para relacionar una variable dependiente nominal con una muestra de valores que se consideran variables continuas. Este modelo de regresión lineal múltiple bien conocido considera una serie de

variables a tomar y afecta cómo se relacionan con los valores de la tercera variable, o reacciones. (Montero, 2016)

A continuación se procede a especificar cada una de las variables, para determinar el modelo multivariante o Regresión múltiple, en el presente documento.

Se definen de la siguiente forma: y=variable dependiente, a= variables independientes cuando son igual a cero, bi = p son los valores de los coeficientes parciales de regresión de cada una de las predictoras, xi= por cada unidad de variación de la variable predictora mantiene las demás predictoras constantes.

Figura 2

fórmula matemática de regresión lineal múltiple

$$y_j = b_0 + b_1 x_{1j} + b_2 x_{2j} + \cdots + b_k x_{kj} + u_j,$$

Tipos de variables

Proponiendo a nuestro MRLM quedaría a la forma siguiente Y hace referencia a variables endógenas dependientes, X son variables exógenas independientes. (Junco, 2021)

La regresión lineal que se pretende aplicar, cuenta con variables categóricas para analizar, se debe realizar un tratamiento adecuado respecto de la naturaleza de este tipo de información.

Las variables en mención permiten darle una interpretación a las n categorías que se tienen en una variable a través de la asignación de n – 1 coeficientes, los cuales indican ausencia o presencia de cada categoría. Se considera n – 1 porque la ausencia de todas las variables igualmente se traduce en la presencia de otra variable, ya que, si estuviera representada como otro coeficiente, se trataría de una colinealidad de variables. (Montero, 2016)

En el análisis de esta regresión llamada múltiple se tiene que tomar en cuenta los mismos aspectos que en una regresión simple: Como la validez y ajuste del modelo, ecuación de

regresión y el análisis de los supuestos. (Portugal, 2020)

Según Baños Ruth (2019) los supuestos que se deben cumplir son: Linealidad: La relación entre variables debe ser lineal, Independencia: Los errores en la medición de las variables explicativas sean independientes entre sí, Homocedasticidad: Los errores tienen varianza constante, Normalidad: Las variables sigan la Ley Normal, No colinealidad: Las variables independientes no estén correlacionadas entre ellas. El objetivo de la regresión múltiple es obtener valores de parámetros de regresión que minimicen la suma de errores al cuadrado o residuos de manera similar a la regresión simple con la finalidad de optimizar las predicciones.

METODOLOGÍA CRISP-DM

Esta metodología es ideal para proyectos de ciencia de datos, cuenta con seis fases iterativas para el desarrollo las cuales son: Entendimiento del negocio, Entendimiento de datos, Preparación de datos, Construcción del modelo, Evaluación del modelo, Despliegue del modelo (Jiménez, 2023).

Utilizada en entorno al proyecto: Recolectar datos: Identifica el objetivo de este artículo, para resolver el problema planteado. Preparación de los datos: Se limpian los datos, se integran y se transforman para prepararlos para su uso en el modelo. Elegir el modelo: En esta fase, se seleccionan y se construyen modelos de regresión lineal múltiple. Entrenar la máquina: Se eligen las variables que utilizaremos para entrenar nuestro modelo. Evaluación: Se evalúan los modelos de regresión lineal múltiple y se determina el mejor modelo.

Tipos de datos

Existen varios tipos de datos que se utilizan en programación y análisis de datos. A continuación, se mencionan algunos de los tipos de datos más comunes: Números enteros (int): Son números enteros, sin decimales. Números de punto flotante (float): Son los números con decimales. Cadenas de texto (string): Es una secuencia de caracteres que se utilizan para representar texto, como palabras y frases. Booleanos (bool): se refiere a valores que pueden ser verdadero o falso (True o False), y se utilizan para representar condiciones lógicas.

Técnicas de preparación de los datos

Las técnicas de preparación de datos son procesos que se realizan para limpiar, transformar y dar formato a los datos para que puedan ser analizados de manera efectiva. Algunas técnicas comunes son: Limpieza de datos: Es la identificación y eliminación de errores y valores atípicos en los datos, así como también la imputación de datos faltantes. Transformación de datos: Se refiere a la conversión de los datos a un formato más adecuado para su análisis. Integración de datos: Combina distintos datos de varias fuentes en un conjunto de datos. Reducción de datos: La eliminación de variables redundantes o irrelevantes para el análisis. (Peña, 2017).

Evaluación del Modelo

Para la evaluación de la calidad del ajuste del modelo se utilizó el MSE (error cuadrático medio) y el R2 (coeficiente de determinación). El MSE mide la diferencia promedio entre los valores observados y los valores predichos por el modelo. El R2 mide la proporción de la varianza total en la variable dependiente que se puede explicar por el modelo. (Roque López, 2021)

Coeficientes e Intersecciones

Los coeficientes representan el cambio en la variable dependiente por unidad de cambio en una variable independiente, manteniendo todas las demás variables constantes. La intersección representa el valor de la variable dependiente cuando todas las variables independientes son iguales a cero. (Roque López, 2021)

Gráfico de residuos

El gráfico de residuos es una herramienta utilizada para evaluar la calidad del ajuste del modelo. Los residuos son las diferencias entre los valores observados y los valores predichos por el modelo. Un gráfico de residuos muestra los residuos en el eje "y" con los valores predichos en el eje x.

III. METODOLOGÍA

Para la metodología se aplicó la metodología de crisp-dm

Paso 1: Recolectar Datos

Para el siguiente paso, se revisaron varios artículos que contenían información sobre IoT de en un repositorio online el cual esta información se presentó en hojas de cálculo, la información en algunos casos estaba incompleta o con errores. Dado este inconveniente, se unificó la data de las diferentes bases de datos en un solo archivo de origen .csv.

Con las bases de datos unificada se procedió a realizar una verificación de los artículos para saber si tenían temática de IoT y en qué áreas del mismo se pertenecían, lo que se realizó para

saber si estos artículos eran sobre Iot se procedió a leer las conclusiones y los resúmenes de los mismo y de esa manera poder identificar cuales trataban del tema de IoT.

Con la data ya unificada se procedió a eliminar los artículos que eran de origen Scopus puesto que no presentaban información muy confiable y de igual manera se eliminaron aquellos que tenían información incompleta. Después de la eliminación de los respectivos artículos, quedaron 280 artículos en total, con los que se procedió a trabajar. Concluido lo anterior se procedió con la etiquetación de la data. Las palabras que se colocaron en la etiquetación fueron seleccionadas del abstract y las conclusiones que presentan los artículos, en la data de IoT se obtuvieron 30 palabras que poseían relevancia sobre el tema a tratar.

Paso 2: Preparar los datos

Carga de la Base de datos

Para empezar nuestro trabajo se realizó la carga de la base de datos de origen csv. en una variable con nombre data y de esta manera poder trabajar de una mejor manera.

Visualización de los Datos

En la exploración de datos, se realizó una exploración de los datos numéricos y categóricos de nuestra data al igual que se realizó una exploración de los tipos de valores.

Preprocesamiento de la data

Para poder realizar un buen preprocesamiento de datos se utilizó diferentes técnicas las cuales fueron limpieza de los datos y adicionalmente la transformación de las variables.

Limpieza de Datos.

Se prepararon los datos, con la finalidad de eliminar aquellos datos nulos y faltantes en la data teniendo con variables con más datos nulos Extracción de datos y Análisis Exploratorio de Datos (EDA).

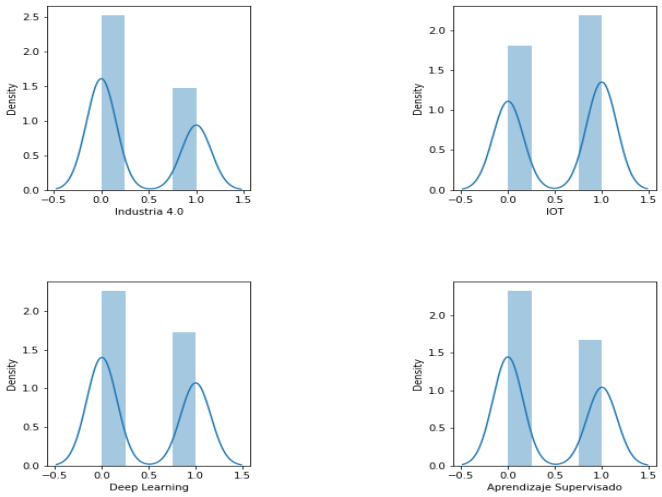
Transformación de variables.

Se realizó una transformación de variables por lo que se procedió a transformar una variable la variable de descarga directa desde fuente oficial (SI/NO) que era de tipo "object". Lo que se realizó fue una codificación numérica dado que la variable que se escogió tenía valores de Si y No dentro de ella, para cambiar esos valores a valores numéricos. Siguiendo en la transformación de variables se realizaron dos cambios en dos variables que eran de origen Float64, las cuales por medio de otro código se transformaron a valores int64.

Una vez se concluyó con el preprocesamiento se realizó un nuevo código con el fin de conocer aquellas variables con más datos dentro de ellas, teniendo como resultado a las siguientes columnas: 'IOT', 'Deep Learning', 'Aprendizaje Supervisado', 'Industria 4.0'. Se realizó una gráfica con las variables antes mencionadas comparándolas con la variable de IOT y de esta manera poder observar cómo se representan con nuestra variable, en la gráfica se muestran dos columnas las cuales corresponden a los valores de si (Pertenece al artículo) o no (No pertenece al artículo) que están representadas como 0 y 1 respectivamente.

Figura 3

Gráfica de las variables con más datos.



Nota. En la gráfica se muestran las variables con más datos en nuestra base de datos. Fuente: Elaborado por: Los autores.

Una vez concluido todo el preprocesamiento se guardaron todos los datos ya procesados en una nueva variable con nombre de data IoT procesada la cual se volvió a cargar en la variable de data para comenzar con el entrenamiento.

Paso 3: Elegir el modelo

En este paso se procedió a escoger un modelo el cual encaje bien con nuestro objetivo planteado para trabajar la data, siendo este el modelo de regresión lineal múltiple.

Paso 4: Entrenar la máquina

Una vez planteado el modelo se procedió a realizar el entrenamiento, como primer paso se definieron las variables a trabajar X, "y" las cuales tendrán como valores las columnas con mayor cantidad de datos.

De esta manera "X" posee los valores de las columnas de Deep Learning, Aprendizaje Supervisado, Industria 4.0 y en cambio la variable "y" posee los valores de la columna de

IOT. Ya con las variables X, y definidas se procedió a realizar el entrenamiento de las variables, en si lo que se realizó fue dividir el conjunto de datos original (‘X’ e ‘y’) en dos conjuntos separados para poder entrenar en el conjunto de entrenamiento y luego evaluar su rendimiento en el conjunto de prueba.

Realizado el entrenamiento se procedió a realizar un escalamiento de los datos con el objetivo de asegurar que todas las características se encuentren en la misma escala.

Por lo tanto, se utiliza una técnica de normalización de características como la estandarización. Esto asegura que las mismas transformaciones se aplican a los datos de entrenamiento y prueba, lo que garantiza una comparación justa de los resultados de la regresión lineal múltiple. Al momento de realizar el entrenamiento del modelo se utilizó un comando para entrenar un modelo de regresión lineal múltiple.

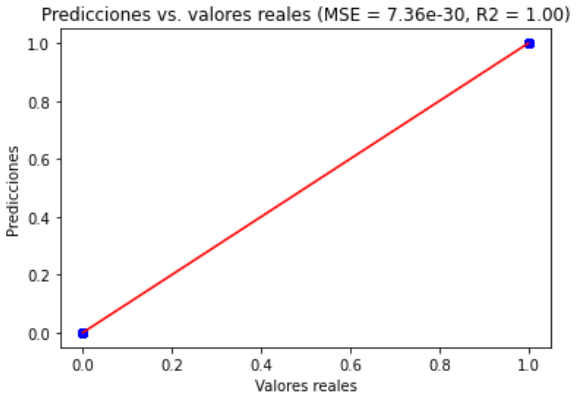
Paso 5: Evaluación

Concluido el entrenamiento se empleó un código que se utiliza para realizar predicciones utilizando un modelo de regresión que ya ha sido entrenado en un conjunto de datos, el código realiza la tarea de hacer predicciones utilizando un modelo de regresión previamente entrenado.

Con estos parámetros sacamos como resultados: MSE: 7.361968094464915e-30 el valor del MSE presentado es extremadamente pequeño, lo que indica que el modelo tiene un ajuste perfecto a los datos de entrenamiento, R2: 1.0 en este caso, lo que indica que el modelo explica el 100% de la variabilidad en la variable objetivo y tiene un ajuste perfecto a los datos de entrenamiento.

Figura 4

Resultados de MSE y R2

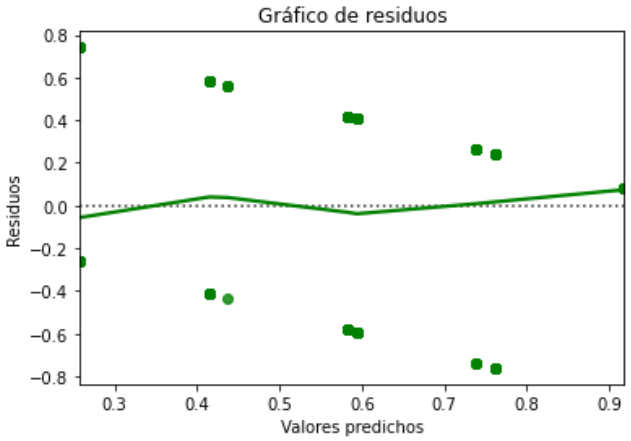


Nota. En la gráfica se puede observar los resultados de MSE y R2. Fuente: Elaborado por: Los autores.

A Continuación, se realizó un código para observar los coeficientes y la intersección en el resultado. Los coeficientes son todos muy cercanos a cero, lo que puede indicar que el modelo no tiene una buena capacidad predictiva. Sin embargo, el valor de la intersección es de 0.528, lo que significa que todas las variables independientes sean o se acerquen a cero. Para finalizar en el análisis del modelo se utilizó la gráfica de residuos que es una herramienta útil de la regresión lineal múltiple.

Figura 5

Gráfico de residuos.



Nota. Gráfica comparativa de valores predichos y residuos. Fuente: Elaborado por: Los autores.

Resultados

Se logró un manejo adecuado del modelo de regresión lineal y un análisis sólido de las variables planteadas durante el proceso de entrenamiento y preprocesamiento. Los resultados del modelo son altamente prometedores, ya que se obtuvo un MSE de 7. Esto sugiere un ajuste prácticamente perfecto a los datos de entrenamiento. Además, el coeficiente de determinación R2 alcanzó el valor de 1.0, lo que indica que el modelo es capaz de explicar el 100% de la variabilidad en la variable objetivo y presenta un ajuste excepcional a los datos de entrenamiento.

Un aspecto importante a destacar es que, en la gráfica de residuos, se observan valores dispersos de manera aleatoria. Esta dispersión aleatoria es una señal positiva que indica que el modelo de regresión se ajusta adecuadamente. En conjunto, estos resultados respaldan la eficacia del modelo y sugieren que es una herramienta confiable para realizar predicciones precisas basadas en las variables consideradas.

Conclusiones

Con los pasos de preprocesamiento de la data se logró obtener datos afines con los que se pudo trabajar con el modelo. En base a esto, se pudo desarrollar el planteamiento de manera correcta en cuanto a la extracción y preparación de los datos. También se analiza en qué casos se aplica la metodología de regresión lineal en este artículo:

Predicción de Temática de Artículos Científicos: Se utiliza la regresión lineal múltiple para predecir la temática de los artículos científicos. En este caso, las variables independientes incluyen características relacionadas con el procesamiento del lenguaje natural, como la presencia de palabras clave específicas, la frecuencia de ciertas palabras, etc. La variable dependiente es la temática del artículo, convirtiéndolo en un problema de regresión donde se busca predecir una variable categórica.

Procesamiento del Lenguaje Natural (PLN): La regresión lineal múltiple se utiliza en el contexto del procesamiento del lenguaje natural, lo que significa que se aplica a datos lingüísticos. Se trata de un caso donde las características lingüísticas se utilizan como variables independientes para predecir resultados relacionados con el procesamiento del lenguaje, como la temática de los artículos.

Análisis de Variables: El artículo describe cómo se seleccionan y utilizan las variables lingüísticas como predictores en el modelo de regresión lineal múltiple. Esto implica analizar y procesar las variables para que sean adecuadas para su uso en el modelo.

Evaluación del Modelo: El artículo también se enfoca en la evaluación del modelo de regresión lineal múltiple mediante medidas como el Error Cuadrático Medio (MSE) y el coeficiente de determinación (R^2). Estas medidas se utilizan para evaluar la calidad del ajuste del modelo.

La implementación del modelo de regresión lineal dio como resultado un rendimiento favorable al momento de la evaluación del modelo, cumpliendo con los parámetros establecidos en el artículo y proporcionando resultados que indican que el modelo implementado se aplicó de manera correcta y bien estructurada. Finalmente, se puede mencionar que se obtuvieron favorables resultados al aplicar el modelo de regresión lineal múltiple basándose en la aplicación del PLN.

Bibliografía

Baños Ruth, F. M. (2019). *Análisis de regresión lineal múltiple con SPSS: un ejemplo práctico*. *REIRE Revista d'Innovació i Recerca en Educació*, 12(2), 1-10.

Cárdenas-Pérez, A. E. (2021). *Explicación del crecimiento económico en la Economía Popular y Solidaria mediante la aplicación del modelo econométrico de Regresión Lineal y Múltiple*. *Revista Publicando*, 8(28), 74-84.

Jiménez, S. A. (2023). *Modelos de Aprendizaje Automático basados en CRISP-DM para el Análisis de los niveles de Depresión en los estudiantes de la Escuela Politécnica Nacional*. *ARTICLE HISTORY LATIN-AMERICAN JOURNAL OF COMPUTING (LAJC)*, 22-43.

Junco, E. E. (2021). *Diseño de un Modelo de Regresión Lineal Múltiple Para Predecir el Rendimiento de Estudiantes de Institutos Superiores Tecnológicos Públicos Frente a la Nueva Normalidad*. *European Scientific Journal ESJ*, 17.

Montero, R. (2016). *Modelos de regresión lineal múltiple* (Documento interno, Departamento de Economía Aplicada, Universidad de Granada).

Portugal, R. T. (2020). *MODELO DE REGRESIÓN LINEAL MÚLTIPLE DE LA GESTIÓN*

DEL CONOCIMIENTO, CON LA CULTURA ORGANIZACIONAL, EL LIDERAZGO Y LAS TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN, EN TRABAJADORES DE UNA EMPRESA DE LA CD. DE DURANGO, DURANGO, MÉXICO. *Hitos de Ciencias Económico Administrativas*, 266-284.

Roque López, J. (2021). *Técnicas de selección de variables en regresión lineal múltiple* (Tesis de maestría, Universidad Internacional de Andalucía). Andalucía, España.

Arellano, A. P. (2020). Modelos de regresión lineal para predecir el consumo de agua potable. *Revista Digital Novasinergia*, 3(1), 27-36.

Avalos Tapia, A. G. (2022). *Estimación de la cantidad de heridos en accidentes de tránsito dentro de la provincia de Lima utilizando modelos de regresión lineal múltiple y PCA*.

Morantes-Quintana, G. R.-P.-S. (2019). Modelo de regresión lineal múltiple para estimar concentración de PM 1. *Revista Internacional de Contaminación Ambiental*, 179-194.

Abuín, J. M., & Rojo, J. M. (2007). *Regresión con variable dependiente cualitativa*. Recuperado de http://humanidades.cchs.csic.es/cchs/web_UAE/tutoriales/PDF/Regresion_variable_dependiente_dicotomica_3.pdf.

Tejada, G. V.-T. (2021). Predicción del contenido de almidón en queso fresco adulterado utilizando Regresión de Mínimos Cuadrados Parciales, Regresión Lineal Múltiple e Imágenes Hiperespectrales. *Journal of Agro-industry Sciences*, 3(1), 15-20.

Branzuela, N. F., San Diego, A. L., & Namoco, S. O. (s/f). *A multiple regression analysis of the factors affecting academic performance of computer-aided designing students during flexible learning program*. Sci-int.com. Recuperado el 16 de marzo de 2023,

de http://www.sci-int.com/pdf/638057079331871453.7%20525-530%20Nicanor%20F.%20Branzuela%20Jr-SARA-MATH%20-28-11-22.pdf

Ouedraogo, I., Defourny, P., & Vanclooster, M. (2018). Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. *Hydrogeology Journal*.

Peña, S. (2017). *Análisis de datos*. Bogotá, Colombia: AREANDINA. Fundación Universitaria del Área Andina.