Artículo original/ Original article
ISSN 2953-6553

Clasificación binaria de artículos científicos basada en Regresión logística.

Binary classification of scientific articles based on logistic regression.

Esteban Pangol-Campoverde [1] 0009-0004-9149-7733, John Criollo-Ramón [1] 0009-0005-8359-3300, David Quezada-Tène [1] 0009-0006-4096-9264 ,Verónica Chimbo-Coronel [1] 0000-0003-3470-739X

esteban.pangol.est@tecazuay.edu.ec, john.criollo.est@tecazuay.edu.ec, david.quezada.est@tecazuay.edu.ec, veronica.chimbo@tecazuay.edu.ec

[1] Instituto Superior Universitario Tecnológico del Azuay, Cuenca, Ecuador

DOI 10.36500/atenas.2.007

Resumen

El objetivo de este proyecto es aplicar un modelo de clasificación binaria basado en el algoritmo de Regresión Logística para el análisis de los artículos científicos. Para comenzar, se recopiló información de diversos repositorios científicos, tales como el Instituto of Electrical and Electronics Engineers Inc, American Society of Civil Engineers (ASCE), entre otros. Se utilizó una hoja de cálculo para llevar a cabo dicha recopilación.

Para la ejecución del proyecto, se llevó a cabo un proceso de filtrado de datos de forma manual en primera instancia, con el fin de eliminar los archivos que no permitían acceder a los repositorios y otros problemas que pudieran afectar el desarrollo del proyecto.

Con base a la información recopilada, se realizó un proceso de etiquetado de los datos en diferentes columnas utilizando valores de 0 y 1. Este procedimiento permitió la creación de variables para llevar a cabo una clasificación binaria que se adecuara a los requerimientos del proyecto. Luego de esta etapa, se seleccionó Python como lenguaje de programación y se utilizó la herramienta Google Colab para trabajar en equipo de manera más eficiente.

Se inició con la aplicación de diversas técnicas de preprocesamiento para refinar la información y prepararla para su posterior uso. Una vez realizados estos cambios, se seleccionaron las variables que se utilizarían en el análisis y se establecieron todos los requisitos necesarios para obtener un análisis favorable.

Finalmente, al obtener los resultados se encontró que el modelo utilizado fue adecuado para trabajar con esta información, obteniéndose datos muy satisfactorios.

Abstract

The objective of this project is to apply a binary classification model based on the Logistic Regression algorithm for the analysis of scientific articles. To begin with, information was collected from various scientific repositories, such as the Institute of Electrical and Electronics Engineers Inc, American Society of Civil Engineers (ASCE), among others. A spreadsheet was used to carry out this compilation.

For the execution of the project, a data filtering process was carried out manually in the first instance, in order to eliminate files that did not allow access to the repositories and other problems that could affect the development of the project.

Based on the information collected, a process of labeling the data in different columns using values of 0 and 1 was carried out. This procedure allowed the creation of variables to carry out a binary classification that suited the requirements of the project. After this stage, Python was selected as the programming language and the Google Collab tool was used to work more efficiently as a team.

Various preprocessing techniques were applied to refine the information and prepare it for later use. Once these changes were made, the variables to be used in the analysis were selected and all the necessary requirements were established to obtain a favorable analysis.

Finally, when the results were obtained, it was found that the model used was adequate for working with this information, obtaining very satisfactory data.

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

Palabras Claves – CRISP-DM, KDD y SEMMA, Minería De Datos, Datos Nulos, Clasificación Binaria.

I. INTRODUCCIÓN

El presente artículo científico se centra en la aplicación de un modelo de clasificación binaria en datos reales, explorando su construcción a partir de un conjunto de datos y la implementación de técnicas para abordar la problemática planteada. La recuperación de información es un tema crucial en el ámbito del análisis de datos, dado el crecimiento constante de la cantidad de información disponible. La clasificación efectiva de datos se vuelve esencial para facilitar la búsqueda de información relevante.

Para este estudio, se dispuso de una base de datos que contenía documentos de diversos repositorios, todos dentro de la base Scopus, relacionados con la industria 4.0 en diferentes áreas. Los repositorios más utilizados incluyeron el Institute of Electrical and Electronics Engineers Inc, la American Society of Civil Engineers (ASCE) y Cambridge University Press, entre otros. A partir de la extracción de información, se realizó un proceso de clasificación de documentos que dio lugar a dos conjuntos de datos específicos: uno relacionado con Machine Learning y otro centrado en el Internet de las cosas (IoT). El conjunto de datos de Machine Learning fue empleado para desarrollar un modelo de clasificación binaria como parte de un enfoque de aprendizaje automático supervisado. Este modelo permite predecir a cuál de las dos clases posibles pertenecía una instancia de datos, lo que resultó en una estrategia eficaz para identificar variables significativas y mejorar la calidad de la evaluación.

La elección de la clasificación binaria se basó en consideraciones relacionadas con la penalización en el procesamiento al modificar algoritmos de cifrado o sus versiones, con el objetivo de mantener la simplicidad y efectividad en el contexto crítico de la seguridad de la información. Esta selección se realizó tras comparar diversas alternativas y determinar que este enfoque era el más óptimo para abordar lo planteado en el artículo científico.

II. MARCO TEÓRICO

2.1 Repositorios Científicos

Los repositorios científicos son plataformas en línea donde se almacenan y comparten investigaciones científicas. Estos repositorios permiten a los investigadores y académicos compartir sus trabajos con una audiencia más amplia, lo que aumenta la visibilidad y la accesibilidad de la investigación.

Los repositorios se clasifican en dos tipos diferentes:

Los repositorios abiertos: tienen la característica de que cualquiera puede leer, descargar, imprimir y distribuir los artículos publicados sin tener que pagar por el acceso. Como afirma Caldera J. (2013):

Las Instituciones deciden poner en acceso abierto su información científica, técnica, cultural y/o social, debido a que es una forma fácil, sencilla y barata de poner a disposición de todos contenidos. (pág. 11)

Los repositorios privados: limitan el acceso a contenido científico a usuarios específicos o grupos de usuarios que han obtenido suscripción o membresía por otro lado, el análisis de repositorios científicos es un proceso más importante para los investigadores que buscan identificar tendencias, tomar en cuenta nuevos hallazgos o recopilar información y documentación importante.

2.3 Metodologías Para El Proceso De Análisis De Datos

Análisis de datos consiste en definir el problema, recolección de datos, preparación, limpieza y transformación, técnicas de visualización, modelado y estadística descriptiva e inferencial, y presentar los resultados. Algunas de las metodologías más utilizadas son CRISP-DM, KDD y SEMMA.

2.4 Crisp-Dm

Cross-Industry Standard Process for Data Mining (CRISP-DM) trata sobre la metodología estándar para proyectos de minería de datos. Muestra de manera ordenada la forma de realizar un proyecto de minería y análisis de datos. Este orienta el modo en el cual será realizado y es ampliamente utilizado en el campo de la ciencia y análisis de datos.

2.5 Minería De Datos

La minería de datos es un proceso de análisis y exploración de grandes cantidades de datos para descubrir patrones, tendencias o relaciones, los cuales pueden ser utilizados para predecir resultados. Se utilizan técnicas y herramientas de inteligencia artificial, estadística y aprendizaje automático para extraer información valiosa de los datos.

2.6 Técnicas De Preparación De Los Datos

2.7 Limpieza De Datos

Dentro del análisis de datos el proceso fundamental para el desarrollo de Xining un análisis de datos es la verificación de valores para conocer si esto está expresado de manera correcta o que al menos se ajusten de manera adecuada con algún conjunto de datos para su posterior modelamiento de información (Oviedo Carrascal, E. A., Vélez Saldarriaga, G. L., (2017

2.8 Transformación De Variables

La transformación de variables es un proceso muy importante ya que ayuda a la mejor visualización de las variables que tenemos en nuestro conjunto de datos y así mismo ver cuáles de estas se pueden transformar y permitir un fortalecimiento del análisis planteado.

2.9 Datos Nulos

Los datos nulos o faltantes son comunes en el análisis de datos y deben ser evaluados para determinar si pueden ser utilizados o eliminados. Existen dos tipos de datos ausentes: aquellos que existen en el mundo real pero no están en nuestro conjunto de datos debido a algún error, y aquellos que no tienen significado para el objeto en cuestión y no existen en el mundo real. Los datos nulos no son lo mismo que cero o una cadena vacía, ya que implican ausencia de información.

2.10 Codificación

La normalización es un proceso importante por lo tanto para realizar una buena codificación que ayude a mejorar el desempeño del modelo, se debe entender qué clase de datos se está manejando para poder utilizar el codificador más adecuado a esta para poder mantener una consistencia de los datos. (Echeverri Giraldo, A. F. (2019)).

2.12 Clasificación Binaria

La clasificación supervisada es comúnmente realizada por Sistemas Inteligentes. En este contexto, diversos paradigmas estadísticos, como la Regresión Logística y el Análisis Discriminante, pueden llevar a cabo estas tareas.

Los modelos de clasificación binaria son esenciales para categorizar observaciones en dos categorías posibles, especialmente cuando se requiere predecir respuestas binarias. En este artículo, nos centramos en la Regresión Logística, que permite predecir una variable categórica binaria a partir de variables predictoras, ya sean continuas o categóricas.

Es fundamental destacar que, durante el desarrollo de este proyecto, se realizaron diversos enfoques utilizando diferentes algoritmos para abordar la problemática planteada. Sin embargo, tras rigurosas evaluaciones, se concluyó que la implementación de un modelo de clasificación binaria demostró ser la opción más óptima. Esta elección se basa en consideraciones como la penalización en el procesamiento al modificar algoritmos de cifrado o sus versiones, que podrían introducir una mayor complejidad en el sistema, especialmente en el contexto crítico de la seguridad de la información. Por lo tanto, la selección de la clasificación binaria se basó en su capacidad para mantener una efectividad y simplicidad sobresalientes en comparación con otros enfoques alternativos.

2.13 Evaluación Del Modelo y Matriz de Confusión

Métricas para evaluar el desempeño en la clasificación La matriz de confusión se considera el punto de partida para el cálculo de la medición del desempeño de un modelo predictivo; en este caso, presenta los resultados de clasificación del Modelo Computacional Social Mining con Naïve Bayes. Donde VP = verdadero positivo, FP = falso positivo, FN = falso negativo y VN = verdadero negativo. En una clasificación binaria, la precisión se calcula dividiendo el número de casos identificados correctamente entre el total de casos. Y con el análisis ROC (Receiver Operating Characteristic por sus siglas en inglés) mide el rendimiento del clasificador binario. La curva ROC se considera una herramienta útil para medir el desempeño de algoritmos clasificadores (Spackman 1989).

2.14 Tablas De Contingencia.

Este es un proceso clave en la identificación de las posibles causas de los problemas de salud, y también de factores que, aun cuando no puedan ser finalmente considerados causales, resulten asociados a estos daños y constituyan importantes elementos prácticos para la identificación de grupos con mayores riesgos de padecer determinado daño.

2.15 Curva De Roc

En términos estadísticos, teniendo un estimador de una variable que cuenta con un parámetro ajustable se pueden especificar sus curvas de *Sensibilidad* y *Especificidad* o Curvas ROC. En la Curva ROC se representa la sensibilidad del modelo frente al valor obtenido de restar la especificidad a la unidad (1-especificidad)

III. MÉTODO

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un marco de trabajo ampliamente utilizado para proyectos de minería de datos y aprendizaje automático.

1. Definición del objetivo: Entrenar un modelo de clasificación binaria para predecir si un ejemplo pertenece a una clase o categoría específica, basado en datos etiquetados previamente.

2. Recopilación de datos: Extracción de documentos relacionados con la Industria 4.0 en áreas de Machine Learning e IOT. Organización de la información en una hoja de cálculo.

3. Análisis del conjunto de datos: Focalización en datos de Machine Learning y exploración de los datos utilizando Google Colab. Visualización de las primeras filas, columnas y dimensiones del conjunto de datos.

4. Limpieza de los datos: Eliminación de columnas innecesarias, dejando solo las columnas numéricas para el análisis. Se conservan dos columnas de tipo string para un proceso posterior.

5. Tratamiento de datos nulos: Identificación de columnas con datos nulos y eliminación de estos valores utilizando el código dropna.

6. Transformación de variables: Utilizar la función "apply" y una función lambda para convertir valores "Si" a 1 y "No" a 0, y la función "lower" para convertir todos los valores a minúsculas.

7. Normalización de los datos: Escalar los datos para que todos estén en una sola escala, facilitando el

funcionamiento de los algoritmos de aprendizaje automático.

8. Clasificación binaria: Aplicar el modelo de Regresión Logística para clasificar elementos en dos grupos según una regla de clasificación.

9. Selección de columnas con mayor cantidad de datos: Identificar las 3 columnas con mayor cantidad de datos para trabajar con ellas en el análisis.

10. Definición de variables para el modelo: Dividir las variables seleccionadas en x e y, y proceder al entrenamiento de las variables.

11. Evaluación del modelo mediante matriz de confusión: Utilizar una tabla que muestre la cantidad de predicciones correctas e incorrectas del modelo para cada categoría.

12. Cálculo de Precision-Recall: Calcular la precisión, exhaustividad y puntaje del modelo para evaluar su efectividad.

13. Tabla de contingencia y demostración de resultados: Calcular la tabla de contingencia a partir de dos columnas de un Dataframe y comparar los resultados obtenidos con las librerías empleadas.

14. Gráfico de Roc: Evaluar la calidad del análisis de clasificación binaria mediante la curva de ROC y comprobar la fiabilidad del trabajo de clasificación binaria.

IV. Resultados y discusión

4.1 El objetivo de estudio de clasificación binaria

En el contexto de la ciencia de datos y el aprendizaje automático (machine learning) es entrenar un modelo que pueda predecir con precisión si un determinado ejemplo pertenece a una clase o categoría específica, basado en un conjunto de datos de entrenamiento previamente etiquetado.

4.2 Recopilar los Datos

- Colectar Datos y Preparar los datos

El recopilamiento de la información se dio mediante la extracción de documentos que están relacionados a la industria 4.0 en diferentes áreas, toda la información se colocó dentro de una hoja de cálculo en el cual se creó una base pequeña de información se decidió que la documentación se dividiría en dos temáticas marcadas las cuales son Machine Learning y IOT ("Internet of Things"/Internet de las Cosas).

4.3 Análisis del Conjunto de Datos

Se determinó que el proyecto estaría fundamentado en relación al conjunto de datos Machine Learning por lo cual se procederá al primer paso para poder trabajar en la modelación de los datos por que como se sabe, los sistemas computacionales no entienden el lenguaje humano sino el numérico por lo que generar estas variables son necesarias para el desarrollo del modelo de clasificación Binaria.

Se cargan los datos que se tienen de la siguiente manera: se guarda en una variable la **data** (nombre usando dentro de Google colab), la cual ser cargada en Google Colab para poder aplicar una exploración, y así apreciar qué datos son los que se tienen. En la exploración, se conoce como *el primer paso para el análisis de datos*. Para ello los códigos a utilizar son: **data. head ()** = Este código permite visualizar los 5 primeros datos del dataframe.

El código da a conocer exactamente qué columnas se tienen dentro del conjunto de datos y se coloca en la siguiente línea: **data. columns** = Este código permite obtener las columnas de la base de datos

El código sirve para saber qué cantidad de datos y se le visualiza en la siguiente línea.

data. shape = Este código muestra las dimensiones totales del conjunto de datos en filas y columnas

Data Shape = (389,55).

El código para saber qué tipo de datos se está trabajando, se colocará en la siguiente línea de código, **data.info ()** = Con el código, se conocerá los tipos de datos que existen en la base de datos.

4.4 Limpieza de los Datos

Se elimina las columnas innecesarias que son más columnas tipo texto para tener que dejar las columnas numéricas para el análisis y se utilizó una función la cual permite conocer todas las variables disponibles en la base de datos, luego de un análisis respectivo se eliminaron las columnas de tipo string a excepción de dos columnas las cuales se usará para un proceso posterior en la preparación de datos.

4.5 Tratamiento de Datos nulos

El modelamiento se tratará del preprocesamiento de los datos por lo que aquí se da el procesamiento de datos nulos con el código el cual permite ver cuál de las columnas del dataframe tiene datos nulos en el que se visualiza que existen 3 columnas con datos nulos estas son: *Descarga indirecta por nSCI-Hub (SI/NO) con 94 datos nulos, seguido por LMS (Learning Management System), Entornos virtuales de aprendizaje con 67 datos nulos y por último Avances científicos con 1 datos nulo, luego de conocer estas variables se procedió con su eliminación por medio del código dropna el cual nos permite eliminar valores nulos del dataframe.*

4.6 Transformación de variables

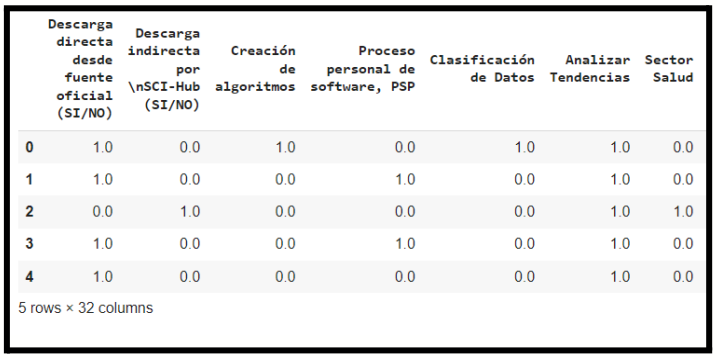
En el proceso de transformación de variables se utiliza la función "apply" la cual se utiliza para aplicar una función a cada elemento de la columna. En este caso, la función es la función lambda que convierte el valor "Si" a 1 y "No" a 0. También se usa la función "lower" para convertir todos los valores a minúsculas y evitar errores debido a mayúsculas o minúsculas.

4.7 Normalización de los Datos

La normalización es un proceso importante porque muchos algoritmos de aprendizaje automático no funcionan bien con datos que tienen diferentes escalas. En este proceso los datos están en una sola escala y se guardan en una nueva variable con la cual ya se ha trabajado con el modelo de clasificación Binaria.

Figura 1

Datos normalizados en el proceso



	Descarga directa desde fuente oficial (SI/NO)	Descarga indirecta por \nSCI-Hub (SI/NO)	Creación de algoritmos	Proceso personal de software, PSP	Clasificación de Datos	Analizar Tendencias	Sector Salud
0	1.0	0.0	1.0	0.0	1.0	1.0	0.0
1	1.0	0.0	0.0	1.0	0.0	1.0	0.0
2	0.0	1.0	0.0	0.0	0.0	1.0	1.0
3	1.0	0.0	0.0	1.0	0.0	1.0	0.0
4	1.0	0.0	0.0	0.0	0.0	1.0	0.0

5 rows × 32 columns

Nota: García, J., Molina, J., Berlanga, A., Patricio, M., Bustamante, A., & Padilla, W. (2018). Ciencia de datos. Técnicas Analíticas y Aprendizaje Estadístico. Bogotá, Colombia. Publicaciones Altaria, SL.

4.8 Clasificación Binaria

Es la tarea de clasificar los elementos de un conjunto en dos grupos sobre la base de una regla de clasificación para lo cual se aplica el modelo de Regresión Logística para clasificación binaria la cual es un modelo que permitirá predecir una variable categórica binaria a partir de una o más variables que se conocen a continuación luego de realizar su selección en base de la mayor cantidad de valores 1 dentro del mismo.

Consideremos que el caso estudiado para este artículo involucró la aplicación de un algoritmo de clasificación binaria basado en Regresión Logística a un conjunto de artículos científicos. Era necesario dividir estos artículos en dos grupos basándose en normas predeterminadas. Esta tarea se abordó utilizando el modelo de Regresión Logística de clasificación binaria. Utilizando un conjunto de variables

predictivas cuidadosamente elegidas, este modelo se utilizó para predecir una variable categórica binaria basada en aquellas que tenían el porcentaje más alto de valores 1 en el conjunto de datos.

Este método funcionó bien para clasificar artículos científicos correctamente según criterios predeterminados, lo que demuestra el valor y la adaptabilidad de la clasificación binaria en contextos científicos. Sin embargo, es fundamental tener en cuenta la importancia de tener en cuenta los posibles costos de procesamiento suplementarios, particularmente al modificar los algoritmos de cifrado o sus versiones, ya que esto podría agregar más complejidad al proceso de clasificación por lo que la obtención de resultado podría verse afectada de forma significativa.

4.9 Columnas con mayor cantidad de Datos

En del proceso de clasificación binaria se realiza un total de los datos de las columnas, en el cual existen ciertas columnas con las que hay que trabajar con ello y realizar un análisis de los datos con otra fórmula la cual se indicó las 3 columnas con mayor cantidad de datos las cuales son Clasificación de Datos, Industria 4.0 y por último Analizar Tendencias

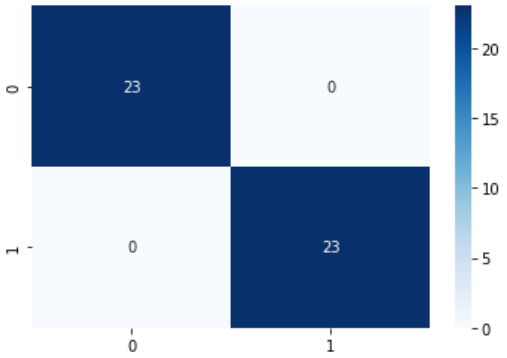
4.10 Variables que se usarán para el modelo

Las variables seleccionadas para el modelo se basaron en su histórico de resultados, priorizando aquellas que mostraron una mayor cantidad de casos con un resultado de 1, lo cual se considera óptimo en nuestro contexto. Estas variables se dividen en dos conjuntos: uno denotado como "x" que incluye características relacionadas con "Industria 4.0" y "Análisis de Tendencias", y otro denotado como "y" que corresponde a la variable de "Clasificación de Datos.1". Este proceso incluyó el entrenamiento de estas variables, preparándose adecuadamente para su inclusión en el modelo. Además, se aplicaron técnicas de tratamiento de datos con el objetivo de mejorar la calidad de las variables y garantizar resultados más confiables en el modelo de clasificación binaria.

4.11 Matriz de Confusión

Figura 2

Matriz de Confusión



Nota: Muñoz, J. M. S. (2016). Análisis de Calidad Cartográfica mediante el estudio de la Matriz de Confusión. *Pensamiento matemático*, *6*(2), 9-26.

La matriz de confusión es una herramienta esencial para la evaluación del desempeño de un modelo de clasificación en un contexto científico. Esta matriz proporciona una representación tabular de las predicciones del modelo, mostrando tanto las clasificaciones correctas como las incorrectas para cada categoría o clase considerada, La grafica muestra 2 de color azul y dos de color blanco los de color azul indican los Falsos Positivos y los blancos indican los Falsos Negativo.

4.12 Precision - Recall

La precisión se define como la fracción de instancias relevantes correctamente identificadas dividida entre todas las instancias obtenidas. La recuperación, por otro lado, representa la fracción de instancias relevantes correctamente identificadas con respecto al total de instancias relevantes presentes en el conjunto de datos. En este trabajo, se evalúan y presentan tres métricas clave para la evaluación del modelo: el puntaje de precisión (accuracy score), el puntaje de precisión (precision score) y el puntaje de recuperación (recall score), que se detallan a continuación:

- accuracy_score: 1.0
- precision_score: 1.0
- recall_score: 1.0

Los resultados obtenidos confirman la eficacia de nuestro algoritmo de clasificación binaria en el contexto de nuestro proyecto. Estos puntajes indican que el modelo es capaz de lograr una precisión del

100% en la identificación de instancias relevantes, lo que respalda su idoneidad para el análisis de las variables consideradas.

4.13 Tabla de contingencia y demostración de resultados con la tabla.

Se calcula la tabla de contingencia a partir de dos columnas de un Dataframe. La tabla de contingencia es similar a la matriz de confusión, y se utiliza para evaluar el rendimiento de un modelo de clasificación dentro del mismo se visualiza 24 ceros y 22 unos no teniendo mucha diferencia por lo cual el modelo podrá tener mayor eficacia.

Las fórmulas empleadas junto con la tabla de contingencia permiten conocer de manera manual los resultados obtenidos anteriormente con las librerías mencionadas es como un método de comprobación que usa los falsos positivos y verdaderos negativos como se observa a continuación con la representación de los resultados obtenidos.

Figura 3

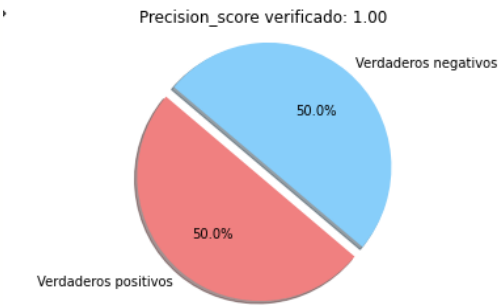


Figura 3. *Gráfica que representa verdaderos Positivos y Verdaderos negativos*

4.14 Gráfico de Roc

Facilita la evaluación de la calidad del análisis realizado mediante el enfoque de clasificación binaria, permitiendo así verificar la confiabilidad de los resultados obtenidos en dicho proceso de clasificación binaria.

Figura 4

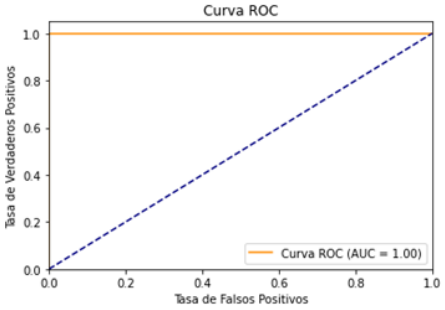


Figura 4. *Gráfico de la curva de ROC 1.0 significa que es perfecto en la discriminación*

V. CONCLUSIONES

La implementación de un modelo de clasificación binaria sobre un conjunto de datos es un proceso importante en el análisis de datos, lo que nosotros logramos demostrar fue que cualquier tarea en la que se necesite dividir los datos en dos categorías distintas se puede abordar con un algoritmo de clasificación binaria.

En resumen, este estudio subraya la importancia crítica de la calidad del conjunto de datos en la construcción de modelos de análisis y clasificación. Se ha demostrado que un conjunto de datos de alta calidad es fundamental para alcanzar niveles óptimos de precisión en el modelo. En nuestro caso, hemos logrado superar los desafíos potenciales relacionados con valores faltantes y datos inconsistentes durante el análisis y clasificación, lo que subraya la eficacia de nuestro enfoque en la gestión de datos y la construcción del modelo. Estos hallazgos refuerzan la importancia de la integridad y calidad de los datos como punto de partida esencial en la investigación y desarrollo de modelos analíticos.

Como se describió el proceso de implementación del modelo de clasificación binaria ha demostrado ser un desafío complejo. Sin embargo, con una preparación minuciosa del conjunto de datos, una selección cuidadosa de los hiper parámetros y una validación rigurosa del modelo, hemos obtenido resultados altamente satisfactorios. Esto nos permite concluir que la elección de implementar este modelo fue apropiada y altamente beneficiosa, ya que nos proporcionó datos relevantes y precisos en nuestro contexto de estudio. Estos resultados respaldan la eficacia y utilidad de nuestro enfoque y subrayan la

importancia de una sólida preparación y validación en la implementación de modelos de clasificación binaria en problemas similares.

En conclusión, una vez que hemos construido el modelo, evaluamos su rendimiento en un conjunto de datos de prueba independiente. Los resultados muestran que la precisión de los datos es óptima para el trabajo y cumple con los requisitos necesarios para el desarrollo de la clasificación en nuestro conjunto de datos. Para lograr un buen rendimiento en la implementación en diversos campos, es esencial contar con datos de alta calidad y seleccionar minuciosamente las características y algoritmos de clasificación adecuados, incluyendo el filtrado y la búsqueda interna de datos que como pudimos constatar un modelo de clasificación binaria puede ser una herramienta poderosa para tomar decisiones informadas en textos de este tipo.

VI. REFERENCIAS BIBLIOGRÁFICAS

Cristina, L., Puerta, C., Jhoana, Lady, & Zuluaga, R. (2022). Modelos de aprendizaje supervisado para la clasificación de riesgo crediticio en la entidad financiera Home Credit. https://bibliotecadigital.udea.edu.co/handle/10495/29124

Echeverri Giraldo, A. F. (2019). *Modelo predictivo de Churn de clientes para el negocio de Telecomunicaciones. Universidad de Antioquia.* https://bibliotecadigital.udea.edu.co/bitstream/10495/15142/1/EcheverriAndres_2019_ModeloPredictivoChurn.pdf

Oviedo Carrascal, E. A., Oviedo Carrascal, A. I., Velez Saldarriaga, G. L., & Universidad Pontificia Bolivariana. (2017). *Minería multimedia: hacia la construcción de una metodología y una herramienta de analítica de datos no estructurados.* Revista Ingenierías Universidad de Medellín, 16(31), 125-142. https://doi.org/10.22395/rium.v16n31a6

Alania Ricaldi, P. F. (2019). *Aplicación de técnicas de minería de datos para predecir la deserción estudiantil de la facultad de ingeniería de la Universidad Nacional Daniel Alcides Carrión*. Universidad Nacional Daniel Alcides Carrión. http://45.177.23.200/bitstream/undac/829/1/T026_40573846_M.pdf

Vallejo Ballesteros, H. F., Guevara Iñiguez, E., & Medina Velasco, S. R. (2018). *Minería de Datos. RECIMUNDO*: Revista Científica de la Investigación y el Conocimiento, 2(1), 339-349. https://dialnet.unirioja.es/servlet/articulo?codigo=6732870&info=resumen&idioma=SPA

Cruz Hurtado, E. J.., y Romero, M. F. (2020). Propuesta de un modelo logístico para la probabilidad de instalación de datáfonos en una empresa ubicada en Bogotá: Proposal for a logistics model for the probability of installing dataphones in a company located in Bogotá. *Noria Investigación Educativa, 1*(5), 41–53. https://doi.org/10.14483/25905791.16452

Kelleher, J. D., & Tierney, B. (2018). *Data science: An introduction.* Chapman and Hall/CRC. https://www.sciencedirect.com/topics/computer-science/binary-classification

López Carvajal, J., y Willian Branch Bedoya, J. (2005). *Comparación De Modelos De Clasificación Automática De Patrones Texturales De Minerales Presentes En Los Carbones Colombianos Comparison Of Models Of Automatic Classification Of Textural Patterns Of Mineral Presents In Colombian Coals*. *Año*, *72*, 115–124. http://www.scielo.org.co/pdf/dyna/v72n146/a09v72n146.pdf

Melillanca, E. (2018). *Evaluación de modelos de clasificación: matriz de confusión y curva ROC. Revista Ingeniería de Sistemas*, 32(2), 139-144. http://www.ericmelillanca.cl/content/evaluaci-n-modelos-clasificaci-n-matriz-confusi-n-y-curva-roc

Servicio Gallego de Salud. (2017). *Ayuda: tablas de contingencia* [Documento de apoyo]. Recuperado de https://www.sergas.es/Saude-publica/Documents/1933/7Ayuda%20Tablas%20de%20contingencia.pdf

Repositorio. (s. f.). En Ecured.cu. Recuperado el 16 de marzo de 2023, de https://www.ecured.cu/index.php/Repositorio

Morales Vargas, A., & Codina, L. (2019). *Atributos de calidad web para repositorios de datos de investigación en universidades*. HIPERTEXT.NET, 14. https://doi.org/10.31009/hipertext.net.2019.i19.04

Hotz, N. (2023). *What is CRISP DM?* Data Science Process Alliance. Recuperado de

https://www.datascience-pm.com/crisp-dm-2/

Caldera Serrano, J. (2018). *Repositorios públicos frente a la mercantilización de la Ciencia: apostando por la ciencia abierta y la evaluación cualitativa.* Métodos de información, 28. https://doi.org/10.5557/IIMEI9-N17-074101

Riquelme Santos, J.C., Ruíz, R. y Gilbert, K. (2006). *Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial.* Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18. Recuperado el 11 de marzo de 2023, de https://idus.us.es/bitstream/handle/11441/43290/Minería%20de%20datos%20conceptos%20y%20tendencias.pdf?sequence=1&isAllowed=y