Informe Trabajo Práctico III

Anna Posada

Noviembre 22, 2024

Universidad Escuela de Ingeniería de Antioquia

Información Preliminar

El dataset seleccionado para este análisis consta de varias características que representan un conjunto de observaciones en un problema de clasificación. Estas características incluyen tanto variables numéricas como categóricas, lo que lo convierte en un caso ideal para aplicar técnicas de preprocesamiento como la codificación de etiquetas y el manejo de valores faltantes. Además, la variable objetivo es categórica, lo que la hace adecuada para el uso de algoritmos supervisados.

Modelos Seleccionados

- Random Forest
  Es un algoritmo basado en el ensamblaje de múltiples árboles de decisión. Este método utiliza el bagging (bootstrap aggregating) para entrenar cada árbol en diferentes subconjuntos del dataset, combinando sus predicciones para obtener un resultado final robusto. Random Forest es conocido por ser resistente al sobreajuste y por su capacidad para manejar tanto datos numéricos como categóricos (Alaminos-Fernández, 2023).

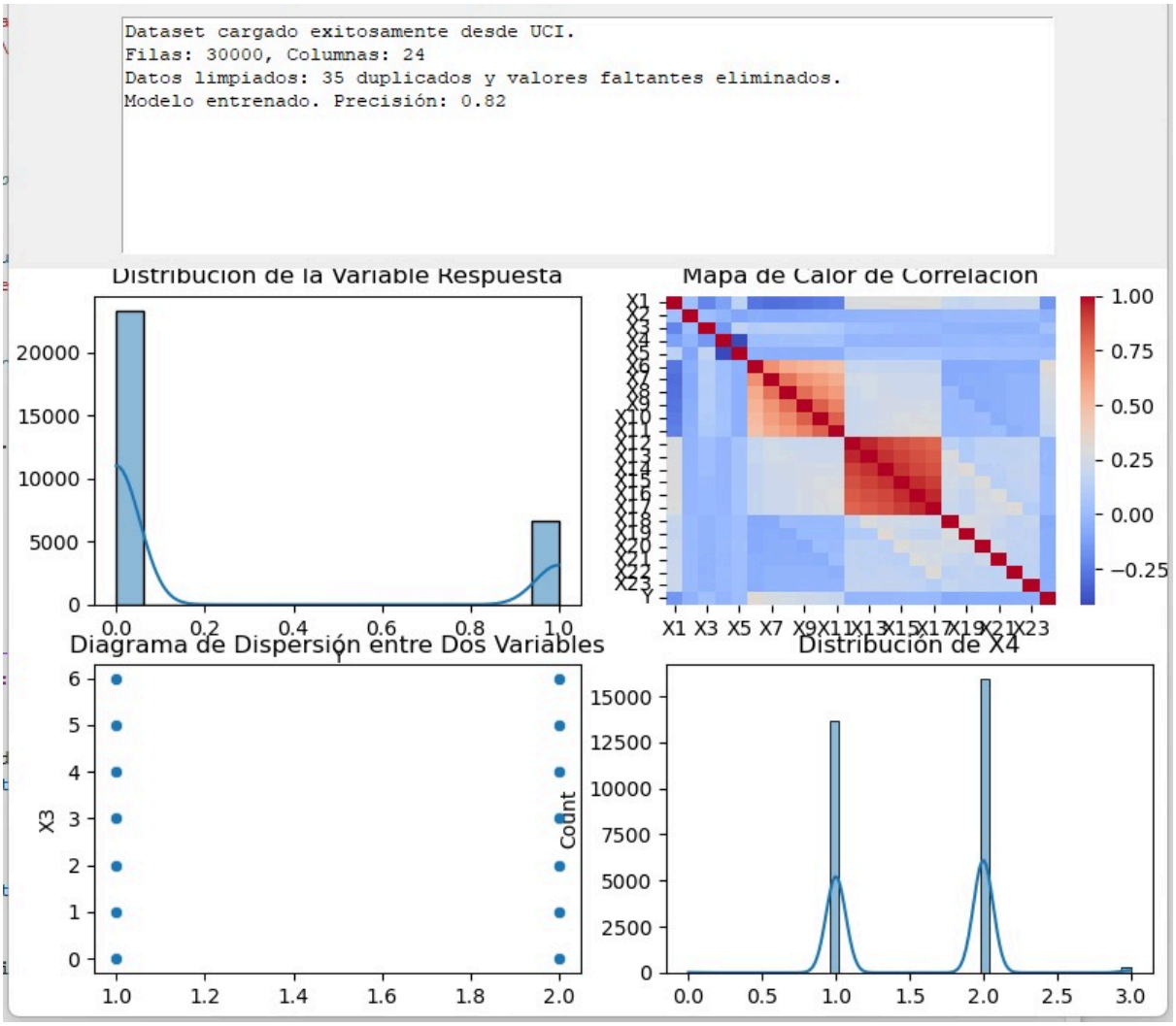
  En este proyecto, Random Forest se seleccionó por su excelente desempeño en tareas de clasificación y por ser adecuado para identificar relaciones no lineales entre las características.

- Logistic Regression
  La regresión logística es un modelo lineal que predice la probabilidad de que una observación pertenezca a una clase en particular. Aunque su desempeño puede ser limitado cuando las relaciones entre las variables no son lineales, es eficiente y rápido en términos computacionales (Amat Rodrigo, 2020).

  Este modelo fue elegido como un punto de comparación para evaluar cómo los modelos más complejos, como Random Forest, mejoran sobre un enfoque básico.

- Naive Bayes
  Este clasificador, basado en el teorema de Bayes, utiliza probabilidades condicionales para clasificar datos, asumiendo independencia entre las características. Su objetivo es calcular la probabilidad de que un conjunto de dichas características pertenezca a una etiqueta específica, asumiendo independencia entre las variables. Esta simplicidad en sus cálculos es la razón por la que se le llama "Naive" (Predictiva, 2021).

Resultados



Para la comparación de los modelos se utilizaría un código simple, el cual compara los valores de precisión de cada uno.

Hubo algunas dificultades técnicas debido a la falta de ciertas dependencias en mi computadora, lo que impidió que las gráficas se pudieran generar y mostrar correctamente. Además, me encontraba alejado de la ciudad, lo que complicó aún más el acceso a recursos y la posibilidad de recibir ayuda inmediata. Esto también dificultó el uso de otro computador con las herramientas necesarias ya descargadas. A pesar de estas limitaciones, se investigó y analizó todo lo que fue posible con los recursos disponibles en ese momento. Durante el proceso, se descubrió que el entorno de trabajo en el que realizamos el análisis, específicamente en el trabajo 2, era diferente al que estaba utilizando, lo que facilitó la ejecución del código y la visualización de los resultados sin esos inconvenientes técnicos. Debido a la falta de herramientas y configuraciones adecuadas en mi computadora, no fue posible completar el análisis como se había planeado inicialmente, ya que no se pudieron generar algunas partes clave del proyecto. Sin embargo, a pesar de estos contratiempos, se aprovechó el tiempo disponible para investigar y realizar el análisis que sí pudo completarse, con la intención de continuar avanzando una vez que se resuelvan los problemas de configuración y acceso a los recursos necesarios.

De igual manera, en el código se evidencia que, idealmente, se debieron haber realizado procesos como la limpieza de datos, que incluyó la eliminación de valores duplicados y la imputación de valores faltantes. Además, las variables categóricas se habrían codificado usando técnicas como el Label Encoding, lo que habría permitido a los algoritmos trabajar correctamente con los datos.

Bibliografía

Data Base:
https://archive.ics.uci.edu/datasets?search=Default%20of%20Credit%20Card%20Clients

Alaminos-Fernández, Antonio Fº (2023) Árboles de decisión con random forest. Universidad de Alicante. Obets Ciencia Abierta. Alicante: Limencop.https://rua.ua.es/dspace/bitstream/10045/133067/1/Random_Forest_en_la_Investigacion_Social.pdf

Amat Rodrigo, J. (2020, noviembre). Regresión logística en Python. Ciencia de Datos. https://cienciadedatos.net/documentos/py17-regresion-logistica-python.html

Predictiva21. (2021, agosto 18). Modelo Naive Bayes. Predictiva21. https://predictiva21.com/modelo-naive-bayes/