Proyecto del segundo parcial

Josué Déley[1] *, Martin Soto[1] *
[1]Escuela de Ingeniería Mecatrónica, Universidad Internacional del Ecuador, Quito, Ecuador
*Corresponding authors, e-mail: pedeleylo@uide.edu.ec, masotomo@uide.edu.ec

***Abstract*—**
The project focuses on the optimization of machine learning algorithms by tuning hyperparameters to obtain the best algorithm with the best metrics in the classification of the dataset. To do this, a case study of a financial institution is used to detect the factors that influence whether a debt will be paid or not. Therefore, data analysis and machine learning techniques are used to process customer data such as personal information, credit history, spending and payment patterns, among others.

***Keywords*—Phyton, accuracy, metrics, datasets, algorithms.**

***Resumen*—** El proyecto se centra en la optimización de algoritmos de aprendizaje automático mediante la sintonización de hiperparámetros para lograr obtener el mejor algoritmo con las mejores métricas en la clasificación del dataset. Para ello se utiliza un caso de estudio de una entidad financiera para detectar los factores que influyen para determinar si una deuda será pagada o no. Por lo que se utiliza técnicas de análisis de datos y aprendizaje automático para procesar los datos de los clientes tales como información personal, historial crediticio, patrones de gasto y pagos, entre otros.

***Palabras Clave*—Phyton, exactitud, métricas, base de datos, algoritmos.**

I. Introducción

El objetivo principal del proyecto es realizar el mejor alritmo para la clasificación de datos del dataset de una entidad financiera. Para ello se utilizan los métodos aprendidos en clase tales como bagging y boosting y se analizan las métricas de: f1 score, exactitud, precisión, sensibilidad y especificidad.

En el ámbito del aprendizaje automático, el uso de técnicas de ensamble se ha convertido en un recurso fundamental para mejorar el rendimiento predictivo de los modelos. Se aplicará una variedad de métricas para la evaluación y comparación de su desempeño.

El objetivo principal de este trabajo es investigar y comparar el impacto de las técnicas de ensamble en la precisión y estabilidad de los modelos de aprendizaje automático.

El uso de estas métricas se justifica por su capacidad para evaluar diferentes aspectos del rendimiento de los modelos, desde su capacidad para clasificar correctamente las instancias positivas y negativas, hasta su habilidad para generalizar correctamente en datos no vistos.

II. Metodología

El caso de estudio a trabajar consta en que el banco Omega ha observado un aumento en el número de clientes que entran en mora en sus pagos de tarjeta de crédito, lo que ha llevado a una disminución de sus ganancias. La entidad financiera desea entender las razones detrás de la mora de sus clientes y detectar los factores que influyen para determinar si una deuda será pagada o no.

Para lograr este objetivo, el banco decide utilizar técnicas de análisis de datos y aprendizaje automático para procesar una gran cantidad de datos de clientes, incluyendo información personal, historial crediticio, patrones de gasto y pagos, entre otros.

A partir de este análisis, el banco busca identificar los factores que podrían estar influyendo en la mora de sus clientes y construir un modelo de clasificación que permita determinar si un cliente en particular pagará su deuda a tiempo o no.

El banco tiene acceso a información personal de los usuarios, recopilada a lo largo de su relación comercial, y la información e historial de pagos de los últimos 6 meses (octubre 2023 – marzo 2024).

A. *Bagging*

Es el método de aprendizaje por conjuntos que se suele utilizar para reducir la varianza dentro de un conjunto de datos ruidoso. En el bagging, se selecciona una muestra aleatoria de datos en un conjunto de entrenamiento con reemplazo, lo que significa que los puntos de datos individuales se pueden elegir más de una vez [1].

Sus principales ventajas son:

- **Facilidad de implementación:** las bibliotecas de Python como scikit-learn (también conocidas como sklearn) permiten combinar fácilmente las predicciones de los aprendices básicos o los estimadores para mejorar el rendimiento del modelo [1].
- **Reducción de varianza:** el bagging puede reducir la varianza de un algoritmo de aprendizaje. Esto resulta especialmente útil con datos de alta dimensión, donde los valores perdidos pueden conducir a una mayor varianza, lo que los hace más propensos a sobreajustes y evita la generalización precisa a nuevos conjuntos de datos [1].

B. Boosting

Es un método de aprendizaje por conjuntos que combina una serie de aprendices débiles en un aprendiz fuerte para minimizar los errores de entrenamiento. En el boosting, se selecciona una muestra aleatoria de datos, que se ajusta a un modelo y luego se entrena de forma secuencial, es decir, cada modelo intenta compensar las debilidades de su predecesor. Con cada iteración, las reglas débiles de cada clasificador individual se combinan para formar una única regla de predicción fuerte [2].

C. F1 Score

Es un estimador de la capacidad de clasificación de una prueba que se usa con frecuencia en la ciencia de datos y en los algoritmos de inteligencia artificial y que puede ser de utilidad para la valoración de las pruebas diagnósticas. Es la media armónica de sensibilidad y valor predictivo positivo, por lo que pondera el valor de ambos en un solo estimador [3].

D. Matriz de confusión

Puede definirse como una tabla que describe el desempeño de un modelo de clasificación en un conjunto de datos de prueba cuyos valores verdaderos son conocidos. Una matriz de confusión es altamente interpretativa y puede ser usada para estimar un número de otras métricas [4].

E. Exactitud

Es la relación entre las predicciones correctas y el número total de predicciones. O más simplemente, con qué frecuencia es correcto el clasificador [4].

F. Precisión

Es la relación entre las predicciones correctas y el número total de predicciones correctas previstas. Esto mide la precisión del clasificador a la hora de predecir casos positivos [4].

G. Sensibilidad

Es la relación entre las predicciones positivas correctas y el número total de predicciones positivas. O más simplemente, cuán sensible es el clasificador para detectar instancias positivas. Esto también se conoce como la tasa verdadera positiva [4].

III. Resultados

Tras revisar el dataset otorgado, se identifico varios datos con probelemas los cuales tenían información erronea o incluso faltante. Para resolver este aspecto, se realizó un análisis exhaustivo de cada uno de las categorias de los datos, sus rangos, tipo de dato y cantidad total en cada una de las columnas del dataset. De esta forma se logró imputar los datos correctamente, remplazando los datos faltantes por datos repetidos o datos aleatorios dependiendo las carcaterísticas y de igual forma se cambiós los datos categóricos de tipo .ºbjeto.ª datos numéricos.

IV. Conclusiones

En este trabajo, hemos explorado el impacto de las medidas de presición, específicamente bagging y boosting, en el contexto del aprendizaje automático. Nuestro objetivo principal fue evaluar cómo estas técnicas pueden mejorar la precisión y estabilidad de los modelos en un entorno dado previamente en el datset.

A lo largo de nuestro estudio, hemos observado que tanto el bagging como el boosting tienen el potencial de mejorar significativamente el rendimiento predictivo de los modelos de aprendizaje automático. Estas técnicas permiten reducir la varianza y el sesgo del modelo, por lo que tiene una mejor capacidad de generalización y una mayor robustez frente a datos desbalanceados.

Al evaluar el desempeño de nuestros modelos utilizando métricas como el F1 Score, la exactitud, la precisión, la sensibilidad y la especificidad, hemos podido obtener una visión completa de su capacidad para clasificar correctamente las instancias positivas y negativas, así como su habilidad para generalizar en datos no vistos.

La elección entre bagging y boosting puede depender del conjunto de datos y del objetivo específico del problema. Mientras que el bagging tiende a funcionar bien en modelos con alta varianza, el boosting puede ser más efectivo en modelos con alto sesgo. Sin embargo, ambas técnicas ofrecen mejoras significativas en términos de rendimiento predictivo en comparación con modelos individuales.

En conclusión, este trabajo destaca la proporción de información valiosa para investigadores y practicantes interesados en mejorar la precisión y estabilidad de sus modelos en entornos similares.

Referencias

[1] "¿Qué es bagging? | IBM," May 2023. [Online]. Available: https://www.ibm.com/es-es/topics/bagging

[2] "¿Qué es boosting? | IBM," Apr. 2023. [Online]. Available: https://www.ibm.com/es-es/topics/boosting

[3] M. Molina, "F1-score - Ciencia," Nov. 2023, section: Epidemiología. [Online]. Available: https://www.cienciasinseso.com/f1$_score$/

[4] L. Gonzalez, "Métricas de Evaluación Clasificación con Scikit Learn," Jun. 2019. [Online]. Available: https://aprendeia.com/metricas-de-evaluacion-clasificacion-con-scikit-learn-machine learning/