Aplicación web para entrenar y predecir datos financieros, basado en algoritmos supervisados de clasificación y regresión (PredictLab)

Jhonatan Stick Gómez Vahos
Sebastián Saldarriaga Arias

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Javier Fernando Botía Valderrama, Doctor en Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(Gómez Vahos & Saldarriaga Arias, 2024)
Referencia **Estilo APA 7 (2020)**	Gómez Vahos, J. S. & Saldarriaga Arias, S. (2024). Aplicación web para entrenar y predecir datos financieros, basado en algoritmos supervisados de clasificación y regresión (PredictLab). Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Danny Alexandro Munera Ramírez

Dedicatoria

Dedicamos este trabajo a nuestras familias, por su amor incondicional, paciencia y apoyo constante a lo largo de este camino de aprendizaje que decidimos tomar.

A nuestros compañeros, que se convierten en grandes amigos, quienes con su ánimo y confianza nos han impulsado a superar cada desafío.

A los profesores que, con su conocimiento y guía, nos han inspirado a alcanzar nuestras metas.

Agradecimientos

En primer lugar, queremos expresar nuestro más sincero agradecimiento a Dios, por darnos la fortaleza, la perseverancia y la sabiduría necesarias para culminar este proyecto.

A nuestro asesor, el Dr. Javier Fernando Botía Valderrama, por su invaluable guía, paciencia y compromiso durante el desarrollo de esta monografía. Su orientación y conocimiento fueron clave para el éxito de este proyecto.

A la Universidad de Antioquia, especialmente a la Facultad de Ingeniería y al programa de Especialización en Analítica y Ciencia de Datos, por brindarnos un entorno académico estimulante, así como los recursos necesarios para llevar a cabo este proyecto.

A nuestros compañeros de la cohorte VII, por su camaradería, apoyo mutuo y valiosas aportaciones, que enriquecieron nuestro aprendizaje y fortalecieron nuestra experiencia durante esta especialización.

Finalmente, a todas las personas e instituciones que, de manera directa o indirecta, contribuyeron con su tiempo, conocimientos y entusiasmo a la culminación de esta etapa académica. A todos, les expresamos nuestro más profundo agradecimiento.

Tabla de contenido

Contenido

Lista de tablas

Lista de figuras

Resumen

El presente trabajo, también llamado PredictLab, desarrolla un programa interactivo basado en una interfaz web que permite a los usuarios realizar análisis predictivo supervisado de manera eficiente. Facilita la carga del conjunto de datos, el preprocesamiento personalizado y la selección del modelo de predicción que mejor se ajusta a los datos en función de métricas específicas.

El enfoque se centra exclusivamente en modelos supervisados de clasificación binaria y regresión. Para la clasificación, se consideran algoritmos como regresión logística, random forest, máquinas de soporte vectorial (SVM), KNN y clasificadores de Bayes ingenuo como GaussianNB y BernoulliNB. En el caso de regresión, se evalúan regresión lineal, ridge, random forest, AdaBoost y gradient boosting. Los parámetros de estos modelos son personalizables y se optimizan mediante un proceso de búsqueda de rejillas (grid search) (Scikit-learn Developers, 2023). La selección del modelo óptimo se basa en el puntaje F1 para tareas de clasificación y el error absoluto medio (MAE) para regresión.

Algunos de los resultados más representativos se obtuvieron al entrenar los conjuntos de datos, "Bank Marketing" obtenido desde UCI Machine Learning Repository (2014) para clasificación y "Credit Card Limit Prediction" disponible en Kaggle (s.f.) para regresión. En el caso de clasificación se obtuvo el mejor puntaje F1 de 0.92 para máquinas de soporte vectorial (SVM).En el caso de la regresión, se obtuvo el mejor valor en el error absoluto medio (MAE) de -0.006 para regresión logística y ridge, -0.02. Estos resultados fueron alcanzados sin configuraciones avanzadas ni depuración previa de los datos, teniendo en cuenta que dichos resultados se obtuvieron cargando los datos tal y como están en los repositorios y con la configuración por defecto de PredictLab.

Palabras clave: Clasificación, regresión, búsqueda de rejillas, puntaje F1, error absoluto medio (MAE).

El enlace del repositorio del proyecto es: https://github.com/jhonatanvahos/AutoML

Abstract

This project, named PredictLab, develops an interactive program with a web-based interface that enables users to perform supervised predictive analysis efficiently. It facilitates dataset uploading, custom preprocessing, and selecting the prediction model that best fits the data based on specific metrics.

The approach focuses exclusively on supervised binary classification and regression models. For classification, algorithms such as logistic regression, random forest, support vector machines (SVM), KNN, and naive Bayes classifiers such as GaussianNB and BernoulliNB are considered. In the case of regression, linear regression, ridge, random forest, AdaBoost, and gradient boosting are evaluated. The parameters of these models are customizable and are optimized using a grid search process (Scikit-learn Developers, 2023). The selection of the optimal model is based on the F1 score for classification tasks and the mean absolute error (MAE) for regression.

Some of the most representative results were obtained by training the data sets, “Bank Marketing” obtained from UCI Machine Learning Repository (2014) for classification and “Credit Card Limit Prediction” available in Kaggle (n.d.) for regression. In the case of classification, the best F1 score of 0.92 was obtained for support vector machines (SVM). In the case of regression, the best value was obtained in the mean absolute error (MAE) of -0.006 for logistic regression and ridge, -0.02. These results were achieved without advanced configurations or prior data cleaning, taking into account that these results were obtained by loading the data as it is in the repositories and with the default configuration of PredictLab.

Keywords: Classification, regression, grid search, F1 score, mean absolute error (MAE).

The project repository is available at: https://github.com/jhonatanvahos/AutoML

1. Descripción del problema

Elegir el modelo predictivo adecuado es uno de los mayores desafíos en el análisis de datos supervisado. Este proceso requiere una comprensión detallada de la variable objetivo, las características del conjunto de datos y las métricas de evaluación. A esto se suma la necesidad de realizar un correcto preprocesamiento de los datos y ajustar los hiperparámetros de los modelos, tareas que pueden resultar complejas y tediosas, especialmente para usuarios con conocimientos limitados en programación o ciencia de datos.

1.1. Problema de negocio

Existe una necesidad en las compañías para tomar decisiones estratégicas a partir de los datos. Sin embargo, muchas empresas enfrentan barreras significativas al intentar aplicar técnicas avanzadas de análisis predictivo. Estas barreras incluyen la falta de personal capacitado en ciencia de datos, la necesidad de herramientas intuitivas para usuarios no técnicos y el tiempo requerido para identificar y configurar modelos adecuados para tareas específicas.

Este proyecto aborda estas limitaciones proporcionando una solución accesible que permite a las empresas automatizar el proceso de análisis predictivo supervisado, simplificando la selección de modelos y la optimización de parámetros.

1.2. Aproximación desde la analítica de datos

Los modelos predictivos desarrollados en este proyecto están diseñados para abordar necesidades específicas del sector financiero, como:

1. **Clasificación**:
    - Identificar clientes con alta probabilidad de incumplimiento en pagos de créditos.
    - Clasificar transacciones como legítimas o fraudulentas para prevenir pérdidas.
    - Detectar perfiles de clientes con mayor probabilidad de adquirir productos financieros específicos.
2. **Regresión**:
    - Estimar la probabilidad de pérdida esperada en carteras de crédito.
    - Proyectar flujos de efectivo de clientes para optimizar la planificación financiera.

- Modelar y predecir variables macroeconómicas que afectan las tasas de interés o inflación.

Este enfoque permite a las organizaciones financieras tomar decisiones más precisas, optimizando la gestión del riesgo, personalizando la experiencia del cliente y aumentando la eficiencia operativa.

1.3. Origen de los datos

Los datos utilizados representan información del sector financiero, como historiales crediticios, características de clientes y encuestas. Estos datos fueron extraídos de UCI Machine Learning Repository y Kaggle.

1.4. Métricas de desempeño

1. **Clasificación**:
    - **Puntaje F1**: Combina precisión y sensibilidad en una sola métrica, lo que resulta ideal para contextos con clases desbalanceadas, como la detección de fraudes. Esta métrica ayuda a minimizar los errores en ambas clases, especialmente cuando los costos asociados a falsos positivos y falsos negativos son significativos.
2. **Regresión**:
    - **Error Absoluto Medio (MAE)**: Calcula el promedio de las diferencias absolutas entre las predicciones y los valores reales. Es fácil de interpretar y refleja de manera clara qué tan lejos están las predicciones de los valores observados, siendo adecuado para estimaciones financieras precisas.

Métricas parametrizables

El sistema permite que el usuario elija otras métricas según el contexto de su análisis. Entre las opciones más comunes se encuentran:

1. **Para clasificación**:
    - **Precisión**: Evalúa qué proporción de las predicciones positivas es correcta, útil en casos donde los falsos positivos tienen un alto costo.
    - **Sensibilidad (Recall)**: Útil cuando el objetivo principal es minimizar los falsos negativos, como en la detección de fraudes o riesgos crediticios.

- **AUC-ROC**: Mide la capacidad del modelo para distinguir entre clases, especialmente relevante cuando las clases están desbalanceadas.

2. **Para regresión**:
    - **RMSE (Root Mean Squared Error)**: Penaliza más los errores grandes, siendo útil cuando estos tienen un impacto significativo en el negocio.
    - **R^2 (Coeficiente de determinación)**: Indica qué tan bien el modelo explica la variabilidad de los datos, adecuado para modelos explicativos.

Valores mínimos esperados

Aunque los valores específicos pueden variar según el contexto del usuario, las métricas predeterminadas están configuradas con valores mínimos recomendados para garantizar un desempeño adecuado:

- **Puntaje F1**: Al menos 0.75, suficiente para asegurar un balance entre precisión y sensibilidad en tareas de clasificación.
- **MAE**: Dentro del 10% del rango promedio de la variable objetivo, lo que asegura predicciones razonablemente precisas para regresión.

Esta flexibilidad permite que la herramienta sea adaptable a diferentes necesidades y contextos del sector financiero, maximizando su utilidad para los usuarios.

2. Objetivos

2.1. Objetivo general

Diseñar un programa interactivo basado en una interfaz web que permita a los usuarios realizar un análisis predictivo supervisado de forma eficiente, facilitando la carga, preprocesamiento y selección del modelo de predicción más adecuado para resolver problemas de clasificación y regresión, utilizando métricas específicas para evaluar el desempeño.

2.2. **Objetivos específicos**

1. Definir los criterios de preprocesamiento de datos que permitan estandarizar las entradas y optimizar el rendimiento de los modelos predictivos, incluyendo técnicas como la imputación de valores faltantes, normalización y codificación de variables categóricas.
2. Desarrollar un sistema que automatice la comparación de modelos de clasificación y regresión utilizando herramientas de hiperparametros para la selección de los mejores modelos.
3. Evaluar la funcionalidad del desarrollo mediante conjuntos de datos del sector financiero, analizando su capacidad para resolver problemas típicos como detección de fraudes, predicción de incumplimientos y estimación de ingresos.
4. Proporcionar al usuario final una herramienta de uso intuitivo para simplificar el entrenamiento y predicción de datos del sector financiero.

3. Datos

3.1. Datos originales

Los conjuntos de datos utilizados para probar la funcionalidad y analizar los resultados fueron obtenidos desde UCI Machine Learning Repository y Kaggle, a continuación, se mencionan los conjuntos de datos utilizados para clasificación y regresión.

1. **Clasificación:**
    - UCI Machine Learning Repository. (2014). Bank marketing dataset. Recuperado de https://archive.ics.uci.edu/dataset/222/bank+marketing
    - Kaggle. Credit Risk Customers. Recuperado de https://www.kaggle.com/datasets/ppb00x/credit-risk-customers
    - Kaggle. Financial Risk for Loan Approval. Recuperado de https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval
    - UCI Machine Learning Repository. Credit Approval. Recuperado de https://archive.ics.uci.edu/dataset/27/credit+approval
    - Kaggle. Credit Card Eligibility Data: Determining Factors. Recuperado de https://www.kaggle.com/datasets/rohit265/credit-card-eligibility-data-determining-factors

2. **Regresión:**
    - Kaggle. Credit Card Limit Prediction. Recuperado de https://www.kaggle.com/datasets/syedasimalishah/credit-card-limit-prediction
    - Kaggle. Credit Score. Recuperado de https://www.kaggle.com/datasets/conorsully1/credit-score.
    - Kaggle. Financial Risk for Loan Approval. Recuperado de https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval

- Kaggle. Crypto Solana Memes Coin. Recuperdo de https://www.kaggle.com/datasets/mafaqbhatti/crypto-solana-memes-coin
- Kaggle. Electronic Store Dataset. Recuperado de https://www.kaggle.com/datasets/rajagrawal7089/electronic-store-dataset

3.2. Datasets

Para la elaboración de los conjuntos de datos en el desarrollo, se siguen una serie de pasos específicos con el fin de preparar adecuadamente los datos para el análisis. Estos pasos incluyen:

1. **Eliminación de Variables:**
   - Se eliminan aquellas variables que poseen un único valor, ya que estas no aportan información relevante para el análisis.
   - Asimismo, se excluyen las variables en las que todos sus valores son distintos, dado que suelen corresponder a identificadores o datos incrementales que no agregan valor al análisis.
2. **Identificación y Tratamiento de Datos:**
   - Se organizan los datos en numéricos, categóricos y mixtos.
   - Para los datos numéricos, se realiza un análisis para verificar la coherencia de los valores dentro del contexto del problema. Por ejemplo, en el caso de representar la edad de una persona, se descartan valores negativos.
   - Los datos categóricos se someten a una transformación básica que incluye la estandarización de caracteres y la eliminación de tildes que podrían afectar la consistencia de los datos. Posteriormente, el usuario puede realizar ajustes manuales si es necesario.
   - Los datos mixtos se separan para que el usuario pueda definir un tratamiento personalizado o la eliminación de los datos.
3. **Imputación de Datos Faltantes:**
   - Se imputan los valores faltantes/nulos de acuerdo con las preferencias del usuario. Este puede establecer un umbral de cantidad de datos faltantes por columna para determinar si se elimina o imputa la información.

- Para los datos numéricos, el usuario puede elegir entre métodos de imputación como media, mediana o moda.

4. **Identificación de Valores Atípicos:**
   - Se utilizan medidas como el puntaje Z y la desviación estándar para identificar valores atípicos en los datos numéricos. Estos valores se eliminan del conjunto de datos.
5. **Selección de la Variable Objetivo:**
   - El usuario selecciona la variable objetivo según el tipo de modelo que se va a aplicar.
   - Para problemas de clasificación, se aplica una codificación de etiquetas (label encoder) a las categorías.
   - Para problemas de regresión, se realiza un escalamiento de los datos según la preferencia del usuario.
6. **Tratamiento de datos:**
   - Para los datos numéricos se aplica un escalamiento utilizando el método de escalamiento estándar (StandardScaler)
   - Para los datos categóricos se utiliza la codificación de un solo paso (OneHotEncoder).

Una vez completadas estas etapas de preparación de datos, se procede a la construcción de los conjuntos principales:

- Conjunto de datos de Entrenamiento: 70% de los datos originales se utilizarán para entrenar el modelo.
- Conjunto de datos de Prueba: 20% de los datos originales se utilizarán para evaluar el modelo.
- Conjunto de datos de Validación: 10% de los datos originales se utilizarán para evaluar el modelo con datos diferentes a los de entrenamiento y prueba.

El tamaño de estas muestras puede ajustarse según las necesidades del usuario, aunque se proporcionan valores predeterminados para facilitar su comprensión y uso.

3.3. Analítica descriptiva

Se realizará un análisis descriptivo preliminar de los datos para comprender mejor su distribución y características principales. Para este análisis se toma como referencia el conjunto de datos llamado “Bank Marketing”. Este análisis incluirá:

- **Descripción general del conjunto de datos:** Nombre de las columnas, cantidad de registros, cantidad de registros no nulos para cada columna y tipo de dato para cada columna.

Figura 1
Descripción general del conjunto de datos

```
--------------------------------------------------
-------- Análisis y visualización de datos --------
--------------------------------------------------
 Información de los datos:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45211 entries, 0 to 45210
Data columns (total 17 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        35210 non-null  float64
 1   job        45211 non-null  object
 2   marital    45209 non-null  object
 3   education  45211 non-null  object
 4   default    45211 non-null  object
 5   balance    45211 non-null  int64
 6   housing    45211 non-null  object
 7   loan       45211 non-null  object
 8   contact    45211 non-null  object
 9   day        45211 non-null  int64
 10  month      45211 non-null  object
 11  duration   45211 non-null  int64
 12  campaign   45211 non-null  int64
 13  pdays      45211 non-null  int64
 14  previous   45211 non-null  int64
 15  poutcome   45211 non-null  object
 16  y          45211 non-null  object
dtypes: float64(1), int64(6), object(10)
```

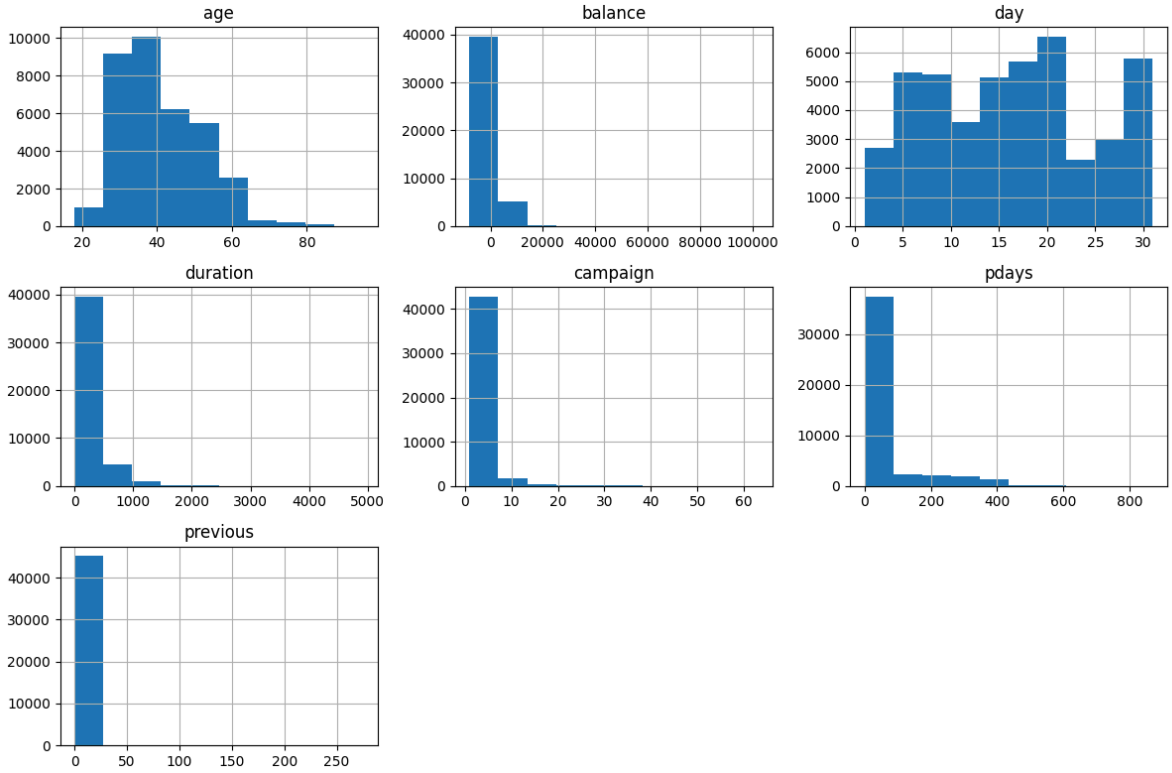
- **Análisis descriptivo para variables numéricas:** Media, desviación estándar, valor mínimo, valor máximo y visualización de los datos para cada columna.

Figura 2
Análisis descriptivo de variables numéricas

```
Análisis descriptivo de variables numéricas:
                age        balance           day      duration      campaign  \
count  35210.000000   45211.000000  45211.000000  45211.000000  45211.000000
mean      40.917211    1362.272058     15.806419    258.163080      2.763841
std       10.577504    3044.765829      8.322476    257.527812      3.098021
min       18.000000   -8019.000000      1.000000      0.000000      1.000000
25%       33.000000      72.000000      8.000000    103.000000      1.000000
50%       39.000000     448.000000     16.000000    180.000000      2.000000
75%       48.000000    1428.000000     21.000000    319.000000      3.000000
max       95.000000  102127.000000     31.000000   4918.000000     63.000000

              pdays      previous
count  45211.000000  45211.000000
mean      40.197828      0.580323
std      100.128746      2.303441
min       -1.000000      0.000000
25%       -1.000000      0.000000
50%       -1.000000      0.000000
75%       -1.000000      0.000000
max      871.000000    275.000000
```

Figura 3
Gráfica de variables numéricas



- **Análisis descriptivo para variables categóricas:** Cantidad de categorías únicas, categoría que más se repite y visualización de los datos para cada columna.

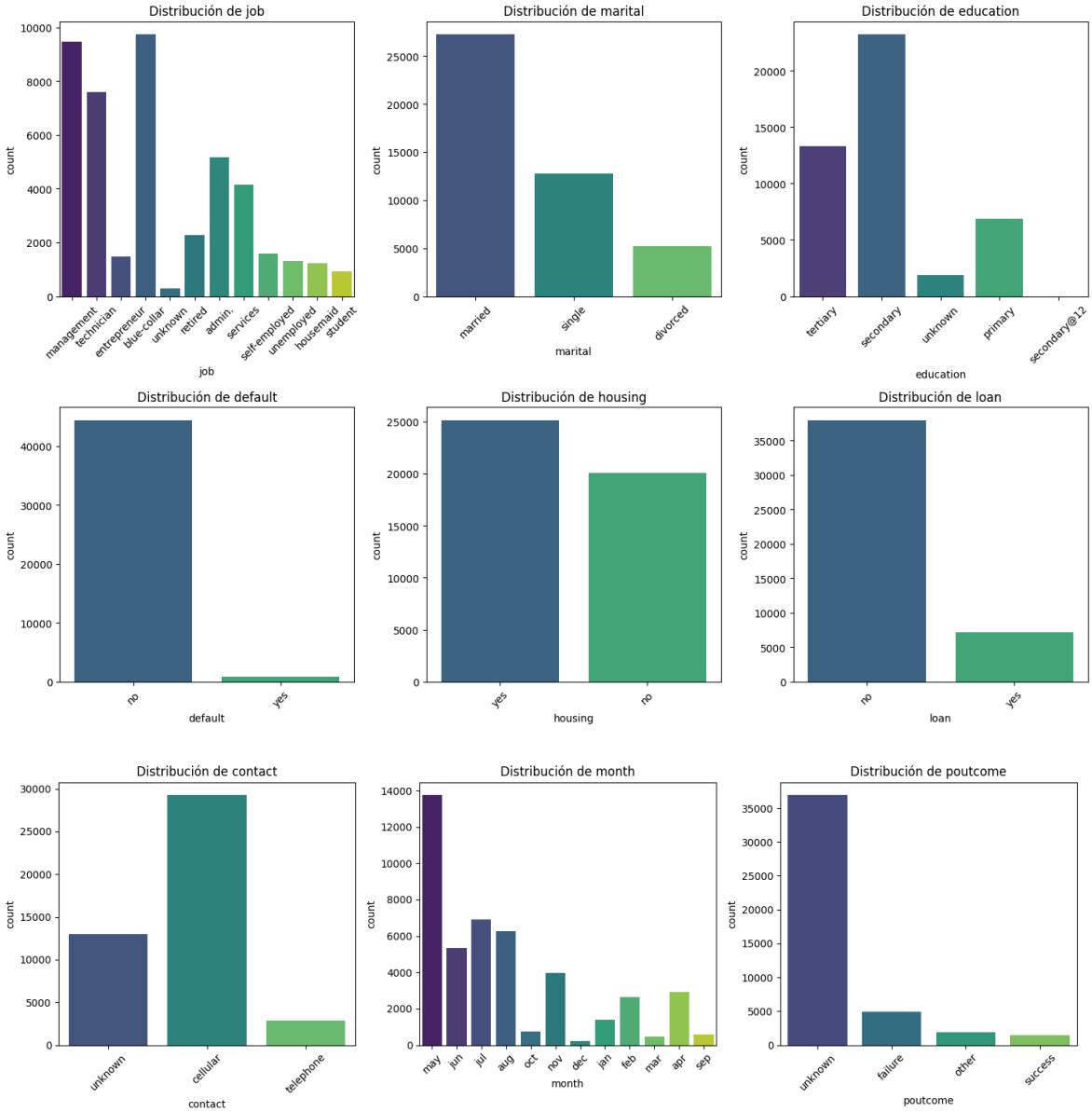
Figura 4
Análisis descriptivo de variables categóricas

```
Análisis descriptivo de variables categóricas:
                job  marital  education default housing   loan   contact  \
count         45211    45209      45211   45211   45211  45211     45211
unique           12        3          5       2       2      2         3
top     blue-collar  married  secondary      no     yes     no  cellular
freq          9732    27212      23201   44396   25130  37967     29285

        month poutcome
count   45211    45211
unique     12        4
top       may  unknown
freq    13766    36959
```

Figura 5

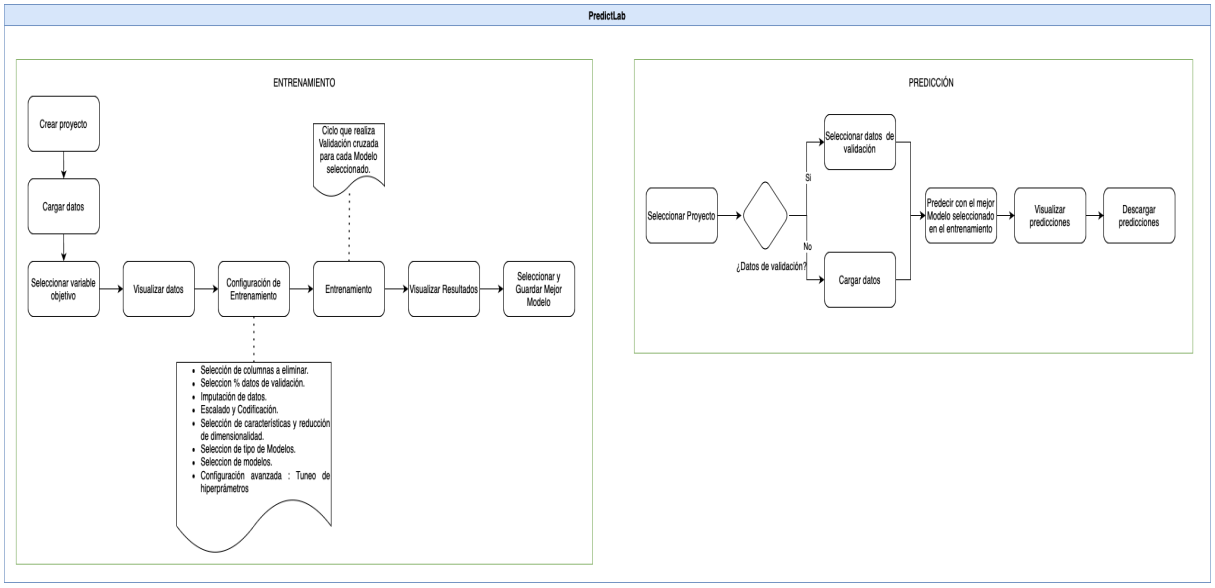
Gráfica de variables categóricas



4. Proceso de analítica

4.1. Pipeline principal

Figura 6
Pipeline principal



4.2. Preprocesamiento

El preprocesamiento de los datos es un paso crucial en cualquier proyecto de modelado predictivo, ya que garantiza que los datos sean adecuados para los modelos a utilizar. En este proyecto, se han considerado y aplicado varias técnicas de preprocesamiento, con el objetivo de mejorar la calidad de los datos y optimizar el rendimiento de los modelos de clasificación y regresión.

1. **Manejo de Valores Atípicos (Outliers):** Para manejar los valores atípicos, se utiliza un enfoque robusto basado en el puntaje Z ajustado, que emplea la mediana y la desviación absoluta mediana (MAD) en lugar de la media y la desviación estándar, lo que lo hace menos sensible a los valores extremos. Este método permite identificar y eliminar los datos atípicos de manera efectiva, asegurando que el modelo no se vea afectado por valores que podrían distorsionar los resultados.
2. **Imputación de Datos Faltantes:** Los valores faltantes en las columnas numéricas son imputados de acuerdo con dos estrategias:

- **Imputación Simple**: Se utiliza la media, la mediana o la moda según el tipo de variable y la cantidad de datos faltantes.
- **Imputación Variable**: Cuando el porcentaje de valores faltantes es significativo, se utiliza el algoritmo de K vecinos más cercanos (KNN) para imputar los valores de forma más precisa, basándose en la información de otras muestras similares. Para las variables categóricas, se emplea la imputación mediante la moda o se utiliza imputación variable después de codificar las categorías.

3. **Escalado de Datos:** Dado que las características numéricas pueden tener escalas muy diferentes, se aplica un escalado estándar (media = 0, desviación estándar = 1) a las variables numéricas. Este paso es esencial para que el modelo no se vea influenciado por la magnitud de las características. Además, la variable objetivo es escalada utilizando diferentes transformaciones, como escalamiento estándar, normalización y transformaciones logarítmicas
4. **Codificación de Variables Categóricas:** Las variables categóricas son codificadas utilizando codificación de un solo paso, una técnica que crea nuevas columnas para cada categoría, representando su presencia con valores binarios. Este proceso permite que los modelos puedan trabajar con variables no numéricas de manera adecuada.
5. **Reducción de Dimensionalidad:** Para optimizar la eficiencia y reducir la complejidad del modelo, se aplica Análisis de Componentes Principales (PCA). PCA permite reducir el número de características, manteniendo la mayor parte de la varianza de los datos, lo que facilita el entrenamiento del modelo sin perder información relevante. Se opta por un enfoque de reducción basado en un porcentaje de varianza explicada.
6. **Selección de Características:** Se emplean técnicas como selección de mejores parámetros y Recursive Feature Elimination (RFE) para seleccionar las características más relevantes. Estas técnicas ayudan a identificar y conservar las variables que tienen un mayor impacto en el rendimiento del modelo, eliminando aquellas que no aportan valor significativo.
7. **Balanceo de Clases:** En los casos donde existe desbalance en las clases, se implementan técnicas de balanceo utilizando SMOTE (Synthetic Minority Over-sampling Technique) y RandomOverSampler (ROS), con el objetivo de mejorar el desempeño de los modelos en situaciones de clasificación desbalanceada.

4.3. Modelos

El proyecto incorpora una variedad de modelos supervisados tanto para tareas de regresión como de clasificación, cada uno configurado con parámetros por defecto que pueden ser modificados por el usuario desde la interfaz web. A continuación, se detallan los modelos utilizados, junto con sus configuraciones predeterminadas:

Modelos de regresión

1. **Regresión Lineal**
    - Parámetros por defecto:
        - fit_intercept: [True, False]
    - Este modelo es útil para problemas lineales simples, ajustando una línea recta que minimiza el error cuadrático medio.
2. **Ridge**
    - Parámetros por defecto:
        - alpha: [0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0]
    - Modelo que introduce regularización L2, lo que reduce el sobreajuste al penalizar coeficientes grandes.
3. **Random Forest Regressor**
    - Parámetros por defecto:
        - n_estimators: [20, 50, 100, 200]
        - max_depth: [5, 10, 20]
        - max_features: ["None", "log2", "sqrt"]
        - criterion: ["squared_error", "absolute_error", "friedman_mse", "poisson"]
    - Un enfoque basado en ensambles que utiliza múltiples árboles de decisión para mejorar la precisión.
4. **AdaBoost Regressor**
    - Parámetros por defecto:
        - n_estimators: [10, 30, 50, 70, 100]
        - learning_rate: [0.001, 0.01, 0.1]
    - Método basado en ensambles que ajusta iterativamente modelos débiles para reducir errores residuales.
5. **Gradient Boosting Regressor**

- Parámetros por defecto:
    - n_estimators: [10, 30, 50, 70, 100]
    - learning_rate: [0.1, 0.01, 0.001]
    - max_depth: [3, 5, 7]
- Algoritmo que optimiza árboles de decisión secuencialmente para minimizar errores.

Modelos de clasificación

1. **Regresión Logística**
    - Parámetros por defecto:
        - multi_class: ["ovr", "multinomial"]
        - solver: ["liblinear", "lbfgs", "newton-cg", "newton-cholesky", "sag", "saga"]
        - class_weight: ["balanced"]
        - max_iter: [1000]
    - Modelo simple y eficiente para problemas de clasificación binaria y multiclase.
2. **Random Forest**
    - Parámetros por defecto:
        - n_estimators: [20, 50, 100, 200, 300]
        - max_features: [5, 7, 9]
        - max_depth: [5, 10, 20, 30, 40, 50]
        - criterion: ["gini", "entropy"]
    - Utiliza múltiples árboles de decisión para mejorar la precisión y manejar datos no lineales.
3. **SVM (Máquinas de Soporte Vectorial)**
    - Parámetros por defecto:
        - kernel: ["linear", "rbf", "poly"]
        - C: [0.1, 1, 10]
        - gamma: ["scale", "auto", 1.0]
    - Es adecuado para problemas de clasificación con un margen claro entre clases.
4. **KNN (K-Nearest Neighbors)**
    - Parámetros por defecto:

        - n_neighbors: [3, 5, 7, 9]
        - weights: ["uniform", "distance"]
        - metric: ["euclidean", "manhattan", "minkowski"]
        - p: [1, 2]
    - Algoritmo sencillo basado en la proximidad de los puntos en el espacio de características.
5. **GaussianNB**
    - No requiere configuración adicional, siendo un modelo rápido para datos con distribución normal (por lo general es un modelo no paramétrico).
6. **BernoulliNB**
    - Similar al GaussianNB, ideal para variables binarias.

4.4. Métricas

Para calcular las métricas de desempeño de los modelos, se utilizaron funciones integradas en la biblioteca scikit-learn. En el caso de los modelos de clasificación, se emplearon las métricas exactitud (accuracy_score), precision (precision_score), exhaustividad (recall_score), puntaje F1 (f1_score), y el área bajo la curva ROC (roc_auc_score). Estas métricas permiten evaluar el balance entre los falsos positivos y falsos negativos, así como el desempeño global del modelo.

Para los modelos de regresión, se calcularon el Error Absoluto Medio (MAE) usando mean_absolute_error, el Error Cuadrático Medio (MSE) mediante mean_squared_error, y el Coeficiente de Determinación (R^2) utilizando r2_score. Estas métricas se eligieron por su capacidad para medir la magnitud de los errores de predicción y explicar la variabilidad de la variable dependiente.

Además, todas las métricas se calcularon con validación cruzada estratificada de 5 pliegues, asegurando una evaluación robusta y generalizable del desempeño de los modelos.

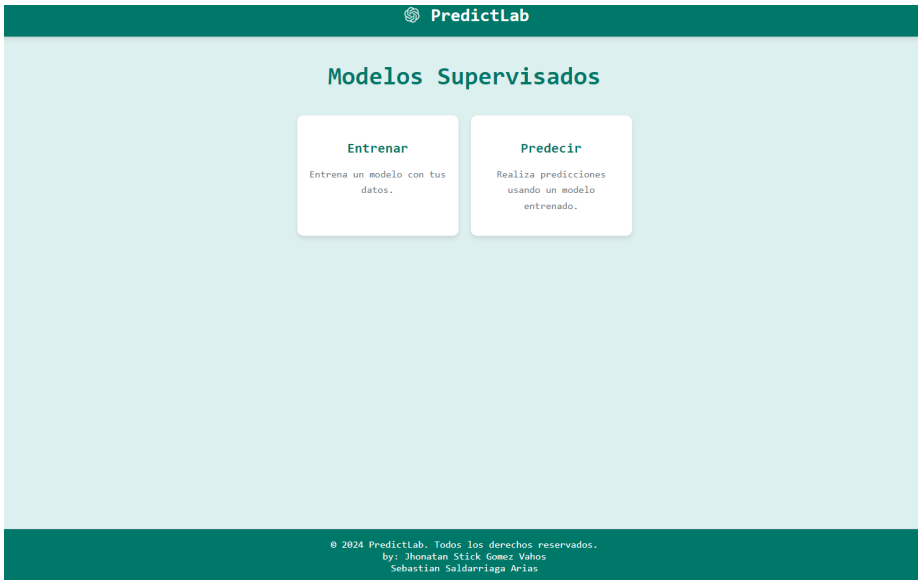
5. Metodología

5.1. Baseline

La solución implementada consta de un front, en el cual el usuario final interactúa con la herramienta y un back, en el cual se realiza el preprocesamiento de los datos y entrenamiento de los modelos.

El front cuenta con dos secciones iniciales en las que el usuario puede interactuar según su necesidad: Entrenar y predecir.

Figura 7
Ventana principal del front

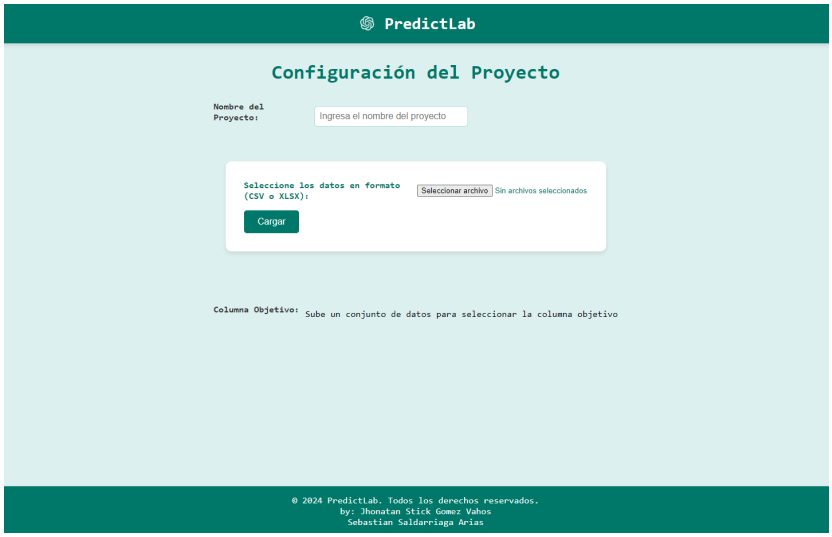


5.2.Validación

Para interactuar por primera vez con PredictLab, el usuario debe dar clic en el botón “Entrenar” (**Figura 7**) y de esta manera podrá configurar todos los hiperparámetros relacionados con su conjunto de datos.

Figura 8

Ventana inicial de la configuración del proyecto para entrenamiento

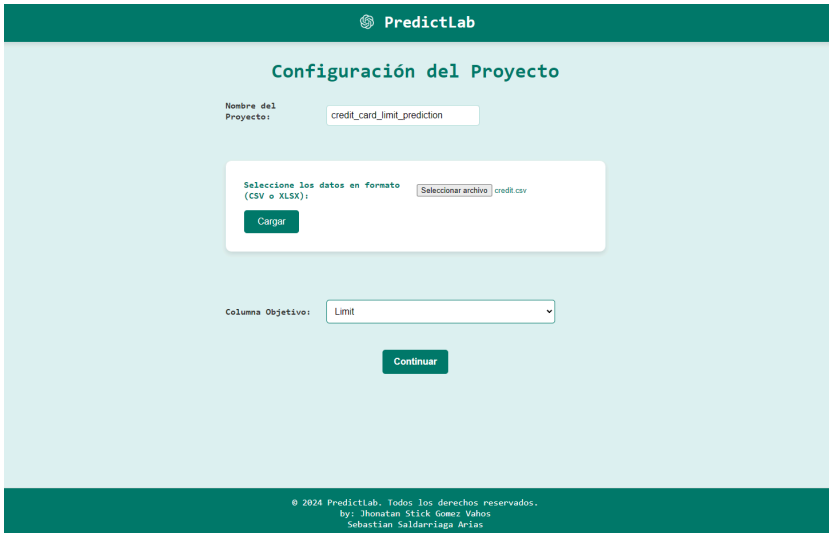


5.3. Iteraciones y evolución

A continuación, se describe como ejemplo una de las iteraciones realizadas, en este caso para el conjunto de datos llamado “Credit Card Limit Prediction” el cual se trata de un problema de regresión.

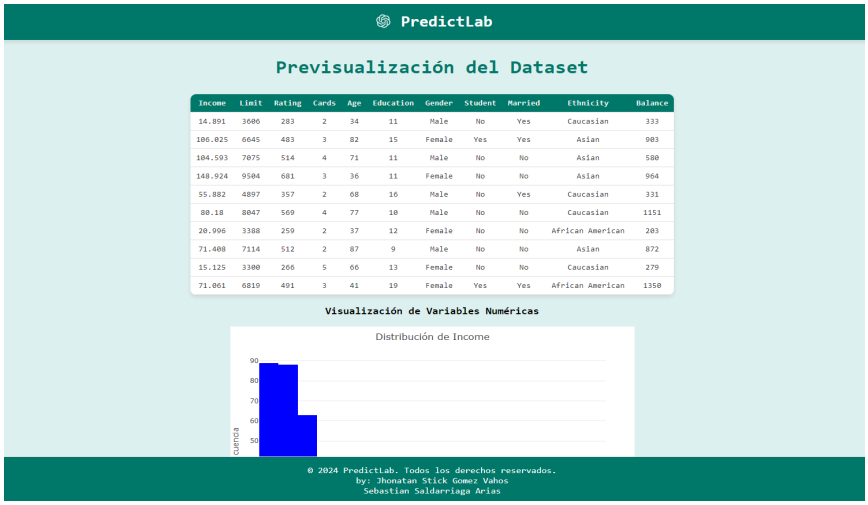
Figura 9

Configuración inicial del proyecto “Credit Card Limit Prediction”



Al dar clic en el botón “Continuar” vemos una previsualización del conjunto de datos y gráficas para cada columna, se muestran algunas.

Figura 10
Previsualización del conjunto de datos



Al final de este apartado hay un botón llamado “Continuar”, el cual nos lleva a configurar los hiperparámetros y los modelos a utilizar.

Figura 11
Primeros parámetros para el entrenamiento

Figura 12
Segundos parámetros para el entrenamiento

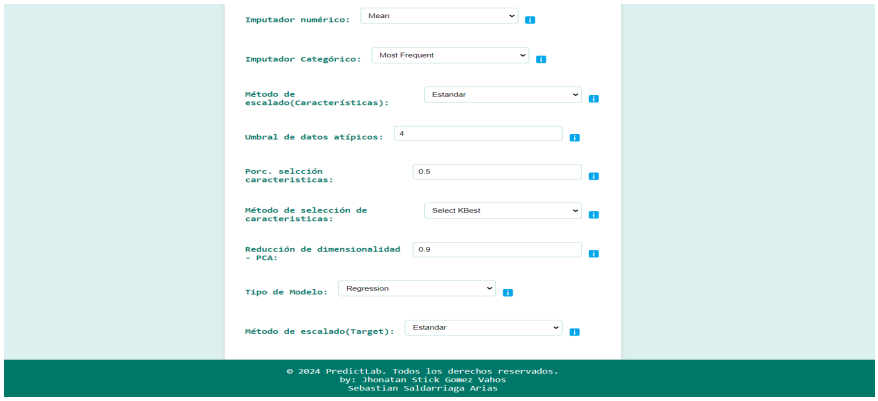
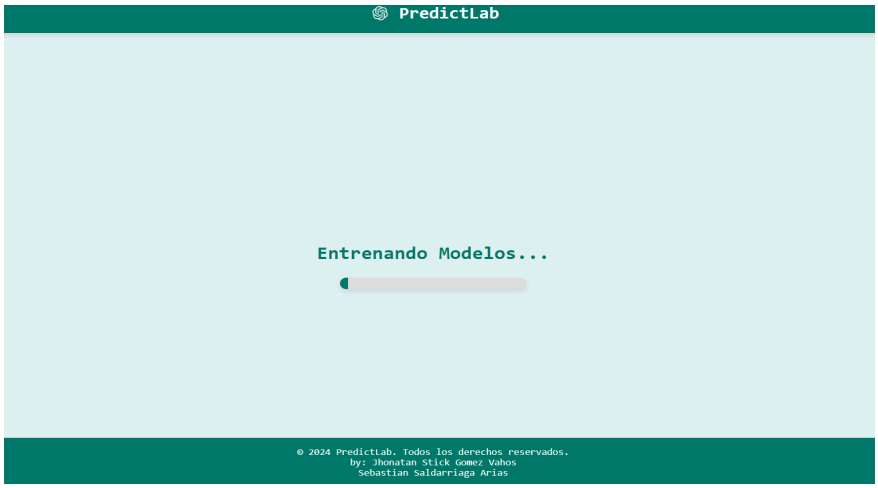


Figura 13
Terceros parámetros para el entrenamiento

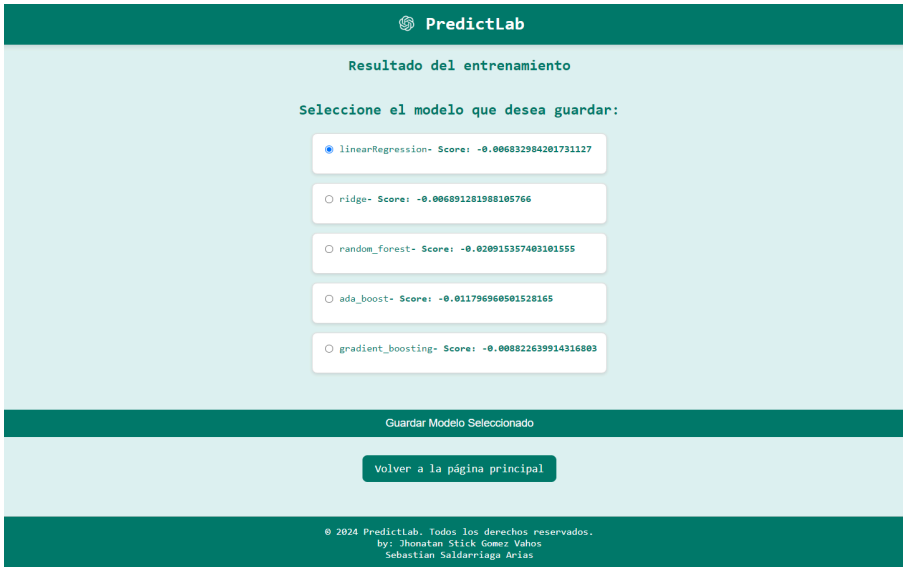
Una vez configurados los parámetros se procede a iniciar el entrenamiento dando clic en el botón “Entrenar modelos”.

Figura 14
Entrenamiento de modelos



Al momento de finalizar la etapa de entrenamiento, se visualizan las métricas obtenidas para cada uno de los modelos seleccionados.

Figura 15
Métricas de entrenamiento y selección del mejor modelo

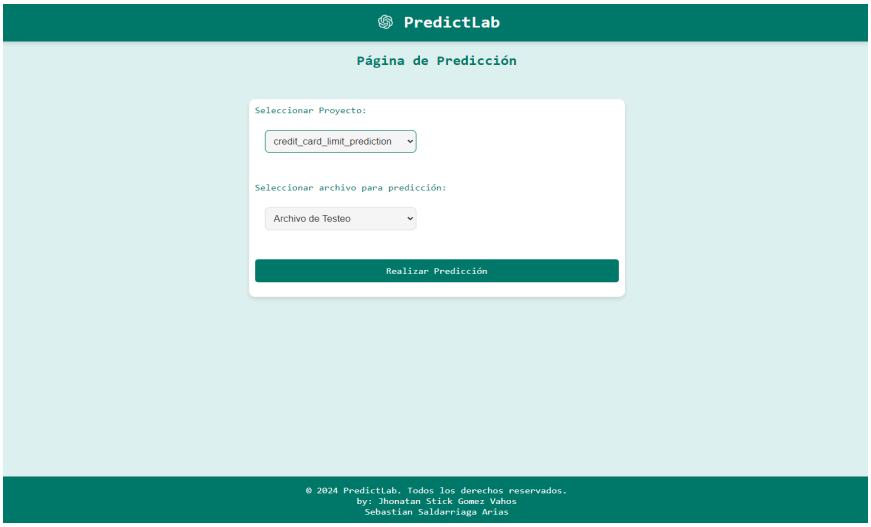


Se selecciona el modelo de preferencia del usuario y se da clic en “Guardar Modelo Seleccionado”.

Para realizar una predicción se da clic en el botón “volver a la página principal” y desde allí se da clic en el botón “Predecir” (**Figura 7**).

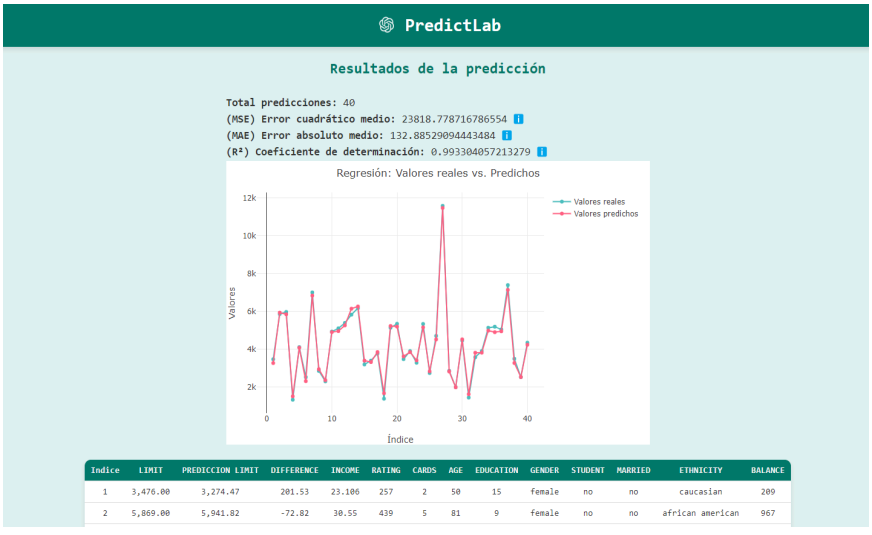
En este apartado se selecciona el proyecto que se está trabajando y está la opción de trabajar con el archivo de testeo generado automáticamente por PredictLab o bien el usuario puede cargar un archivo de testeo, el cual debe cumplir con la misma estructura del conjunto de datos de entrenamiento.

Figura 16
Configuración para realizar una predicción



Al dar clic en el botón “Realizar Predicción” se muestran las métricas de la predicción, gráfica comparativa de los datos originales y su respectiva predicción y una previsualización del conjunto de datos con su predicción.

Figura 17
Resultados de la predicción



Al final de este apartado hay un botón llamado “Descargar Resultados (CSV)” el cual sirve para exportar en formato csv el conjunto de datos con su respectiva predicción.

5.4 Herramientas

Para el desarrollo de PredictLab, se emplearon las siguientes herramientas:

1. **Lenguajes de programación**:
    - **Python**: Para la implementación del backend, incluyendo el procesamiento de datos, la optimización de modelos y la lógica del programa.
2. **Bibliotecas**:
    - **Scikit-learn**: Biblioteca clave para la implementación de los algoritmos de aprendizaje supervisado y el proceso de búsqueda de rejillas (grid search) (Pedregosa et al., 2011).
3. **Entornos de desarrollo**:
    - **VS Code**: Para la edición del código y pruebas locales del frontend y backend.
    - **Frontend:** El desarrollo de la interfaz de usuario se realizó con React (Meta Platforms, Inc., 2013), un framework de JavaScript que permite construir interfaces web interactivas y dinámicas de manera eficiente.
    - **Backend:** Se utilizó FastAPI (FastAPI, s. f.), un framework moderno de Python, que facilita la creación de aplicaciones web y APIs de alto rendimiento y escalabilidad.

6. Resultados y discusión

6.1. Métricas

Los resultados obtenidos por PredictLab para las tareas de clasificación y regresión en los conjuntos de datos mencionados en la sección **3.1. Datos originales** se observan en la **Tabla 1** para clasificación y en la **Tabla 2** para regresión.

Tabla 1
Puntaje F1 para algoritmos de clasificación

Conjunto de datos	Regresión Logística	Random Forest	SVM	KNN	GaussianNB	BernoulliNB
Bank marketing dataset	0.81	0.88	**0.92**	0.91	0.64	0.74
Credit Risk Customers	0.76	0.84	**0.91**	0.80	0.48	0.75
Financial Risk for Loan Approval	**0.99**	**0.99**	**0.99**	0.98	0.97	0.91
Credit Approval	0.85	**0.89**	0.87	0.86	0.80	0.86
Credit Card Eligibility Data	0.72	**0.86**	0.82	0.81	0.71	0.74

Tabla 2
Error Absoluto Medio (MAE) para algoritmos de regresión

Conjunto de datos	Regresión Lineal	Ridge	Random Forest	AdaBoost	Gradient Boosting
Credit Card Limit Prediction	**-0.006**	**-0.006**	0.021	-0.118	-0.009
Credit Score	-0.210	**-0.205**	-0.282	-0.299	-0.252
Financial Risk for Loan Approval	-0.193	-0.193	-0.161	-0.209	**-0.148**
Crypto Solana Memes Coin	-3.724	-0.776	-0.459	-0.057	**-0.004**
Electronic Store Dataset	-0.125	-0.125	-0.109	-0.102	**-0.097**

6.2. Evaluación cualitativa

Los resultados obtenidos por PredictLab muestran un desempeño sólido en la mayoría de las tareas de clasificación y regresión, pero también revelan algunos desafíos relacionados con el ajuste de los modelos en ciertos conjuntos de datos.

Clasificación

1. **Desempeño general:**

    - Los puntajes F1 son consistentemente altos para modelos como SVM y Random Forest, alcanzando valores de hasta 0.99 en el conjunto de datos *Financial Risk for Loan Approval*. Esto indica que estos modelos lograron capturar patrones relevantes en los datos sin comprometer la precisión o la sensibilidad.
    - Por el contrario, los modelos basados en Bayes (GaussianNB y BernoulliNB) presentan resultados inferiores en todos los conjuntos de datos, lo que podría estar relacionado con la falta de cumplimiento de sus supuestos como lo es la distribución normal de los datos.

2. **Casos de sobreajuste y ajuste insuficiente:**

    - El desempeño casi perfecto de los modelos en el conjunto *Financial Risk for Loan Approval* (puntajes F1 de 0.99 para la mayoría de los algoritmos) puede ser indicativo de sobreajuste, especialmente si el conjunto de datos contiene un desbalance en las clases o características redundantes.
    - Modelos como GaussianNB muestran signos de ajuste insuficiente, particularmente en *Credit Risk Customers*, con un puntaje F1 de solo 0.48. Esto sugiere que no lograron capturar adecuadamente las relaciones entre las características y las etiquetas.

3. **Utilidad de los resultados:**

    Los puntajes F1 altos en contextos financieros, como la aprobación de préstamos, son cruciales, ya que aseguran decisiones bien informadas al reducir tanto los falsos positivos como los falsos negativos.

Regresión

1. **Desempeño general:**

    - En tareas de regresión, algoritmos como Gradient Boosting y Random Forest se destacan por su precisión, obteniendo los valores de MAE más bajos en la mayoría de los casos, como en el conjunto de datos *Crypto Solana Memes Coin* (MAE de -0.004). Esto evidencia una excelente capacidad de predicción en datos complejos y posiblemente no lineales.
    - Por otro lado, modelos lineales como Regresión Lineal y Ridge tienen un rendimiento competitivo en algunos conjuntos de datos como por ejemplo en *Credit Card Limit Prediction*, pero presentan limitaciones en escenarios más complejos como *Credit Score*.

2. **Casos de sobreajuste y ajuste insuficiente:**

    - El rendimiento extremadamente bajo de algunos modelos, como AdaBoost en *Credit Card Limit Prediction* (MAE de -0.118), podría ser un indicativo de sobreajuste si el modelo se ajusta demasiado a determinados patrones en los datos de entrenamiento.
    - En conjuntos de datos como *Credit Score*, el rendimiento más bajo de todos los algoritmos sugiere que los datos pueden no tener una relación lineal clara o que requieren características adicionales para mejorar el ajuste.

3. **Utilidad de los resultados:**

    En tareas financieras, como la predicción de límites de crédito, un MAE bajo asegura predicciones confiables y precisas, lo cual es esencial para evitar riesgos financieros.

En conclusión, aunque PredictLab logra resultados notables, algunos casos indican la necesidad de ajustes adicionales para evitar problemas de sobreajuste o ajuste insuficiente. Sin embargo, el desempeño global demuestra que la herramienta tiene el potencial de proporcionar métricas útiles y directamente aplicables en contextos de negocio.

7. Conclusiones

El desarrollo de PredictLab como una herramienta interactiva para análisis predictivo supervisado representó un avance significativo en la simplificación y accesibilidad del proceso de construcción de modelos de clasificación y regresión. A través de una interfaz intuitiva, el sistema permite a usuarios con distintos niveles de experiencia cargar datos, realizar preprocesamientos personalizados y seleccionar modelos óptimos basados en métricas relevantes como el puntaje F1 y el MAE.

Entre los principales logros del proyecto se destacan:

1. **Resultados robustos en clasificación y regresión**:
    - En el caso de clasificación, se obtuvieron puntajes F1 elevados (superiores a 0.9) con algoritmos como SVM y Random Forest en datasets relevantes, demostrando su capacidad para capturar patrones complejos.
    - En tareas de regresión, los modelos generados alcanzaron errores absolutos medios bajos, destacando Gradient Boosting como el modelo más consistente.
2. **Optimización de modelos sin intervención manual**:
    - La implementación de la búsqueda de rejillas para la optimización de hiperparámetros permitió mejorar el desempeño de los modelos de manera automatizada, eliminando la necesidad de ajustes manuales complejos.
3. **Facilidad de uso y enfoque práctico**:
    - PredictLab ofrece una solución amigable para usuarios no técnicos, eliminando barreras asociadas al desarrollo de modelos predictivos desde cero, lo que contribuye a la democratización del análisis de datos.

En relación con los objetivos planteados, PredictLab cumple su propósito como una herramienta práctica y eficiente para el análisis predictivo de datos financieros.

8. Recomendaciones

A partir del desarrollo y los resultados obtenidos con PredictLab, se identifican diversas oportunidades de mejora y expansión que podrían enriquecer su utilidad y alcance en futuras investigaciones:

1. **Incorporación de modelos no supervisados**
    - Ampliar la funcionalidad de PredictLab para incluir algoritmos de agrupamiento, reducción de dimensionalidad y detección de anomalías, lo que permitiría abordar un rango más amplio de problemas analíticos, especialmente en casos donde no se cuenta con etiquetas predefinidas.
2. **Desarrollo de capacidades para datos en tiempo real**
    - Implementar la integración con flujos de datos en tiempo real que puedan apalancar aplicaciones en entornos dinámicos como detección de fraudes financieros o análisis de transacciones bancarias.
3. **Evaluación en escenarios del mundo real**
    - Realizar pruebas en empresas del sector financiero para evaluar la eficacia del desarrollo en condiciones reales. Esto podría ayudar a identificar posibles limitaciones y adaptar el sistema a requisitos específicos del negocio.
4. **Optimización del desempeño computacional**
    - Implementar el uso de frameworks de computación distribuida, como Apache Spark, para manejar conjuntos de datos más grandes y mejorar la velocidad de procesamiento.
5. **Creación de documentación y tutoriales accesibles**
    - Desarrollar materiales educativos, como guías paso a paso y videos explicativos, que faciliten el aprendizaje y uso de PredictLab, especialmente para usuarios novatos en ciencia de datos.

Referencias

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://scikit-learn.org

FastAPI. (s. f.). *FastAPI*. https://fastapi.tiangolo.com

Meta Platforms, Inc. (2013). *React: A JavaScript Library for Building User Interfaces*. https://react.dev

Scikit-learn Developers**.** (2023). *GridSearchCV: Exhaustive search over specified parameter values for an estimator*. Recuperado de https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

UCI Machine Learning Repository. (2014). Bank marketing dataset. Recuperado de https://archive.ics.uci.edu/dataset/222/bank+marketing

Kaggle. Credit Risk Customers. Recuperado de https://www.kaggle.com/datasets/ppb00x/credit-risk-customers

Kaggle. Financial Risk for Loan Approval. Recuperado de https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval

UCI Machine Learning Repository. Credit Approval. Recuperado de https://archive.ics.uci.edu/dataset/27/credit+approval

Kaggle. Credit Card Eligibility Data: Determining Factors. Recuperado de https://www.kaggle.com/datasets/rohit265/credit-card-eligibility-data-determining-factors

Kaggle. Credit Card Limit Prediction. Recuperado de https://www.kaggle.com/datasets/syedasimalishah/credit-card-limit-prediction

Kaggle. Credit Score. Recuperado de https://www.kaggle.com/datasets/conorsully1/credit-score.

Kaggle. Financial Risk for Loan Approval. Recuperado de https://www.kaggle.com/datasets/lorenzozoppelletto/financial-risk-for-loan-approval

Kaggle. Crypto Solana Memes Coin. Recuperdo de https://www.kaggle.com/datasets/mafaqbhatti/crypto-solana-memes-coin

Kaggle. Electronic Store Dataset. Recuperado de https://www.kaggle.com/datasets/rajagrawal7089/electronic-store-dataset

Anexos

Anexo 1. Descripción del conjunto de datos “Bank marketing”

Este conjunto de datos contiene información de una campaña telefónica de marketing de un banco portugués. La variable objetivo es la respuesta a si un cliente suscribirá un depósito a plazo.

Las variables del conjunto de datos son:

- **age:** Edad del cliente
- **job:** Tipo de trabajo (categórico: 'admin.','blue-collar','empresario','housemaid','gestión','retirado','autónomo','servicios','estudiante','técnico','desempleado','desconocido')
- **marital:** Estado civil (categórico: 'divorciado','casado','soltero','desconocido'; nota: 'divorciado' significa divorciado o viudo)
- **education:** Nivel de Educación (categórico: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
- **default:** ¿Tiene crédito en incumplimiento?
- balance: Saldo medio anual en euros
- **housing:** ¿Tiene préstamo de vivienda?
- **loan:** ¿Tiene préstamo personal?
- **contact:** Tipo de comunicación de contacto (categórico: 'celular','telefono')
- **day_of_week:** Último día de contacto de la semana
- **month:** último mes de contacto del año (categórico: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
- **duration:** Duración del último contacto, en segundos (numérico)
- **campaign:** Número de contactos realizados durante esta campaña y para este cliente (numérico, incluye último contacto)
- **pdays:** Número de días que pasaron después de que el cliente fue contactado por última vez de una campaña anterior (numérico; - 1 significa que el cliente no fue contactado previamente)
- **previous:** Número de contactos realizados antes de esta campaña y para este cliente
- **poutcome:**Resultado de la campaña de marketing anterior (categorical: 'failure','nonexistent','success')
- **y:** ¿El cliente ha suscrito un depósito a plazo? (Variable objetivo)

El conjunto de datos tiene un peso de 4,547 KB, 45211 registros y en el rango de tiempo comprendido entre mayo del 2008 y noviembre 2010.

Anexo 2. Descripción del conjunto de datos "Credit Risk Customers"

El conjunto de datos *Credit Risk Customers* fue obtenido de la plataforma Kaggle y está diseñado para tareas de análisis y predicción del riesgo crediticio. Este conjunto de datos contiene información detallada sobre clientes de una institución financiera, lo que permite evaluar el comportamiento crediticio y la probabilidad de incumplimiento.

Las variables del conjunto de datos son:

- customer_id: Identificador único del cliente.
- age: Edad del cliente (en años).
- gender: Género del cliente (Male/Female).
- annual_income: Ingreso anual del cliente (en dólares).
- credit_score: Puntuación crediticia asignada al cliente.
- debt: Cantidad de deuda actual del cliente (en dólares).
- loan_amount: Monto del préstamo solicitado.
- loan_approval_status: Estado de aprobación del préstamo (Approved/Rejected).
- employment_status: Estado laboral del cliente (Employed/Unemployed).
- account_balance: Saldo en la cuenta del cliente (en dólares).
- marital_status: Estado civil del cliente (Single/Married).
- dependents: Número de dependientes a cargo del cliente.
- loan_purpose: Propósito del préstamo (e.g., Personal Loan, Mortgage).
- risk_level: Nivel de riesgo asignado al cliente (Low, Medium, High).

El conjunto de datos consta de 10,000 registros y tiene un peso de 153,02 KB

Anexo 3. Descripción del conjunto de datos "Financial Risk for Loan Approval"

Este conjunto de datos, disponible en Kaggle, está diseñado para analizar y predecir riesgos financieros relacionados con la aprobación de préstamos. Contiene múltiples características que permiten realizar tanto tareas de clasificación como de regresión, dependiendo de la variable objetivo seleccionada.

El conjunto de datos incluye información financiera y demográfica de solicitantes de préstamos, como ingresos, historial crediticio, puntuaciones de riesgo, entre otras características relevantes.

Las variables clave utilizadas son:

- **LoanApproved**: Variable objetivo utilizada para tareas de clasificación. Indica si un préstamo fue aprobado (1) o rechazado (0).
- **RiskScore**: Variable objetivo utilizada para tareas de regresión. Representa una puntuación numérica que mide el riesgo asociado a cada solicitante.

En términos generales el conjunto de datos cuenta con las siguientes características:

- Información demográfica, como edad y género.
- Detalles financieros, incluyendo ingresos mensuales, deuda existente y monto solicitado del préstamo.
- Historial crediticio, como la cantidad de pagos atrasados y la relación deuda-ingreso.

En el proyecto PredictLab, este conjunto de datos fue utilizado para:

- **Clasificación**: Predecir la aprobación del préstamo (LoanApproved), con métricas como el puntaje F1 para evaluar el desempeño de los modelos.
- **Regresión**: Estimar la puntuación de riesgo (RiskScore), evaluada mediante el error absoluto medio (MAE).

Anexo 4. Descripción del conjunto de datos "Credit Approval"

Este conjunto de datos, disponible en el repositorio UCI Machine Learning, se utiliza para el análisis y predicción de aprobaciones crediticias. Su estructura permite realizar tareas de clasificación, enfocándose en determinar si una solicitud de crédito será aprobada o no, basándose en diversas características del solicitante.

El conjunto de datos contiene información sobre características demográficas, financieras y de comportamiento de los solicitantes. Incluye tanto variables numéricas como categóricas, algunas de las cuales contienen valores faltantes.

Las variables clave del conjunto de datos son:

- **Clase objetivo**: Indica si la solicitud de crédito fue aprobada (+) o denegada (-).
- **Atributos predictivos**: Incluyen características como edad, ingresos, estado civil, historial crediticio, cantidad de crédito solicitado, entre otros.

Algunas columnas están codificadas con valores numéricos o alfanuméricos para proteger la privacidad de los solicitantes.

El conjunto de datos contiene 690 registros con 15 columnas, incluyendo la variable objetivo.

Anexo 5. Descripción del conjunto de datos "Credit Card Eligibility Data"

Este conjunto de datos, disponible en Kaggle, se utiliza para analizar los factores que influyen en la elegibilidad de una persona para obtener una tarjeta de crédito. Su estructura permite abordar tareas de clasificación relacionadas con la aprobación o rechazo de solicitudes de tarjetas de crédito.

El conjunto de datos contiene información demográfica, financiera y comportamental de los solicitantes, proporcionando un panorama integral de los factores que impactan la decisión de otorgar una tarjeta de crédito.

La variable objetivo del conjunto de datos es "Target", la cual indica si el solicitante es elegible (1) o no (0) para recibir una tarjeta de crédito.

Otras variables incluyen características como ingresos mensuales, deudas existentes, historial de crédito, edad, y otros factores relacionados con el perfil financiero y demográfico del solicitante.

Anexo 6. Descripción del conjunto de datos "Credit Card Limit Prediction"

Este conjunto de datos, disponible en Kaggle, se utiliza para predecir los límites de crédito asignados a los clientes en función de sus características financieras y demográficas. Su enfoque principal es la construcción de modelos de regresión para estimar el monto del límite de crédito.

El conjunto de datos contiene información detallada sobre diversos atributos que influyen en la asignación de límites de crédito, incluyendo datos financieros, demográficos y de comportamiento del cliente.

La variable objetivo "CreditLimit" representa el límite de crédito asignado al cliente.

El conjunto de datos incluye variables como ingresos anuales, gastos totales, nivel de deuda, edad, puntaje crediticio, entre otras características relevantes para evaluar la solvencia del cliente.

Este conjunto de datos fue utilizado exclusivamente para tareas de regresión en el proyecto PredictLab, enfocándose en predecir la variable "CreditLimit". El desempeño de los modelos se evaluó mediante el error absoluto medio (MAE).

Anexo 7. Descripción del conjunto de datos “Credit Score”

Este conjunto de datos, disponible en Kaggle, se enfoca en predecir el puntaje crediticio (Credit Score) de los clientes, un indicador esencial en la evaluación de riesgos financieros. Su aplicación principal está dirigida a construir modelos de regresión que estimen este puntaje con base en las características financieras y demográficas del cliente.

El conjunto de datos contiene múltiples variables relacionadas con los ingresos, gastos, hábitos financieros y otros factores determinantes del puntaje crediticio. Es ideal para desarrollar modelos predictivos que respalden la toma de decisiones en instituciones financieras.

La variable objetivo “Credit Score” representa el puntaje crediticio del cliente, utilizado como medida de su capacidad de pago y confiabilidad financiera.

El conjunto de datos incluye variables nivel de deuda, historial de pagos, proporción de utilización del crédito, ingresos, entre otras características relevantes.

Este conjunto de datos fue utilizado exclusivamente para tareas de regresión, enfocándose en la predicción precisa del puntaje crediticio. Los resultados fueron evaluados utilizando el error absoluto medio (MAE).

Anexo 8. Descripción del conjunto de datos “Crypto Solana Memes Coin”

El conjunto de datos *Crypto Solana Memes Coin* contiene información sobre monedas meme basadas en la blockchain de Solana. Este conjunto de datos está diseñado para analizar factores que afectan el mercado de criptomonedas, con un enfoque particular en las monedas memes, que son conocidas por sus movimientos erráticos en el mercado. Proporciona datos clave sobre el comportamiento de las monedas y sus características asociadas.

En este trabajo, el conjunto de datos fue utilizado exclusivamente para tareas de regresión, con el objetivo de predecir la variable “Returns” (rendimientos) de las monedas meme. Esta variable se refiere a la variación porcentual en el precio de la moneda durante un período determinado, lo cual es fundamental para realizar predicciones en el mercado de criptomonedas.

Las principales variables del conjunto de datos son:

- **CoinName:** Nombre de la moneda meme.
- **MarketCap:** Capitalización de mercado de cada moneda.
- **Volume24h:** Volumen de transacciones en las últimas 24 horas.
- **PriceChange24h:** Cambio porcentual del precio en las últimas 24 horas.
- **CirculatingSupply:** Suministro circulante de la moneda.
- **TotalSupply:** Suministro total de la moneda.
- **DateAdded:** Fecha en que la moneda fue listada.

El conjunto de datos cuenta con 2,346 registros y 16 columnas.

Anexo 9. Descripción del conjunto de datos "Electronic Store Dataset"

El conjunto de datos *Electronic Store Dataset* contiene información detallada sobre transacciones realizadas en una tienda electrónica. Cada registro representa una venta individual, con diversas características de los clientes, productos vendidos y el comportamiento de compra. Este conjunto de datos es valioso para predecir la "SatisfactionScore" (puntaje de satisfacción) de los clientes con base en variables como el tipo de producto comprado, el método de pago, la región y otros factores de la transacción.

En este trabajo, el conjunto de datos fue utilizado para tareas de regresión con el objetivo de predecir la variable "SatisfactionScore", que mide la satisfacción del cliente con su compra.

El conjunto de datos cuenta con 5,000 registros y 15 columnas.

Las variables del conjunto de datos son:

- **Age:** Edad del cliente.
- **Items Purchased:** Número de productos comprados en la transacción.
- **Total Spent:** Total gastado por el cliente.
- **Discount:** Descuento aplicado en la transacción.
- **SatisfactionScore:** Puntaje de satisfacción del cliente.
- **Warranty Extension:** Si el cliente extendió la garantía (1: sí, 0: no).
- **Gender:** Género del cliente.
- **Region:** Región donde se realizó la venta.
- **Product Category:** Categoría del producto (por ejemplo, tablet, teléfono, televisión, etc.).
- **Payment Method:** Método de pago utilizado por el cliente (por ejemplo, tarjeta de crédito, efectivo, etc.).
- **Revenue:** Ingresos generados por la venta.
- **Store Rating:** Calificación de la tienda en la transacción.
- **Loyalty Score:** Puntaje de lealtad del cliente.
- **Membership Status:** Estado de membresía del cliente (1: miembro, 0: no miembro).
- **Preferred Visit Time:** Hora preferida de visita (por ejemplo, mañana, tarde, noche).