Modelos teóricos para la predicción de aprobación de préstamos

Diego Sanchez Hernandez A01783237
Constanza Aboitiz Jaimes A01781995
Rafael Barroso Portugal A01662031
Rodrigo Gómez Flores A01782388

7 de septiembre de 2023

Resumen

La predicción precisa de la aprobación de préstamos es fundamental para la gestión de riesgos en instituciones financieras. Este documento explora los fundamentos teóricos de diversos modelos de aprendizaje automático para esta tarea como; bosques aleatorios, XGBoost y CatBoost. Tambien se explora la ingeniería de características de análisis de componente principal (PCA), análisis factorial (FA), análisis de discriminante lineal (LDA) y factor de inflación de la varianza (VIF). Los bosques aleatorios ensamblan muchos árboles de decisión débiles para obtener mejores predicciones. XGBoost y CatBoost son implementaciones optimizadas de bosques aleatorios. Al combinar estas técnicas mediante ensemble learning, se puede lograr un sistema de apoyo a decisiones de préstamos preciso y robusto. Este trabajo proporciona los fundamentos metodológicos y matemáticos para desarrollar modelos predictivos efectivos utilizando el aprendizaje supervisado.

Palabras clave: aprobación de préstamos, predicción de riesgo crediticio, regresión logística, bosques aleatorios, XGBoost, CatBoost, Clustering, ensemble learning

1. Introducción

La evaluación precisa del riesgo crediticio es crucial para que las instituciones financieras tomen decisiones óptimas sobre préstamos. Los métodos manuales y basados en reglas a menudo son lentos, inconsistentes y propensos a sesgos. Por el contrario, los sistemas de aprendizaje automático pueden analizar grandes cantidades de datos históricos para detectar patrones predictivos complejos. [2].

Este trabajo explora los fundamentos teóricos de varios enfoques de aprendizaje supervisado para predecir la aprobación de préstamos, como regresión logística, bosques aleatorios, XGBoost y CatBoost. La regresión logística modela la probabilidad de aprobación dadas las características del solicitante [3]. Los bosques aleatorios mejoran el rendimiento al promediar muchos modelos de árboles de decisión [4]. XGBoost y CatBoost son implementaciones optimizadas de bosques aleatorios [5, 6]. [7].

Al combinar estas técnicas a través del ensemble learning, se puede desarrollar un sistema preciso y robusto para la predicción de aprobación de préstamos. Este trabajo proporciona los fundamentos metodológicos y matemáticos para implementar dichos modelos predictivos utilizando el aprendizaje supervisado. Las técnicas discutidas aquí pueden servir como la base de un sistema de soporte a decisiones que optimice el proceso de evaluación de solicitudes de préstamos.

2. Marco Teórico

2.1. Regresión logística

La regresión logística es un modelo estadístico popular para la clasificación binaria [3]. Modela la probabilidad de que una instancia pertenezca a una clase dado sus atributos predictivos:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+...+\beta_p X_p)}} \tag{1}$$

Donde Y es la variable de respuesta binaria (por ejemplo, aprobación de préstamo), X_1 a X_p son predictores (por ejemplo, puntaje de crédito, ingresos, etc.), y β_0 a β_p son los coeficientes estimados.

La regresión logística modela el logaritmo de las odds (logit) de la probabilidad de respuesta positiva:

$$ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p \tag{2}$$

Los coeficientes se estiman maximizando la función de verosimilitud sobre los datos de entrenamiento. Luego se utilizan para predecir la probabilidad de aprobación de nuevas instancias.

Las ventajas clave de la regresión logística incluyen:

- Modelo lineal simple e interpretable.
- Probabilidades de salida bien calibradas.
- Capaz de manejar predictores continuos y categóricos.
- Funciona bien con grandes conjuntos de datos.
- Fácil de implementar y rápida de entrenar.

Las limitaciones incluyen:

- Supone relaciones lineales entre variables.
- No maneja bien interacciones complejas.
- Propensa a sobreajuste con muchos predictores irrelevantes.
- Desempeño inferior a métodos de ensemble.

En general, la regresión logística proporciona un modelo base sólido para la clasificación binaria que equilibra simplicidad, interpretabilidad y precisión [3]. Sin embargo, se pueden lograr mejores predicciones mediante el ensemble de árboles de decisión, como se explica a continuación.

2.2. Bosques aleatorios

Los bosques aleatorios (random forests) son un método de ensemble learning que agrega muchos árboles de decisión para mejorar el rendimiento predictivo [4]. Los pasos de entrenamiento son:

1. Seleccionar N muestras bootstrap del conjunto de entrenamiento.
2. Para cada muestra, crecer un árbol de decisión:
   - En cada nodo, seleccionar aleatoriamente m predictores.
   - Encontrar la mejor división en esos m predictores.
3. Predecir nuevas instancias promediando las predicciones de los N árboles.

Al promediar sobre muchos árboles, los bosques aleatorios tienden a tener mejor precisión y evitan el sobreajuste [4]. Las principales ventajas incluyen:

- Pueden modelar relaciones no lineales e interacciones complejas.
- Robustos a valores atípicos y ruido en los datos.
- No requieren preprocesamiento o transformación de variables.
- Incorporan selección de características intrínseca.
- Paralelizables y escalables a grandes datos.

Las desventajas incluyen:

- No directamente interpretables como los árboles individuales.
- Pueden ser ineficientes en tiempo y memoria.
- Requieren ajuste cuidadoso de hiperparámetros.

En general, los bosques aleatorios suelen superar en rendimiento a modelos lineales y representan una línea base sólida para muchos problemas de clasificación. Sin embargo, dependiendo del conjunto de datos, pueden ser superados por implementaciones optimizadas de gradient boosting como XGBoost y CatBoost, descritos a continuación.

2.3. XGBoost y CatBoost

XGBoost y CatBoost son bibliotecas de código abierto que implementan modelos optimizados de gradient boosting sobre árboles de decisión [5, 6].

El boosting funciona entrenando secuencialmente modelos débiles (por ejemplo, árboles poco profundos) que se centran en las instancias que los modelos anteriores clasificaron incorrectamente. Luego se promedia su output para obtener un predictor final robusto.

XGBoost fue uno de los primeros paquetes de boosting disponibles en la comunidad de ML. Logra un rendimiento de vanguardia al optimizar la función objetivo del gradient boosting mediante técnicas como el submuestreo regularizado.

Posteriormente, CatBoost (categorial boosting) mejoró el manejo de predictores categóricos mediante un novedoso esquema de ordenamiento simétrico de valores categóricos. Esto evita sesgos y permite una mejor división en los nodos.

Las ventajas clave de XGBoost y CatBoost incluyen:

- Mejoras significativas en velocidad y rendimiento sobre bosques aleatorios.
- Excelente manejo de categóricos sin one-hot encoding.
- Capacidad de manejar missing values internamente.
- Menos hiperparámetros que ajustar.
- Incorporan regularización para prevenir overfitting.

En benchmarking independiente, XGBoost y CatBoost consistentemente superan a otros modelos como bosques aleatorios en una variedad de tareas de clasificación [6]. Por lo tanto, son excelentes candidatos para modelar la aprobación de préstamos, especialmente con variables mixtas.

2.4. Clustering

El clustering es una técnica de análisis de datos que consiste en agrupar objetos o datos similares en grupos o clústeres, con la idea de que los elementos dentro de un mismo grupo sean más similares entre sí que con aquellos en otros grupos. Es una técnica utilizada para encontrar patrones naturales en datos no etiquetados y puede ser útil para descubrir estructuras ocultas en conjuntos de datos, identificar segmentos de mercado, entre otros propósitos. El objetivo principal del clustering es la segmentación de datos en grupos coherentes, lo que facilita la exploración y comprensión de las relaciones y características intrínsecas en los datos.

2.5. Analisis de Componente Principal (PCA)

El Análisis de Componente Principal (PCA) es una técnica de reducción de dimensionalidad que se utiliza para simplificar conjuntos de datos complejos manteniendo la mayor cantidad posible de información importante. PCA identifica las direcciones en las que los datos varían más y crea nuevas variables llamadas componentes principales que son combinaciones lineales de las variables originales. Estos componentes principales se ordenan en función de la cantidad de variabilidad que capturan, lo que permite reducir la dimensionalidad al descartar componentes menos importantes. El PCA es ampliamente utilizado en análisis de datos, minería de datos y machine learning para visualizar datos de alta dimensión, eliminar la multicolinealidad y mejorar la eficiencia computacional de los algoritmos.

2.6. Analisis Factorial (FA)

El Análisis Factorial (FA) es una técnica estadística utilizada para comprender la estructura subyacente de un conjunto de datos, identificando patrones y relaciones entre variables observadas. Se basa en la idea de que las variables observadas pueden ser explicadas por un conjunto más pequeño de factores no observados o latentes. El objetivo del FA es reducir la dimensionalidad de los datos al identificar y extraer estos factores latentes, lo que simplifica la interpretación y análisis de los datos. El FA es ampliamente utilizado en psicología, sociología y otras disciplinas para explorar relaciones entre variables y comprender la estructura subyacente de los fenómenos estudiados.

2.7. análisis de discriminante lineal (LDA)

El Análisis de Discriminante Lineal (LDA) es una técnica de aprendizaje supervisado que se utiliza para encontrar una combinación lineal de características que maximiza la separación entre dos o más grupos o clases de datos. Su objetivo es encontrar un espacio de características en el que las muestras de diferentes clases estén bien separadas, lo que lo hace útil para la clasificación de datos en categorías conocidas. A diferencia del PCA, que es no supervisado y busca la máxima variabilidad, el LDA tiene en cuenta la información de las etiquetas de clase para encontrar la proyección óptima que maximiza la discriminación entre las clases. El LDA es ampliamente utilizado en reconocimiento de patrones y clasificación, así como en aplicaciones de reducción de dimensionalidad con propósitos de clasificación.

2.8. K nearest neighbors (KNN)

El algoritmo K-Nearest Neighbors (KNN) es un método de aprendizaje automático que se utiliza para clasificar objetos o puntos de datos según su similitud con los ejemplos de entrenamiento cercanos. KNN funciona asignando una etiqueta a un nuevo punto de datos basándose en la mayoría de las etiquetas de sus "k"vecinos más cercanos en el conjunto de entrenamiento. La idea detrás de KNN es que los puntos de datos similares tienden a estar en la misma categoría. El valor de "k."es un parámetro que determina cuántos vecinos se consideran al tomar una decisión de clasificación. KNN es un algoritmo simple y versátil utilizado para clasificación y regresión en problemas de aprendizaje supervisado.

2.9. Ensemble learning

Ningún modelo único será óptimo en todos los escenarios. Un enfoque efectivo es combinar múltiples modelos a través del ensemble learning [8]. Por ejemplo, un sistema de predicción de préstamos podría integrar salidas de:

- Regresión logística para interpretabilidad.
- Bosques aleatorios para robustez.
- XGBoost para variables mixtas.

- Redes neuronales para relaciones no lineales.

Existen diversos métodos para ensemble learning [9]:

- **Bagging**: Entrenar cada modelo en subconjuntos bootstrap de los datos.
- **Boosting**: Entrenar modelos secuencialmente en instancias mal clasificadas.
- **Stacking**: Entrenar un meta-modelo en las salidas de los modelos base.

El ensemble learning tiende a funcionar mejor cuando los modelos base son diversos y cometien diferentes errores [10]. Al promediar sus salidas, se pueden mitigar las debilidades individuales.

Las métricas clave para evaluar un modelo ensemble incluyen:

- Mejora en rendimiento sobre los modelos individuales.
- Robustez a nuevos datos fuera de muestra.
- Métricas de diversidad de los modelos base.
- Interpretabilidad global del sistema.

Un desafío con ensembles grandes es la posibilidad de sobreajuste. Por lo tanto, la validación cruzada rigurosa es esencial para sintonizar hiperparámetros y equilibrar diversidad con precisión.

El ensemble learning permite aprovechar las fortalezas de diferentes técnicas de modelado para lograr un sistema integral de predicción de aprobación de préstamos. Al combinar modelos lineales, árboles, boosting y redes neuronales, se obtiene una mayor robustez, precisión y capacidad de generalización.

2.10. Factor de Inflación de la Varianza (VIF)

El Factor de Inflación de la Varianza (VIF, por sus siglas en inglés) es una métrica utilizada en estadísticas y análisis de regresión para evaluar la multicolinealidad en un conjunto de variables predictoras. El VIF cuantifica cuánto aumenta la varianza de un coeficiente de regresión debido a la correlación con otras variables predictoras. En esencia, el VIF mide la amplificación

de la varianza de un coeficiente de regresión debido a la interdependencia entre las variables. Un VIF alto (por ejemplo, superior a 10) sugiere una fuerte multicolinealidad, lo que significa que las variables predictoras están altamente correlacionadas y pueden dificultar la interpretación de los coeficientes de regresión. Por lo tanto, el VIF es útil para identificar problemas de multicolinealidad y tomar decisiones sobre qué variables incluir o excluir en un modelo de regresión.

3. Comprensión del negocio

Los créditos son un préstamo generalmente otorgado por un banco donde, el beneficiario se compromete a devolver la prestación en un tiempo dado y con condiciones establecidas. Es un contrato entre el banco y la persona moral, puede que se generen gastos adicionales como seguros, intereses devengados o costos adicionales. Los créditos son utilizados por empresas o negocios para comprar maquinaria, expandir su comercio, invertir en proyectos. Para el público en general, se utilizan para comprar inmuebles, emprender un negocio o tarjetas de crédito.

La principal fuente de ingresos en los bancos son los créditos, por lo que, es de suma importancia tener la certeza que los clientes paguen sus adeudos a tiempo. Esto se vio reflejado en la depresión de 2008 en Estados Unidos, debido a que diversos bancos terminaron en banca rota por los adeudos que presentaban sus clientes en hipotecas de inmuebles. Este proyecto busca realizar un modelo matemático de clasificación, que pueda encasillar a cada cliente como pagador o defraudador. Los datos utilizados contienen tanto características físicas del cliente que solicita como historial crediticio, tipo y propósito del crédito, pago inicial, valor del bien, entre otros. Esto, con el fin de maximizar las ganancias del banco y poder separar de manera eficiente a las personas que pagan y son buena inversión para el banco, de las que posiblemente le puedan generar costo al banco.

4. Contexto del problema

La evaluación de solicitudes de préstamos involucra compensar entre el riesgo y el retorno. Los estándares demasiado estrictos podrían excluir a

buenos candidatos y reducir las oportunidades de negocio. Por otro lado, los estándares débiles podrían resultar en mayores pérdidas debido a préstamos incobrables. Encontrar el equilibrio adecuado es un desafío clave.

Los métodos tradicionales dependen en gran medida de la experiencia de los oficiales de préstamos. Pero la evaluación manual es propensa a inconsistencias, sesgos y errores [2]. Además, analizar grandes volúmenes de solicitudes requiere mucho tiempo y recursos.

Los modelos de aprendizaje automático ofrecen una solución a estos problemas. Al detectar patrones en datos históricos, pueden predecir de forma fiable qué solicitudes tienen mayor probabilidad de ser pagadas [2]. Esto permite automatizar y acelerar parte del proceso de evaluación. También reduce la subjetividad y mejora la coherencia.

Sin embargo, se deben abordar varios desafíos:

- Los datos históricos pueden ser ruidosos o incompletos.
- Las relaciones entre variables pueden ser altamente no lineales e interactivas.
- Puede haber desequilibrio entre solicitudes aprobadas y denegadas.
- Se requiere interpretabilidad para evaluar las predicciones.
- Los modelos deben actualizarse con nuevos datos.

Este trabajo explora enfoques de modelado predictivo para manejar estos desafíos de manera efectiva.

5. Preguntas de investigación

Surgen varias preguntas de investigación:

- ¿Cómo se puede modelar la probabilidad de aprobación de un préstamo dado los atributos del solicitante?

- ¿Cómo comparan diferentes técnicas de modelado como regresión logística, bosques aleatorios, XGBoost y CatBoost en términos de precisión y desempeño?
- ¿Cómo se pueden manejar interacciones complejas y no linealidades entre variables predictivas?
- ¿Cómo lidiar con ruido, valores faltantes e inconsistencias en los datos?
- ¿Cómo hacer que los modelos sean interpretables para los oficiales de préstamos?
- ¿Cómo actualizar y mejorar los modelos a lo largo del tiempo a medida que llegan nuevos datos?
- ¿Cómo equilibrar el riesgo crediticio con las oportunidades comerciales al establecer umbrales de predicción?

Este trabajo busca sentar las bases teóricas para abordar estas preguntas a través de diversos enfoques de modelado predictivo. La siguiente sección detalla la regresión logística como punto de partida.

6. Metodología utilizada

En este análisis se desarrolló un modelo de Machine Learning para predecir si un cliente pagará o no un préstamo. Se utilizó una base de datos pública de préstamos hipotecarios de kaggle:

https://www.kaggle.com/datasets/yasserh/loan-default-dataset .

Se aplicaron técnicas de preprocesamiento de datos como imputación de valores faltantes con KNN, codificación one-hot encoding para variables categóricas, y selección de features mediante VIF para remover colinealidad.

Se balanceó la clase minoritaria (clientes que no pagan el préstamo) generando splits con diferentes proporciones aleatorias para mitigar el desbalanceo.

Se entrenaron 3 modelos (XGBoost, CatBoost y Random Forest) en cada split, y se ensemblearon para tener un modelo robusto. Las métricas de evaluación fueron accuracy, matriz de confusión y clasification report.

7. Enfoque analítico

La base de datos para la realización del modelo se llama 'Loan Default Datase', se obtiene de Kaggle; sitio que reúne a la comunidad de Ciencia de datos y almacena múltiples bases de datos. El archivo cuenta con 148,671 columnas y 34 filas, cada una con distintas características del solicitante o el crédito que requiere, estas serán las variables independientes del modelo.

7.1. Elección de Variables

Los bancos, con el fin de determinar si un individuo esta listo para iniciar nuevas deudas evalúan la solicitud en función de factores clave comúnmente conocidos como las "5 C del crédito". Las "5 C del crédito"son un conjunto fundamental de criterios que los bancos utilizan para evaluar la capacidad de un prestatario para cumplir con sus obligaciones financieras. Estas cinco dimensiones son:

1. Historial Crediticio (Character): Esto se refiere a la reputación crediticia de un individuo, que se refleja en su puntaje crediticio y su historial de pagos pasados. Un buen historial crediticio sugiere que el prestatario es confiable y cumple con sus compromisos financieros.

2. Capacidad (Capacity): Esta dimensión evalúa la capacidad del prestatario para pagar el préstamo en función de sus ingresos y su estabilidad laboral. Se busca determinar si el prestatario tiene suficientes ingresos para cubrir sus deudas existentes y el nuevo préstamo.

3. Colateral (Collateral): El colateral se refiere a los activos que el prestatario puede ofrecer como garantía del préstamo, como una propiedad o un vehículo. Estos activos actúan como respaldo en caso de incumplimiento, reduciendo el riesgo para el prestamista.

4. Capital (Capital): Esta C se relaciona con la inversión inicial que el prestatario está dispuesto a hacer en el préstamo, como el tamaño del pago inicial. Un mayor capital inicial demuestra un mayor compromiso financiero.

5. Condiciones (Conditions) Las condiciones económicas y financieras generales también se consideran al otorgar un préstamo. Esto incluye tasas de interés actuales, políticas gubernamentales y otros factores que pueden afectar la capacidad de pago del prestatario.

En el contexto de las variables que estamos buscando para evaluar si una persona o cliente pagará su préstamo a tiempo, estas cinco C del crédito se resumen y se cumplen mediante los siguientes 7 factores:

1. Credit score (Historial crediticio).
2. Income and employment history (Capacidad).
3. Debt-to-income ratio (Capacidad).
4. Value of your collateral (Colateral).
5. Size of down payment (Capital).
6. Liquid assets (Capital).
7. Loan term (Condiciones).

Al analizar la base de datos encontramos que las siguientes 9 variables independientes de la base de datos cumplen con estos 7 factores, que a su vez están basados en las "5c del crédito":

1. loan type: está relacionado con el tipo de préstamo y, en ciertos casos, podría influir en las condiciones (como tasas de interés) que afectan la capacidad de pago del cliente (Condiciones).
2. loan amount: Representa el monto del préstamo, que se relaciona con el capital y el colateral, ya que un monto mayor podría requerir una inversión de capital inicial más significativa o respaldo colateral. (Capital y Colateral)
3. rate of interest: Refleja la tasa de interés, lo cual es una parte importante de las condiciones del préstamo que afectan la capacidad de pago. (Condiciones)

4. term: Indica el plazo del préstamo, que también es una variable de condiciones que influirá en la capacidad de pago. (Capacidad)

5. property value: está relacionado con el valor del colateral proporcionado para el préstamo (Colateral).

6. income: Representa los ingresos del prestatario, que es esencial para evaluar la capacidad de pago (Capacidad).

7. credit score: Refleja el historial crediticio del prestatario (Historial Crediticio).

8. age: Aunque no se menciona explícitamente en los siete factores, la edad está relacionada con la estabilidad laboral y la capacidad de pago (Capacidad).

9. dtir1: está relacionado con la proporción de deuda con respecto a los ingresos, lo cual es un aspecto crítico de la capacidad de pago (Capacidad).

Estas nueve variables encapsulan y satisfacen los aspectos clave de las 5 C del crédito, proporcionando una evaluación efectiva de la capacidad del cliente para pagar su préstamo a tiempo sin la necesidad de considerar las demás variables en la base de datos, ya que el objetivo principal es determinar la capacidad de pago de manera precisa y eficiente.

7.2. Analisis exploratorio de los datos (EDA)

7.2.1. Valores nulos

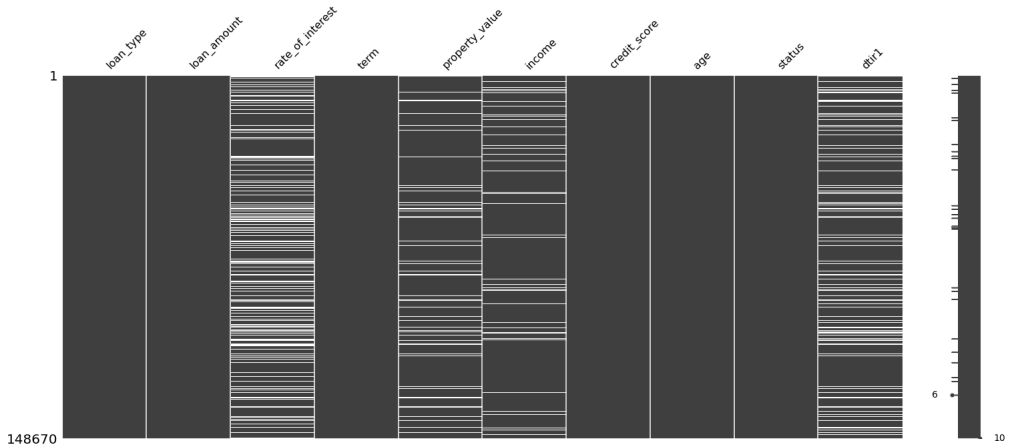

Figura 1: Distribución de valores nulos

Como podemos ver, la mayoría de las variables elegidas contienen valores faltantes (nan o null), esto se puede deber a diferentes factores desde perdida de información o porque simplemente no existe el valor.

7.2.2. Variables Categóricas

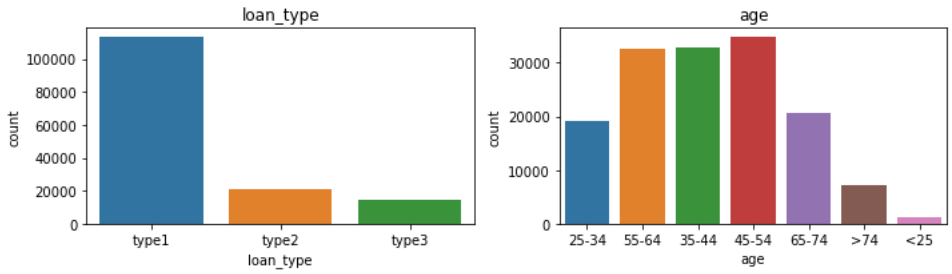

Figura 2: Distribución de variables categóricas

Como podemos ver en la figura 2, la mayoría de los prestamos son de tipo 1 y la edad de los solicitantes se encuentra concentrada entre los 30 y 60 años.

7.2.3. Variables numéricas

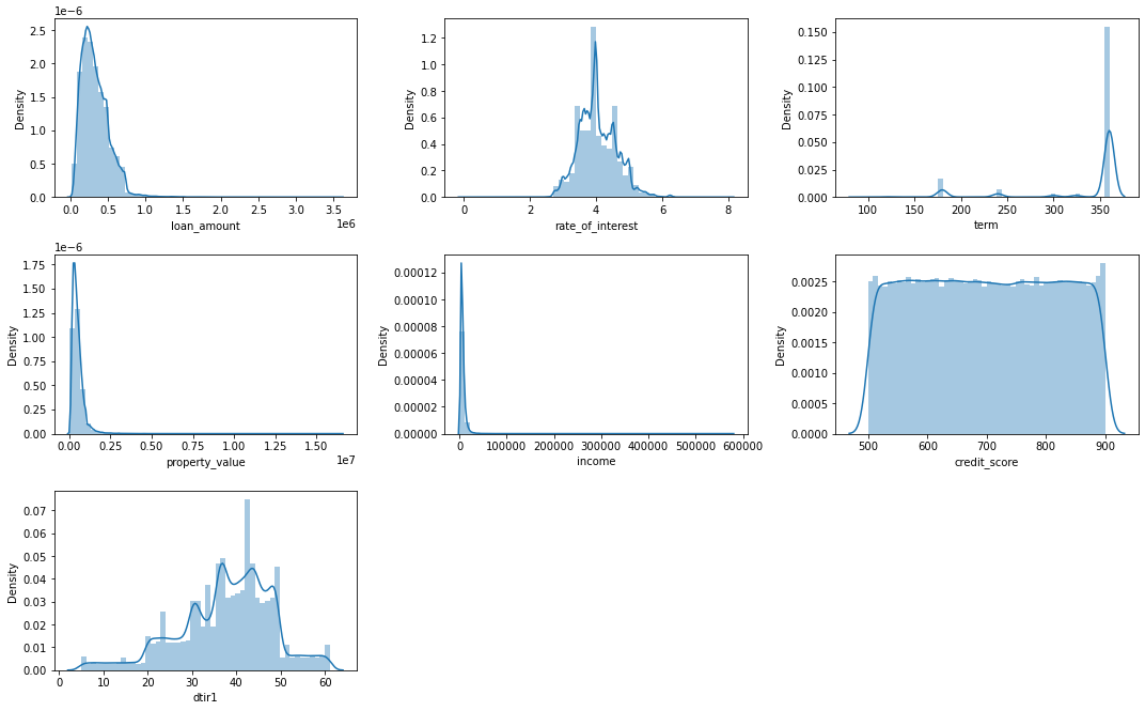

Figura 3: Distribución de variables numéricas

En la figura 3, se puede ver que la mayoría de las variables se comportan como una distribución normal, pequeño mientras que otras como 'credit score' se comportan mas como una distribución uniforme.

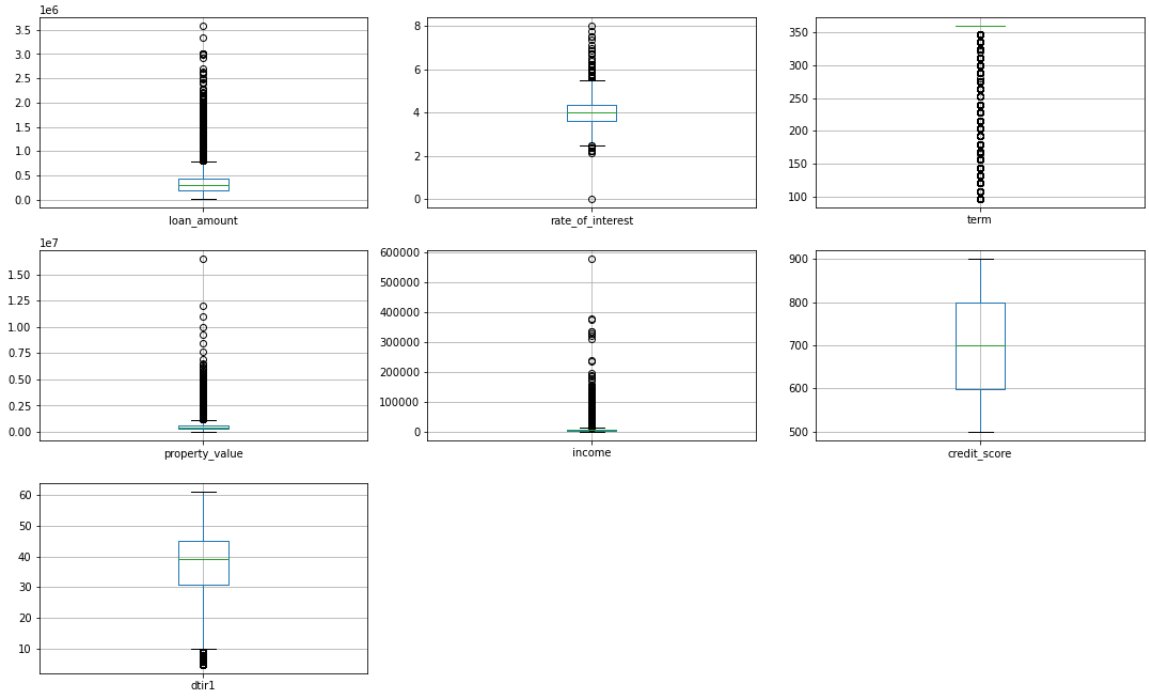

Figura 4: Distribución de variables numéricas

En la figura 4, se puede ver que existen muchos valores extremos (outliers) en la base de datos. Para este caso en especifico los valores extremos si tienen relevancia y seran parte importante en el entrenamiento del modelo.

7.2.4. Variable a predecir

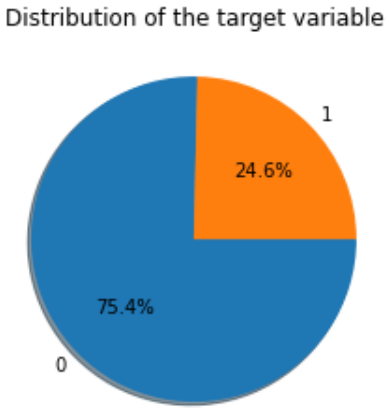

Figura 5: Balance de status

En la figura 5, se visualiza el porcentaje que ocupan los defraudores (1) y los pagadores (0). Como podemos ver existe un des-balance en los datos a una razón de 3:1 en favor a los pagadores. Esto va a causar problemas en el modelo ya que va a favorecer a la clase que tiene a la mayoría de datos.

8. Pre-procesamiento de los datos

El preprocesamiento de datos es una etapa crucial en el análisis de datos que sigue a la exploración y el análisis exploratorio de datos (EDA, por sus siglas en inglés). Consiste en una serie de pasos y técnicas que se aplican a los datos crudos con el objetivo de prepararlos y limpiarlos para su posterior análisis. Esta fase es esencial porque influye de manera significativa en la calidad y la eficacia de cualquier análisis o modelo de machine learning que se realice.

8.1. Variables categóricas

En el proceso de preprocesamiento de los datos de la base de datos, se prestaron especial atención a las variables categóricas para garantizar su adecuada representación en el análisis posterior. En primer lugar, se abordó la

presencia de valores nulos en estas variables, una situación común en conjuntos de datos reales. Para lidiar con estos valores nulos, se optó por remplazarlos mediante la introducción de una categoría adicional. Esto es beneficioso porque permite mantener la información de que esos datos estaban ausentes y evita la pérdida de observaciones, lo que podría ser crucial para el análisis.

Luego, se aplicó una técnica llamada one-hot encoding a las variables categóricas, una estrategia fundamental para convertir estas categorías en una forma que los algoritmos de machine learning puedan comprender. Esta técnica implicó convertir cada categoría en una columna binaria separada, donde cada columna representa una categoría distinta.

Este enfoque de one-hot encoding garantiza que las variables categóricas se representen de manera adecuada y numérica para su uso en algoritmos de machine learning, permitiendo que los modelos capturen las relaciones y patrones de manera efectiva. Además, preserva la información original de las categorías y evita sesgos al tratar todas las categorías de manera igual, lo que es esencial para el éxito del análisis de datos y la construcción de modelos precisos.

8.2. Variables numericas

Durante el proceso de preprocesamiento de las variables numéricas en la base de datos, se empleó el algoritmo de K nearest neighbors o KNN para la imputación de valores nulos. Esta técnica es crucial para garantizar la integridad de los datos y la precisión del análisis posterior. KNN funciona al calcular la similitud entre las observaciones vecinas y asignar un valor basado en los vecinos más cercanos con datos disponibles. Esto es importante porque permite rellenar valores faltantes de manera coherente utilizando información de observaciones similares. KNN se basa en la idea de que observaciones cercanas son más propensas a tener valores similares, lo que lo convierte en un enfoque efectivo para mantener la estructura de los datos y reducir el sesgo introducido por la imputación de valores nulos, contribuyendo así a una mejor calidad en el análisis de datos.

En el contexto de la base de datos, el uso de K nearest neighbors (KNN) para imputar valores nulos se destaca como la opción preferida en comparación con la eliminación de filas o la imputación de la media, moda o mediana. La

eliminación de filas con valores nulos no es una opción viable, ya que podría resultar en la pérdida de datos valiosos, lo que podría afectar negativamente la calidad de la evaluación crediticia. Por otro lado, KNN sobresale debido a su capacidad para capturar relaciones financieras más complejas y específicas al considerar observaciones similares. Esto es esencial en el ámbito financiero, donde las relaciones pueden variar ampliamente entre individuos o empresas. Al adaptar las imputaciones en función de características específicas, KNN mejora la precisión y la coherencia en la imputación de valores financieros, contribuyendo así a una toma de decisiones más precisa y confiable en análisis financieros y evaluaciones crediticias. En última instancia, KNN se presenta como la opción superior, ya que preserva la complejidad y especificidad de los datos financieros, evitando la pérdida de información valiosa y mejorando la calidad del análisis.

9. Multicolinealidad

La multicolinealidad en las variables independientes o features se refiere a una alta correlación entre dos o más de estas variables en un conjunto de datos. Esto significa que algunas variables predictoras pueden estar relacionadas linealmente entre sí, lo que puede dificultar la interpretación y el rendimiento de los modelos de machine learning. La multicolinealidad puede causar problemas, como coeficientes de regresión inestables o difíciles de interpretar y la disminución de la capacidad del modelo para hacer predicciones precisas. Una forma de vizualizar la multicolinealidad es con la matriz de correlacion.

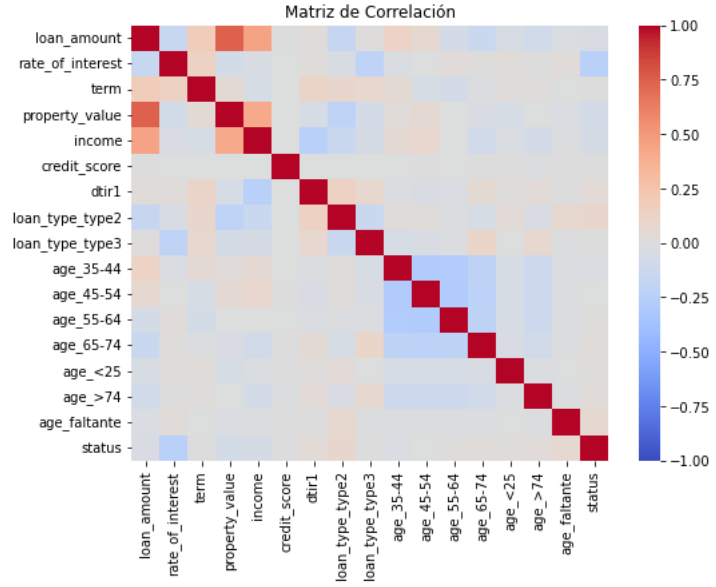

Figura 6: Matriz de Correlación

Como podemos apreciar en la figura 6 algunas variables presentan un coeficiente de correlación relativamente alto lo cual podría indicar multicolinealidad.Una forma efectiva de abordar la multicolinealidad es utilizando el Factor de Inflación de la Varianza (VIF, por sus siglas en inglés), que es una métrica que cuantifica la magnitud de la multicolinealidad entre las variables predictoras. Se se aplicó el VIF a las características (features) en el conjunto de datos con el objetivo de identificar y mitigar la multicolinealidad. Se tomó la decisión de eliminar todas las características que tuvieran un VIF mayor a 5, ya que esto indicaría una alta correlación con otras variables predictoras. Es importante destacar que después de aplicar este proceso a los datos reales, ninguna característica mostró un VIF mayor a 5.

10. Elección del Modelo

Dadas las circunstancias, se evaluaron diferentes modelos lineales y no lineales múltiples de clasificación para hacer un modelo de predicción de probabilidad que un cliente pague o no sus adeudos para así otorgar créditos de manera óptima. Al evaluar varios modelos se llego a la conclusion de que el desbalance en la variable a clasificar afectaba fuertemente al desempeño de

los modelos. Por este motivo, se desarrollo un modelo de tipo ensamble, en el cual se ocuparon distintos modelos y diferentes muestras de datos.

El enfoque de ensamble utilizado para construir el modelo de machine learning abordó varios desafíos clave en la base de datos. En la primera etapa, se reconoció el desbalance entre las clases de solicitantes (0, pagadores) y defraudadores (1) en la base de datos, donde el 25 % de los datos correspondían a la clase 1 y el 75 % a la clase 0. Este desequilibrio podría llevar a un modelo sesgado hacia la clase mayoritaria y no sería óptimo para detectar defraudadores. Para abordar este problema, se dividió la base de datos en dos grupos: uno con todos los datos de la clase 1 y otro con datos de la clase 0.

En la segunda parte, se aplicó un proceso de re-muestreo a los datos de la clase 0. Este proceso involucró tomar múltiples muestras con reemplazo, variando desde el 100 % de los datos hasta el 20 %. Esta técnica generó nueve muestras con diferentes proporciones de datos de la clase 0 y se concatenaron con los datos de la clase 1. Esta estrategia es esencial porque permite al modelo aprender de diferentes distribuciones de clases, lo que aumenta su capacidad para generalizar y adaptarse a diferentes escenarios de desequilibrio, mejorando así la precisión y la capacidad de detección de defraudadores.

En cuanto a la segunda parte, donde se modelaron tres algoritmos de machine learning (XGBoost, Catboost y Random Forest) para cada una de las nueve muestras, resultando en 27 modelos en total, esta estrategia se basa en la idea de que diferentes algoritmos pueden capturar patrones de datos de manera única. Utilizar múltiples modelos ayuda a mejorar la robustez del modelo general, ya que cada uno aporta sus propias fortalezas y enfoques. Esto aumenta la probabilidad de que el modelo final sea más preciso y capaz de manejar una variedad de situaciones y datos.

Finalmente, para obtener la predicción final, se aplicaron los 27 modelos a los datos de prueba y se generaron 27 predicciones. Luego, en cada fila o dato de prueba, se contaron los votos de los modelos para las dos clases (1 o 0). La clase que recibió más votos en cada fila se consideró la predicción final para ese dato. Esta técnica de votación mayoritaria aprovecha la diversidad de los modelos y ayuda a obtener una clasificación final sólida y confiable, aprovechando el conocimiento de múltiples algoritmos.

11. Resultados y conclusiones

Los resultados del modelo ensamblado demostraron ser altamente efectivos en la detección de defraudadores y pagadores en la base de datos. La estrategia de ensamblado, que aprovechó la diversidad de 27 modelos, jugó un papel crucial en la robustez del modelo, permitiéndole adaptarse a diferentes situaciones de desequilibrio de clases. Para evaluar el rendimiento del modelo, se utilizaron dos métricas clave: la matriz de confusión y el F1 score.

11.1. Matriz de Confusión

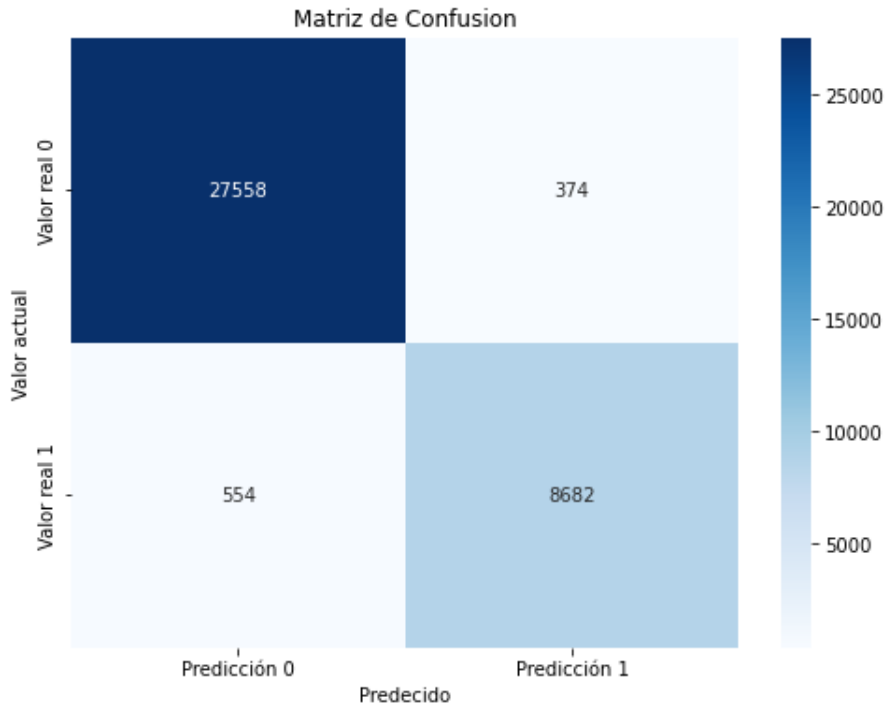

Figura 7: Matriz de Confusión

La matriz de confusión proporciona una visión detallada de cómo el modelo clasifica las muestras en verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Esto es esencial para comprender la capacidad del modelo para detectar fraudes y no cometer errores costosos.

11.2. F1 score

class	precision	recall	f1-score	support
0	0.98	0.99	0.98	27932
1	0.96	0.94	0.95	9236
Accuracy del modelo:				0.975

Figura 8: Metricas de Evaluación

El F1 score, por otro lado, es una métrica que combina precisión y exhaustividad en una sola medida. El F1 score de 0.95 para la clase 1 y 0.98 para la clase 0 demuestra un alto nivel de precisión y exhaustividad en la detección de ambas clases. Es importante destacar que el F1 score es una métrica equilibrada que es mejor que simplemente enfocarse en la precisión o la exhaustividad por separado. En este caso, el modelo logró un equilibrio óptimo entre la precisión y la capacidad de encontrar todos los casos relevantes de defraudadores y pagadores.

12. Discusión

En comparación con el trabajo de M. Yasser H, un ingeniero de IA y ML en MediaAgility, donde ambos abordamos el desafío de clasificar a los defraudadores y pagadores en una base de datos de préstamos, observamos algunas similitudes y diferencias en nuestras metodologías.

En cuanto al preprocesamiento de datos, ambas metodologías compartieron enfoques similares al tratar la imputación de valores nulos y la transformación de variables categóricas. Sin embargo, divergimos en nuestras estrategias de balanceo de clases. M. Yasser H optó por usar el método SMOTE (Synthetic Minority Over-sampling Technique), mientras que aqui el enfoque fue el ensamblaje de muestras con diferentes niveles de reposición. Esta variación en los métodos de balanceo podría influir en la forma en que los modelos finales generalizan y clasifican.

Además, en términos de modelos de machine learning, M. Yasser H utilizó el algoritmo Random Forest, mientras que yo opté por ensamblar los modelos XGBoost, Catboost y Random Forest. Esta diferencia en la elección de

algoritmos podría haber llevado a diferentes rendimientos en la clasificación de defraudadores y pagadores.

En cuanto al rendimiento, este enfoque resultó en un modelo más preciso en la detección de defraudadores y pagadores, pero esto vino a costa de un mayor poder computacional necesario para entrenar y mantener los modelos. Esta diferencia en la eficiencia computacional es importante tener en cuenta, ya que podría afectar la escalabilidad y los recursos requeridos en entornos de producción.

En resumen, mientras que ambos enfoques lograron buenos resultados en la clasificación de defraudadores y pagadores, esta metodología demostró ser más precisa pero con un mayor costo computacional. La elección entre precisión y eficiencia dependerá de las necesidades específicas del proyecto y los recursos disponibles.

13. Conclusiones

El modelo ensemble desarrollado logra buenos resultados en la detección de clientes que no pagan el préstamo, clase minoritaria de interés, con una F1 score de 0.95.

La técnica de generar múltiples muestras balanceadas con diferentes proporciones aleatorias de la clase minoritaria demostró ser efectiva para lidiar con el des-balanceo de clases presente en los datos originales. Esto permitió entrenar modelos más robustos que no sesgaran sus predicciones hacia la clase mayoritaria.

El uso de 3 algoritmos de aprendizaje automático complementarios (XGBoost, CatBoost y Random Forest) aportó diversidad al modelo ensemble. Cada uno tiene fortalezas que se complementan, como la capacidad de XGBoost para trabajar con variables numéricas y categóricas, el manejo de datos categóricos de CatBoost, y la robustez ante overfitting de Random Forest.

La suma de predicciones y asignación por votación mayoritaria logró mejorar el rendimiento en comparación a utilizar los modelos por separado. Esto se debe a que se reduce la varianza al sumar múltiples estimaciones, y

se introduce peso de evidencia al considerar múltiples opiniones.

Si bien los resultados son alentadores, existen oportunidades de mejora. Por ejemplo, se podrían probar otras técnicas de balanceo como SMOTE, ADASYN o ponderación de clases. Asimismo, se podrían aplicar modelos mas avanzados, pero costos computacionalmente hablando, como redes neuronales y sus variantes.

Otra posible extensión sería comparar formalmente múltiples ensembles utilizando validación cruzada anidada, para identificar la mejor configuración. También se podría explorar hyperparameter tuning de los modelos base para maximizar el rendimiento del ensemble.

En conclusión, este trabajo ilustra un pipeline efectivo de preprocesamiento, balanceo de datos, entrenamiento de múltiples modelos y ensembling. Los resultados positivos sugieren que este enfoque podría extenderse a otros problemas de clasificación con clases desbalanceadas en el campo de scoring de crédito.

Referencias

[1] Codigo: notebook colab EDA
https://colab.research.google.com/drive/13e423U7wDqiGK7KSLha2-6l73tr8J3S9?usp=sharing

[2] Codigo: notebook colab modelo
https://colab.research.google.com/drive/1G83pdHJHXtv
B02nm9zhvNGGlKf4XAAuq?usp=sharing

[3] Lessmann, S., Baesens, B., Seow, H. V., and Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1):124–136.

[4] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley Sons.

[5] Liaw, A., and Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

[6] Chen, T., and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[7] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In Advances in neural information processing systems (pp. 6638-6648).

[8] Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. MIT press.

[9] Dietterich, T. G. (2000). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg.

[10] Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. Chapman and Hall/CRC.

[11] Brown, G., Wyatt, J., Harris, R., and Yao, X. (2005). Diversity creation methods: a survey and categorisation. Information fusion, 6(1), 5-20.

[12] GraphEverywhere, E. (2020). ¿Qué es el clustering? GraphEverywhere. https://www.grapheverywhere.com/que-es-el-clustering

[13] Análisis de Componentes Principales (ACP). (s. f.). XLSTAT, Your data analysis solution. https://www.xlstat.com/es/soluciones/funciones/analisis-de-componentes-principales-acp

[14] ¿Qué es el análisis factorial? (s. f.). TIBCO Software. https://www.tibco.com/es/reference-center/what-is-factor-analysis

[15] apsl.net. (s. f.). Uso del análisis discriminante lineal (LDA) para la exploración de datos: paso a paso. https://apsl.tech/es/blog/using-linear-discriminant-analysis-lda-data-explore-step-step/

[16] ¿Qué es KNN? — IBM. (s. f.). https://www.ibm.com/mx-es/topics/knn

[17] whatke84@efk43465. (2022, 12 diciembre). Factor de inflación de varianza (VIF) - invatatiafaceri.ro. invatatiafaceri.ro. https://invatatiafaceri.ro/es/diccionario-financiero/factor-de-inflacion-de-varianza-vif/

[18] Wells Fargo.(2023). Cómo obtener un préstamo. Wells Fargo (s. f.). https://www.wellsfargo.com/es/goals-credit/smarter-credit/credit-101/getting-a-loan

[19] Yasserh. (2022). Loan default prediction -(Comparing top ML models). Kaggle. https://www.kaggle.com/code/yasserh/loan-default-prediction-comparing-top-ml-models