Aplicación de Modelos de Aprendizaje Automático en Clasificación Binaria de Créditos

Diego Adrian Cruz Martinez
Maestría en Ciencia de Datos

13/11/2023

---

1 Introducción

El presente articulo aborda el análisis de datos originados de LendingClub, un líder en el sector de préstamos entre pares en los Estados Unidos y reconocida como una de las plataformas más grandes a nivel mundial en este ámbito. Con sede en San Francisco, California, LendingClub se ha establecido como un pilar fundamental en el mercado financiero, proporcionando un contexto amplio para el estudio de patrones en el comportamiento de préstamos y créditos.

El conjunto de datos utilizado, cubre un periodo desde el comienzo de 2008 hasta el cierre de 2018, englobando 2,260,701 registros, con un total de 151 variables. Dicha compilación de datos ofrece una perspectiva amplia y detallada sobre las tendencias en los préstamos otorgados.

El objetivo principal de la investigación es el desarrollo de un modelo predictivo que pueda clasificar si un préstamo resultara "charge off", situación en la que los deudores no cumplen con la devolución del préstamo. Dado que los préstamos no reembolsados constituyen un desafío significativo para las entidades financieras, al representar fuentes considerables de pérdidas. Por lo tanto, se busca crear un modelo eficiente para predecir la capacidad de pago de los solicitantes de préstamos. Este estudio se propone aportar una herramienta para la gestión de riesgos y optimización de estrategias de crédito en prestamos personales.

2 Marco teórico

El aprendizaje automático (AA) se enfoca en el desarrollo de sistemas capaces de aprender y mejorar a partir de la experiencia, sin necesidad de ser programados de manera explícita. En el ámbito de la clasificación, el AA se emplea para categorizar datos en distintos grupos o clases, basándose en patrones y características inherentes a los datos.

Este campo se divide en tres categorías principales: aprendizaje supervisado, no supervisado y por refuerzo. Dentro de estas, el aprendizaje supervisado es el enfoque más común en tareas de clasificación. Aquí, un modelo se entrena utilizando datos previamente etiquetados, con el objetivo de predecir las etiquetas o categorías de nuevos conjuntos de datos.

En el contexto de la evaluación del riesgo de crédito, los métodos de clasificación desempeñan un papel crucial. Los modelos predictivos basados en AA permiten a las instituciones financieras identificar con mayor precisión aquellos préstamos que presentan un alto riesgo de incumplimiento. Esta capacidad de predicción no solo contribuye a minimizar las pérdidas derivadas de créditos no reembolsados, sino que también optimiza la asignación de recursos y mejora la gestión de carteras de préstamos. Estos avances representan un cambio significativo en la manera en que las entidades financieras abordan la gestión del riesgo crediticio.

2.1 Modelos de clasificación en aprendizaje supervisado

Los **árboles de decisión** son modelos predictivos que utilizan un conjunto de reglas binarias para calcular una decisión. Matemáticamente, un árbol de decisión se puede describir como una serie de funciones de decisión, que se aplican secuencialmente para llegar a una predicción.

La estructura de un árbol de decisión consta de nodos y ramas. Cada nodo representa una característica del conjunto de datos y cada rama representa una decisión que lleva al siguiente nodo. Un nodo que no tiene ramas descendentes se llama nodo hoja y representa una decisión final o una predicción.

La decisión en cada nodo se basa en la selección de la característica y un umbral que mejor divide el conjunto de datos según un criterio de pureza, como la entropía en el caso de la clasificación (para árboles ID3 o C4.5) o el error cuadrático medio en el caso de la regresión (para árboles de regresión). La entropía, por ejemplo, se calcula con la fórmula:

$$H(T) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

donde $H(T)$ es la entropía del conjunto T, p_i es la proporción de la clase i en el conjunto T, y la suma se realiza sobre todas las clases. El objetivo es maximizar la ganancia de información, que se mide como la disminución de la entropía.

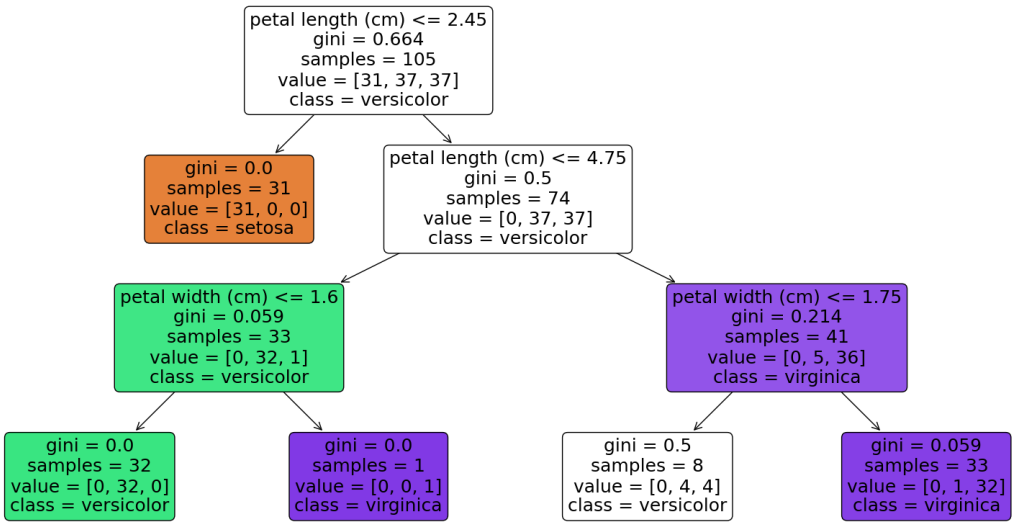


Figura 1: Ejemplo árbol de decisión

La **regresión logística** es un modelo estadístico que, a pesar de su nombre, se utiliza para tareas de clasificación binaria. Matemáticamente, la regresión logística modela la probabilidad de que una instancia pertenezca a una clase en particular (usualmente denotada como clase "1") en función de una o más variables independientes.

La fórmula de la regresión logística es:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

donde:

- $P(Y = 1)$ es la probabilidad de que la instancia pertenezca a la clase 1.
- e es la base del logaritmo natural.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ son los coeficientes del modelo, determinados durante el entrenamiento.
- $X_1, X_2, \ldots, X_n$ son las variables independientes.

El lado derecho de la ecuación es una función logística (o sigmoide) que toma una combinación lineal de las variables independientes y produce un valor entre 0 y 1, interpretado como una probabilidad.

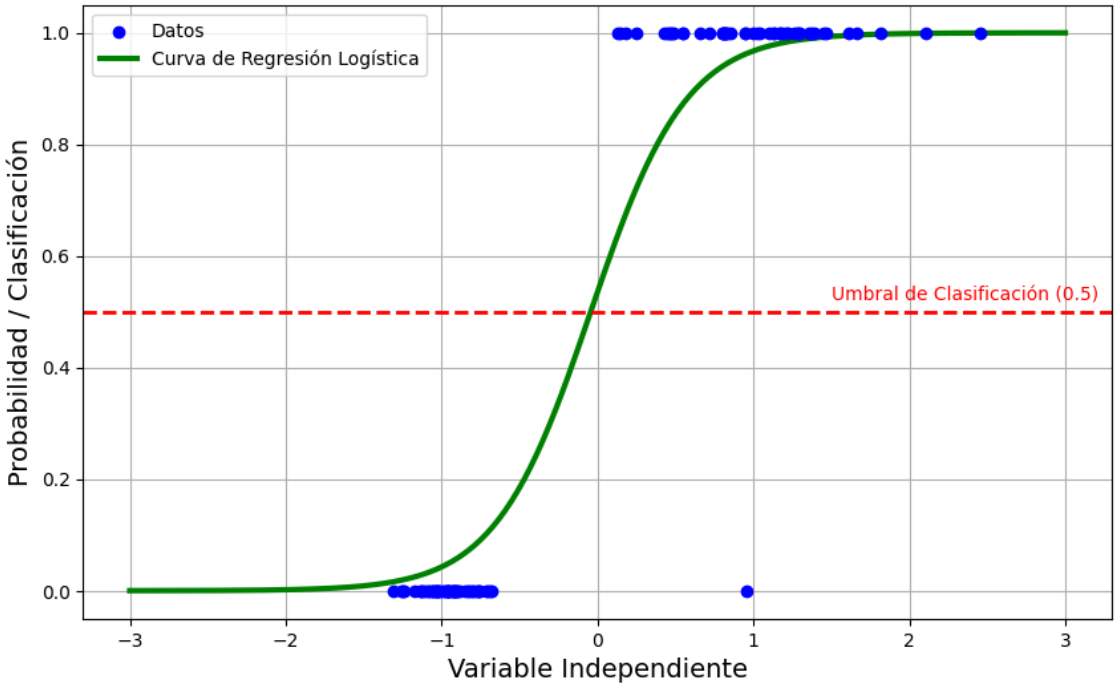


Figura 2: Ejemplo de regresión logística

Las **Máquinas de Soporte Vectorial** (SVM, por sus siglas en inglés) son un conjunto de métodos de aprendizaje supervisado utilizados para la clasificación y la regresión. En el contexto de la clasificación, el objetivo de una SVM es encontrar un hiperplano en un espacio de N dimensiones (N es el número de características) que clasifique claramente las clases de datos.

Un hiperplano se puede describir con la ecuación:

$$\mathbf{w} \cdot \mathbf{x} - b = 0$$

donde

- $\mathbf{w}$ es un vector normal al hiperplano.
- $\mathbf{x}$ es un vector de características.
- b es el sesgo.

La idea es elegir los parámetros $\mathbf{w}$ y b de tal manera que el hiperplano tenga la máxima distancia (margen) a los puntos de datos más cercanos de cada clase, que son conocidos como vectores de soporte.

En problemas de clasificación no lineales, las SVM utilizan lo que se conoce como el "truco del kernel" para transformar el espacio de características a una dimensión superior donde es posible realizar una separación lineal.

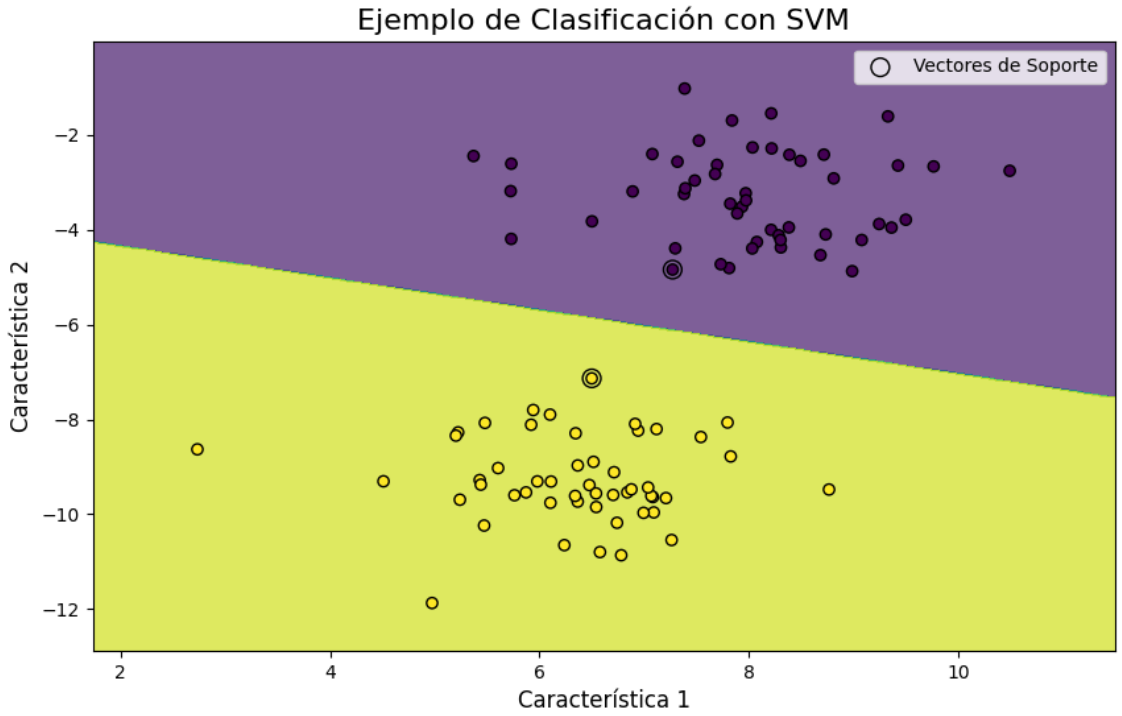


Figura 3: Ejemplo SVM

Las **redes neuronales artificiales** están inspiradas en las redes neuronales biológicas del cerebro humano. Consisten en una serie de capas de nodos, también conocidos como neuronas, interconectados a través de enlaces ponderados. Estas estructuras son particularmente eficaces para modelar relaciones complejas y no lineales en los datos.

Una red neuronal típica incluye una capa de entrada, una o más capas ocultas y una capa de salida. Cada neurona en una capa está conectada a varias neuronas en la siguiente capa a través de pesos, que se ajustan durante el entrenamiento del modelo. La información se mueve a través de la red, desde la entrada hasta la salida, transformándose en cada capa mediante una función de activación.

Las redes neuronales se utilizan en una amplia gama de aplicaciones, desde la clasificación y regresión hasta el procesamiento de lenguaje natural y la visión por computadora.

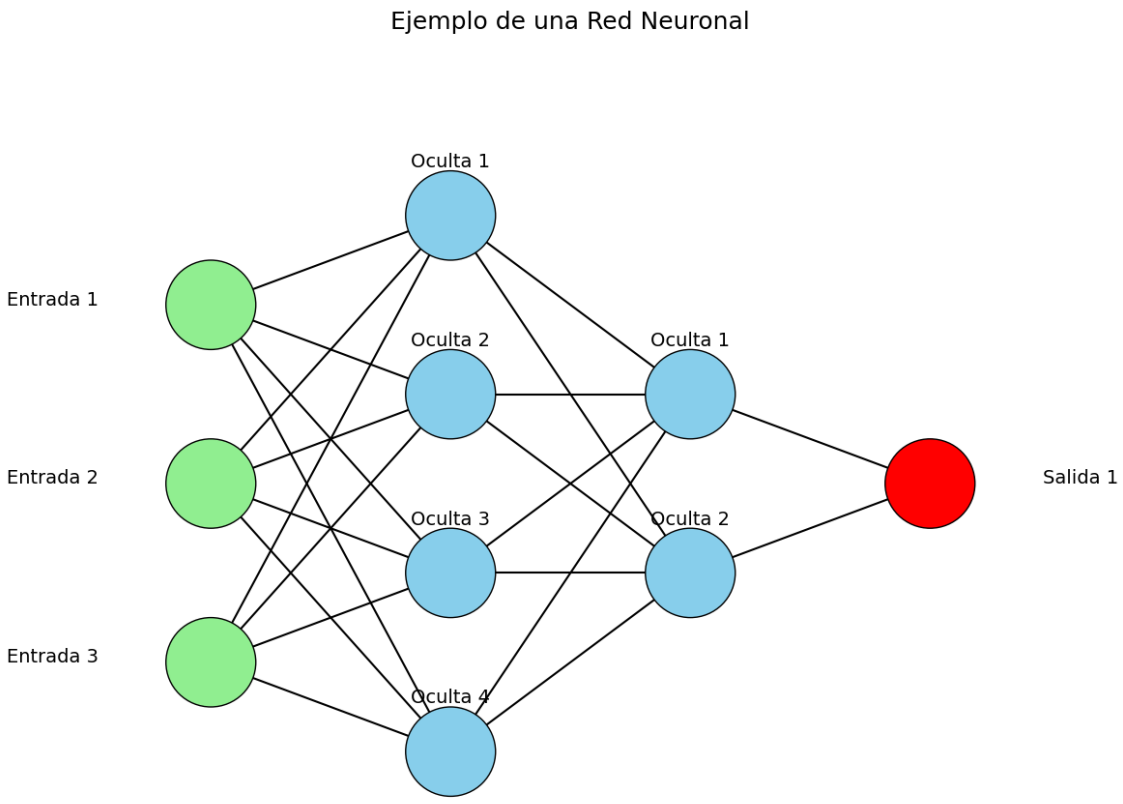


Figura 4: Ejemplo red nueronal

Métodos de Ensamblaje: Técnicas como Bagging, Boosting y Random Forest combinan múltiples modelos para mejorar la robustez y precisión.

Una de los métodos mas utilizados es el **eXtreme Gradient Boosting**, comúnmente conocido como XGBoost, ha ganado una relevancia significativa en tiempos recientes. Esta técnica de ensamblaje basada en arboles de decisión, es especialmente reconocida por su alta eficiencia y rendimiento en una amplia variedad de problemas de aprendizaje automático. XGBoost se destaca por su capacidad de manejar grandes volúmenes de datos y por su flexibilidad, permitiendo su aplicación en tareas de clasificación y regresión, entre otras.

XGBoost es un algoritmo de aprendizaje supervisado que se implementa como un conjunto optimizado de árboles de decisión. Es una extensión del algoritmo Gradient Boosting Machine (GBM), diseñada para mejorar la velocidad y la eficiencia del modelo.

Para ajustar un dataset de entrenamiento utilizando XGBoost, se realiza una predicción inicial. Los residuales se calculan en función del valor predicho y de los valores observados. Se crea un árbol de decisión con los residuales utilizando una puntuación de similitud de los residuales. Se calcula la similitud de los datos de una hoja, así como la ganancia de similitud

de la división posterior. Se comparan las ganancias para determinar una entidad y un umbral para un nodo. El valor de salida de cada hoja también se calcula mediante los residuales. Para la clasificación, los valores se calculan generalmente utilizando el registro de momios y probabilidades. La salida del árbol se convierte en el nuevo residual para el dataset, que se utiliza para construir otro árbol. Este proceso se repite hasta que los residuales dejan de reducirse, o bien el número de veces especificado. Cada árbol subsiguiente aprende a partir de los árboles anteriores y no tiene asignado el mismo peso.

En general XGBoost mejora los árboles de decisión al utilizar el principio de "boosting", un enfoque de aprendizaje de conjunto donde nuevos modelos se añaden para corregir los errores cometidos por los modelos existentes.

2.2 Métricas de error

En el campo de la clasificación binaria existen múltiples métricas de error. Con base en [1] la prueba KS puede utilizarse para evaluar modelos basados en la separación de la función de distribución respectiva de cada clase, en el contexto de la calificación crediticia, el porcentaje de solicitantes "buenos" se maximiza mientras que el porcentaje de solicitantes "malos" se minimiza, sin tener en cuenta los costes relativos. Cuando no se dispone de los costes relativos de la información, la curva ROC y la medida AUC representan el método de evaluación más adecuado.

Dentro del enfoque de riesgo de crediticio nos centraremos específicamente en el indice KS. que indica si modelo es eficaz en minimizar el riesgo de crédito al rechazar una alta proporción de solicitantes de alto riesgo, al tiempo que aprueba la mayoría de los solicitantes de bajo riesgo. Esta capacidad de discriminación es crucial en la gestión de riesgos de crédito, donde la identificación precisa de los solicitantes de alto riesgo puede tener un impacto sustancial en la rentabilidad y la sostenibilidad de las operaciones de crédito.

Índice KS (Kolmogorov-Smirnov)

El índice KS es una medida estadística que cuantifica la distancia entre las funciones de distribución acumulativa de dos grupos. En el contexto de la clasificación binaria, se utiliza para determinar qué tan bien el modelo distingue entre las dos clases.

Se calcula como la máxima distancia entre las curvas de distribución acumulativa de las dos clases. Matemáticamente, se expresa como:

$$\text{KS} = \max_t |FPR(t) - TPR(t)|$$

donde t es el umbral de decisión que varía.

[1] Nargundkar & Priestley, 2003, "Assessment of Model Development Techniques and Evaluation Methods for Binary Classification in the Credit Industry" [1]

En términos prácticos, el KS es una medida de cuán separadas están las distribuciones de las probabilidades predichas para las dos clases (por ejemplo, préstamos que resultarán en "charge off" frente a los que no). Un valor alto de KS indica que el modelo es bueno para distinguir entre las dos clases.

El KS es 1 si las puntuaciones dividen la población en dos grupos separados en los que un grupo contiene todos los positivos y el otro todos los negativos.Por otro lado, si el modelo no puede diferenciar entre positivos y negativos, entonces es como si el modelo seleccionara los casos al azar de la población. El KS sería 0. En la mayoría de los modelos de clasificación, el KS estará comprendido entre 0 y 1, y cuanto mayor sea el valor, mejor será el modelo para separar los casos positivos de los negativos.

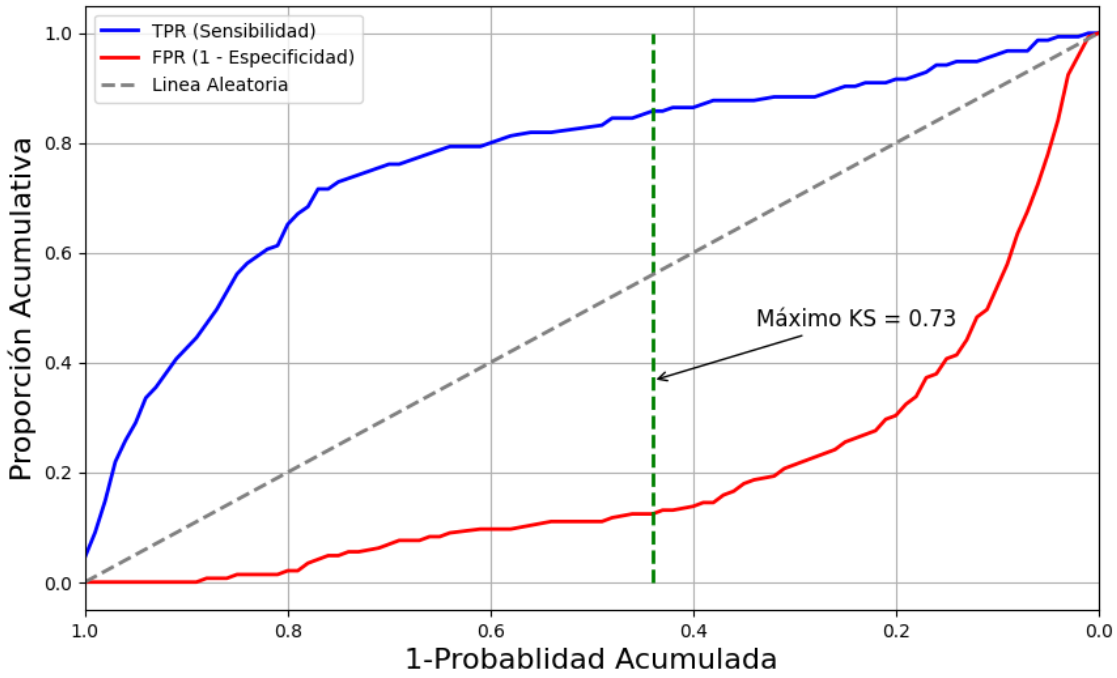


Figura 5: Ejemplo índice KS

Curva AUC-ROC

ROC (Receiver Operating Characteristic): Es un gráfico que muestra el rendimiento de un modelo de clasificación en todos los umbrales de clasificación.Este gráfico traza dos parámetros:

- Tasa de Verdaderos Positivos (TPR): También conocida como sensibilidad o recall. Se calcula como TPR = $\frac{TP}{TP+FN}$, donde TP son los verdaderos positivos y FN los falsos negativos.
- Tasa de Falsos Positivos (FPR): Se calcula como FPR = $\frac{FP}{FP+TN}$, donde FP son los falsos positivos y TN los verdaderos negativos.

AUC (Area Under the Curve)Es el área bajo la curva ROC. Un modelo con un AUC de 1.0 es un modelo perfecto, mientras que un AUC de 0.5 indica un rendimiento no mejor que el azar.

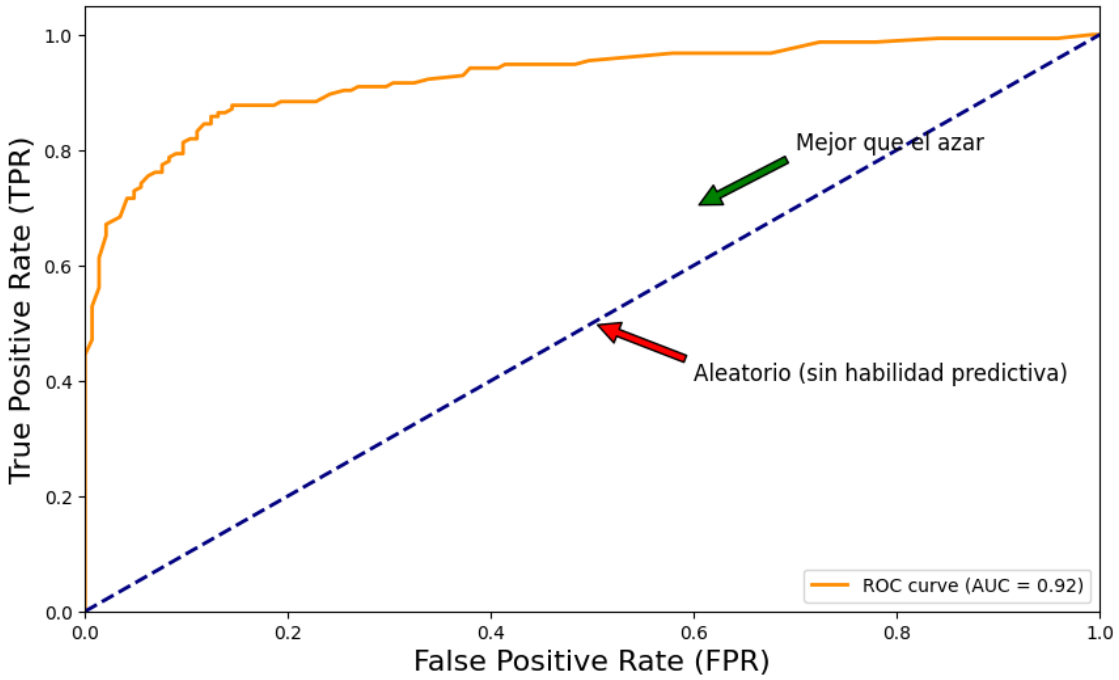


Figura 6: Ejemplo curva AUC-RUC

2.3 Métodos de remuestreo en el contexto del desbalance de clases

El desbalance de clases es una problemática común en conjuntos de datos, especialmente en problemas de clasificación, donde una clase tiene una presencia significativamente menor que la otra. Esta disparidad puede afectar negativamente la capacidad de los modelos de aprendizaje automático para generalizar y predecir con precisión la clase minoritaria.

Para abordar este problema, se recurre a métodos de remuestreo que buscan equilibrar la distribución de las clases en el conjunto de datos. Uno de estos métodos es el de "Tomek Links", que se centra en identificar instancias ambiguas cercanas entre las clases mayoritaria y minoritaria.
El método Tomek Links se basa en la premisa de que las instancias que forman pares cercanos y de clases opuestas son propensas a ser ruido o instancias mal etiquetadas. Al eliminar la instancia de la clase mayoritaria en estos pares, se busca mejorar la capacidad del modelo para discriminar entre clases, reduciendo la interferencia de instancias ambiguas.

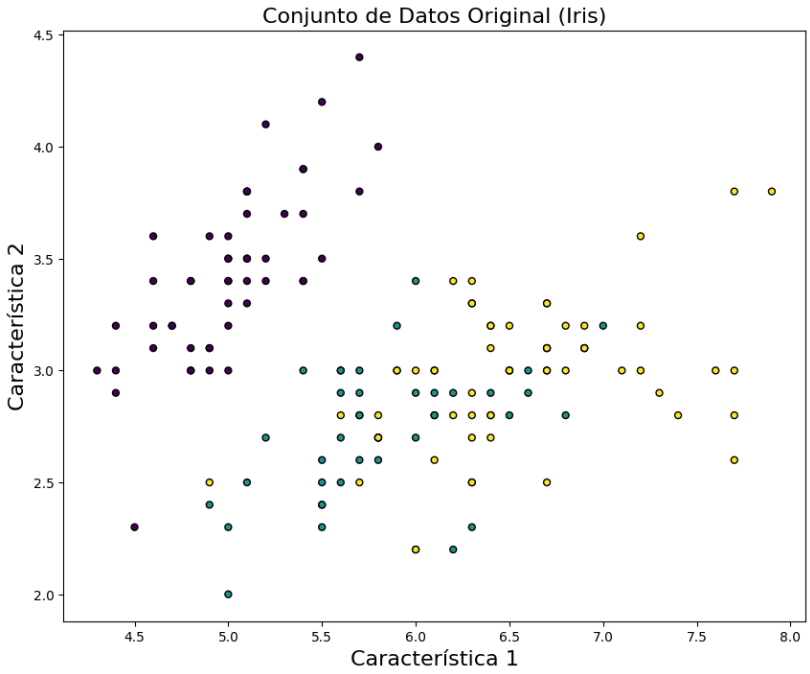


Figura 7: Datos iris

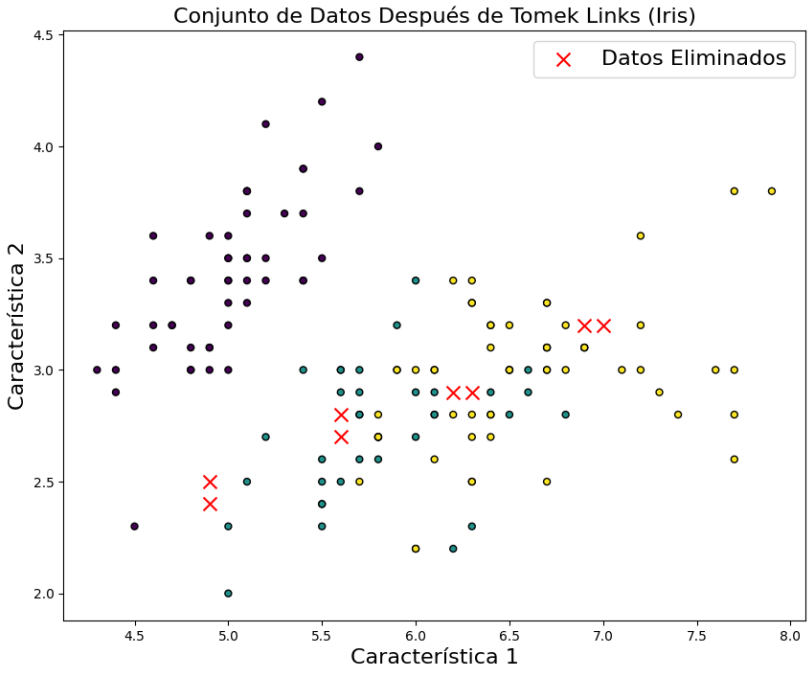


Figura 8: Ejemplo Tomek Links en datos iris

3 Metodología

Esta sección describe en detalle la metodología empleada para el desarrollo y evaluación del modelo de clasificación de créditos para la base de datos empleada.

3.1 Preparación y limpieza de datos

La preparación de los datos constituye un paso fundamental antes de proceder con cualquier análisis. Esta fase garantiza que los datos utilizados sean representativos para entrenar el modelo.

3.1.1 Tratamiento de datos nulos

En la fase inicial de la metodología, se realiza un análisis de las variables para identificar la presencia de valores nulos. Aquellas variables que exhiban más del 50% de sus datos como nulos serán excluidas. Esta decisión se fundamenta en la premisa de que la alta proporción de datos faltantes podría introducir más ruido que información útil, comprometiendo la calidad del modelo.
Para las variables que superen este filtro inicial, se implementarán técnicas de imputación. En el caso de las variables numéricas, los valores nulos se sustituirán por la media de la variable en cuestión. Por otro lado, para las variables categóricas, se optará por la imputación mediante la moda.

3.1.2 Selección y exclusión de variables no representativas o inaccesibles

El propósito este proceso es identificar las variables que no estarían disponibles en un escenario de toma de decisiones crediticias en tiempo real. Por ejemplo, ciertos datos pueden ser retrospectivos y solo conocidos después de que se haya determinado el resultado del préstamo, como ciertos indicadores de desempeño post-crédito. Estas variables, aunque potencialmente informativas, podrían introducir un sesgo retrospectivo si se incluyen en el modelo.

El proceso de esta exclusión es bajo una revisión crítica de las variables basada en el criterio experto, para determinar su relevancia con un sentido de negocio. Además este enfoque garantiza que el modelo se alinee con las necesidades y realidades del entorno empresarial.

3.1.3 Creación de la variable objetivo

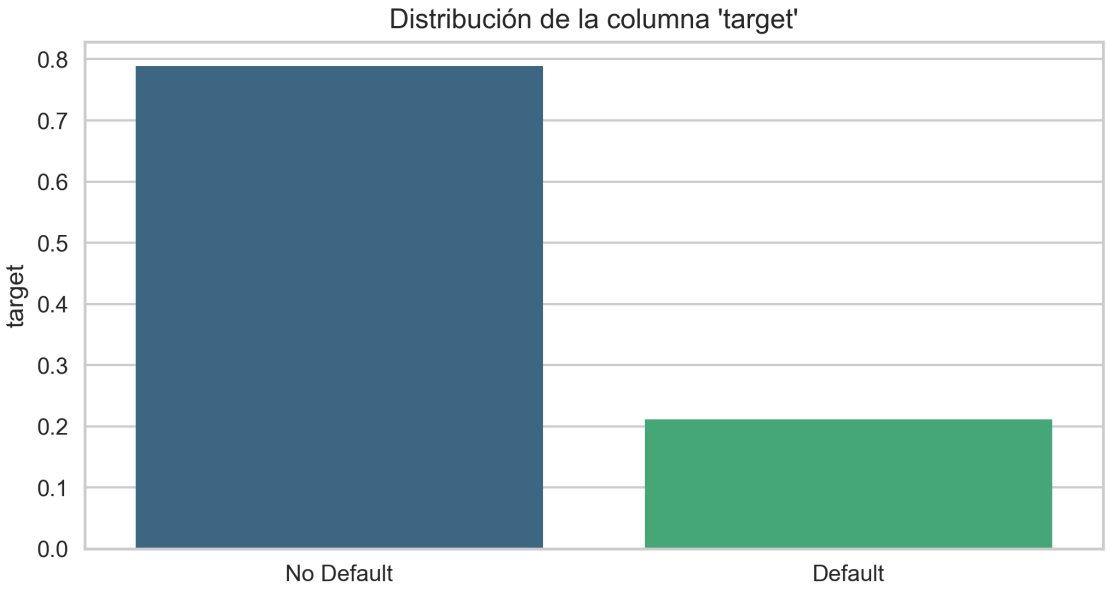
La construcción de una variable objetivo binaria; que actúa como la dependiente en el modelo, se define con el objetivo de categorizar los préstamos en dos estados. Para lograr una segmentación de clases más efectiva y precisa, se considerarán únicamente aquellos préstamos que han entrado en incumplimiento (Default, Late (31-120 días) y Charged Off) y aquellos que se han pagado en su totalidad, es decir, los clasificados como Fully Paid, que representan préstamos sin exposición a riesgo.

En consecuencia, la variable objetivo binaria se definirá de la siguiente manera: asignaremos el valor 0 a los préstamos clasificados como Fully Paid, indicando que no presentan riesgo de incumplimiento, y el valor 1 a los préstamos que se encuentran en alguna de las categorías de incumplimiento.

Los registros catalogado en otro estado del prestamos serán suprimidos de la base de datos.

Estado del Préstamo	**Descripción**
Default	El prestatario ha fallado en realizar pagos por al menos 120 días.
Late (31-120 días)	Hay un pago pendiente por un período de 31 a 120 días.
In Grace Period	El pago se ha retrasado entre 1 y 15 días; esto no se considera de alto riesgo inicialmente, pero puede escalar si no se realiza el pago.
Late (16-30 días)	El pago está pendiente por un período de 16 a 30 días.
Charged Off	El préstamo ha sido clasificado como irrecuperable (NPL) y se ha eliminado del activo, lo cual ocurre generalmente entre 90 y 180 días según la institución.
Current	Todos los pagos están al día, pero aún hay pagos futuros pendientes.
Issued	El préstamo ha sido aprobado recientemente y está a punto de comenzar el ciclo de pagos.
Fully Paid	El préstamo ha sido completamente saldado por el prestatario.

Una vez establecida la variable dependiente y tras descartar categorías superfluas asi como la variable original de estatus del prestamo, el análisis procede con un total de 1,373,886 registros, los cuales presentan un notable desequilibrio entre las clases.



3.1.4 Manejo de variables categóricas

En el proceso de análisis de datos, uno de los pasos críticos es el adecuado tratamiento de las variables categóricas. Estas variables, a menudo presentes en formas no numéricas como etiquetas o categorías, requieren una transformación adecuada para ser utilizadas eficazmente por los algoritmos de aprendizaje automático. Para este fin, se emplea la técnica de codificación "One-Hot Encoding". Esta decisión se fundamenta en el hecho de que las categorías en nuestro conjunto de datos no exhiben una secuencia o jerarquía inherente que justifique la aplicación de un encoding ordinal. En consecuencia, la totalidad de las variables categóricas serán transformadas mediante esta metodología.

One-Hot Encoding [2] es un método para convertir variables categóricas en un formato que puede ser interpretado por algoritmos de AA. En esencia, este proceso crea nuevas columnas para cada categoría posible de la variable original. Cada una de estas nuevas columnas contiene un valor de 0 o 1, indicando la ausencia o presencia de la categoría correspondiente en cada registro.

3.1.5 Creación de conjuntos de entrenamiento y prueba

Una fase crítica en la preparación de datos para el AA supervisado es la división del conjunto de datos en subconjuntos de entrenamiento y prueba. Esta separación es fundamental para evaluar la eficacia y la generalización de los modelos de AA de manera objetiva.
En el proceso de preparación de datos para nuestro análisis, se optó por una división de 80-20 del conjunto de datos al ser empíricamente el mejor división de los conjuntos de entrenamiento y prueba [3]. Esta proporción, ampliamente reconocida en el campo del apren-

[2] Müller, A. C. y Guido, S., 2016. Introduction to Machine Learning with Python. O'Reilly [3]

[3] Gholamy, A., Kreinovich, V., & Kosheleva, O. 2018. "Why 70/30 or 80/20 Relation Between Training and Testing Sets" [4]

dizaje automático, asegura un equilibrio entre una cantidad sustancial de datos para el entrenamiento del modelo y un conjunto adecuado para su evaluación. Para preservar la integridad de la variable objetivo en ambos conjuntos, se implementó un método de estratificación. Esta técnica garantiza que la distribución de la variable objetivo sea consistente en el conjunto de entrenamiento y en el de prueba, lo cual es esencial para evitar sesgos en el modelo, especialmente en presencia de un desequilibrio de clases.

3.1.6 Transformación de datos

Una etapa crucial en la preparación de los datos para el análisis mediante modelos de aprendizaje automático es la normalización de las características. Para este propósito, se empleará el método de escalado MinMax. Esta técnica ajusta los valores de las variables a una escala común, específicamente en el rango de 0 a 1, lo que es necesario para modelos que son sensibles a la magnitud de las variables, como las redes neuronales y los algoritmos basados en gradientes.

El escalador MinMax transforma cada característica del conjunto de datos de manera que su mínimo y máximo correspondan a 0 y 1 respectivamente. La fórmula utilizada para esta transformación es:

$$x_{\text{escalado}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

donde x es el valor original, $x_{\min}$ es el valor mínimo de la característica, y $x_{\max}$ es el valor máximo.

3.1.7 Remuestreo de limpieza

En esta fase de nuestra metodología, implementaremos el algoritmo de Tomek Links como una estrategia de remuestreo de limpieza. El objetivo de este enfoque es mejorar la calidad de nuestro conjunto de datos de entrenamiento.Tomek Links identifican pares de instancias cercanas pero de clases opuestas y eliminan aquellas que contribuyen al ruido y a la ambigüedad en la frontera de decisión entre las clases.

El uso de Tomek Links se espera que refine nuestro conjunto de datos, eliminando ejemplos que son posiblemente ruidosos o mal etiquetados y, por ende, mejore la capacidad del modelo de aprendizaje automático para generalizar a partir de datos más limpios y representativo.

3.1.8 Selección de características

Dado que nuestro conjunto original de datos presenta una alta dimensionalidad en las variables, es posible que algunas de ellas contribuyan más al ruido que a la información útil para el modelo. Por esta razón, seleccionaremos el subconjunto de características que ofrezca el mejor desempeño.

Para llevar a cabo esta selección, se optará por el método de Eliminación Recursiva de Características (Recursive Feature Elimination, RFE). Esta técnica ha demostrado ser comparable

a diversas otras estrategias de selección de características en modelos de clasificación binaria de créditos[4]. Este método consiste en eliminar las características que muestran menor importancia. El proceso de selección recursiva se basa en clasificar las características según su importancia en un modelo definido. De manera recursiva, el modelo se reconstruye utilizando las características restantes (inicialmente todas), y se eliminan los atributos menos importantes.En este trabajo, implementamos un algoritmo RFE denominado lrRFECV, basado en el modelo de regresión logística y utilizando técnicas de validación cruzada "kfold". Como algoritmo recursivo, el lrRFECV consume tiempo dependiendo del número de características evaluadas. Por lo tanto, en este estudio, el lrRFECV está precedido por el cálculo de la matriz de correlación de las características y la eliminación de aquellas altamente correlacionadas.

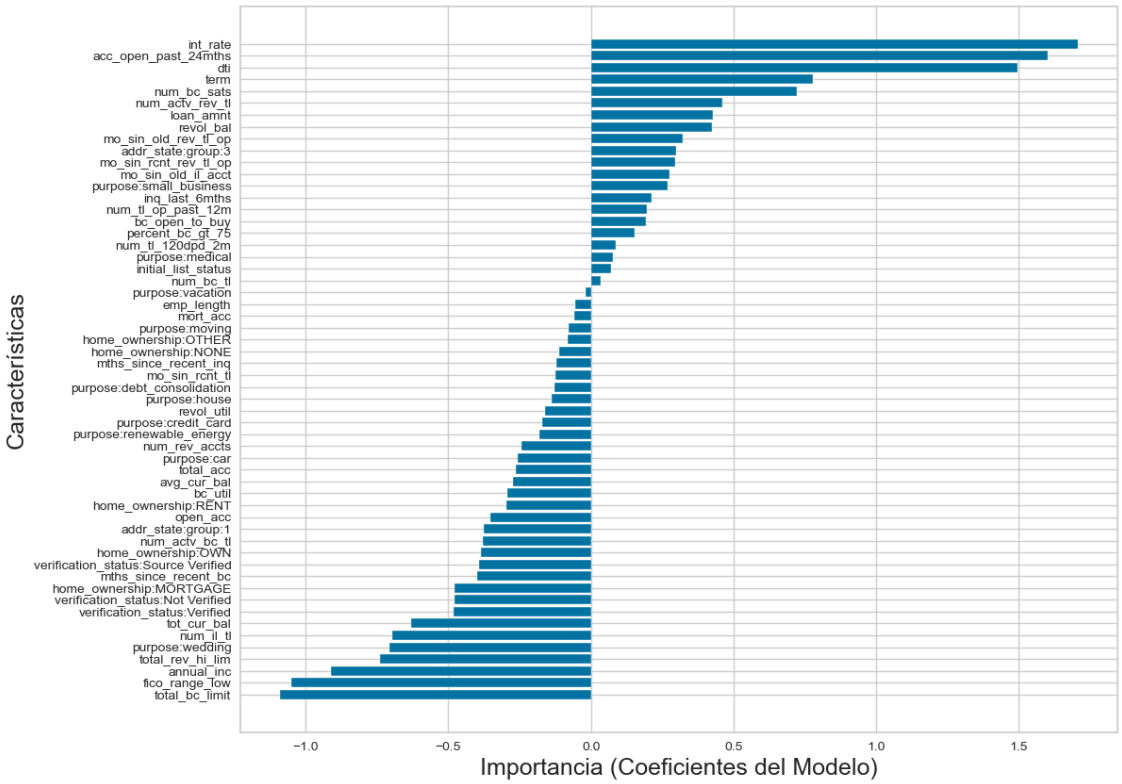


Figura 9: Variables seleccionadas

3.2 Diseño de experimentos

Tras completar las etapas de limpieza y preparación de la base de datos, el siguiente paso en nuestra metodología consiste en implementar una serie de pruebas estructuradas, conocidas como diseño de experimentos. Este proceso permite determinar el modelo más eficaz.

Considerando la alta demanda de recursos computacionales y el tiempo significativo requerido para la evaluación de estos experimentos, se ha decidido emplear una muestra aleatoria que

[4]Jemai, J., & Zarrad, A. 2023. "Feature Selection Engineering for Credit Risk Assessment in Retail Banking. Information"[5]

mantenga la proporción inicial de la variable objetivo, dicha muestra es correspondiente al 20% de nuestros datos. Esta medida se adopta con el objetivo principal de acelerar el proceso de evaluación. Sin embargo, es relevante mencionar que, en condiciones ideales, la utilización del conjunto completo de datos para obtener un análisis más preciso.

3.2.1 Selección de modelo

En este proceso, nuestro objetivo principal es encontrar un modelo con un equilibrio óptimo entre varios factores clave: el tiempo de ejecución, la eficiencia (medida mediante la métrica ROC-AUC) y la interpretabilidad del modelo.
Para asegurar una evaluación precisa de la eficiencia y el costo computacional de cada modelo, adoptaremos un enfoque en el que evaluaremos cada modelo en $n = 30$ ocasiones diferentes. Esta repetición garantiza la obtención de medidas representativas para cada resultado, proporcionando así una base de datos confiable para el análisis comparativo.

El análisis de los resultados, como se ilustra en el gráfico adjunto, indica que el valor máximo de ROC-AUC obtenido en los modelos más eficientes es de aproximadamente 0.73, con tiempos de ejecución variados. Basándonos en esta observación, hemos decidido seleccionar la regresión logística para una evaluación más detallada. La elección de la regresión logística se debe a su reconocimiento como un método clásico y robusto en la clasificación binaria además de presentar uno de los mejores rendimientos.

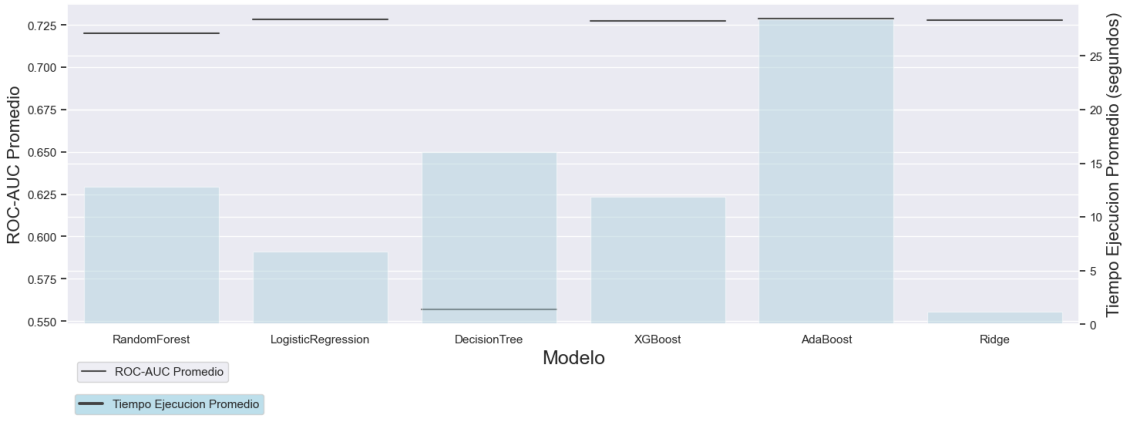


Figura 10: Distribución de ROC-AUC y Tiempo de Ejecución por Modelo

3.2.2 Selección de hiperparametros

Dada la elección del modelo de regresión logística, se procederá a realizar una búsqueda exhaustiva de los hiperparámetros más adecuados. Los hiperparámetros que serán objeto de estudio y optimización incluyen:
penalty (Penalización): Determina la norma utilizada para la regularización del modelo. Las opciones a considerar son:

- **l1**: Aplica la norma L1, conduciendo a modelos con un menor número de coeficientes distintos de cero, favoreciendo la selección de características.

- **l2**: Emplea la norma L2, que penaliza los coeficientes de mayor magnitud, resultando en modelos más equilibrados y con una mejor generalización.
- **elasticnet**: Combina las normas L1 y L2, ofreciendo un equilibrio entre la selección de características y la regularización.
- **none**: No aplica regularización, omitiendo el término de penalización en el modelo.

C (Fuerza de Regularización): Controla la inversa de la fuerza de regularización, con implicaciones directas en la capacidad del modelo para ajustarse a los datos. Un "C" bajo incrementa la regularización, lo que puede prevenir el sobreajuste pero a riesgo de un ajuste insuficiente, en cambio un "C" alto disminuye la regularización, permitiendo un ajuste más estrecho a los datos de entrenamiento pero con mayor riesgo de sobreajuste.

Experimento 1:
Se realizó una validación cruzada con `kfold=5` y se exploraron diversos hiperparámetros. Los valores probados fueron:

`penalty`: ["l1", "l2", "elasticnet", "none"]

`C`: [0.001, 0.01, 0.1, 1, 10, 100]

El resultado óptimo de este primer experimento identificó un `"C"` de 0.1 y una regularización de tipo `"l2"` como la combinación más eficaz para el modelo.
Experimento 2:
En este experimento, se procedió a una validación cruzada utilizando `kfold=5`. El foco estuvo en afinar el valor del parámetro de regularización `"C"`, explorando valores cercanos al obtenido en el Experimento 1, manteniendo constante la regularización `"l2"`.

`penalty`: ["l2"]

`C`: [0.05, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.2]

Como resultado de este análisis detallado, se identificó que el valor `"C"` de 0.12 ofrecía el mejor rendimiento para el modelo.

4 Resultados

A continuación se presentan los resultados del modelo final obtenido de calorificación créditos.

La importancia de las principales variables se puede resumir:

Métrica	Valor (Test)	Valor (Train)
Exactitud (Accuracy):	0.821	0.820
Precisión (Precision):	0.519	0.511
Sensibilidad (Recall):	0.115	0.116
Puntuación F1 (F1-Score):	0.188	0.190
Área bajo la curva ROC (ROC AUC):	0.728	0.731
Estadístico KS (KS):	0.334	0.339

Tabla 1: Resultados de métricas de rendimiento del modelo

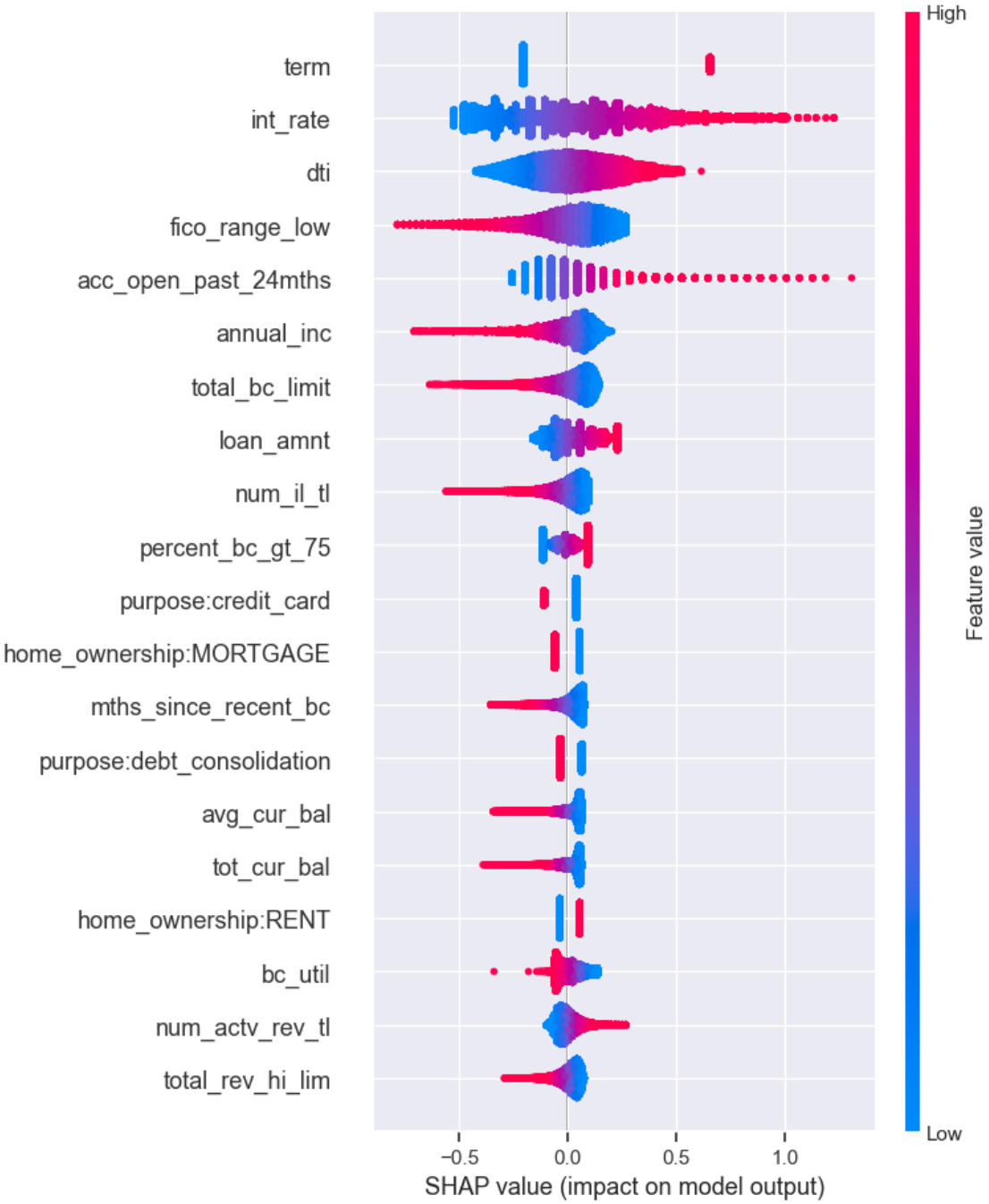


Figura 11: Importancia de las variables

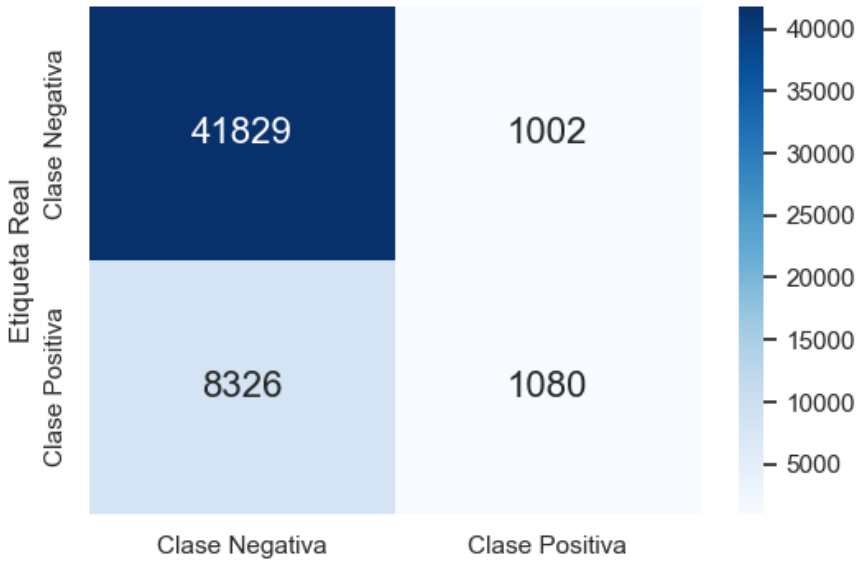


Figura 12: Matriz de confusión modelo final con remuestreo

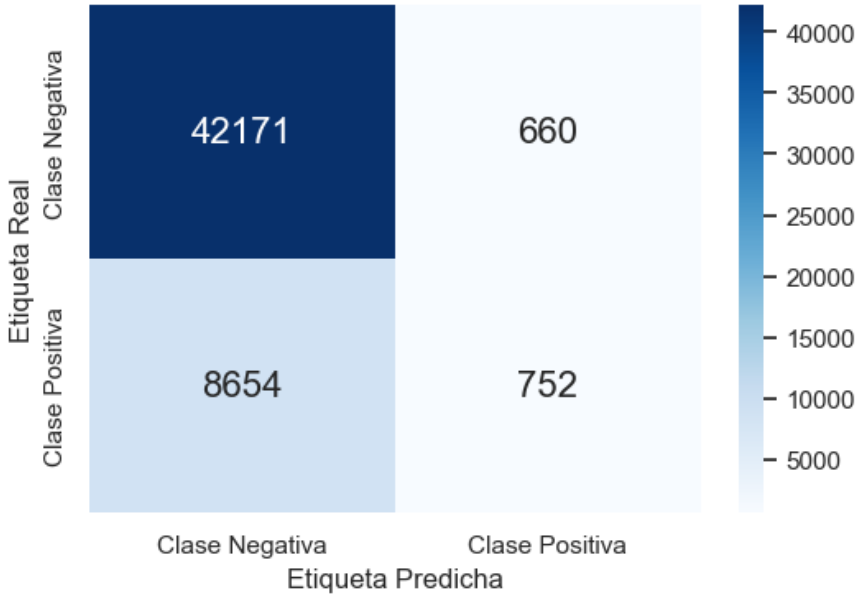


Figura 13: Matriz de confusión modelo final sin remuestreo

Aplicando un corte de clases personalizado

La selección de un umbral óptimo (threshold) en modelos de clasificación permite mejorar su rendimiento y aplicabilidad. Este proceso implica ajustar el punto de corte en las probabilidades predichas que determina la clasificación en diferentes categorías. Una técnica efectiva para esto es el uso del índice de Kolmogorov-Smirnov (KS), que identifica el umbral donde la diferencia entre las distribuciones acumulativas de las clases positivas y negativas es máxima.

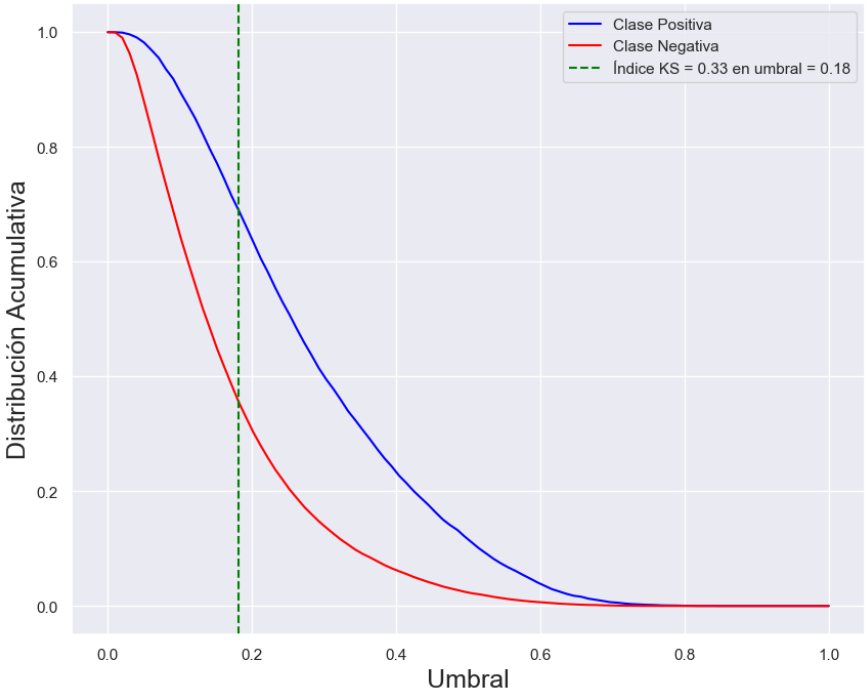


Figura 14: CDF de clases con índice KS

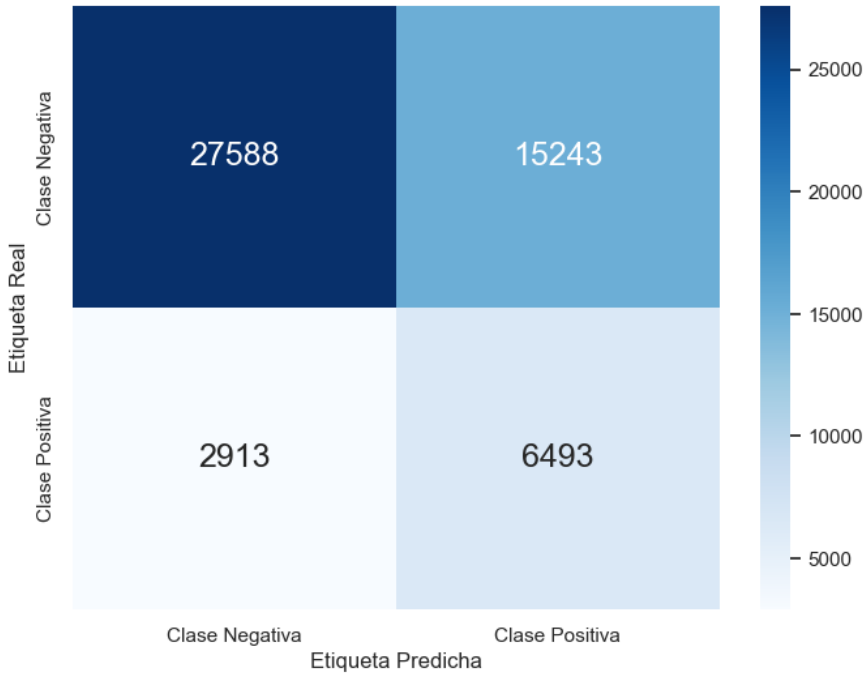


Figura 15: Matriz de confusión modelo final corte de clases basado en KS

5 Conclusiones

El análisis de los modelos de clasificación binaria aplicados en este estudio revela que, en general, todos los modelos evaluados mostraron un desempeño comparable en términos de sus métricas de rendimiento. Sin embargo, una excepción notable fueron los árboles de decisión, los cuales no alcanzaron la eficacia de sus contrapartes. En cuanto al modelo final seleccionado, la regresión logística con un ajuste de hiperparámetros, se observó que no hubo un incremento significativo en el desempeño en comparación con la configuración inicial del modelo.

No obstante, la implementación de técnicas de remuestreo, en particular el uso de Tomek Links, resultó ser un factor determinante para el mejoramiento en la distinción entre las clases. Este hallazgo subraya la importancia de una preparación de datos adecuada, más allá de la selección y ajuste de modelos, en la mejora de la capacidad predictiva en contextos de clasificación desbalanceados.

De manera adicional, al ajustar el umbral de decisión para personalizar el corte, el modelo ha demostrado su capacidad para rechazar aproximadamente el 80% de los créditos en default, aunque esto también resulta en el rechazo de alrededor de un tercio de los créditos buenos.

Ambos modelos, con y sin el umbral de corte personalizado, tienen potencial para ser herramientas valiosas, dependiendo del contexto específico de la institución y su apetito de riesgo. Las entidades que priorizan la reducción del riesgo de default podrían favorecer el modelo con el umbral de corte personalizado, mientras que aquellas que buscan maximizar la aceptación de créditos sin incrementar significativamente el riesgo podrían optar por el modelo estándar.

Bibliografía

[1] Nargundkar & Priestley, 2003, "Assessment of Model Development Techniques and Evaluation Methods for Binary Classification in the Credit Industry"

[2] Beyene & Babo, 2023, "Bank Loan Classification of Imbalanced Dataset Using Machine Learning Approach"

[3] Müller, A. C. y Guido, S., 2016. Introduction to Machine Learning with Python. O'Reilly.

[4] Gholamy, A., Kreinovich, V., & Kosheleva, O. 2018. "Why 70/30 or 80/20 Relation Between Training and Testing Sets"

[5] Jemai, J., & Zarrad, A. 2023. "Feature Selection Engineering for Credit Risk Assessment in Retail Banking. Information"