UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Maestría en Ciencias Matemáticas

Desarrollo de un Modelo Predictivo de Clasificación para la Evaluación del Riesgo Crediticio.

TÉSIS
QUE PARA OPTAR POR EL GRADO DE:
MAESTRÍA EN CIENCIAS MATEMÁTICAS

PRESENTA:
FRANCINE OCHOA FERNÁNDEZ

TUTOR O TUTORES PRINCIPALES
Erick Treviño Aguilar
Unidad Cuernavaca del Instituto de Matemáticas UNAM

Ciudad Universitaria,CD. MX, ENERO 2025

Agradecimientos

Trabajo realizado gracias al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM, proyecto TA101322.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Presentación

El presente trabajo de investigación tiene como objetivo principal explorar y aplicar modelos de aprendizaje de maquina (en inglés machine learning) para predecir el riesgo crediticio en clientes de un conjunto de datos específico. El desarrollo y aplicación de un modelo de calificación crediticia se abordará mediante una metodología que comprende diversas etapas fundamentales en la gestión del riesgo crediticio.

El proceso inicia con una detallada revisión de los datos y la definición de los parámetros del proyecto. Esto implica determinar la viabilidad del desarrollo de un modelo de calificación crediticia y establecer los objetivos, ventanas de muestreo y rendimiento, garantizando la disponibilidad y calidad de los datos necesarios. Esta fase es crucial, ya que se busca la selección de cuentas representativas que reflejen el comportamiento financiero esperado.

La recopilación de datos es un paso subsiguiente y crucial en la definición de los parámetros del proyecto. Se enfoca en la obtención de información relevante, como el historial crediticio, indicadores financieros y demográficos, entre otros, fundamentales para la construcción de un modelo de calificación crediticia efectivo.

La definición de cuentas "buenas" y "malas" es un aspecto esencial en el proceso. Esta etapa incluye la exclusión de ciertos tipos de cuentas que no representan el perfil de cliente normal o que podrían distorsionar el análisis. Además, se establecen las ventanas de rendimiento y de muestra, donde se evaluará el comportamiento de las cuentas para clasificarlas como "buenas" o "malas".

La definición final de cuentas "malas" se valida mediante métodos analíticos y de consenso,

garantizando que éstas sean coherentes con los objetivos organizacionales y con los productos específicos para los que se está construyendo el modelo de calificación crediticia.

Asimismo, se aborda la segmentación como una estrategia que podría proporcionar una mejor diferenciación del riesgo entre diferentes subgrupos de clientes. Esto se lleva a cabo mediante la identificación de características o comportamientos diferenciales entre segmentos, permitiendo la elaboración de un modelo de calificación crediticia más preciso y adaptado a cada grupo.

En un paso posterior, se ejecuta un minucioso proceso de minería de datos, el cual abarca varias etapas fundamentales, cada uno crucial para la construcción de modelos de clasificación efectivos. Estos pasos se distribuyen en diversas fases que abarcan desde la comprensión inicial del problema hasta la validación del modelo construido.

- Comprensión del Problema y Recopilación de Datos: En esta fase inicial, se realiza una inmersión profunda en el problema de predicción del riesgo crediticio. Se identifican y analizan las variables relevantes que impactan en la evaluación del riesgo financiero. Posteriormente, se recopilan los datos necesarios para abordar este problema, asegurando la representatividad y calidad de los mismos.
- Exploración y Preprocesamiento de Datos: Una vez recopilados los datos, se inicia un análisis exploratorio exhaustivo. Se busca comprender la distribución de las características, detectar posibles valores atípicos o anomalías, y comprender las relaciones entre las variables. Aquí se realiza el preprocesamiento de los datos, incluyendo la limpieza, la sustitución de valores faltantes y la codificación de variables categóricas. El objetivo principal es garantizar la calidad, coherencia y consistencia de los datos para su posterior utilización en los modelos.
- Selección de Características y División de Datos: Una vez completada la fase de exploración, se procede a la selección de las características más relevantes para la predicción del riesgo crediticio. Este paso implica la identificación de aquellas variables que mejor contribuyen a la capacidad predictiva de los modelos. Luego, se divide el conjunto de datos en conjuntos de entrenamiento y prueba, asegurando que los modelos se construyan sobre datos independientes y se evalúen con datos no vistos.

- Elección del Modelo y Evaluación: En esta etapa, se seleccionan los algoritmos de clasificación más adecuados para el problema de predicción del riesgo crediticio. Se entrenan varios modelos, utilizando algoritmos como regresión logística, árboles de decisión, máquinas de soporte vectorial (SVM) o redes neuronales, según la idoneidad de cada técnica. Luego, se evalúa el rendimiento de estos modelos utilizando métricas específicas para la clasificación, como el área bajo la curva ROC (abreviado AUC), precisión, recall, F1-score, entre otras. En el caso de la tesis, se utilizará el AUC, matriz de confusión y una métrica basada en el estadístico de Kolmogorov-Smirnov.
- Ajuste y Validación del Modelo: El ajuste fino de los modelos seleccionados se realiza para mejorar su precisión y capacidad de generalización. Esta fase implica la optimización de hiperparámetros, la validación cruzada y la selección del mejor modelo. Además, se valida el rendimiento final del modelo elegido utilizando datos de prueba, asegurando su capacidad para generalizar patrones y realizar predicciones precisas en situaciones reales.

Estos pasos de la minería de datos son fundamentales para la construcción de modelos de clasificación efectivos en la predicción del riesgo crediticio.

1.1. Introducción

El acceso al crédito juega un papel fundamental en la vida de millones de personas en todo el mundo. Desde la compra de una vivienda hasta el financiamiento de un negocio, la capacidad de obtener crédito de manera eficiente y justa es crucial para el desarrollo económico y social. En este contexto, el desarrollo de modelos de calificación crediticia se presenta como una herramienta indispensable para evaluar el riesgo asociado a la concesión de créditos y préstamos.

Este trabajo se adentra en el proceso de desarrollo de un modelo de calificación crediticia, una tarea compleja que abarca múltiples etapas y aspectos clave. Comenzamos revisando la disponibilidad y calidad de los datos, reconociendo que la calidad de los datos es fundamental para la precisión y confiabilidad de cualquier modelo. La recopilación de datos para la definición de los parámetros del proyecto se convierte así en un ejercicio crucial, donde la exhaustividad y la precisión son muy importantes.

Una vez establecidos los parámetros del proyecto, nos adentramos en la definición de cuentas "malas" y "buenas", una tarea que requiere una cuidadosa segmentación y análisis. Es importante tener en cuenta que ciertos tipos de cuentas deben excluirse de la muestra de desarrollo, y que las ventanas de rendimiento y de muestra deben definirse con claridad para garantizar la validez del modelo. La definición de clientes "malos" y la confirmación de esta definición son procesos que requieren un análisis detallado y estratégico, donde la segmentación juega un papel fundamental.

Con las bases bien establecidas, nos adentramos en la experimentación con una variedad de modelos de clasificación. Utilizando los modelos XGBoost, Random Forest, Regresión Logístic, Máquinas de Soporte Vectorial, Análisis Discriminante Lineal, k-Vecinos Más Cercanos y Naive Bayes, exploramos el rendimiento de diferentes algoritmos utilizando métricas clave, el Área bajo la Curva ROC, el Estadístico de Kolmogorov-Smirnov (KS) y la Matriz de Confusión. Este análisis comparativo nos brinda una visión integral del desempeño de cada modelo y nos permite identificar las fortalezas y debilidades de cada enfoque.

La implementación del código se lleva a cabo siguiendo una estructura clara y organizada, lo que facilita el análisis y la visualización de resultados. A través de un análisis descriptivo y una visualización profunda, exploramos la distribución de métricas clave en función de variables relevantes y modelos, resaltando los factores que influyen en el rendimiento del modelo.

Este trabajo se sumerge en el desarrollo de modelos de calificación crediticia, explorando cada etapa del proceso con rigor. A través de una combinación de análisis estadístico, exploración de datos y experimentación con modelos, se ofrece una contribución significativa al campo de la evaluación del riesgo crediticio y sentar las bases para futuras investigaciones en esta área crucial de las finanzas modernas.

1.2. Proceso de desarrollo de un modelo de calificación crediticia

La presente sección describe el proceso de desarrollo de un modelo de calificación crediticia, el cual se basa en los principios y metodologías presentadas en el libro "Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring" de Naeem Siddiqi[12]. Este libro proporciona una guía integral para la creación e implementación de sistemas de puntuación crediticia inteligentes, abordando aspectos clave como la selección de variables, la construcción de modelos predictivos y la validación de la capacidad predictiva del modelo. A lo largo de este capítulo, se aplicarán los conceptos y técnicas presentados en este libro para desarrollar un modelo robusto y efectivo para la evaluación del riesgo crediticio.

1.2.1. Revisión de datos y parámetros del proyecto

Esta etapa está diseñada para determinar si el desarrollo de un modelo de calificación crediticia es viable y también para definir los parámetros del proyecto. Los parámetros incluyen las exclusiones, la definición de objetivos, la ventana de muestreo y la ventana de rendimiento.

1.2.2. Disponibilidad y calidad de los datos

Es necesario comprobar la disponibilidad de los datos, en los términos de calidad y cantidad. Para el desarrollo un modelo de calificación crediticia se necesitan datos fiables y limpios, con un número mínimo aceptable de cuentas "buenas" y "malas". Normalmente, es más difícil encontrar un número suficiente de cuentas "malas" que de cuentas "buenas". En general, la cantidad de datos necesarios debe cumplir los requisitos de significancia estadística y aleatoriedad.

Los datos demográficos y otros datos de la aplicación que no se verifican, como los ingresos, son más susceptibles de ser distorsionados, en cambio elementos como los datos del buró de crédito, los datos inmobiliarios, los índices financieros, etc, son más fiables y pueden utilizarse.

1.2.3. Recopilación de datos para la definición de los parámetros del proyecto

Los parámetros del proyecto incluyen la determinación de las definiciones de cuentas "buenas" y "malas", el establecimiento de las ventanas de rendimiento y de muestreo, y la definición de las exclusiones de datos para la creación de la muestra de desarrollo y el propio proceso de desarrollo.

El número de cuenta/identificación, la fecha de aplicación, el historial de atrasos/reclamaciones a lo largo de la vida de la cuenta, el indicador de aceptación/rechazo, el producto/canal y otros identificadores de segmento, el estado actual de la cuenta son algunos de los datos que suelen recopilarse para las aplicaciones.

Para el desarrollo de un modelo de calificación crediticia de comportamiento, las cuentas se eligen en un momento determinado y se analiza su comportamiento durante, normalmente, un periodo de 6 ó 12 meses. En nuestro caso, los créditos son muy cortos, de cuatro semanas, y este análisis no es tan relevante. Si es necesario, pueden añadirse otros elementos de datos relevantes, como datos demográficos, geográficos y cualquier otro criterio que ayude a construir el perfil de la cartera de clientes.

1.2.4. Definición de los parámetros del proyecto

Los análisis se realizan no sólo para definir los parámetros del proyecto, sino también para comprender el negocio a través de los datos.

1.2.5. Exclusiones

Ciertos tipos de cuentas deben excluirse de la muestra de desarrollo. Las cuentas utilizadas para el desarrollo son las que se puntuarían durante las operaciones cotidianas normales de otorgamiento de créditos, y las que constituirían la cartera de clientes prevista. Las cuentas que tienen un rendimiento anormal no deben formar parte de ninguna muestra de desarrollo, como las cuentas de empleados, VIP, cuentas fuera del país, cuentas preaprobadas, tarjetas perdidas/robadas, cuentas de fallecidos, cuentas de menores de edad y cuentas canceladas voluntariamente. Algunos desarrolladores incluyen las cuentas

canceladas como "indeterminadas". Si hay zonas geográficas o mercados en los que la empresa ya no opera, también deben excluirse los datos de estos mercados; del mismo modo, no debe incluirse ningún tipo de cuenta o solicitante que no vaya a puntuarse, o que no sea un cliente normal.

1.2.6. Ventanas de rendimiento y de muestra y definición de cuentas "malas"

El modelo de calificación crediticia se elabora partiendo del supuesto de que "el rendimiento futuro reflejará el rendimiento pasado". El rendimiento de las cuentas abiertas en el pasado se analiza para predecir el rendimiento de las cuentas futuras. Tenemos que recopilar datos de las cuentas abiertas durante un periodo de tiempo concreto y, a continuación, supervisar su rendimiento durante otro periodo de tiempo concreto para determinar si fueron buenas o malas.

Los datos recopilados (las variables) junto con la clasificación bueno/malo (el objetivo) constituyen la muestra de desarrollo a partir de la cual se elaboran el modelo de calificación crediticia. "Ventana de rendimiento" es la ventana temporal en la que se supervisa el rendimiento de las cuentas abiertas durante un periodo de tiempo determinado (es decir, la ventana de muestra) para asignar la clase (objetivo). "Ventana de muestra" es el marco temporal a partir del cual se seleccionarán los casos buenos y malos conocidos, para la muestra de desarrollo.

Una forma sencilla de establecer las ventanas de rendimiento y de muestra consiste en analizar el rendimiento de los pagos o la morosidad de la cartera, y trazar la evolución de los casos "malos" definidos a lo largo del tiempo. La selección de muestras de desarrollo a partir de un conjunto maduro se hace para minimizar las posibilidades de clasificar erróneamente el rendimiento (es decir, todas las cuentas han tenido tiempo suficiente para volverse deficientes), y para asegurar que la definición de " malo " que resulta de una muestra poco madura no subestimará los índices finales de malos resultados esperados. El tiempo que tardan las cuentas en madurar varía en función del producto y de la definición de "malo" seleccionada.

El modelo de calificación crediticia de comportamiento para uso operativo se construyen

normalmente para ventanas de rendimiento de 6 o 12 meses. Los modelos de cobro se construyen normalmente para ventanas de rendimiento de un mes. Cuando se desarrollan modelos predictivos para requisitos normativos específicos, la ventana de rendimiento puede venir dictada por la normativa.

Cuando se construye un modelo de calificación crediticia de morosidad, este análisis debe repetirse para varias definiciones de morosidad relevantes. Esto se hace porque las diferentes definiciones producirán diferentes recuentos de muestras. Factores como la ventana de muestreo y la definición de bueno/malo deben combinarse para obtener una muestra suficientemente grande.

En la medida de lo posible, el análisis debe realizarse utilizando la definición "alguna vez mala " (es decir, se considera que la cuenta es " mala " si alcanza el estado de morosidad definido en cualquier momento durante la ventana de rendimiento). Si esto no es posible, bastará con una definición "actual" de "malo", en la que el estado de morosidad de las cuentas se tome del rendimiento más reciente de final de mes.

1.2.7. Efectos de la estacionalidad

En este punto también debe establecerse la variación de los índices de solicitudes y aprobaciones a lo largo del tiempo, así como el efecto de cualquier efecto de temporalidad. Se trata de garantizar que la muestra de desarrollo (de la ventana de muestreo) no incluya ningún dato de periodos "anormales", de modo que la muestra utilizada para el desarrollo se ajuste a los periodos comerciales normales, representando a la población típica "que pasa por la puerta". El objetivo es ajustarse a la suposición de que "el futuro es como el pasado", de modo que la muestra de desarrollo sea representativa de los futuros solicitantes previstos, lo que también ayuda a generar predicciones precisas de la tasa de aprobación/tasa de morosos y produce un modelo de calificación crediticia sólido que supera el paso del tiempo. Con ello podemos detectar comportamientos extremos, ya que establecer una referencia a lo "normal" es difícil.

Hay varias formas de contrarrestar los efectos de los periodos anormales cuando la población de solicitantes no representa la población normal "que pasa por la puerta". En primer lugar, deben establecerse las razones de la anormalidad; la mejor forma de hacerlo es comparando las características del cliente normal con las de la ventana de muestreo.

Otra técnica para "normalizar" los datos consiste en filtrar la fuente de anormalidad. La muestra de desarrollo resultante (y las estadísticas de la cartera) se ajustarán entonces a las operaciones cotidianas normales de esta empresa.

Los efectos de la estacionalidad también pueden contrarrestarse tomando múltiples ventanas de muestreo, pero teniendo cada una de ellas una ventana de rendimiento igual. En los casos en que no sea posible tomar muestras escalonadas o ampliar la ventana de muestreo, y se sepa y entienda que las razones de la anormalidad se limitan a un mes concreto, también es posible crear una muestra excluyendo los registros atípicos. Para ello es necesario disponer de información detallada sobre las distribuciones existentes de las características durante los periodos normales de actividad; se recomienda analizar una muestra de los registros excluidos para detectar tendencias antes de descartarlos.

1.2.8. Definición de "Malo"

Esta fase clasifica el rendimiento de las cuentas en tres grupos principales: "mala", "buena" e "indeterminada". La definición de lo que constituye una cuenta "mala" depende de varias consideraciones:

- La definición debe estar en sintonía con los objetivos de la organización. Si el objetivo es aumentar la rentabilidad, la definición debe fijarse en un punto de morosidad en el que la cuenta deje de ser rentable.
- La definición debe estar en línea con el producto o el objetivo para el que se está construyendo el modelo de calificación crediticia.
- Una definición "más estricta" proporciona una diferenciación más extrema (y precisa), pero en algunos casos puede dar lugar a tamaños de muestra pequeños.
- Una definición "menos estricta" proporcionará un mayor número de cuentas para la muestra, pero puede no ser un diferenciador suficientemente bueno entre cuentas buenas y malas, y por lo tanto producirá un modelo de calificación crediticia débil.
- La definición debe ser fácilmente interpretable y rastreable.
- Puede ser beneficioso tener definiciones comunes de cuentas "malas" en varios segmentos.

1.2.9. Confirmación de la definición de "malo"

Una vez identificada una definición inicial de cuenta "mala", se puede realizar un análisis posterior para confirmarla, para asegurarse de que los identificados son realmente malos. La confirmación puede hacerse mediante el juicio de expertos, análisis o una combinación de ambos.

Método de consenso El método de juicio o consenso implica que varios desarrolladores se reúnan y lleguen a un consenso sobre la mejor definición de una cuenta "mala", basándose en la experiencia y en consideraciones operativas.

Métodos analíticos

Dos métodos analíticos para confirmar las definiciones de "malos" son el análisis de la tasa de rotación y la comparación de la morosidad actual con la peor. También se puede realizar un análisis de rentabilidad para confirmar que los definidos como malos no son rentables o producen un valor actual neto (VAN) negativo. Los siguientes análisis para determinar y confirmar las definiciones de "malo" podrían realizarse tanto para el modelo de calificación crediticia de aplicación como para el de comportamiento.

Análisis de la tasa de rotación compara la peor morosidad en los "x" meses anteriores especificados con la de los "x" meses siguientes y, a continuación, calcula el porcentaje de cuentas que mantienen su peor morosidad, mejoran o "rotan" a los siguientes niveles de morosidad. El objetivo es identificar un "punto de no retorno" (es decir, el nivel de morosidad en el que la mayoría de las cuentas se vuelven incurables). Por lo general, la inmensa mayoría de las cuentas que alcanzan los 90 días de mora no se curan, sino que empeoran (rotación).

Comparación de la morosidad actual con la peor concepto similar al del análisis de la tasa de rotación, pero es más sencillo. Compara el peor estado de morosidad (de la historia) de las cuentas con su estado de morosidad más actual. El objetivo aquí también es buscar un "punto de no retorno".

1.2.10. "Buenas" e "Indeterminadas"

El mismo análisis realizado anteriormente para las cuentas "malas" puede utilizarse para definir una cuenta "buena". Definir las cuentas "buenas" es menos analítico, y normalmente evidente. Algunas características de una buena cuenta son:

- Nunca morosa o morosa hasta un punto en el que la tasa de rotación es inferior a $x\,\%$.
- Rentable, o VAN positivo.
- Sin reclamaciones.
- Nunca en situación de quiebra.
- Sin fraude.
- Tasa de recuperación de $y\,\%$.

Mientras que las cuentas buenas deben mantener su estatus durante toda la ventana de resultados, una cuenta mala puede definirse por alcanzar la fase de morosidad especificada en cualquier momento de la ventana de resultados (según la definición de " alguna vez ").

Las cuentas indeterminadas son aquellas que no entran de forma concluyente ni en la categoría de "buenas" ni en la de "malas". Se trata de cuentas que no tienen un historial de resultados suficiente para ser clasificadas, o que presentan una morosidad leve con tasas de rotación ni lo suficientemente bajas para ser clasificadas como buenas, ni lo suficientemente altas para ser malas. Las cuentas indeterminadas sólo se utilizan cuando la definición de "mala" puede establecerse de varias maneras, y no suelen ser necesarias cuando la definición es clara. Como regla general, las indeterminaciones no deben superar entre el 10 % y el 15 % de la cartera. En los casos en que la proporción de indeterminados sea muy elevada, deben realizarse análisis para abordar las causas profundas de la inactividad. En la elaboración del modelo de calificación crediticia sólo se utilizan las cuentas definidas como "buenas" y "malas". Las cuentas indeterminadas se añaden al hacer las proyecciones para reflejar adecuadamente la verdadera población "de entrada".

1.2.11. Segmentación

En algunos casos, el uso de varios modelos de calificación crediticia para una cartera proporciona una mejor diferenciación del riesgo que el uso de un solo modelo de calificación crediticia para todos. Este suele ser el caso cuando una población se compone de distintas subpoblaciones, donde una modelo de calificación crediticia no funcionará eficazmente para todas ellas.

El proceso de identificación de estas subpoblaciones se denomina segmentación. Hay dos formas principales de realizar la segmentación:

1. Generando ideas de segmentación basadas en la experiencia y el conocimiento del sector, y validándolas después mediante análisis.
2. Generando segmentos únicos mediante técnicas estadísticas como la clusterización o los árboles de decisión.

En cualquier caso, los segmentos seleccionados deben ser lo suficientemente grandes como para permitir un muestreo significativo para el desarrollo modelos de calificación crediticia independientes.

En el desarrollo de modelos de calificación crediticia, una población "distinta" no se reconoce como tal en función de sus características definitorias, sino más bien en función de su rendimiento. El objetivo es definir segmentos basados en el rendimiento en función del riesgo, no sólo en el perfil de riesgo.

Detectar un comportamiento diferente no es razón suficiente para la segmentación. La diferencia debe traducirse en efectos mensurables sobre el negocio (por ejemplo, menores pérdidas, mayores tasas de aprobación para ese segmento, etc). La segmentación también debe hacerse teniendo en cuenta los planes futuros, los modelos de calificación crediticia deben aplicarse en el futuro, en futuros segmentos de solicitantes.

1. *Segmentación basada en la experiencia (heurística)* La segmentación basada en la experiencia incluye ideas generadas a partir del conocimiento y la experiencia empresarial, consideraciones operativas y prácticas del sector. Las áreas de segmentación típicas utilizadas en la industria incluyen aquellas basadas en:
    - Datos demográficos.

- Tipo de producto.
- Fuentes de negocio.
- Datos disponibles.
- Tipo de solicitante.

Una vez generadas las ideas sobre segmentación, es necesario confirmarlas al menos con alguna prueba empírica. Un método sencillo para confirmar las ideas de segmentación y establecer la necesidad de segmentación consiste en analizar el comportamiento de riesgo de la misma característica en diferentes segmentos predefinidos. Si la misma característica predice de forma diferente en segmentos únicos, esto puede suponer un argumento a favor de los modelos de calificación crediticia segmentados. Sin embargo, si la característica predice el riesgo de la misma manera en los distintos segmentos, no se necesitan modelos de calificación crediticia adicionales, ya que no hay diferenciación.

Otra forma de confirmar las ideas iniciales de segmentación y de identificar segmentos únicos consiste en examinar los índices de morosidad observados en las distintas subpoblaciones seleccionadas. Analizar los malos índices para diferentes atributos en características seleccionadas, y luego identificar segmentos apropiados basados en un rendimiento significativamente diferente.

2. *Segmentación basada en estadísticas*

La clusterización es una técnica ampliamente utilizada para identificar grupos similares entre sí con respecto a las variables de entrada. Puede utilizarse para segmentar bases de datos, colocando objetos en grupos o " clústeres ". Dos de los métodos utilizados para formar clusteres son la clusterización de K-means y los mapas autoorganizados (SOM). También se pueden realizar otros análisis, como la distribución de características dentro de cada clúster, para obtener un conjunto de reglas que definan cada grupo único. La agrupación identifica grupos similares en función de sus características, no de su rendimiento. Por lo tanto, las agrupaciones pueden parecer diferentes, pero pueden tener un rendimiento de riesgo similar. Por lo tanto, las agrupaciones deben analizarse más a fondo para garantizar que la segmentación producida corresponde a grupos con diferentes perfiles de riesgo.

Los árboles de decisión aíslan segmentos basándose en criterios de rendimiento (es decir, diferencian entre "bueno" y "malo"), identifican puntos de ruptura óptimos para cada característica y resultan un método muy potente de segmentación.

1.2.12. Comparación de los beneficios

Los análisis anteriores no cuantifican los beneficios de la segmentación. Por lo tanto, debemos estimar en qué medida la segmentación resulta provechosa. El primer paso consiste en medir la eficacia predictiva de la segmentación. Para ello se pueden utilizar varios estadísticos, como el de Kolmogorov-Smirnov (KS), el estadístico c, etcétera. A continuación, el usuario debe decidir qué nivel de mejora es lo suficientemente significativo como para justificar el esfuerzo adicional de desarrollo y aplicación. La mejor forma de responder a esta pregunta es utilizar los criterios de negocio, no los estadísticos.

1.2.13. Estrategia de trabajo

Existen varias técnicas matemáticas para elaborar modelos de calificación crediticia de predicción de riesgos, por ejemplo, regresión logística, redes neuronales, árboles de decisión, etc. La técnica más adecuada a utilizar puede depender de cuestiones como:

- Calidad de los datos disponibles.
- Tipo de resultado objetivo, es decir, binario (bueno/malo) o continuo (beneficio/pérdida en dólares).
- Tamaño de las muestras disponibles.
- Plataformas de aplicación
- Interpretabilidad de los resultados
- Seguimiento y diagnóstico del rendimiento del modelo de calificación crediticia.

Al final de esta fase, todos los requisitos de datos y la documentación del plan del proyecto están completos, y se puede empezar a trabajar en la construcción de la base de datos.

1.3. Estadístico Kolmogorov-Smirnov (KS)

La prueba de Kolmogorov-Smirnov desempeña un papel crucial en la evaluación del éxito de los modelos de calificación crediticia debido a su capacidad para determinar si dos conjuntos de datos provienen de la misma distribución subyacente. En el contexto de la calificación crediticia, esta prueba se utiliza para comparar la distribución empírica de ciertas variables (como puntajes de crédito) con la distribución esperada o teórica, como la distribución normal estándar.

La importancia de esta prueba radica en su capacidad para detectar desviaciones significativas entre las distribuciones observadas y las distribuciones esperadas. Si un modelo de calificación crediticia está bien ajustado, se esperaría que las distribuciones de los puntajes de crédito de la muestra y la distribución teórica sean similares. Por lo tanto, al aplicar la prueba de Kolmogorov-Smirnov, se puede determinar si el modelo de calificación crediticia está capturando adecuadamente la estructura subyacente de los datos y si proporciona predicciones precisas.

El estadístico de Kolmogorov-Smirnov es fundamental para evaluar la idoneidad de un modelo de calificación crediticia al comparar la distribución de los puntajes de crédito observados con la distribución teórica esperada. Su capacidad para identificar diferencias significativas entre estas distribuciones ayuda a los analistas a mejorar la precisión y la fiabilidad de los modelos de calificación crediticia, lo que a su vez contribuye a la toma de decisiones financieras más informadas y efectivas.

La presente sección se basa en gran medida en el capítulo "VII Stochastic analysis and its applications in statistics", sección 3 y 4 del libro "Mathematical Statistics" escrito por Wiebe R. Pestman [9]. Este libro proporciona una sólida base teórica en estadística matemática, incluyendo una exhaustiva cobertura del Estadístico Kolmogorov-Smirnov y sus propiedades. Las definiciones, teoremas, lemas y demostraciones presentados en esta sección se derivan directamente de los contenidos expuestos en dicho texto. Reconocemos la invaluable exposición de [9] del estadístico Kolmogorov-Smirnov en el contexto de la evaluación de modelos estadísticos. Esta referencia ha sido fundamental para el desarrollo y la fundamentación teórica del estadístico Kolmogorov-Smirnov utilizado en este estudio.

Es necesario introducir notación. Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias definidas en un espacio de probabilidad $(\Omega, \mathfrak{U}, \mathbb{P})$. Las variables de la sucesión son independientes con distribución idéntica F. La función F es no decreciente, continua por la derecha y con límites por la izquierda (esta última propiedad se abrevia en francés como càdlàg y así es comúnmente conocida). Está definida por $F(t) := \mathbb{P}(X_1 \leq t)$. La distribución empírica de la sucesión se define como:

$$\hat{F}(X_1, X_2, \ldots)(x) := \frac{\#\{i \mid i \leq n, X_i \leq x)}{n}$$

Para una función $f : A \subset \mathbb{R} \to \mathbb{R}$ se define la norma del supremo denotada $\|f\|_\infty$ como

$$\|f\|_\infty := \sup_{x \in A} |f(x)|$$

Lema 1.3.1. *Sea $f : (a, b) \to \mathbb{R}$ una función arbitraria creciente. Entonces:*

1. *El conjunto de todos los puntos donde f es discontinua es contable.*
2. *Todo intervalo abierto no vacío (a, b) contiene al menos un punto en el que f es continua.*

Demostración.

1. Si $f : (a, b) \to \mathbb{R}$ es creciente, entonces para cada $x \in (a, b)$ las expresiones $f(x-)$ y $f(x+)$ existen. Entonces el conjunto de todos los puntos de discontinuidad se puede caracterizar como

   $$A := \{x \in (a, b) : f(x-) \neq f(x+)\} = \{x \in (a, b) : f(x-) < f(x+)\}$$

   Después, se selecciona para cada $x \in A$ un numero racional q_x en el intervalo abierto no vacío $(f(x-), f(x+))$. Ahora el mapeo $x \mapsto q_x$ es uno a uno de A hacia $\mathbb{Q}$. Ya que $\mathbb{Q}$ es contable, esto implica que A tambien es contable, probandose asi 1.

2. Un intervalo abierto no vacío siempre contiene un número incontable de elementos. Por esta razón la afirmación 2 es una consecuencia inmediata de la afirmación 1.

□

La demostración del siguiente resultado es directa por lo cual se omitirá.

Lema 1.3.2. *Sea F y G funciones de distribución arbitrarias y sea C un conjunto denso arbitrario en $\mathbb{R}$. Entonces se tiene*

$$\|F - G\|_\infty = \sup_{x \in C} |F(x) - G(x)|.$$

Lema 1.3.3. *Sea F una función de distribución arbitraria. Entonces para todo $\varepsilon > 0$ fijo, el conjunto $\{x : F(x) - F(x-) \geq \varepsilon\}$ es finito.*

Demostración.

Haciendo uso del Lema 1.3.1, el conjunto $\{x : F(x) - F(x-) > 0\}$ es a lo sumo contablemente infinito. Si este conjunto es finito, la afirmación del lema es inmediata. Si no lo es, entonces existe una enumeración $x_1, x_2, \ldots$ del conjunto en cuestión. Para todo n se tiene entonces

$$\sum_{i=1}^{n} (F(x_i) - F(x_i-)) \leq 1.$$

Por lo tanto la serie

$$\sum_{i=1}^{\infty} (F(x_i) - F(x_i-)).$$

es convergente. De ello se deduce que se tiene

$$F(x_i) - F(x_i-) < \varepsilon,$$

para todos los i excepto un numero finito de ellos. □

Lema 1.3.4. *Sea ε cualquier numero positivo y sea F cualquier función de distribución. Si para todos los elementos x en el intervalo (a, b) se tiene*

$$F(x) - F(x-) < \varepsilon,$$

entonces existe una sucesión finita $x_0, x_1, \ldots, x_k$ con las siguientes propiedades:

(i) $a = x_0 < x_1 < \ldots < x_k = b$,

(ii) $F(x_i) - F(x_{i-1}) < \varepsilon$ *para* $i = 1, \ldots, k-1$,

(iii) $F(x_k-) - F(x_{k-1}) < \varepsilon$.

Demostración.

Bajo las premisas del lema, existe para cada elemento $c \in (a, b)$ un numero $\delta(c) > 0$ tal que

$$|F(x) - F(y)| < \varepsilon \text{ para todo } x, y \in (c - \delta(c), c + \delta(c)).$$

Para ver esto, se fija cualquier $c \in \mathbb{R}$. Entonces

$$F(c) - F(c-) = F(c+) - F(c-) < \varepsilon.$$

Se sigue que existe un numero $\delta > 0$, dependiente de c, tal que

$$F(c + \delta) - F(c - \delta) < \varepsilon.$$

Debido a la monotonía de F lo anterior implica:

$$F(y) - F(x) < \varepsilon \text{ para todo } x, y \in (c - \delta, c + \delta).$$

Además, hay un número $\delta(a) > 0$ tal que

$$|F(x) - F(a)| < \varepsilon \text{ para toda } x \in [a, a + \delta(a))$$

y un número $\delta(b) > 0$ tal que

$$|F(b-) - F(x)| < \varepsilon \text{ para toda } x \in (b - \delta(b), b).$$

Debido a que $[a, b]$ es un conjunto compacto, existen numeros $c_1 < \ldots < c_n$ tales que

$$[a, b] \subset [a, a+\tfrac{1}{2}\delta(a)) \cup (c_1 - \tfrac{1}{2}\delta(c_1), c_1 + \tfrac{1}{2}\delta(c_1)) \cup \ldots \cup (c_n - \tfrac{1}{2}\delta(c_n), c_n + \tfrac{1}{2}\delta(c_n)) \cup (b - \tfrac{1}{2}\delta(b), b].$$

A continuación, se define el número $\delta > 0$ como se muestra:

$$\delta := \tfrac{1}{2} \text{mín}\{\delta(a), \delta(c_1), \ldots, \delta(c_n), \delta(b)\}.$$

Ahora bien, si x y y son elementos del intervalo abierto (a, b) y si $|x - y| < \delta$, entonces en los intervalos

$$[a, \delta(a)), (c_1 - \delta(c_1), c_1 + \delta(c_1)), \ldots, (c_n - \delta(c_n), c_n + \delta(c_n)), (b - \delta(b), b).$$

hay por lo menos uno, que contiene a ambos x y y. De esto se deduce que se tiene para tales x y y la siguiente desigualdad:

$$|F(x) - F(y)| < \varepsilon \text{ tan pronto como } |x - y| < \delta.$$

Es fácil ver que cada secuencia

$$a = x_0 < x_1 < \ldots < x_n = b$$

para las cuales $x_i - x_{i-1} < \delta$ satisface las propiedades $(i), (ii)$ y (iii) del lema. □

Teorema 1.3.5. *(V. Glivenko, F.P. Cantelli)*

Sea $X_1, X_2, \ldots$ una muestra infinita de una población con función de distribución F. Entonces

$$\lim_{n\to\infty} \|\hat{F}(X_1, \ldots, X_n) - F\|_\infty = 0$$

fuertemente.

Demostración.

Sea $(\Omega, \mathfrak{U}, \mathbb{P})$ el espacio de probabilidad subyacente de $X_1, X_2, \ldots$. Escogiendo cualquier $\varepsilon > 0$ fija. Se probará que entonces

$$\overline{\lim_{n\to\infty}} \|\hat{F}(X_1, \ldots, X_n) - F\|_\infty \leq \varepsilon \tag{1.3.6}$$

casi seguramente.

Primero, notar que por el Lema 1.3.3, junto con el hecho de que

$$\lim_{a\to-\infty} F(a) = 0 \text{ y } \lim_{a\to+\infty} F(a) = 1,$$

se sigue que existe una secuencia $a_0 < a_1 < \ldots < a_m$ tal que

$$F(a_0) < \varepsilon, \quad F(a_m) > 1 - \varepsilon \tag{1.3.7}$$

y

$$F(x) - F(x-) < \varepsilon \text{ si } x \in (a_{i-1}, a_i) \tag{1.3.8}$$

para $i = 1, \ldots, m$. Para simplificar la notación, se usará $\hat{F}_n(\omega) := \hat{F}(X_1(\omega), \ldots, X_n(\omega))$. Así mismo, se fija

$$\begin{aligned}
\|\hat{F}_n(\omega) - F\|_0 &:= \sup_{x\leq a_0} |\hat{F}_n(\omega)(x) - F(x)|, \\
\|\hat{F}_n(\omega) - F\|_i &:= \sup_{a_{i-1}<x<a_i} |\hat{F}_n(\omega)(x) - F(x)|, \\
\|\hat{F}_n(\omega) - F\|_{m+1} &:= \sup_{x\geq a_m} |\hat{F}_n(\omega)(x) - F(x)|.
\end{aligned}$$

Aplicando el Lema 1.3.2, se verifica facilmente que

$$\|\hat{F}(\omega) - F\|_\infty := \max_{0\leq i\leq m+1} \|\hat{F}(\omega) - F(x)\|_i \tag{1.3.9}$$

Para probar la ecuación (1.3.6), es suficiente mostrar que para cada i hay un conjunto

$A_i \in \mathfrak{U}$ tal que $\mathbb{P}(A_i) = 1$ tal que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_i \leq \varepsilon \ \ \forall \ \ \omega \in A_i$$

Una vez que se ha verificado lo anterior se puede definir lo siguiente

$$A := A_0 \cap A_1 \cap \ldots \cap A_m \cap A_{m+1}.$$

Para este conjunto A se tiene que $\mathbb{P}(A) = 1$ y por (1.3.9)

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_\infty \leq \varepsilon \ \forall \ \ \omega \in A$$

que es lo mismo que en (1.3.6). El resto de la prueba se dividirá en 3 partes.

Parte 1. Existe un conjunto $A_0 \in \mathfrak{U}$ tal que $\mathbb{P}(A_0) = 1$ y tal que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_\infty \leq \varepsilon \ \forall \ \ \omega \in A_0$$

Para ver esto, notar que, por la fuerte ley de los grandes números, existe un conjunto $A_0 \in \mathfrak{U}$ tal que $\mathbb{P}(A_0) = 1$ y tal que

$$\hat{F}_n(\omega)(a_0) \to F(a_0) \ \ \forall \ \ \omega \in A_0 \tag{1.3.10}$$

Ahora, si $x \leq a_0$, se tiene que

$$\hat{F}_n(\omega)(x) - F(x) \leq \hat{F}_n(\omega)(x) \leq \hat{F}_n(\omega)(a_0)$$

y

$$F(x) - \hat{F}_n(\omega)(x) \leq F(x) \leq F(a_0) \leq \varepsilon.$$

Por lo tanto

$$\|\hat{F}(\omega) - F\|_0 \leq \text{máx}(\hat{F}_n(\omega)(a_0), \varepsilon).$$

Por la ecuación (1.3.10), lo anterior implica que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_\infty \leq \varepsilon \ \forall \ \ \omega \in A$$

probandose así la *Parte 1.*

Parte 2. Para cada $i = 1, \ldots, m$ fija, existe un conjunto $A_i \in \mathfrak{U}$ tal que $\mathbb{P}(A_i) = 1$ y tal que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_i \leq \varepsilon \ \forall \ \ \omega \in A_i$$

Para demostrar esta afirmación, se fija cualquier $i = 1, \ldots, m$. Posteriormente, por el Lema 1.3.4 existe una secuencia $x_0, x_1, \ldots, x_k$ tal que $a_{i-1} = x_0 < x_1 < \ldots < x_k = a_i$ y

$$F(x_j) - F(x_{j-1}) < \varepsilon \text{ para } j = 1, \ldots, k-1,$$
$$F(x_k-) - F(x_{k-1}) < \varepsilon.$$

Para cada $j = 0, 1, \ldots, k-1$ fija, por la fuerte ley de los grandes numeros, hay un conjunto $B_j \in \mathfrak{U}$ tal que $\mathbb{P}(B_j) = 1$ y tal que

$$\lim_{n\to\infty} \hat{F}_n(\omega)(x_j) = F(x_j) \ \forall \ \omega \in B_k$$

Fijando

$$A_i := B_0 \cap B_1 \cap \ldots \cap B_k$$

se tiene que $\mathbb{P}(A_i) = 1$. Ahora, para todo n, la funcion $M_n : \Omega \to \mathbb{R}$ esta definida por

$$M_n(\omega) := \text{máx}(|\hat{F}_n(\omega)(x_0) - F(x_0)|, |\hat{F}_n(\omega)(x_1) - F(x_1)|, \ldots, |\hat{F}_n(\omega)(x_{k-1}) - F(x_{k-1})|, |\hat{F}_n(\omega)(x_k-) - F(x_k-)|)$$

Por construccioón, se obtiene que

$$\lim_{n\to\infty} M_n(\omega) = 0 \ \forall \ \omega \in A_i \tag{1.3.11}$$

Resulta que la expresión $\|\hat{F}_n(\omega) - F\|_i$ se puede denominar en terminos de $M_n(\omega)$. Para ver esto, sea x cualquier numero real, estrictamente entre x_{j-1} y x_j. Entonces se tiene por un lado

$$F(x_{j-1}) \leq F(x) \leq F(x_j) \leq F(x_{j-1}) + \varepsilon \tag{1.3.12}$$

y por otro lado

$$\hat{F}_n(\omega)(x_{j-1}) \leq \hat{F}_n(\omega)(x) \leq \hat{F}_n(\omega)(x_j) \tag{1.3.13}$$

En las ecuaciones (1.3.12) y (1.3.13), y tambien en la siguiente expresion abajo, se debe remplazar x_j por x_j- si $j = k$. Las ecuaciones (1.3.12) y (1.3.13) juntas, implican que para toda x estrictamente entre x_{j-1} y x_j se tiene que

$$\hat{F}_n(\omega)(x) - F(x) \leq \hat{F}_n(\omega)(x_j) - F(x_{j-1}) \leq \hat{F}_n(\omega)(x_j) - F(x_j) + \varepsilon$$

y

$$F(x) - \hat{F}_n(\omega)(x) \leq F(x_{j-1}) + \varepsilon - \hat{F}_n(\omega)(x_{j-1}).$$

Las dos desigualdades anteriores pueden resumirse con la afirmacion de que para toda x entre x_{j-1} y x_j se tiene que

$$|\hat{F}_n(\omega)(x) - F(x)| \leq \text{máx}(|\hat{F}_n(\omega)(x_{j-1}) - F(x_{j-1})|, |\hat{F}_n(\omega)(x_j) - F(x_j)|) + \varepsilon$$

Al aplicar el Lema 1.3.2, se sigue que para toda $x \in (a_{i-1}, a_i)$ la siguiente desigualdad se cumple

$$|\hat{F}_n(\omega)(x) - F(x)| \leq \max_{j=0,1,\ldots,k} |\hat{F}_n(\omega)(x_j) - F(x_j)| + \varepsilon = M_n(\omega) + \varepsilon$$

Por lo tanto

$$\|\hat{F}_n(\omega) - F\|_i \leq M_n(\omega) + \varepsilon.$$

Por la ecuación (1.3.11) se llega a la conclusión de que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_i \leq \varepsilon \text{ para toda } \omega \in A_i.$$

probandose así la *Parte 2.*

Parte 3. Existe un conjunto $A_{m+1} \in \mathfrak{U}$ tal que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_{m+1} \leq \varepsilon \text{ para toda } \omega \in A_{m+1}$$

La verificación de esto, es similar a las partes 1 y 2. Al juntar las partes 1, 2 y 3, se sigue que para toda $\varepsilon > 0$ existe un conjunto $A(\varepsilon) \in \mathfrak{U}$ tal que $\mathbb{P}(A(\varepsilon)) = 1$ y tal que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_\infty \leq \varepsilon \text{ para toda } \omega \in A(\varepsilon).$$

Se define el conjunto A como sigue

$$A := \bigcap_{p=1}^{\infty} A(\tfrac{1}{p}).$$

Entonces $A \in \mathfrak{U}, \mathbb{P}(A) = 1$ y por $\omega \in A$ se tiene que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_\infty \leq \tfrac{1}{p} \text{ para toda } p = 1, 2, \ldots.$$

Evidentemente, esto significa que

$$\overline{\lim_{n\to\infty}} \|\hat{F}_n(\omega) - F\|_\infty = 0 \text{ para toda } \omega \in A.$$

Probandose así todas las partes.

□

1.3.1. Definición de estadístico Kolmogorov-Smirnov (KS)

Sea $X_1, X_2, X_3, \ldots, X_n$ una muestra de una población que tiene una función de distribución F. Se quiere comprobar si la función F es (o no) igual a una determinada función de distribución predeterminada F_0. Más precisamente, se desea probar la hipótesis nula

$$H_0 : F = F_0 \tag{1.3.14}$$

contra la hipótesis alternativa

$$H_1 : F \neq F_0 \ . \tag{1.3.15}$$

Como estadístico de prueba, se utilizara la cantidad D_n, definida como

$$D_n = D_n(F_0) := \|\hat{F}(X_1, \ldots, X_n) - F_0\|_\infty \ . \tag{1.3.16}$$

Bajo la hipótesis nula, haciendo uso del Teorema de Glivenko-Cantelli 1.3.5, se tiene que $D_n \to 0$ con probabilidad 1 si $n \to \infty$. Por lo anterior, si H_0 es verdadera, es probable que se obtengan resultados pequeños de D_n. El procedimiento de decisión se basa en las siguientes observaciones:

- Si D_n da resultados pequeños, entonces fallar en rechazar a H_0.
- Si D_n da resultados grandes, entonces rechazar a H_0.

Para precisar qué se entiende por un resultado pequeño (o grande) es necesario conocer la distribución de probabilidad del estadístico de prueba D_n bajo la hipótesis nula. Para esto, el siguiente teorema nos es de ayuda.

Teorema 1.3.17. *La distribución de probabilidad del estadístico D_n es, bajo la hipótesis nula, la misma para todas las funciones de distribución continuas F_0.*

Demostración.

Sea U_0 una función de distribución que pertenece a la distribución uniforme de probabilidad en el intervalo $(0, 1)$. Entonces,

$$U_0(u) = \begin{cases} 0 & \text{si} \quad u \leq 0 \\ u & \text{si} \quad 0 < u < 1 \\ 1 & \text{si} \quad u \geq 1 \end{cases} \tag{1.3.18}$$

Se afirma que siempre que F_0 sea continua, los estadísticos $D_n(F_0)$ y $D_n(U_0)$ están, bajo las hipótesis nulas asociadas a F_0 y U_0, idénticamente distribuidos. Esto se demostrará para el caso especial en el que F_0 es estrictamente creciente. Es decir, se supondrá que

$$x_1 < x_2 \Longrightarrow F_0(x_1) < F_0(x_2).$$

Bajo el supuesto anterior la función $F_0 : \mathbb{R} \to (0, 1)$ es biyectiva y por esta razón existe una función inversa $F_0^{-1} : (0, 1) \to \mathbb{R}$. Ahora bien, si $X_1, \ldots, X_n$ es una muestra de una población que tiene la función de distribución F_0, entonces las variables $F_0(X_1), \ldots, F_0(X_n)$ pueden considerarse como una muestra de una población que tiene la función de distribución uniforme U_0. Para ver esto, se nota que $F_0(X_1), \ldots, F_0(X_n)$ forman un sistema estadísticamente independiente (Proposición I.4.2) y que

$$\mathbb{P}(F_0(X_i) \leq u) = 0 = U_0(u) \quad \text{si} \quad u \leq 0 \ . \tag{1.3.19}$$

Ahora bien

$$F_0(x) \leq u \iff x \leq F_0^{-1}(u),$$

y entonces se tiene que para $0 < u < 1$:

$$\mathbb{P}(F_0(X_i) \leq u) = \mathbb{P}(X_i \leq F_0^{-1}(u)) = F_0(F_0^{-1}(u)) = u = U_0(u) \tag{1.3.20}$$

y para $u \geq 1$

$$\mathbb{P}(F_0(X_i) \leq u) = 1 = U_0(u) \ . \tag{1.3.21}$$

Las funciones de distribución empíricas de la muestra $X_1, \ldots, X_n$ y de la muestra transformada distribuida aparentemente de manera uniforme $F_0(X_1), \ldots, F_0(X_n)$ están relacionadas como sigue. Para $0 < u < 1$ se tiene lo siguiente

$$\hat{F}(F_0(X_1),\ldots,F_0(X_n))(u) = \frac{\#\{i:F_0(X_i)\leq u\}}{n} = \frac{\#\{i:X_i\leq F_0^{-1}(u)\}}{n} = \hat{F}(X_1,\ldots,X_n)(F_0^{-1}(u))$$

y es fácil checar que para $u \leq 0$ o $u \geq 1$ se tiene que

$$\hat{F}(F_0(X_1),\ldots,F_0(X_n))(u) = U_0(u).$$

Para finalizar la prueba, tenemos lo siguiente:

$$\begin{aligned}
D_n(U_0) &= \|\hat{F}(F_0(X_1),\ldots,F_0(X_n)) - U_0\|_\infty \\
&= \sup_{u\in\mathbb{R}} |\hat{F}(F_0(X_1),\ldots,F_0(X_n))(u) - U_0(u)| \\
&= \sup_{u\in(0,1)} |\hat{F}(F_0(X_1),\ldots,F_0(X_n))(u) - U_0(u)| \\
&= \sup_{u\in(0,1)} |\hat{F}(F_0(X_1),\ldots,F_0(X_n))(u) - u| \\
&= \sup_{u\in(0,1)} |\hat{F}(X_1,\ldots,X_n)(F_0^{-1}(u)) - F_0(F_0^{-1}(u))| \\
&= \sup_{x\in\mathbb{R}} |\hat{F}(X_1,\ldots,X_n)(x) - F_0(x)| \\
&= \|\hat{F}(X_1,\ldots,X_n) - F_0\|_\infty = D_n(F_0)
\end{aligned}$$

Probandose asi el teorema. □

El estadístico de Kolmogorov-Smirnov, denotado como D_n, rinde homenaje a dos influyentes matemáticos rusos: Andrej N. Kolmogorov (1903-1987) y Nikolai V. Smirnov (1900-1966). Kolmogorov, un destacado teórico de la probabilidad, realizó contribuciones fundamentales a diversos campos matemáticos, incluida la teoría de la medida, la teoría de la probabilidad y la mecánica cuántica. Sus trabajos sentaron las bases para el desarrollo de la teoría moderna de la probabilidad y la estadística. Por otro lado, Smirnov fue un destacado matemático, conocido por su trabajo en teoría de la probabilidad y estadística matemática. Sus contribuciones incluyen el desarrollo de métodos estadísticos y la aplicación de la teoría de la probabilidad en diversos contextos prácticos. La denominación conjunta del estadístico de prueba Kolmogorov-Smirnov reconoce la importancia y el legado de ambos matemáticos en el campo de la estadística y la probabilidad.

1.3.1.0.1 Análisis del Rendimiento de Modelos Predictivos Utilizando scorecardpy: Enfoque en el Estadístico de Kolmogorov-Smirnov

El código Python implementado en esta tesis utiliza la biblioteca *scorecardpy* para evaluar el rendimiento de un modelo predictivo. En este proceso de evaluación, intervienen varios componentes clave. En primer lugar, se importa la biblioteca scorecardpy y se invoca la función $perf\_eva$ para evaluar el rendimiento del modelo. Dentro de esta llamada a la función, se pasan varios parámetros para habilitar el proceso de evaluación. Uno de estos parámetros representa los datos de prueba, que contienen los resultados correspondientes al conjunto de datos de prueba, proporcionando una base de comparación con las predicciones del modelo.

Otro parámetro crucial es el subconjunto de probabilidades predichas generadas por el modelo para la clase positiva. Al seleccionar las probabilidades asociadas con la clase positiva, la evaluación puede centrarse en la capacidad del modelo para discriminar entre instancias positivas y negativas dentro del conjunto de datos de prueba.

Como parte del proceso de evaluación, se extrae el estadístico de Kolmogorov-Smirnov (KS) de los resultados de la evaluación. El estadístico de Kolmogorov-Smirnov cuantifica la diferencia máxima entre las funciones de distribución acumulativa de las etiquetas verdaderas y las probabilidades predichas por los modelos de clasificación. Al examinar esta estadística, se obtiene información sobre la capacidad del modelo para distinguir entre diferentes clases dentro del conjunto de datos de prueba.

De esta manera, se orquesta la evaluación del rendimiento de un modelo predictivo, aprovechando las funcionalidades ofrecidas por la biblioteca scorecardpy y centrándose en el estadístico de Kolmogorov-Smirnov.

1.4. Descripción de Datos y Lógica de Negocio

Para llevar a cabo la presente investigación, se estableció una colaboración estratégica con un banco digital cuyo nombre se mantendrá confidencial y se referenciará como abc. A través de esta colaboración, se obtuvo acceso autorizado a su base de datos, lo que permitió realizar un análisis exhaustivo de las operaciones financieras y datos relevantes asociados a la gestión de riesgo crediticio. La disponibilidad de esta información directa

proveniente de la plataforma digital del banco ha sido fundamental para profundizar en el estudio de los patrones transaccionales, la evaluación crediticia basada en algoritmos de aprendizaje de maquina, y otros aspectos cruciales que configuran el marco de este trabajo de tesis. Este acceso privilegiado a los datos del banco digital ha brindado una perspectiva detallada y precisa, garantizando la calidad y relevancia de los hallazgos presentados en el desarrollo de la investigación. Cabe destacar que el manejo de dicha información se realizó de conformidad con los protocolos de privacidad y seguridad establecidos por ambas partes, asegurando la confidencialidad y la integridad de los datos durante todo el proceso de investigación.

En esta sección, exploraremos detalladamente los datos y operaciones financieras que caracterizan la plataforma digital del banco. Esto incluye la apertura de cuentas, el manejo de fondos, las opciones de pago y las distintas modalidades de préstamos disponibles. Se realizará un análisis basado en datos reales de abc, cuyo principal canal de interacción con los usuarios es a través de una aplicación móvil. Este banco se especializa en la provisión de préstamos personales de corto plazo, entre otros servicios financieros. Cabe destacar que el importe máximo de préstamo ofrecido por dicha aplicación asciende a 300,000 nairas y el importe mínimo es de 3,000 nairas.

1.4.1. Creación de Cuentas

La creación de una cuenta en la aplicación del banco digital requiere la provisión de datos indispensables, tales como número telefónico, dirección de correo electrónico, nombre completo, fecha de nacimiento, BVN (número de verificación bancaria) y género. Posteriormente, estos datos son sometidos a un proceso de revisión, seguido de un breve periodo de espera para la verificación. Una vez que la solicitud es aprobada, la cuenta abc se encuentra operativa, generando simultáneamente una cuenta de banco abc asociada al usuario. La interacción del usuario con esta cuenta de banco se realiza a través de la aplicación abc, específicamente en la sección denominada "monedero", donde se pueden llevar a cabo diversas transacciones.

La información suministrada durante este proceso se almacena en la tabla "Usuario", asignándose un identificador único (ID) para este usuario, junto con todos sus datos. Una vez que la información ha sido verificada satisfactoriamente, se genera un ID adicional en

la tabla "Cuenta", específicamente para el usuario en cuestión.

1.4.2. Añadir Fondos a la Cuenta abc

El banco abc ofrece diversas opciones para cargar fondos. Cada vez que se incorpora una nueva cuenta bancaria, es necesario verificar la información mediante una fotografía del rostro del usuario. Al agregar una cuenta de banco externa, el banco abc realiza un análisis estadístico exhaustivo de la información proveniente de todas las cuentas bancarias proporcionadas por el usuario. Los resultados de este análisis se reflejan en la tabla "Características del riesgo de crédito". Con cada adición de cuenta bancaria, se asigna un identificador único (ID) en la tabla "Características del riesgo de crédito", indicando que se ha llevado a cabo un análisis con la información acumulada hasta ese momento, proveniente de todas las cuentas vinculadas con la aplicación.

La información empleada para este estudio comprende el historial de transacciones existentes en una o varias cuentas de banco vinculadas con el banco abc. Este enfoque permite evaluar de manera integral el comportamiento financiero del usuario, contribuyendo así a la evaluación y gestión eficaz del riesgo crediticio.

1.4.3. Añadir dinero con tarjeta

Se puede añadir dinero a la cuenta abc fácilmente vinculando una tarjeta de débito de algún banco externo. Si el usuario tiene una cuenta en otro banco, se puede añadir esa tarjeta a la cuenta abc. Se puede vincular más de una tarjeta a la cuenta abc.

1.4.4. Añadir dinero a través del número de cuenta del banco digital

Para añadir o recibir dinero usando la cuenta bancaria abc, el usuario necesita compartir el número de cuenta virtual que se genera en su perfil. Con el número de cuenta virtual se puede iniciar una transferencia desde cualquier banco a la cuenta abc utilizando su número de cuenta.

1.4.5. Añadir dinero mediante peticiones de dinero P2P

El usuario puede recibir o enviar una solicitud de dinero entre sus contactos en el banco abc. Se puede mandar una petición de cierta cantidad de dinero a sus contactos que tengan cuenta en el banco abc, una vez que el contacto haya aceptado o declinado la petición, el usuario será notificado y el dinero será enviado o recibido en la cuenta.

1.4.6. Pagos en la plataforma del banco

Cada modalidad de pago se registrará en la tabla "Transacción", incluyendo referencias al canal correspondiente (ID para el tipo de transacción) y al ID del usuario que llevó a cabo dicha transacción, junto con otros detalles informativos.

1.4.7. Envío de dinero a otros bancos

Se puede transferir dinero a cualquier persona que tenga una cuenta bancaria válida. Es necesario conocer el nombre de la persona a la que se envía el dinero y su número de cuenta, para poder confirmarlo antes de enviar el dinero.

1.4.8. Retirar dinero de la cuenta abc

Se necesita haber vinculado una tarjeta bancaria a la cuenta abc para posteriormente retirar dinero de la cuenta de banco abc hacia una cuenta de banco externa.

1.4.9. Envío de dinero a personas fuera de abc

Si se requiere enviar dinero a alguien que no tiene abc, y que no tiene cuenta de banco, todo lo que se necesita es su número de teléfono. El destinatario recibirá una notificación por SMS en la que se le informa que su contacto le ha enviado dinero. El receptor deberá descargar abc para acceder al dinero que se le ha enviado.

1.4.10. Pago de recibos

El servicio Bill Payment permite pagar cualquier factura desde la cuenta del banco digital. El usuario puede agregar un comerciante y pagar su Internet, tiempo aire, e incluso suscripciones de TV.

1.4.11. Préstamos desde la plataforma

La totalidad de la información concerniente a los préstamos se almacena en la tabla "Préstamo", la cual cuenta con un identificador (ID) único para cada préstamo, facilitando su posterior referencia y seguimiento. Esta tabla contiene todos los detalles esenciales sobre los préstamos. Por otro lado, en la tabla "Transacción" se reflejarán todos los movimientos que realice el usuario en relación con su préstamo. Esta información incluirá referencias al canal correspondiente, al ID del préstamo, al ID del usuario, y otros datos relevantes.

1.4.12. ¿Que son los préstamos en el banco abc?

abcLoan es el tipo de préstamo que otorga el banco digital abc, el dinero del préstamo aparece recargado en el monedero abc. Puede servir para pagar facturas dentro de la aplicación o retirar el dinero a una cuenta bancaria externa.

1.4.12.1. ¿Cómo solicitar un préstamo abc?

Antes de solicitar un préstamo, es requisito contar con una cuenta bancaria externa vinculada al banco abc, y se debe verificar la información mediante una foto del rostro del usuario. Una vez validada esta información, el usuario puede seleccionar el monto que la aplicación sugiere como oferta tentativa para el préstamo. En algunos casos, es posible que el usuario no sea elegible y no reciba ninguna oferta.

La solicitud de préstamo pasa por un proceso de revisión, y una vez aprobada, se notifica al usuario. Si el préstamo es aprobado, el monto correspondiente aparecerá en su cuenta abc en el monedero. El proceso de aprobación puede tardar entre 24 y 48 horas desde la vinculación de la cuenta bancaria.

Cuando el usuario solicita un préstamo, el banco abc emplea un algoritmo de aprendizaje

de maquina para evaluar al usuario y calcular el monto definitivo que se le puede ofrecer. Este algoritmo utiliza el historial de transacciones en las cuentas de banco vinculadas con el banco.

El monto del préstamo concedido se determina según la puntuación crediticia del usuario, calculada considerando el historial de transacciones asociado a la cuenta bancaria vinculada al monedero en la aplicación. Es esencial destacar que cada usuario solo puede acceder al importe máximo de préstamo visualizado en la oferta de préstamo. Si se desea solicitar un monto mayor, es necesario fortalecer la puntuación crediticia cumpliendo con el cronograma de pagos establecido.

1.4.12.2. ¿Qué ocurre si el usuario paga de más?

El usuario tiene la posibilidad de efectuar pagos precisos de su préstamo al realizar la transacción desde su aplicación. Sin embargo, en caso de realizar un pago excesivo a la cuenta abc, el excedente se abonará automáticamente en su monedero.

1.4.12.3. Datos importantes acerca de los prestamos

- Al usuario se le muestran sus pagos programados en la aplicación para que este al tanto de las fechas.
- Un usuario no puede tener más de un préstamo activo a la vez.

1.4.13. Descripción de Datos

En el marco de esta investigación, se trabajará con diversas tablas, tales como "Transacción", "Canales", "Préstamo", "Usuario", "Características del Riesgo de Crédito", "Cuentas de Banco" y "Cuentas". Estas tablas proporcionarán la base de datos necesaria para el análisis y la comprensión de los procesos y factores involucrados en la gestión crediticia de este banco digital.

1.4.13.1. Tabla de Transacciones

La tabla de transacciones constituye un registro detallado de los datos vinculados a las operaciones financieras realizadas por los clientes. Esta recopilación abarca tanto

información general acerca de las transacciones como detalles específicos de cada una. Entre los campos contenidos en esta tabla, se encuentran:

- Id: Un número único o código de identificación asignado a cada transacción, facilitando una referencia clara para consultas posteriores.
- Id_canal: Este campo representa un número único o código de identificación asociado a cada tipo de transacción, vinculándose con los diversos productos ofrecidos por el banco. Funciona como una llave foránea que hace referencia a la tabla "Canales".
- $Estado$: Indica el estado actual de la transacción, pudiendo ser pendiente, en proceso, completada, cancelada o fallida.
- $Cantidad$: Refiere a la cantidad de dinero involucrada en la transacción, expresada en la moneda específica correspondiente.
- $Divisa$: Proporciona información sobre la divisa utilizada en la transacción, determinando la unidad monetaria en la cual se efectúa.
- Id_emisor: Un número único que se relaciona con el identificador del usuario que realiza la transacción, actuando como una llave foránea que referencia la tabla "Usuarios".
- $Id\_receptor$: Similar al campo anterior, este número único se asocia al identificador del usuario receptor de la transacción. También es una llave foránea que referencia la tabla "Usuarios".
- $Id\_externo\_1$, $Id\_externo\_2$, $Id\_externo\_3$, $Id\_externo\_4$, $Id\_externo\_5$, $Id\_externo\_6$: Estos campos almacenan números únicos que hacen referencia a cuentas bancarias, negocios o entidades externas relacionadas con la generación de la transacción. El significado y uso de estas columnas dependen del Id_canal (tipo de transacción) en particular, pudiendo contener valores nulos en casos específicos.
- $Id\_prestamo$: Un número único relacionado con el préstamo asociado a la transacción, en caso de aplicar. Funciona como una llave foránea que hace referencia a la tabla "Préstamos".

1.4.13.2. Tabla de Préstamos

Dentro de la tabla de Préstamos, se registran datos esenciales relacionados con los préstamos ofrecidos, concedidos o rechazados a los clientes. Esta estructura de datos proporciona un marco valioso para el análisis y seguimiento detallado de los préstamos otorgados por el banco digital, permitiendo una comprensión más profunda de los aspectos financieros y temporales asociados con cada préstamo en particular.

La tabla de Préstamos contiene información específica acerca de cada préstamo, así como su estado actual. Entre los campos destacados en esta tabla se encuentran:

- Id: Un número único o código de identificación asignado a cada préstamo, facilitando su referencia y seguimiento posterior.
- $Id\_usuario$: Este campo constituye un número único o código de identificación que hace referencia al usuario de la aplicación móvil que ha solicitado el préstamo. Actúa como una llave foránea vinculada con la tabla "Usuarios".
- $Creacion$: La fecha y hora en las cuales el usuario realizó su solicitud de préstamo por primera vez. En este punto, aún no se ha evaluado el historial crediticio del usuario.
- $Fecha\_inicio$: La fecha en la cual el usuario recibió los fondos del préstamo en su cuenta bancaria dentro de la aplicación.
- $Fecha\_final$: Una fecha teórica que indica cuándo se espera que el préstamo sea pagado en su totalidad, calculada sumando la fecha de inicio con la duración establecida para el préstamo.
- $Estado\_prestamo$: Un indicador que refleja el estado actual del préstamo, pudiendo ser activo, aprobado, pagado, rechazado, en proceso, pendiente de aprobación o cancelado.
- $Fecha\_pagado$: La fecha en la cual el préstamo ha sido pagado en su totalidad, incluyendo los intereses generados. Este campo solo aparecerá para aquellos préstamos que se encuentren en estado pagado.
- $Oferta$: Almacena información en formato JSON acerca de la cantidad máxima de

dinero ofrecida al usuario como préstamo, una vez que se ha evaluado su historial crediticio.

- *Cantidad*: Refiere a la cantidad de dinero en nairas que el usuario ha tomado prestado, la cual puede ser menor que la oferta inicial realizada, y no incluye los intereses que el usuario debe pagar.
- *Tasa_interes*: Representa la tasa de interés acordada para el préstamo.
- *Id_cuenta*: Un número único o código de identificación relacionado con la cuenta bancaria o de ahorros del usuario dentro de la aplicación, actuando como una llave foránea vinculada con la tabla "Cuenta".
- *Ultima_modificacion*: La fecha del último cambio realizado en el indicador de *Estado_prestamo*.

1.4.13.3. Tabla de Usuarios

La tabla de Usuarios almacena de manera integral la información general correspondiente a cada usuario registrado en la aplicación móvil, independientemente de si han solicitado un préstamo o vinculado alguna cuenta bancaria a la aplicación. Esta tabla abarca datos personales, información de contacto y detalles relacionados con las cuentas bancarias de los usuarios. Proporciona una estructura robusta para el almacenamiento y gestión de la información personal y de contacto, así como los detalles asociados con las cuentas bancarias. Este conjunto de datos es esencial para realizar un análisis exhaustivo de los perfiles de los usuarios y asegurar la integridad y seguridad de la información en la aplicación móvil.

- Id: Número único o código de identificación asignado a cada usuario con el propósito de facilitar su referencia y seguimiento posterior.
- *Codigo_pais*: Texto que indica el país de origen del usuario.
- *Creado*: Fecha y hora en las cuales el usuario fue creado en la aplicación.
- *Msisdn*: Número de teléfono celular al cual se envía un código de verificación cuando el usuario ingresa por primera vez a la aplicación.
- *Id_nacional*: Código numérico, un BVN (número de verificación bancaria) que ofrece

una identidad única abarcando todo el sector bancario nigeriano para facilitar la identificación y verificación en diversas operaciones bancarias. Utiliza características biométricas y números de identificación personal.

- $Verificacion\_estado$: Refleja el estado de verificación de los datos personales del usuario mediante la base de datos gubernamental del país.
- $Genero$: Referencia al sexo biológico de la persona.
- DOB: Fecha de nacimiento de la persona.

1.4.13.4. Tabla de Cuentas Bancarias

La presente tabla almacena información relacionada con las cuentas bancarias ofrecidas por el banco digital dentro de su aplicación. La tabla de Cuentas proporciona una estructura organizada para almacenar y gestionar los detalles de las cuentas bancarias ofrecidas por el banco digital en su aplicación móvil. Este conjunto de datos es esencial para garantizar la integridad y la seguridad en las operaciones financieras realizadas por los usuarios dentro de la plataforma.

- Id: Número único o código de identificación asignado a cada cuenta bancaria de los usuarios que han sido aprobados y verificados. Es importante destacar que un usuario de la aplicación puede poseer múltiples tipos de cuentas, como cuentas de ahorro y cuentas transaccionales. Actualmente, la aplicación solo hace uso de las cuentas transaccionales para los usuarios.
- $Id\_usuario$: Número único o código de identificación que se vincula con el usuario de la aplicación móvil al cual pertenece la cuenta bancaria. Es una llave foránea que se enlaza con la tabla "Usuarios".
- $Numero\_cuenta$: Representa un número de cuenta perteneciente a un banco externo, utilizado para realizar transferencias interbancarias entre la cuenta de este banco digital y la cuenta externa de otro banco.
- Id_tipo_cuenta: Número entero que indica el nivel de privilegios que un usuario de la aplicación posee en su cuenta bancaria, estableciendo así diferentes niveles de acceso y funcionalidades.

- *Ultima_modificacion*: Fecha y hora de la última modificación realizada en el campo *Id_tipo_cuenta*.

1.4.13.5. Características del riesgo de crédito

En esta tabla se detalla el cálculo de parámetros estadísticos que describen el historial de transacciones de las cuentas bancarias que el usuario ha agregado y vinculado con su cuenta de banco. Cada vez que un usuario vincula una cuenta de banco externa, el banco ejecuta un algoritmo que realiza un estudio estadístico llenando esta tabla con la información acumulada sobre las cuentas de banco del usuario.

- *Id*: Número único o código de identificación que indica que se ha ejecutado el algoritmo para evaluar el historial de transacciones del usuario específico.
- *Id_usuario*: Número único o código de identificación que hace referencia al usuario de la aplicación móvil al cual pertenece la cuenta bancaria externa analizada. Es una llave foránea que se vincula con la tabla "Usuarios".
- *Id_prestamo*: Número único o código de identificación asignado a cada préstamo, facilitando su referencia y seguimiento. Hace referencia al id del último préstamo del usuario al que corresponde el análisis estadístico, pudiendo tener valor nulo. Es una llave foránea vinculada con la tabla "Prestamo".
- *Tx_encontradas*: Cantidad de transacciones acumuladas en todas las cuentas de banco que el usuario ha vinculado en la aplicación abc.
- *Edad_actividad*: Fecha más antigua de las transacciones existentes en las cuentas de banco del usuario.
- *Mediana_balance*: Mediana del balance que tiene el usuario en sus cuentas de banco a lo largo del tiempo.
- *Balance_promedio*: Promedio del balance que tiene el usuario en sus cuentas de banco a lo largo del tiempo.
- *Balance_Maximo*: Balance máximo que tiene el usuario en sus cuentas de banco a lo largo del tiempo.
- *Balance_Minimo*: Balance mínimo que tiene el usuario en sus cuentas de banco a

lo largo del tiempo.

- *Salario_mensual_estimado*: Cantidad estimada de salario mensual que el usuario percibe según sus transacciones bancarias.
- *Dias_edad_actividad*: Cantidad de días entre el valor de la columna *Edad_actividad* y el día en que el usuario agregó la cuenta de banco.
- *Exito_parametros*: Parámetro booleano que indica si se logró calcular todos los parámetros de esta tabla. Razones para no calcular los parámetros se detallan en el campo Razón, como datos incorrectos o menos de 90 días de información de transacciones bancarias.
- *Razon*: Explicación de por qué no se pudieron calcular los parámetros de esta tabla, en caso de que *Exito_parametros* sea false.
- $Dispositivos_{a}ctivos$: Cantidad de dispositivos (celulares) con los que el usuario utiliza la aplicación.
- *Cantidad_maxima*: Monto de la transacción más grande registrada entre todas las cuentas de banco del usuario.
- $Oferta$: Dato calculado por el banco que indica si se le podría hacer una oferta de préstamo al usuario.
- *Creado*: Fecha y hora de la creación de los parámetros de esta tabla para la información de las cuentas de banco agregadas por el usuario.
- Versión: Indica la versión del código utilizada para el cálculo de los parámetros estadísticos de esta tabla.
- *Numero_prestamo*: Cantidad de préstamos que ha solicitado el usuario al banco.
- *Codigo_pais*: País de origen del usuario.

La infraestructura de datos presentada mediante las tablas, en la aplicación de abc proporciona un robusto sistema para gestionar y analizar la información relacionada con los usuarios, sus transacciones bancarias, préstamos solicitados y parámetros estadísticos derivados. En conjunto, estas tablas forman un sistema integral que permite al banco ofrecer servicios financieros personalizados, garantizando la seguridad, integridad y eficiencia en

la gestión de datos. La aplicación se posiciona así como una herramienta valiosa para los usuarios que buscan una experiencia financiera completa y adaptada a sus necesidades.

Capítulo 2

Exploración y Evaluación de Modelos de Clasificación

En este capítulo hacemos un análisis extenso de modelación de datos. Exploramos los siguientes aspectos: Definimos tres diferentes variables a explicar Y que son dicotomicas y para las que el conjunto $Y = 1$ representan la presencia de un evento crediticio. Definimos tres diferentes conjuntos de variables X que representan selecciones de todas las variables disponibles para explicar a la variable Y. Lo anterior da un total de nueve combinaciones. Para cada una de estas combinaciones exploramos el ajuste de los siguientes modelos de aprendizaje de maquina: KNN, Naive Bayes, LDA, SVM, Regresión Logistica, Bosques aleatorios y XGBoost. Las nueve combinaciones en conjunto con los siete modelos referidos da un total de sesenta y tres ejercicios de modelación. A su vez cada uno de estos se repetirán cien veces para controlar por la variabilidad en los datos ya que se hace una partición aleatoria 70-30 para datos de entrenamiento y validación. Para controlar el desempeño de cada ejercicio se estiman las métricas: AUC, componentes de la diagonal de la matriz de confusión y métrica basada en el estadístico de Kolmogorov-Smirnov.

En este capítulo nos adentramos en el estudio de distintos modelos en la predicción del riesgo crediticio. Este análisis se lleva a cabo con el objetivo de explorar y comparar el rendimiento de diversos modelos, así como evaluar la influencia de diferentes variables y criterios de evaluación en el resultado final de la clasificación crediticia.

2.1. Exploración Estadística

El análisis detallado de las variables provenientes de la información proporcionada por el banco abc ha sido un componente esencial en nuestra investigación. La aplicación de técnicas de estadística descriptiva nos ha permitido examinar la distribución y las propiedades fundamentales de las variables clave, proporcionando una visión holística de los datos. Al emplear medidas resumen, como la media, la mediana y la desviación estándar, hemos obtenido una comprensión más profunda de la tendencia central, la dispersión y la variabilidad de las variables analizadas.

Además, hemos llevado a cabo un análisis gráfico exhaustivo para visualizar de manera efectiva los patrones y las relaciones entre las variables. El uso de gráficos, como histogramas, diagramas de dispersión y diagramas de caja, ha facilitado la identificación de posibles tendencias, patrones atípicos y la correlación entre variables. Estas representaciones visuales han enriquecido nuestra capacidad para interpretar los datos y comunicar hallazgos de manera clara y efectiva.

El proceso de análisis también incluyó la exploración de variables mediante la segmentación y comparación de subconjuntos de datos. Esto nos ha permitido identificar posibles disparidades y patrones específicos en diferentes categorías, lo que contribuye a una comprensión más completa y contextualizada de la información proporcionada por el banco abc.

A través del empleo de técnicas estadísticas descriptivas y herramientas visuales, hemos llevado a cabo un análisis minucioso de las variables derivadas de los datos del banco abc. Este enfoque nos ha proporcionado una base sólida para extraer conclusiones significativas y tomar decisiones informadas en el desarrollo de nuestra investigación.

2.2. Preparación Estratégica de Datos

En el proceso de preparación de los datos proporcionados por el banco abc, se llevó a cabo una exhaustiva revisión para asegurar la calidad y coherencia de la información. Esto incluyó la identificación y manejo de valores atípicos, la corrección de datos faltantes y la estandarización de formatos en todas las tablas relacionadas con cuentas bancarias,

usuarios y préstamos.

Además, se procedió a la creación de variables nuevas que se consideraron fundamentales para abordar los objetivos específicos de la investigación. Estas variables se diseñaron de manera estratégica para capturar aspectos relevantes del comportamiento financiero de los usuarios y facilitar análisis posteriores.

La transformación de variables numéricas en cualitativas fue una parte esencial del proceso de preparación de datos. Se utilizaron técnicas como la discretización para convertir variables continuas en categóricas, permitiendo así una mejor interpretación y análisis de los resultados.

En el caso de variables categoricas que requerían representación numérica, se implementó la técnica de "one-hot encoding". Esta estrategia convierte variables categóricas en un arreglo de 0 y 1, facilitando la incorporación de información cualitativa en modelos de aprendizaje automático.

El proceso de preparación de datos no solo se centró en la corrección y transformación de variables existentes, sino también en la creación de nuevas características que enriquecieron la perspectiva de la información. Este enfoque proactivo garantizó que los datos entregados al modelo estuvieran optimizados para la generación de conocimientos significativos y la toma de decisiones informadas.

2.3. Experimentación con Variables Crediticias

Durante esta etapa, se experimenta con diversas definiciones de la variable a explicar. Se consideran los siguientes tres eventos crediticios:

- *ydeudor* denota el estado de un préstamo si es que está completamente pagado o no, independientemente de su historial de morosidad.
- *yimpuntual* indica si es que existió morosidad en los pagos programados o no, independientemente de si el usuario se puso al corriente en sus pagos extemporáneamente o no.

- *yrecuperacionbaja* denota si el usuario ha efectuado un pago inferior al esperado en comparación con el préstamo concedido.

2.3.1. Volumen de la información basado en las Variables Crediticias

Se cuenta con un conjunto de datos que comprende 27,349 observaciones correspondientes al año 2022 de las operaciones del banco digital del cual se obtuvo acceso a los datos. Inicialmente, la cantidad de datos disponibles se redujo a 27,349 después de aplicar procesos de formateo, limpieza y transformación de variables.

A continuación se presenta una tabla que muestra tanto la cantidad absoluta de observaciones para cada variable crediticia, como su proporción respecto al total inicial de observaciones mencionado.

Variables Crediticias	Cantidad Absoluta	Cantidad Relativa 1	Cantidad Relativa 0
yimpuntual	13,864	0.5069	0.4931
yrecuperacionbaja	1,843	0.0673	0.9327
ydeudor	1,933	0.0706	0.9294

2.4. Exploración de Combinaciones de Variables Explicativas

Se lleva a cabo un análisis exhaustivo experimentando con diversas combinaciones de variables explicativas, explorando cómo estas influencian el rendimiento de los modelos predictivos. A continuacion se describen todas las posibles variables explicativas de nuestro conjunto de datos:

- $loan\_amount$: Se refiere al monto total del prestamo otorgado al cliente.
- $number\_of\_payments$: Esta variable hace referencia a la cantidad de cuotas programadas que el cliente deberá abonar para saldar su préstamo.
- $term$: Esta variable indica la duración total del préstamo en días.

- *interest_rate*: Representa la tasa de interés que se aplicará al préstamo otorgado.
- *tx_found*: La cantidad de transacciones registradas en las cuentas bancarias externas vinculadas a la aplicación.
- *max_amount*: El máximo monto de dinero que ha sido transaccionado en las cuentas bancarias externas asociadas a la aplicación.
- *activity_age_days*: Representa la cantidad de días transcurridos desde la última transacción registrada en las cuentas bancarias externas asociadas a la aplicación.
- *estimated_monthly_income*: Es una aproximación del ingreso mensual del usuario.
- *average_balance*: Es el saldo promedio de las cuentas bancarias externas vinculadas a la aplicación.
- *min_balance*: Es el saldo mínimo encontrado en las cuentas bancarias externas vinculadas a la aplicación.
- *max_balance*: Es el saldo máximo encontrado en las cuentas bancarias externas vinculadas a la aplicación.
- *loan_number*: Representa el número de préstamos concedidos al usuario a través de la aplicación móvil. Es un número entero.
- *max_weekly_payment*: Es una estimación del máximo que el usuario podría pagar semanalmente.
- *gender*: Indica el genero del usuario, toma el valor de 1 si el usuario es femenino y 0 si es masculino.
- *edad*1: Esta variable toma el valor de 1 si el usuario tiene entre 18 y 25 años de edad, de lo contrario, su valor es 0.
- *edad*2: Esta variable toma el valor de 1 si el usuario tiene entre 26 y 35 años de edad, de lo contrario, su valor es 0.
- *edad*3: Esta variable toma el valor de 1 si el usuario tiene entre 36 y 45 años de edad, de lo contrario, su valor es 0.
- *edad*4: Esta variable toma el valor de 1 si el usuario tiene entre 46 y 55 años de

edad, de lo contrario, su valor es 0.

- *edad*5: Esta variable toma el valor de 1 si el usuario tiene entre 56 y 65 años de edad, de lo contrario, su valor es 0.

- *edad*6: Esta variable toma el valor de 1 si el usuario tiene mas de 65 años de edad, de lo contrario, su valor es 0.

Variables Explicativas	Tipo Variable	Min	Max	Promedio
$loan\_amount$	Real positiva	2,000.0	300,000.0	33,516.7
$number\_of\_payments$	Entera positiva	1	12	1.8
$term$	Entera positiva	28	92	28.3
$interest\_rate$	Real positiva	5	23	12.6
tx_found	Entera positiva	1	98,807	1,223.8
max_amount	Real positiva	20	300,000.0	23,368.03
$activity\_age\_days$	Entera positiva	89	4,525	407.7
$estimated\_monthly\_income$	Real positiva	0	999,983.0	55,415.8
$average\_balance$	Real positiva	0	999,973.72	18,538.09
$min\_balance$	Real positiva	0	987,214.0	7,887.28
$max\_balance$	Real positiva	0	2.352954657E7	88,840.99
$loan\_number$	Entera positiva	1	47	3.24

Cuadro 2.4.1: Tabla descriptiva de las variables explicativas utilizadas en este trabajo.

2.4.1. Conjuntos que agrupan variables explicativas

En los modelos y análisis que se presentan en las secciones subsiguientes, se emplean las variables *opcion*1, *opcion*2 y *opcion*3, las cuales son conjuntos que agrupan variables explicativas. Esto facilita la gestión y organización de este conjunto de datos. A continuación, se detalla el contenido de cada variable:

- Para este primer conjunto se utilizaron todas las variables explicativas disponibles.
  opcion1: $loan\_amount, number\_of\_payments, interest\_rate, tx\_found,$ $max\_amount, activity\_age\_days, estimated\_monthly\_income, average\_balance,$ $min\_balance, max\_balance, loan\_number, gender, cocienteingresodeuda, edad1,$ $edad2, edad3, edad4, edad5.$

- En este conjunto se agruparon las características del individuo.
  opcion2: $loan\_amount, number\_of\_payments, interest\_rate, activity\_age\_days,$ $estimated\_monthly\_income, average\_balance, loan\_number, gender,$ $cocienteingresodeuda, edad1, edad2, edad3.$

- En este conjunto se agruparon las caracteríscas de crédito y cuentas bancarias, sin utilizar informacion sociodemográfica.
  opcion3: $loan\_amount, number\_of\_payments, interest\_rate, tx\_found, max\_amount, activity\_age\_days, average\_balance, min\_balance, max\_balance, loan\_number$

2.5. Evaluación de Modelos

Se experimenta con una variedad de modelos de clasificación utilizando herramientas disponibles en Python:

- XGBoost (xgb) XGBoost es un algoritmo de ensamblaje que utiliza árboles de decisión. Destaca por su eficiencia y precisión en conjuntos de datos grandes. La técnica de boosting secuencial permite corregir errores en las predicciones anteriores, mejorando gradualmente la precisión del modelo. Es especialmente útil en conjuntos de datos complejos con características no lineales y relaciones no triviales. En el capítulo 7, en la sección "Gradient Boosting", la referencia [3] proporciona una explicación detallada sobre las bases del modelo XGBoost.

- Random Forest (rf) Este modelo también se basa en árboles de decisión, pero utiliza múltiples árboles (un "bosque") para tomar decisiones de predicción. Cada árbol se entrena independientemente y la predicción final se determina por votación o promedio. Random Forest es conocido por su capacidad para manejar sobreajuste y generalizar bien a nuevos datos. En la sección 8.2.2, la referencia [6] proporciona una explicación detallada sobre las bases del modelo de Random Forest.

- Máquinas de Soporte Vectorial (SVM) Las SVM buscan encontrar el hiperplano que mejor separa las clases en un espacio dimensional superior. Puede manejar tanto problemas de clasificación lineal como no lineal utilizando "kernels" para mapear los datos a espacios de características más complejos. SVM se destaca en la identificación de límites de decisión óptimos en conjuntos de datos complejos.

  El clasificador de Soporte Vectorial encuentra límites lineales en el espacio de características de entrada. Al igual que con otros métodos lineales, podemos flexibilizar el procedimiento ampliando el espacio de características mediante

expansiones de bases, como polinomios. Por lo general, los límites lineales en el espacio ampliado consiguen una mejor separación en las clases de entrenamiento, y se traducen en límites no lineales en el espacio original.

Los datos de entrenamiento consisten de N pares $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$, donde $x_i \in \mathbb{R}^p$ y $y_i \in \{-1, 1\}$. Se define un hiperplano por

$$\{x : f(x) = x^T\beta + \beta_0 = 0\}, \tag{2.5.1}$$

donde β es un vector unitario $\|\beta\| = 1$. Una regla de clasificación inducida por $f(x)$ es

$$G(x) = sign[x^T\beta + \beta_0] \tag{2.5.2}$$

(Sección 12.2 de referencia [5], pagina 372.)

Una vez se seleccionan las funciones de la base $h_m(x), m = 1, \ldots, M$, entrenamos el clasificador SV utilizando las características de entrada
$h(x_i) = (h_1(x_i), h_2(x_i), \ldots, h_M(x_i))$, $i = 1, \ldots, N$, produciendo la función (no lineal) $\hat{f}(x) = h(x)^T\hat{\beta} + \hat{\beta}_0$. El clasificador es $\hat{G}(x) = sign(\hat{f}(x))$. *(Sección 12.3 de referencia [5], pagina 377.)*

El clasificador de máquinas de soporte vectorial es una extensión de esta idea, en la que se permite que la dimensión del espacio ampliado sea muy grande, infinita en algunos casos. Podría parecer que los cálculos se volverían muy grandes. También podría parecer que, con suficientes funciones de base, los datos serían separables y se produciría un sobreajuste. [5]

La función de Lagrange primal se utiliza en la formulación inicial del problema de optimización en SVM. Esta función combina la función de costo y las restricciones del problema en una sola expresión. El objetivo es minimizar esta función sujeta a las restricciones que aseguran que las instancias de datos se clasifiquen correctamente.

Definimos la función de Lagrange primal de la siguiente forma:

$$L_P = \frac{1}{2}\|\beta\|^2 + \gamma\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i\left[y_i\left(x_i^T\beta + \beta_0\right) - (1 - \xi_i)\right] - \sum_{i=1}^{N}\mu_i\xi_i \tag{2.5.3}$$

la cual minimizamos con respecto a β, β_0 y ξ_i. Fijando las derivadas respectivas a cero, obtnemos

$$\begin{aligned} \beta &= \sum_{i=1}^{N} \alpha_i y_i x_i, \\ 0 &= \sum_{i=1}^{N} \alpha_i y_i, \\ \alpha_i &= \gamma - \mu_i, \quad \forall i, \end{aligned} \tag{2.5.4}$$

así como las restricciones de positividad $\alpha_i, \mu_i, \xi_i \geq 0 \, \forall i$. *(Sección 12.2.1 de referencia [5], pagina 374.)*

Podemos representar el problema de optimización para la función de Lagrange primal 2.5.3 y su solución de una manera que sólo implica las características de entrada a través de productos internos. Lo hacemos directamente para los vectores $h(x_i)$. A continuación, vemos que para determinadas elecciones de h, estos productos internos pueden calcularse de forma muy sencilla. La función dual de Lagrange tiene la forma

$$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} \left\langle h\left(x_i\right), h\left(x_{i'}\right) \right\rangle \tag{2.5.5}$$

De 2.5.4, la función de solución o decisión $f(x)$ se puede escribir como

$$\begin{aligned} f(x) &= h(x)^T \beta + \beta_0 \\ &= \sum_{i=1}^{N} \alpha_i y_i \left\langle h(x), h\left(x_i\right) \right\rangle + \beta_0 \end{aligned} \tag{2.5.6}$$

Dadas α_i, β_0 se puede determinar al resolver $y_i f(x_i) = 1$ en 2.5.6 para cualquier x_i para la cual $0 < \alpha_i < \gamma$. *(Sección 12.3.1 de referencia [5], pagina 377.)*

Ambas 2.5.5 y 2.5.6 implican $h(x)$ sólo mediante productos internos. En realidad, no necesitamos especificar la transformación h(x) en absoluto, sino que sólo necesitamos conocer la función kernel

$$K\left(x, x'\right) = \left\langle h(x), h\left(x'\right) \right\rangle \tag{2.5.7}$$

que calcula productos internos en el espacio transformado. K debe ser una función simétrica positiva (semi) definida. Tres opciones populares para K en la literatura

SVM son

- Polinomio de grado d: $K(x, x') = (1 + \langle x, x' \rangle)^d$
- Base radial: $K(x, x') = \exp\left(- \|x - x'\|^2 / c\right)$
- Red neuronal: $K(x, x') = \tanh\left(\kappa_1 \langle x, x' \rangle + \kappa_2\right)$

(Sección 12.3.1 de referencia [5], pagina 378.)

La función de Lagrange dual se deriva de la función primal mediante la eliminación de las variables β y β_0 y reformulando el problema en términos de los multiplicadores de Lagrange α_i. La formulación dual es más conveniente para resolver debido a la complejidad reducida y la facilidad de manejo de las restricciones.

La función de solución o decisión es la que se utiliza para clasificar nuevas instancias de datos. Una vez que se han encontrado los valores óptimos de los multiplicadores de Lagrange, la función de decisión se puede expresar en términos de estos valores y los vectores de soporte.

- K-Vecinos Más Cercanos (KNN)

KNN clasifica una instancia basándose en las clases de sus vecinos más cercanos. No realiza entrenamiento propiamente dicho; su predicción se basa en la mayoría de las clases de los vecinos más cercanos. Es robusto y fácil de entender, pero puede ser sensible a valores atípicos y requiere una elección adecuada del parámetro k (número de vecinos). Los clasificadores K-Vecinos Más Cercanos se basan en la memoria y no requieren el ajuste de ningún modelo.

Dado un punto de consulta x_0, encontramos los k puntos de entrenamiento $x(r), r = 1, \ldots, k$ más cercanos en distancia a x_0, y a continuación, clasificamos utilizando el voto mayoritario entre los k vecinos. Los empates se rompen al azar. Por simplicidad, supondremos que las características son de valor real y utilizaremos la distancia euclidiana en el espacio de características

$$d_{(i)} = \left\| x_{(i)} - x_0 \right\| \tag{2.5.8}$$

Normalmente, primero estandarizamos cada una de las características para que tengan media cero y varianza 1, ya que es posible que se midan en unidades

diferentes. A pesar de su simplicidad, k-Vecinos Más Cercanos ha tenido éxito en un gran número de problemas de clasificación, incluyendo dígitos escritos a mano, imágenes de satélite y patrones de electrocardiograma. Suele tener éxito cuando cada clase tiene muchos prototipos posibles y el límite de decisión es muy irregular. *(Sección 13.3 de referencia [5], pagina 417.)*

- Naive Bayes (NB) Los clasificadores Naive Bayes se basan en el teorema de Bayes y asumen independencia condicional entre las características. A pesar de su suposición simplificada, funcionan bien en conjuntos de datos grandes y de alta dimensionalidad. Son rápidos y eficientes, especialmente en problemas de clasificación de texto y categorización.

  El clasificador Naive Bayes es una técnica especialmente apropiada cuando la dimensión p del espacio de características es alta, lo que hace poco atractiva la estimación de la densidad. El modelo de Naive Bayes supone que, dada una clase $G = j$, las características X_k son independientes:

$$f_j(X) = \prod_{k=1}^{p} f_{jk}\left(X_k\right) \tag{2.5.9}$$

  donde f_{jk} es la función de densidad del predictor j entre las observaciones de la clase k. *(Sección 6.6.3 de referencia [5], pagina 184-185.)*

  Esencialmente, estimar una función de densidad p-dimensional es un reto porque debemos considerar no sólo la distribución marginal de cada predictor, es decir, la distribución de cada predictor por sí mismo, sino también la distribución conjunta de los predictores, es decir, la asociación entre los distintos predictores. En el caso de una distribución normal multivariante, la asociación entre los distintos predictores se resume en los elementos no diagonales de la matriz de covarianza. *(Sección 4.4.4 de referencia [6], pagina 155.)*

- Análisis Discriminante Lineal (LDA) LDA es un método estadístico que busca encontrar combinaciones lineales de variables predictoras que mejor discriminen entre clases. Funciona bien cuando las clases están separadas linealmente y asume una distribución normal de los datos.

  La teoría de decisión para clasificación nos dice que necesitamos conocer los resultados

posteriores de la clase $Pr(G|X)$ para una clasificación óptima. Suponer que $f_k(x)$ es la densidad condicional de clase de X en la clase $G = k$, y sea π_k la probabilidad anterior de la clase k, con $\sum_{k=1}^{K} \pi_k = 1$. Una simple aplicación del teorema de Bayes nos da

$$\Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell} \tag{2.5.10}$$

Vemos que en términos de capacidad de clasificación, tener la $f_k(x)$ es casi equivalente a tener la cantidad $Pr(G = k|X = x)$.

Supongamos que modelamos la densidad de cada clase como una gaussiana multivariante

$$f_k(x) = \frac{1}{(2\pi)^{p/2} \left|\mathbf{\Sigma}_k\right|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}_k^{-1}(x-\mu_k)} \tag{2.5.11}$$

El análisis lineal discriminante (LDA) surge en el caso especial en que suponemos que las clases tienen una matriz de covarianza común $\sum_k = \sum \forall k$. Para comparar dos clases k y l, basta con observar la relación logarítmica, y vemos que

$$\begin{aligned} \log \frac{\Pr(G = k \mid X = x)}{\Pr(G = \ell \mid X = x)} &= \log \frac{f_k(x)}{f_\ell(x)} + \log \frac{\pi_k}{\pi_\ell} \\ &= \log \frac{\pi_k}{\pi_\ell} - \frac{1}{2} \left(\mu_k + \mu_\ell\right)^T \mathbf{\Sigma}^{-1} \left(\mu_k - \mu_\ell\right) \\ &\quad + x^T \mathbf{\Sigma}^{-1} \left(\mu_k - \mu_\ell\right), \end{aligned} \tag{2.5.12}$$

una ecuación lineal en x. Las matrices de covarianza iguales hacen que los factores de normalización se cancelen, así como la parte cuadrática en los exponentes. *(Sección 4.3 de referencia [5], pagina 84-86.)*

- Regresión Logística (logit) Aunque su nombre sugiere regresión, la Regresión Logística se utiliza comúnmente para problemas de clasificación binaria. Utiliza la función logística para modelar la probabilidad de que una instancia pertenezca a una clase particular. Es simple, rápido y útil para comprender la importancia relativa de cada variable predictora. Generalmente es el modelo de base para comparar desempeño con otros algoritmos como los mencionados anteriormente.

  El modelo de regresión logística surge del deseo de modelar las probabilidades posteriores de las K clases mediante funciones lineales en x, mientras que al mismo tiempo tiempo asegurándose de que suman uno y permanecen en $[0, 1]$. El modelo

tiene la forma:

$$\begin{aligned}
\log \frac{\Pr(G=1 \mid X=x)}{\Pr(G=K \mid X=x)} &= \beta_{10} + \beta_1^T x \\
\log \frac{\Pr(G=2 \mid X=x)}{\Pr(G=K \mid X=x)} &= \beta_{20} + \beta_2^T x \\
&\vdots \\
\log \frac{\Pr(G=K-1 \mid X=x)}{\Pr(G=K \mid X=x)} &= \beta_{(K-1)0} + \beta_{K-1}^T x
\end{aligned} \tag{2.5.13}$$

El modelo se especifica en términos de $K-1$ transformaciones logit (reflejando la restricción de que las probabilidades suman uno). *(Sección 4.4 de referencia [5], pagina 95-96.)*

En el caso de este trabajo de investigación, contamos con dos clases a modelar, por lo que el modelo utilizado tiene la siguiente forma:

$$\log \frac{\Pr(G=1 \mid X=x)}{\Pr(G=2 \mid X=x)} = \beta_0 + \beta^T x \tag{2.5.14}$$

El modelo logit, originario de la econometría, ha sido ampliamente adoptado en ciencia de datos, sobre todo para clasificación, debido a su capacidad para modelar decisiones o resultados binarios, como la predicción de si un cliente comprará un producto o no. La simplicidad y robustez del modelo lo han convertido en una herramienta esencial tanto en econometría como en machine learning.

Los modelos econométricos son herramientas estadísticas que permiten analizar y cuantificar relaciones entre variables, especialmente en la economía e incluso fuera de esta área de conocimiento. Se basan en la idea de modelar fenómenos observables (por ejemplo en la economía pueden ser el consumo o la producción) a través de regresiones que permiten identificar cómo una variable depende de otras. Los modelos más comunes incluyen regresiones lineales, donde se asume una relación lineal entre variables, pero también existen modelos no lineales que se utilizan para capturar comportamientos más complejos, como los modelos probit o logit para variables dependientes cualitativas. Estos modelos permiten hacer predicciones.

2.6. Manejo del Desequilibrio de Clases en Modelos de Clasificación: Técnicas y Parámetros

El problema de clasificar dos grupos o clases con tamaños significativamente diferentes es un desafío bien conocido en el ámbito del aprendizaje automático y la minería de datos. Este problema, a menudo denominado desequilibrio de clases, surge cuando uno de los grupos, denominado la clase minoritaria, es considerablemente más pequeño que el otro, la clase mayoritaria. Este desequilibrio puede llevar a diversos problemas en la clasificación, afectando negativamente la precisión y la utilidad del modelo predictivo.

2.6.1. Problemas Causados por el Desequilibrio de Clases

- **Sesgo hacia la Clase Mayoritaria**

  Cuando los algoritmos de clasificación se entrenan en conjuntos de datos con clases desequilibradas, tienden a estar sesgados hacia la clase mayoritaria. Esto se debe a que los clasificadores intentan minimizar el error global, y pueden lograr una alta precisión simplemente clasificando la mayoría de las instancias como pertenecientes a la clase mayoritaria. Como resultado, la clase minoritaria es frecuentemente clasificada erróneamente como parte de la clase mayoritaria.

- **Baja Tasa de Clasificación Correcta**

  Debido a este sesgo, la tasa de clasificación correcta para la clase minoritaria puede ser extremadamente baja, frecuentemente menor al 50 % e incluso acercándose a cero en casos extremos. Esto significa que el modelo tiene un desempeño muy pobre en la identificación correcta de instancias pertenecientes a la clase minoritaria, lo que es especialmente problemático en aplicaciones críticas como la detección de fraudes, el diagnóstico de enfermedades raras y la detección de defectos en manufactura.

- **Pérdida de Información Importante**

El desequilibrio de clases puede llevar a la pérdida de información crucial contenida en la clase minoritaria. En muchos casos, las instancias de la clase minoritaria son precisamente las más importantes desde el punto de vista de la aplicación del modelo (por ejemplo, transacciones fraudulentas en detección de fraudes). La incapacidad del modelo para identificar correctamente estas instancias puede llevar a decisiones no óptimas y a la pérdida de oportunidades para intervenir o actuar de manera efectiva.

2.6.2. Estrategias para Manejar el Desequilibrio de Clases

Para mitigar los problemas asociados con el desequilibrio de clases, se pueden utilizar diversas estrategias:

- Recolección de Datos Adicionales: Aumentar el número de instancias de la clase minoritaria puede ayudar a balancear el conjunto de datos.
- Rebalanceo de Datos: Técnicas como el sobremuestreo de la clase minoritaria (e.g., SMOTE) o el submuestreo de la clase mayoritaria pueden ayudar a equilibrar la proporción de clases.
- Ajuste de Pesos en el Modelo: Modificar los pesos del modelo para penalizar más los errores en la clasificación de la clase minoritaria puede mejorar el desempeño del clasificador.
- Uso de Algoritmos Sensibles al Desequilibrio: Algunos algoritmos, como los árboles de decisión y los métodos de ensemble (e.g., Random Forest, Gradient Boosting) pueden ser más robustos frente a datos desequilibrados cuando se configuran adecuadamente.

Existen varias formas para medir el nivel de error en la clasificación y comparar resultados. Estas medidas comparan el número de casos buenos y malos reales respecto al número de casos buenos y malos previstos para un determinado umbral. Por "buenos" y "malos" se entienden los casos por encima y por debajo del límite propuesto. Las medidas se basan en una matriz de confusión. Un resultado ideal seria aquel donde se maximizaran los casos "verdaderos" y, a la inversa, se minimizaran los "falsos". Hay cuatro medidas principales

utilizadas para medir la clasificación errónea: [12]

- Precisión: (verdaderos positivos y negativos) / (total de casos)
- Tasa de error: (falsos positivos y negativos) / (total de casos)
- Sensibilidad: (verdaderos positivos) / (total de positivos reales)
- Especificidad: (verdaderos negativos) / (total de negativos reales)

Estas estadísticas pueden interpretarse de la siguiente manera:

- Falso Positivo: Aceptación de malos
- Verdadero positivo: Aceptación de buenos
- Falso negativo: Declinación de buenos
- Verdadero negativo: Declinación de malos

2.6.3. Métodos de Balanceo en Python

- Submuestreo (Under-sampling)

  Esta técnica implica reducir el número de instancias de la clase mayoritaria para igualar el número de instancias de la clase minoritaria. Aunque esto puede equilibrar el conjunto de datos, puede también llevar a la pérdida de información significativa. Función *RandomUnderSampler* de librería *imblearn.under_sampling*.

- Sobremuestreo (Over-sampling)

  En contraste con el submuestreo, el sobremuestreo aumenta el número de instancias de la clase minoritaria replicando ejemplos existentes o creando nuevos ejemplos sintéticos. Una técnica común es el SMOTE (Synthetic Minority Over-sampling Technique). Función *SMOTE* de librería *imblearn.over_sampling*.

- Combinación de Submuestreo y Sobremuestreo

  Algunas técnicas combinan tanto el sobremuestreo de la clase minoritaria como el submuestreo de la clase mayoritaria para equilibrar el conjunto de datos. Función *SMOTETomek* de librería *imblearn.combine*.

- Ponderación de Clases

Algunos algoritmos de clasificación permiten asignar un peso mayor a la clase minoritaria durante el entrenamiento, penalizando más los errores cometidos en estas clases. Este método es particularmente útil cuando se utilizan modelos como la regresión logística (Logit) y el Random Forest (RF). En los parametros de la funcion del modelo de clasificación de python se puede agregar lo siguiente $class\_weight =' balanced'$. Donde el modo $'balanced'$ utiliza los valores de y para ajustar automáticamente las ponderaciones de forma inversamente proporcional a las frecuencias de clase en los datos de entrada como $n\_samples/(n\_classes*np.bincount(y))$. Éste fue el método utilizado en la realización del codigo para el entrenamiento de modelos. [11]

2.6.4. Penalización en Modelos de Aprendizaje Automático y sus Implementaciones en Python

La penalización es una técnica crucial en el aprendizaje automático y la estadística, utilizada para prevenir el sobreajuste de los modelos. El sobreajuste ocurre cuando un modelo se ajusta demasiado bien a los datos de entrenamiento, capturando el ruido junto con la señal y, por lo tanto, desempeñándose mal en datos no vistos. La penalización agrega un término de regularización a la función de costo, que penaliza los coeficientes de los modelos para promover la simplicidad y evitar el sobreajuste.

- Lasso (L1) Regularización

  La regularización Lasso (Least Absolute Shrinkage and Selection Operator) agrega la suma de los valores absolutos de los coeficientes a la función de costo. Esto puede llevar a que algunos coeficientes se reduzcan a cero, lo que permite la selección automática de características y resulta en un modelo más simple.

- Ridge (L2) Regularización:

  La regularización Ridge agrega la suma de los cuadrados de los coeficientes a la función de costo. Esto no fuerza a los coeficientes a ser cero, pero sí los reduce, promoviendo un modelo más estable y menos susceptible a multicolinealidad.

En la regresión logística, se puede agregar una penalización $L1$ o $L2$ para evitar el

sobreajuste y mejorar la generalización del modelo. La penalización se especifica mediante el parámetro *penalty*. En este trabajo utilizamos el parámetro $L2$ como penalización para el modelo de clasificación LOGIT.

La regresión Ridge (L2) reduce los coeficientes de regresión imponiendo una penalización a su tamaño. Los coeficientes Ridge minimizan una suma de cuadrados residual penalizada,

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{2.6.1}$$

Aquí $\lambda \geq 0$ es un parámetro de complejidad que controla la reducción: cuanto mayor sea el valor de λ, mayor será la reducción. Los coeficientes se encogen hacia cero (y entre sí). La idea de penalizar por la suma de los cuadrados de los parámetros también se utiliza en redes neuronales, donde se conoce como decaimiento del peso. [5]

Notar que la regresión Lasso (L1) añade el "valor absoluto de magnitud" del coeficiente como término de penalización a la función de pérdida, en la ecuación 2.6.1 en lugar de tener el término β_j^2, se tiene $|\beta_j|$. Es decir:

$$\hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{2.6.2}$$

La penalización L2 ayuda a reducir el sobreajuste al imponer un costo adicional para grandes valores de los coeficientes, forzándolos a ser más pequeños y, por lo tanto, creando un modelo más simple que generaliza mejor en datos no vistos. Al restringir los coeficientes, la penalización L2 puede hacer que el modelo sea más estable y menos sensible a las pequeñas variaciones en los datos de entrenamiento.

2.7. Comparativo de Rendimiento

El análisis comparativo se basa en múltiples métricas de evaluación, en el presente trabajo de tesis nos enfocamos en el uso de las siguientes: el Área bajo la Curva ROC, la matriz de confusión y la prueba de Kolmogorov-Smirnov (KS).

2.7.1. El Área bajo la Curva ROC

El Área bajo la Curva ROC es una métrica crítica utilizada en la evaluación de modelos de clasificación que representa la capacidad de un modelo para distinguir entre clases positivas y negativas. Este valor se encuentra entre 0 y 1, donde un AUC de 1 indica un modelo perfecto que puede discriminar perfectamente entre clases, mientras que un valor de 0.5 sugiere un rendimiento similar al azar, sin capacidad de discriminación.

La Curva ROC (Receiver Operating Characteristic) es el fundamento para calcular el AUC. Esta curva gráfica la tasa de verdaderos positivos (Sensibilidad) frente a la tasa de falsos positivos (1 - Especificidad) para diversos umbrales de clasificación del modelo. Cuanto mayor sea el AUC, mayor será la capacidad del modelo para diferenciar entre clases.

La Sensibilidad representa la proporción de verdaderos positivos correctamente identificados, lo que implica una menor cantidad de falsos negativos. Por otro lado, la Especificidad indica la proporción de verdaderos negativos correctamente identificados, reduciendo así la cantidad de falsos positivos. Estas dos métricas, Sensibilidad y Especificidad, son clave en la evaluación de modelos de clasificación y se reflejan en la Curva ROC y su AUC asociado.

El AUC es beneficioso en la comparación de múltiples modelos de clasificación, permitiendo seleccionar el más adecuado para un problema específico. Es robusto incluso cuando existe desequilibrio entre las clases de los datos, lo que lo convierte en una métrica valiosa en escenarios donde una clase es significativamente más numerosa que otra.

A pesar de su utilidad, el AUC no proporciona información sobre el umbral de clasificación óptimo. Un modelo con un AUC alto puede tener un desempeño pobre en un umbral específico. Por lo tanto, aunque el AUC sea una métrica crucial para evaluar el rendimiento general de un modelo de clasificación, se debe complementar con otras métricas y consideraciones al tomar decisiones sobre el modelo a utilizar.

2.7.2. Estadístico de Kolmogorov-Smirnov (KS)

El estadístico de Kolmogorov-Smirnov (KS), denotado por D_n en (1.3.16), es una métrica estadística utilizada para evaluar la capacidad de un modelo de clasificación para distinguir entre dos clases diferentes en un conjunto de datos. Se enfoca específicamente en la distribución acumulativa de probabilidad de las predicciones del modelo para las clases positiva y negativa.

Esta métrica se deriva de la Curva de Ganancia Acumulada (Cumulative Gain Curve), que compara la tasa de verdaderos positivos (TPR) con la tasa de falsos positivos (FPR) a través de varios puntos de corte. El KS mide la máxima distancia vertical entre estas dos curvas, lo que indica la máxima discrepancia entre las distribuciones acumulativas de las clases positiva y negativa.

En un contexto de evaluación de modelos de clasificación, el KS ofrece una medida de la habilidad del modelo para clasificar correctamente las instancias positivas por encima de las negativas. Un KS más alto sugiere que el modelo puede separar de manera más efectiva las dos clases, lo que se traduce en una mejor capacidad de discriminación entre ellas.

Cuando se emplea el KS para evaluar modelos de riesgo crediticio, por ejemplo, se utiliza para analizar la capacidad del modelo para ordenar a los individuos según su riesgo crediticio. Un KS alto indica que el modelo puede distinguir eficientemente a aquellos individuos con mayor riesgo de aquellos con menor riesgo de incumplimiento crediticio.

Es importante tener en cuenta que, al igual que otras métricas de evaluación, el KS no proporciona información sobre la precisión absoluta del modelo, sino que se enfoca en su capacidad de discriminación entre clases. Una desventaja es que el KS tiende a ser sensible al tamaño del conjunto de datos y puede ser influenciado por desequilibrios entre las clases, aunque es menos afectado por desequilibrios en comparación con el AUC (Área bajo la Curva ROC).

Nos preguntamos si existe una relación entre el AUC y el KS, es decir, ¿es cierto que un alto valor de AUC corresponde a un alto valor de KS? Este análisis nos permite entender la posible relación entre estas métricas y su influencia en la precisión del modelo de clasificación.

Es razonable anticipar una relación entre el Área bajo la Curva ROC y el estadístico de Kolmogorov-Smirnov (KS), pero esta relación no es necesariamente directa o lineal en todos los casos. En muchos casos, un modelo con un alto AUC también puede tener un alto KS, ya que ambos evalúan la habilidad del modelo para distinguir entre clases. Sin embargo, no siempre es una relación directa y puede haber excepciones. Por ejemplo, si el AUC es alto pero hay un desbalance significativo entre las clases, el KS puede no reflejar adecuadamente la capacidad del modelo para discriminar entre las clases.

En conjunto, la fase de modelación representa un paso crucial en la selección y evaluación de modelos para la predicción del riesgo crediticio. La experimentación meticulosa y la comparación de diversos enfoques nos proporcionarán una comprensión más profunda del rendimiento de los modelos y su capacidad para abordar este desafío específico de clasificación.

2.8. Implementación de Código

En el código implementado, se realiza un experimento completo de modelado de aprendizaje supervisado utilizando varios algoritmos y diferentes combinaciones de variables de entrada. El código sigue el flujo estándar de un proyecto de aprendizaje de maquina, desde la preparación de datos hasta la evaluación del modelo. Las decisiones tomadas en cada sección se basan en la teoría de aprendizaje de maquina y en las mejores prácticas para abordar desafíos comunes en la construcción de modelos predictivos. El código utilizado se encuentra en el apéndice B.1.

2.8.1. Objetivo General

El objetivo principal del código es ajustar varios modelos de aprendizaje supervisado: KNN, Naive Bayes, LDA, SVM, Logistic Regression, Random Forest y XGBoost. Para predecir diferentes variables objetivo. Además, explora diversas combinaciones de variables explicativas para cada modelo y variable objetivo, evaluando su desempeño mediante las métricas: AUC, KS y matrices de confusión.

2.8.2. Estructura del Código

1. Importación de Librerías: Se importan librerías fundamentales como pandas, numpy, matplotlib, y diversas herramientas de scikit-learn y scorecardpy para el modelado y evaluación de desempeño.
2. Preliminares:
   - Configuración de listas para experimentar con diferentes estrategias de balanceo de clases y penalizaciones en los modelos.
   - Definición de modelos a ser reportados, variables objetivo y diferentes conjuntos de variables explicativas.
3. Datos: Lectura de datos desde un archivo CSV y división del conjunto de datos en conjuntos de entrenamiento y prueba.
4. Funciones: Definición de funciones para el entrenamiento de modelos y la generación de informes de desempeño. Cada función se centra en un conjunto específico de modelos y variables objetivo.
   - Se entrenan modelos para diferentes combinaciones de variables y estrategias.
   - Se utilizan los siguientes modelos de clasificación
     - KNN: utilizando el modelo de Python *KNeighborsClassifier* con los parámetros *n_ neighbors=4*
     - NB: se empleó el modelo de Python *GaussianNB* con sus parámetros por defecto para realizar la clasificación.
     - LDA: se empleó el modelo de Python *LinearDiscriminantAnalysis* utilizando el parámetro *solver='svd'*
     - RF: se empleó el modelo de Python *RandomForestClassifier* utilizando los parámetros *n_ estimators=500* y *max_ depth=6*
     - XGB: se empleó el modelo de Python *XGBClassifier* utilizando los parámetros *n_ estimators=300* y *max_ depth=6*
     - LOGIT: se empleó el modelo de Python *LogisticRegression* utilizando

los parámetros *solver= 'newton-cg'*, *class_weight='balanced'* o *class_weight=None* así como *penalty='l2'*

- Para asegurar la robustez y confiabilidad de los resultados obtenidos, se realizaron 100 repeticiones para cada uno de los modelos evaluados. Este enfoque se implementó para mitigar el impacto de la variabilidad inherente a la división de los datos y proporcionar una estimación más precisa del desempeño de los modelos.
- Se almacenan los modelos entrenados en archivos pickle para su uso futuro.
- Se mide y registra el tiempo de entrenamiento.
- Se evalúa el desempeño de los modelos en el conjunto de prueba.
- Se calculan las métricas AUC, KS y matrices de confusión.

5. Main:
   - Ciclo principal que itera a través de repeticiones y diferentes combinaciones de variables.
   - Para cada repetición, se selecciona un conjunto de entrenamiento, y se realizan iteraciones para diferentes combinaciones de modelos, variables objetivo, penalidades y estrategias de balanceo.
   - Utiliza la función de entrenamiento para ajustar modelos a los datos de entrenamiento.
   - Explora diferentes modelos, variables objetivo, variables explicativas y estrategias de balanceo y penalización.
   - Entrenamiento de modelos y guardado de los modelos resultantes en archivos pickle.
   - Reporte de desempeño de los modelos, calculando métricas como AUC, KS y matrices de confusión.
6. Reporte de Desempeño: Consolidación de los resultados en un dataframe y exportación a un archivo CSV para un análisis más detallado. Filtra modelos basándose en umbrales AUC y guarda resultados para análisis posterior.

2.8.3. Consideraciones Importantes.

Se realiza un preprocesamiento mínimo y ajuste básico de los modelos para mantener el número de combinaciones en un limite razonable. No obstante, experimentos posteriores se beneficiarán de la información adquirida. En particular, se descartarán algunos modelos por haber tenido un bajo desempeño.

2.8.4. Resultados:

- El código genera un informe completo del desempeño de los modelos para diferentes combinaciones de variables y estrategias.
- Se utilizan las métricas AUC y KS para evaluar la capacidad predictiva de los modelos.

2.8.5. Notas Finales:

El código es estructurado y modular, lo que facilita la comprensión y la extensión para futuros experimentos y mejoras.

2.9. Calibración de hiperparametros

Los modelos considerados en la tesis además de los coeficientes que definen cada uno de ellos, incluyen hiperparametros que determinan el algoritmo de estimación. Estos hiperparametros se especifican mediante un procedimiento conocido como calibración (o tunning en inglés) y para el que existen diferentes alternativas. Aquí nos enfocaremos al procedimiento conocido como validación cruzada.

La validación cruzada es una técnica utilizada para evaluar el rendimiento de un modelo de machine learning y al mismo tiempo evitar el sobreajuste. Consiste en dividir el conjunto de datos en múltiples subconjuntos, entrenar el modelo en varios de estos subconjuntos y evaluarlo en el subconjunto restante. Este proceso se repite varias veces, de modo que cada subconjunto se utiliza tanto para entrenamiento como para evaluación en diferentes iteraciones del proceso. La validación cruzada es fundamental porque proporciona una

evaluación más robusta del rendimiento del modelo al promediar los resultados de múltiples particiones de datos, lo que ayuda a mitigar el sesgo de estimación que puede surgir al evaluar el modelo en un único conjunto de datos de prueba.

Algunos de los hiperparametros son categóricos y en la sección 2.10.3 se reportan ejercicios de estimación para todos los modelos variando los hiperparametros de regularización y de balanceo. Adicional a dicho ejercicio, en esta sección se describirá la validación cruzada mediante el cual se exploraron los valores optimos de hiperparametros para los modelos XGBoost, Bosques aleatorios y KNN.

Los modelos adicionales considerados en esta tesis (SVM, NB, LDA y Regresión Logística) también incluyen hiperparámetros, algunos de naturaleza numérica, los cuales se mantendrán en sus valores predeterminados (como el parámetro "tol" para el modelo SVC lineal). Esto se debe, por un lado, a los costos computacionales asociados con la optimización de estos parámetros y, por otro, a que se realizó una validación cruzada exhaustiva con malla para los modelos XGBoost, bosques aleatorios y vecinos próximos. Esta validación no solo proporciona evidencia concreta de los costos computacionales involucrados, sino que también asegura un resultado óptimo para comparación.

2.9.1. Validación cruzada para XGBoost

Para el modelo XGBoost se realizó una validación cruzada para la que se consideró la siguiente malla:

```
grid={'n_estimators':range(20,1001,20),'max_depth':range(1,21),
'learning_rate':[.001,.01,.1,0.2],'eval_metric':['auc'],
'n_jobs':[-1], 'reg_alpha':[0.01, 0.5,1,5],
'reg_lambda':[0.5,1,5],'objective':['binary:logistic']}
```

La validación cruzada se realizó en Python mediante la instrucción:

```
GridSearchCV(XGBClassifier(),grid,cv=5,scoring='roc_auc')
```

El proceso de validación cruzada requirió 727,291 segundos (aproximadamente 8.5 días) ejecutandose en 32-1 hilos de procesamiento. Los mejores hiperparametros fueron:

```
{'learning_rate': 0.01, 'max_depth': 1, 'n_estimators': 1000,
```

```
'reg_alpha': 0.01, 'reg_lambda': 1}
#Best score:  0.7868118293599966
```

El modelo resultante de la validación cruzada tuvo el siguiente desempeño:

Datos	AUC	DIAG 1	DIAG 2
Train	0.7809	0.6505	0.7918
Test	0.7693	0.6503	0.7775

También se realizó un segundo ejercicio de validación cruzada en el que se exploró con una malla mas fina el número de estimadores y en el que los valores de otros hiperparametros se fijaron a los valores obtenidos en el primer ejercicio. Se consideró la siguiente malla:

```
grid_cpu_short1={'n_estimators':range(10,1001,10), 'max_depth':range(1,11),
'eval_metric':['auc'],'n_jobs':[-1], 'objective':['binary:logistic'],
'learning_rate':[.01],'reg_lambda':[1],'reg_alpha':[0.01],
'subsample':[1]}
```

El proceso de validación cruzada requirió 37,302 segundos (aproximadamente 10.4 horas) ejecutandose en 32-1 hilos de procesamiento. Los mejores hiperparametros fueron:

```
Tiempo de ejecución: 37,302 segundos (aproximadamente 10.4 horas)
#Best set of hyperparameters:  {'learning_rate': 0.01, 'max_depth': 5,
'n_estimators': 580,'reg_alpha': 0.01, 'reg_lambda': 1}
Best score:  0.7830151673767398
```

El modelo resultante de la validación cruzada tuvo el siguiente desempeño:

Datos	AUC	DIAG 1	DIAG 2
Train	0.8184	0.6810	0.7990
Test	0.7733	0.6609	0.7630

Comparando las tablas que reportan los desempeños vemos que el segundo ejercicio presenta mayor evidencia de sobreajuste en comparación con el primer ejercicio. Sin embargo, el desempeño en los datos de prueba es superior para el segundo ejercicio. Motivo por el cual, para el ejercicio en la sección 2.8 se utilizan los parametros: $max\_depth = 6$ para la profundidad maxima y $n\_estimators = 300$ para la cantidad de estimadores.

2.9.2. Validación cruzada para bosques aleatorios

Para el modelo de bosques aleatorios se realizó una validación cruzada para la que se consideró la siguiente malla:

```
grid={'n_estimators':range(20,1001,20),'criterion':['entropy'],
'max_depth':range(1,21), 'max_features':['sqrt'],
'bootstrap':[False]}
```

El proceso de validación cruzada requirió 9095 segundos (aproximadamente 2.5 horas) en 16-1 hilos. Los mejores hiperparametros fueron:

```
{'max_depth': 13, 'n_estimators': 640}
Best score:  0.7858939239134216
```

El modelo resultante de la validación cruzada tuvo el siguiente desempeño:

Datos	AUC	DIAG 1	DIAG 2
Train	0.9312	0.7706	0.8865
Test	0.7771	0.6756	0.7618

La tabla muestra cierta evidencia de sobreajuste. Motivo por el cual, en el ejercicio en la sección 2.8 bajamos la complejidad del modelo y se utilizan los parametros: $max\_depth = 6$ para la profundidad maxima y $n\_estimators = 500$ para la cantidad de estimadores.

2.9.3. Validación cruzada para vecinos próximos

Para el modelo de vecinos próximos (KNN) se realizó una validación cruzada para la que se consideró la siguiente malla:

```
grid={'n_neighbors':range(1,11,1),'p':[1,2],'n_jobs':[-1]}
```

El proceso de validación cruzada requirió 9 segundos ejecutandose en 16-1 hilos de procesamiento. Los mejores hiperparametros fueron:

```
Best set of hyperparameters:  { 'n_neighbors': 10, 'p': 1}
Best score:  0.5691465794757966
```

El modelo resultante de la validación cruzada tuvo el siguiente desempeño:

Datos	AUC	DIAG 1	DIAG 2
Train	0.7170	0.5258	0.7765
Test	0.5758	0.4305	0.6797

La tabla muestra evidencia de sobreajuste. Motivo por el cual, en el ejercicio en la sección 2.8 bajamos la complejidad del modelo y utilizaremos la especificación $n\_neighbors = 4$ para el número de vecinos próximos.

2.10. Análisis y Visualización de Resultados

El análisis y visualización de resultados desempeñan un papel fundamental en la evaluación y comprensión de los modelos desarrollados en el ámbito de la investigación y la ciencia de datos. Esta sección se centra en la presentación efectiva de los resultados obtenidos durante el curso de este estudio, empleando una variedad de herramientas visuales y tabulares para revelar patrones, tendencias y métricas clave.

La visualización de datos es una herramienta poderosa que va más allá de la mera presentación estética; constituye un medio para comunicar de manera efectiva la complejidad inherente en los conjuntos de datos y resultados experimentales. En este contexto, se explorarán diversas técnicas gráficas, tales como diagramas de caja, gráficos de barras, y otros tipos de visualizaciones, que permitirán destacar características específicas y proporcionar una interpretación intuitiva de los hallazgos.

Además de las representaciones gráficas, se emplearán tablas informativas para organizar y presentar de manera concisa los resultados cuantitativos. Estas tablas ofrecerán un resumen detallado de las métricas de rendimiento y otras estadísticas relevantes, proporcionando a los lectores una visión completa y estructurada de los resultados obtenidos.

En particular, se hará uso de las capacidades de programación en Python, aprovechando bibliotecas como Matplotlib, Seaborn y Pandas, para crear visualizaciones dinámicas y personalizadas. Esto permitirá una presentación visual atractiva y una exploración detallada de los resultados desde diferentes perspectivas.

Esta sección no solo se limita a exponer los resultados, sino que se busca la interpretación y extracción de conclusiones valiosas a partir de las representaciones visuales y tabulares presentadas. A través de este análisis visual, se proporcionará una visión más completa y

accesible de la complejidad de los datos y del impacto de las decisiones tomadas durante el proceso de investigación.

Se abordará el análisis detallado y la visualización de los resultados obtenidos, utilizando un enfoque integral que combina gráficos especializados y tablas informativas para proporcionar una representación clara y significativa del trabajo llevado a cabo en este estudio.

2.10.1. Análisis Descriptivo de Tablas

A continuación, se examinarán las tablas descriptivas que presentan los resultados de las métricas de los modelos de clasificación. Estas tablas se encuentran en el apéndice del documento A1, A2, A3, A4, A5. Las tablas en cuestión fueron generadas utilizando código Python, el cual fue discutido en la sección anterior. Además, se han generado gráficos basados en estas tablas, los cuales serán analizados en la siguiente sección. En los análisis que se presentan a continuación, se emplean las variables *opcion*1, *opcion*2 y *opcion*3, las cuales son conjuntos que agrupan variables explicativas. El contenido de cada variable es explicado en la sección 2.4.1.

La realización de un análisis detallado de las tablas descriptivas que contienen métricas fundamentales como AUC (Area Under the Curve) para los conjuntos de prueba (AUC test) y entrenamiento (AUC train), la métrica de Kolmogorov-Smirnov (KS), y las matrices de confusión, proporciona una visión profunda y significativa de la calidad y el rendimiento de los modelos predictivos.

El análisis descriptivo de estas tablas revela información clave sobre la variabilidad y consistencia de los resultados, expresada a través de medidas como el promedio, la desviación estándar, y los valores mínimo y máximo. Estos indicadores estadísticos no solo ofrecen una comprensión central de la distribución de los AUC, sino que también proporcionan perspectivas sobre la estabilidad del rendimiento del modelo en diferentes configuraciones. Al explorar las estadísticas descriptivas, se puede obtener información crítica sobre la uniformidad de las predicciones y la capacidad del modelo para adaptarse a diferentes distribuciones de probabilidad.

Las matrices de confusión, que comprenden los elementos verdaderos positivos, falsos

positivos, verdaderos negativos y falsos negativos, ofrecen una visión detallada del rendimiento del modelo en términos de clasificación. Es crucial destacar que la diagonal principal de la matriz representa los casos correctamente clasificados, es decir, los verdaderos positivos y verdaderos negativos. Una diagonal con valores cercanos a 1 es indicativa de un modelo preciso y confiable, mientras que las desviaciones de la diagonal pueden señalar áreas de mejora en el rendimiento del modelo. Por tanto, el análisis descriptivo de estas matrices, proporciona información sobre la precisión y la robustez del modelo en la identificación de diferentes clases.

En conjunto, este análisis descriptivo no solo sirve como un medio para comprender la variabilidad en las métricas de rendimiento, sino que también permite la identificación de patrones, tendencias y posibles áreas de mejora en los modelos predictivos. Es un paso esencial en el proceso de evaluación de modelos, proporcionando una base cuantitativa sólida para la toma de decisiones informada en contextos predictivos y analíticos.

A continuación se explica con mayor detalle las tablas generadas.

En la Tabla A1, se evidencia que los modelos con un AUC-test promedio superior a 0.75 incluyen LDA con balanceo nulo y penalidad nula, XGB con balanceo nulo y penalidad nula, RF con balanceo nulo y penalidad nula, RF con balanceo aplicado y penalidad nula, LOGIT con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad "l2", LOGIT con balanceo aplicado y penalidad nula, LOGIT con balanceo aplicado y penalidad "l2". Este umbral superior a 0.75 se aplica exclusivamente a estos modelos y se refiere específicamente a la variable "yimpuntual" en todas las configuraciones de la variable "x". La combinación de todos estos modelos muestra una desviación estándar promedio de 0.00422. Los resultados más destacados, con un valor promedio de 0.77, corresponden a los modelos XGB con balanceo nulo y penalidad nula, RF con balanceo nulo y penalidad nula, RF con balanceo aplicado y penalidad nula. Por otro lado, los resultados menos favorables, con un valor promedio de 0.54, se atribuyen al modelo KNN con balanceo nulo y penalidad nula. Estos resultados se observaron para las variables "ydeudor", "yimpuntual" y "yrecuperacionbaja" en todas las opciones de la variable "x".

En la Tabla A2, se destaca que los modelos con un AUC-train promedio superior a 0.75 incluyen KNN con balanceo nulo y penalidad nula, XGB con balanceo nulo y penalidad

nula, RF con balanceo nulo y penalidad nula, RF con balanceo aplicado y penalidad nula. Estos resultados aplican específicamente a las variables "ydeudor", "yimpuntual" y "yrecuperacionbaja" en todas las configuraciones de la variable "x". El conjunto de estos modelos exhibe una desviación estándar promedio de 0.00263. Otros de los modelos con un AUC-train promedio superior a 0.75 comprende LDA con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad "l2", LOGIT con balanceo aplicado y penalidad nula, LOGIT con balanceo aplicado y penalidad "l2". Limitados exclusivamente a la variable "yimpuntual" en todas las configuraciones de la variable "x". La desviación estándar promedio de este conjunto es de 0.00186. Los resultados más notables, con un valor promedio de 0.88, corresponden al modelo KNN con equilibrio nulo y penalización nula. Asimismo, uno de los resultados destacados, con un valor promedio de 0.82, pertenece a XGB con equilibrio nulo y penalización nula.

Por otra parte, se registran algunos de los resultados menos favorables, evidenciando un valor promedio de 0.69 en el modelo LDA con equilibrio nulo y penalización nula, así como en el modelo LOGIT en todas sus variantes. Estos resultados son específicos para las variables "ydeudor" y "yrecuperacionbaja" en todas las configuraciones de la variable "x", especialmente cuando se elige la "opcion3". Otros resultados menos alentadores, con un valor promedio de 0.61, se presentan en el modelo NB con equilibrio nulo y penalización nula para las variables "ydeudor", "yimpuntual" y "yrecuperacionbaja" en todas las opciones de la variable "x".

Se evidencia una disparidad significativa entre los valores promedio de AUC test y AUC train en el modelo KNN. Mientras que el AUC test muestra un valor promedio de 0.54, situándose entre los más bajos, el AUC train exhibe valores promedio de 0.88, figurando entre los más altos. **Esta discrepancia indica un caso de sobreajuste para el modelo KNN, dado que la disparidad entre ambos valores es considerablemente elevada.**

En la Tabla A3, se resalta que los modelos que muestran un valor promedio superior a 0.95 en el componente uno de la diagonal de la matriz de confusión son LDA con balanceo nulo y penalidad nula, KNN con balanceo nulo y penalidad nula, NB con balanceo nulo y penalidad nula, XGB con balanceo nulo y penalidad nula, RF con balanceo nulo y

penalidad nula, LOGIT con balanceo nulo y penalidad nula, y LOGIT con balanceo nulo y penalidad "l2". Estos resultados son específicos para las variables "yrecuperacionbaja" y "ydeudor" en todas las configuraciones de la variable "x". El conjunto de estos modelos presenta una desviación estándar promedio de 0.00069.

Entre los resultados más destacados, con un valor promedio de 1.0, se encuentran LOGIT con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad "l2", específicamente para las variables "yrecuperacionbaja" y "ydeudor" en todas las configuraciones de la variable "x". Además, el modelo LDA con balanceo nulo y penalidad nula muestra un valor de 1.0 para la variable "ydeudor" en la opción "opcion3", y el modelo XGB con balanceo nulo y penalidad nula obtiene un valor de 1.0 para la variable "ydeudor" en todas las configuraciones de la variable "x".

Por otro lado, se registran algunos de los resultados menos favorables, con un valor promedio de 0.1596, en el modelo NB con balanceo nulo y penalidad nula. Estos resultados se observaron para la variable "yimpuntual" en las opciones de la variable "x" "opcion1" y "opcion2".

En la Tabla A4, se destaca que los modelos que exhiben un valor promedio superior a 0.7 en el componente dos de la diagonal de la matriz de confusión son LDA con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad "l2", NB con balanceo nulo y penalidad nula, RF con balanceo nulo y penalidad nula, RF con penalidad nula y balanceo, y XGB con balanceo nulo y penalidad nula. Esto se aplica específicamente a la variable "yimpuntual" en todas las configuraciones de la variable "x". El conjunto de estos modelos presenta una desviación estándar promedio de 0.0095.

De manera similar, con un promedio superior a 0.7 en el componente dos de la diagonal de la matriz de confusión, se encuentran LOGIT con balanceo aplicado y penalidad nula, y LOGIT con balanceo aplicado y penalidad "l2", referenciándose a las variables "yimpuntual", "ydeudor" y "yrecuperacionbaja" en todas las configuraciones de la variable "x". El conjunto de estos modelos presenta una desviación estándar promedio de 0.0128.

Entre los resultados más destacados, con un valor promedio de 0.89, se encuentra NB

con balanceo nulo y penalidad nula, específicamente para la variable "yimpuntual" en las configuraciones de la variable "x" "opcion1" y "opcion2".

Por otro lado, se registran algunos de los resultados menos favorables, con un valor promedio de 0.0013, en el modelo KNN con balanceo nulo y penalidad nula, LDA con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad "l2", NB con balanceo nulo y penalidad nula, RF con balanceo nulo y penalidad nula, y XGB con balanceo nulo y penalidad nula. Estos resultados se observaron para las variables "yrecuperacionbaja" y "ydeudor" en todas las opciones de la variable "x".

Al analizar la diagonal de la matriz de confusión, se observan resultados positivos y negativos en el rendimiento de los modelos evaluados en la clasificación de distintas clases. El componente uno de la diagonal de la matriz de confusión, asociada a las variables "yrecuperacionbaja" y "ydeudor," revela que varios modelos, como LDA, KNN, NB, XGB, RF y LOGIT, logran una clasificación excepcionalmente precisa, con valores promedio superiores a 0.95. **Es importante destacar que, a pesar de que KNN y NB obtuvieron un valor de 1 en el componente uno de la diagonal, indicativo de una alta precisión en la clasificación de la variable "yrecuperacionbaja" y "ydeudor", fallaron en el componente dos de la diagonal. Esta discrepancia resalta la importancia de evaluar el desempeño del modelo de manera integral, ya que el hecho de que el componente uno de la diagonal tenga un valor alto no necesariamente garantiza un rendimiento óptimo.** En este sentido, es crucial considerar los valores de la diagonal en conjunto para una evaluación completa y precisa del modelo. En particular, LOGIT y XGB destacan al alcanzar valores perfectos de 1.0 en ciertas configuraciones y variables. Sin embargo, se identifica un rendimiento menos favorable del modelo NB con balanceo nulo y penalidad nula para la variable "yimpuntual."

En cuanto al componente dos de la diagonal de la matriz de confusión, relacionada con la variable "yimpuntual," se destaca un rendimiento sólido en modelos como LDA, LOGIT, NB, RF y XGB, con valores promedio superiores a 0.7. Notablemente, el modelo NB muestra un rendimiento destacado con un valor promedio de 0.89 en las configuraciones de "opcion1" y "opcion2." Sin embargo, se encuentran resultados desfavorables en modelos como KNN, LDA, LOGIT, NB, RF y XGB, con un valor promedio de 0.0013,

especialmente en las variables "yrecuperacionbaja" y "ydeudor".

En la Tabla A5, se destaca que los modelos con un valor de la metrica KS promedio superior a 0.4 incluyen LDA con balanceo nulo y penalidad nula, XGB con balanceo nulo y penalidad nula, RF con balanceo nulo y penalidad nula, RF con balanceo aplicado y penalidad nula, LOGIT con balanceo nulo y penalidad nula, LOGIT con balanceo nulo y penalidad "l2", LOGIT con balanceo aplicado y penalidad nula, LOGIT con balanceo aplicado y penalidad "l2". Estos resultados aplican específicamente a la variable "yimpuntual" en todas las configuraciones de la variable "x". El conjunto de estos modelos exhibe una desviación estándar promedio de 0.0082. Los resultados más notables, con un valor promedio de 0.44, corresponden a XGB con balanceo nulo y penalidad nula, RF con balanceo nulo y penalidad nula, RF con balanceo aplicado y penalidad nula. Por otra parte, se registran algunos de los resultados menos favorables, evidenciando un valor promedio de 0.082 en el modelo KNN con balanceo nulo y penalidad nula. Estos resultados se observaron para las variables "ydeudor", "yimpuntual" y "yrecuperacionbaja" en todas las opciones de la variable "x".

2.10.2. Visualización Profunda: Distribución de Métricas Clave en Función de Variables Relevantes y Modelos

Se utilizan diagramas de caja (boxplots) para visualizar la distribución de diversas métricas de rendimiento de modelos en función de factores clave, como el tipo de modelo y la variable "y". La esencia de los boxplots radica en su capacidad para representar de manera gráfica la estadística descriptiva de un conjunto de datos, lo que facilita la interpretación de la variabilidad y las tendencias centrales.

En este contexto, el subconjunto de datos seleccionado incluye información crítica, como el modelo específico utilizado, el tipo de balanceo, la penalidad aplicada, la variable "y", la variable "x" y varias métricas de evaluación del rendimiento del modelo, tales como el área bajo la curva ROC en los conjuntos de entrenamiento y prueba, el índice KS y las componentes diagonales de la matriz de confusión.

La generación de boxplots para cada métrica permite realizar múltiples análisis visuales. Primero, estos gráficos permiten la comparación de la distribución de las métricas entre diferentes modelos, proporcionando una comprensión inmediata de las diferencias en rendimiento y posibles puntos atípicos. Además, al utilizar la variable "y" como un componente clave de la visualización, se explora cómo estas métricas varían en función de categorías específicas de la variable de interés.

La elección de boxplots como herramienta visual es especialmente útil debido a su capacidad para resumir la información estadística de una manera compacta. Cada caja representa el rango intercuartílico (IQR), proporcionando una medida de la dispersión de los datos. Las líneas dentro de las cajas indican las medianas, y los puntos atípicos se destacan, lo que facilita la identificación de posibles valores extremos.

En términos prácticos, esta visualización es esencial para entender cómo las diferentes configuraciones de modelos y las variaciones en la variable "y" influyen en el rendimiento de los modelos. Facilita la interpretación de patrones, tendencias y anomalías que podrían ser imperceptibles en una presentación de datos más convencional.

El código utilizado para la generación de estas gráficas se encuentra en el apendice B.3.

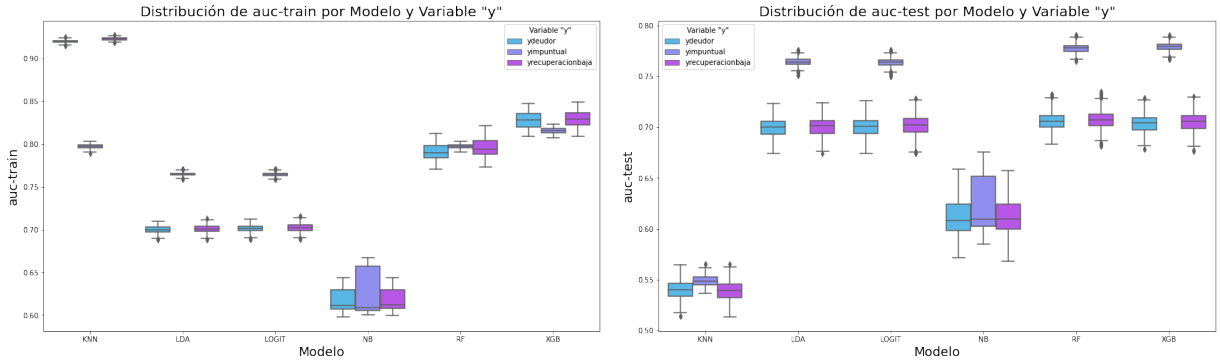

Figura 2.10.1: Gráficas de caja para las métricas AUC train y AUC test para cada modelo y variable "y"

En la Figura 2.10.1, se muestra una representación visual de la distribución de las métricas AUC-ROC para los conjuntos de datos de entrenamiento y prueba. En esta gráfica, se utiliza un diagrama de caja para mostrar la distribución de los valores.

En el diagrama de caja, el rectángulo central representa el rango intercuartílico (IQR), que abarca desde el primer cuartil (Q1) hasta el tercer cuartil (Q3). La línea central dentro

del rectángulo indica la mediana (Q2) del conjunto de datos.

Las líneas que se extienden desde el rectángulo, conocidas como "bigotes", muestran la extensión del conjunto de datos, excluyendo los valores atípicos. La longitud de estos bigotes puede variar según la distribución de los datos y la definición de valores atípicos.

Al analizar la distribución de los datos de entrenamiento, se observa que los modelos NB, RF y XGB muestran un rango más amplio en comparación con los demás modelos. Por otro lado, al examinar los datos de prueba, se aprecia que los modelos KNN y NB presentan un rango más amplio en comparación con los otros modelos.

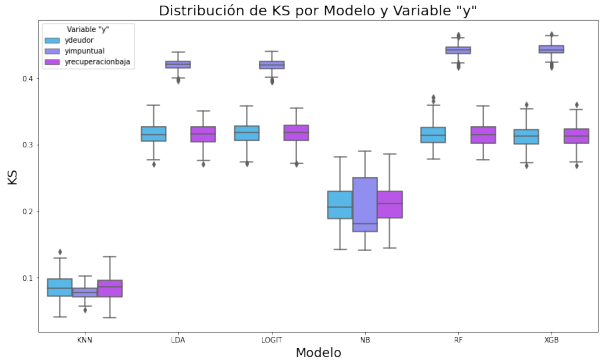

Figura 2.10.2: Gráfica de caja para la métrica Kolmogorov-Smirnov para cada modelo y variable "y"

En la Figura 2.10.2, se muestra una representación visual de la distribución de la métrica de Kolmogorov-Smirnov mediante un diagrama de caja. Mostrando una visión detallada de la distribución de los valores, incluyendo información sobre la mediana, los cuartiles y los valores atípicos.

Al analizar la distribución de los datos, se observa que el modelo NB muestra un rango más amplio en comparación con los demás modelos, lo que sugiere una mayor variabilidad en el desempeño del modelo en diferentes conjuntos de datos. Esta variabilidad puede ser importante para comprender la robustez y la estabilidad del modelo en diferentes escenarios y condiciones.

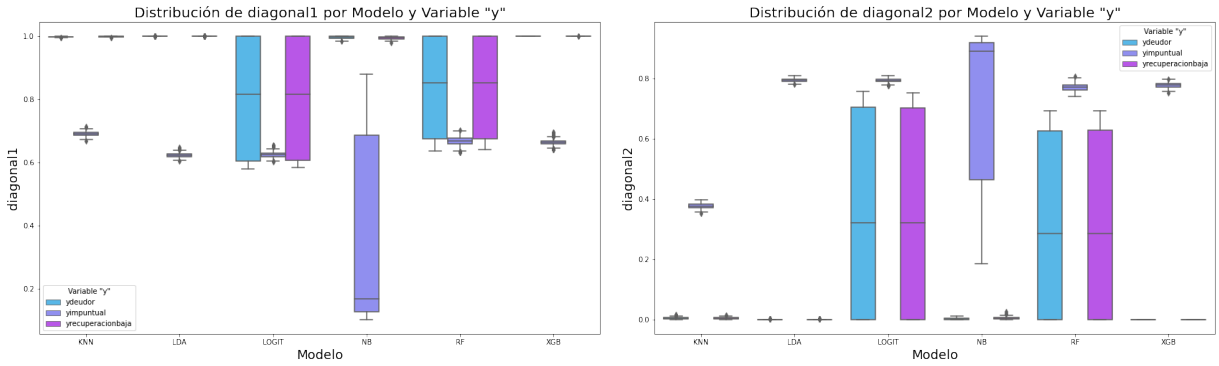

Figura 2.10.3: Gráficas de caja para los valores de los componentes uno y dos de la diagonal de la matriz de confusión para cada modelo y variable "y"

En la Figura 2.10.3, se presenta una representación visual de la distribución de los valores correspondientes a los componentes uno y dos de la diagonal de la matriz de confusión a través de un diagrama de caja. Este gráfico ofrece una visión detallada de cómo se distribuyen los valores, incluyendo información sobre la mediana, los cuartiles y los valores atípicos.

Al examinar la distribución de los datos, se observa que para el componente uno de la diagonal de la matriz de confusión, el modelo NB exhibe un rango más amplio en comparación con los demás modelos, específicamente para la variable "yimpuntual". Además, se nota que los modelos LOGIT y RF muestran un rango más amplio y cercano al valor de 1 para las variables "ydeudor" y "yrecuperacionbaja".

En cuanto al componente dos de la diagonal de la matriz de confusión, se aprecia un patrón similar. Los modelos LOGIT y RF muestran un rango amplio, especialmente para las variables "ydeudor" y "yrecuperacionbaja". A continuación, el modelo NB también exhibe un rango amplio en comparación con los demás modelos, pero únicamente para la variable "yimpuntual".

Basándonos en los resultados y observaciones presentadas en las gráficas mencionadas, podemos extraer varias conclusiones importantes sobre el desempeño de los diferentes modelos de clasificación:

1. Distribución de AUC-ROC:

    - Los modelos NB, RF y XGB muestran una mayor variabilidad en sus métricas AUC-ROC tanto en el conjunto de entrenamiento como en el de prueba, lo que

sugiere que estos modelos pueden ser más sensibles a cambios en los datos o a diferentes configuraciones de entrenamiento.

- Por otro lado, los modelos KNN y NB exhiben un rango más amplio en los datos de prueba, lo que podría indicar una mayor variabilidad en su desempeño cuando se enfrentan a datos no vistos.

2. Distribución de la métrica de Kolmogorov-Smirnov (KS)

   - El modelo NB muestra una distribución más amplia de los valores de KS en comparación con los demás modelos, lo que sugiere una mayor variabilidad en su desempeño en diferentes conjuntos de datos. Esta variabilidad es crucial para comprender la robustez y la estabilidad del modelo en diversas condiciones.

3. Componentes de la matriz de confusión

   - Para el componente uno de la diagonal de la matriz de confusión, el modelo NB exhibe un rango más amplio, especialmente para la variable "yimpuntual". Por otro lado, los modelos LOGIT y RF muestran un rango más amplio para las variables "ydeudor" y "yrecuperacionbaja".

   - En cuanto al componente dos de la diagonal de la matriz de confusión, se observa un patrón similar, donde los modelos LOGIT y RF muestran un rango amplio para las variables "ydeudor" y "yrecuperacionbaja", seguidos por el modelo NB para la variable "yimpuntual".

En general, estos hallazgos sugieren que cada modelo tiene sus propias fortalezas y debilidades en diferentes aspectos de la clasificación de datos crediticios. La variabilidad en el desempeño de los modelos, especialmente observada en el modelo NB, destaca la importancia de evaluar y comparar detenidamente múltiples modelos para seleccionar el más adecuado para una tarea específica de clasificación de riesgo crediticio. Además, el análisis detallado de las métricas y la distribución de los resultados proporciona una comprensión más profunda del comportamiento de los modelos en diferentes situaciones y escenarios, lo que puede ser crucial para la toma de decisiones informadas en el ámbito crediticio.

2.10.3. Análisis Comparativo de Métricas de Desempeño por Modelo, Balanceo, Penalidad, y Variables de Interés

En esta sección se realiza un análisis y visualización de métricas promedio para diferentes combinaciones de variables en un conjunto de datos. En los análisis que se presentan a continuación, se emplean las variables *opcion*1, *opcion*2 y *opcion*3, las cuales son conjuntos que agrupan variables explicativas. El contenido de cada variable es explicado en la sección 2.4.1.

Se seleccionan las columnas relevantes del conjunto de datos original, incluyendo información sobre el modelo, el balanceo, la penalidad, las variables "y" y "x", y las métricas de interés (AUC Train, AUC Test, KS, diagonal1, diagonal2). Esto ayuda a reducir la cantidad de datos y enfocarse en las variables clave.

Se utiliza la función "groupby" del lenguaje de programación "python", para agrupar los datos por combinaciones únicas de 'Modelo', 'balanceo', 'penalidad', 'y' y 'x'. Luego, se calcula el promedio de las métricas para cada grupo. Esto proporciona un resumen estadístico para cada combinación única de variables.

Se especifica una lista de métricas ('auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2') que se utilizarán para la visualización.

Se realiza un bucle sobre las combinaciones únicas de 'balanceo', 'penalidad' y 'x'. Dentro de este bucle, se realiza otro bucle sobre las métricas especificadas.

Para cada métrica, se crea un gráfico de barras utilizando la biblioteca Seaborn. Cada gráfico muestra el promedio de la métrica para diferentes modelos y para cada valor único de 'y'. Los colores en el gráfico representan las categorías de 'y'. La leyenda y el título del gráfico proporcionan información detallada sobre las variables utilizadas en la agrupación.

El propósito principal de estas gráficas es visualizar de manera efectiva las diferencias en las métricas promedio entre modelos, teniendo en cuenta diversas condiciones como 'balanceo', 'penalidad' y 'x'. La visualización permite una comprensión más profunda de cómo estas variables afectan las métricas de rendimiento para diferentes modelos

y categorías de 'y'. Esto puede ayudar en la toma de decisiones y la identificación de patrones en los datos. Además, la presentación detallada en los títulos y leyendas facilita la interpretación de los resultados.

A continuación se presentan algunas de las gráficas más relevantes derivadas de los resultados. El código utilizado para la generación de estas gráficas se encuentra en el apendice B.4.

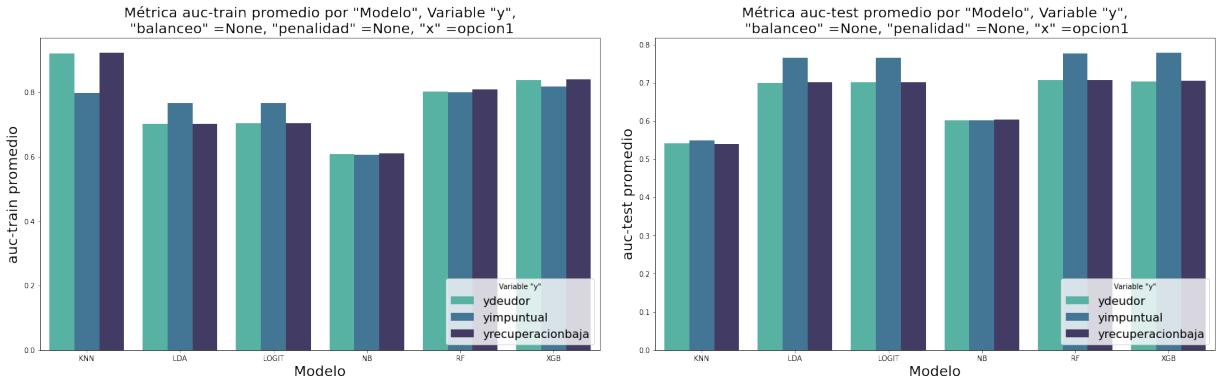

Figura 2.10.4: Gráficas de barras para las metricas AUC train y AUC test para diferentes modelos y variables "y". El balanceo y penalidad tienen valor nulo y se tienen valores de "x" de acuerdo a la lista de variables **opcion1**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, estimated-monthly-income, average-balance, min-balance, max-balance, loan-number, gender, cocienteingresodeuda, edad1,edad2, edad3, edad4, edad5. 2.4.1.

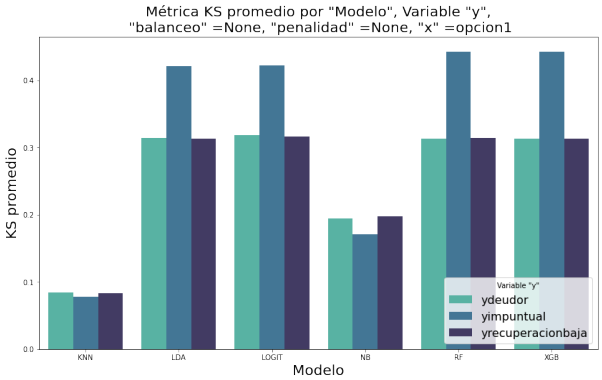

Figura 2.10.5: Gráfica de barras para la metrica KS para diferentes modelos y variables "y". El balanceo y penalidad tienen valor nulo y se tienen valores de "x" de acuerdo a la lista de variables **opcion1**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, estimated-monthly-income, average-balance, min-balance, max-balance, loan-number, gender, cocienteingresodeuda, edad1,edad2, edad3, edad4, edad5. 2.4.1.

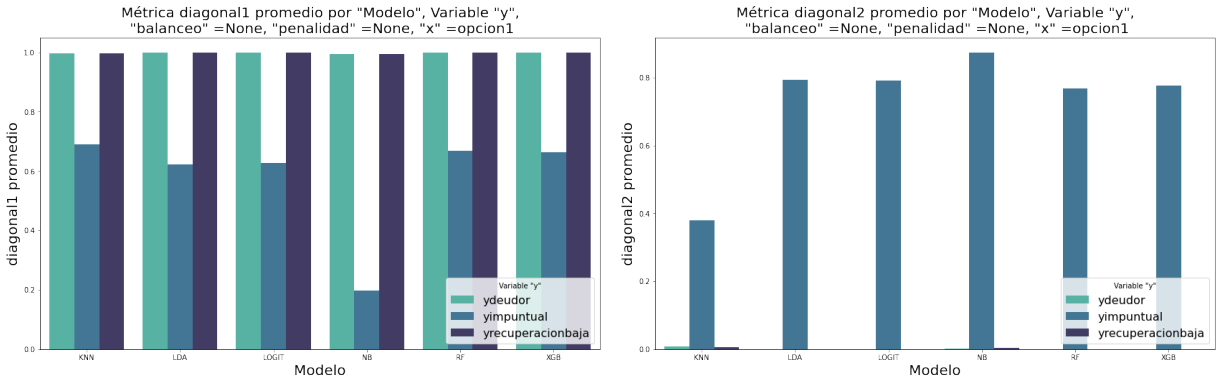

Figura 2.10.6: Gráficas de barras para los valores de los componentes uno y dos de la diagonal de la matriz de confusión para diferentes modelos y variables "y". El balanceo y penalidad tienen valor nulo y se tienen valores de "x" de acuerdo a la lista de variables **opcion1**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, estimated-monthly-income, average-balance, min-balance, max-balance, loan-number, gender, cocienteingresodeuda, edad1,edad2, edad3, edad4, edad5. 2.4.1.

Se seleccionaron las Figuras 2.10.4, 2.10.5, y 2.10.6 debido a su destacada representatividad en el escenario donde tanto el balanceo como la penalidad tienen valores nulos. Los distintos valores de la variable "x", que abarcan las opciones "opcion1", "opcion2", y "opcion3", exhiben un comportamiento sumamente similar al evidenciado en estas gráficas correspondientes a la opción "opcion1" de la variable "x".

En la Figura 2.10.4, que representa los valores de la métrica AUC Train, se evidencia que los valores más bajos corresponden al modelo NB, alcanzando aproximadamente un valor de 0.6. Los valores más altos se registran en el modelo KNN, especialmente en las variables "ydeudor" y "yrecuperacionbaja", alcanzando un valor cercano a 0.95, mientras que "yimpuntual" alcanzó un valor de 0.8. Los demás modelos se sitúan alrededor de 0.7. En la misma Figura se presentan los valores de la métrica AUC Test, se puede notar que los valores más bajos corresponden al modelo KNN, alcanzando aproximadamente un valor de 0.55. Seguido por el modelo NB, alcanzando aproximadamente un valor de 0.6. Los demas modelos muestran un comportamiento muy similar entre si, en las variables "ydeudor" y "yrecuperacionbaja", se alcanzan valores cercanos al 0.7, mientras que "yimpuntual" alcanzó un valor de 0.76. Se evidencia una disparidad significativa entre los valores promedio de AUC test y AUC train en el modelo KNN. Mientras que el AUC test muestra un valor promedio de 0.54, situándose entre los más bajos, el AUC train exhibe valores promedio de 0.88, figurando entre los más altos. Esta discrepancia indica

un caso de sobreajuste para el modelo KNN, dado que la disparidad entre ambos valores es considerablemente elevada.

En la Figura 2.10.5, que representa los valores de la métrica Kolmogorov-Smirnov, se evidencia que los valores más bajos corresponden al modelo KNN, alcanzando aproximadamente un valor de 0.09. Seguido por el modelo NB, alcanzando aproximadamente un valor de 0.2. Los demas modelos muestran un comportamiento muy similar entre si, en las variables "ydeudor" y "yrecuperacionbaja", se alcanzan valores cercanos al 0.31, mientras que "yimpuntual" alcanzó un valor de 0.43.

En la Figura 2.10.6 que representa los valores del componente uno de la diagonal de la matriz de confusión, los modelos exhiben un comportamiento muy similar entre sí en las variables "ydeudor" y "yrecuperacionbaja", alcanzando valores cercanos a 0.99. Se resalta la importancia de evaluar el desempeño del modelo de manera integral, ya que el hecho de que el componente uno de la diagonal tenga un valor alto no necesariamente garantiza un rendimiento óptimo. En este sentido, es crucial considerar los valores de la diagonal en conjunto para una evaluación completa y precisa del modelo. La variación se presenta en la variable "yimpuntual," donde el modelo NB muestra los valores más bajos, aproximadamente alrededor de 0.2, mientras que en los demás modelos se mantienen entre los valores de 0.6 a 0.7. En la misma Figura, se presentan los valores del componente dos de la diagonal de la matriz de confusión. Se observa un comportamiento peculiar, ya que para ninguno de los modelos en las variables "ydeudor" y "yrecuperacionbaja" se supera el valor de 0.01. En cuanto a los valores para "yimpuntual," se nota que el modelo KNN exhibe los valores más bajos, aproximadamente 0.4, mientras que el modelo NB alcanza valores más altos, alrededor de 0.9. Finalmente, los demás modelos se mantienen en el valor de 0.8.

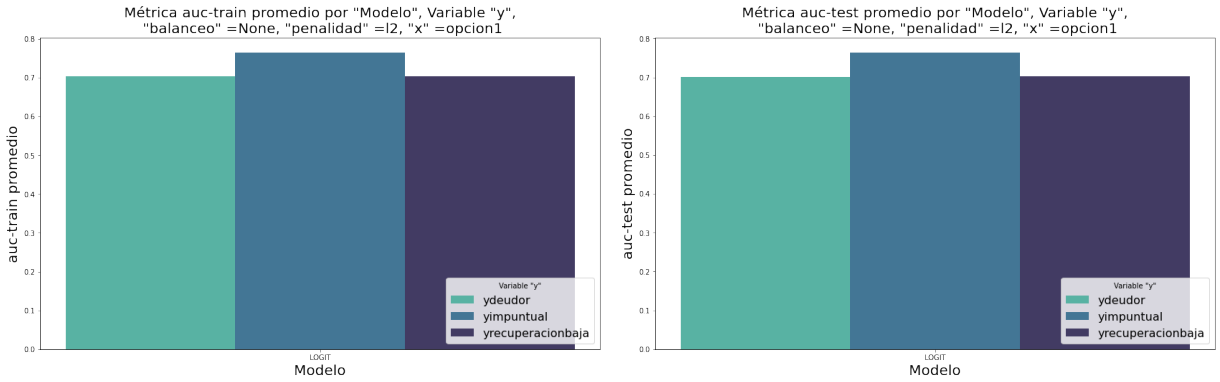

Figura 2.10.7: Gráficas de barras para las metricas AUC train y AUC test para el modelo Logit con diferentes variables "y". El balanceo tiene valor nulo, la penalidad utiliza la métrica "l2" y se tienen valores de "x" de acuerdo a la lista de variables **opcion1**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, estimated-monthly-income, average-balance, min-balance, max-balance, loan-number, gender, cocienteingresodeuda, edad1,edad2, edad3, edad4, edad5. 2.4.1.

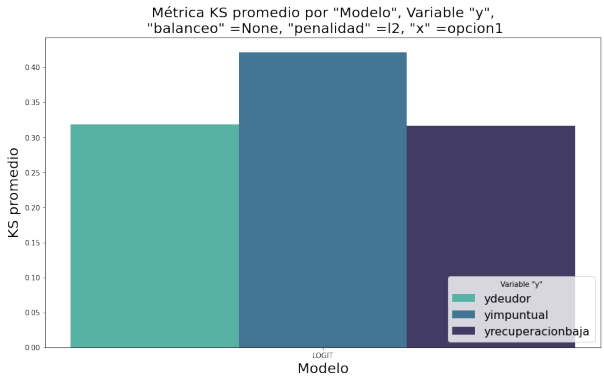

Figura 2.10.8: Gráfica de barras para la metrica KS para el modelo Logit con diferentes variables "y". El balanceo tiene valor nulo, la penalidad utiliza la métrica "l2" y se tienen valores de "x" de acuerdo a la lista de variables **opcion1**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, estimated-monthly-income, average-balance, min-balance, max-balance, loan-number, gender, cocienteingresodeuda, edad1,edad2, edad3, edad4, edad5. 2.4.1.

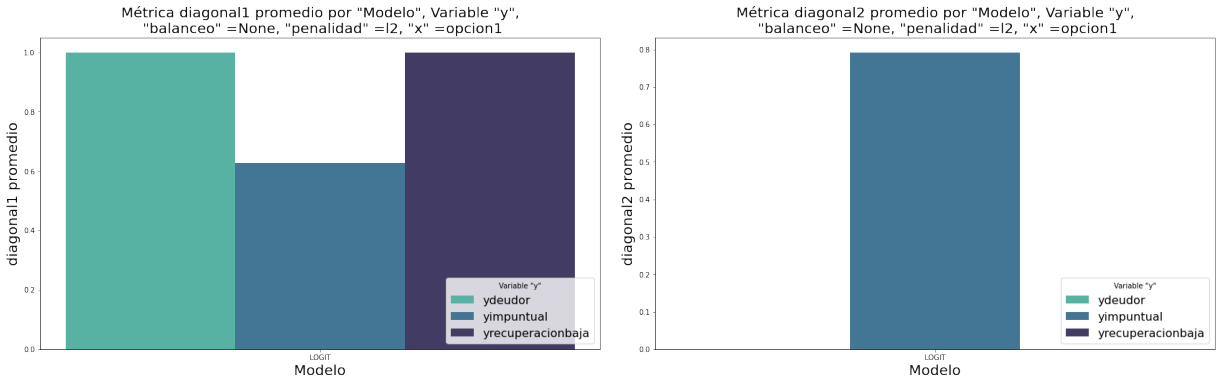

Figura 2.10.9: Gráficas de barras para las metricas de las componentes uno y dos de la diagonal de la matriz de confusión para el modelo Logit con diferentes variables "y". El balanceo tiene valor nulo, la penalidad utiliza la métrica "l2" y se tienen valores de "x" de acuerdo a la lista de variables **opcion1**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, estimated-monthly-income, average-balance, min-balance, max-balance, loan-number, gender, cocienteingresodeuda, edad1,edad2, edad3, edad4, edad5.2.4.1.

Se seleccionaron las Figuras 2.10.7, 2.10.8, y 2.10.9 debido a que son muy representativas para el escenario donde el balanceo tiene valores nulos y la penalidad utiliza la métrica "l2". Los distintos valores de la variable "x", que abarcan las opciones "opcion1", "opcion2", y "opcion3", exhiben un comportamiento bastante similar al evidenciado en estas gráficas correspondientes a la opción "opcion1" de la variable "x". Es importante destacar que solo se presenta el modelo Logit en estas gráficas, ya que es el único modelo que incluye la opción del valor "l2" en el hiperparametro de penalidad.

En la Figura 2.10.7, que refleja los resultados de la métrica AUC Train, se observa claramente que los valores más elevados se asocian con la variable "yimpuntual", llegando a alcanzar aproximadamente 0.75. Por otro lado, los valores más bajos se encuentran en las variables "ydeudor" y "yrecuperacionbaja", rondando alrededor de 0.7. En esta misma Figura, se presentan los resultados de la métrica AUC Test, y se aprecia la similitud con los datos observados en la métrica AUC Train.

En la Figura 2.10.8, que refleja los valores de la métrica Kolmogorov-Smirnov, se observa claramente que los valores más elevados se asocian con la variable "yimpuntual", llegando a alcanzar aproximadamente 0.43. Por otro lado, los valores más bajos se encuentran en las variables "ydeudor" y "yrecuperacionbaja", rondando alrededor de 0.32.

En la Figura 2.10.9 que representa los valores del componente uno de la diagonal de la

matriz de confusión, se observa claramente que los valores más elevados se asocian con las variables "ydeudor" y "yrecuperacionbaja," alcanzando valores cercanos a 0.99. Por otro lado, los valores más bajos se encuentran en la variable "yimpuntual", llegando a alcanzar aproximadamente 0.6. En la misma Figura, se presentan los valores del componente dos de la diagonal de la matriz de confusión. Se observa un comportamiento peculiar, ya que en las variables "ydeudor" y "yrecuperacionbaja" no se supera el valor de 0.01. En cuanto a los valores para "yimpuntual" se mantiene en el valor de 0.8.

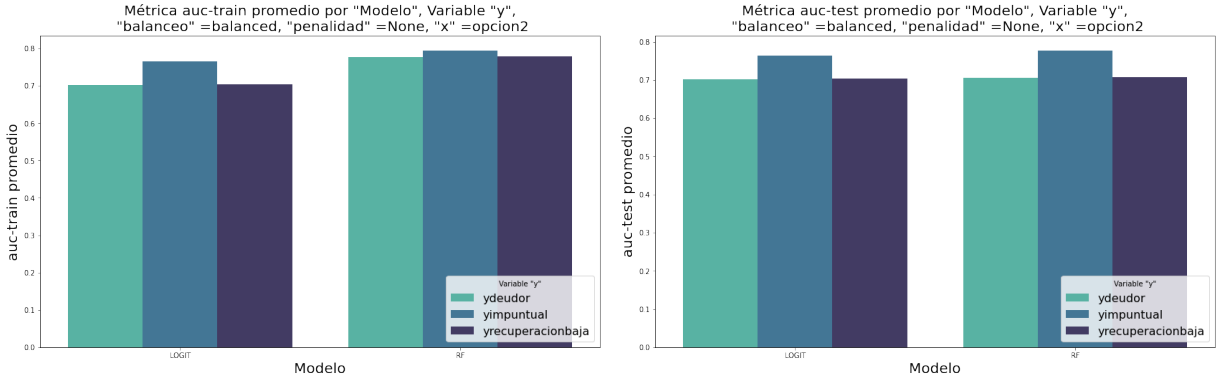

Figura 2.10.10: Gráficas de barras para las metricas AUC train y AUC test para los modelos Logit y RF con diferentes variables "y". La penalidad tiene valor nulo y se cuenta con balanceo. Se tienen valores de "x" de acuerdo a la lista de variables **opcion2**: loan-amount, number-of-payments, interest-rate, activity-age-days, estimated-monthly-income, average-balance, loan-number, gender, cocienteingresodeuda, edad1, edad2, edad3. 2.4.1.

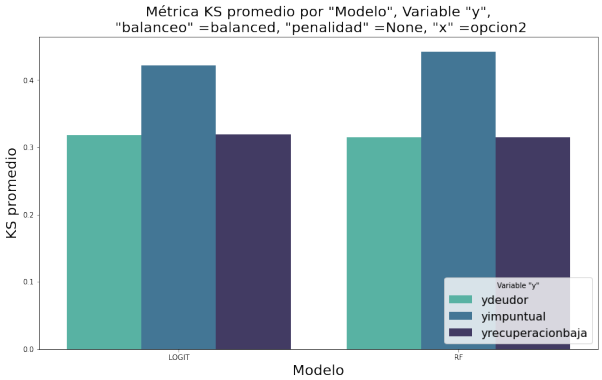

Figura 2.10.11: Gráfica de barras para la metrica KS para los modelos Logit y RF con diferentes variables "y". La penalidad tiene valor nulo y se cuenta con balanceo. Se tienen valores de "x" de acuerdo a la lista de variables **opcion2**: loan-amount, number-of-payments, interest-rate, activity-age-days, estimated-monthly-income, average-balance, loan-number, gender, cocienteingresodeuda, edad1, edad2, edad3. 2.4.1.

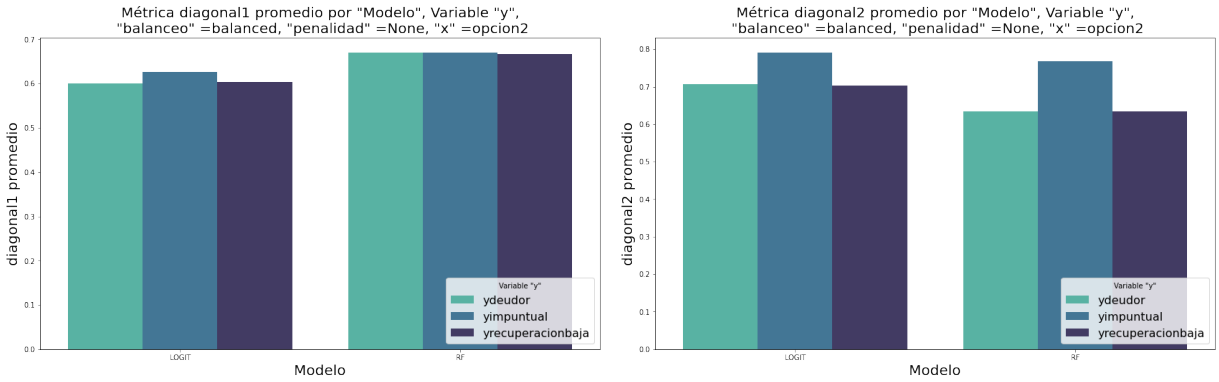

Figura 2.10.12: Gráficas de barras para los valores de los componentes uno y dos de la diagonal de la matriz de confusión para los modelos Logit y RF con diferentes variables "y". La penalidad tiene valor nulo y se cuenta con balanceo. Se tienen valores de "x" de acuerdo a la lista de variables **opcion2**: loan-amount, number-of-payments, interest-rate, activity-age-days, estimated-monthly-income, average-balance, loan-number, gender, cocienteingresodeuda, edad1, edad2, edad3. 2.4.1.

Se seleccionaron las Figuras 2.10.10, 2.10.11, y 2.10.12 debido a que son muy representativas para el escenario donde la penalidad tiene valor nulo y se cuenta con balanceo. Los distintos valores de la variable "x", que abarcan las opciones "opcion1", "opcion2", y "opcion3", exhiben un comportamiento bastante similar al evidenciado en estas gráficas correspondientes a la opción "opcion2" de la variable "x". Es importante destacar que solo se presentan los modelos Logit y RF en estas gráficas, ya que son los únicos modelos que incluyen la opción del parámetro balanceo.

En la Figura 2.10.10, que exhibe los resultados de la métrica AUC Train, se destaca que los valores más altos están asociados al modelo RF, donde la variable "yimpuntual" alcanza aproximadamente 0.8. Asimismo, las variables "ydeudor" y "yrecuperacionbaja" muestran valores cercanos a 0.78. Notablemente, el modelo Logit presenta un comportamiento muy similar al RF, con la variable "yimpuntual" alcanzando alrededor de 0.76, mientras que las variables "ydeudor" y "yrecuperacionbaja" rondan alrededor de 0.7. En esta misma Figura, se presentan los resultados de la métrica AUC Test, evidenciando la similitud entre los modelos Logit y RF. Nuevamente, la variable "yimpuntual" muestra los valores más elevados, aproximadamente 0.76, mientras que las variables "ydeudor" y "yrecuperacionbaja" se sitúan alrededor de 0.7.

En la Figura 2.10.11 que refleja los valores de la métrica Kolmogorov-Smirnov, se observa claramente una similitud en el comportamiento de ambos modelos. Los valores más

elevados se asocian con la variable "yimpuntual", llegando a alcanzar aproximadamente 0.45. Por otro lado, los valores más bajos se encuentran en las variables "ydeudor" y "yrecuperacionbaja", rondando alrededor de 0.32.

En la Figura 2.10.12, que ilustra los valores del componente uno de la diagonal de la matriz de confusión, se evidencia claramente que los valores más elevados están asociados al modelo RF, donde las variables "y" presentan valores cercanos a 0.68. En contraste, los valores más bajos se encuentran en el modelo Logit, donde las variables "y" llegan a alcanzar aproximadamente un valor de 0.6. Asimismo, en la misma Figura se presentan los valores del componente dos de la diagonal de la matriz de confusión, evidenciándose un comportamiento similar entre ambos modelos. Para el modelo Logit, las variables "ydeudor" y "yrecuperacionbaja" tienen valores de 0.7, mientras que "yimpuntual" llega hasta el valor de 0.79. En el caso del modelo RF, las variables "ydeudor" y "yrecuperacionbaja" presentan valores de 0.65, mientras que "yimpuntual" alcanza hasta el valor de 0.75.

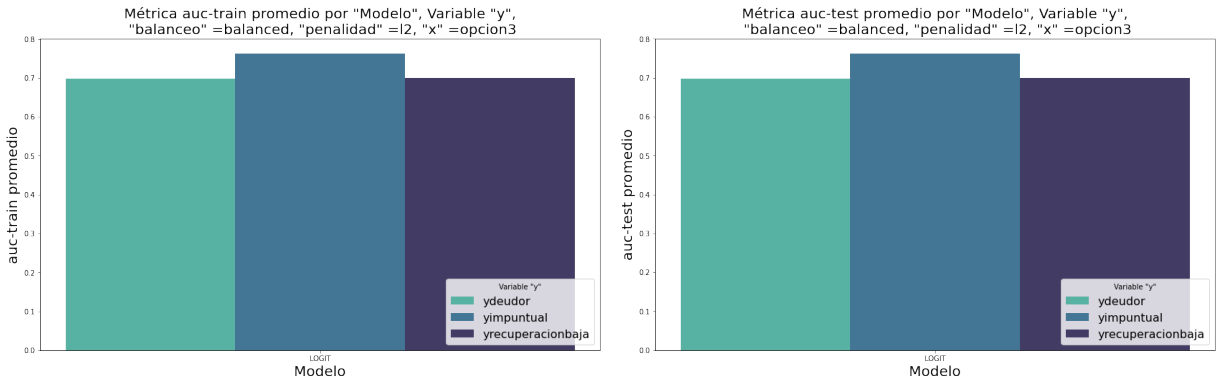

Figura 2.10.13: Gráficas de barras para las metricas AUC train y AUC test para el modelo Logit con diferentes variables "y". La penalidad utiliza la métrica "l2" y se cuenta con balanceo. Se tienen valores de "x" de acuerdo a la lista de variables **opcion3**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, average-balance, min-balance, max-balance, loan-number. 2.4.1.

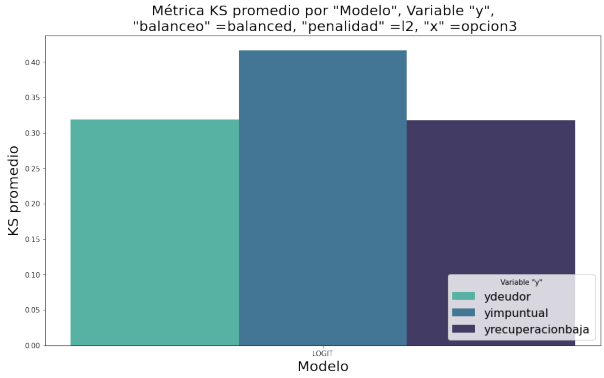

Figura 2.10.14: Gráfica de barras para la metrica KS para el modelo Logit con diferentes variables "y". La penalidad utiliza la métrica "l2" y se cuenta con balanceo. Se tienen valores de "x" de acuerdo a la lista de variables **opcion3**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, average-balance, min-balance, max-balance, loan-number. 2.4.1.

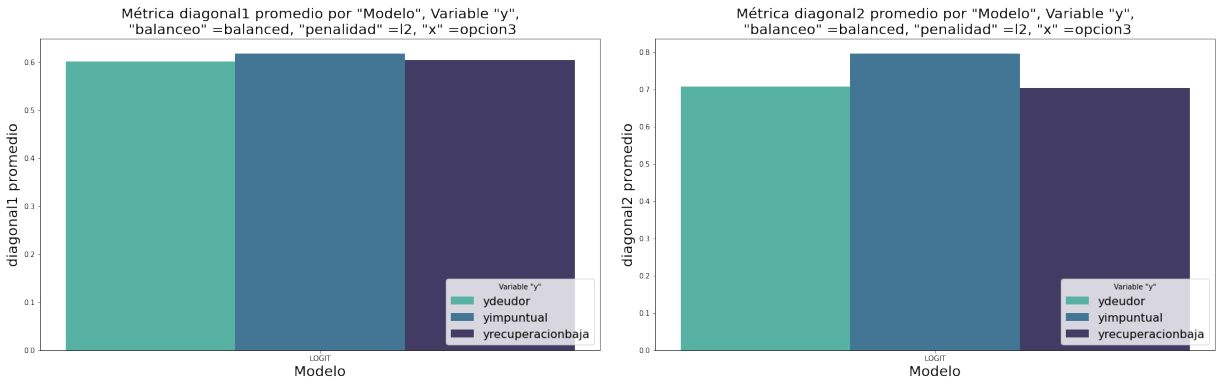

Figura 2.10.15: Gráficas de barras para los valores de los componentes uno y dos de la diagonal de la matriz de confusión para el modelo Logit con diferentes variables "y". La penalidad utiliza la métrica "l2" y se cuenta con balanceo. Se tienen valores de "x" de acuerdo a la lista de variables **opcion3**: loan-amount, number-of-payments, interest-rate, tx-found, max-amount, activity-age-days, average-balance, min-balance, max-balance, loan-number. 2.4.1.

Se seleccionaron las Figuras 2.10.13, 2.10.14, y 2.10.15 debido a que son muy representativas para el escenario donde la penalidad utiliza la métrica "l2" y se cuenta con balanceo. Los distintos valores de la variable "x", que abarcan las opciones "opcion1", "opcion2", y "opcion3", exhiben un comportamiento bastante similar al evidenciado en estas gráficas correspondientes a la opción "opcion3" de la variable "x". Es importante destacar que solo se presenta el modelo Logit en estas gráficas, ya que es el único modelo que incluye la opción del valor "l2" en el hiperparametro de penalidad.

En la Figura 2.10.13, que refleja los resultados de la métrica AUC Train, se observa

claramente que los valores más elevados se asocian con la variable "yimpuntual", llegando a alcanzar aproximadamente 0.75. Por otro lado, los valores más bajos se encuentran en las variables "ydeudor" y "yrecuperacionbaja", rondando alrededor de 0.7. En esta misma Figura, se presentan los resultados de la métrica AUC Test, y se aprecia la similitud con los datos observados en la métrica AUC Train.

En la Figura 2.10.14, que refleja los valores de la métrica Kolmogorov-Smirnov, se observa claramente que los valores más elevados se asocian con la variable "yimpuntual", llegando a alcanzar aproximadamente 0.43. Por otro lado, los valores más bajos se encuentran en las variables "ydeudor" y "yrecuperacionbaja", rondando alrededor de 0.32.

En la Figura 2.10.15 que representa los valores del componente uno de la diagonal de la matriz de confusión, se observa claramente que los valores más elevados se asocian con la variable "yimpuntual", alcanzando valores cercanos a 0.61. Por otro lado, los valores más bajos se encuentran en las variables "ydeudor" y "yrecuperacionbaja", llegando a alcanzar aproximadamente 0.6. En la misma Figura, se presentan los valores del componente dos de la diagonal de la matriz de confusión. Se observa un comportamiento en el que los valores más elevados se asocian con la variable "yimpuntual", alcanzando valores cercanos a 0.8. Por otro lado, los valores más bajos se encuentran en las variables "ydeudor" y "yrecuperacionbaja", llegando a alcanzar aproximadamente 0.71.

El modelo Naive Bayes muestra consistentemente un rendimiento inferior en términos de AUC y métrica Kolmogorov-Smirnov en comparación con otros modelos. Esto sugiere que el modelo NB tiene dificultades para discriminar eficazmente entre las clases positivas y negativas en los conjuntos de datos evaluados. Aunque es conocido por su simplicidad y eficiencia computacional, el NB puede no ser la mejor opción cuando se requiere una alta precisión en la clasificación.

El modelo KNN exhibe una discrepancia significativa entre los valores de AUC en los conjuntos de entrenamiento y prueba, lo que sugiere un caso de sobreajuste. Aunque el KNN puede adaptarse bien a los datos de entrenamiento, su capacidad para generalizar a datos nuevos y no vistos parece ser limitada en este caso. Esto destaca la importancia de ajustar adecuadamente los hiperparámetros del modelo KNN y aplicar técnicas de validación cruzada para evitar el sobreajuste y mejorar su capacidad predictiva.

La regresión logística muestra un rendimiento relativamente consistente en términos de AUC y métrica Kolmogorov-Smirnov en comparación con otros modelos. Aunque no exhibe los valores más altos de AUC, la regresión logística es conocida por su interpretabilidad y facilidad de implementación. Su rendimiento estable en diferentes configuraciones sugiere que puede ser una opción sólida para problemas de clasificación en los que se requiere una comprensión clara de las relaciones entre las variables predictoras y la variable objetivo.

El modelo Random Forest muestra un rendimiento sólido en términos de AUC, matriz de confusión y métrica Kolmogorov-Smirnov en varias configuraciones. La naturaleza en conjunto de los árboles de decisión en el bosque aleatorio permite que el modelo capture relaciones complejas entre las variables predictoras y la variable objetivo, lo que resulta en un buen rendimiento predictivo en una variedad de escenarios. Sin embargo, es importante tener en cuenta que el RF puede ser propenso al sobreajuste, especialmente en conjuntos de datos pequeños o altamente desequilibrados. A pesar de esto, no identificamos problemas serios de sobreajuste en nuestros análisis, y no encontramos un criterio objetivo con el cual identificar problemas severos de sobreajuste. Esto sugiere que, bajo las condiciones y configuraciones utilizadas, el modelo Random Forest mantiene su capacidad de generalización sin incurrir en un sobreajuste significativo.

En general, estos modelos representan una gama diversa de enfoques en el aprendizaje automático y tienen sus propias fortalezas y debilidades en diferentes contextos y aplicaciones. El análisis detallado de su rendimiento en términos de métricas clave como AUC y Kolmogorov-Smirnov proporciona información valiosa para la selección y optimización de modelos.

Capítulo 3

Análisis comparativo entre el comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC

En el ámbito de la evaluación de modelos de clasificación el análisis basado en el estadístico de Kolmogorov-Smirnov (KS) y el Área Bajo la Curva ROC (AUC-ROC) es frecuente. Ambos son métodos estadísticos fundamentales utilizados para evaluar el rendimiento predictivo de los modelos en diversas disciplinas. Sin embargo, no han sido suficientemente discutidas la relación y diferencias entre ambos.

El estadístico de Kolmogorov-Smirnov se basa en la comparación entre dos distribuciones de probabilidad, proporcionando una medida de la distancia máxima entre las funciones de distribución acumulativa empírica de dos muestras.

A pesar de sus enfoques diferentes, estos dos métodos están interrelacionados en la evaluación del rendimiento del modelo. En este trabajo de investigación se observa que existe una relación entre el comportamiento del estadístico KS y el AUC-ROC en el contexto del ajuste del modelo y la generalización a nuevos datos. Por ejemplo, durante el análisis de modelos predictivos, se ha notado que cuando tanto el AUC-ROC de entrenamiento como el AUC-ROC de prueba aumentan o disminuyen, la métrica de

Kolmogorov-Smirnov también sigue la misma tendencia. Esta observación sugiere una relación entre la capacidad discriminativa del modelo y la diferencia máxima entre las funciones de distribución de probabilidad acumulativa empírica de las clases positiva y negativa. A medida que el modelo es más capaz de distinguir entre las clases positiva y negativa (mayor capacidad discriminativa), se espera que la diferencia entre las funciones de distribución acumulativa empírica sea mayor. En otras palabras, cuando el modelo es más preciso en sus predicciones, la separación entre las distribuciones de probabilidad acumulativa de las clases positiva y negativa tiende a ser más pronunciada.

En esta sección, exploraremos en detalle estas relaciones y su implicación en la evaluación y comparación de modelos predictivos, proporcionando una visión integral del uso tanto del estadístico de Kolmogorov-Smirnov así como del AUC-ROC en la evaluación del rendimiento de los modelos de clasificación.

A continuación, se presentarán una serie de gráficas que ilustran el comportamiento discutido anteriormente. Estas visualizaciones ofrecen una representación visual clara de cómo varían tanto el estadístico de Kolmogorov-Smirnov así como el Área Bajo la Curva ROC (AUC) en función de diferentes condiciones y escenarios.

Las gráficas permitirán una comparación directa entre el comportamiento de ambos estadísticos en distintas situaciones, lo que ayudará a identificar posibles patrones o tendencias. Además, estas representaciones visuales facilitarán la interpretación de los resultados y proporcionarán una visión más intuitiva de la relación entre la capacidad discriminativa del modelo y la diferencia entre las funciones de distribución acumulativa empírica.

Cada gráfica estará acompañada de una descripción detallada que destacará los puntos clave y las conclusiones derivadas de la observación de los patrones presentados. Esta visualización será fundamental para comprender mejor la relación entre el estadístico de Kolmogorov-Smirnov y el AUC-ROC y su utilidad en la evaluación del rendimiento de modelos de clasificación en diversos contextos. El código utilizado para la generación de estas gráficas se encuentra en el apendice B.5. Las siguientes gráficas, desde la Figura 3.1 hasta la Figura 3.8, comparan el estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC, abarcando todas las variaciones de los modelos evaluados, tanto con penalización como sin ella, y tanto con balanceo como sin balanceo.

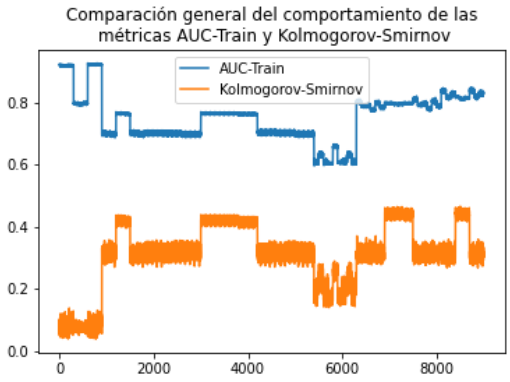

Figura 3.1: Gráfica del comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC para los datos de entrenamiento, para los modelos KNN, LDA, Logit, NB, RF y XGB. Con las variables objetivo ydeudor, yimpuntual y yrecuperacionbaja.

En la Figura 3.1 se muestra el comportamiento del estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC para los datos de entrenamiento, abarcando todos los modelos evaluados tanto con penalización como sin ella, y tanto con balanceo como sin balanceo, tomando en cuenta los tres conjuntos de variables x (opcion1, opcion2 y opcion3 definidas en la sección 2.4.1) asi como las tres variables y (ydeudor, yimpuntual y yrecuperacionbaja definidas en 2.3). Se puede observar que, a pesar de presentar ligeras desviaciones, existe una consistencia notable entre ambas métricas, independientemente de los parámetros de entrada que puedan tomar los modelos. Esta coherencia sugiere una relación cercana entre la capacidad discriminativa de los modelos, evaluada a través del estadístico de Kolmogorov-Smirnov, y su capacidad predictiva global, evaluada mediante el Área Bajo la Curva ROC. La comparación de estas dos métricas proporciona una comprensión más completa del rendimiento de los modelos y ayuda a identificar posibles inconsistencias o áreas de mejora en su capacidad de clasificación.

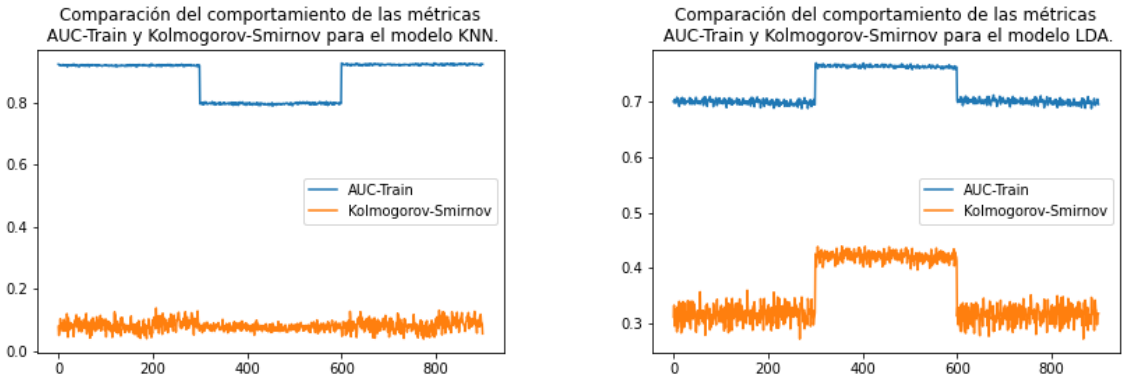

Figura 3.2: Gráfica del comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC para los datos de entrenamiento, para los modelos KNN y LDA. Con las variables objetivo ydeudor, yimpuntual y yrecuperacionbaja.

En la Figura 3.2, se analiza el comportamiento del estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC para los datos de entrenamiento, centrándose en los modelos KNN y LDA. Para el modelo LDA, se observan ligeras desviaciones entre ambas métricas, lo que sugiere variaciones en su capacidad discriminativa y predictiva. En contraste, el modelo KNN exhibe una consistencia notable entre el comportamiento de ambas métricas, indicando una relación estrecha entre su capacidad discriminativa y el Área Bajo la Curva ROC.

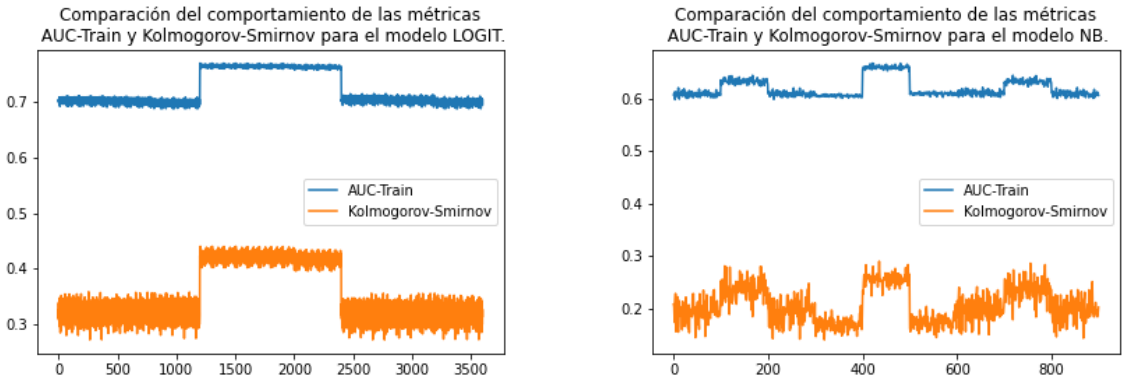

Figura 3.3: Gráfica del comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC para los datos de entrenamiento, para los modelos LOGIT y NB. Con las variables objetivo ydeudor, yimpuntual y yrecuperacionbaja.

En la Figura 3.3, se analiza el comportamiento del estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC para los datos de entrenamiento, centrándose en los modelos LOGIT y NB. Para ambos modelos LOGIT y NB, se muestra una consistencia notable

entre el comportamiento de ambas métricas, indicando una relación estrecha entre su capacidad discriminativa y el Área Bajo la Curva ROC.

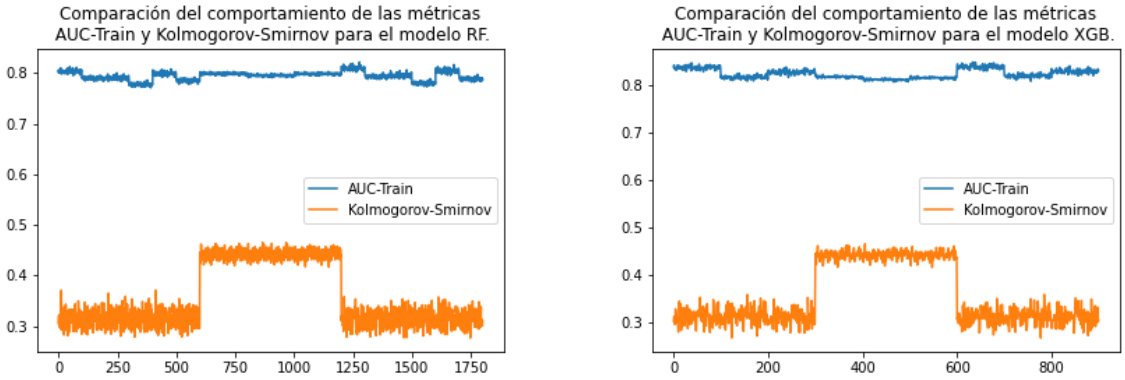

Figura 3.4: Gráfica del comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC para los datos de entrenamiento, para los modelos RF y XGB. Con las variables objetivo ydeudor, yimpuntual y yrecuperacionbaja.

En la Figura 3.4, se analiza el comportamiento del estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC para los datos de entrenamiento, centrándose en los modelos RF y XGB. Para ambos modelos RF y XGB, se observan desviaciones entre ambas métricas, lo que sugiere variaciones en su capacidad discriminativa y predictiva.

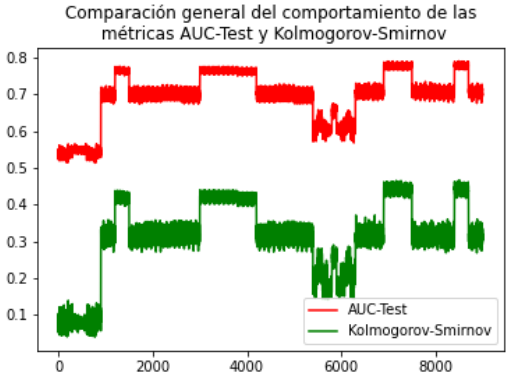

Figura 3.5: Gráfica del comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC para los datos de prueba, para los modelos KNN, LDA, Logit, NB, RF y XGB. Con las variables objetivo ydeudor, yimpuntual y yrecuperacionbaja.

En la Figura 3.5 se muestra el comportamiento del estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC para los datos de prueba, abarcando todos los modelos evaluados tanto con penalización como sin ella, y tanto con balanceo como sin balanceo, tomando en cuenta los tres conjuntos de variables x (opcion1, opcion2 y opcion3 definidas en la sección

2.4.1) asi como las tres variables y (ydeudor, yimpuntual y yrecuperacionbaja definidas en 2.3). Se puede observar que, existe una consistencia muy notable entre ambas métricas, de manera casi identica, independientemente de los parámetros de entrada que puedan tomar los modelos. Esta coherencia sugiere una relación cercana entre la capacidad discriminativa de los modelos, evaluada a través del estadístico de Kolmogorov-Smirnov, y su capacidad predictiva global, evaluada mediante el Área Bajo la Curva ROC. La comparación de estas dos métricas proporciona una comprensión más completa del rendimiento de los modelos y ayuda a identificar posibles inconsistencias o áreas de mejora en su capacidad de clasificación.

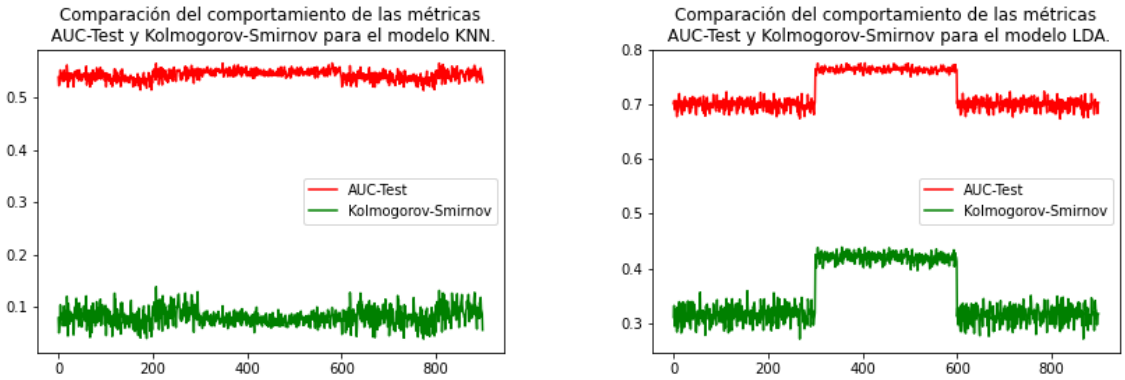

Figura 3.6: Gráfica del comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC para los datos de prueba, para los modelos KNN y LDA. Con las variables objetivo ydeudor, yimpuntual y yrecuperacionbaja.

En la Figura 3.6, se analiza el comportamiento del estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC para los datos de prueba, centrándose en los modelos KNN y LDA. Para ambos modelos, se exhibe una consistencia bastante similar entre el comportamiento de ambas métricas, indicando una relación estrecha entre su capacidad discriminativa y el Área Bajo la Curva ROC.

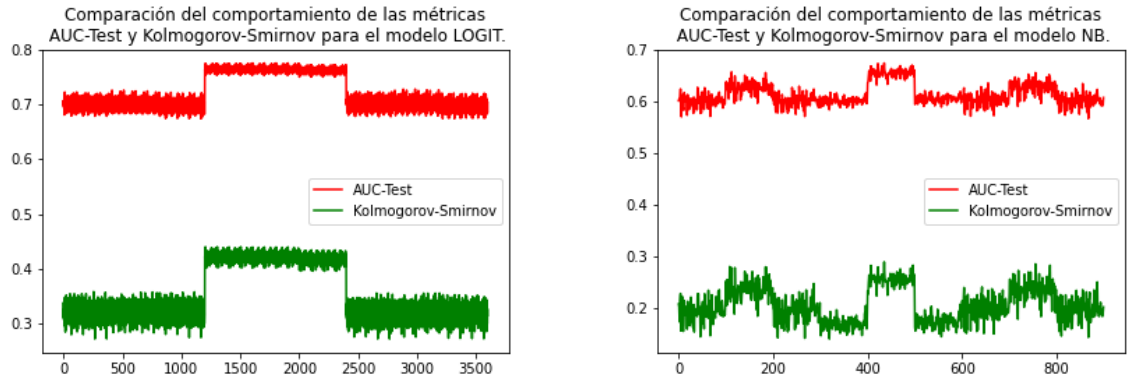

Figura 3.7: Gráfica del comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC para los datos de prueba, para los modelos LOGIT y NB. Con las variables objetivo ydeudor, yimpuntual y yrecuperacionbaja.

En la Figura 3.7, se examina el desempeño del estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC para los datos de prueba, específicamente enfocándose en los modelos LOGIT y NB. Ambos modelos muestran una coherencia notable entre el comportamiento de estas métricas, lo que sugiere una relación cercana entre su capacidad discriminativa. Este hallazgo subraya la consistencia en el rendimiento predictivo de los modelos LOGIT y NB en este contexto particular de evaluación.

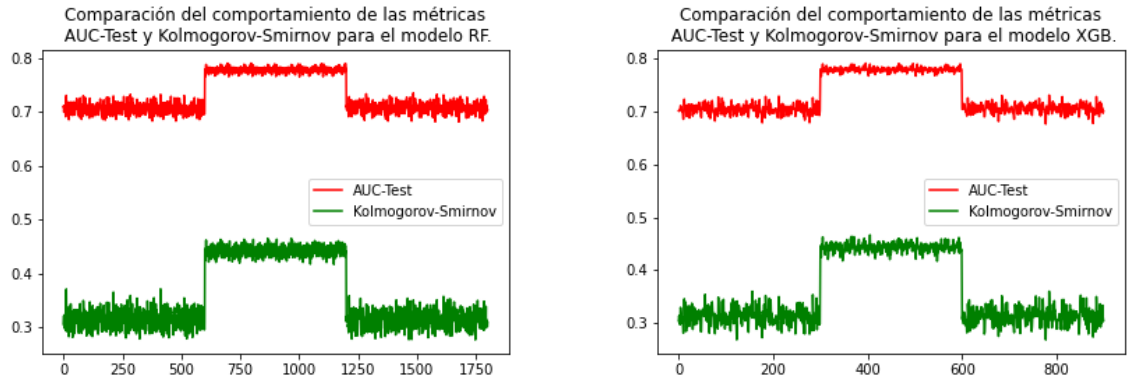

Figura 3.8: Gráfica del comportamiento del estadístico de Kolmogorov-Smirnov y el Area Bajo la Curva ROC para los datos de prueba, para los modelos RF y XGB. Con las variables objetivo ydeudor, yimpuntual y yrecuperacionbaja.

En la Figura 3.8, se analiza el comportamiento del estadístico de Kolmogorov-Smirnov y el Área Bajo la Curva ROC para los datos de prueba, centrándose en los modelos RF y XGB. Para ambos modelos, se exhibe una consistencia bastante similar entre el comportamiento de ambas métricas, indicando una relación estrecha entre su capacidad discriminativa y el Área Bajo la Curva ROC.

El análisis de las figuras que muestran el comportamiento del estadístico de Kolmogorov-Smirnov (KS) y el Área Bajo la Curva ROC (AUC-ROC) ofrece una visión detallada del rendimiento de los modelos de clasificación en diferentes conjuntos de datos y condiciones. A partir de estas observaciones, se pueden extraer varias conclusiones importantes:

Relación entre KS y AUC-ROC: En las figuras que muestran los resultados tanto para los datos de entrenamiento como para los de prueba, se observa una consistencia notable entre el comportamiento de las métricas KS y AUC-ROC. Esta coherencia sugiere una relación cercana entre la capacidad discriminativa de los modelos, evaluada mediante el estadístico KS, y su capacidad predictiva global, evaluada a través del AUC-ROC. Esta relación es crucial ya que proporciona una comprensión más completa del rendimiento de los modelos y ayuda a identificar posibles áreas de mejora.

Variabilidad entre modelos: Se observan ligeras desviaciones entre las métricas KS y AUC-ROC para diferentes modelos. Por ejemplo, en la Figura 3.2 se muestra que el modelo LDA exhibe ligeras variaciones entre ambas métricas, lo que sugiere variaciones en su capacidad discriminativa y predictiva. En contraste, el modelo KNN muestra una consistencia notable entre ambas métricas. Este análisis resalta la importancia de evaluar la consistencia del rendimiento de los modelos en diferentes contextos.

Comparación entre modelos: Al comparar el comportamiento de las métricas KS y AUC-ROC para diferentes modelos, se pueden identificar patrones de rendimiento distintivos. Por ejemplo, en la Figura 3.4 se observa que los modelos RF y XGB muestran desviaciones entre ambas métricas, lo que sugiere variaciones en su capacidad discriminativa y predictiva. Esta comparación permite identificar modelos que pueden ser más robustos en diferentes escenarios y condiciones.

Consistencia en los datos de prueba: En las figuras que muestran los resultados para los datos de prueba, se observa una consistencia muy notable entre las métricas KS y AUC-ROC. Esta coherencia indica una capacidad predictiva sólida de los modelos en un conjunto de datos no visto previamente. Esto sugiere que los modelos están generalizando bien a datos nuevos y desconocidos, lo cual es fundamental para su aplicabilidad en la práctica.

El análisis conjunto de las métricas KS y AUC-ROC proporciona una evaluación integral del

rendimiento de los modelos de clasificación. Estas métricas permiten entender la capacidad discriminativa y predictiva de los modelos en diferentes conjuntos de datos y condiciones, lo que es fundamental para tomar decisiones informadas sobre su implementación y mejora.

Capítulo 4

Conclusiones

En este trabajo de tesis partiendo de una base de datos con información real de una cartera de crédito de un banco cuyo nombre se mantiene confidencial, se ha explorado el desempeño de siete modelos de aprendizaje de maquina para la predicción de tres eventos crediticios. Se definieron tres diferentes eventos crediticios que dieron lugar a tres diferentes variables dicotómicas a explicar y que llamamos “yimpuntual”, "yrecuperacionbaja" y "ydeudor". Se delimitaron tres conjuntos de variables explicativas que llamamos opcion1, opcion2 y opcion3. Para medir el rendimiento de los modelos se utilizaron el AUC, los componentes de la diagonal de la matriz de confusión y el estadístico de Kolmogorov-Smirnov. En total se investigaron sesenta y tres combinaciones. De este ejercicio encontramos los siguientes hallazgos:

- Con respecto a las variables a explicar: Solo para una de las tres variables se obtuvieron buenos resultados, esta fue la variable que llamamos “yimpuntual”.
- Profundizando en el punto anterior, se encontró evidencia de que todos los modelos son vulnerables al desbalance de datos. Lo anterior se detecta en el hecho de que todos los modelos presentan un bajo rendimiento para las variables “yrecuperacionbaja” y “ydeudor” que tienen un desbalance grande (6.7 % y 7.06 % respectivamente de presencia del evento crediticio) contra el mejor desempeño en la variable “yimpuntual” (altamente equilibrada con 50.69 % de presencia del evento).
- Sistemáticamente los modelos NB y KNN dieron los peores resultados.
- Se detectó una fuerte presencia del fenómeno de sobreajuste para el modelo KNN

visualizado en gran disparidad en AUC de entrenamiento y prueba.

- En cambio los dos modelos basados en arboles de decisión tuvieron sistemáticamente un buen desempeño para la variable “yimpuntual”, como se aprecia en la Tabla 4.0.1. Estos son los modelos de bosques aleatorios y XGBoost.

-	Prediccion Falso	Prediccion Verdadero
Real Falso	0.663	0.337
Real Verdadero	0.223	0.777
AUC Test	0.779	

Cuadro 4.0.1: Matriz de confusión para el modelo XGB. Se utilizan valores de "x" de acuerdo a la lista de variables **opcion1**: *loan_amount*, *number_of_payments*, *interest_rate*, *tx_found*, *max_amount*, *activity_age_days*, *estimated_monthly_income*, *average_balance*, *min_balance*, *max_balance*, *loan_number*, *gender*, *cocienteingresodeuda*, *edad*1, *edad*2, *edad*3, *edad*4, *edad*5. Los resultados presentados son para la variable **"yimpuntual"**. Se realizaron cien repeticiones.

- El modelo LOGIT obtuvo resultados en la matriz de confusión y valores de AUC bastante consistentes para las diferentes variables "ydeudor", "yimpuntual" y "yrecuperacionbaja". Esto se presentó, tanto con balanceo como sin él, y con penalización L2 o nula. Esto se puede ver en las Tablas 4.0.2, 4.0.3, 4.0.4, donde se presentan los resultados para el modelo balanceado y con penalidad l2.

-	Prediccion Falso	Prediccion Verdadero
Real Falso	0.608	0.392
Real Verdadero	0.302	0.698
AUC Test	0.7	

Cuadro 4.0.2: Matriz de confusión para el modelo LOGIT, balanceado y con penalidad "l2". Se utilizan valores de "x" de acuerdo a la lista de variables **opcion1**: *loan_amount*, *number_of_payments*, *interest_rate*, *tx_found*, *max_amount*, *activity_age_days*, *estimated_monthly_income*, *average_balance*, *min_balance*, *max_balance*, *loan_number*, *gender*, *cocienteingresodeuda*, *edad*1, *edad*2, *edad*3, *edad*4, *edad*5. Los resultados presentados son para la variable **"ydeudor"**. Se realizaron cien repeticiones.

-	Prediccion Falso	Prediccion Verdadero
Real Falso	0.628	0.372
Real Verdadero	0.209	0.791
AUC Test	0.764	

Cuadro 4.0.3: Matriz de confusión para el modelo LOGIT, balanceado y con penalidad "l2". Se utilizan valores de "x" de acuerdo a la lista de variables **opcion1**: *loan_amount*, *number_of_payments*, *interest_rate*, *tx_found*, *max_amount*, *activity_age_days*, *estimated_monthly_income*, *average_balance*, *min_balance*, *max_balance*, *loan_number*, *gender*, *cocienteingresodeuda*, *edad*1, *edad*2, *edad*3, *edad*4, *edad*5. Los resultados presentados son para la variable **"yimpuntual"**. Se realizaron cien repeticiones.

-	Prediccion Falso	Prediccion Verdadero
Real Falso	0.612	0.388
Real Verdadero	0.307	0.693
AUC Test	0.702	

Cuadro 4.0.4: Matriz de confusión para el modelo LOGIT, balanceado y con penalidad "l2". Se utilizan valores de "x" de acuerdo a la lista de variables **opcion1**: *loan_amount*, *number_of_payments*, *interest_rate*, *tx_found*, *max_amount*, *activity_age_days*, *estimated_monthly_income*, *average_balance*, *min_balance*, *max_balance*, *loan_number*, *gender*, *cocienteingresodeuda*, *edad*1, *edad*2, *edad*3, *edad*4, *edad*5. Los resultados presentados son para la variable **"yrecuperacionbaja"**. Se realizaron cien repeticiones.

Los resultados detallados de la matriz de confusión, estadístico de Kolmogorov-Smirnov, AUC de prueba y entrenamiento, para todos los tipos de modelos, variables y parámetros utilizados se encuentran en las Tablas A1, A2, A3, A4, A5, del apendice.

En el ejercicio de explorar sistemáticamente el desempeño de siete modelos encontramos empíricamente una fuerte relación monótona entre el AUC y el estadístico KS. Esta es una relación que apriori se anticipa razonable pues ambas metricas miden la calidad predictiva de un modelo, sin embargo, no encontramos referencias bibliograficas en donde se haga esta observación. Consideramos que éste es un fenómeno interesante que vale la pena estudiarse más, tanto empírica como teóricamente.

Al explorar siete diferentes modelos de aprendizaje de maquina encontramos que los modelos de bosques aleatorios y XGBoost tuvieron el mejor rendimiento, en concordancia con el resultado en [1] quienes recomiendan el uso de dicho modelo. Para estos modelos se obtuvó la marca de 0.77 AUC en promedio, lo cual indica buenos resultados. Ahora

bien, para buscar mejorarlos y convertirlos en muy buenos y superar la marca de 0.8 AUC se deben explorar algunas técnicas adicionales que para mantener el presente trabajo en un tamaño razonable no se han desarrollado aquí. No obstante este es una tarea que vale la pena realizar, y que si bien se deja fuera en la tesis, se abordará en investigación futura. En específico, se pueden hacer las siguientes acciones: investigar sistemáticamente la definición de hiperparametros, investigar el efecto de balancear datos en especial con técnicas avanzadas (GAN's), investigar el efecto de interacción de variables, investigar el desempeño de otros modelos basados en arboles de decisión como por ejemplo el modelo lightGBM que en la literatura se reporta da muy buenos resultados. [7]

Para calibrar los resultados obtenidos y probar que la metodología propuesta funciona, se realizaron evaluaciones exhaustivas utilizando datos de prueba y datos de entrenamiento. Los modelos de clasificación fueron evaluados mediante las métricas de Área Bajo la Curva ROC (AUC), el estadístico de Kolmogorov-Smirnov (KS) y la matriz de confusión.

Los resultados mostraron que los modelos lograron un desempeño consistente y robusto. El análisis del AUC indicó una alta capacidad discriminativa de los modelos, con valores cercanos a 1 tanto en los datos de entrenamiento como en los datos de prueba, lo que sugiere una fuerte capacidad predictiva y una adecuada generalización a datos no vistos. Además, el estadístico de Kolmogorov-Smirnov mostró diferencias mínimas entre las distribuciones acumuladas de las predicciones y las observaciones reales, lo cual respalda la fiabilidad de los modelos.

La matriz de confusión proporcionó una visión detallada del desempeño de cada modelo. Mostrarón que el modelo XGBoost tuvo el mejor desempeño, como se puede ver en la tabla 4.0.1 para la variable "yimpuntual". Se obtuvieron valores en la diagonal de la matriz de confusión de 0.663 (Falsos correctamente predichos) y 0.777 (Verdaderos correctamente predichos) con un AUC Test de 0.779. Estos valores fueron el promedio obtenido de las 100 repeticiones hechas. Para este modelo se obtuvieron los resultados mas altos comparados con los otros modelos usados. A continuación para referencia se presentan los resultados del modelo LOGIT.

Para el modelo LOGIT, se puede ver en la tabla 4.0.2 para la variable "ydeudor", los valores obtenidos en la diagonal de la matriz de confusión de 0.608 (Falsos correctamente predichos) y 0.698 (Verdaderos correctamente predichos) con un AUC Test de 0.7. Estos

valores fueron el promedio obtenido de las 100 repeticiones hechas.

Para el modelo LOGIT, se puede ver en la tabla 4.0.3 para la variable "yimpuntual", los valores obtenidos en la diagonal de la matriz de confusión de 0.628 (Falsos correctamente predichos) y 0.791 (Verdaderos correctamente predichos) con un AUC Test de 0.764. Estos valores fueron el promedio obtenido de las 100 repeticiones hechas.

Para el modelo LOGIT, se puede ver en la tabla 4.0.4 para la variable "yrecuperacionbaja", los valores obtenidos en la diagonal de la matriz de confusión de 0.612 (Falsos correctamente predichos) y 0.693 (Verdaderos correctamente predichos) con un AUC Test de 0.702. Estos valores fueron el promedio obtenido de las 100 repeticiones hechas.

También evidenció que los modelos LDA y LOGIT presentaron una notable coherencia entre las métricas KS y AUC, indicando una relación estrecha entre su capacidad discriminativa y el Área Bajo la Curva ROC; ver el Capítulo 3.

El uso de estas métricas para evaluar tanto los datos de entrenamiento como los de prueba confirmó que la metodología propuesta no solo es funcional, sino también robusta y precisa. La coherencia en los resultados de las diferentes métricas asegura que los modelos están bien calibrados y son capaces de ofrecer predicciones fiables y precisas en escenarios prácticos.

Bibliografía

[1] Björn Rafn Gunnarsson and Seppe vanden Broucke and Bart Baesens and María Óskarsdóttir and Wilfried Lemahieu (2021). Deep learning for credit scoring: Do or don't? *European Journal of Operational Research*.

[2] Boschetti, A. and Massaron, L. (2015). *Python data science essentials*. Packt Publishing Ltd.

[3] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.".

[4] Greenacre, M. J. (1984). Theory and applications of correspondence analysis.

[5] Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

[6] James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.

[7] Lextrait, B. (2023). Scaling up smes' credit scoring scope with lightgbm. *Applied Economics*, 55(9):925–943.

[8] Pearson, E. S. (1931). The test of significance for the correlation coefficient. *Journal of the American Statistical Association*, 26(174):128–134.

[9] Pestman, W. R. (2009). *Mathematical statistics*. Walter de Gruyter.

[10] R Core Team (2017). R: A language and environment for statistical computing. https://www.R-project.org/.

[11] Scikit-learn (2024). Logistic regression. Accessed: 2024-06-23 https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

[12] Siddiqi, N. (2012). *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons.

Apéndice

Apéndice A

Tablas de métricas comparativas en el desempeño de los modelos

El código utilizado para la generación de estas tablas descriptivas se encuentra en el apendice B.2.

Cuadro A1: Tabla Métrica AUC Test.

Modelo	balanceo	penalidad	y	x	mean	std	min	25 %	50 %	75 %	max
KNN	None	None	ydeudor	opcion1	0.540646	0.008426	0.519708	0.536265	0.540399	0.546192	0.561695
KNN	None	None	ydeudor	opcion2	0.534824	0.008149	0.514382	0.530369	0.534851	0.539330	0.554296
KNN	None	None	ydeudor	opcion3	0.544381	0.008537	0.524021	0.538940	0.543485	0.550983	0.564856
KNN	None	None	yimpuntual	opcion1	0.549602	0.005473	0.539157	0.546387	0.549065	0.553771	0.565186
KNN	None	None	yimpuntual	opcion2	0.547127	0.005310	0.536575	0.543185	0.546576	0.551201	0.559075
KNN	None	None	yimpuntual	opcion3	0.550112	0.005557	0.537234	0.546444	0.549954	0.554038	0.565609
KNN	None	None	yrecuperacionbaja	opcion1	0.539056	0.008385	0.520051	0.533987	0.539741	0.543965	0.560465
KNN	None	None	yrecuperacionbaja	opcion2	0.534513	0.008269	0.513353	0.530449	0.533952	0.539921	0.554494
KNN	None	None	yrecuperacionbaja	opcion3	0.544371	0.008581	0.525019	0.538384	0.544692	0.550272	0.565304
LDA	None	None	ydeudor	opcion1	0.699401	0.009200	0.677946	0.693976	0.699975	0.706080	0.719972
LDA	None	None	ydeudor	opcion2	0.700068	0.009080	0.680810	0.694095	0.700714	0.706344	0.723394
LDA	None	None	ydeudor	opcion3	0.697889	0.009660	0.674286	0.691409	0.698926	0.703988	0.720356
LDA	None	None	yimpuntual	opcion1	0.764693	0.004428	0.753736	0.762850	0.764665	0.767100	0.775512
LDA	None	None	yimpuntual	opcion2	0.764638	0.004364	0.753371	0.762858	0.764566	0.766787	0.776012
LDA	None	None	yimpuntual	opcion3	0.763135	0.004347	0.750834	0.761315	0.763376	0.765103	0.773187
LDA	None	None	yrecuperacionbaja	opcion1	0.700562	0.009386	0.678159	0.695114	0.701600	0.707188	0.720453
LDA	None	None	yrecuperacionbaja	opcion2	0.701196	0.009269	0.677256	0.695165	0.702073	0.707688	0.724156
LDA	None	None	yrecuperacionbaja	opcion3	0.698605	0.009863	0.673832	0.692023	0.700637	0.704508	0.719709
LOGIT	None	None	ydeudor	opcion1	0.700640	0.009437	0.679112	0.694271	0.701266	0.706975	0.722815
LOGIT	None	None	ydeudor	opcion2	0.701519	0.009347	0.683158	0.694642	0.702042	0.707883	0.726194
LOGIT	None	None	ydeudor	opcion3	0.698338	0.009776	0.674541	0.691224	0.699864	0.705063	0.722267
LOGIT	None	None	yimpuntual	opcion1	0.764442	0.004420	0.753293	0.762538	0.764310	0.766856	0.775416

LOGIT	None	None	yimpuntual	opcion2	0.764375	0.004349	0.753069	0.762555	0.764211	0.766577	0.775877
LOGIT	None	None	yimpuntual	opcion3	0.762692	0.004360	0.750134	0.760777	0.762891	0.764640	0.772817
LOGIT	None	None	yrecuperacionbaja	opcion1	0.702134	0.009693	0.678381	0.696430	0.703153	0.708307	0.723802
LOGIT	None	None	yrecuperacionbaja	opcion2	0.703067	0.009607	0.679770	0.697347	0.704600	0.709114	0.728147
LOGIT	None	None	yrecuperacionbaja	opcion3	0.698705	0.009952	0.674258	0.691743	0.700441	0.705217	0.720949
LOGIT	None	l2	ydeudor	opcion1	0.700663	0.009440	0.679085	0.694289	0.701325	0.706985	0.722888
LOGIT	None	l2	ydeudor	opcion2	0.701538	0.009347	0.683149	0.694709	0.702083	0.707898	0.726209
LOGIT	None	l2	ydeudor	opcion3	0.698338	0.009776	0.674542	0.691211	0.699869	0.705095	0.722250
LOGIT	None	l2	yimpuntual	opcion1	0.764444	0.004420	0.753292	0.762541	0.764319	0.766859	0.775413
LOGIT	None	l2	yimpuntual	opcion2	0.764378	0.004348	0.753067	0.762553	0.764208	0.766574	0.775874
LOGIT	None	l2	yimpuntual	opcion3	0.762693	0.004360	0.750151	0.760778	0.762889	0.764634	0.772856
LOGIT	None	l2	yrecuperacionbaja	opcion1	0.702152	0.009689	0.678389	0.696432	0.703157	0.708313	0.723813
LOGIT	None	l2	yrecuperacionbaja	opcion2	0.703090	0.009615	0.679775	0.697402	0.704652	0.709128	0.728225
LOGIT	None	l2	yrecuperacionbaja	opcion3	0.698703	0.009954	0.674256	0.691756	0.700439	0.705226	0.720942
LOGIT	balanced	None	ydeudor	opcion1	0.700488	0.009320	0.678661	0.693427	0.700835	0.706322	0.721128
LOGIT	balanced	None	ydeudor	opcion2	0.701128	0.009189	0.682938	0.694405	0.701768	0.706856	0.724823
LOGIT	balanced	None	ydeudor	opcion3	0.698303	0.009683	0.674096	0.690754	0.699310	0.704493	0.721441
LOGIT	balanced	None	yimpuntual	opcion1	0.764441	0.004421	0.753297	0.762537	0.764315	0.766852	0.775425
LOGIT	balanced	None	yimpuntual	opcion2	0.764375	0.004349	0.753047	0.762544	0.764205	0.766575	0.775871
LOGIT	balanced	None	yimpuntual	opcion3	0.762695	0.004357	0.750212	0.760817	0.762894	0.764629	0.772828
LOGIT	balanced	None	yrecuperacionbaja	opcion1	0.702384	0.009596	0.677653	0.696966	0.703422	0.708322	0.723447
LOGIT	balanced	None	yrecuperacionbaja	opcion2	0.703081	0.009476	0.679185	0.697723	0.704087	0.708998	0.726728
LOGIT	balanced	None	yrecuperacionbaja	opcion3	0.698949	0.009829	0.675343	0.692265	0.700511	0.705272	0.720686
LOGIT	balanced	l2	ydeudor	opcion1	0.700487	0.009321	0.678654	0.693460	0.700847	0.706296	0.721173
LOGIT	balanced	l2	ydeudor	opcion2	0.701133	0.009191	0.682926	0.694472	0.701778	0.706842	0.724834

LOGIT	balanced	l2	ydeudor	opcion3	0.698301	0.009686	0.674093	0.690757	0.699315	0.704518	0.721439
LOGIT	balanced	l2	yimpuntual	opcion1	0.764444	0.004420	0.753299	0.762543	0.764320	0.766855	0.775422
LOGIT	balanced	l2	yimpuntual	opcion2	0.764378	0.004348	0.753044	0.762549	0.764214	0.766586	0.775873
LOGIT	balanced	l2	yimpuntual	opcion3	0.762693	0.004361	0.750167	0.760773	0.762890	0.764626	0.772879
LOGIT	balanced	l2	yrecuperacionbaja	opcion1	0.702397	0.009596	0.677671	0.696972	0.703436	0.708387	0.723459
LOGIT	balanced	l2	yrecuperacionbaja	opcion2	0.703086	0.009478	0.679176	0.697718	0.704102	0.709000	0.726810
LOGIT	balanced	l2	yrecuperacionbaja	opcion3	0.698954	0.009830	0.675346	0.692248	0.700507	0.705257	0.720702
NB	None	None	ydeudor	opcion1	0.601097	0.012056	0.571413	0.593666	0.602463	0.608916	0.628542
NB	None	None	ydeudor	opcion2	0.629016	0.012072	0.597824	0.622893	0.629541	0.637004	0.658343
NB	None	None	ydeudor	opcion3	0.602116	0.012196	0.572156	0.594483	0.603504	0.608787	0.630741
NB	None	None	yimpuntual	opcion1	0.602446	0.006748	0.585032	0.598432	0.602588	0.606332	0.618722
NB	None	None	yimpuntual	opcion2	0.656453	0.008390	0.633194	0.651407	0.656533	0.661804	0.675659
NB	None	None	yimpuntual	opcion3	0.606740	0.006935	0.590291	0.601812	0.607082	0.611174	0.622219
NB	None	None	yrecuperacionbaja	opcion1	0.604046	0.012569	0.572715	0.595620	0.604547	0.611815	0.636934
NB	None	None	yrecuperacionbaja	opcion2	0.629655	0.012736	0.591655	0.622763	0.630467	0.638874	0.657197
NB	None	None	yrecuperacionbaja	opcion3	0.602084	0.012439	0.568113	0.594464	0.602780	0.609410	0.635893
RF	None	None	ydeudor	opcion1	0.706296	0.009142	0.685435	0.699588	0.706363	0.711929	0.730401
RF	None	None	ydeudor	opcion2	0.707012	0.009090	0.685203	0.700790	0.707308	0.712224	0.732275
RF	None	None	ydeudor	opcion3	0.706425	0.009250	0.683010	0.699415	0.706075	0.712232	0.732221
RF	None	None	yimpuntual	opcion1	0.777444	0.004588	0.765302	0.774848	0.777549	0.780375	0.790234
RF	None	None	yimpuntual	opcion2	0.777332	0.004436	0.765928	0.775115	0.777640	0.780002	0.788962
RF	None	None	yimpuntual	opcion3	0.777674	0.004595	0.764956	0.774754	0.777753	0.780346	0.789676
RF	None	None	yrecuperacionbaja	opcion1	0.707835	0.008712	0.684368	0.701496	0.708458	0.712928	0.733394
RF	None	None	yrecuperacionbaja	opcion2	0.708252	0.008816	0.682888	0.702361	0.708571	0.713488	0.735036
RF	None	None	yrecuperacionbaja	opcion3	0.707619	0.008783	0.681379	0.701629	0.708540	0.712377	0.731475

RF	balanced	None	ydeudor	opcion1	0.704825	0.008722	0.684796	0.699183	0.704294	0.710630	0.727081
RF	balanced	None	ydeudor	opcion2	0.705733	0.008475	0.683031	0.700325	0.705601	0.710974	0.728726
RF	balanced	None	ydeudor	opcion3	0.704966	0.008622	0.685724	0.699790	0.704786	0.710196	0.730206
RF	balanced	None	yimpuntual	opcion1	0.777417	0.004582	0.765142	0.774695	0.777613	0.780150	0.790307
RF	balanced	None	yimpuntual	opcion2	0.777317	0.004433	0.765756	0.775113	0.777622	0.779972	0.788987
RF	balanced	None	yimpuntual	opcion3	0.777651	0.004581	0.765238	0.774718	0.777849	0.780179	0.789907
RF	balanced	None	yrecuperacionbaja	opcion1	0.706365	0.008463	0.682988	0.701258	0.706059	0.711619	0.730131
RF	balanced	None	yrecuperacionbaja	opcion2	0.706898	0.008459	0.681628	0.701975	0.706524	0.712688	0.732420
RF	balanced	None	yrecuperacionbaja	opcion3	0.706366	0.008394	0.683240	0.701337	0.706222	0.711407	0.728931
XGB	None	None	ydeudor	opcion1	0.703907	0.008983	0.681785	0.697730	0.703765	0.710680	0.727020
XGB	None	None	ydeudor	opcion2	0.704246	0.008705	0.678342	0.697888	0.705448	0.709459	0.728010
XGB	None	None	ydeudor	opcion3	0.702819	0.009213	0.678459	0.696226	0.703319	0.708325	0.728635
XGB	None	None	yimpuntual	opcion1	0.779096	0.004489	0.766371	0.776771	0.779342	0.781691	0.790801
XGB	None	None	yimpuntual	opcion2	0.778785	0.004353	0.766786	0.776483	0.779280	0.781230	0.790150
XGB	None	None	yimpuntual	opcion3	0.778552	0.004438	0.766750	0.776109	0.778591	0.780965	0.789513
XGB	None	None	yrecuperacionbaja	opcion1	0.705544	0.008674	0.680936	0.698887	0.705649	0.712208	0.729656
XGB	None	None	yrecuperacionbaja	opcion2	0.705918	0.008216	0.675976	0.702063	0.706416	0.710232	0.730498
XGB	None	None	yrecuperacionbaja	opcion3	0.704258	0.009171	0.677287	0.697407	0.704960	0.711373	0.727688

Cuadro A2: Tabla Métrica AUC Train.

Modelo	balanceo	penalidad	y	x	mean	std	min	25 %	50 %	75 %	max
KNN	None	None	ydeudor	opcion1	0.920318	0.001657	0.915236	0.919478	0.920350	0.921308	0.924659
KNN	None	None	ydeudor	opcion2	0.919729	0.001583	0.915274	0.918459	0.919655	0.920774	0.924994
KNN	None	None	ydeudor	opcion3	0.920161	0.001665	0.916061	0.919027	0.920086	0.921420	0.923837
KNN	None	None	yimpuntual	opcion1	0.797600	0.002409	0.791866	0.796349	0.797599	0.799322	0.803310
KNN	None	None	yimpuntual	opcion2	0.796011	0.002210	0.789196	0.794755	0.796167	0.797793	0.800687
KNN	None	None	yimpuntual	opcion3	0.798332	0.002620	0.791791	0.796603	0.798836	0.800088	0.803439
KNN	None	None	yrecuperacionbaja	opcion1	0.922893	0.001658	0.918664	0.921987	0.922871	0.923972	0.927437
KNN	None	None	yrecuperacionbaja	opcion2	0.922795	0.001573	0.918775	0.921835	0.922697	0.923650	0.927732
KNN	None	None	yrecuperacionbaja	opcion3	0.923342	0.001767	0.919138	0.921990	0.923519	0.924584	0.926849
LDA	None	None	ydeudor	opcion1	0.701565	0.003943	0.691724	0.698914	0.701499	0.704407	0.709974
LDA	None	None	ydeudor	opcion2	0.700517	0.003991	0.689723	0.697906	0.700699	0.703111	0.709030
LDA	None	None	ydeudor	opcion3	0.698173	0.004062	0.688083	0.695530	0.698180	0.701227	0.707867
LDA	None	None	yimpuntual	opcion1	0.765549	0.001880	0.760791	0.764607	0.765526	0.766397	0.770739
LDA	None	None	yimpuntual	opcion2	0.765231	0.001860	0.760442	0.764386	0.765164	0.766017	0.770547
LDA	None	None	yimpuntual	opcion3	0.763627	0.001872	0.758899	0.762761	0.763565	0.764451	0.769016
LDA	None	None	yrecuperacionbaja	opcion1	0.702472	0.004027	0.692229	0.700054	0.702024	0.705013	0.712701
LDA	None	None	yrecuperacionbaja	opcion2	0.701256	0.004085	0.689821	0.698585	0.700963	0.703846	0.711324
LDA	None	None	yrecuperacionbaja	opcion3	0.698738	0.004149	0.688328	0.696073	0.698298	0.701546	0.709723
LOGIT	None	None	ydeudor	opcion1	0.702894	0.004036	0.692654	0.700288	0.702837	0.705453	0.711931
LOGIT	None	None	ydeudor	opcion2	0.701992	0.004083	0.690864	0.699361	0.702107	0.704571	0.711060
LOGIT	None	None	ydeudor	opcion3	0.698863	0.004153	0.688404	0.696158	0.698722	0.701794	0.708997
LOGIT	None	None	yimpuntual	opcion1	0.765388	0.001873	0.760566	0.764559	0.765418	0.766268	0.770635

LOGIT	None	None	yimpuntual	opcion2	0.765050	0.001848	0.760213	0.764180	0.764967	0.765853	0.770380
LOGIT	None	None	yimpuntual	opcion3	0.763238	0.001867	0.758495	0.762377	0.763148	0.764062	0.768665
LOGIT	None	None	yrecuperacionbaja	opcion1	0.704216	0.004160	0.693408	0.701689	0.703804	0.706689	0.715502
LOGIT	None	None	yrecuperacionbaja	opcion2	0.703184	0.004206	0.691278	0.700538	0.702876	0.705613	0.714172
LOGIT	None	None	yrecuperacionbaja	opcion3	0.699107	0.004231	0.688454	0.696273	0.698715	0.702234	0.710503
LOGIT	None	l2	ydeudor	opcion1	0.702894	0.004036	0.692661	0.700287	0.702835	0.705447	0.711918
LOGIT	None	l2	ydeudor	opcion2	0.701998	0.004084	0.690874	0.699363	0.702109	0.704571	0.711083
LOGIT	None	l2	ydeudor	opcion3	0.698863	0.004154	0.688389	0.696154	0.698714	0.701781	0.708993
LOGIT	None	l2	yimpuntual	opcion1	0.765390	0.001873	0.760563	0.764560	0.765418	0.766270	0.770635
LOGIT	None	l2	yimpuntual	opcion2	0.765053	0.001848	0.760220	0.764183	0.764970	0.765856	0.770379
LOGIT	None	l2	yimpuntual	opcion3	0.763236	0.001868	0.758488	0.762368	0.763142	0.764050	0.768680
LOGIT	None	l2	yrecuperacionbaja	opcion1	0.704219	0.004160	0.693416	0.701691	0.703805	0.706690	0.715510
LOGIT	None	l2	yrecuperacionbaja	opcion2	0.703191	0.004205	0.691305	0.700544	0.702882	0.705609	0.714175
LOGIT	None	l2	yrecuperacionbaja	opcion3	0.699107	0.004231	0.688458	0.696269	0.698714	0.702232	0.710502
LOGIT	balanced	None	ydeudor	opcion1	0.702750	0.004005	0.692661	0.700205	0.702543	0.705515	0.711498
LOGIT	balanced	None	ydeudor	opcion2	0.701593	0.004050	0.690554	0.699109	0.701694	0.704277	0.710278
LOGIT	balanced	None	ydeudor	opcion3	0.698754	0.004125	0.688806	0.695910	0.698594	0.701738	0.709227
LOGIT	balanced	None	yimpuntual	opcion1	0.765388	0.001873	0.760567	0.764559	0.765417	0.766269	0.770646
LOGIT	balanced	None	yimpuntual	opcion2	0.765050	0.001848	0.760216	0.764179	0.764968	0.765856	0.770384
LOGIT	balanced	None	yimpuntual	opcion3	0.763240	0.001870	0.758496	0.762379	0.763135	0.764040	0.768720
LOGIT	balanced	None	yrecuperacionbaja	opcion1	0.704557	0.004136	0.693858	0.702098	0.704024	0.707068	0.715838
LOGIT	balanced	None	yrecuperacionbaja	opcion2	0.703199	0.004182	0.691366	0.700529	0.702941	0.705533	0.714075
LOGIT	balanced	None	yrecuperacionbaja	opcion3	0.699284	0.004232	0.688675	0.696519	0.698849	0.702135	0.711210
LOGIT	balanced	l2	ydeudor	opcion1	0.702754	0.004005	0.692692	0.700191	0.702534	0.705491	0.711500
LOGIT	balanced	l2	ydeudor	opcion2	0.701595	0.004049	0.690557	0.699120	0.701693	0.704277	0.710272

LOGIT	balanced	l2	ydeudor	opcion3	0.698755	0.004125	0.688806	0.695922	0.698599	0.701727	0.709242
LOGIT	balanced	l2	yimpuntual	opcion1	0.765390	0.001873	0.760566	0.764562	0.765419	0.766269	0.770647
LOGIT	balanced	l2	yimpuntual	opcion2	0.765053	0.001848	0.760220	0.764182	0.764972	0.765857	0.770388
LOGIT	balanced	l2	yimpuntual	opcion3	0.763238	0.001868	0.758495	0.762372	0.763148	0.764077	0.768682
LOGIT	balanced	l2	yrecuperacionbaja	opcion1	0.704560	0.004135	0.693859	0.702107	0.704036	0.707067	0.715837
LOGIT	balanced	l2	yrecuperacionbaja	opcion2	0.703201	0.004180	0.691377	0.700532	0.702939	0.705536	0.714073
LOGIT	balanced	l2	yrecuperacionbaja	opcion3	0.699283	0.004231	0.688701	0.696519	0.698873	0.702163	0.711187
NB	None	None	ydeudor	opcion1	0.607906	0.004397	0.597793	0.604841	0.608110	0.610388	0.620180
NB	None	None	ydeudor	opcion2	0.632444	0.004667	0.623511	0.629343	0.631940	0.635684	0.643789
NB	None	None	ydeudor	opcion3	0.608481	0.004209	0.598190	0.605472	0.608720	0.611152	0.619778
NB	None	None	yimpuntual	opcion1	0.604896	0.001935	0.600416	0.603435	0.604776	0.606133	0.609037
NB	None	None	yimpuntual	opcion2	0.658856	0.003147	0.651777	0.656851	0.658719	0.660633	0.666875
NB	None	None	yimpuntual	opcion3	0.608870	0.002315	0.602057	0.607516	0.609040	0.610245	0.614081
NB	None	None	yrecuperacionbaja	opcion1	0.610070	0.004430	0.602112	0.606725	0.609976	0.613154	0.620831
NB	None	None	yrecuperacionbaja	opcion2	0.632042	0.004138	0.622026	0.629388	0.631996	0.634699	0.643849
NB	None	None	yrecuperacionbaja	opcion3	0.608579	0.004476	0.599871	0.605427	0.608451	0.611660	0.620470
RF	None	None	ydeudor	opcion1	0.802830	0.003213	0.794805	0.800432	0.802564	0.804637	0.812298
RF	None	None	ydeudor	opcion2	0.790578	0.003302	0.783012	0.788305	0.790724	0.792582	0.800057
RF	None	None	ydeudor	opcion3	0.798824	0.003317	0.791615	0.796890	0.798501	0.800957	0.807672
RF	None	None	yimpuntual	opcion1	0.798616	0.001835	0.794194	0.797382	0.798644	0.799876	0.802859
RF	None	None	yimpuntual	opcion2	0.795025	0.001759	0.790761	0.793970	0.794901	0.796079	0.799513
RF	None	None	yimpuntual	opcion3	0.798165	0.001807	0.793729	0.796967	0.798080	0.799244	0.802362
RF	None	None	yrecuperacionbaja	opcion1	0.809376	0.003471	0.801367	0.807144	0.809265	0.811279	0.821564
RF	None	None	yrecuperacionbaja	opcion2	0.795478	0.003575	0.788190	0.792877	0.795762	0.797708	0.806169
RF	None	None	yrecuperacionbaja	opcion3	0.804276	0.003790	0.795008	0.801942	0.803891	0.806657	0.816503

RF	balanced	None	ydeudor	opcion1	0.789317	0.002680	0.783584	0.787289	0.789372	0.791175	0.795697
RF	balanced	None	ydeudor	opcion2	0.777184	0.002885	0.770917	0.775102	0.776568	0.779297	0.785864
RF	balanced	None	ydeudor	opcion3	0.784812	0.002928	0.778006	0.782860	0.784425	0.787201	0.791895
RF	balanced	None	yimpuntual	opcion1	0.798597	0.001837	0.794025	0.797351	0.798632	0.799859	0.802856
RF	balanced	None	yimpuntual	opcion2	0.795046	0.001773	0.790714	0.793850	0.794985	0.796120	0.799529
RF	balanced	None	yimpuntual	opcion3	0.798162	0.001800	0.793898	0.796967	0.798004	0.799295	0.802379
RF	balanced	None	yrecuperacionbaja	opcion1	0.792554	0.002601	0.786759	0.790797	0.792301	0.794169	0.801043
RF	balanced	None	yrecuperacionbaja	opcion2	0.779508	0.002866	0.772697	0.777719	0.779089	0.781518	0.788808
RF	balanced	None	yrecuperacionbaja	opcion3	0.788015	0.002921	0.781560	0.785772	0.787755	0.790177	0.797004
XGB	None	None	ydeudor	opcion1	0.837790	0.003901	0.828926	0.835374	0.837657	0.840556	0.847745
XGB	None	None	ydeudor	opcion2	0.817691	0.003588	0.809247	0.814817	0.817886	0.820069	0.826248
XGB	None	None	ydeudor	opcion3	0.828335	0.004181	0.817101	0.825573	0.827790	0.831154	0.838768
XGB	None	None	yimpuntual	opcion1	0.818343	0.001888	0.813772	0.817187	0.818478	0.819500	0.823397
XGB	None	None	yimpuntual	opcion2	0.812369	0.001843	0.807346	0.811079	0.812585	0.813559	0.816851
XGB	None	None	yimpuntual	opcion3	0.816092	0.001852	0.810950	0.814930	0.816218	0.817498	0.820646
XGB	None	None	yrecuperacionbaja	opcion1	0.839615	0.004255	0.830339	0.836195	0.839295	0.842651	0.849436
XGB	None	None	yrecuperacionbaja	opcion2	0.820194	0.003734	0.809133	0.817553	0.820151	0.822576	0.828705
XGB	None	None	yrecuperacionbaja	opcion3	0.829917	0.004245	0.820405	0.826965	0.829158	0.832840	0.841223

Cuadro A3: Tabla Componente uno de la diagonal, Matriz de Confusión.

Modelo	balanceo	penalidad	y	x	mean	std	min	25 %	50 %	75 %	max
KNN	None	None	ydeudor	opcion1	0.997270	0.000693	0.995417	0.996851	0.997252	0.997772	0.998819
KNN	None	None	ydeudor	opcion2	0.997441	0.000702	0.995938	0.996989	0.997506	0.997903	0.998952
KNN	None	None	ydeudor	opcion3	0.997293	0.000814	0.995023	0.996725	0.997384	0.997904	0.998688
KNN	None	None	yimpuntual	opcion1	0.690000	0.007750	0.670855	0.685994	0.690132	0.695322	0.706845
KNN	None	None	yimpuntual	opcion2	0.688057	0.007966	0.672103	0.682896	0.688801	0.693902	0.705058
KNN	None	None	yimpuntual	opcion3	0.689231	0.008397	0.666585	0.684048	0.689059	0.694634	0.712587
KNN	None	None	yrecuperacionbaja	opcion1	0.997546	0.000630	0.996073	0.997129	0.997515	0.997915	0.998826
KNN	None	None	yrecuperacionbaja	opcion2	0.997796	0.000653	0.995956	0.997391	0.997777	0.998204	0.999214
KNN	None	None	yrecuperacionbaja	opcion3	0.997603	0.000739	0.995564	0.997125	0.997647	0.998168	0.998954
LDA	None	None	ydeudor	opcion1	0.999988	0.000038	0.999868	1.000000	1.000000	1.000000	1.000000
LDA	None	None	ydeudor	opcion2	0.999990	0.000036	0.999868	1.000000	1.000000	1.000000	1.000000
LDA	None	None	ydeudor	opcion3	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LDA	None	None	yimpuntual	opcion1	0.623638	0.006356	0.607527	0.619432	0.623984	0.627325	0.646192
LDA	None	None	yimpuntual	opcion2	0.622554	0.006281	0.605816	0.618167	0.623083	0.626548	0.644944
LDA	None	None	yimpuntual	opcion3	0.618866	0.006366	0.603372	0.615068	0.619253	0.622342	0.641448
LDA	None	None	yrecuperacionbaja	opcion1	0.999979	0.000048	0.999869	1.000000	1.000000	1.000000	1.000000
LDA	None	None	yrecuperacionbaja	opcion2	0.999983	0.000044	0.999869	1.000000	1.000000	1.000000	1.000000
LDA	None	None	yrecuperacionbaja	opcion3	0.999993	0.000029	0.999869	1.000000	1.000000	1.000000	1.000000
LOGIT	None	None	ydeudor	opcion1	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	None	ydeudor	opcion2	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	None	ydeudor	opcion3	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	None	yimpuntual	opcion1	0.626274	0.006702	0.608993	0.621803	0.626505	0.630125	0.652185

LOGIT	None	None	yimpuntual	opcion2	0.625294	0.006516	0.607771	0.621062	0.625585	0.628556	0.649938
LOGIT	None	None	yimpuntual	opcion3	0.617232	0.006393	0.602639	0.613678	0.617449	0.621122	0.639451
LOGIT	None	None	yrecuperacionbaja	opcion1	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	None	yrecuperacionbaja	opcion2	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	None	yrecuperacionbaja	opcion3	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	l2	ydeudor	opcion1	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	l2	ydeudor	opcion2	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	l2	ydeudor	opcion3	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	l2	yimpuntual	opcion1	0.626269	0.006693	0.608993	0.621638	0.626505	0.630125	0.651935
LOGIT	None	l2	yimpuntual	opcion2	0.625296	0.006528	0.607771	0.621062	0.625585	0.628556	0.650187
LOGIT	None	l2	yimpuntual	opcion3	0.617219	0.006397	0.602395	0.613714	0.617338	0.621122	0.639451
LOGIT	None	l2	yrecuperacionbaja	opcion1	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	l2	yrecuperacionbaja	opcion2	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	None	l2	yrecuperacionbaja	opcion3	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
LOGIT	balanced	None	ydeudor	opcion1	0.608393	0.007842	0.589488	0.603331	0.607911	0.613282	0.628508
LOGIT	balanced	None	ydeudor	opcion2	0.599768	0.007879	0.579106	0.594272	0.599095	0.604990	0.622243
LOGIT	balanced	None	ydeudor	opcion3	0.601832	0.007912	0.583706	0.596555	0.600973	0.606671	0.621590
LOGIT	balanced	None	yimpuntual	opcion1	0.627980	0.006706	0.610215	0.623848	0.627998	0.632026	0.652934
LOGIT	balanced	None	yimpuntual	opcion2	0.626805	0.006682	0.609238	0.622561	0.627458	0.630175	0.651935
LOGIT	balanced	None	yimpuntual	opcion3	0.618178	0.006345	0.604350	0.614627	0.618332	0.621563	0.640200
LOGIT	balanced	None	yrecuperacionbaja	opcion1	0.612085	0.008084	0.593448	0.606643	0.612410	0.616830	0.631930
LOGIT	balanced	None	yrecuperacionbaja	opcion2	0.603739	0.008039	0.583628	0.599101	0.602672	0.608009	0.621904
LOGIT	balanced	None	yrecuperacionbaja	opcion3	0.603842	0.008017	0.585069	0.598180	0.604045	0.608613	0.622335
LOGIT	balanced	l2	ydeudor	opcion1	0.608362	0.007900	0.589225	0.602915	0.608041	0.613270	0.628639
LOGIT	balanced	l2	ydeudor	opcion2	0.599770	0.007879	0.579106	0.594211	0.598979	0.604828	0.622112

LOGIT	balanced	l2	ydeudor	opcion3	0.601792	0.007977	0.583574	0.596328	0.600866	0.606346	0.621981
LOGIT	balanced	l2	yimpuntual	opcion1	0.627955	0.006701	0.610215	0.623848	0.627998	0.632026	0.652934
LOGIT	balanced	l2	yimpuntual	opcion2	0.626808	0.006671	0.609238	0.622561	0.627456	0.630113	0.651935
LOGIT	balanced	l2	yimpuntual	opcion3	0.618166	0.006350	0.604350	0.614627	0.618209	0.621563	0.640200
LOGIT	balanced	l2	yrecuperacionbaja	opcion1	0.612024	0.008094	0.593318	0.606635	0.612299	0.616958	0.631930
LOGIT	balanced	l2	yrecuperacionbaja	opcion2	0.603727	0.008064	0.583890	0.599187	0.602889	0.608376	0.621945
LOGIT	balanced	l2	yrecuperacionbaja	opcion3	0.603838	0.008050	0.585069	0.597989	0.604046	0.609177	0.622335
NB	None	None	ydeudor	opcion1	0.995038	0.004072	0.983322	0.992572	0.996456	0.997709	1.000000
NB	None	None	ydeudor	opcion2	0.995711	0.003669	0.986797	0.992619	0.996401	0.998982	1.000000
NB	None	None	ydeudor	opcion3	0.995109	0.004138	0.983892	0.992076	0.996596	0.998196	1.000000
NB	None	None	yimpuntual	opcion1	0.198051	0.080012	0.122721	0.142242	0.165355	0.222540	0.481303
NB	None	None	yimpuntual	opcion2	0.121297	0.010113	0.101079	0.113944	0.120737	0.127010	0.163329
NB	None	None	yimpuntual	opcion3	0.735157	0.100652	0.355727	0.687086	0.765721	0.811647	0.879056
NB	None	None	yrecuperacionbaja	opcion1	0.993488	0.004844	0.979316	0.989893	0.994898	0.996727	1.000000
NB	None	None	yrecuperacionbaja	opcion2	0.994841	0.003707	0.986562	0.991471	0.995881	0.997771	1.000000
NB	None	None	yrecuperacionbaja	opcion3	0.994530	0.004261	0.982909	0.991782	0.995753	0.997159	1.000000
RF	None	None	ydeudor	opcion1	0.999986	0.000041	0.999868	1.000000	1.000000	1.000000	1.000000
RF	None	None	ydeudor	opcion2	0.999999	0.000013	0.999869	1.000000	1.000000	1.000000	1.000000
RF	None	None	ydeudor	opcion3	0.999974	0.000056	0.999738	1.000000	1.000000	1.000000	1.000000
RF	None	None	yimpuntual	opcion1	0.668518	0.012036	0.630083	0.663614	0.669355	0.675928	0.697714
RF	None	None	yimpuntual	opcion2	0.666041	0.012233	0.631553	0.659264	0.667041	0.673812	0.692385
RF	None	None	yimpuntual	opcion3	0.658207	0.012270	0.632288	0.650003	0.658752	0.666106	0.688140
RF	None	None	yrecuperacionbaja	opcion1	0.999991	0.000034	0.999869	1.000000	1.000000	1.000000	1.000000
RF	None	None	yrecuperacionbaja	opcion2	0.999999	0.000013	0.999869	1.000000	1.000000	1.000000	1.000000
RF	None	None	yrecuperacionbaja	opcion3	0.999975	0.000052	0.999869	1.000000	1.000000	1.000000	1.000000

RF	balanced	None	ydeudor	opcion1	0.681335	0.010823	0.650111	0.675193	0.680483	0.688833	0.704716
RF	balanced	None	ydeudor	opcion2	0.669272	0.011679	0.636090	0.662455	0.669645	0.677541	0.696906
RF	balanced	None	ydeudor	opcion3	0.671567	0.010473	0.648267	0.665232	0.672563	0.677288	0.704190
RF	balanced	None	yimpuntual	opcion1	0.673188	0.011338	0.634248	0.668077	0.674405	0.679040	0.700632
RF	balanced	None	yimpuntual	opcion2	0.670543	0.011676	0.633758	0.663368	0.672678	0.676868	0.699660
RF	balanced	None	yimpuntual	opcion3	0.663978	0.011810	0.636943	0.656143	0.664029	0.670636	0.697878
RF	balanced	None	yrecuperacionbaja	opcion1	0.680759	0.010600	0.652923	0.672985	0.681006	0.687456	0.704439
RF	balanced	None	yrecuperacionbaja	opcion2	0.667170	0.011355	0.641160	0.658828	0.666514	0.674739	0.691628
RF	balanced	None	yrecuperacionbaja	opcion3	0.674115	0.010640	0.647443	0.667483	0.674226	0.680554	0.703704
XGB	None	None	ydeudor	opcion1	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
XGB	None	None	ydeudor	opcion2	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
XGB	None	None	ydeudor	opcion3	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
XGB	None	None	yimpuntual	opcion1	0.662763	0.008999	0.640762	0.657597	0.663679	0.666872	0.694382
XGB	None	None	yimpuntual	opcion2	0.662680	0.008791	0.642229	0.657755	0.662308	0.666831	0.695131
XGB	None	None	yimpuntual	opcion3	0.660981	0.009043	0.638988	0.656210	0.660819	0.665348	0.691635
XGB	None	None	yrecuperacionbaja	opcion1	0.999999	0.000013	0.999869	1.000000	1.000000	1.000000	1.000000
XGB	None	None	yrecuperacionbaja	opcion2	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000
XGB	None	None	yrecuperacionbaja	opcion3	0.999999	0.000013	0.999870	1.000000	1.000000	1.000000	1.000000

Cuadro A4: Tabla Componente dos de la diagonal, Matriz de Confusión.

Modelo	balanceo	penalidad	y	x	mean	std	min	25 %	50 %	75 %	max
KNN	None	None	ydeudor	opcion1	0.007506	0.003297	0.000000	0.005275	0.007086	0.010118	0.015762
KNN	None	None	ydeudor	opcion2	0.004456	0.002455	0.000000	0.003398	0.003676	0.005698	0.010753
KNN	None	None	ydeudor	opcion3	0.005386	0.002549	0.000000	0.003513	0.005137	0.006885	0.012153
KNN	None	None	yimpuntual	opcion1	0.379307	0.007439	0.362424	0.374737	0.379479	0.384415	0.395142
KNN	None	None	yimpuntual	opcion2	0.374425	0.009114	0.352528	0.368036	0.373640	0.380113	0.397713
KNN	None	None	yimpuntual	opcion3	0.377169	0.007364	0.362806	0.372292	0.376683	0.382529	0.396991
KNN	None	None	yrecuperacionbaja	opcion1	0.006646	0.003172	0.000000	0.003790	0.007080	0.009042	0.014388
KNN	None	None	yrecuperacionbaja	opcion2	0.004458	0.002460	0.000000	0.003518	0.003728	0.005558	0.011215
KNN	None	None	yrecuperacionbaja	opcion3	0.004991	0.002449	0.000000	0.003541	0.005226	0.007121	0.011152
LDA	None	None	ydeudor	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LDA	None	None	ydeudor	opcion2	0.000155	0.000495	0.000000	0.000000	0.000000	0.000000	0.001802
LDA	None	None	ydeudor	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LDA	None	None	yimpuntual	opcion1	0.794201	0.005330	0.780552	0.790837	0.794370	0.798149	0.805970
LDA	None	None	yimpuntual	opcion2	0.794942	0.005434	0.781513	0.791508	0.795045	0.798533	0.807174
LDA	None	None	yimpuntual	opcion3	0.796173	0.005555	0.781513	0.792211	0.796337	0.800305	0.808618
LDA	None	None	yrecuperacionbaja	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LDA	None	None	yrecuperacionbaja	opcion2	0.000178	0.000538	0.000000	0.000000	0.000000	0.000000	0.001880
LDA	None	None	yrecuperacionbaja	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	None	ydeudor	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	None	ydeudor	opcion2	0.000018	0.000179	0.000000	0.000000	0.000000	0.000000	0.001789
LOGIT	None	None	ydeudor	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	None	yimpuntual	opcion1	0.792140	0.005356	0.777514	0.788487	0.792199	0.795688	0.804285

LOGIT	None	None	yimpuntual	opcion2	0.793118	0.005357	0.779653	0.790365	0.793182	0.796939	0.805007
LOGIT	None	None	yimpuntual	opcion3	0.796833	0.005488	0.781273	0.792523	0.796927	0.801178	0.808377
LOGIT	None	None	yrecuperacionbaja	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	None	yrecuperacionbaja	opcion2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	None	yrecuperacionbaja	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	l2	ydeudor	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	l2	ydeudor	opcion2	0.000018	0.000179	0.000000	0.000000	0.000000	0.000000	0.001789
LOGIT	None	l2	ydeudor	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	l2	yimpuntual	opcion1	0.792152	0.005360	0.777514	0.788627	0.792199	0.795749	0.804285
LOGIT	None	l2	yimpuntual	opcion2	0.793110	0.005353	0.779653	0.790365	0.793182	0.796939	0.805007
LOGIT	None	l2	yimpuntual	opcion3	0.796836	0.005498	0.781273	0.792523	0.796927	0.801178	0.808618
LOGIT	None	l2	yrecuperacionbaja	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	l2	yrecuperacionbaja	opcion2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	None	l2	yrecuperacionbaja	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
LOGIT	balanced	None	ydeudor	opcion1	0.697648	0.018118	0.645796	0.685927	0.698490	0.709471	0.737113
LOGIT	balanced	None	ydeudor	opcion2	0.706706	0.017405	0.669052	0.692254	0.708034	0.719220	0.756803
LOGIT	balanced	None	ydeudor	opcion3	0.706504	0.019065	0.642218	0.694257	0.706038	0.720085	0.753401
LOGIT	balanced	None	yimpuntual	opcion1	0.790523	0.005407	0.775374	0.787047	0.790648	0.794453	0.803322
LOGIT	balanced	None	yimpuntual	opcion2	0.791532	0.005359	0.778464	0.788575	0.791627	0.795589	0.803804
LOGIT	balanced	None	yimpuntual	opcion3	0.796079	0.005468	0.780792	0.791769	0.796231	0.800000	0.807656
LOGIT	balanced	None	yrecuperacionbaja	opcion1	0.693349	0.017706	0.642857	0.681416	0.694781	0.705987	0.725777
LOGIT	balanced	None	yrecuperacionbaja	opcion2	0.703157	0.018113	0.654122	0.690829	0.706364	0.715500	0.738574
LOGIT	balanced	None	yrecuperacionbaja	opcion3	0.703956	0.018922	0.648496	0.690443	0.703541	0.718034	0.751371
LOGIT	balanced	l2	ydeudor	opcion1	0.697770	0.018159	0.645796	0.685927	0.698714	0.709471	0.737113
LOGIT	balanced	l2	ydeudor	opcion2	0.706707	0.017346	0.669052	0.692711	0.707311	0.718921	0.756803

LOGIT	balanced	l2	ydeudor	opcion3	0.706592	0.018916	0.642218	0.694704	0.706280	0.720085	0.753401
LOGIT	balanced	l2	yimpuntual	opcion1	0.790504	0.005406	0.775137	0.787047	0.790648	0.794453	0.803322
LOGIT	balanced	l2	yimpuntual	opcion2	0.791546	0.005377	0.778464	0.788575	0.791627	0.795656	0.803804
LOGIT	balanced	l2	yimpuntual	opcion3	0.796072	0.005472	0.780792	0.791769	0.796231	0.799964	0.807656
LOGIT	balanced	l2	yrecuperacionbaja	opcion1	0.693387	0.017786	0.642857	0.682518	0.695064	0.706043	0.725777
LOGIT	balanced	l2	yrecuperacionbaja	opcion2	0.703195	0.018286	0.654122	0.690829	0.706364	0.715500	0.738574
LOGIT	balanced	l2	yrecuperacionbaja	opcion3	0.704046	0.018863	0.648496	0.690443	0.703796	0.718034	0.751371
NB	None	None	ydeudor	opcion1	0.002641	0.002742	0.000000	0.000000	0.001747	0.003565	0.009509
NB	None	None	ydeudor	opcion2	0.003800	0.003205	0.000000	0.000000	0.003490	0.005759	0.010545
NB	None	None	ydeudor	opcion3	0.002471	0.002753	0.000000	0.000000	0.001738	0.003475	0.012132
NB	None	None	yimpuntual	opcion1	0.875032	0.046072	0.679310	0.868057	0.890378	0.904082	0.928832
NB	None	None	yimpuntual	opcion2	0.923753	0.007213	0.904556	0.918626	0.924322	0.927602	0.941247
NB	None	None	yimpuntual	opcion3	0.385897	0.128061	0.186125	0.290182	0.343162	0.464024	0.791493
NB	None	None	yrecuperacionbaja	opcion1	0.004836	0.004326	0.000000	0.001791	0.003683	0.007181	0.025097
NB	None	None	yrecuperacionbaja	opcion2	0.005543	0.003665	0.000000	0.003077	0.005401	0.008673	0.019305
NB	None	None	yrecuperacionbaja	opcion3	0.003229	0.003088	0.000000	0.000000	0.001927	0.005229	0.013645
RF	None	None	ydeudor	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
RF	None	None	ydeudor	opcion2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
RF	None	None	ydeudor	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
RF	None	None	yimpuntual	opcion1	0.769262	0.010843	0.743104	0.761556	0.768834	0.776258	0.795102
RF	None	None	yimpuntual	opcion2	0.771966	0.011098	0.747421	0.763311	0.771446	0.780143	0.797397
RF	None	None	yimpuntual	opcion3	0.779220	0.011646	0.743584	0.772304	0.779561	0.786709	0.808128
RF	None	None	yrecuperacionbaja	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
RF	None	None	yrecuperacionbaja	opcion2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
RF	None	None	yrecuperacionbaja	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

RF	balanced	None	ydeudor	opcion1	0.616579	0.020876	0.571186	0.603481	0.617970	0.630329	0.668367
RF	balanced	None	ydeudor	opcion2	0.634476	0.021334	0.579737	0.616927	0.632277	0.651714	0.692177
RF	balanced	None	ydeudor	opcion3	0.629389	0.020753	0.581614	0.616705	0.627942	0.645356	0.685374
RF	balanced	None	yimpuntual	opcion1	0.764923	0.010357	0.741425	0.757510	0.764607	0.770507	0.791012
RF	balanced	None	yimpuntual	opcion2	0.767515	0.010594	0.743104	0.759521	0.767281	0.774566	0.793959
RF	balanced	None	yimpuntual	opcion3	0.773832	0.011352	0.739746	0.765903	0.773675	0.781416	0.798216
RF	balanced	None	yrecuperacionbaja	opcion1	0.618713	0.020332	0.573451	0.604410	0.617197	0.632341	0.670232
RF	balanced	None	yrecuperacionbaja	opcion2	0.634985	0.022598	0.578761	0.619163	0.635053	0.650000	0.693405
RF	balanced	None	yrecuperacionbaja	opcion3	0.627781	0.020472	0.571429	0.613084	0.629277	0.641003	0.676525
XGB	None	None	ydeudor	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
XGB	None	None	ydeudor	opcion2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
XGB	None	None	ydeudor	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
XGB	None	None	yimpuntual	opcion1	0.776880	0.007857	0.757496	0.772239	0.777331	0.782593	0.794841
XGB	None	None	yimpuntual	opcion2	0.776804	0.007983	0.758455	0.771856	0.777288	0.781955	0.795074
XGB	None	None	yimpuntual	opcion3	0.778525	0.007891	0.753178	0.773237	0.778513	0.783445	0.798497
XGB	None	None	yrecuperacionbaja	opcion1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
XGB	None	None	yrecuperacionbaja	opcion2	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
XGB	None	None	yrecuperacionbaja	opcion3	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Cuadro A5: Tabla Métrica Kologorov Smirnov.

Modelo	balanceo	penalidad	y	x	mean	std	min	25 %	50 %	75 %	max
KNN	None	None	ydeudor	opcion1	0.084416	0.016493	0.0427	0.074425	0.08520	0.095900	0.1216
KNN	None	None	ydeudor	opcion2	0.076355	0.017420	0.0409	0.064750	0.07630	0.087050	0.1268
KNN	None	None	ydeudor	opcion3	0.092956	0.017670	0.0526	0.079650	0.09330	0.104425	0.1392
KNN	None	None	yimpuntual	opcion1	0.078391	0.009079	0.0596	0.072875	0.07715	0.083875	0.1028
KNN	None	None	yimpuntual	opcion2	0.074889	0.008675	0.0518	0.068750	0.07360	0.081400	0.1003
KNN	None	None	yimpuntual	opcion3	0.079053	0.009446	0.0569	0.073450	0.07970	0.085600	0.1007
KNN	None	None	yrecuperacionbaja	opcion1	0.083340	0.017584	0.0426	0.070925	0.08550	0.093675	0.1283
KNN	None	None	yrecuperacionbaja	opcion2	0.076105	0.017703	0.0398	0.064550	0.07755	0.087225	0.1223
KNN	None	None	yrecuperacionbaja	opcion3	0.093692	0.017291	0.0532	0.083950	0.09205	0.104700	0.1320
LDA	None	None	ydeudor	opcion1	0.313741	0.015966	0.2777	0.302750	0.31455	0.324850	0.3562
LDA	None	None	ydeudor	opcion2	0.315934	0.015556	0.2824	0.305575	0.31675	0.327825	0.3590
LDA	None	None	ydeudor	opcion3	0.317464	0.016426	0.2709	0.306950	0.31610	0.329800	0.3554
LDA	None	None	yimpuntual	opcion1	0.421671	0.007997	0.4003	0.417550	0.42175	0.426625	0.4389
LDA	None	None	yimpuntual	opcion2	0.421140	0.007932	0.4013	0.416100	0.42160	0.426050	0.4384
LDA	None	None	yimpuntual	opcion3	0.417587	0.007921	0.3960	0.412925	0.41810	0.422700	0.4354
LDA	None	None	yrecuperacionbaja	opcion1	0.312717	0.016179	0.2758	0.301175	0.31350	0.325775	0.3447
LDA	None	None	yrecuperacionbaja	opcion2	0.315827	0.015208	0.2792	0.306375	0.31815	0.326200	0.3506
LDA	None	None	yrecuperacionbaja	opcion3	0.317189	0.016781	0.2709	0.304150	0.31990	0.328700	0.3491
LOGIT	None	None	ydeudor	opcion1	0.318150	0.016396	0.2792	0.306650	0.31770	0.329500	0.3540
LOGIT	None	None	ydeudor	opcion2	0.317666	0.015679	0.2831	0.305150	0.31780	0.330050	0.3535
LOGIT	None	None	ydeudor	opcion3	0.318262	0.016575	0.2799	0.306400	0.31820	0.328550	0.3554
LOGIT	None	None	yimpuntual	opcion1	0.421816	0.007878	0.4017	0.417275	0.42215	0.426625	0.4400

LOGIT	None	None	yimpuntual	opcion2	0.421532	0.007945	0.4014	0.416450	0.42195	0.426550	0.4394
LOGIT	None	None	yimpuntual	opcion3	0.416766	0.007944	0.3954	0.411575	0.41660	0.422000	0.4342
LOGIT	None	None	yrecuperacionbaja	opcion1	0.316724	0.016585	0.2751	0.303900	0.31745	0.328425	0.3472
LOGIT	None	None	yrecuperacionbaja	opcion2	0.317153	0.015701	0.2748	0.307225	0.31785	0.328775	0.3518
LOGIT	None	None	yrecuperacionbaja	opcion3	0.317241	0.016861	0.2728	0.304650	0.31815	0.329050	0.3523
LOGIT	None	l2	ydeudor	opcion1	0.318136	0.016379	0.2794	0.306300	0.31780	0.329375	0.3542
LOGIT	None	l2	ydeudor	opcion2	0.317735	0.015694	0.2831	0.304950	0.31775	0.330125	0.3532
LOGIT	None	l2	ydeudor	opcion3	0.318224	0.016537	0.2795	0.306825	0.31810	0.328350	0.3556
LOGIT	None	l2	yimpuntual	opcion1	0.421812	0.007859	0.4019	0.417275	0.42215	0.426625	0.4400
LOGIT	None	l2	yimpuntual	opcion2	0.421530	0.007936	0.4014	0.416450	0.42195	0.426500	0.4394
LOGIT	None	l2	yimpuntual	opcion3	0.416783	0.007943	0.3954	0.411600	0.41665	0.422025	0.4342
LOGIT	None	l2	yrecuperacionbaja	opcion1	0.316763	0.016628	0.2755	0.304025	0.31740	0.328425	0.3471
LOGIT	None	l2	yrecuperacionbaja	opcion2	0.317172	0.015791	0.2749	0.306875	0.31810	0.329225	0.3524
LOGIT	None	l2	yrecuperacionbaja	opcion3	0.317246	0.016829	0.2724	0.304650	0.31800	0.329150	0.3521
LOGIT	balanced	None	ydeudor	opcion1	0.317248	0.015616	0.2715	0.305475	0.31835	0.327300	0.3580
LOGIT	balanced	None	ydeudor	opcion2	0.318649	0.015124	0.2826	0.307050	0.31905	0.328300	0.3569
LOGIT	balanced	None	ydeudor	opcion3	0.318413	0.016189	0.2740	0.307450	0.31870	0.328150	0.3566
LOGIT	balanced	None	yimpuntual	opcion1	0.421819	0.007853	0.4017	0.417275	0.42210	0.426625	0.4397
LOGIT	balanced	None	yimpuntual	opcion2	0.421528	0.007947	0.4016	0.416300	0.42195	0.426500	0.4394
LOGIT	balanced	None	yimpuntual	opcion3	0.416780	0.007936	0.3952	0.411775	0.41675	0.422125	0.4342
LOGIT	balanced	None	yrecuperacionbaja	opcion1	0.317659	0.016690	0.2715	0.306400	0.31840	0.330150	0.3553
LOGIT	balanced	None	yrecuperacionbaja	opcion2	0.319576	0.015843	0.2822	0.310000	0.32190	0.330200	0.3551
LOGIT	balanced	None	yrecuperacionbaja	opcion3	0.317769	0.016380	0.2721	0.306300	0.31930	0.329650	0.3513
LOGIT	balanced	l2	ydeudor	opcion1	0.317282	0.015639	0.2716	0.305300	0.31850	0.327300	0.3566
LOGIT	balanced	l2	ydeudor	opcion2	0.318627	0.015104	0.2825	0.306950	0.31915	0.328400	0.3571

LOGIT	balanced	l2	ydeudor	opcion3	0.318406	0.016139	0.2740	0.307775	0.31840	0.328475	0.3575
LOGIT	balanced	l2	yimpuntual	opcion1	0.421798	0.007839	0.4017	0.417275	0.42210	0.426625	0.4400
LOGIT	balanced	l2	yimpuntual	opcion2	0.421527	0.007946	0.4016	0.416450	0.42195	0.426425	0.4396
LOGIT	balanced	l2	yimpuntual	opcion3	0.416791	0.007934	0.3952	0.411775	0.41675	0.422050	0.4342
LOGIT	balanced	l2	yrecuperacionbaja	opcion1	0.317673	0.016751	0.2714	0.307125	0.31820	0.330575	0.3553
LOGIT	balanced	l2	yrecuperacionbaja	opcion2	0.319581	0.015831	0.2822	0.310050	0.32175	0.330125	0.3553
LOGIT	balanced	l2	yrecuperacionbaja	opcion3	0.317788	0.016382	0.2717	0.306175	0.31920	0.329825	0.3512
NB	None	None	ydeudor	opcion1	0.194154	0.020504	0.1453	0.183450	0.19395	0.205650	0.2363
NB	None	None	ydeudor	opcion2	0.237967	0.019671	0.1852	0.225600	0.23700	0.250575	0.2814
NB	None	None	ydeudor	opcion3	0.194893	0.021811	0.1425	0.180900	0.19650	0.211625	0.2483
NB	None	None	yimpuntual	opcion1	0.170492	0.011637	0.1411	0.163050	0.16995	0.178000	0.2059
NB	None	None	yimpuntual	opcion2	0.256555	0.011898	0.2245	0.250000	0.25570	0.264650	0.2904
NB	None	None	yimpuntual	opcion3	0.175618	0.012047	0.1455	0.167650	0.17590	0.184100	0.2084
NB	None	None	yrecuperacionbaja	opcion1	0.197833	0.021695	0.1510	0.184325	0.19660	0.213525	0.2418
NB	None	None	yrecuperacionbaja	opcion2	0.236411	0.020403	0.1793	0.222400	0.23675	0.249425	0.2865
NB	None	None	yrecuperacionbaja	opcion3	0.197440	0.021746	0.1441	0.184150	0.19640	0.212425	0.2515
RF	None	None	ydeudor	opcion1	0.313412	0.016489	0.2816	0.300450	0.31325	0.323650	0.3711
RF	None	None	ydeudor	opcion2	0.316227	0.016502	0.2805	0.304050	0.31485	0.327400	0.3655
RF	None	None	ydeudor	opcion3	0.315400	0.015960	0.2810	0.306025	0.31405	0.327000	0.3713
RF	None	None	yimpuntual	opcion1	0.442142	0.008753	0.4208	0.437100	0.44280	0.447200	0.4632
RF	None	None	yimpuntual	opcion2	0.442569	0.008703	0.4214	0.438175	0.44315	0.447200	0.4653
RF	None	None	yimpuntual	opcion3	0.442140	0.008530	0.4175	0.438075	0.44240	0.447375	0.4611
RF	None	None	yrecuperacionbaja	opcion1	0.314514	0.016458	0.2786	0.301825	0.31625	0.324975	0.3587
RF	None	None	yrecuperacionbaja	opcion2	0.315545	0.016462	0.2773	0.302400	0.31545	0.326500	0.3540
RF	None	None	yrecuperacionbaja	opcion3	0.316627	0.015970	0.2840	0.304775	0.31715	0.326850	0.3576

RF	balanced	None	ydeudor	opcion1	0.313411	0.015303	0.2831	0.300750	0.31405	0.323525	0.3473
RF	balanced	None	ydeudor	opcion2	0.315101	0.015935	0.2787	0.304600	0.31555	0.326075	0.3550
RF	balanced	None	ydeudor	opcion3	0.314535	0.015799	0.2789	0.301850	0.31480	0.323850	0.3546
RF	balanced	None	yimpuntual	opcion1	0.442108	0.008823	0.4195	0.436375	0.44275	0.447575	0.4634
RF	balanced	None	yimpuntual	opcion2	0.442531	0.008628	0.4214	0.438050	0.44300	0.446825	0.4653
RF	balanced	None	yimpuntual	opcion3	0.442214	0.008460	0.4170	0.438075	0.44245	0.447575	0.4608
RF	balanced	None	yrecuperacionbaja	opcion1	0.313847	0.015172	0.2792	0.302225	0.31410	0.325925	0.3410
RF	balanced	None	yrecuperacionbaja	opcion2	0.314514	0.015244	0.2810	0.302500	0.31415	0.327100	0.3511
RF	balanced	None	yrecuperacionbaja	opcion3	0.314977	0.015685	0.2770	0.302550	0.31385	0.327050	0.3499
XGB	None	None	ydeudor	opcion1	0.313298	0.016374	0.2804	0.300700	0.31410	0.323025	0.3543
XGB	None	None	ydeudor	opcion2	0.316078	0.016933	0.2689	0.306150	0.31840	0.324275	0.3601
XGB	None	None	ydeudor	opcion3	0.310695	0.016040	0.2795	0.300400	0.31065	0.322025	0.3539
XGB	None	None	yimpuntual	opcion1	0.443036	0.009313	0.4178	0.438425	0.44325	0.448475	0.4640
XGB	None	None	yimpuntual	opcion2	0.443554	0.008781	0.4203	0.438925	0.44330	0.449350	0.4667
XGB	None	None	yimpuntual	opcion3	0.442714	0.008774	0.4171	0.438125	0.44200	0.448050	0.4618
XGB	None	None	yrecuperacionbaja	opcion1	0.312701	0.016030	0.2754	0.300100	0.31315	0.322400	0.3516
XGB	None	None	yrecuperacionbaja	opcion2	0.315027	0.016284	0.2688	0.306375	0.31550	0.325225	0.3603
XGB	None	None	yrecuperacionbaja	opcion3	0.311928	0.016574	0.2741	0.301925	0.31070	0.322300	0.3528

Apéndice B

Códigos para la generación de tablas y gráficas

B.1. Códigos para el entrenamiento y prueba de los modelos de clasificación.

A continuación se muestra el código de Python utilizado para hacer el entrenamiento y prueba de los modelos de clasificación utilizados.

```
#Se hace el siguiente experimento:
#1 Se ajustan varios modelos de aprendizaje: KNN, NB, LDA, SVM, LOGIT, RF,
    XGB
#2 Para cada modelo se experimenta con tres variables a explicar:
    yimpuntual, yrecuperacionbaja, ydeudor.
#3 Para cada modelo y para cada variable <y> se experimenta con subgrupos
    de variables explicativas <x>
#4 Se calcula para cada combinacion el desempeno  con AUC, KS,  y matriz
    de confusion.
#5 Se hace un preprocesamiento y un  de ajuste en los modelos.

#%%
#*************************************************************************
#LIBRERIAS
#*************************************************************************
```

```
#Cargar paquetes
import pandas as pd
import numpy as np
#import itertools
import matplotlib.pyplot as plt
import scorecardpy as sc

from pickle import dump, load
from xgboost import XGBClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import LinearSVC
from sklearn.feature_selection import SelectKBest, chi2
from sklearn import metrics
from time import time, strftime



#*****************************************************************************
#SECCION PRELIMINARES
#*****************************************************************************
lista_balanceo=[None,'balanced']
lista_penalidad=[None,'l1','l2']

#Lista de modelos a reportar desempeno
lista_modelNames = ['KNN','NB','LDA', 'LinearSVCDual','LinearSVCPrimal',
    'LOGIT','RF', 'XGB']

#Eligiendo columnas X y columnas Y
lista_variables_y=['yimpuntual','yrecuperacionbaja','ydeudor']

#todas las variables x disponibles
opcion_1=["loan_amount",'number_of_payments',"interest_rate","tx_found",
"max_amount","activity_age_days","estimated_monthly_income","average_balance",
"min_balance","max_balance","loan_number","gender", 'cocienteingresodeuda',
'edad1','edad2', 'edad3', 'edad4','edad5']
```

```
#variables del individuo, caracteristicas del credito.
opcion_2= ["loan_amount", 'number_of_payments' ,
    "interest_rate","activity_age_days",
"estimated_monthly_income","average_balance","loan_number","gender",
'cocienteingresodeuda' , 'edad1','edad2', 'edad3']

#variables de caracteristicas del credito, y cuentas.
opcion_3= ["loan_amount", 'number_of_payments' , "interest_rate",
    'tx_found', 'max_amount',  "activity_age_days",
    "average_balance",'min_balance',
'max_balance',  "loan_number"]

#Lista de modelos a entrenar
seleccionX = []
seleccionX.append(('opcion1', opcion_1))
seleccionX.append(('opcion2', opcion_2))
seleccionX.append(('opcion3', opcion_3))

N=100
num_repeticiones=range(0,N,1)
#*****************************************************************************
#SECCION DATOS
#*****************************************************************************
#%%Dividir los datos en conjuntos de entrenamiento y prueba (70% - 30%)
#Cargar datos
data=pd.read_csv('informacionFinalcolSelectas.csv')
train_ratio = 0.7
#*****************************************************************************
#SECCION DE FUNCIONES
#*****************************************************************************
#%%
def funcionEntrenamiento(par_i,par_train_data,par_x_columns,par_yname,
    par_opcion,par_models,par_balanceo,par_penalidad,par_prefix,par_suffix):
    x_train = par_train_data[par_x_columns]
    y_col=[par_yname]
    y_train = par_train_data[y_col]

    #Entrenando los modelos
    for name, model in par_models:
```

```
        inicio = time()
        print(f'\n Entrenando el modelo: {name}, repeticion: {par_i},
            seleccionY: {par_yname}, seleccionX: {par_opcion},
            weight-class: {par_balanceo}, penalidad: {par_penalidad}')
        y_train_flat = y_train.values.ravel() # Aplana y_train para
            convertirlo en un array 1D
        model.fit(x_train, y_train_flat)
        dump(model, open(f'modelos/{par_prefix}_{par_i}_{name}_
        {par_yname}_{par_opcion}_{par_balanceo}_{par_penalidad}_
        {par_suffix}.pkl', 'wb'))
        tiempo_final=time() - inicio
        print(f'{name}: el entrenamiento duro {tiempo_final} segundos')

def funcionReportarDesempenio(par_i, par_train_data, par_test_data,
    par_x_columns, par_y_col, par_yname, par_opcion, par_models, par_balanceo,
    par_penalidad, par_prefix, par_suffix, par_a, par_b, par_umbral):

    x_train = par_train_data[par_x_columns]
    y_train= par_train_data[par_y_col].values.ravel() # Aplana y_train
        para convertirlo en un array 1D
    x_test = par_test_data[par_x_columns]
    y_test = par_test_data[par_y_col].values.ravel() # Aplana y_train para
        convertirlo en un array 1D
    lista_reporte=[]
    #Generar metricas a reportar
    for name in par_models:
        #Cargando el modelo
        model = load(open(f'modelos/{par_prefix}_{par_i}_{name}
        _{par_yname}_{par_opcion}_{par_balanceo}_{par_penalidad}_
        {par_suffix}.pkl', 'rb'))
        #Calculo de AUC
        if(not name in ["LinearSVCDual",'LinearSVCPrimal']):
            #Haciendo predicciones de probabilidades y clases
            probest_test = model.predict_proba(x_test)
            probest_train = model.predict_proba(x_train)
            y_predict_test=np.where(probest_test[:,1]>par_umbral,1,0)
            AUC_test=metrics.roc_auc_score(y_test, probest_test[:, 1])
            AUC_train=metrics.roc_auc_score(y_train, probest_train[:, 1])
        if(name in ["LinearSVCDual",'LinearSVCPrimal']):
```

```
        AUC_test=-1
        AUC_train=-1
        y_predict_test=model.predict(x_test)
        ks_test=-1
    if ((AUC_test>=par_a and AUC_test<=par_b) or name in
        ["LinearSVCDual",'LinearSVCPrimal']):
        '''if name=="KNN":
            #Seleccionar las mejores caracteristicas utilizando
                chi-cuadrado
            selector = SelectKBest(score_func=chi2, k=5)  #
                Seleccionar las 5 mejores caracteristicas
            print(selector)
        elif name=="RF":
            coefficients = model.feature_importances_
            feature_coefficient_mapping = dict(zip(x_columns,
                coefficients))
            print("\n Coeficientes que indican la importancia de las
                caracteristicas de mayor a menor:")
            diccionario_ordenado =
                dict(sorted(feature_coefficient_mapping.items(),
                key=lambda item: item[1], reverse=True))
            print(dict(itertools.islice(diccionario_ordenado.items(),
                4)))

        elif name in ["LOGIT","LDA",'LinearSVCDual','LinearSVCPrimal']:
            coefficients = model.coef_[0]
            feature_coefficient_mapping = dict(zip(x_columns,
                coefficients))
            print("\n Coeficientes que indican la importancia de las
                caracteristicas de mayor a menor:")
            diccionario_ordenado =
                dict(sorted(feature_coefficient_mapping.items(),
                key=lambda item: item[1], reverse=True))
            print(dict(itertools.islice(diccionario_ordenado.items(),
                4)))

        print("\n_____")
        print("_____")
```

```
            print(f"ID: MODELO:{name}, VARIABLE OBJETIVO:{var_obj},
                VARIABLES INPUT:{par_opcion}, weight-class:{par_balanceo},
                penalidad:{par_penalidad}")
            print(f"\n AUC_TRAIN:{AUC_train}, AUC_TEST:{AUC_test}")
            print("_____________________________________________________________")
            '''
            #print(f'\n Parametros {model.get_params}')
            if(not name in ["LinearSVCDual",'LinearSVCPrimal']):
                #sc.perf_eva(y_train, probest_train[:, 1], title = "train")
                ks_test=sc.perf_eva(y_test, probest_test[:, 1], title =
                    "test",show_plot=False)['KS']
                print(ks_test)
                #KS
                #plot_ks_statistic(y_test,probest_test,figsize=(10,4))
                #ROC
                #plot_roc(y_test,probest_test,plot_micro=False,plot_macro=False,
                figsize=(10,4))
            # Matriz de confusion normalizada
            cm = metrics.confusion_matrix(y_test, y_predict_test,
                normalize="true")
            lista_reporte.append((f'{name}',f'{i}',f'{par_yname}',
            f'{par_opcion}', f'{par_balanceo}',f'{par_penalidad}',
            AUC_train,AUC_test,ks_test,cm[0,0],cm[1,1]))
            #disp=metrics.ConfusionMatrixDisplay(confusion_matrix=cm)
            #disp.plot()
            #plt.show()
    return(lista_reporte)
#*****************************************************************************



#*****************************************************************************
#MAIN
#*****************************************************************************
#%%Ciclo principal de entrenamiento
lista_reporte=[]
for i in num_repeticiones:
    print("\n")
    print(strftime("%c"))
```

```
inicio = time()
train_data = data.sample(frac=train_ratio)  # 70% para entrenamiento
test_data = data.drop(train_data.index)  # El resto (30%) para prueba
for opcionx, x_columns in seleccionX:
    for var_obj in lista_variables_y:
        for ppenalizacion in lista_penalidad:
            for pbalanceo in lista_balanceo:
                #Lista de modelos a entrenar
                modelsFull = []
                if(pbalanceo==None and ppenalizacion==None):
                    modelsFull.append(('KNN',
                        KNeighborsClassifier(n_jobs=-1,n_neighbors=4)))
                    modelsFull.append(('NB', GaussianNB()))
                        #Posiblemente no haga sentido utilizar este
                        modelo ya que las variables deben ser
                        categoricas
                    modelsFull.append(('LDA',
                        LinearDiscriminantAnalysis(solver='svd')))
                if(pbalanceo in ['balanced',None] and
                    ppenalizacion==None):
                    modelsFull.append(('RF',
                        RandomForestClassifier(bootstrap=False,
                         random_state=56,n_jobs=-1,
                         n_estimators=500,
                         class_weight = pbalanceo,max_depth=6)))
                if(pbalanceo==None and ppenalizacion==None):
                    modelsFull.append(('XGB',
                        XGBClassifier(random_state=56, n_jobs=-1,
                        n_estimators=300,max_depth=6,learning_rate=.01)))
                if(ppenalizacion in ['l2',None] and pbalanceo in
                    ['balanced',None]):
                    modelsFull.append(('LOGIT',
                        LogisticRegression(random_state=56, n_jobs=-1,
                        tol=1e-3,max_iter=90000,
                        solver='newton-cg',class_weight=pbalanceo,
                        penalty=ppenalizacion)))
                if(pbalanceo in [None,'balanced'] and ppenalizacion in
                    ['l1','l2']):
```

```
                        #Disminuye  en calidad seleccionar dual=True y no
                            mejora con class-weight='balanced'. Costo alto
                            de computo.
                        #modelsFull.append(('LinearSVCDual',
                            LinearSVC(random_state=56,
                            tol=1e-4,max_iter=60000,
                             dual=True,class_weight=pbalanceo,
                             penalty=ppenalizacion)))
                        #Mejora con class-weight='balanced' para ydeudor,
                            yrecuperacionbaja.
                        modelsFull.append(('LinearSVCPrimal',
                            LinearSVC(random_state=56,
                             tol=1e-5,max_iter=10000,
                             dual=False,class_weight=pbalanceo,
                             penalty=ppenalizacion)))
                    funcionEntrenamiento(par_i=i,par_train_data=train_data,
                    par_x_columns=x_columns,par_yname=var_obj,
                    par_opcion=opcionx,par_models=modelsFull,
                    par_balanceo=pbalanceo,par_penalidad=ppenalizacion,
                    par_prefix='01',par_suffix='ronda1')
fin = time()
tiempoF=fin-inicio
#261 segundos sin lsvcDual en laptop-workstation
print("\n_____")
print("_____")
print(f'El tiempo del ciclo {i} de entrenamiento fue {tiempoF}
    segundos')
print("_____")
print("_____")

#Ciclo {i} del reporte de desempeno
#Observar que AUC, KS de LinearSVC no tiene ningun significado.
inicio2 = time()
for opcionx, x_columns in seleccionX:
        for var_obj in lista_variables_y:
            for ppenalizacion in lista_penalidad:
                for pbalanceo in lista_balanceo:
                    y_col=[var_obj]
                    modelNames = []
```

```
                    if(pbalanceo==None and ppenalizacion==None):
                        modelNames.append('KNN')
                        modelNames.append('NB')
                        modelNames.append('LDA')
                    if(pbalanceo in ['balanced',None] and
                       ppenalizacion==None):
                        modelNames.append('RF')
                    if(pbalanceo==None and ppenalizacion==None):
                        modelNames.append('XGB')
                    if(ppenalizacion in ['l2',None] and pbalanceo in
                       ['balanced',None]):
                        modelNames.append('LOGIT')
                        ppenalizacion='none'
                    if(pbalanceo in [None,'balanced'] and
                       ppenalizacion in ['l1','l2']):
                        modelNames.append('LinearSVCPrimal')
                    regresar=funcionReportarDesempenio(par_i=i,
                        par_train_data=train_data,par_test_data=test_data,
                        par_x_columns=x_columns,par_y_col=y_col,
                        par_yname=var_obj,par_opcion=opcionx,
                        par_models=modelNames,par_balanceo=pbalanceo,
                        par_penalidad=ppenalizacion,par_prefix='01',
                        par_suffix='ronda1',par_a=.5,par_b=.9,par_umbral=.5)
                    lista_reporte.extend(regresar)
                    fin2 = time()
    tiempoF2=fin2-inicio2
    print("\n_____")
    print("_____")
    print(f'El reporte {i} duro {tiempoF2} segundos') #296 segundos
    print("_____")
    print("_____")
#FIN DEL CICLO for i in num_repeticiones:

reporte=pd.DataFrame(lista_reporte)
reporte.columns=['Modelo','repeticion','y','x','balanceo','penalidad',
'auc-train','auc-test','KS','diagonal1','diagonal2']
reporte=reporte.sort_values(by=['Modelo','y','x','balanceo',
'penalidad','repeticion'])
reporte.to_csv('Reporte.csv', sep=',', index=False, encoding='utf-8')
```

```
#%%
```

B.2. Códigos para la generación de tablas descriptivas de resultados

Este código toma un DataFrame llamado data que contiene información sobre varios modelos, variables, y métricas de evaluación; 'auc-train', 'auc-test', 'KS', 'diagonal1' y 'diagonal2'. Agrupa estos datos según algunas columnas específicas de 'Modelo', 'balanceo', 'penalidad', 'y', y 'x'. Luego, calcula estadísticas resumidas: la media, desviación estándar, mínimo, máximo y cuartiles, para estas métricas agrupadas. Finalmente, crea tablas LaTeX para cada métrica, excluyendo la columna 'count' de las estadísticas, y guarda estas tablas en archivos .tex en una carpeta llamada 'metricas/'.

```
# Agrupar los datos por Modelo, Variable "y" y Metrica
data_subset = data[['Modelo', 'balanceo', 'penalidad', 'y', 'x',
    'auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2']]

grouped_data = data_subset.groupby(['Modelo', 'balanceo', 'penalidad',
    'y', 'x'])

# Calcular la media, desviacion estandar, minimo, maximo y los cuartiles
summary_stats = grouped_data['auc-train', 'auc-test', 'KS', 'diagonal1',
    'diagonal2'].describe()

#%%Guardar el resultado
for metric in ['auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2']:
    individual_table=summary_stats[metric].reset_index().drop('count',
        axis=1)
    individual_table.to_latex('metricas/'+metric+'.tex', index=False)
```

B.3. Código para la generación de gráficas de caja

El código utiliza boxplots para representar gráficamente la distribución de las métricas de rendimiento ('auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2'). Estos boxplots permiten visualizar la variabilidad y

la distribución de estas métricas. La separación por color en los boxplots se realiza según los distintos valores de la variable 'y', lo que facilita la comparación entre las diferentes categorías.

```
# Filtrar y ordenar los datos
data_subset = data[['Modelo', 'balanceo', 'penalidad', 'y', 'x', 'auc-train',
                    'auc-test', 'KS', 'diagonal1', 'diagonal2']]

# Crear graficos de boxplots para cada metrica
metrics = ['auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2']

for metric in metrics:
    plt.figure(figsize=(12, 6))
    sns.boxplot(x='Modelo', y=metric, hue='y', palette='cool', data=data_subset)
    plt.title(f'Distribucion de {metric} por Modelo y Variable "y"')
    plt.xlabel('Modelo')
    plt.ylabel(metric)
    plt.legend(title='Variable "y"')
    plt.show()
```

B.4. Código para la generación de gráficas de barras

El código tiene como objetivo principal explorar y visualizar la relación entre diferentes métricas de rendimiento ('auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2') y diferentes modelos en un conjunto de datos. Este conjunto de datos esta relacionado con los modelos predictivos de aprendizaje automático, que han sido evaluados en varias condiciones.

1. Selección y Preparación de Datos: Primero, se realiza una selección de las columnas relevantes del DataFrame original, lo cual incluye información sobre el modelo ('Modelo'), las condiciones experimentales ('balanceo', 'penalidad', 'y', 'x'), y las métricas de rendimiento.
2. Agrupación y Promedio: Luego, se agrupan los datos por diferentes combinaciones de condiciones ('Modelo', 'balanceo', 'penalidad', 'y', 'x'). El propósito de esta agrupación es calcular el promedio de las métricas numéricas para cada combinación única de condiciones. Esto simplifica la visualización y proporciona una visión agregada del rendimiento del modelo bajo diferentes escenarios.
3. Definición de Métricas: Se eligen específicamente las métricas de rendimiento que se van a visualizar

('auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2').

4. Creación de Gráficos: El código utiliza la biblioteca Seaborn y Matplotlib para crear gráficos de barras que muestran el rendimiento promedio de diferentes modelos para cada métrica. Cada gráfico está diferenciado por la variable 'y' (la categoría objetivo), y los modelos se muestran en el eje x. La paleta de colores se utiliza para resaltar diferentes categorías de 'y'.

5. Exploración Detallada: El bucle adicional itera sobre las combinaciones únicas de 'balanceo', 'penalidad' y 'x', creando gráficos de barras específicos para cada conjunto de condiciones. Esto permite una exploración más detallada del rendimiento del modelo en diferentes situaciones experimentales.

```
# Filtrar y ordenar los datos
data_subset = data[['Modelo', 'balanceo', 'penalidad', 'y', 'x',
    'auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2']]

# Calcular el promedio por cada combinacion de variables
data_avg = data_subset.groupby(['Modelo', 'balanceo', 'penalidad', 'y',
    'x']).mean().reset_index()

# Definir las metricas a graficar
metrics = ['auc-train', 'auc-test', 'KS', 'diagonal1', 'diagonal2']

# Iterar sobre las combinaciones unicas de 'balanceo', 'penalidad' y 'x'
for _, subset in data_avg.groupby(['balanceo', 'penalidad', 'x']):
    for metric in metrics:
        plt.figure(figsize=(12, 6))
        sns.barplot(x='Modelo', y=metric, hue='y', palette='mako_r',
            dodge=True, data=subset, ci=None)
        plt.title(f'Metrica {metric} promedio por "Modelo", Variable "y",
            "balanceo" ={subset.iloc[0]["balanceo"]}, "penalidad"
            ={subset.iloc[0]["penalidad"]}, "x" ={subset.iloc[0]["x"]}')
        plt.xlabel('Modelo')
        plt.ylabel(f'{metric} promedio')
        plt.legend(title='Variable "y"', bbox_to_anchor=(1.05, 1),
        loc='upper left')
        plt.show()
```

B.5. Código para la generación de gráficas comparativas entre las métricas Kolmogorov-Smirnov y AUC Train y AUC Test

Este código genera gráficos comparativos entre las métricas AUC-Train, AUC-Test y Kolmogorov-Smirnov para diferentes modelos de un análisis o experimento.

- KS vs AUC Train General: Este bloque de código crea un gráfico que muestra las curvas AUC-Train y KS para todo el conjunto de datos. Primero, traza las curvas AUC-Train y KS en el mismo gráfico.
- KS vs AUC Train por Modelo: Este bloque de código itera sobre cada modelo único en los datos y crea un gráfico para cada uno de ellos. Para cada modelo, selecciona las métricas AUC-Train y KS correspondientes a ese modelo, las traza en un gráfico.
- KS vs AUC Test General: Este bloque de código crea un gráfico que muestra las curvas AUC-Test y KS para todo el conjunto de datos. Primero, traza las curvas AUC-Test y KS en el mismo gráfico.
- KS vs AUC Test por Modelo: Este bloque de código itera sobre cada modelo único en los datos y crea un gráfico para cada uno de ellos. Para cada modelo, selecciona las métricas AUC-Test y KS correspondientes a ese modelo, las traza en un gráfico.

```
#%%KS vs Auc train

#%%Comparacion general

string_name_train='_auc-train_vs_ks.png'
plt.plot( data['auc-train'], label = "AUC-Train")
plt.plot( data['KS'], label = "Kolmogorov-Smirnov")
plt.title(f'Comparacion general del comportamiento de las\n metricas
    AUC-Train y Kolmogorov-Smirnov')
plt.legend()
plt.show()

#%%Comparacion entre cada modelo
```

```
lista_modelos=data["Modelo"].unique()

for model in lista_modelos:
    new_data=data[['Modelo','auc-train','KS']][data["Modelo"]==model].reset_index()
    plt.plot( new_data['auc-train'], label = "AUC-Train")
    plt.plot( new_data['KS'], label = "Kolmogorov-Smirnov")
    plt.title(f'Comparacion del comportamiento de las metricas\n AUC-Train
        y Kolmogorov-Smirnov para el modelo {model}.')
    plt.legend()
    plt.show()

# %%KS vs Auc Test

#%%Comparacion general
string_name_test='_auc-test_vs_ks.png'
plt.plot( data['auc-test'], label = "AUC-Test", color='red')
plt.plot( data['KS'], label = "Kolmogorov-Smirnov", color='green')
plt.title(f'Comparacion general del comportamiento de las\n metricas
   AUC-Test y Kolmogorov-Smirnov')
plt.legend()
plt.show()

#%%Comparacion entre cada modelo

for model in lista_modelos:
    new_data=data[['Modelo','auc-test','KS']][data["Modelo"]==model].reset_index()
    plt.plot( new_data['auc-test'], label = "AUC-Test", color='red')
    plt.plot( new_data['KS'], label = "Kolmogorov-Smirnov", color='green')
    plt.title(f'Comparacion del comportamiento de las metricas\n AUC-Test
        y Kolmogorov-Smirnov para el modelo {model}.')
    plt.legend()
    plt.show()
```