Análisis predictivo del score de riesgo crediticio mediante machine learning: una herramienta para la toma de decisiones financieras

Oscar Andrés Urrutia Dorado

Johan Andrés Franco Vanegas

Asesor

Isaac Esteban Camargo Freile

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Nota de Aceptación

Isaac Esteban Camargo Freile

Director de Trabajo de Grado

_____ _____

Jurado

2024

Dedicatoria

Dedicamos este trabajo a nuestras esposas y padres, quienes siempre han sido nuestra mayor fuente de inspiración y apoyo incondicional. Su amor y sacrificio nos han permitido perseguir mis sueños y alcanzar las metas académicas. A nuestros profesores y mentores, que con su guía y conocimientos han enriquecido nuestra formación profesional en el campo de la ciencia de datos y analítica. Su dedicación ha sido fundamental en el desarrollo de este proyecto.

Agradecimientos

Queremos expresar nuestros más sinceros agradecimientos al Dr. Isaac Esteban Camargo Freile, director de proyecto, por su invaluable orientación y apoyo a lo largo de este proyecto. Su experiencia en ciencia de datos y analítica ha sido fundamental para el desarrollo de esta investigación.

Agradezco también a los profesores del programa de especialización en Ciencia de Datos y Analítica, quienes nos han brindado las herramientas necesarias para abordar este desafío. Su dedicación y pasión por la enseñanza han dejado una huella profunda en nuestra formación profesional.

Nuestra gratitud se extiende a mis compañeros de estudio, con quienes compartí momentos de aprendizaje y colaboración. Gracias por sus valiosas aportaciones y por hacer de este camino una experiencia enriquecedora.

Finalmente, quiero agradecer a nuestras familias por su amor y apoyo incondicional. Su confianza en nosotros y su estímulo constante han sido el motor que me impulsó a alcanzar esta meta.

Resumen

Al analizar el sistema financiero en Colombia se puede evidenciar que la gestión del riesgo bancario se divide en diferentes tipologías como lo sería el riesgo crediticio, de mercado, operativo y de liquidez, siendo el primer elemento el riesgo más importante para el sector financiero, teniendo en cuenta que uno de sus principales ingresos son el otorgamiento de créditos, bajo este contexto el presente trabajo de grado tiene como objetivo implementar un modelo de predicción del score de riesgo de crediticio a partir de la evaluación de algoritmos de Machine Learning que sean implementado en el campo financiero, obteniendo una herramienta optimizada que permita a la banca tomar decisiones con un alto nivel de asertividad referente al otorgamiento o no de un crédito basado en el análisis de la viabilidad financiera y la capacidad de pago del siente o deudor.

El contexto este proyecto aplicado, se realizó sobre el conjunto de datos públicos "credibilidad – Datos crediticios alemanes" obtenido de la plataforma Kaggle, efectuando un análisis a 1000 registros y 20 variables de información suministrada por esta plataforma web; implementando y entrenando diferentes algoritmos de machine learning para construir el modelo de score de riesgo crediticio que permitieran evaluar el rendimiento y desempeño de las diferentes métricas utilizadas como precisión, recall y F1-score.

Los resultados obtenidos en este trabajo demuestran el nivel de precisión del modelo, destacando significativamente su rendimiento, para ayudar a las instituciones financieras al momento de evaluar y analizar la posibilidad que tiene la entidad de sufrir pérdidas derivadas de un impago parcial o total de los créditos que pretenden otorgar o fueron concedidos a sus clientes o deudores.

Palabras claves: Decisión, machine learning, modelo, predicción, score.

Abstract

When analyzing the financial system in Colombia, it can be seen that banking risk management is divided into different typologies such as credit, market, operational and liquidity risk, the first element being the most important risk for the financial sector, taking into account that one of its main revenues is the granting of loans, In this context, the objective of this degree work is to implement a predictive model of credit risk score from the evaluation of Machine Learning algorithms that are implemented in the financial field, obtaining an optimized tool that allows the bank to make decisions with a high level of assertiveness regarding the granting or not of a credit based on the analysis of the financial viability and the payment capacity of the customer or debtor.

The context of this applied project was carried out on the data set “credibility - German credit data” Kaggle, performing an analysis of 1000 records and 20 variables of information provided by this web platform; implementing and training different machine learning algorithms to build the credit risk score model to evaluate the performance of the different metrics used such as accuracy, recall and F1-score.

The results obtained in this work demonstrate the level of accuracy of the model, highlighting significantly its effectiveness in helping financial institutions when evaluating and analyzing the possibility of the entity suffering losses derived from a partial or total default of the loans that they intend to grant or were granted to their clients or debtors.

***Keywords*:** Decision, machine learning, model, prediction, score.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Planteamiento del Problema

El score de riesgo crediticio es una instrumento con gran importancia en las entidades financieras colombianas, porque con ello determina un rol determinante al momento de asignar un crédito o de gestionar cartera en una entidad financiera, siendo así un elemento que permite a las personas de acuerdo a su puntaje crediticio acceder y obtener beneficios en este mercado, generando inclusión o exclusión de los productos y servicios financieros para el individuo; problemática que se abordara en este documento, destacando que la inclusión finiera como un mecanismo que facilita la consecución de 7 de los 17 objetivos de desarrollo sostenible (banco mundial, 2022), así mismo según lo mencionado por Lesmes (2023), se evidencian diferentes inconvenientes en el cálculo del score de riesgo crediticio generando que los consumidores del sector financiero obtén por otras opciones de financiamiento como lo son los créditos informales, que de acuerdo a Ocho y Villares (2016), son debido al limitado acceso de fuentes formales por las barreras que presenta una persona al solicitar un crédito en una entidad financiera (papeleos y garantías).

En atención a lo anterior y de acuerdo al contexto que escribe nuestro conjunto de datos ubicado en la plataforma web de kaggle (2024), señala que el banco al recibir una solicitud de préstamo analiza el perfil del solicitante para tomar una decisión encaminada a conceder o no el préstamo, basándose en los datos tradicionales como cuenta, saldo, pagos, ahorros, empleo, estado civil, créditos y garantía entre otras variables relevantes de la capacidad económica de la persona, para predecir el riesgo potencial que tiene el individuo de pagar o entrar en mora en el crédito o servicio financiero que se otorga, por lo cual el presente estudio aborda la relevancia del score de riesgo crediticio, teniendo como eje central la estimación del modelo con los datos que permiten analizar el nivel de riesgo, puntaje e historial crediticio del solicitante, como

elementos de información que influencia en la credibilidad que tiene ante una entidad financiera para desembolsarle un crédito. Como bien se menciona en un estudio de la financiera Bayport Colombia y la firma Experian (El Tiempo, 2021), “en promedio el 79 por ciento de los créditos de libranza otorgados por entidades financieras tradicionales se desembolsaron a personas con puntaje superior a 630 (riesgo moderado)”, analizando una escala de riesgo de 150 a 1000 puntos.

Esta problemática ha venido siendo más notoria al paso del tiempo, esto por que incluye varios aspectos que impactan al consumidor financiero, como la menciona Asobancaria (2019) con la demostración de transparencia donde no se dispone de un score conocido por el usuario, ni se conoce que determina su cálculo ante posibles errores de las entidades financieras. Adicional por sus variables de cálculo al obtener un puntaje bajo limita al consumidor a obtener productos o acotar su capacidad ante emergencias financieras Saavedra (2023).

La literatura ha evidenciado que existen métodos alternativos que dependen de aprendizajes automáticos basados en ingesta de datos y modelado predictivo que incluye una variedad de información y variables que permitan mostrar una solución para que las minorías poblacionales tengan una inclusión favorable en el sistema financiero, lo cual hace que decrezca el riesgo de crédito e incremente el acceso al mismo (García, 2020). Estos métodos no son tan comunes ya que existe en la actualidad un monopolio usado por los establecimientos financieros que a través de compañías privadas calculan el score crediticio utilizando como insumo información histórica de los deudores entregada por las entidades autorizadas, brindando un puntaje proyectado según su comportamiento y habito de pago que garantice un cálculo asertivo del riesgo crediticio como lo destaca Espinosa (2014).

Es por ello, que la central de riesgo Data Crédito (2021), señala que para poder entender cómo funciona el puntaje del crédito, hay que calcular el score de riesgo crediticio sobre cuatro factores; hábito de pago, endeudamiento, composición de su portafolio y experiencia crediticia por años, siendo estas variables posibles detonantes que dinamizan la inclusión o exclusión de las personas en el sistema financiero, las cuales se deben analizar de una manera más detallada para encontrar la solución mediante técnicas de aprendizaje automatizado que permita al individuo tener mayor probabilidad de obtener los servicios financiero finales y de banca tradicional (Pucha, 2022), sin importar edad, estrato o sus actividades financieras pasadas, sino teniendo en cuenta una proyección respaldada por su capacidad de pago.

Lo anterior nos lleva a la pregunta ¿Es posible implementar una herramienta de análisis de score crediticio que permita la toma de decisiones en el sistema financiero, por medio de técnicas y evaluación de modelos de machine learning?

Justificación

El score de riesgo de crédito debe ser un puntaje calculado para la inclusión financiera, donde sus variables de cálculo tengan la generalidad de poder brindar a cualquier sector de una población acceder a productos financieros como préstamos personales, hipotecas y tarjetas de créditos (Espinosa, 2014). Este puntaje crediticio tiene una gran relevancia ya que puede impulsar el crecimiento económico para los consumidores, aumenta el acceso a recursos y genera una asignación más eficiente del riesgo con sus provisiones de cartera. Lo anterior como lo indican algunos autores en este tema que señalan "cuando no hay acceso a la información o el acceso es asimétrico, obtener préstamos se vuelve más difícil, más costoso y menos eficiente" (transunion White paper, 2007, p. 3).

Disponer de un buen cálculo de score financiero ayuda a las compañías privadas y a los deudores a que se pueda acceder a diferentes productos y servicios, desde el ámbito financiero hasta en procesos de contrato de un inmueble en arriendo u obtener beneficios en solicitudes de créditos con mejores tasas de interés de créditos (Grupo Bancolombia, 2021). Pero dicho score no es público, y dispone de un método basado en la vida crediticia del cliente, por tal motivo existen algunos proyectos que experimentales que están realizando un cálculo de score financiero basado en variables que se acomoden a cada necesidad Giraldo (2021).

Como mencionamos anteriormente el riesgo crediticio puede ser el mayor riesgo al que se enfrenta el sistema financiero, teniendo en cuenta que esta herramienta evita perdidas y maximiza ganancias, elementos claves de cualquier empresa para tener éxito, solvencia y rentabilidad en el tiempo. Si analizamos este contexto en nuestro país se puede evidenciar que las entidades bancarias para otorgar un producto o servicio como ahorro, inversión y financiación entre otros, tiene en cuenta el score de riesgo crediticio, que permite evaluar de la persona el

historial de pago, saldos pendientes, deudas no pagadas, antigüedad de su historial de crédito y reporte a centrales de riesgo, posteriormente el establecimiento bancario crea un perfil del individuo sobre su capacidad económica para aprobarle o no el servicio financiero que desea adquirir con esa entidad (El Tiempo, 2021). De acuerdo con el artículo de este medio de comunicación escrito, en Colombia el Score se mide de 0 a 1000 puntos que dependen de varios factores como los mencionamos anteriormente, permitiendo a las entidades construir una escala de riesgo, así:

Tabla 1

Escala de Riesgo Crediticio

Ítem	Puntaje	Nivel de Riesgo
1	Mayor a 746	Muy Bajo
2	Entre 646 y 745	Bajo
3	Entre 476 y 475	Moderado
4	Entre 421 y 475	Medio
5	Entre 386 y 420	Medio
6	Entre 341 y 385	Medio
7	Entre 301 y 340	Medio
8	Entre 261 y 300	Alto
9	Entre 150 y 260	Muy Alto

Nota. Esta tabla representa el puntaje o nivel de riesgo para otorgarle a una persona un préstamo u otros productos financieros en Colombia. Tomado de Periódico El Tiempo (2021).

En atención al cuadro anterior, se puede mencionar que, si la puntuación de una persona se encuentra en el rango de los 646 puntos, le representaría al banco un riesgo bajo y más probabilidades de acceder a créditos u otros productos de este sector. Por lo cual al analizar el estudio de la Financiera Bayport Colombia y la firma Experian (2021), indican que el 79% de los

créditos de libranza otorgados por las financieras tradicionales fueron desembolsados a personas con puntaje superior a 630 puntos, que representan un riesgo moderado.

De esta manera al analizar la importancia del score crediticio en nuestro país y la información que utilizan para otorgar o no un crédito a una persona, encaminamos el presente estudio en realizar un modelo practico enfocado en variables que permitan estandarizar un modelo genérico dando como resultado un score de crédito financiero que permita identificar un cliente con capacidad de pago según su monto adeudar (Saavedra, 2023). Este cálculo del riesgo de crédito es base de cada entidad financiera que otorga obligaciones a deudores y está es fundamentada bajo la circular externa 52 de 2011 dirigida por Superfinanciera (2011) indica que "Elementos mínimos a considerar en la evaluación del riesgo de crédito de las carteras colectivas, así como para la valoración de inversiones en títulos valores y demás derechos de contenido económico".

Por tal motivo, este modelo de aprendizaje automático con diversas técnicas pretende contribuir en el cálculo del riesgo de crédito cumpliendo con las normas establecidas para su control desde la superintendencia financiera y brindando un cálculo más asertivo con el score de crédito que genera gran relevancia en empresas privadas (Flores y Malca, 2017), dando una tranquilidad de solvencia económica y a sus deudores en obtener beneficios en sus créditos o en adquirir servicios o como menciona en la plataforma web kaggle en su concurso "credibilidad – Datos crediticios alemanes" (2024), este modelo permitiría a los proveedores de entidades financieras aceptar más solicitudes de préstamos, ayudando a las personas que históricamente se le ha negado algún crédito debido a la falta de historial crediticio.

Objetivos

Objetivo General

Implementar un modelo de machine learning para la predicción precisa del score de riesgo crediticio, facilitando así una herramienta sólida que mejore la toma de decisiones financieras en instituciones crediticias.

Objetivos Específicos

Interpretar un conjunto de datos de públicos sobre credibilidad – Datos crediticios alemanes, para obtener las características de información financiera de un individuo que aporten a disminuir el riesgo del perfil del solicitante para obtener un crédito en una entidad bancaria.

Identificar las variables más significativas de la base de datos credibilidad – Datos crediticios alemanes, para estimar y evaluar el modelo de aprendizaje automático, que permita tener un nivel de asertividad optimo en el cálculo del score financiero.

Validar los resultados del modelo con los factores que influyen en la calificación crediticia, basado en el conjunto de datos tratados.

Marco Conceptual y Teórico

Uno de los principales dolores de cabeza de la economía mundial es el gasto excesivo financiado con crédito donde no hay garantías de cumplimientos de las obligaciones, problemática rebasa el actuar de los gobiernos para contener la fuente de la crisis, un ejemplo internacional de esta situación se presentó con la burbuja del mercado inmobiliario en Estados Unidos en los años de 2008 y 2009, tuvo su origen principalmente por una bajada de las de interés en sistema financiero estadunidense generando exceso del gasto y sobreendeudamiento en la mayoría de sus ciudadanos norteamericanos, lo cual llevo aun gran incremento del gasto principalmente en el sector inmobiliario, donde muchas personas adquirieron hipotecas por encima de su capacidad de pago (González, 2009). En estas problemáticas de gran impacto histórico para el sector financiero se evidencio una serie de deficiencias en el desarrollo del análisis y estudio del score de riesgo crediticio llevado por los analistas del sistema bancario para el otorgamiento de créditos hipotecarios a individuos o empresas que no contaban con las garantías necesarias para cumplir con sus compromisos, como señala García (2019) “se entregaban préstamos a los clientes con supuestas calificaciones AAA sin riesgo, pero en realidad no se tenía garantías de que los prestamistas lograran pagar el crédito en los términos acordados con su banco, llevándolos a entrar en Credit Default Swap”.

Para analizar desde el contexto local una experiencia económica similar a la Estados Unidos, debemos tener presente la crisis que ocurrió en 1999, fue la de mayor impacto en la historia de la economía colombiana, con una caída del PIB del 4,5%, que genero el cierre así como venta de bancos privados y públicos entre las entidades financieras de renombre para esta época esta Bancafé, Banco del Estado y Granahorrar, dejando un sistema financiero más conservador, un estatuto regulatorio para el crédito de vivienda y cambios drásticos en la política

monetaria, escenario que endureció los requisitos de ingreso al crédito, llevando consigo a que varias familias colombinas que tenían hipotecas vigentes no pudieran refinanciar sus créditos imposibilitando los pagos de sus obligaciones por los altos intereses del momento, perdiendo sus hogares junto con las cuotas de los pagos que habían realizado al banco sobre sus préstamos.

Una nota editorial del Banco de la República (2023) señala que "luego de la crisis financiera de finales de la década de los noventa este sector sea venido fortaleciendo gracias a la regulación del Gobierno Nacional, reflejado en buenos indicadores de rentabilidad, riesgo y solvencia", de esta manera se afianzo el sistema financiero en Colombia, que esta conforma por entidades de crédito conocidas como (ED), establecimientos de servicios financieros (ESF) y otras instituciones financieras, que en su gran mayoría se agrupan en una figura denominada conglomerados financieros, las cuales están bajo la vigilancia de la Superintendencia Financiera de Colombia, de igual forma, así como tienen presencia en nuestro país algunas también lo hacen en el sistema financiero exterior, esta situación conllevo a que los organismos reguladores de nuestras políticas económicas y monetarias restructuran el sistema económico y financiero de Colombia (David, 2017; Banco de la Republica 2017), lo cual permitió sobrellevar la crisis internacional de 2008 y 2009 antes mencionada.

En atención a esa problemática que marco la historia financiera y económica de la región, este proyecto aplicado detalla conceptos del sector financiero colombiano, que se enfoca en construir un modelo que permita calcular el score de riesgo crediticio en nuestro país, empleando técnicas de aprendizaje automático y diversos métodos de análisis estadísticos, que sustenten la metodología utilizada en el desarrollo y validación de este modelo para el "Análisis Predictivo del Score de Riesgo Crediticio mediante Machine Learning", teniendo en cuenta que el riesgo crediticio es una variable fundamental de las entidades financieras para conocer el nivel de riesgo

que puede enfrentar un banco al generar un préstamo de dinero incobrable o de activo moroso, como lo señala Castro (2022), “la gestión del riesgo de crédito en los bancos es crucial para evitar el riesgo de perder dinero debido a los malos prestamos” (p.168).

De esta manera al sumergimos en el desarrollo de nuestro modelo de machine learning hay diferentes estudios e investigaciones que se han enfocado en la predicción del score financiero usando métodos de aprendizaje automatizado, entre ellos resaltamos el trabajo realizado por Ocampo Ortin & Orrego Quintero, 2023, para realizar su investigación utilizando también una base de datos de la plataforma de Kaggle, focalizando su proyecto en la potenciación de modelos predictivos mediante la implementación de técnicas como regresión logística, clasificado K-NN, random forest, gradient bossting y xgbosst classifier, esta dos últimas técnicas después de rigurosas evaluaciones y ajustes al modelo lograron obtener una presión del 89% y 90% para que las entidades bancarias definan de manera eficiente el riesgo financiero asociado a la solicitud de préstamos.

El machine learning o aprendizaje automático es una disciplina del campo de la inteligencia artificial, la cual mediante algoritmos permite a los computadores tener la capacidad para identificar patrones en datos masivos y construir predicciones “análisis predictivo”. este modelo de aprendizaje logra que los ordenadores lleven a cabo tareas específicas de manera autónoma, es decir que es capaz de reconocer la situación problema que se está tratando y reaccionar utilizando la estrategia aprendida en la cual de modela esta herramienta con las funciones o técnicas programadas. “El aspecto central del aprendizaje automático es la creación y aplicación de un algoritmo de programación que en secuencia de pasos y actividades permita dar solución a un problema determinado” (Pucha, 2022, p. 30). Es de tener en cuenta que el aprendizaje automático esta dividió por dos campos, según se presente o no una

retroalimentación del proceso que está realizando, los cuales serían Aprendizaje supervisado siendo un conjunto de algoritmos que utiliza variables etiquetadas para realizar su entrenamiento, este modelo principalmente se asocia con problemas de clasificación, regresión y "ranking problema" como lo señala Castro (2022). Aprendizaje No Supervisado este algoritmo no presenta un conocimiento previo enfrentándose a un caos de datos con el fin de encontrar patrones que permitan organizarlos de alguna manera. Por lo cual en el contexto dirigido al "Análisis Predictivo del Score de Riesgo Crediticio mediante Machine Learning", utilizaremos esta herramienta aplicando aprendizaje supervisado no supervisado con algoritmos como método chi cuadrado, para poder observar los cálculos que estamos realizando en el modelo; correlación de person -1 a 1, 1 para medir la relación estadística entre variables continuas; peso de la evidencia de WOE que logra recodificar automáticamente los valores de las variables predictoras y categóricas; p-value la cual ayuda a la hora de seleccionar los parámetros y modelos que mejor se adaptan a los datos de nuestra matriz; análisis de variables esta técnica analiza de manera simultánea las variables que están siendo sometidas a la investigación que realizamos; índice de Gini nos permitiría disminuir las impurezas desde los nodos de la raíz si trabajamos este modelo con un método denominado árbol de decisiones. Tal como señala en su investigación aplicada Cepeda (2022) al utilizar técnicas de machine learning en los modelos del score crediticio para realizar el análisis predictivo del riesgo crediticio en las entidades financieras, se evidencia gran variedad de metodologías y técnicas en la literatura para obtener y comparar el mejor modelo predictivo que ayude a minimizar el riesgo y optimizar las ganancias de las entidades financieras al generar créditos para prestatarios.

De acuerdo con el proyecto de investigación de Cepeda (2022), en cual señala que los modelos de score crediticio son técnicas muy manejadas por las instituciones financieras para

apoyarse en las decisiones de concesión de créditos al consumo. Se debe saber también que este score de riesgo crediticio en nuestro sistema financiero se analiza como la incertidumbre que involucra en primera medida la capacidad que tiene un prestatario de pagar o cumplir una obligación crediticia, según Castro (2022), señala que “el prestaría puede ser una persona u organización”. Por tal motivo se hace necesario llevar a cabo un proceso que permita analizar de manera detallada el puntaje crediticio de un prestador para conocer o predecir las probabilidades que pague una obligación financiera (prestamos) a tiempo. Por lo cual la elaboración e implementación de un modelo de aprendizaje automático disciplina de inteligencia artificial, es una herramienta importante para determinar el score como un elemento de clasificación crediticia y de análisis para identificar a que persona u organización se le podría otorgar un crédito.

Según un artículo del diario La República (2024), titulado ¿Por es importante el score crediticio?, se puede mencionar que este puntaje crediticio se desempeña como herramienta que ayuda al prestamista a poder acceder a un mayor número de productos en el sistema financiero y poder apalancar su endeudamiento en el mismo, de igual forma incide en gran medida en las condiciones favorables o desfavorables que el sector bancario le otorgue un crédito al prestatario por ejemplo en términos de acceso al monto solicitado, interés, plazo y formas de pago. Por ello toda entidad dedicada a las finanzas o microfinanzas debe contar con un sistema o metodología de gestión del riesgo para tener una forma de medir el riesgo crediticio (Aceituno, 2019).

Metodología

Para el desarrollo de este proyecto aplicado se seleccionó la metodología Crisp-DM (Cross-Industry Standard Process for Data Mining) como lo indica IBM (2021) es una metodología utilizada en proyectos de data mining y machine learning, esto porque en su metodología incluye descripciones de las fases comunes de un proyecto y como modelo brinda un resumen del ciclo vital de minería de datos.

Esta metodología flexible fue seleccionada para este proyecto ya que garantiza que este se lleve a cabo de manera eficiente, donde los resultados obtenidos sean valiosos para el enfoque de negocio objetivo.

Esta metodología nos permite revisar y ajustar las fases del proyecto según sea necesario. Con esta, se pretende tener un énfasis en los requisitos del negocio en cada etapa o paso del proyecto logrando así cumplir con los objetivos planteados Rodríguez (2023).

Para el cumplimento de la metodología seleccionada para este proyecto de Análisis Predictivo del Score de Riesgo Crediticio mediante Machine Learning, se seguirán las siguientes fases:

Comprensión del Negocio

En esta fase se detalla el proceso de entendimiento a nivel estructural donde se alinearán los objetivos del proyecto con los objetivos del negocio, brindando efecto real en uso del modelo. Para ello se definen los objetivos del negocio con su respectiva evaluación de la situación actual brindando un panorama más amplio de cuáles son los problemas o limitantes actuales, con ello se fijan los planes del proyecto y se planifican por fases o pasos.

1. Objetivo de negocio: Para ello se realizó un análisis de la problemática actual, consultando cómo funciona en la actualidad el cálculo del score financiero, con ello analizar un

modelo de machine learning para predecir la probabilidad que un cliente se le pueda brindar un producto de crédito basado en el score de riesgo crediticio.

2. Valoración de la situación: El uso del score de crédito en entidades financieras es de uso común basado en información histórica suministrada por cada ente a las compañías buro de crédito quienes generan un puntaje según variables como el comportamiento de pago de anteriores obligaciones o cantidad de deudas activas de la persona consultada.

3. Objetivos de la minería de Datos: Los objetivos del negocio en función de la minería de datos que se realizará van guiados a emplear un modelo machine learning usando técnicas de aprendizaje automático para el cálculo del score crediticio de uso del sector financiero colombiano.

4. Plan de proyecto: En el proyecto se contemplan las siguientes etapas generales con el estudio del marco teórico profundizando en el esquema actual que disponen las entidades financieras en Colombia, y profundizando en métodos de machine learning como la prueba Chi-Cuadrado, P-Valor, correlación de Pearson y WOE. Adicional con la exploración y limpieza de los datos suministrados de un concurso de kaggle por su completitud y su dificultad en obtención de otras fuentes ya que se manejan datos sensibles no públicos. Se realizará la división de los datos, seleccionando el modelo de aprendizaje automático que más se adapte a la necesidad del negocio. Con el modelo y las variables seleccionadas se genera un entrenamiento con el objetivo de validar el nivel de predicción y garantizar las fases finales en la evaluación y despliegue del modelo.

Compresión de los Datos

En esta fase se hará la recolección, comprensión del set de datos, exploración y limpieza garantizando la verificación de la calidad de estos.

1. Recolección de los datos**:** Los datos que se usarán en dicho proyecto son datos públicos de una competencia kaggle la cual brinda un set de datos completos con diferentes variables que permiten tener información de cada tercero. La información crediticia en Colombia privada y solo se puede usar con permiso de cada ciudadano, por tal motivo las entidades financieras no pueden brindar estos sets de Datos.

2. Exploración y Verificación de los datos**:** La exploración de los datos recolectados contempla bases de entrenamiento y prueba del modelo, con tres niveles de profundidad en formato CSV. Contiene información de proveedores fiscales A, B, C y proveedores de buró de crédito A, B con bases de depósitos, personas, tarjetas débito, etc. Los datos no se encuentran estandarizados conteniendo formatos errados e información que debe ser procesada para su limpieza.

Preparación de los Datos

La preparación de los datos se da en las fases de recolección y exploración, al implementar el análisis sobre las variables seleccionadas, generando Dataframes enfocados en los tipos de datos correspondientes. Realizando limpieza de Datos con el tratamiento de valores faltantes, eliminación de datos duplicados y estandarizando valores atípicos, adicional se realizará la clasificación de variables categóricas, normalización de variables numéricas, e ingeniería de características.

Modelado

En el modelado se hará la implementación del algoritmo de machine learning necesario para cumplir con los objetivos establecidos en la fase de comprensión del negocio.

1. Selección de Modelo**:** Basados en la evaluación del conjunto de datos y la compresión del negocio el algoritmo de machine learning adecuado para el problema es la

regresión logística. Esto porque dicho modelo genera coeficientes fácilmente interpretables y los cuales ayudan a identificar factores influyentes en la determinación del riesgo crediticio. Adicional este algoritmo genera un eficiente manejo de variables categóricas disponiendo de una capacidad de proporcionar estimaciones más eficaces.

2. Entrenamiento del Modelo**:** Se hará la división de los datos en conjuntos de entrenamiento y prueba (80% - 20%) dando así un ajuste del modelo a los datos de entrenamiento.

3. Validación del Modelo**:** Con en Análisis y la ejecución del modelo según datos de entrenamiento y prueba se hace la evaluación del rendimiento del modelo utilizando métricas como el Accuracy o precisión, el recall y F1-score, entre otros.

Evaluación

Para dar cumplimiento a los objetivos y necesidades identificadas en la fase de conocimiento del negocio según el algoritmo empleado, se determina si el modelo aplicado genera un score de crédito más estandariza permitiendo inclusión financiera a todos los sectores del país.

Despliegue

Con los resultados del modelo pueden surgir recomendaciones de aplicación de los resultados, o de rendimiento que requiera de actualizaciones periódicas Buenaño y Fernández (2016).

Información Disponible y Compresión de Datos

Información Disponible

El conjunto de datos utilizado se denomina Datos Crediticios Alemanes y fue obtenido de una fuente de datos público de riesgo crediticio alemán actualizado en el año 2016, este análisis exploratorio de datos se centra en 1000 registros contenidos cada uno de ellos en 20 variables con características y atributos de importancia financiera que permiten la implementación de nuestro modelo, teniendo en cuenta que cada registro representa una persona que solicitó un crédito en un banco alemán. A continuación, se listan las variables disponibles para la realización del análisis predictivo del score de riesgo crediticio mediante machine learning, así:

Tabla 2

Variables Utilizadas como Insumos del Modelo

Variable	Descripción	Valores únicos
Credibilidad	Clasifica a las personas en función de su riesgo crediticio	2
Saldo de Cuenta	Monto disponible en la cuenta como capacidad del cliente para pagar un crédito	4
Duración del Credito Mensual	Cantidad de tiempo acordado para el pago del crédito	33
Estado de pago del Credito Anterior	Muestra cómo el cliente ha manejado sus deudas previas	5

Finalidad	Razón o propósito específico por el cual el cliente solicita un crédito	10
Importe Credito	Monto total que el cliente está solicitando como préstamo o crédito	923
Valor de ahorro	Nivel específico de ahorro	5
Duracion del Empleo Actual	Diferentes niveles de duración en su empleo actual	5
Porcentaje_por_cuota	Porcentaje del monto total del crédito que se debe pagar como cuota mensual	4
Sexo Estado Marital	Combinación del género y el estado civil del cliente	4
Garantes	Codeudores o avalistas en el contexto de un crédito	3
Duracion_en_la_direccion_actual	Tiempo que el cliente ha estado viviendo en su residencia actual	4
Activo_disponible_mas_valioso	Activo más valioso que el cliente tiene disponible como garantía para un crédito	4
Años de edad	Rango de las edades exactas de los clientes	53

Creditos_concurrentes	Cantidad de créditos adicionales que el cliente tiene en otras instituciones financieras	3
Tipo de apartamento	Categoría de vivienda en la que reside el cliente	3
No_creditos_en_este_banco	Cantidad de créditos que el cliente ha solicitado anteriormente en el mismo banco	4
Ocupación	Tipo de empleo o sector laboral del solicitante	4
No de dependientes	Cantidad de personas que dependen económicamente del solicitante	2
Teléfono	El solicitante tiene un teléfono registrado o no	2
Trabajor_extranjero	El solicitante es extranjero o local	2

Nota. Esta tabla contiene las variables disponibles con la información financiera para la implementación de nuestro modelo de machine learning. Tomado de Elaboración Propia (2024).

Una vez descrita la información disponible, se procede a explorar los datos para determinar nuestra variable objetivo que en este estudio es la Credibilidad, la cual se utiliza para predecir o clasificar el riesgo de crédito del cliente, esta variable se determina a partir de diversas características y atributos de otras columnas del conjunto de datos, tales como el saldo de cuenta, la duración del crédito, el estado de pago anterior, finalidad del crédito y el importe solicitado,

entre otros, es de tener en cuenta que la credibilidad refleja la capacidad del solicitante para cumplir con sus obligaciones crediticias, lo que la convierte en la variable clave que se desea predecir o clasificar en este modelo, en este contexto, las variables predictores que incluyen columnas como saldo en la cuenta, la duración mensual del crédito, y el estado de pago del crédito anterior, entre otras, son fundamentales para modelar y prever la variable objetivo.

Preprocesamiento de los Datos

Con el propósito de llevar a cabo un análisis de datos predictivo del score de riesgo crediticio, en primer lugar, se importan un conjunto de herramientas de software requeridas para facilitar la implementación de este modelo, incorporando librerías esenciales como Pandas para la manipulación de datos, Numpy para operaciones matriciales y estadísticas, Matplotlid para gráficos y visualización básica, Seaborn para visualización estadística simplificada, Sklearn para modelado y aprendizaje automático, Scipy para cálculos científicos y técnicos, Pycaret para automatización de modelos, Joblib para guardar y cargar modelos y por ultimo Random para realizar operaciones con aleatoriedad; por lo cual al ejecutar los primeros bloques de código con 'Helper' que encapsula varias funciones para el análisis preliminar del *dataset* y 'load', que permite obtener información clave sobre el conjunto de datos importados como se visualiza en la tabla No. 3 que se relaciona a continuación, donde obtenemos un resumen detallado que incluye los siguientes parámetros, así:

Tabla 3

Visualización de los Datos Importados para facilitar su Compresión

	count	mean	std	min	25%	50%	75%	max
Saldo de cuenta	1000.0	2.577	1.257.638	1.0	1.0	2.0	4.00	4.0

Duración del Crédito_mensual	1000.0	20.903	12.058.814	4.0	12.0	18.0	24.00	72.0
Estado de pago del crédito_anterior	1000.0	2.545	1.083.120	0.0	2.0	2.0	4.00	4.0
Finalidad	1000.0	2.828	2.744.439	0.0	1.0	2.0	3.00	10.0
Importe_Crédito	1000.0	3.271.248	2.822.751.760	250.0	1365.5	2319.5	3972.25	18424.0
Valor de ahorro	1000.0	2.105	1.580.023	1.0	1.0	1.0	3.00	5.0
Duración del Empleo Actual	1000.0	3.384	1.208.306	1.0	3.0	3.0	5.00	5.0
porcentaje_por_cuota	1000.0	2.973	1.118.715	1.0	2.0	3.0	4.00	4.0
Sexo Estado Marital	1000.0	2.682	0.708080	1.0	2.0	3.0	3.00	4.0
Garantes	1000.0	1.145	0.477706	1.0	1.0	1.0	1.00	3.0
Duración_en_la_dirección_actual	1000.0	2.845	1.103.718	1.0	2.0	3.0	4.00	4.0
Activo_disponible_más_valioso	1000.0	2.358	1.050.209	1.0	1.0	2.0	3.00	4.0
Años de edad	1000.0	35.542	11.352.670	19.0	27.0	33.0	42.00	75.0
Créditos_concurrentes	1000.0	2.675	0.705601	1.0	3.0	3.0	3.00	3.0

Tipo de apartamento	1000.0	1.928	0.530186	1.0	2.0	2.0	2.00	3.0
No de créditos_en_este_Banco	1000.0	1.407	0.577654	1.0	1.0	1.0	2.00	4.0
Ocupación	1000.0	2.904	0.653614	1.0	3.0	3.0	3.00	4.0
No de dependientes	1000.0	1.155	0.362086	1.0	1.0	1.0	1.00	2.0
Teléfono	1000.0	1.404	0.490943	1.0	1.0	1.0	2.00	2.0
Trabajador_extranjero	1000.0	1.037	0.188856	1.0	1.0	1.0	1.00	2.0
Credibilidad	1000.0	0.700	0.458487	0.0	0.0	1.0	1.00	1.0

Nota. Esta tabla contiene información clave sobre el dataset, facilitando su comprensión antes de aplicar modelos de aprendizaje automático. Tomado de Elaboración Propia (2024).

1. Dimensiones del conjunto de datos: muestra el número de observaciones que son 1000 y las columnas que son 21 presentes en este conjunto de datos.

2. Tipos de variables: identifica el tipo de datos que contienen las columnas que en este caso son variables con datos enteros.

3. Uso de memoria: imprime un informe sobre el consumo de memoria del conjunto de datos que estamos trabajando que es de 164.2 KB, lo cual nos ayuda a entender su eficiencia y el impacto en el sistema, analizando que no vamos a requerir recursos adicionales para el procesamiento.

4. Estadística descriptiva: muestra un resumen de las estadísticas descriptivas para las columnas numéricas, incluyendo medidas de tendencia central, dispersión, asimetría y curtosis, que permiten evaluar la distribución de los datos, algunas interpretaciones:

La media en este caso, el valor promedio del Importe_Crédito 3271.25 indica que, en general, los clientes suelen solicitar créditos cercanos a este monto.

La mediana del Importe_Crédito 2319.5 es menor que la media, lo que sugiere que la mayoría de los valores están por debajo del promedio, pero existen algunos créditos muy altos que incrementan la media.

La count de 1000 indica que no hay valores faltantes en la variable Credibilidad, lo que es ideal para realizar la implementación de nuestro modelo sin necesidad de imputaciones o ajustes.

La media de Credibilidad que es de 0.7 muestra que, en promedio, el 70% de los registros tiene una credibilidad alta o aceptable (siendo 1 el indicador de credibilidad positiva). Esto sugiere que la mayoría de los solicitantes en el dataset tiene un buen perfil de riesgo.

La std de Credibilidad que es de 0.458 indica que existe una variación moderada en los valores de esta variable.

El máximo de 1 confirma que el mayor nivel de Credibilidad asignado es el valor ideal o positivo, consistente con la definición binaria de la variable; 0 = no confiable, 1 = confiable.

5. Duplicados: se evidencia que no hay filas duplicadas en el *dataset*, lo que es crucial para garantizar la calidad de los datos.

6. Valores faltantes: detecta que no hay ningún valor faltante en nuestro conjunto de datos, lo cual es esencial para la limpieza de datos antes de realizar un análisis predictivo.

De acuerdo con lo descrito y al visualizar las métricas de la variable objeto Credibilidad, se puede mencionar que está bien definida y es homogénea, con valores mayoritariamente concentrados en torno a 0.7, permitiéndonos analizar en los datos que gran parte de los individuos tiene un perfil de riesgo aceptable o confiable, con una variabilidad moderada que podría ser analizada más a fondo para identificar patrones específicos en grupos particulares.

Siguiendo con la etapa del análisis de datos y la construcción de nuestro modelo de machine learning se implementó la función 'missing_variable_summary' para evaluar la calidad de los datos, la magnitud de los mismos y diagnosticar valores faltantes en este conjunto de datos, que como podemos observar en la siguiente tabla No. 4, en nuestro dataset hay frecuencias completas en de cada variable y no se evidencia valores faltantes, permitiendo avanzar en el análisis sin tener preocupación de ausencia de datos en el desarrollo del modelo.

Tabla 4

Visualización de la Calidad de los Datos

	Variable	n_missing	n_cases	pct_missing
0	Saldo de cuenta	0	1000	0.0
1	Duración del Crédito_mensual	0	1000	0.0
2	Estado de pago del crédito_anterior	0	1000	0.0
3	Finalidad	0	1000	0.0
4	Importe_Crédito	0	1000	0.0
5	Valor de ahorro	0	1000	0.0
6	Duración del Empleo Actual	0	1000	0.0
7	porcentaje_por_cuota	0	1000	0.0
8	Sexo Estado Marital	0	1000	0.0
9	Garantes	0	1000	0.0
10	Duración_en_la_dirección_actual	0	1000	0.0
11	Activo_disponible_más_valioso	0	1000	0.0

12	Años	0	1000	0.0
13	Créditos_concurrentes	0	1000	0.0
14	Tipo de apartamento	0	1000	0.0
15	No de créditos_en_este_Banco	0	1000	0.0
16	Ocupación	0	1000	0.0
17	No de dependientes	0	1000	0.0
18	Teléfono	0	1000	0.0
19	Trabajador_extranjero	0	1000	0.0
20	Credibilidad	0	1000	0.0

Nota. Esta tabla contiene información completa del dataset, que presenta un buen punto de partida para el desarrollo del modelo. Tomado de Elaboración Propia (2024).

Seguidamente al clasificar todas las variables del dataset como **cuantitativas**, se puede interpretar que los datos son numéricos y representan magnitudes continuas o discretas que pueden ser medidas y analizadas matemáticamente, así mismo para garantizar la interpretabilidad y efectividad del modelo de machine learning en el análisis del score de riesgo crediticio, se llevó a cabo la transformación de algunas variables cuantitativas a cualitativas las cuales pudieran influir en la variable objeto Credibilidad, cuyo proceso consiste en reclasificar valores numéricos en categorías que representen características específicas, por ejemplo, la variable "Saldo de cuenta", inicialmente cuantitativa, fue transformada en categorías como 'Sin cuenta', 'Ninguna (sin saldo)', 'Algún saldo' y 'Saldo superior', facilitando que el modelo interprete patrones relacionados con estas categorías de manera más efectiva y aporta contexto adicional al análisis. Este tipo de manejo es especialmente útil en datasets donde las magnitudes absolutas tienen menor relevancia que las clasificaciones que describen el comportamiento de los individuos, proceso que se demuestra en la siguiente tabla No. 5, que fue aplicado a las variables que contienen información relevante la cual pueden ayudar a determinar el nivel de riesgo

crediticio, ya que su inclusión permite que el modelo identifique patrones y relaciones entre estas características y la probabilidad de cumplimiento crediticio, mejorando la precisión de las predicciones.

Tabla 5

Visualización de Variable Cuantitativas a Variables Cualitativas

	Saldo de cuenta	Duración del Crédito_mensual	Estado de pago del crédito_anterior	Finalidad	Importe_Crédito	Valor de ahorro	Duración del Empleo Actual	porcentaje_por_cuota
0	Sin cuenta	18	Créditos anteriores pagados	Carro Usado	1049	Ninguno	<1 Año	abajo 20%
1	Sin cuenta	9	Créditos anteriores pagados	Otro	2799	Ninguno	[1, 4)	(25%, 35%)
2	Ninguna (sin saldo)	12	Pagado	Reentrenamiento	841	Menos de 100 DM	[4, 7)	(25%, 35%)
3	Sin cuenta	12	Créditos anteriores pagados	Otro	2122	Ninguno	[1, 4)	[20%, 25%)
4	Sin cuenta	12	Créditos anteriores pagados	Otro	2171	Ninguno	[1, 4)	abajo 20%

Sexo Estado Marital	Garantes	Duración_en_la_dirección_actual	Activo_disponible_más_valioso	Años de edad	Créditos_concurrentes
Hombre, soltero	Ninguno	por encima de 7	Carro	21	Ninguno
Hombre, Casado/Viudo	Ninguno	[1, 4)	Ninguno	36	Ninguno
Hombre, soltero	Ninguno	por encima de 7	Ninguno	23	Ninguno
Hombre, Casado/Viudo	Ninguno	[1, 4)	Ninguno	39	Ninguno
Hombre, Casado/Viudo	Ninguno	por encima de 7	Carro	38	Otros bancos

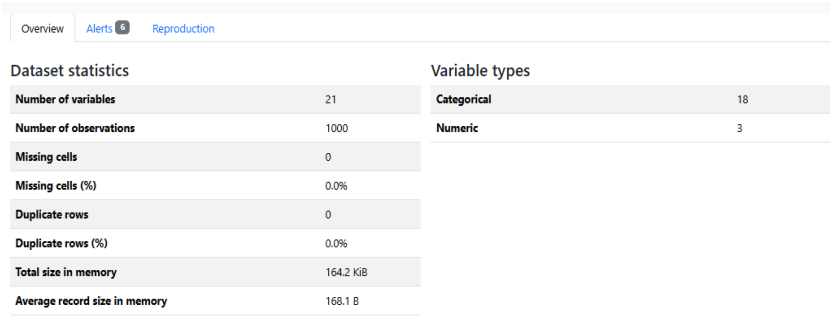
Tipo de apartamento	No de créditos_en_este_Banco	Ocupación	No de dependientes	Teléfono	Trabajador_extranjero	Credibilidad
Libre	1	Expertos	3 o mas	No	No	1
Libre	2 or 3	Expertos	menos de 3	No	No	1
Libre	1	Residente permanente no cualificado	3 o mas	No	No	1
Libre	2 or 3	Residente permanente no cualificado	menos de 3	No	Si	1
Rentado	2 or 3	Residente permanente no cualificado	3 o mas	No	Si	1

Nota. Esta tabla contiene información de la transformación de variables, reclasificando sus valores numéricos en categóricos con características específicas. Tomado de Elaboración Propia (2024).

Continuando con el análisis extendido del conjunto de datos, utilizamos la función 'ProfileReport()' de pandas-profiling, que permite generar automáticamente un informe detallado sobre este dataset, ofreciendo un análisis exploratorio de datos (EDA) eficiente y visualmente intuitivo que incluye un resumen general del dataset, análisis individual de las variables, correlaciones, alertas y calidad de datos como se evidencia en las siguientes graficas:

Figura 1

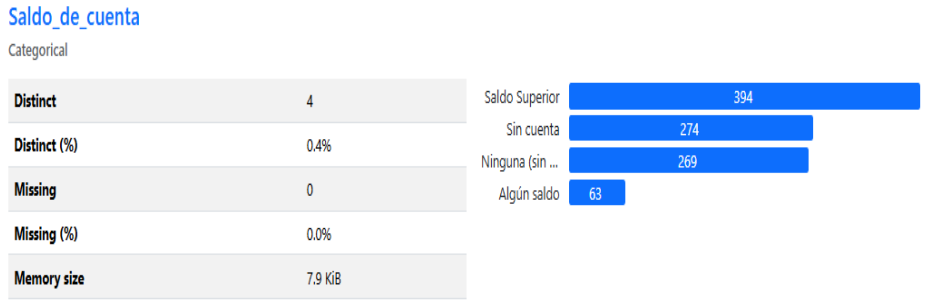
Visión General del Dataset y Tipos de Variables que Contiene



El conjunto de datos mantiene su estructura original con 21 variables (18 categóricas y 3 numéricas) y 1000 observaciones, no presenta datos faltantes, lo que asegura integridad en la información, además, el consumo de memoria se mantiene en 164.2 KB, indicando eficiencia en su manejo, esta información proporciona una base sólida para el análisis y modelado.

Figura 2

Análisis de la Variable Saldo de Cuenta



La variable Saldo de cuenta es categórica y presenta 4 categorías distintas, representando una distribución relativamente equitativa de las observaciones en su conjunto, no hay datos faltantes, lo que indica que la variable está completamente disponible para el análisis.

Las categorías con mayor frecuencia son "Saldo Superior" (394 observaciones) y "Sin cuenta" (274 observaciones), mientras que las categorías "Ninguna (sin saldo)" (269 observaciones) y "Algún saldo" (63 observaciones) son menos frecuentes.

La distribución muestra que la mayoría de las personas en el conjunto de datos tienen un saldo de cuenta, y una menor proporción no tiene saldo o cuenta en absoluto.

Esta información es útil para el análisis de factores relacionados con el riesgo crediticio, ya que el estado de la cuenta de un individuo podría influir en su capacidad para acceder al crédito.

Figura 3

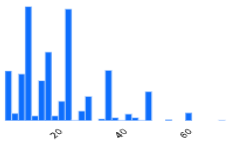
Análisis de la Variable Duración del Crédito Mensual

Duración_del_Crédito_mensual

Real number (ℝ)

High correlation

Distinct	33	**Minimum**	4
Distinct (%)	3.3%	**Maximum**	72
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	20.903	**Memory size**	7.9 KiB

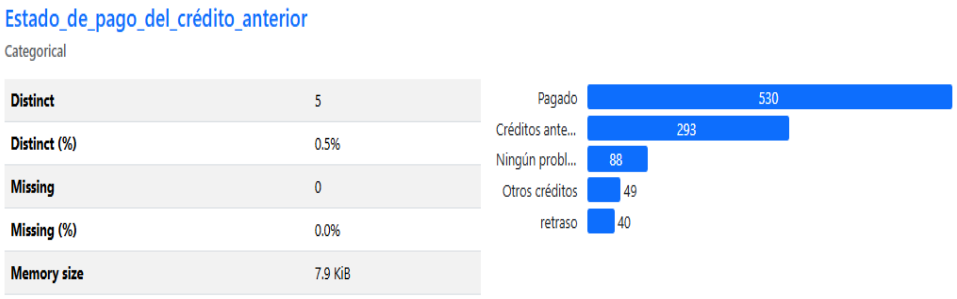


La variable Duración del Crédito mensual es numérica con una media de 20.903 meses, con valores que van desde un mínimo de 4 meses hasta un máximo de 72 meses, la distribución muestra 33 valores distintos, lo que indica cierta diversidad en la duración de los créditos, además, no presenta valores faltantes ni valores infinitos, lo cual es positivo para el análisis, ya que no se requiere tratamiento de datos faltantes, esta variable tampoco presenta valores negativos ni ceros, lo que también es relevante, ya que sugiere que los valores son consistentes con el tipo de datos esperados, así mismo se puede observar que los plazos más comunes para la concesión de créditos oscilan entre 20, 40 y 60 meses, sin embargo, el plazo de 20 meses destaca

como el preferido por los clientes para el pago de sus créditos, esta tendencia sugiere que los prestatarios optan por un período intermedio que equilibre la accesibilidad en las cuotas mensuales con la gestión eficiente del crédito.

Figura 4

Análisis de la Variable Estado de Pago del Crédito Anterior

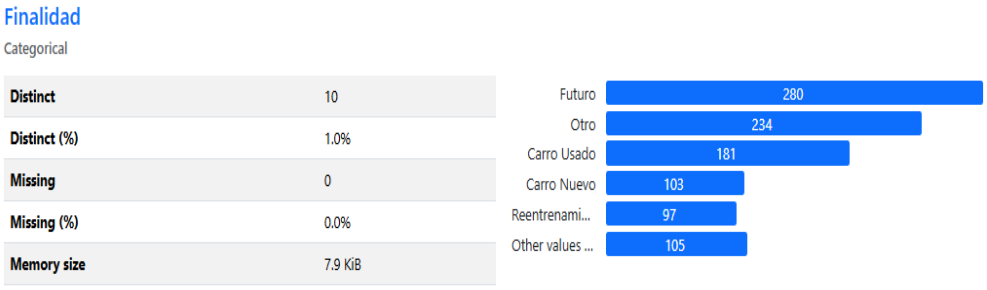


La variable Estado de pago del crédito anterior revela que la mayoría de los casos tienen una clasificación favorable, por lo cual, de los 1000 registros analizados, la mayor parte de los clientes tiene antecedentes de crédito pagado (530 casos), mientras que 293 han tenido créditos anteriores pagados sin inconvenientes, solo un pequeño porcentaje enfrenta situaciones más complejas, como problemas con créditos actuales o retrasos en pagos (40 casos), además, se observa que la variable tiene cinco categorías distintas, pero ninguna muestra valores faltantes, lo que indica que la información está completa y es relevante para el análisis.

Esta distribución sugiere que la mayoría de los clientes tienen un historial de crédito positivo, lo que podría correlacionarse con una mayor probabilidad de ser considerados confiables en el modelo de riesgo crediticio, sin embargo, las categorías con problemas de pago, aunque pocas, podrían ser indicativas de un mayor riesgo de impago, lo que las hace relevantes para la predicción de la credibilidad en el contexto de evaluación del riesgo crediticio.

Figura 5

Análisis de la Variable Finalidad



La variable Finalidad muestra que existen 10 categorías distintas, lo que indica una diversidad de objetivos para los créditos solicitados, por lo cual, de los 1000 registros, las finalidades más frecuentes son para "Futuro" (280 casos), "Otro" (234 casos), y "Carro Usado" (181 casos), así mismo otras finalidades, como la compra de "Carro Nuevo" (103 casos) y "Reentrenamiento" (97 casos), son menos comunes, es interesante mencionar que 105 registros se encuentran en una categoría de "Otros valores", lo que puede representar necesidades o razones menos definidas para la solicitud de crédito.

La variable no tiene valores faltantes, lo que la hace completa y útil para el análisis. La distribución sugiere que los clientes tienden a solicitar créditos para propósitos comunes, como la compra de vehículos o inversiones futuras, mientras que los casos más atípicos (como "reentrenamiento" o "otros") podrían reflejar objetivos específicos o menos convencionales.

Figura 6

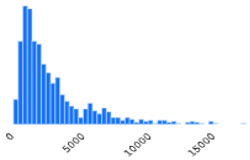
Análisis de la Variable Importe Crédito

Importe_Crédito

Real number (ℝ)

High correlation

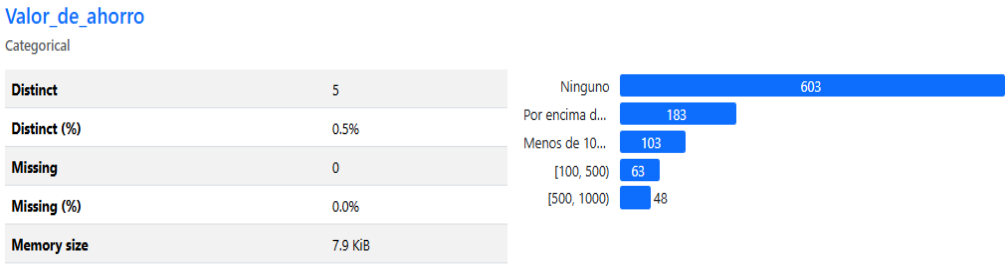
Distinct	923	**Minimum**	250
Distinct (%)	92.3%	**Maximum**	18424
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	3271.248	**Memory size**	7.9 KiB



La variable Importe_Crédito muestra que es numérica con una alta correlación, lo que indica que puede estar relacionada estrechamente con otras variables en el modelo, así mismo tiene 923 valores distintos, lo que refleja una gran variabilidad en los importes solicitados, se logra analizar que el valor mínimo registrado es de 250, mientras que el máximo es 18424, lo que sugiere que los créditos pueden variar significativamente en su monto, la media es de 3271.248, lo que indica que el promedio de los créditos solicitados está dentro de un rango moderado, por ultimo podemos detallar en la gráfica que los importes solicitados aumentan desde 0 hasta 5000, lo que indica que la mayoría de los clientes tienden a solicitar montos bajos y a medida que los montos solicitados se acercan a 15000, se presenta una disminución en la frecuencia de solicitudes, lo que sugiere que menos clientes optan por solicitar créditos de mayor valor.

Figura 7

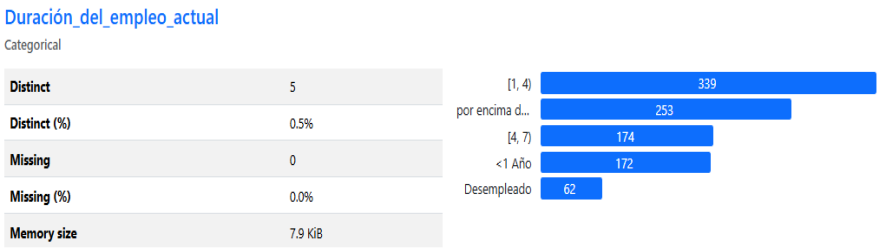
Análisis de la Variable Valor de Ahorro



La variable Valor de ahorro muestra que la mayoría de los clientes no tienen ahorros (603 casos, representando el 60.3% del total), lo cual es relevante para entender el perfil financiero de los solicitantes del crédito, las categorías con valores de ahorro más altos, como "Por encima de 1000" (183 casos) y "Menos de 100 DM" (103 casos), representan una menor proporción de la muestra, esto sugiere que un segmento considerable de clientes tiene ahorros limitados o nulos, lo que podría influir en la capacidad de pago y, por lo tanto, en el riesgo crediticio, es de tener en cuenta que la baja frecuencia en los rangos de ahorro superiores indica que los ahorros no son un factor significativo para la mayoría de los solicitantes.

Figura 8

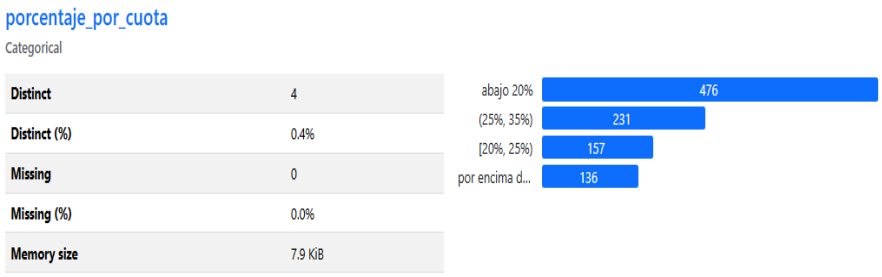
Análisis de la Variable Duración del Empleo Actual



La variable Duración del Empleo Actual muestra que la mayoría de los registros se distribuyen en cinco categorías distintivas, la categoría con más registros es la de personas con entre 1 y 4 años de empleo (339 casos), seguida por aquellos con más de 7 años de empleo (253 casos), otros rangos de duración incluyen entre 4 y 7 años (174 casos), menos de 1 año (172 casos) y personas desempleadas (62 casos), esta distribución sugiere que la mayoría de los clientes tienen una estabilidad laboral moderada a alta, lo cual puede ser relevante para el análisis de riesgo crediticio, ya que una mayor estabilidad laboral podría estar asociada a una mayor capacidad de pago.

Figura 9

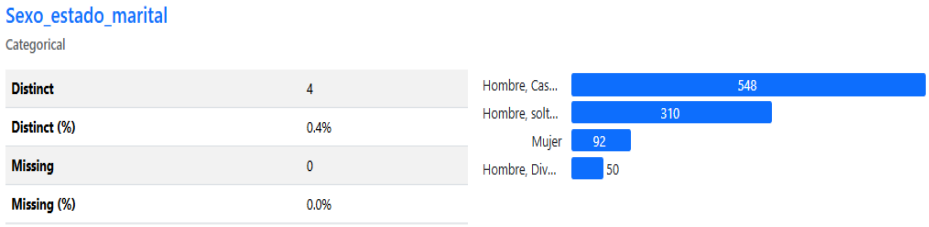
Análisis de la Variable Porcentaje por Cuota



La variable porcentaje_por_cuota muestra que la mayoría de los registros tienen un porcentaje de cuota de pago inferior al 20% (476 casos), lo que sugiere que una gran parte de los clientes mantiene una cuota de pago baja en relación con el crédito, así mismo un número considerable de clientes se encuentra en el rango de 25% a 35% (231 casos), mientras que una menor cantidad de clientes se ubica en los rangos de 20% a 25% (157 casos) y por encima de 35% (136 casos), estos datos pueden indicar que la mayoría de los solicitantes prefieren mantener una cuota relativamente baja, lo cual podría tener implicaciones en la capacidad de pago y en la gestión del riesgo crediticio.

Figura 10

Análisis de la Variable Sexo Estado Marital



La variable Sexo Estado Marital muestra que la mayoría de los registros corresponden a hombres casados o viudos (548 casos), seguidos de hombres solteros (310 casos), un número menor de mujeres está representado en el dataset (92 casos), y los hombres divorciados son los menos numerosos (50 casos), estos resultados sugieren una distribución desigual entre géneros y estados civiles, con una mayor presencia de hombres casados o viudos, lo que podría influir en los patrones de riesgo y en las decisiones de crédito.

Figura 11

Análisis de la Variable Garantes

La variable Garantes muestra que la mayoría de los registros en el dataset no cuentan con un garante, con un 90.7% de observaciones sin garante (907 casos), solo una pequeña fracción de los datos tiene un garante (52 casos) o un codeudor (41 casos), por lo cual esta distribución desbalanceada de esta variable puede indicar que, en la mayoría de los casos, los solicitantes de

crédito no requieren un garante, lo que podría influir en el perfil de riesgo asociado a la variable, esta información es importante, ya que el tipo de garantía o codeudor podría afectar la evaluación del riesgo crediticio.

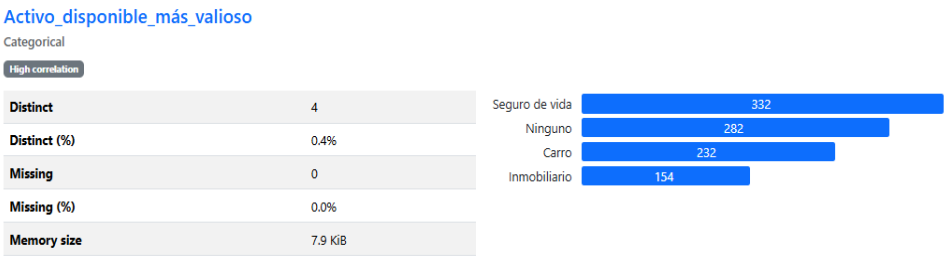
Figura 12

Análisis de la Variable Duración en la Dirección Actual

La variable Duración_en_la_dirección_actual refleja el tiempo que los individuos han permanecido en su residencia actual, estos resultados muestran que la mayoría ha vivido por más de 7 años en su dirección actual (413 casos, 41.3%), seguida de quienes han residido entre 1 y 4 años (308 casos, 30.8%), aquellos con una permanencia de entre 4 y 7 años representan el 14.9% (149 casos), y los residentes de menos de un año constituyen el grupo más pequeño, con 13% (130 casos), por lo cual esta distribución indica estabilidad residencial en la mayoría de los casos, lo que podría asociarse a una menor percepción de riesgo en el análisis crediticio.

Figura 13

Análisis de la Variable Activo Disponible Mas Valioso



La variable Activo_disponible_más_valioso identifica los principales activos de los clientes, cuyos resultados muestran que el seguro de vida es el activo más común, representando el 33.2% de los casos (332 observaciones), seguido por la categoría "Ninguno" con 28.2% (282 casos), lo que podría indicar una ausencia de bienes valiosos declarados, el carro es el tercer activo más frecuente (23.2%, 232 casos), y los bienes inmobiliarios son los menos comunes (15.4%, 154 casos), la alta correlación de esta variable sugiere que el tipo de activo declarado puede tener una influencia importante en la evaluación de riesgo crediticio, al asociarse con la capacidad financiera del cliente.

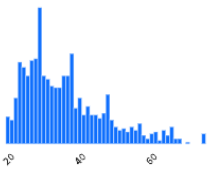
Figura 14

Análisis de la Variable Años

Años_de_edad

Real number (ℝ)

Distinct	53	**Minimum**	19
Distinct (%)	5.3%	**Maximum**	75
Missing	0	**Zeros**	0
Missing (%)	0.0%	**Zeros (%)**	0.0%
Infinite	0	**Negative**	0
Infinite (%)	0.0%	**Negative (%)**	0.0%
Mean	35.542	**Memory size**	7.9 KiB

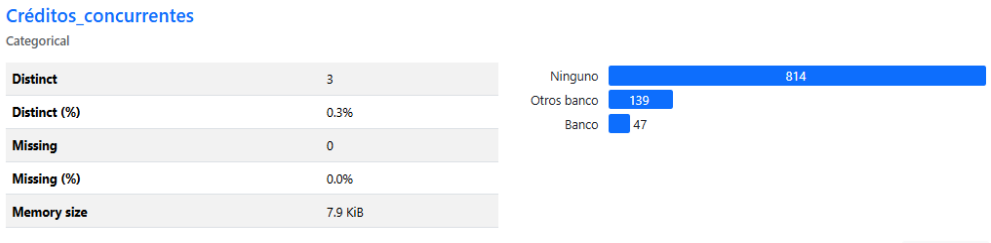


La variable Años de edad muestra un rango de edades de los clientes entre los 19 y los 75 años, con una media de 35.54 años, esto indica que la mayoría de los clientes se encuentran en una etapa laboral activa, lo cual es relevante para evaluar su capacidad de pago y estabilidad financiera, así mismo se puede evidenciar que no hay datos faltantes ni valores fuera de rango (como negativos o ceros), lo que garantiza la calidad de esta variable para su análisis, por último se observa en la gráfica que los picos más altos se concentran entre 20 y 40 años, lo que refleja

que la mayoría de los clientes pertenece a este rango, siendo menos frecuentes los valores superiores a 60 años.

Figura 15

Análisis de la Variable Créditos Concurrentes



La variable Créditos_concurrentes muestra que la gran mayoría de los clientes (814 casos, 81.4%) no tienen créditos concurrentes en otros lugares, una proporción menor, pero significativa (139 casos, 13.9%), tiene créditos en otros bancos, y solo un pequeño porcentaje (47 casos, 4.7%) tiene créditos concurrentes en el mismo banco, por lo cual este análisis refleja que la mayoría de los solicitantes no están sobrecargados con múltiples obligaciones financieras, lo que podría indicar un menor riesgo crediticio, aunque es necesario evaluar su impacto en el modelo.

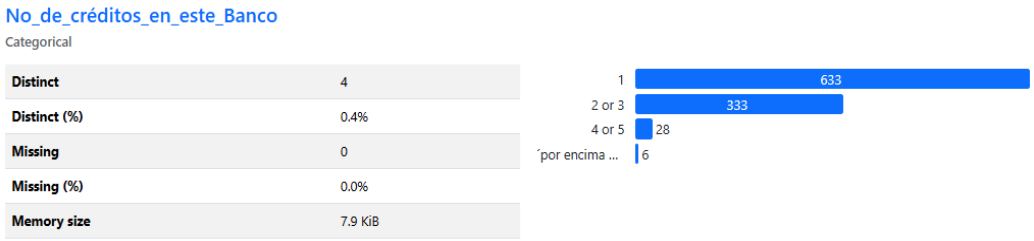
Figura 16

Análisis de la Variable Años

La variable Tipo de apartamento refleja que la mayoría de los clientes (714 casos, 71.4%) vive en propiedades rentadas, mientras que una menor proporción vive en viviendas libres de costo (179 casos, 17.9%) o en propiedades propias (107 casos, 10.7%), la alta correlación asociada con esta variable sugiere que las condiciones de vivienda podrían estar relacionadas significativamente con otros factores en el modelo de riesgo crediticio, posiblemente reflejando estabilidad económica o financiera.

Figura 17

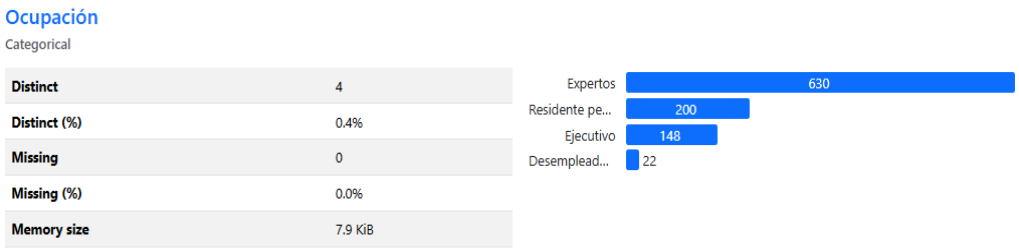
Análisis de la Variable No. De Créditos en este Banco



La variable No de créditos en este banco muestra que la mayoría de los clientes (633 casos, 63.3%) tienen un solo crédito en este banco, mientras que una proporción considerable (333 casos, 33.3%) tiene entre dos y tres créditos y solo una pequeña fracción de los clientes posee entre cuatro y cinco créditos (28 casos, 2.8%) o más de seis créditos (6 casos, 0.6%), esto sugiere que la mayoría de los clientes tienen una relación relativamente sencilla con el banco, con pocos clientes acumulando múltiples créditos, por lo cual este comportamiento podría indicar un perfil de clientes con una menor exposición financiera en este banco en particular.

Figura 18

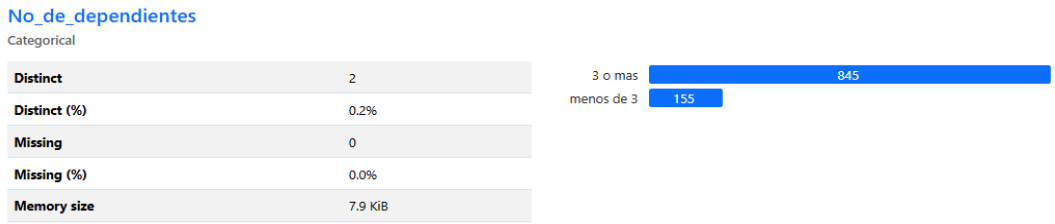
Análisis de la Variable Ocupación



La variable Ocupación revela que la mayoría de los clientes (630 casos, 63%) están clasificados como "Expertos", lo que sugiere que estos clientes tienen un nivel profesional especializado, le sigue un grupo considerable de clientes que son "Residentes permanentes no cualificados" (200 casos, 20%), lo que podría indicar personas con una permanencia más estable en el país, pero con cualificaciones laborales limitadas, en menor medida, se encuentran "Ejecutivos" (148 casos, 14.8%) y un pequeño número de "Desempleados no cualificados" (22 casos, 2.2%), en conclusión esta métrica refleja una diversidad ocupacional, con una alta concentración de personas con una ocupación estable y calificada, lo que podría influir en la capacidad de pago y en el perfil de riesgo crediticio.

Figura 19

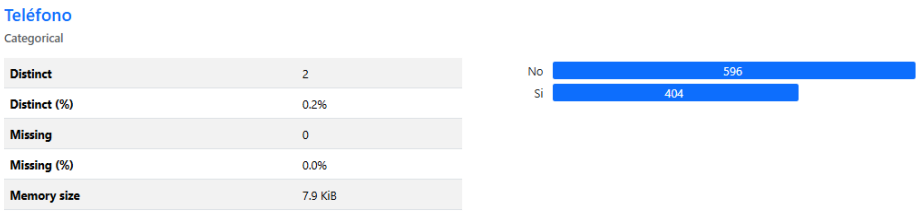
Análisis de la Variable No. De Dependientes



La variable No de dependientes muestra que la mayoría de los clientes (845 casos, 84.5%) tienen tres o más dependientes, lo que podría indicar una mayor carga financiera familiar, por otro lado, un grupo más pequeño (155 casos, 15.5%) tiene menos de tres dependientes, estos resultados sugieren que la mayoría de los clientes con un número elevado de dependientes pueden enfrentar desafíos adicionales en su capacidad de pago, lo que podría influir en su riesgo crediticio, mientras que aquellos con menos dependientes podrían tener menos cargas familiares.

Figura 20

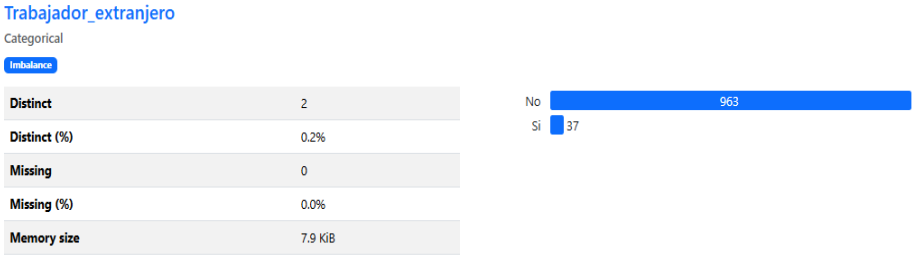
Análisis de la Variable Teléfono



La variable Teléfono muestra que la mayoría de los clientes (596 casos, 59.6%) no tienen teléfono registrado, mientras que un grupo menor (404 casos, 40.4%) sí lo tiene, esto puede indicar que los clientes sin teléfono podrían representar un segmento más difícil de contactar o verificar, lo que podría afectar la evaluación de su confiabilidad crediticia, ya que la falta de teléfono podría estar asociada con una mayor falta de accesibilidad o estabilidad, lo que podría aumentar el riesgo para los prestamistas.

Figura 21

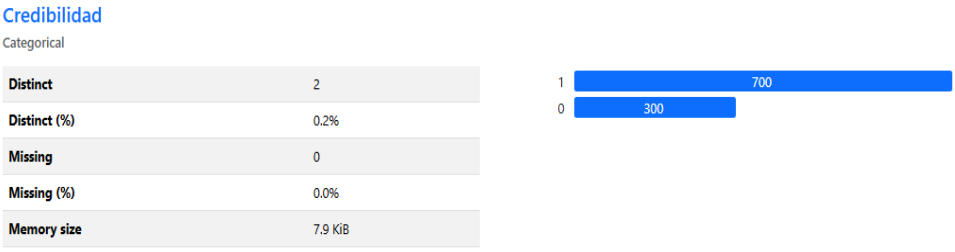
Análisis de la Variable Trabajador Extranjero



La variable Trabajador_extranjero muestra un desequilibrio en los datos, ya que la mayoría de los registros (963 casos, 96.3%) corresponden a personas que no son trabajadoras extranjeras, mientras que un pequeño grupo (37 casos, 3.7%) sí lo son, este desbalance puede influir en el modelo de riesgo crediticio, ya que una representación tan pequeña de trabajadores extranjeros podría afectar la capacidad del modelo para generalizar correctamente en este grupo. Además, este desequilibrio podría implicar que los trabajadores extranjeros tengan características o riesgos distintos a los de los demás, lo que es importante considerar en el análisis.

Figura 22

Análisis de la Variable Credibilidad



La variable Credibilidad tiene una distribución binaria con dos posibles valores 1 y 0, por lo cual, de los 1000 registros, 700 (70%) corresponden a un valor de "1", lo que podría indicar

que la mayoría de los solicitantes en este conjunto de datos tienen una credibilidad positiva, mientras que 300 (30%) tienen un valor de "0", lo que sugiere una credibilidad negativa, este desbalance en los datos, aunque no extremo, puede influir en la forma en que el modelo aprenda a predecir la credibilidad de los solicitantes, ya que el mayor número de casos con credibilidad positiva podría hacer que el modelo sea más propenso a clasificar correctamente a los individuos de ese grupo, siendo crucial abordar este desequilibrio durante la fase de modelado para evitar sesgos en las predicciones.

Conclusiones del Análisis de las Variables

En el contexto de un modelo de score de riesgo crediticio, este análisis a cada variable permitió entender los resultados obtenidos, donde se observan varias correlaciones y desequilibrios importantes entre las variables, por ejemplo, la alta correlación entre "Activo_disponible_más_valioso" y "Tipo de apartamento" sugiere que ciertos tipos de activos están fuertemente relacionados con el tipo de vivienda de los individuos, lo cual podría tener un impacto en su perfil crediticio.

De manera similar, la correlación entre "Duración del Crédito mensual" e "Importe_Crédito" también indica que los montos de crédito y los plazos de pago están asociados, lo que podría influir en el comportamiento de pago de los solicitantes.

Por otro lado, se identifican dos variables con desequilibrios significativos en su distribución, como "Garantes" y "Trabajador_extranjero", en el caso de "Garantes", el 66% de los casos corresponden a personas sin ningún garante, lo que podría reflejar una menor capacidad de respaldo en términos de solvencia, mientras tanto, la variable "Trabajador_extranjero" presenta un desequilibrio del 77.2%, donde la mayoría de los solicitantes no son trabajadores

extranjeros, lo que podría influir en la estabilidad y riesgo asociado a ciertos perfiles de solicitantes.

Por último, se puede mencionar que estos resultados son cruciales para el análisis del riesgo crediticio, ya que permiten identificar patrones de relación entre variables y posibles factores de riesgo o seguridad en el comportamiento crediticio de los solicitantes.

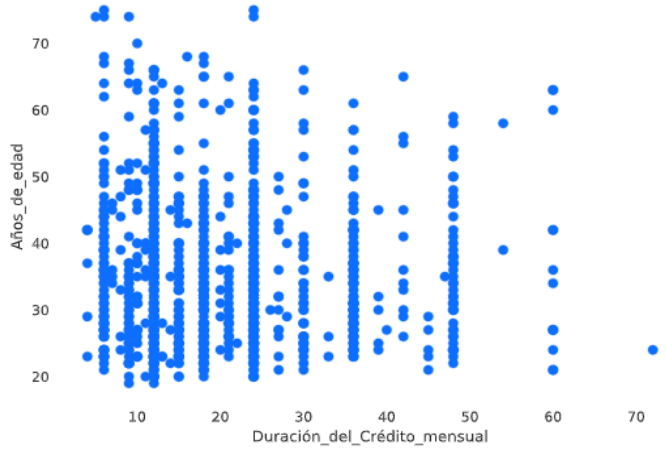
Análisis Gráfico de Interacciones, Correlaciones y Datos faltantes

Teniendo en cuenta el análisis realizado a cada una de las variables presentes en nuestro dataset procedemos a visualizar y explorar las relaciones entre las mismas, identificando patrones de interacción para evaluar su grado de correlación, dado que no se encontraron valores faltantes ni duplicados en el conjunto de datos, el análisis se centrará en comprender la estructura de los datos y su distribución a través de los siguientes gráficos representativos, ya que este enfoque permitirá interpretar de manera intuitiva las conexiones y dependencias entre las variables, sentando las bases para un modelado predictivo eficiente.

Interacciones

Figura 23

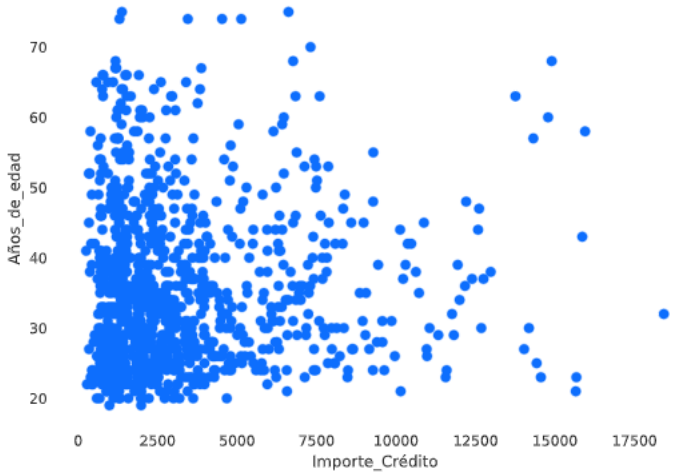
Análisis de las Variables Años y Duración del Crédito Mensual



El gráfico de dispersión indica que la mayoría de los créditos tienen una duración concentrada entre 10 y 40 meses, independientemente de la edad del cliente, con una ligera densidad en duraciones más cortas menos de 20 meses, Los clientes jóvenes, en el rango de 20 a 40 años, están más agrupados en créditos de menor duración, lo que podría estar relacionado con una menor capacidad de compromiso financiero o preferencia por liquidar el crédito rápidamente. Por otro lado, los clientes de mayor edad 40 años en adelante muestran una mayor dispersión en la duración del crédito, incluyendo valores más altos, como créditos de hasta 60 meses, lo que puede reflejar una mayor estabilidad o acceso a diferentes condiciones crediticias.

Figura 24

Análisis de las Variables Años e Importe Crédito



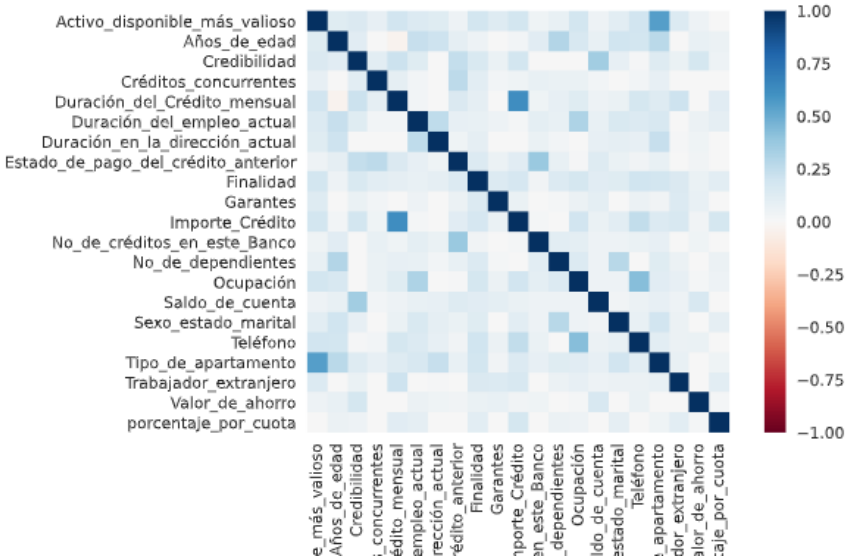
El gráfico de dispersión refleja la relación entre la edad de los clientes (eje Y) y el importe del crédito solicitado (eje X), la mayoría de los importes solicitados se concentran en valores inferiores a 10,000 unidades monetarias, mientras que las edades de los clientes se encuentran mayormente entre los 20 y 60 años, se observa una concentración significativa de créditos entre 0 y 2,500 unidades monetarias para clientes de entre 20 y 40 años, lo que sugiere

que estos grupos suelen optar por importes pequeños, posiblemente debido a limitaciones financieras o menor capacidad de endeudamiento, no se aprecia una tendencia clara entre las variables, ya que personas de distintas edades solicitan importes similares, los créditos superiores a 12,500 unidades monetarias son menos frecuentes y están distribuidos de forma dispersa, principalmente en clientes de entre 30 y 50 años, lo que indica una mayor diversidad en los importes más altos.

Correlaciones

Figura 25

Análisis y Evaluación de la Correlación de las Variables del Datset



Al interpretar esta matriz de correlación que está evaluando las diferentes variables que permitirán la construcción de nuestro modelo predictivo del score de riesgo crediticio, se logra evidenciar la relación que hay entre las diferentes variables representadas por valores que van de -1 (correlación negativa perfecta, rojo oscuro) a 1 (correlación positiva perfecta, azul oscuro), de lo cual se describe los análisis más relevantes que se observan, así:

Variables con Alta Correlación Positiva

“Tipo de apartamento" y "Activo disponible más valioso" (0.548), existe una relación positiva significativa, lo que sugiere que el tipo de apartamento está relacionado con el activo disponible más valioso, por lo cual quienes tienen apartamentos de mayor categoría también posean activos más valiosos.

"Importe Crédito" y "Duración del Crédito mensual" (0.625), esta fuerte correlación positiva indica que los créditos de mayor importe tienden a tener una duración más larga, esto tiene sentido ya que los montos más altos suelen requerir plazos extendidos para ser pagados.

"Teléfono" y "Ocupación" (0.422), esta correlación moderada sugiere que ciertas ocupaciones están asociadas con una mayor probabilidad de tener un teléfono registrado.

Variables con Alta Correlación Negativa

"Duración del Crédito mensual" y "Años" (-0.039), aunque la correlación es débil, la relación negativa podría indicar que los solicitantes mayores tienden a solicitar créditos con menor duración.

Variables con Correlaciones Moderadas

"Saldo de cuenta" y "Credibilidad" (0.348), existe una correlación positiva moderada, lo que indica que un saldo de cuenta más alto está asociado con una mayor credibilidad, este análisis es consistente con la práctica bancaria de considerar los saldos como indicadores de solvencia.

"Años" y "No. de dependientes" (0.291), esta correlación moderada sugiere que las personas mayores tienden a tener más dependientes, esto podría estar relacionado con el tamaño de las familias en diferentes etapas de la vida.

"Tipo de apartamento" y "Años" (0.271), esta relación moderada implica que las personas mayores tienden a vivir en ciertos tipos de apartamentos, posiblemente relacionados con su estabilidad financiera o social.

"Duración del empleo actual" y "Ocupación" (0.305), esta correlación moderada entre estas variables sugiere que ciertas ocupaciones están asociadas con una mayor estabilidad laboral.

Variables con Correlaciones Bajas o Nulas

"Trabajador extranjero" y la mayoría de las variables (0.0), la falta de correlaciones significativas sugiere que ser trabajador extranjero no tiene una relación directa con otras variables financieras o demográficas.

"Garantes" y otras variables (0.0), esta variable no muestra relaciones significativas, lo que indica que podría no ser un factor clave en el análisis.

Tabla 6

Resumen Correlaciones Claves

Variables	Correlación	Interpretación Clave
Tipo de apartamento - Activo disponible más valioso	0.548	Los apartamentos de mayor categoría están relacionados con activos más valiosos
Importe Crédito - Duración del Crédito mensual	0.625	Créditos más grandes suelen tener plazos más largos.
Saldo de cuenta - Credibilidad	0.348	Saldos más altos están asociados con mayor credibilidad.

Años - No. de dependientes	0.291	Las personas mayores tienden a tener más dependientes.
Duración del empleo actual - Ocupación	0.305	Ciertas ocupaciones están asociadas con mayor estabilidad laboral.

Nota. Esta tabla un resumen de correlaciones claves para construcción de nuestro de acuerdo con los valores representativos de cada variable. Tomado de Elaboración Propia (2024).

Así mismo se realizó un análisis utilizando el método estadístico de correlación de Pearson para evaluar la relación lineal entre las variables, Años de edad, Importe crédito y Duración del Crédito mensual, ya que son fundamentales para analizar el comportamiento crediticio de los clientes porque contienen información clave sobre sus características y decisiones financieras, así mismo al analizar estas variables juntas, se busca identificar patrones y relaciones que permitan a las instituciones financieras diseñar productos más ajustados a las necesidades de los clientes, optimizar la gestión del riesgo y mejorar la experiencia del usuario al ofrecer créditos personalizados, de lo cual se obtuvieron los siguientes resultados, así:

Correlación entre Años e Importe de Crédito: La correlación de 0.0323 indica una relación muy débil y positiva entre los años y el importe del crédito. Esto sugiere que no hay una relación lineal significativa entre estas dos variables. El p-valor de 0.3080 es mayor que el umbral común de 0.05, lo que significa que no hay evidencia suficiente para rechazar la hipótesis nula. En otras palabras, no se puede concluir que exista una correlación significativa entre los años y el importe del crédito.

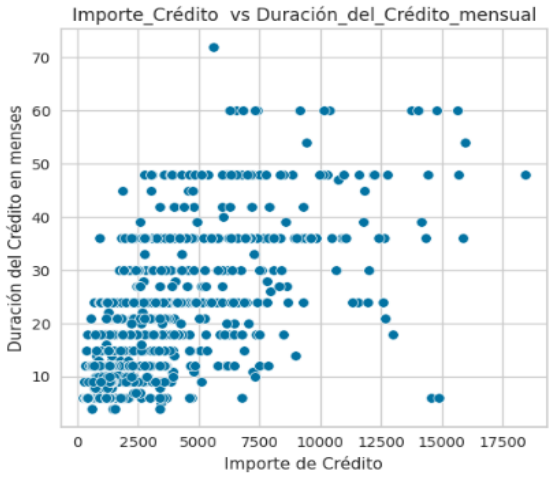
Correlación entre Años y Duración del Crédito Mensual: La correlación de -0.0375 indica una relación muy débil y negativa entre los años y la duración del crédito mensual. Esto

sugiere que, en general, no hay una relación lineal significativa entre estas dos variables. El p-valor de 0.2355, al igual que en el caso anterior, es mayor que 0.05, lo que implica que no hay evidencia suficiente para afirmar que existe una correlación significativa entre los años y la duración del crédito mensual.

Correlación entre Importe de Crédito y Duración del Crédito Mensual: La correlación de 0.6250 indica una relación moderadamente fuerte y positiva entre el importe del crédito y la duración del crédito mensual. Esto sugiere que a medida que aumenta el importe del crédito, también tiende a aumentar la duración del crédito mensual. El p-valor extremadamente bajo (1.8547e-109) es significativamente menor que 0.05, lo que proporciona evidencia sólida para rechazar la hipótesis nula. Esto indica que existe una correlación estadísticamente significativa entre el importe del crédito y la duración del crédito mensual.

Figura 26

Análisis de Valores Faltantes en el Dataset por Variable



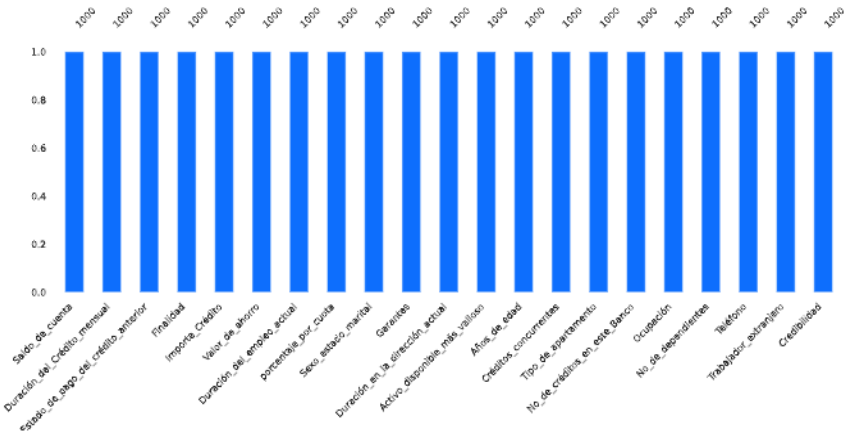
Con base en el análisis de correlación de Pearson, que mostró una relación moderada a fuerte entre el importe del crédito y su duración en meses, se elaboró en la figura 26, los

resultados muestran que los créditos de mayor importe suelen estar asociados con plazos más largos, lo cual es consistente con las prácticas financieras comunes, teniendo en cuenta que la mayoría de los créditos con importe bajo (entre 0 y 5,000) están asociados con duraciones más cortas, concentrándose entre 10 y 30 meses, créditos de importe más alto (mayores a 10,000) suelen tener duraciones mayores, alcanzando hasta 70 meses, aunque se observa una tendencia general, existen puntos dispersos que indican variación, como créditos altos con duraciones cortas o créditos pequeños con duraciones largas.

Valores Perdidos

Figura 27

Análisis de Valores Perdidos en el Dataset por Variable



Este gráfico confirma que no hay valores faltantes en ninguna de las columnas del conjunto de datos, siendo ideal para seguir realizando nuestro análisis estadísticos y entrenamiento de los modelados, ya que no será necesario realizar imputaciones o eliminar filas/columnas debido a datos faltantes.

Posteriormente al realizar el análisis de las variables y su representación gráfica, utilizamos la función “tipos de variables = clasificar_variables(data_train)” para la clasificación

de las variables del conjunto de datos en dos categorías principales Nominales y Cuantitativas. Esta clasificación permite preparar los datos de manera adecuada, asegurando que las variables nominales y cuantitativas sean tratadas correctamente según su naturaleza, además, facilita la selección de técnicas de modelado y análisis exploratorio, como gráficos de frecuencias para variables nominales o histogramas para variables cuantitativas, este análisis inicial es crucial para garantizar la calidad del modelo y su capacidad para predecir el riesgo crediticio de manera precisa y confiable.

Figura 28

Clasificación y Visualización de Tipos de Variables

```
{'Saldo_de_cuenta': 'Nominal',
 'Duración_del_Crédito_mensual': 'Cuantitativa',
 'Estado_de_pago_del_crédito_anterior': 'Nominal',
 'Finalidad': 'Nominal',
 'Importe_Crédito': 'Cuantitativa',
 'Valor_de_ahorro': 'Nominal',
 'Duración_del_empleo_actual': 'Nominal',
 'porcentaje_por_cuota': 'Nominal',
 'Sexo_estado_marital': 'Nominal',
 'Garantes': 'Nominal',
 'Duración_en_la_dirección_actual': 'Nominal',
 'Activo_disponible_más_valioso': 'Nominal',
 'Años_de_edad': 'Cuantitativa',
 'Créditos_concurrentes': 'Nominal',
 'Tipo_de_apartamento': 'Nominal',
 'No_de_créditos_en_este_Banco': 'Nominal',
 'Ocupación': 'Nominal',
 'No_de_dependientes': 'Nominal',
 'Teléfono': 'Nominal',
 'Trabajador_extranjero': 'Nominal',
 'Credibilidad': 'Cuantitativa'}
```

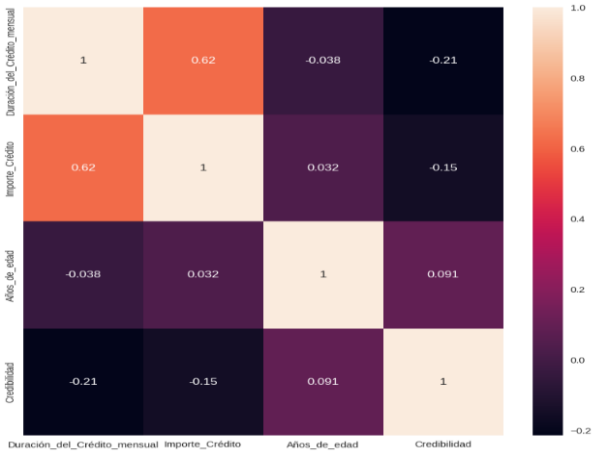
Las variables nominales representan atributos categóricos que no tienen un orden o valor matemático intrínseco, como "Saldo de cuenta", "Finalidad" o "Sexo Estado Marital", estas variables contienen información cualitativa, como categorías o etiquetas, que deben transformarse en un formato numérico antes de ser utilizadas en modelos predictivos, para ello, se pueden emplear técnicas como One-Hot Encoding o Label Encoding, por ejemplo, la variable "Finalidad" puede transformarse en columnas binarias que indiquen si el crédito tiene como propósito "Educación", "Vivienda", etc.

Por otro lado, las variables cuantitativas son aquellas que contienen valores numéricos continuos o discretos, como "Importe_Crédito", "Años" y "Duración_Crédito_mensual",

estas variables tienen significado matemático y pueden utilizarse directamente en análisis estadísticos de nuestro modelo predictivo, sin embargo, en algunos casos, es necesario aplicar técnicas de normalización o estandarización para garantizar que todas las variables estén en la misma escala y evitar que alguna domine el modelo debido a su magnitud.

Figura 29

Matriz de Correlación entre Variables Cuantitativas



Tras clasificar las variables en nominales y cuantitativas, se generó una matriz de correlación para analizar la relación entre las variables cuantitativas seleccionadas: "Duración del Crédito mensual", "Importe_Crédito", "Años" y "Credibilidad". Este análisis permite identificar posibles problemas de multicolinealidad, así como relaciones significativas entre las variables.

Del gráfico se observa que la correlación más alta (0.62) se da entre "Duración del Crédito mensual" e "Importe_Crédito". Esto indica que, en general, créditos de mayor importe suelen estar asociados con plazos más largos. Sin embargo, esta correlación no es lo

suficientemente alta como para representar un problema de multicolinealidad, ya que no supera el umbral crítico de 0.8.

Por otro lado, la variable objetivo "Credibilidad" muestra correlaciones bajas con las demás variables, siendo la más significativa la relación negativa con "Duración del Crédito mensual" (-0.21). Esto sugiere que plazos más largos podrían estar asociados con un mayor riesgo de incumplimiento, aunque la relación es débil. Asimismo, la correlación con "Importe_Crédito" (-0.15) y "Años" (0.091) es baja, lo que indica que estas variables tienen un impacto limitado en la credibilidad de los solicitantes de crédito.

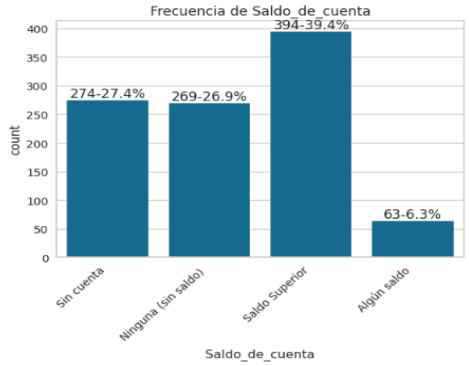
En conclusión, el análisis de correlación confirma que las variables cuantitativas seleccionadas no presentan problemas graves de multicolinealidad y pueden ser utilizadas en el modelo.

Análisis de las Tablas de Distribución de Frecuencias

El análisis de las tablas de distribución de frecuencias proporciona información valiosa sobre las características predominantes de los datos, permitiendo identificar patrones clave y posibles desequilibrios en las variables categóricas, a continuación, se detallan los hallazgos más relevantes, en las siguientes gráficas, así:

Figura 30

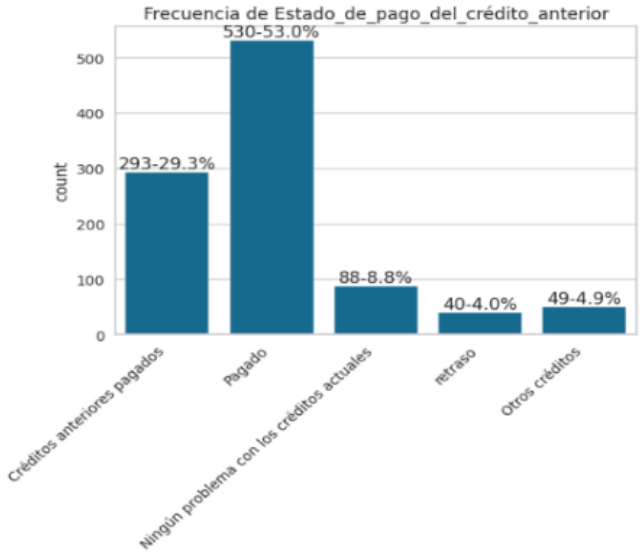
Distribución de Frecuencia del Saldo en la Cuenta



- Saldo de cuenta: la categoría más frecuente es "Saldo Superior" (39.4%), seguida de "Sin cuenta" (27.4%) y "Ninguna (sin saldo)" (26.9%), esto indica que casi el 54% de los clientes no tiene una cuenta activa o saldo significativo, lo que podría influir en su capacidad crediticia y en la evaluación de riesgos.

Figura 31

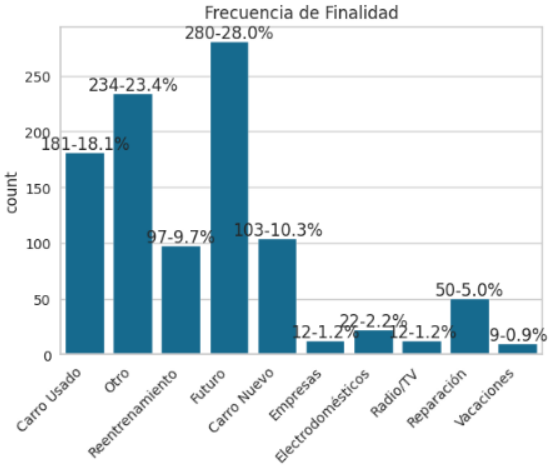
Distribución de Frecuencia de Estado de Pago del Crédito Anterior



Estado de pago del crédito anterior, más de la mitad de los clientes (53%) tiene un historial positivo ("Pagado"), mientras que solo el 4% presenta retrasos, este resultado resalta una base de clientes con buen comportamiento crediticio, aunque un pequeño porcentaje puede representar un mayor riesgo.

Figura 32

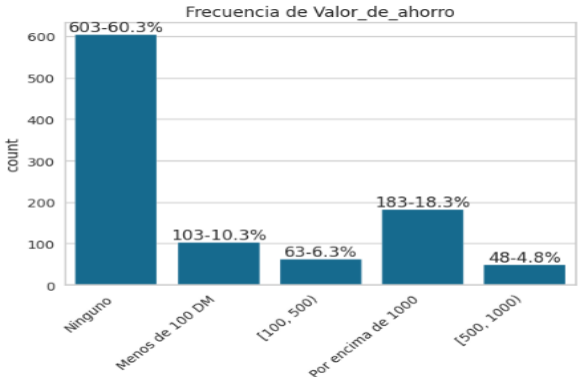
Distribución de Frecuencia de Finalidad



Finalidad, las principales razones para solicitar crédito son "Futuro" (28%), "Otro" (23.4%) y "Carro Usado" (18.1%), esto muestra que los clientes priorizan metas a largo plazo y necesidades inmediatas, como la compra de vehículos, lo cual es útil para segmentar el mercado y diseñar productos financieros específicos.

Figura 33

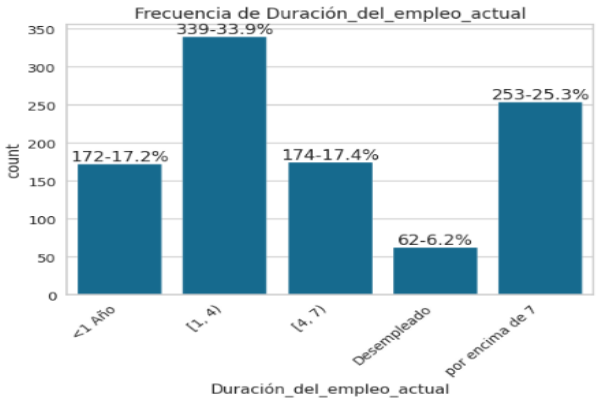
Distribución de Frecuencia de Valor de Ahorro



Valor de ahorro, la mayoría de los clientes (60.3%) no tiene ahorros disponibles, lo que podría indicar una mayor dependencia del crédito para satisfacer sus necesidades financieras, solo un 18.3% cuenta con ahorros superiores a 1000, lo que sugiere un segmento más solvente.

Figura 34

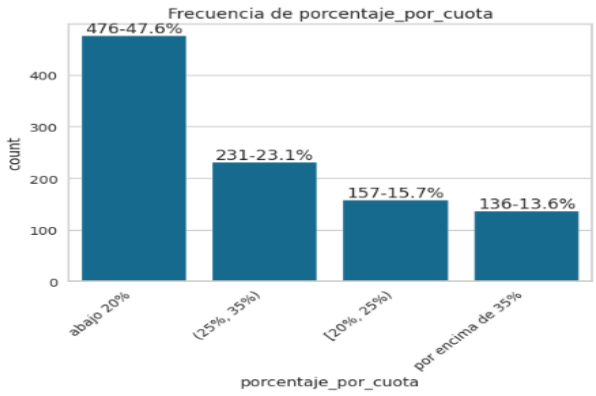
Distribución de Frecuencia de Duración del Empleo Actual



Duración del Empleo Actual, la categoría más frecuente es "[1, 4)" años (33.9%), seguida de "por encima de 7" años (25.3%), esto refleja que una parte importante de los clientes tiene estabilidad laboral moderada, aunque un 17.2% lleva menos de un año en su empleo, lo que podría ser un indicador de mayor riesgo.

Figura 35

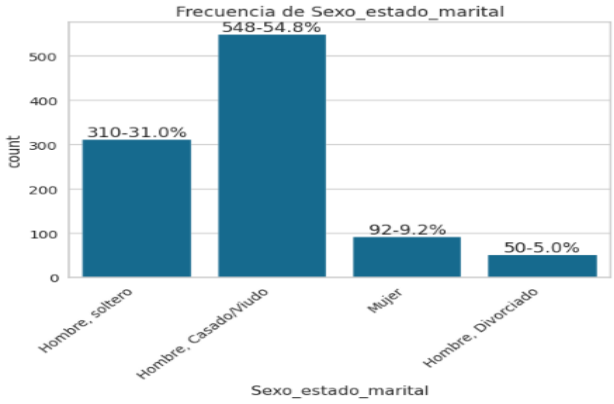
Distribución de Frecuencia de Porcentaje por Cuota



Porcentaje_por_cuota, el 47.6% de los clientes destina menos del 20% de sus ingresos al pago de cuotas, lo que indica una carga financiera manejable, sin embargo, un 13.6% supera el 35%, lo que podría representar un mayor riesgo de incumplimiento.

Figura 36

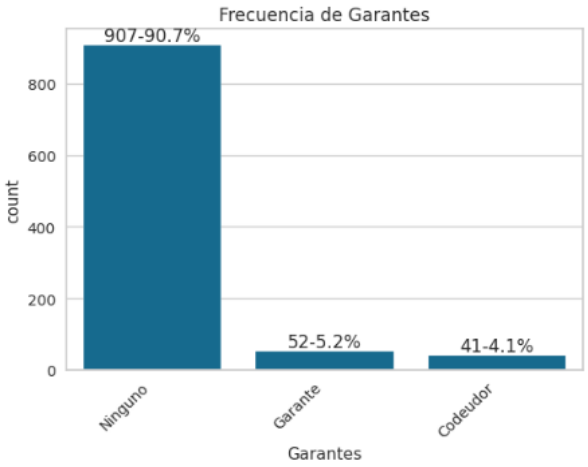
Distribución de Frecuencia de Sexo Estado Marital



Sexo Estado Marital, la mayoría de los clientes son "Hombres, Casados/Viudos" (54.8%), seguidos de "Hombres, Solteros" (31%), esto sugiere que los hombres representan un segmento dominante en la base de datos, con diferentes niveles de estabilidad familiar.

Figura 37

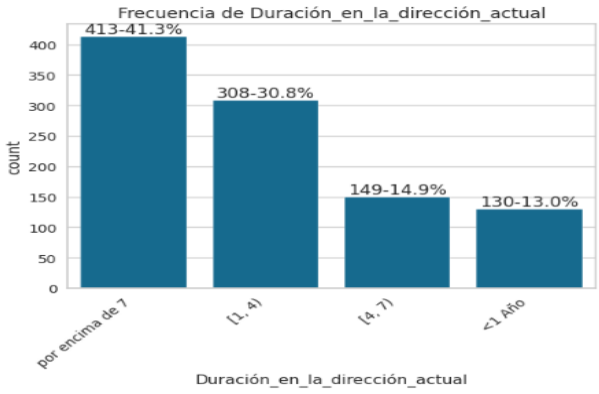
Distribución de Frecuencia de Garantes



Garantes, el 90.7% de los clientes no cuenta con garantes, lo que podría limitar las garantías colaterales disponibles para respaldar los préstamos.

Figura 38

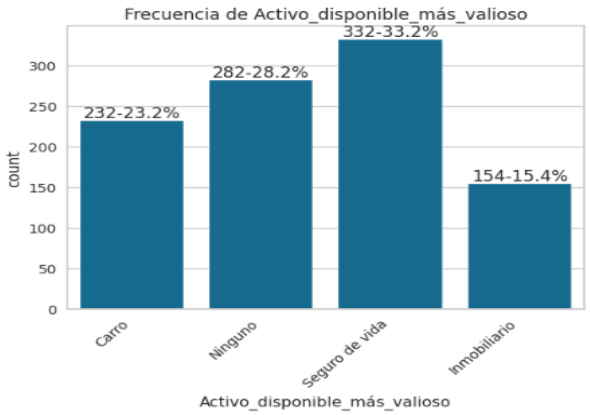
Distribución de Frecuencia de Duración en la Dirección Actual



Duración_en_la_dirección_actual, el 41.3% de los clientes ha vivido más de 7 años en su dirección actual, lo que indica estabilidad residencial, sin embargo, un 13% ha residido menos de un año, lo que podría ser un factor de mayor riesgo.

Figura 39

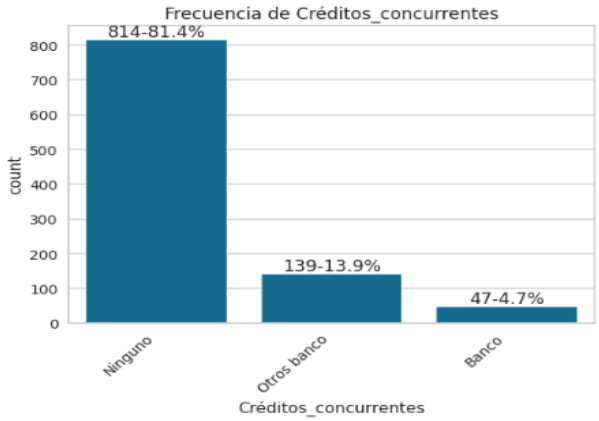
Distribución de Frecuencia de Activo Disponible más Valioso



Activo_disponible_más_valioso, la categoría más común es "Seguro de vida" (33.2%), seguida de "Ninguno" (28.2%), esto sugiere que una proporción significativa de los clientes no posee activos valiosos que puedan ser utilizados como respaldo financiero.

Figura 40

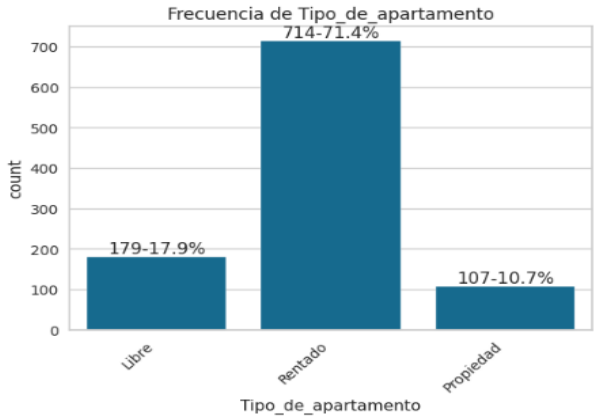
Distribución de Frecuencia de Activo Disponible más Valioso



Créditos_concurrentes, el 81.4% de los clientes no tiene créditos concurrentes, lo que indica que la mayoría no está sobreendeudada, sin embargo, un 13.9% tiene créditos con otros bancos, lo que podría requerir mayor análisis.

Figura 41

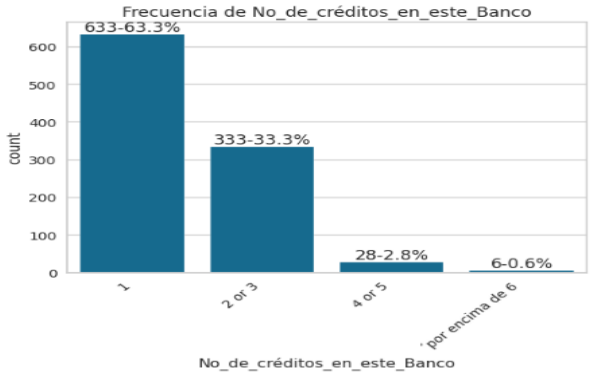
Distribución de Frecuencia de Tipo de Apartamento



Tipo de apartamento, la mayoría de los clientes vive en propiedades rentadas (71.4%), lo que podría influir en su capacidad de ahorro y estabilidad financiera.

Figura 42

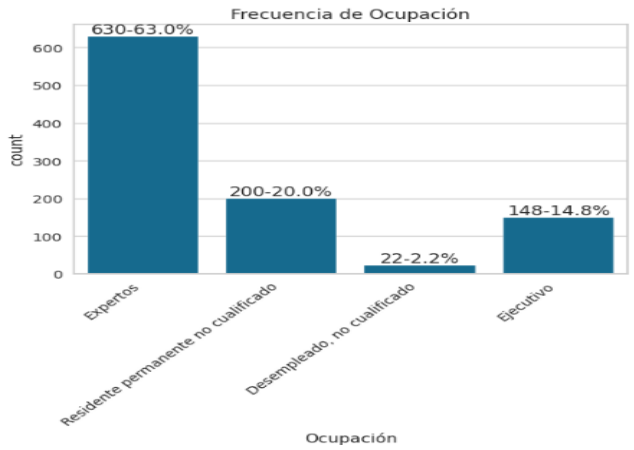
Distribución de Frecuencia del No. de Créditos en este Banco



No de créditos en este banco, el 63.3% de los clientes tiene solo un crédito en el banco, mientras que solo un 3.4% tiene más de 3 créditos, esto sugiere que la mayoría de los clientes no está altamente endeudada en la institución.

Figura 43

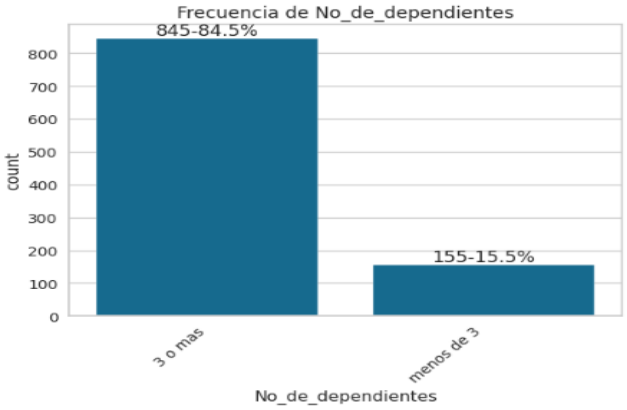
Distribución de Frecuencia de Ocupación



Ocupación: La categoría predominante es "Expertos" (63%), lo que indica que la mayoría de los clientes tiene ocupaciones calificadas. Sin embargo, un 22% pertenece a grupos no cualificados o desempleados, lo que podría representar un mayor riesgo.

Figura 44

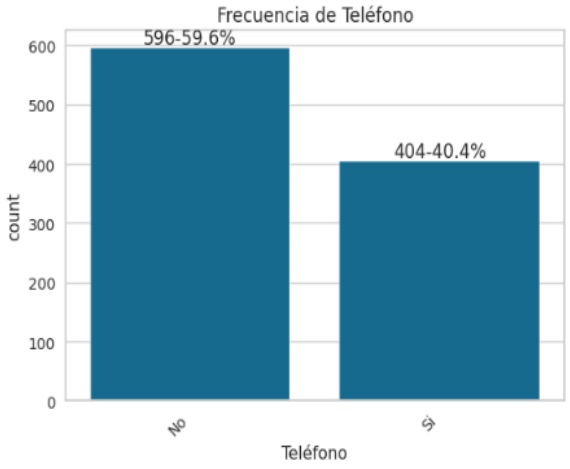
Distribución de Frecuencia de No. de Dependiente



No de dependientes: La mayoría de los clientes (84.5%) tiene "3 o más" dependientes, lo que podría aumentar su carga financiera y afectar su capacidad de pago.

Figura 45

Distribución de Frecuencia del Teléfono



Teléfono, el 59.6% de los clientes no tiene teléfono registrado, lo que podría dificultar la comunicación y el seguimiento de los préstamos.

Importancia del Análisis

Este análisis permite comprender la distribución y las características predominantes de las variables, lo que es crucial para la toma de decisiones, por ejemplo, identificar perfiles de riesgo basados en el historial de pago, estabilidad laboral, activos disponibles y carga financiera ayuda a segmentar a los clientes y diseñar estrategias adecuadas, además, este enfoque permite detectar posibles desequilibrios o sesgos en los datos, como la predominancia de ciertas categorías ("Ninguno" en "Valor de ahorro"), que podrían afectar el rendimiento del modelo predictivo ya que el modelo puede aprender a enfocarse demasiado en esa categoría al momento de hacer las predicciones.

Prueba Chi-Cuadrado para Evaluar la Asociación entre Variables Categóricas

En el presente análisis de score de riesgo crediticio, se aplicó la prueba Chi-Cuadrado con el objetivo de evaluar la independencia o asociación entre dos variables categóricas relevantes para el modelo predictivo, este procedimiento permitió determinar si existe una relación estadísticamente significativa entre dichas variables, lo que resulta fundamental para identificar patrones en el comportamiento crediticio de los clientes.

Tabla 7

Tabla de Contingencia 1

No de dependientes	3 o mas	menos de 3
Tipo de apartamento		
Libre	160	19
Propiedad	77	30
Rentado	608	106

Nota. tabla muestra la distribución de los tipos de apartamento (Libre, Propiedad, Rentado) según el número de dependientes (menos de 3 o 3 o más). Tomado de Elaboración Propia (2024).

Predominancia del Tipo Rentado: La mayoría de las personas que viven en apartamentos rentados tienen tres o más dependientes (60.8%). Esto podría indicar que las familias más grandes tienden a optar por opciones rentadas, posiblemente debido a la flexibilidad que ofrecen.

Menor Proporción en Propiedad y Libre: Tanto en apartamentos libres como en propiedad, hay menos personas con tres o más dependientes (16.0% y 7.7%, respectivamente). Esto sugiere que las familias más pequeñas pueden preferir estos tipos de vivienda.

Tendencias Generales: La mayoría (845) tiene tres o más dependientes, lo que indica una tendencia hacia familias numerosas en comparación con aquellas con menos de tres dependientes (155).

Resultado de Chi-cuadrado tipo de apartamento según los dependientes: 16.336471904397705, p-valor: 0.0002835177143698381, esto sugiere que el tipo de apartamento está significativamente asociado con el número de dependientes. Esto implica que la elección del tipo de apartamento podría depender del número de dependientes en un hogar, lo cual podría ser relevante para la decisión.

Tabla 8

Tabla de Contingencia 2

No de dependientes	3 o mas	menos de 3
Saldo de cuenta		
Ninguna (sin saldo)	238	31
Algún saldo	54	9
Saldo Superior	331	63
Sin cuenta	222	52

Nota. tabla muestra la distribución de saldo en la cuenta (Ninguna, Algún Saldo, Saldo, Saldo Superior, Sin Cuenta) según el número de dependientes (menos de 3 o 3 o más). Tomado de Elaboración Propia (2024).

Predominancia del Grupo sin Saldo: La categoría con mayor número de personas es "Ninguna (sin saldo)", especialmente entre aquellos con tres o más dependientes (23.8%). Esto sugiere que muchas personas con familias numerosas no tienen ahorros o cuentas bancarias.

Saldo Superior y Familias Grandes: Un porcentaje significativo (33.1%) de las personas con tres o más dependientes tiene un "Saldo Superior". Esto puede indicar que las familias grandes tienen una mayor capacidad para ahorrar o gestionar sus finanzas

Bajo Porcentaje en "Algún Saldo": Tanto en el grupo con "Algún saldo" como en "Sin cuenta", los porcentajes son relativamente bajos, lo que sugiere que hay una gran parte de la población que no tiene ahorros significativos ni acceso a cuentas bancarias.

Tendencias Generales: La mayoría (845) tiene tres o más dependientes, lo que indica una tendencia hacia familias numerosas en comparación con aquellas con menos de tres dependientes (155). Esto puede tener implicaciones sobre cómo se gestionan las finanzas familiares.

Oportunidades para Servicios Financieros: Esta información puede ser útil para instituciones financieras al identificar segmentos del mercado que podrían beneficiarse de productos diseñados para fomentar el ahorro y la gestión financiera, especialmente entre familias numerosas.

Resultado de Chi-cuadrado validar saldo de la cuenta según los dependientes: 5.920383215375646, p-valor: 0.11554866201401238, siendo un valor más alto sugiere una mayor diferencia entre lo que se observa y lo que se esperaría si no hubiera relación entre las variables (en este caso, el saldo de la cuenta y el número de dependientes).

En general, un p-valor menor a 0.05 indica que hay suficiente evidencia para rechazar la hipótesis nula (que establece que no hay relación entre las variables).

En este caso, un p-valor de 0.1155 es mayor que 0.05, lo que sugiere que no hay evidencia suficiente para rechazar la hipótesis nula. Esto implica que no se puede concluir que exista una relación significativa entre el saldo de la cuenta y el número de dependientes.

Tabla 9

Tabla de Contingencia 3

No de dependientes	3 o mas	menos de 3
Finalidad		
Carro Nuevo	81	22
Carro Usado	165	16
Electrodomésticos	17	5
Empresas	10	2
Futuro	250	30
Otro	182	52
Radio/TV	12	0
Reentrenamiento	82	15
Reparación	39	11
Vacaciones	7	2

Nota. La tabla muestra cómo se distribuyen las finalidades de uso (Carro Nuevo, Carro Usado, Electrodomésticos, etc.) según el número de dependientes (menos de 3 o 3 o más). Tomado de Elaboración Propia (2024).

Predominancia en Finalidades Específicas

La finalidad del crédito con mayor porcentaje entre aquellos con "3 o más" dependientes es "Futuro" con un notable 34.72%, lo que sugiere que las familias numerosas están interesadas en inversiones a largo plazo, posiblemente para educación o ahorro.

"Carro Usado" también tiene un porcentaje significativo con 22.92%, indicando que muchas familias grandes optan por vehículos usados.

Bajo Interés en Algunas Finalidades

Las finalidades como "Vacaciones" y "Empresas" tienen porcentajes muy bajos, lo que podría indicar que las familias numerosas priorizan necesidades básicas y ahorro sobre gastos recreativos o empresariales.

Comparación entre Grupos

En general, las finalidades del crédito tienden a ser más variadas entre aquellos con "3 o más" dependientes en comparación con el grupo "menos de tres", donde las proporciones son notablemente menores en todas las categorías.

Oportunidades para Instituciones Financieras

Esta información puede ser útil para diseñar productos financieros que se alineen mejor con las necesidades específicas según el número de dependientes, enfocándose en opciones que ayuden a las familias numerosas a planificar su futuro y gestionar sus recursos.

Los resultados Chi-cuadrado validar la finalidad del crédito según los dependientes: 26.81400328265878, p-valor: 0.0015013461175697313, es un valor tan alto sugiere que hay una diferencia significativa entre lo que se observa y lo que se esperaría si no hubiera relación entre las variables.

El p-valor es extremadamente bajo, lo que indica que la probabilidad de observar los resultados obtenidos (o más extremos) bajo la hipótesis nula (que no hay relación entre la finalidad del crédito y el número de dependientes) es muy baja.

La evidencia sugiere que el número de dependientes influye en la finalidad del crédito solicitado. Esto podría implicar que las personas con diferentes cantidades de dependientes

tienen distintas necesidades o prioridades al solicitar un crédito. Es importante considerar qué tipos de finalidades son más comunes entre los diferentes grupos de dependientes, lo cual puede ofrecer información valiosa para instituciones financieras al diseñar productos crediticios adaptados a las necesidades específicas de estos grupos.

Tabla 10

Tabla de Contingencia 4

Estado de pago_ del_crédito_anterior	Créditos anteriores pagados	Ningún problema con los créditos actuales	Otros créditos	Pagado	retraso
Valor de ahorro					
Menos de 100 DM	20	14	8	56	5
Ninguno	180	50	25	318	30
Por encima de 1000	56	16	8	99	4
[100, 500)	17	4	5	36	1
[500, 1000)	20	4	3	21	0

Nota. La tabla muestra la relación entre el estado de pago del crédito anterior y el valor de ahorro, los datos se dividen en varias categorías de ahorro y su correspondiente estado de crédito). Tomado de Elaboración Propia (2024).

Predominancia del Estado "Pagado": La categoría "Pagado" es la más alta, representando 53.0% del total. Esto sugiere que una gran parte de las personas en el estudio han cumplido con sus obligaciones crediticias.

Problemas con Créditos Actuales y Retrasos Bajos: La categoría "Ningún problema con los créditos actuales" es relativamente baja (8.8%), lo que puede indicar que muchos tienen problemas o no están al día con sus pagos.

La categoría "Retraso" tiene un porcentaje bajo (4.0%), lo que puede ser positivo si se considera que solo una pequeña fracción tiene retrasos en sus pagos.

Valor de Ahorro y Estado de Pago Relacionados: Aquellos con "Ninguno" como valor de ahorro son los que más han pagado sus créditos (60.3%), lo que podría indicar que las personas sin ahorros tienden a estar más enfocadas en cumplir con sus obligaciones crediticias.

Los que tienen ahorros por encima de 1000 DM también muestran un buen estado en el pago del crédito anterior.

Los resultados de Chi-cuadrado validar el valor de ahorro según el estado de crédito anterior: 20.4543620055045, p-valor: 0.20044971817169938, en este caso, el p-valor de 0.2004 es mayor que el umbral comúnmente utilizado de 0.05. Esto significa que no hay suficiente evidencia para rechazar la hipótesis nula, que establece que no hay relación entre el valor de ahorro y el estado del crédito anterior. En términos prácticos, un p-valor superior a 0.05 sugiere que cualquier diferencia observada podría ser atribuida al azar y no a una relación real entre las variables.

La falta de significancia en esta prueba sugiere que otros factores podrían estar influyendo en el estado del crédito, o que la relación entre estas variables es débil o inexistente.

Tabla 11

Tabla de Contingencia 5

Duración del Empleo Actual	<1 Año	Desempleado	[1, 4)	[4, 7)	por encima de 7
Sexo Estado Marital					
Hombre, Casado/Viudo	52	32	175	108	181
Hombre, Divorciado	11	3	19	4	13
Hombre, soltero	86	23	107	47	47
Mujer	23	4	38	15	12

Nota. La tabla muestra la distribución de la duración del empleo actual de las personas según su sexo y estado marital. Tomado de Elaboración Propia (2024).

La categoría "Hombre, Casado/Viudo" es la más alta con 54.8% del total. Esto sugiere que este grupo tiene una representación significativa en el mercado laboral.

La categoría "Desempleado" es relativamente baja (6.2%), lo que puede indicar una buena situación laboral general entre los encuestados. La duración "Menos de un Año" también es baja (17.2%), sugiriendo que muchos tienen empleos más estables.

Aquellos con "Más de siete años" en el empleo representan un 25.3% significativo. Esto sugiere que hay un número considerable de personas con estabilidad laboral a largo plazo.

Los hombres solteros tienen una representación notable en las categorías con menos duración de empleo (Menos de un año y Desempleados), lo que podría indicar que este grupo enfrenta más desafíos laborales. Las mujeres tienen una representación menor en comparación con los hombres en todas las categorías.

El resultado de Chi-cuadrado validar el estado civil según duración empleo: 82.84379436077376, p-valor: 1.1803034107604528e-12, es un valor tan alto sugiere que hay diferencias importantes entre lo observado y lo que se esperaría si no hubiera relación entre las variables.

El p-valor es extremadamente bajo, lo que indica que la probabilidad de observar los resultados obtenidos (o más extremos) bajo la hipótesis nula (que no hay relación entre el estado civil y la duración del empleo) es prácticamente cero.

La evidencia sugiere que el estado civil influye en la duración del empleo. Esto podría implicar que diferentes estados civiles (soltero, casado, divorciado, etc.) tienen diferentes patrones o comportamientos en relación con la duración del empleo.

Tabla 12

Tabla de Contingencia 6

Activo_disponible_más_valioso	Carro	Inmobiliario	Ninguno	Seguro de vida
Duración del Empleo Actual				
<1 Año	53	18	52	49
Desempleado	15	19	10	18
[1, 4)	69	34	122	114
[4, 7)	43	18	42	71
por encima de 7	52	65	56	80

Nota. La tabla analiza la relación entre la duración del empleo actual y el activo más valioso disponible de las personas. Tomado de Elaboración Propia (2024).

Predominancia del Activo "Ninguno" y "Seguro de Vida": La categoría "Ninguno" es la más alta con 29.2% del total, lo que sugiere que muchas personas no tienen un activo significativo como su posesión más valiosa.

"Seguro de vida" también tiene una representación significativa con 33.2%, lo que podría indicar que es considerado un activo valioso para muchas personas.

Estabilidad en el Empleo a Largo Plazo y Activos Disponibles: Aquellos con "Más de siete años" en el empleo tienen una buena distribución en activos como "Inmobiliario" (65) y "Seguro de Vida" (80), lo que sugiere que las personas con mayor estabilidad laboral tienden a poseer activos más valiosos.

Desempleo y Activos Disponibles Bajos: El grupo desempleado tiene una baja representación en activos significativos, especialmente en "Ninguno" (10), lo que puede indicar que aquellos sin empleo pueden depender menos de activos valiosos.

Distribución por Duración del Empleo en Diferentes Activos Disponibles: La duración del empleo entre aquellos con activos como "Carro" y "Inmobiliario" es notablemente alta en las categorías [1,4) y más de siete años, sugiriendo que tener un empleo estable puede estar asociado con la posesión de activos valiosos.

El resultado de Chi-cuadrado validar empleo actual según Activos disponibles: 64.93787921616979, p-valor: 2.7993196967883535e-09, su p-valor es mucho menor que el umbral comúnmente utilizado de 0.05, se puede rechazar la hipótesis nula. Esto sugiere que existe una relación estadísticamente significativa entre el empleo actual y los activos disponibles.

La evidencia sugiere que el estado del empleo actual está influenciado por la cantidad de activos disponibles. Esto puede implicar que las personas con diferentes niveles de activos tienen diferentes patrones o comportamientos en relación con su situación laboral.

Tabla 13

Tabla de Contingencia 7

Ocupación	Desempleado, no cualificado	Ejecutivo	Expertos	Residente permanente no cualificado
Créditos_concurrentes				
Banco	0	12	25	10
Ninguno	18	110	532	154
Otros banco**s**	4	26	73	36

Nota. La tabla analiza la relación entre la duración del empleo actual y el activo más valioso disponible de las personas. Tomado de Elaboración Propia (2024).

Predominancia del Crédito "Ninguno": La categoría "Ninguno" es la más alta con 81.4% del total, lo que sugiere que una gran mayoría de las personas en el estudio no tienen créditos concurrentes.

Bajo Uso de Créditos Bancarios: La categoría "Banco" tiene una representación muy baja con solo 4.7%, lo que podría indicar que pocos individuos están utilizando créditos a través de bancos.

Distribución de Créditos según Ocupación: Los "Expertos" tienen la mayor cantidad de créditos "Ninguno" (532), lo que sugiere que este grupo puede tener menos necesidad o acceso a créditos.

Los "Ejecutivos" y "Residentes permanentes no cualificados" también muestran un uso significativo de créditos "Ninguno", pero tienen más diversidad en el uso de otros tipos de crédito.

Desempleo y Acceso a Créditos: El grupo de "Desempleado, no cualificado" tiene un bajo uso de créditos, lo que puede reflejar dificultades en el acceso a financiamiento debido a su situación laboral.

El resultado de Chi-cuadrado validar créditos concurrentes según ocupación: 14.116615749042458, p-valor: 0.028360284678683602, su p-valor es menor que el umbral comúnmente utilizado de 0.05, lo que indica que hay suficiente evidencia para rechazar la hipótesis nula. Esto significa que existe una relación estadísticamente significativa entre los créditos concurrentes y la ocupación.

Existencia de Asociación Significativa: La evidencia sugiere que el tipo de ocupación influye en la posibilidad de tener créditos concurrentes. Esto podría implicar que diferentes ocupaciones tienen diferentes patrones o comportamientos en relación con la obtención de créditos.

Implicaciones para el Análisis: Esta información puede ser relevante para instituciones financieras, ya que entender cómo la ocupación afecta el acceso a créditos puede ayudar a diseñar productos financieros más adecuados para diferentes grupos ocupacionales.

Procesamiento de Datos

En el contexto del análisis predictivo del score de riesgo crediticio mediante técnicas de machine learning, el procesamiento de datos constituye una etapa fundamental para garantizar la calidad y representatividad de los datos utilizados en el modelo, esta fase implica la preparación, limpieza y segmentación de los datos, asegurando que la información esté estructurada de manera óptima para el entrenamiento y la evaluación del modelo predictivo.

El código presentado implementa una división de los datos disponibles en dos conjuntos principales:

Figura 46

Segmentación de Datos para Modelado y Evaluación

```
data = data_train.sample(frac=0.95, random_state=786)
data_unseen = data_train.drop(data.index)
data.reset_index(inplace=True, drop=True)
data_unseen.reset_index(inplace=True, drop=True)
print('Data for Modeling: ' + str(data.shape))
print('Unseen Data For Predictions: ' + str(data_unseen.shape))
```

- Datos para Modelado (Training Data): Este conjunto, que representa el 95% de los datos originales, se utiliza para entrenar y validar el modelo predictivo. La selección aleatoria de los datos mediante frac=0.95 asegura que el conjunto de entrenamiento sea representativo y reduzca posibles sesgos en el modelo.
- Datos No Vistos (Unseen Data): El 5% restante de los datos se reserva como un conjunto independiente, destinado exclusivamente a la evaluación final del modelo. Este enfoque permite medir el desempeño del modelo en datos completamente nuevos, emulando escenarios del mundo real.

Resultados del Proceso

El resultado de este procedimiento de acuerdo con la implementación del código antes descrito garantiza una segmentación adecuada de los datos en dos subconjuntos clave:

- Datos para Modelado: Contienen el 95% de los datos originales, con un tamaño de (n, m) (donde n es el número de muestras - filas y m el número de características - columnas).
- Datos No Vistos: Incluyen el 5% restante, con un tamaño de (k, m) (donde k es el número de muestras – filas y m el número de características - columnas).

Por ejemplo, si los datos originales contienen 1000 muestras y 20 características, los tamaños resultantes serían:

- Datos para Modelado: (950, 20)
- Datos No Vistos: (50, 20)

Importancia de esta Actividad

La segmentación de los datos es fundamental para evitar problemas como el *overfitting* (sobreajuste) y garantizar que el modelo generalice correctamente. Además, el uso de datos no vistos en la evaluación final proporciona una métrica objetiva del desempeño del modelo, lo que facilita la toma de decisiones informadas en el contexto financiero.

Este enfoque metodológico asegura que el modelo predictivo sea robusto, confiable y capaz de generar predicciones precisas del score de riesgo crediticio, apoyando así la toma de decisiones financieras estratégicas.

Tabla 14

Resumen del Preprocesamiento y Configuración del Modelo

	Description	Value
0	Session id	123
1	Target	Credibilidad
2	Target type	Binary
3	Original data shape	(950, 21)
4	Transformed data shape	(950, 69)
5	Transformed train set shape	(665, 69)
6	Transformed test set shape	(285, 69)
7	Numeric features	3
8	Categorical features	17
9	Preprocess	True
10	Imputation type	simple

11	Numeric imputation	mean
12	Categorical imputation	mode
13	Maximum one-hot encoding	25
14	Encoding method	None
15	Fold Generator	StratifiedKFold
16	Fold Number	10
17	CPU Jobs	-1
18	Use GPU	False
19	Log Experiment	False
20	Experiment Name	clf-default-name
21	USI	008b

Nota. La Tabla que detalla el tamaño de los datos, las transformaciones aplicadas, la división en conjuntos de entrenamiento y prueba. Tomado de Elaboración Propia (2024).

La tabla presentada resume de manera efectiva el flujo de trabajo del código y los resultados del preprocesamiento aplicado, el aumento en el número de características, la imputación de valores faltantes y la división de los datos en conjuntos de entrenamiento y prueba destacan la importancia de un preprocesamiento adecuado para garantizar el éxito del modelo, además, la configuración de validación cruzada y el uso óptimo de los recursos computacionales refuerzan la robustez del análisis. Este enfoque asegura que los resultados obtenidos sean confiables y puedan ser utilizados para evaluar la credibilidad de manera precisa y reproducible.

Interpretación de la Tabla

Según Configuración General

- La variable Session ID (123) es un identificador único que permite rastrear la sesión de trabajo actual. Esto es útil para reproducir los resultados en futuras ejecuciones.
- La variable Target (Credibilidad) es la variable objetivo que se busca predecir. En

este caso, representa la "Credibilidad", probablemente una etiqueta binaria que clasifica a los clientes según su riesgo crediticio.

- La variable Target Type (Binary) indica que la variable objetivo tiene dos posibles valores (por ejemplo, "Aprobado" o "Rechazado").

Según Tamaño de los Datos

- Original Data Shape (950, 21): el conjunto de datos inicial contiene 950 muestras (filas) y 21 características (columnas).
- Transformed Data Shape (950, 69): después del preprocesamiento, el número de características aumenta a 69 debido a transformaciones como la codificación de variables categóricas (por ejemplo, one-hot encoding).
- Transformed Train Set Shape (665, 69): el conjunto de entrenamiento incluye 665 muestras y 69 características, lo que corresponde aproximadamente al 70% de los datos transformados.
- Transformed Test Set Shape (285, 69): el conjunto de prueba incluye 285 muestras y 69 características, lo que corresponde aproximadamente al 30% de los datos transformados.

Características de los Datos

- Numeric Features (3): hay tres características numéricas en el conjunto de datos original.
- Categorical Features (17): hay 17 características categóricas que han sido transformadas mediante técnicas como one-hot encoding.
- Maximum One-Hot Encoding (25): se establece un límite de 25 categorías para la codificación one-hot, lo que ayuda a evitar una explosión en el número de columnas.

Preprocesamiento

- Preprocess (True): indica que los datos han sido preprocesados automáticamente.

Esto incluye tareas como la imputación de valores faltantes y la codificación de variables categóricas.

- Imputation Type (Simple): el método de imputación utilizado es simple.
- Numeric Imputation (Mean): los valores faltantes en las características numéricas se reemplazan por la media.
- Categorical Imputation (Mode): los valores faltantes en las características categóricas se reemplazan por la moda.

Configuración del Modelo

- Fold Generator (StratifiedKFold): se utiliza una validación cruzada estratificada para garantizar que la proporción de clases en cada subconjunto sea representativa del conjunto completo.
- Fold Number (10): se realizan 10 particiones (folds) durante la validación cruzada.
- CPU Jobs (-1): se utilizan todos los núcleos disponibles del procesador para optimizar el rendimiento.
- Use GPU (False): indica que no se está utilizando una GPU para acelerar el procesamiento.

Información Adicional

- Log Experiment (False): no se está registrando automáticamente la experimentación.
- Experiment Name (clf-default-name): nombre predeterminado del experimento.
- USI (008b): identificador único del experimento.

Comparación de Modelos

En el proyecto de análisis predictivo del score de riesgo crediticio, la etapa de Comparación de Modelos es fundamental para identificar el algoritmo de clasificación que mejor se ajuste a los datos y proporcione las predicciones más precisas sobre la credibilidad crediticia de los clientes. Durante este proceso, se evalúan múltiples modelos de aprendizaje automático utilizando métricas estándar para determinar su rendimiento.

- Funciones y Procesos Clave**:** Esta función de PyCaret permite comparar automáticamente diversos algoritmos de clasificación (como árboles de decisión, bosques aleatorios, regresión logística, entre otros). Se evalúan con base en métricas como la precisión (*accuracy*), el área bajo la curva ROC (*AUC*), sensibilidad (*recall*), y F1-score.
- Selección del Mejor Modelo: A partir de los resultados de compare_models, se selecciona el modelo con el mejor desempeño según la métrica principal definida (por ejemplo, AUC o precisión).
- Uso de tune_model**:** Una vez identificado el modelo más prometedor, se utiliza esta función para optimizar sus hiperparámetros, mejorando su rendimiento en el conjunto de prueba.
- Validación Cruzada: Durante la comparación y el ajuste, se emplea validación cruzada estratificada para garantizar que los resultados sean consistentes y generalizables, minimizando el riesgo de sobreajuste.
- Evaluación Final: El modelo seleccionado se evalúa en el conjunto de prueba para confirmar su desempeño y generar métricas finales que respalden su uso en la predicción del *score crediticio*.
- Importancia del Proceso: Este proceso asegura que el modelo elegido no solo sea

preciso, sino también robusto y confiable para la clasificación del riesgo crediticio, la automatización de PyCaret permite explorar rápidamente múltiples algoritmos y configuraciones, optimizando el flujo de trabajo y facilitando la toma de decisiones basada en datos.

Tabla 15

Configuración del Modelo de Clasificación

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.7713	0.7841	0.9100	0.7968	0.8483	0.3879	0.4085	0.7010
et	Extra Trees Classifier	0.7638	0.7761	0.8952	0.7985	0.8419	0.3752	0.3947	0.4680
gbc	Gradient Boosting Classifier	0.7593	0.7743	0.8737	0.8048	0.8366	0.3800	0.3886	0.7490
lightgbm	Light Gradient Boosting Machine	0.7519	0.7767	0.8631	0.8024	0.8304	0.3672	0.3744	0.5590
ridge	Ridge Classifier	0.7503	0.7647	0.8610	0.8023	0.8289	0.3634	0.3724	0.3860
lr	Logistic Regression	0.7473	0.7677	0.8525	0.8041	0.8260	0.3620	0.3690	0.9920
ada	Ada Boost Classifier	0.7473	0.7318	0.8610	0.7979	0.8275	0.3554	0.3611	0.4240
lda	Linear Discriminant Analysis	0.7412	0.7637	0.8418	0.8033	0.8207	0.3527	0.3572	0.2890

nb	Naive Bayes	0.7112	0.7315	0.7712	0.8108	0.7891	0.3293	0.3327	0.3510
dummy	Dummy Classifier	0.7038	0.5000	1.0000	0.7038	0.8261	0.0000	0.0000	0.2820
knn	K Neighbors Classifier	0.6813	0.5751	0.8697	0.7309	0.7936	0.1154	0.1214	0.4930
dt	Decision Tree Classifier	0.6798	0.6372	0.7414	0.7896	0.7638	0.2660	0.2683	0.2830
svm	SVM - Linear Kernel	0.6377	0.5230	0.8037	0.7549	0.7010	0.0414	0.0530	0.2990
qda	Quadratic Discriminant Analysis	0.4557	0.6106	0.3289	0.6597	0.3832	0.0794	0.0664	0.2790

Nota. La Tabla muestra los parámetros del experimento de clasificación y preprocesamiento, destacando a Random Forest como el modelo más robusto. Tomado de Elaboración Propia (2024).

Interpretación de la Tabla Configurada y el Mejor Modelo

Contexto de la Configuración del Modelo

La tabla presentada muestra los parámetros y configuraciones utilizadas en el experimento de clasificación para el proyecto de score crediticio. Estos son algunos puntos importantes:

- Datos Originales y Transformados: Los datos originales tienen una forma de (950, 21), lo que indica 950 registros y 21 características. Después del preprocesamiento (como codificación y tratamiento de valores nulos), los datos se transformaron a (950, 69), lo que

sugiere que se añadieron nuevas columnas debido a la codificación de variables categóricas y otras transformaciones.

- División del Conjunto de Datos: El conjunto de datos se dividió en entrenamiento (665 registros) y prueba (285 registros) con las mismas características transformadas (69 columnas).
- Características Numéricas y Categóricas: Hay 3 características numéricas y 17 categóricas, lo que confirma la necesidad de codificación para las categóricas.
- Preprocesamiento: El preprocesamiento está activado (Preprocess = True), con imputación de valores faltantes para las características numéricas (media) y categóricas (moda).
- Validación Cruzada: Se utilizó un generador de pliegues estratificado (StratifiedKFold) con 10 pliegues, lo que asegura una evaluación robusta del modelo al mantener la proporción de clases en cada pliegue.

Identificación del Mejor Modelo

Según la interpretación proporcionada, el Random Forest Classifier se destacó como el mejor modelo basado en las métricas clave:

- Accuracy: proporción de predicciones correctas sobre el total de predicciones. Un valor más alto indica un mejor rendimiento general. Mejor modelo: Random Forest Classifier (0.7713). Peor modelo: Quadratic Discriminant Analysis (0.4557).
- AUC (Area Under the Curve): Medida que refleja la capacidad del modelo para distinguir entre clases. Un AUC de 1 indica un modelo perfecto, mientras que 0.5 indica un modelo sin capacidad discriminativa. Mejor modelo, Random Forest Classifier (0.7841). Peor modelo: Dummy Classifier (0.5000).

- Recall: Proporción de verdaderos positivos sobre el total de positivos reales. Indica la capacidad del modelo para identificar correctamente las instancias positivas. Mejor modelo: Random Forest Classifier (0.9100). Peor modelo: Quadratic Discriminant Analysis (0.3289).
- Precision (Prec.): Proporción de verdaderos positivos sobre el total de positivos predichos. Mide la exactitud de las predicciones positivas. Mejor modelo: K Neighbors Classifier (0.8108). Peor modelo: Dummy Classifier (0.7038).
- F1 Score: Media armónica entre precisión y recall, proporcionando un balance entre ambas métricas. Es útil en situaciones donde hay una clase desbalanceada. Mejor modelo: Random Forest Classifier (0.8483). Peor modelo: Quadratic Discriminant Analysis (0.3832).
- Kappa: Estadística que mide la concordancia entre las predicciones y las observaciones, ajustando por la posibilidad de que las coincidencias ocurran por azar. Mejor modelo: Random Forest Classifier (0.3879). Peor modelo: Dummy Classifier (0.0000).
- MCC (Matthews Correlation Coefficient): Mide la calidad de las clasificaciones binarias, teniendo en cuenta verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Mejor modelo: Random Forest Classifier (0.4085). Peor modelo: Dummy Classifier (0.0000).
- TT (Sec): Tiempo total en segundos que tomó entrenar el modelo. Modelo más rápido: Dummy Classifier (0.2680). Modelo más lento: Light Gradient Boosting Machine (1.1210).

Cierre

El Random Forest Classifier sobresale como el modelo más robusto y confiable para este conjunto de datos, ya que domina en múltiples métricas clave. Esto sugiere que es el más

adecuado para predecir el score crediticio, ofreciendo precisión y generalización en sus predicciones. Se recomienda utilizar este modelo para la implementación final, asegurándose de ajustar sus hiperparámetros para maximizar aún más su rendimiento.

Tabla 16

Desempeño del Modelo Random Forest

Fold	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.7761	0.7867	0.9149	0.7963	0.8515	0.4057	0.4222
1	0.7910	0.8011	0.9362	0.8000	0.8627	0.4363	0.4609
2	0.7612	0.7644	0.9362	0.7719	0.8462	0.3342	0.3675
3	0.8209	0.8468	0.9574	0.8182	0.8824	0.5168	0.5459
4	0.7761	0.7734	0.9574	0.7759	0.8571	0.3651	0.4126
5	0.7879	0.8371	0.9149	0.8113	0.8600	0.4289	0.4424
6	0.7727	0.7273	0.9149	0.7963	0.8515	0.3774	0.3944
7	0.7727	0.8046	0.8511	0.8333	0.8421	0.4369	0.4372
8	0.7576	0.8364	0.8043	0.8409	0.8222	0.4419	0.4429
9	0.6970	0.6636	0.9130	0.7241	0.8077	0.1361	0.1592
Mean	0.7713	0.7841	0.9100	0.7968	0.8483	0.3879	0.4085
Std	0.0300	0.0533	0.0454	0.0321	0.0199	0.0965	0.0944

Nota. La Tabla muestra los resultados de validación cruzada destacando la robustez del modelo con un recall de 0.91 y un AUC de 0.7841. Tomado de Elaboración Propia (2024).

Con se evidencia en la presente grafica se utilizó el algoritmo Random Forest para evaluar su desempeño en la clasificación del conjunto de datos, la tabla muestra los resultados obtenidos en 10 pliegues de validación cruzada, donde se calcularon métricas clave como accuracy, AUC, recall, precisión, F1 score, Kappa y MCC, estas métricas permiten evaluar la capacidad del modelo para realizar predicciones precisas y equilibradas.

El promedio de los resultados (resaltado en amarillo) indica un excelente desempeño general del modelo, destacando un recall de 0.91 y un AUC de 0.7841, lo que sugiere que el modelo tiene alta sensibilidad y capacidad para diferenciar entre clases, esto lo convierte en una herramienta confiable para predecir el score crediticio en nuestro proyecto, además, los valores bajos de desviación estándar reflejan consistencia en los resultados obtenidos.

Dado que Random Forest demostró ser el modelo más robusto y confiable, el siguiente paso es su implementación práctica en el sistema, esto incluye ajustar sus hiperparámetros para maximizar su desempeño, evaluar su efectividad en datos reales y analizar su capacidad de generalización. También se explorará cómo integrarlo en un sistema automatizado que facilite la toma de decisiones en el contexto del score crediticio.

Tabla 17

Comparación de Desempeño: Decisión Tree vs Random Forest

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.6716	0.6367	0.7234	0.7907	0.7556	0.2586	0.2609
1	0.6119	0.5941	0.6383	0.7692	0.6977	0.1689	0.1747
2	0.6567	0.6261	0.7021	0.7857	0.7416	0.2352	0.2386
3	0.7463	0.7186	0.7872	0.8409	0.8132	0.4192	0.4214
4	0.6567	0.6117	0.7234	0.7727	0.7473	0.2142	0.2153
5	0.7879	0.7256	0.8723	0.8367	0.8542	0.4659	0.4673
6	0.7273	0.6517	0.8298	0.7959	0.8125	0.3133	0.3142
7	0.6212	0.5929	0.6596	0.7750	0.7126	0.1675	0.1723
8	0.7273	0.6913	0.7826	0.8182	0.8000	0.3721	0.3730

9	0.5909	0.5228	0.6957	0.7111	0.7033	0.0450	0.0450
Mean	0.6798	0.6372	0.7414	0.7896	0.7638	0.2660	0.2683
Std	0.0614	0.0594	0.0710	0.0358	0.0508	0.1219	0.1215

Nota. La gráfica muestra métricas clave (Accuracy, AUC, Recall, F1 Score) donde el color amarillo representa el desempeño de Decision Tree. Tomado de Elaboración Propia (2024).

La evaluación del modelo Decision Tree Classifier tras aplicar Random Forest permite comparar un modelo básico con uno avanzado y justificar la selección del más adecuado, la gráfica muestra métricas clave (Accuracy, AUC, Recall, F1 Score) donde el color amarillo representa el desempeño de Decision Tree, evidenciando su menor precisión frente a Random Forest, que logra mejores resultados gracias a su robustez y estabilidad.

Este análisis no solo valida la superioridad de Random Forest, sino que también documenta la evolución metodológica y refuerza la justificación del modelo final.

Tabla 18

Rendimiento del Modelo Optimizado

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold							
0	0.7015	0.7915	0.6809	0.8649	0.7619	0.3767	0.3965
1	0.7463	0.8000	0.7660	0.8571	0.8090	0.4347	0.4409
2	0.6866	0.7266	0.6809	0.8421	0.7529	0.3373	0.3517
3	0.7910	0.8617	0.8085	0.8837	0.8444	0.5282	0.5330
4	0.7015	0.7691	0.7447	0.8140	0.7778	0.3260	0.3290
5	0.8030	0.8443	0.8298	0.8864	0.8571	0.5412	0.5442
6	0.6061	0.6641	0.5957	0.8000	0.6829	0.1913	0.2062
7	0.7273	0.8309	0.6809	0.9143	0.7805	0.4402	0.4744

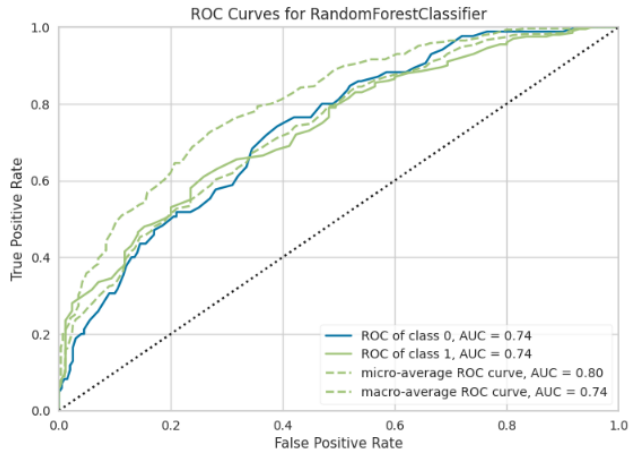
8	0.6667	0.8174	0.6087	0.8750	0.7179	0.3412	0.3758
9	0.6667	0.7293	0.7391	0.7727	0.7556	0.2326	0.2331
Mean	0.7097	0.7835	0.7135	0.8510	0.7740	0.3749	0.3885
Std	0.0566	0.0585	0.0741	0.0416	0.0505	0.1086	0.1085

Nota. La tabla muestra las métricas promedio del modelo Random Forest Classifier tras su optimización. Tomado de Elaboración Propia (2024).

Se realizó la optimización del modelo Random Forest Classifier utilizando validación cruzada para evaluar su desempeño mediante métricas clave, los resultados muestran un Accuracy promedio de 0.7097, indicando un buen rendimiento general, y un AUC de 0.7835, reflejando una sólida capacidad para distinguir entre clases, además, destaca un Recall de 0.7135, que asegura la detección de la mayoría de los casos positivos, y una Precisión de 0.8510, que garantiza una alta proporción de predicciones correctas, el F1 Score de 0.7740 confirma un equilibrio adecuado entre precisión y recall. Este análisis es crucial para el proyecto, ya que, valida la eficacia del modelo optimizado, asegurando que cumpla con los requisitos de predicción y toma de decisiones en escenarios reales.

Figura 47

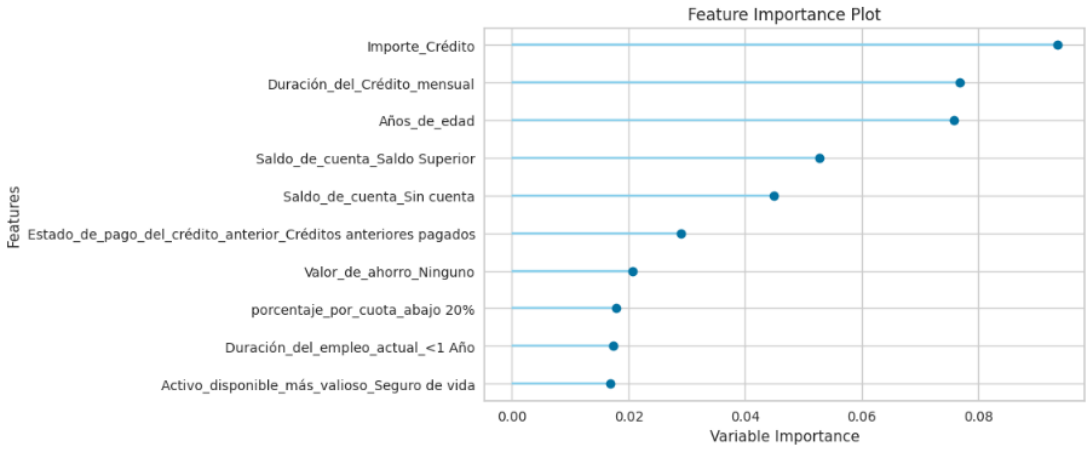
Visualización Resultados Modelo Curva ROC para Clasificación del Riesgo



La gráfica de curvas ROC del modelo Random Foret Classifier muestra un desempeño moderado en la clasificación del riesgo crediticio. Las curvas ROC para ambas clases (0 y 1) presentan un AUC de 0.74, lo que indica una capacidad aceptable para distinguir entre clientes de bajo y alto riesgo. Además, el micro-promedio AUC de 0.80 refleja un buen rendimiento global al considerar todas las instancias, mientras que el macro-promedio AUC de 0.74 confirma un desempeño consistente entre las clases. Este análisis es fundamental para evaluar la eficacia del modelo en la toma de decisiones crediticias, asegurando una clasificación confiable y precisa de los perfiles de riesgo.

Figura 48

Distribución de Factores de Importancia de Factores en el Modelo

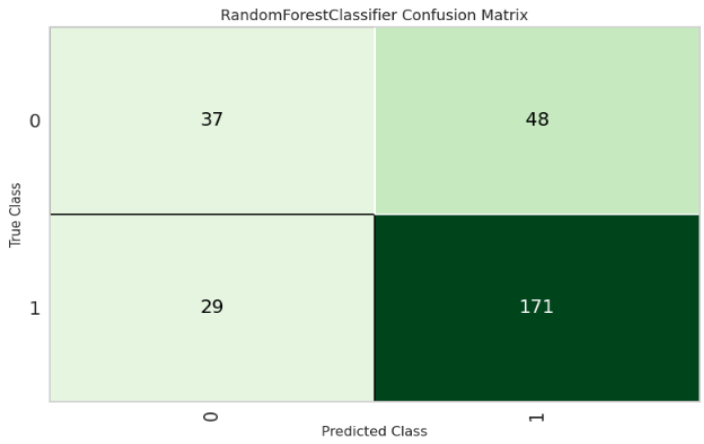


La gráfica de distribución de factores importantes destaca las variables más relevantes utilizadas por el modelo Random Forest Classifier para predecir el riesgo crediticio. Entre las principales variables se encuentra el Importe crédito, que tiene la mayor influencia en el modelo, lo que indica que la cantidad solicitada es un factor clave para evaluar el riesgo. Le siguen la Duración del Crédito mensual, que refleja la relación entre plazos más largos y el riesgo, y los Años, que probablemente están asociados con la estabilidad financiera y la capacidad de pago.

Otras variables importantes incluyen el estado del saldo de la cuenta (Saldo de cuenta Saldo Superior y Saldo de cuenta Sin cuenta) y el Estado de pago del crédito anterior, que mide el historial de pagos previos como un fuerte indicador del comportamiento crediticio futuro. Por otro lado, factores como Activo_disponible_más_valioso_Seguro de vida y Duración del Empleo Actual_<1 Año tienen un impacto menor en el modelo.

Figura 49

Desempeño del Modelo: Matriz de Confusión



La matriz de confusión presentada corresponde a un modelo Random Forest Classifier utilizado para evaluar su desempeño en la clasificación de dos categorías: Clase 0 (por ejemplo, clientes con bajo riesgo crediticio) y Clase 1 (clientes con alto riesgo crediticio).

Los resultados muestran que el modelo identificó correctamente 37 casos como Clase 0 (verdaderos negativos) y 171 casos como Clase 1 (verdaderos positivos). Sin embargo, también cometió errores, clasificando incorrectamente 48 casos como Clase 1 cuando pertenecían a Clase 0 (falsos positivos) y 29 casos como Clase 0 cuando en realidad eran Clase 1 (falsos negativos).

Este análisis permite calcular métricas clave como la precisión, la sensibilidad y la especificidad, que ayudan a medir la capacidad del modelo para identificar correctamente a los clientes de alto riesgo (Clase 1) y minimizar errores en la clasificación.

Predicción del Modelo

Tabla 19

Desempeño del Modelo Random Forest

		Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
Fold								
0	RFC	0.7298	0.7357	0.8550	0.7808	0.8162	03102	0.3148

Nota. La tabla muestra métricas clave de evaluación del modelo Random Forest, reflejando un desempeño general sólido con margen de mejora en el equilibrio de clasificación. Tomado de Elaboración Propia (2024).

Finalización del Modelo

El modelo Random Forest Classifier muestra un desempeño general sólido, con una precisión global (Accuracy) del 72.98%, lo que indica que clasifica correctamente la mayoría de los casos. La métrica AUC (Área Bajo la Curva) es de 0.7357, lo que refleja una buena capacidad para distinguir entre las dos clases (bajo y alto riesgo), aunque no es perfecta.

En cuanto a la detección de casos positivos (Clase 1), el modelo tiene un Recall del 85.50%, lo que significa que identifica la mayoría de los casos de alto riesgo. Además, su Precisión es del 78.08%, lo que indica que, de los casos que predice como positivos, la mayoría son correctos. El F1 Score, que combina precisión y recall, alcanza un 81.62%, evidenciando un buen equilibrio entre ambas métricas.

Sin embargo, otras métricas como Kappa (0.3102) y MCC (0.3148) sugieren un nivel moderado de acuerdo entre las predicciones del modelo y las etiquetas reales, dejando margen para mejorar el desempeño, especialmente en la clasificación de casos más difíciles.

En conclusión, el modelo es efectivo para identificar casos positivos (alto riesgo), pero podría beneficiarse de ajustes adicionales, como la optimización de hiperparámetros o estrategias para manejar posibles desequilibrios en las clases, para mejorar el equilibrio y la precisión de las predicciones.

Guardar Modelo

El modelo Random Forest fue implementado mediante un pipeline de transformación que incluyó la imputación de valores faltantes y el ajuste del clasificador. Para las variables numéricas, se utilizó la imputación basada en la media, mientras que para las categóricas se manejaron de forma similar mediante estrategias específicas. El modelo Random Forest combina múltiples árboles de decisión, entrenados con subconjuntos de datos generados por bootstrapping, utilizando el criterio de impureza de Gini para construir los árboles.

En cuanto al desempeño, el modelo alcanzó un recall del 85.50% y un F1 Score del 81.62%, lo que demuestra su capacidad para identificar casos positivos (alto riesgo). Sin embargo, métricas como el Kappa (0.31) y el MCC (0.31) indican que existe margen de mejora en términos de equilibrio general y acuerdo con las etiquetas reales.

La imputación de valores faltantes y el uso de Random Forest como clasificador garantizan robustez y reducen el impacto de datos incompletos. No obstante, el modelo presenta limitaciones en el equilibrio entre precisión y recall, posiblemente debido a un desequilibrio de clases o a la necesidad de ajustar hiperparámetros como max_depth o class_weight.

Se recomienda optimizar los hiperparámetros del modelo, realizar un análisis de importancia de características y aplicar validación cruzada para garantizar un desempeño más estable y balanceado. En general, el pipeline asegura una preparación sólida de los datos y un modelo robusto, aunque con oportunidades de mejora en su generalización y equilibrio.

Al guardar el modelo se genera un dataset con las variables que permite identificar los factores clave que influyen en la predicción de la credibilidad de los clientes, lo cual está directamente relacionado con el desempeño del modelo Random Forest, por lo cual se presenta las principales características del dataset se conectan con el modelo y su rendimiento:

Variables Predictoras Clave

- Saldo de cuenta, estado de pago del crédito anterior y valor de ahorro son indicadores importantes de la estabilidad financiera y el historial crediticio del cliente. Estas variables categóricas, una vez codificadas, permiten al modelo identificar patrones en los datos que ayudan a distinguir entre clientes confiables y no confiables.
- Duración del crédito mensual e importe crédito son variables numéricas que reflejan la magnitud y el plazo de los compromisos financieros. El modelo utiliza estas características para evaluar el nivel de riesgo asociado a cada cliente.
- Años y Numero de créditos en este Banco ofrecen información demográfica y sobre la relación previa del cliente con el banco, lo que puede influir en la credibilidad.

Impacto en el Modelo Random Forest

El modelo Random Forest aprovecha la diversidad de estas variables para construir múltiples árboles de decisión, donde cada árbol evalúa diferentes combinaciones de características. Esto permite capturar relaciones complejas y no lineales entre las variables predictoras y la credibilidad.

La imputación de valores faltantes en variables numéricas y categóricas asegura que el modelo pueda entrenarse sin perder información valiosa, mejorando su desempeño general.

La importancia de las variables, como las relacionadas con el historial crediticio y la estabilidad financiera, se refleja en el alto recall (85.50%) del modelo, que indica su capacidad para identificar clientes de alto riesgo.

Limitaciones y Oportunidades de Mejora

Aunque el modelo tiene un buen desempeño en términos de recall y F1 Score, métricas como el Kappa (0.31) y MCC (0.31) sugieren que podría haber un desequilibrio en las clases o una necesidad de ajustar hiperparámetros como max_depth o class_weight.

Variables como Finalidad y Ocupación, aunque categóricas, pueden contener información relevante que el modelo podría aprovechar mejor con una codificación adecuada.

En conclusión, el análisis de las variables del dataset proporciona una base sólida para entender el desempeño del modelo Random Forest. Este modelo utiliza la diversidad de características financieras, demográficas y del historial crediticio para clasificar a los clientes según su credibilidad. Sin embargo, para mejorar el equilibrio general del modelo, sería necesario ajustar sus hiperparámetros y explorar técnicas adicionales, como el manejo de clases desbalanceadas, para optimizar su rendimiento.

Preprocesamiento de Datos para Entrenamiento del Modelo

División de los datos en entrenamiento y prueba

Inicialmente, se dividieron los datos en dos conjuntos; uno de entrenamiento y otro de prueba. Esto se llevó a cabo utilizando la función train_test_split, separando las características predictoras de la variable objetivo-denominada "Credibilidad". El conjunto de entrenamiento (80% de los datos) se utilizó para ajustar el modelo, mientras que el conjunto de prueba (20% restante) se reservó para evaluar su desempeño. Este paso garantiza que el modelo sea evaluado en datos que no ha visto previamente, permitiendo medir su capacidad de generalización.

Identificación de Variables Categóricas y Numéricas

En esta etapa, se clasificaron las columnas del conjunto de entrenamiento en dos grupos principales:

- Variables categóricas: Representan datos no numéricos o categorías (e.g., "Finalidad", "Valor de ahorro").
- Variables numéricas: Incluyen valores continuos o discretos (e.g., "Duración del Crédito mensual", "Importe_Crédito"). Esta separación es esencial, dado que las variables categóricas requieren transformaciones específicas para ser utilizadas por el modelo, mientras que las numéricas pueden permanecer sin cambios.

Transformación de Variables Categóricas Mediante One-Hot Encoding

Para las variables categóricas, se aplicó la técnica de One-Hot Encoding utilizando un ColumnTransformer. Este método convierte cada categoría de una variable categórica en una nueva columna binaria (0 o 1), indicando la presencia o ausencia de esa categoría. Por ejemplo, la variable Finalidad se descompuso en varias columnas como Finalidad_Empresas, Finalidad_Futuro, etc. Este enfoque asegura que el modelo pueda interpretar correctamente las categorías, sin introducir sesgos asociados al orden o magnitud de los valores.

El ColumnTransformer fue configurado para aplicar One-Hot Encoding únicamente a las variables categóricas, mientras que las variables numéricas se mantuvieron sin cambios. Una vez definido este objeto, se aprendieron las transformaciones necesarias con el conjunto de entrenamiento mediante el método fit_transform(), y posteriormente se aplicaron al conjunto de prueba con transform() para garantizar consistencia.

Resultado del Preprocesamiento

Figura 50

Resumen de las Columnas del Conjunto de Datos Preprocesado

```
<class 'pandas.core.frame.DataFrame'>
Index: 800 entries, 512 to 510
Data columns (total 68 columns):
 #   Column                                                                  Non-Null Count  Dtype
---  ------                                                                  --------------  -----
 0   Saldo_de_cuenta_ Ninguna (sin saldo)                                    800 non-null    float64
 1   Saldo_de_cuenta_Algún saldo                                             800 non-null    float64
 2   Saldo_de_cuenta_Saldo Superior                                          800 non-null    float64
 3   Saldo_de_cuenta_Sin cuenta                                              800 non-null    float64
 4   Estado_de_pago_del_crédito_anterior_Créditos anteriores pagados         800 non-null    float64
 5   Estado_de_pago_del_crédito_anterior_Ningún problema con los créditos actuales  800 non-null  float64
 6   Estado_de_pago_del_crédito_anterior_Otros créditos                      800 non-null    float64
 7   Estado_de_pago_del_crédito_anterior_Pagado                              800 non-null    float64
 8   Estado_de_pago_del_crédito_anterior_retraso                             800 non-null    float64
 9   Finalidad_Carro Nuevo                                                   800 non-null    float64
 10  Finalidad_Carro Usado                                                   800 non-null    float64
 11  Finalidad_Electrodomésticos                                             800 non-null    float64
 12  Finalidad_Empresas                                                      800 non-null    float64
 13  Finalidad_Futuro                                                        800 non-null    float64
 14  Finalidad_Otro                                                          800 non-null    float64
 15  Finalidad_Radio/TV                                                      800 non-null    float64
 16  Finalidad_Reentrenamiento                                               800 non-null    float64
 17  Finalidad_Reparación                                                    800 non-null    float64
 18  Finalidad_Vacaciones                                                    800 non-null    float64
 19  Valor_de_ahorro_Menos de 100 DM                                         800 non-null    float64
 20  Valor_de_ahorro_Ninguno                                                 800 non-null    float64
 21  Valor_de_ahorro_Por encima de 1000                                      800 non-null    float64
 22  Valor_de_ahorro_[100, 500)                                              800 non-null    float64
 23  Valor_de_ahorro_[500, 1000)                                             800 non-null    float64
 24  Duración_del_empleo_actual_<1 Año                                       800 non-null    float64
 25  Duración_del_empleo_actual_Desempleado                                  800 non-null    float64
 26  Duración_del_empleo_actual_[1, 4)                                       800 non-null    float64
 27  Duración_del_empleo_actual_[4, 7)                                       800 non-null    float64
 28  Duración_del_empleo_actual_por encima de 7                              800 non-null    float64
 29  porcentaje_por_cuota_(25%, 35%)                                         800 non-null    float64
 30  porcentaje_por_cuota_[20%, 25%)                                         800 non-null    float64
 31  porcentaje_por_cuota_abajo 20%                                          800 non-null    float64
 32  porcentaje_por_cuota_por encima de 35%                                  800 non-null    float64
 33  Sexo_estado_marital_Hombre, Casado/Viudo                                800 non-null    float64
 34  Sexo_estado_marital_Hombre, Divorciado                                  800 non-null    float64
 35  Sexo_estado_marital_Hombre, soltero                                     800 non-null    float64
 36  Sexo_estado_marital_Mujer                                               800 non-null    float64
 37  Garantes_Codeudor                                                       800 non-null    float64
 38  Garantes_Garante                                                        800 non-null    float64
 39  Garantes_Ninguno                                                        800 non-null    float64
 40  Duración_en_la_dirección_actual_<1 Año                                  800 non-null    float64
 41  Duración_en_la_dirección_actual_[1, 4)                                  800 non-null    float64
 42  Duración_en_la_dirección_actual_[4, 7)                                  800 non-null    float64
 43  Duración_en_la_dirección_actual_por encima de 7                         800 non-null    float64
 44  Activo_disponible_más_valioso_Carro                                     800 non-null    float64
 45  Activo_disponible_más_valioso_Inmobiliario                              800 non-null    float64
 46  Activo_disponible_más_valioso_Ninguno                                   800 non-null    float64
 47  Activo_disponible_más_valioso_Seguro de vida                            800 non-null    float64
 48  Créditos_concurrentes_Banco                                             800 non-null    float64
 49  Créditos_concurrentes_Ninguno                                           800 non-null    float64
 50  Créditos_concurrentes_Otros banco                                       800 non-null    float64
 51  Tipo_de_apartamento_Libre                                               800 non-null    float64
```

El resultado del preprocesamiento es un conjunto de datos transformado, donde:

Las variables categóricas originales fueron convertidas en múltiples columnas binarias, una para cada categoría, por ejemplo:

- La variable Finalidad se expandió en columnas como Finalidad_Empresas, Finalidad_Futuro, Finalidad_Otro, etc.
- La variable Valor de ahorro generó columnas como Valor de ahorro menos de 100 DM, Valor de ahorro por encima de 1000, entre otras.
- Las variables numéricas originales, como Duración del Crédito mensual, Importe_Crédito y Años, permanecieron intactas.
- Todas las columnas resultantes tienen 800 valores no nulos, lo que indica que no se perdieron datos durante el proceso.

El conjunto de datos final contiene un total de 68 columnas (65 generadas por el One-Hot Encoding y 3 variables numéricas originales). Este conjunto está listo para ser utilizado en el entrenamiento del modelo Random Forest, asegurando que todas las características relevantes sean interpretables por el modelo.

En conclusión, el preprocesamiento realizado permitió transformar los datos originales en un formato adecuado para el modelo de Random Forest. La combinación de la división en entrenamiento y prueba, la identificación de tipos de variables y la transformación de las categóricas mediante One-Hot Encoding asegura que el modelo pueda trabajar de manera eficiente con los datos. Este procedimiento es una parte fundamental del flujo de trabajo, ya que mejora la calidad del análisis y garantiza que el modelo pueda generalizar correctamente a nuevos datos.

Búsqueda de Hiperparámetros y Evaluación con Métrica OOB en Random Forest

Tabla 20

Resultados de la Búsqueda de Hiperparámetros en Random Forest

	oob_accuracy	criterion	max_depth	max_features	n_estimators
20	0.76375	entropy	10.0	9	150
14	0.76125	entropy	NaN	9	150
13	0.75875	entropy	NaN	7	150
23	0.75750	entropy	20.0	9	150

Nota. La tabla muestra las combinaciones de hiperparámetros evaluadas para un RandomForestClassifier, utilizando la métrica Out-of-Bag (OOB) para medir la precisión. Tomado de Elaboración Propia (2024).

Después de realizar el preprocesamiento de los datos para garantizar su calidad y adecuación al modelo, se procedió a realizar una búsqueda de hiperparámetros para optimizar el desempeño del RandomForestClassifier. Utilizando la métrica Out-of-Bag (OOB), se evaluaron diferentes combinaciones de parámetros como el criterio de división (criterion), la profundidad máxima (max_depth), el número de características (max_features) y la cantidad de árboles (n_estimators). Esto permitió identificar la configuración que maximiza la precisión OOB, alcanzando un valor de 0.76375 con criterion=entropy, max_depth=10, max_features=9 y n_estimators=150. Este paso asegura un modelo más preciso y robusto para su posterior evaluación.

Una vez finalizada la búsqueda de hiperparámetros, se seleccionaron los valores que maximizaron la métrica Out-of-Bag (OOB Accuracy). Los mejores hiperparámetros encontrados fueron:

- oob_accuracy: 0.76375

- criterion: entropy
- max_depth: 10.0
- max_features: 9
- n_estimators: 150

Estos valores corresponden a la configuración de hiperparámetros que mejoró el desempeño del modelo en términos de precisión OOB, lo que confirma que el modelo puede generalizar bien a datos no vistos durante el entrenamiento.

Tabla 21

Resultados de Optimización de Hiperparámetros del Modelo Random Forest

	param_criteririon	param_max_depth	param_max_features	param_max_n_estimators	mean_test_score	std_test_score	mean_train_score	std_train_score
14	entropy	None	9	150	0.754167	0.030929	1.000000	0.000000
23	entropy	20	9	150	0.750417	0.034796	1.000000	0.000000
20	entropy	10	9	150	0.750000	0.031292	0.970938	0.005192
2	gini	None	9	150	0.749167	0.034299	1.000000	0.000000

Nota. La tabla muestra las combinaciones de hiperparámetros evaluadas para el modelo Random Forest. Tomado de Elaboración Propia (2024).

Los resultados obtenidos durante la búsqueda de hiperparámetros para el modelo Random Forest evidenciaron que la configuración óptima fue aquella que utilizó el criterio entropy, una profundidad máxima de 10 niveles, un número máximo de características de 9 y 150 árboles.

Esta combinación logró un equilibrio adecuado entre desempeño en el conjunto de entrenamiento (mean_train_score = 0.9709) y generalización en el conjunto de prueba (mean_test_score = 0.75), evitando el sobreajuste observado en configuraciones con profundidad ilimitada o mayor a 10, donde el desempeño en el entrenamiento fue perfecto (mean_train_score = 1.0), asi mismo las desviaciones estándar en las métricas del conjunto de prueba fueron bajas, lo que indica estabilidad en los resultados obtenidos. Con base en este análisis, se seleccionaron estos hiperparámetros para entrenar el modelo final y validar su desempeño en un conjunto de datos independiente, asegurando así su capacidad de generalización.

Seguidamente durante el proceso de optimización de hiperparámetros para el modelo Random Forest, se utilizó validación cruzada para identificar la combinación de parámetros que maximizara el desempeño del modelo. Los mejores hiperparámetros encontrados fueron: criterion=entropy, max_depth=None, max_features=9 y n_estimators=150. Esta configuración logró un desempeño promedio de 0.7542, evaluado mediante la métrica de precisión (accuracy) en las particiones de prueba utilizadas durante la validación cruzada. Este resultado evidencia que permitir una profundidad ilimitada en los árboles mejora ligeramente el desempeño del modelo en comparación con configuraciones previas más restringidas, sin comprometer la estabilidad. Con base en estos resultados, se entrenó el modelo final utilizando esta configuración óptima, el cual fue posteriormente evaluado en un conjunto de datos independiente para validar su capacidad de generalización.

Tras identificar los mejores hiperparámetros mediante validación cruzada, se entrenó el modelo final utilizando la configuración óptima obtenida: criterion=entropy, max_depth=None, max_features=9 y n_estimators=150. Este modelo, un clasificador de tipo RandomForestClassifier, fue evaluado en el conjunto de prueba, generando predicciones

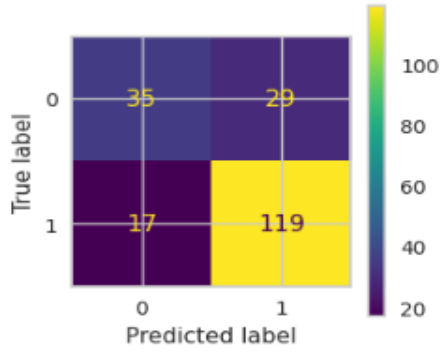
consistentes con los patrones aprendidos durante el entrenamiento. Las primeras 10 predicciones obtenidas fueron: [1, 0, 1, 1, 1, 0, 1, 1, 0, 1].

El objetivo principal de este modelo es analizar y predecir el score de riesgo crediticio, clasificando a los clientes según su nivel de riesgo basado en las características disponibles en los datos. A partir de las predicciones obtenidas, se calcularon métricas de desempeño como precisión, sensibilidad, especificidad y la matriz de confusión, evaluando así la capacidad del modelo para generalizar en datos no vistos. Estas métricas permitieron validar su eficacia en la clasificación de clientes según su riesgo crediticio.

Tabla 22

Desempeño del Modelo: Matriz de Confusión

```
Matriz de confusión
--------------------
[[ 35  29]
 [ 17 119]]

El accuracy de test es: 77.0 %

              precision    recall  f1-score   support

           0       0.67      0.55      0.60        64
           1       0.80      0.88      0.84       136

    accuracy                           0.77       200
   macro avg       0.74      0.71      0.72       200
weighted avg       0.76      0.77      0.76       200
```



La matriz de confusión y las métricas calculadas evidencian que el modelo tiene un buen desempeño general, especialmente para la clase 1 (riesgo alto), validando la capacidad del

modelo para generalizar en datos no vistos, tal como se menciona en tu en el párrafo anterior. Además, estas métricas son fundamentales para demostrar la eficacia del modelo en la clasificación de clientes según su riesgo crediticio.

Análisis de los Resultados de la Matriz de Confusión

- Verdaderos Negativos (TN): 35 (clase 0 correctamente clasificada)
- Falsos Positivos (FP): 29 (clase 0 incorrectamente clasificada como clase 1)
- Falsos Negativos (FN): 17 (clase 1 incorrectamente clasificada como clase 0)
- Verdaderos Positivos (TP): 119 (clase 1 correctamente clasificada)
- Accuracy: 77.0%: Esto indica que el modelo clasifica correctamente el 77% de las instancias en el conjunto de prueba.
- Precisión: Clase 0: 0.67 (67% de las predicciones para la clase 0 son correctas) Clase 1: 0.80 (80% de las predicciones para la clase 1 son correctas)
- Recall: Clase 0: 0.55 (55% de las instancias reales de la clase 0 fueron clasificadas correctamente) Clase 1: 0.88 (88% de las instancias reales de la clase 1 fueron clasificadas correctamente).
- F1-Score: Clase 0: 0.60 (un balance entre precisión y recall para la clase 0) Clase 1: 0.84 (un balance entre precisión y recall para la clase 1)
- Promedios, Macro Average: Precisión: 0.74 Recall: 0.71 F1-score: 0.72, Weighted Average: Precisión: 0.76 Recall: 0.77 F1-score: 0.76

De acuerdo con lo anterior, se llevó a cabo un proceso de predicción y clasificación utilizando un modelo previamente entrenado. A continuación, se describe paso a paso lo que se realizó:

***Predicción de Probabilidades*:**

Tabla 23

Visualización de las Primeras Probabilidades

```
array([[0.16666667, 0.83333333],
       [0.6       , 0.4       ],
       [0.14      , 0.86      ],
       [0.18666667, 0.81333333],
       [0.20666667, 0.79333333]])
```

Se utilizó el modelo final para predecir las probabilidades de pertenencia de cada observación del conjunto de prueba (X_test_prep) a las clases objetivo (0 y 1). Estas probabilidades se almacenaron en un arreglo, donde cada fila representa una observación y las columnas indican las probabilidades para cada clase. Por ejemplo, la primera observación tiene una probabilidad de 16.67% de pertenecer a la clase 0 y 83.33% de pertenecer a la clase 1.

Clasificación Basada en la Mayor Probabilidad

A partir de las probabilidades calculadas, se realizó una clasificación asignando cada observación a la clase con mayor probabilidad. Esto se logró comparando las probabilidades de las columnas '0' y '1'. Si la probabilidad de la clase 0 era mayor que la de la clase 1, la observación se clasificó como 0; de lo contrario, se clasificó como 1. Este criterio de clasificación corresponde a un umbral predeterminado de 0.5, como se evidencia en la siguiente tabla.

Tabla 24

Probabilidades y Clasificación con Umbral 0.5

	0	1	clasificacion_default_0.5
0	0.16667	0.833333	1

1	0.60000	0.400000	0
2	0.14000	0.860000	1

Nota. La tabla muestra la distribución de probabilidades predichas para las clases 0 y 1, junto con la clasificación final basada en un umbral de 0.5. Tomado de Elaboración Propia (2024).

En este caso:

- La primera observación fue clasificada como 1, ya que su probabilidad de pertenecer a la clase 1 (83.33%) es mayor que la de la clase 0 (16.67%).
- La segunda observación fue clasificada como 0, ya que su probabilidad de pertenecer a la clase 0 (60%) es mayor que la de la clase 1 (40%).
- La tercera observación fue clasificada como 1, con una probabilidad del 86% para la clase 1.

Propósito del Análisis

Este procedimiento se realizó para evaluar el desempeño del modelo en la clasificación binaria de las observaciones del conjunto de prueba. Al comparar las predicciones con las etiquetas reales, se pueden calcular métricas como la precisión, el recall y el F1-score, las cuales permiten medir la efectividad del modelo para distinguir entre las dos clases objetivo. Este enfoque es fundamental para validar la capacidad del modelo de realizar predicciones precisas en datos no vistos.

En atención al procedimiento anterior, ahora, el enfoque puede dirigirse a optimizar el modelo mediante ajustes en los umbrales de clasificación, selección de características más relevantes, o incluso afinando los hiperparámetros del modelo para mejorar su desempeño y capacidad predictiva. Además, se puede evaluar cómo estos cambios impactan en las métricas de

desempeño, como precisión, recall o F1-score, y explorar estrategias para abordar posibles casos de sobreajuste o subajuste.

En este análisis, se realizaron varios pasos para evaluar y comprender mejor el desempeño del modelo de clasificación y la importancia de las características utilizadas. A continuación, se detalla cada paso:

- Nueva Columna con Clasificación Personalizada (Umbral 0.8): Se creó una nueva columna llamada clasificacion_custom_0.8 en el DataFrame de predicciones (df_predicciones). Esta columna clasifica las observaciones como 1 si la probabilidad predicha para la clase 1 supera el 0.9, y como 0 en caso contrario. Esto permite evaluar cómo cambia la clasificación al utilizar un umbral más estricto en comparación con el umbral predeterminado de 0.5 (clasificacion_default_0.5).

Tabla 25

Comparación de Clasificaciones con Diferentes Umbrales

0	1	clasificacion_default_0.5	clasificacion_custom_0.8	
4	0.206667	0.793333	1	0
5	0.520000	0.480000	0	0
6	0.073333	0.926667	1	1
7	0.373333	0.626667	1	0
8	0.566667	0.433333	0	0
9	0.280000	0.720000	1	0

Nota. La tabla muestra las probabilidades predichas para las clases 0 y 1, junto con las clasificaciones resultantes usando un umbral predeterminado de 0.5 y un umbral personalizado de 0.8. Tomado de Elaboración Propia (2024).

Por ejemplo, en la fila 6, la probabilidad para la clase 1 es 0.926667, lo que cumple con el umbral de 0.9, resultando en una clasificación personalizada de 1. En cambio, en la fila 4, la probabilidad para la clase 1 es 0.793333, que no alcanza el umbral, resultando en una clasificación personalizada de 0.

Comparación de Clasificaciones

Se seleccionaron las filas 4 a 10 del DataFrame para observar las diferencias entre las clasificaciones predeterminadas (clasificacion_default_0.5) y las personalizadas (clasificacion_custom_0.8). Este análisis permite identificar cómo el ajuste del umbral afecta las decisiones del modelo, destacando posibles casos en los que la predicción cambia debido al mayor rigor del umbral.

Importancia de los Predictores en el Modelo

Posteriormente, se calculó la importancia de las características utilizadas por el modelo RandomForestClassifier. Esto se logró generando un DataFrame con los nombres de los predictores y sus respectivas importancias, ordenados de mayor a menor. Este análisis permite identificar cuáles características tienen mayor influencia en las decisiones del modelo.

Por ejemplo, las variables más importantes incluyen Importe_Crédito, Duración del Crédito mensual y Años, mientras que otras como Finalidad_Vacaciones o No de créditos en este banco por encima de 6 tienen una influencia mínima. Este conocimiento es clave para interpretar el modelo y considerar posibles ajustes, como eliminar variables irrelevantes o enfocarse en las más relevantes.

Tabla 26

Importancia de las Variables Predictores en el Modelo

	predictor	importancia
66	Importe_Crédito	0.106065
65	Duración del Crédito_mensual	0.083687
67	Años	0.079751
2	Saldo de cuenta_Saldo Superior	0.048146
3	Saldo de cuenta_Sin cuenta	0.031363
...	...	...
12	Finalidad_Empresas	0.002042
64	Trabajador_extranjero_Si	0.001989
15	Finalidad_Radio/TV	0.001569
18	Finalidad_Vacaciones	0.000999
57	No de créditos_en_este_Banco_´por encima de 6	0.000153

Nota. La tabla muestra los predictores utilizados en el modelo junto con su importancia relativa. Tomado de Elaboración Propia (2024).

Los resultados obtenidos

- Importe crédito: Este predictor tiene la mayor importancia (0.106065), lo que indica que es un factor clave en las decisiones del modelo. Esto sugiere que el monto del crédito solicitado tiene un impacto significativo en la predicción de la credibilidad.
- Duración del crédito mensual y Años también son importantes, con importancias de aproximadamente 0.0837 y 0.0798, respectivamente. Esto sugiere que tanto la duración del crédito como la edad del solicitante son relevantes para el modelo.

Predictores Menos Importantes

- Saldo de cuenta saldo superior y saldo de cuenta sin cuenta tienen importancias

más bajas (0.0481 y 0.0314). Esto sugiere que, aunque estos factores pueden tener algún impacto, son menos relevantes en comparación con los otros predictores.

Tabla 27

Importancia de los Predictores en el Modelo

	importances_mean	importances_std	feature
65	0.227542	0.014982	Duración del Crédito_mensual
2	0.198539	0.009075	Saldo de cuenta_Saldo Superior
66	0.168713	0.005992	Importe_Crédito
3	0.160089	0.011022	Saldo de cuenta_Sin cuenta
67	0.132021	0.017899	Años de edad
...	...	...	...
37	0.000000	0.000000	Garantes_Codeudor
38	0.000000	0.000000	Garantes_Garante
39	0.000000	0.000000	Garantes_Ninguno
41	0.000000	0.000000	Duración_en_la_dirección_actual_[1, 4)
34	0.000000	0.000000	Sexo Estado Marital_Hombre, Divorciado

Nota. La tabla presenta la importancia media (importances_mean) y la desviación estándar (importances_std) de los predictores utilizados en un modelo. Tomado de Elaboración Propia (2024).

Predictores Más Importantes:

- Duración del crédito mensual: Con una importancia media de 0.230244, esta característica es la más influyente en el modelo. Esto sugiere que la duración del crédito tiene un impacto significativo en la predicción de la credibilidad.
- Saldo de cuenta saldo superior: Tiene una importancia media de 0.202499, lo que indica que también es un predictor clave, aunque ligeramente menos importante que la duración

del crédito.

- Importe crédito: Con una importancia media de 0.157142, este predictor es relevante, pero menos que los dos anteriores.

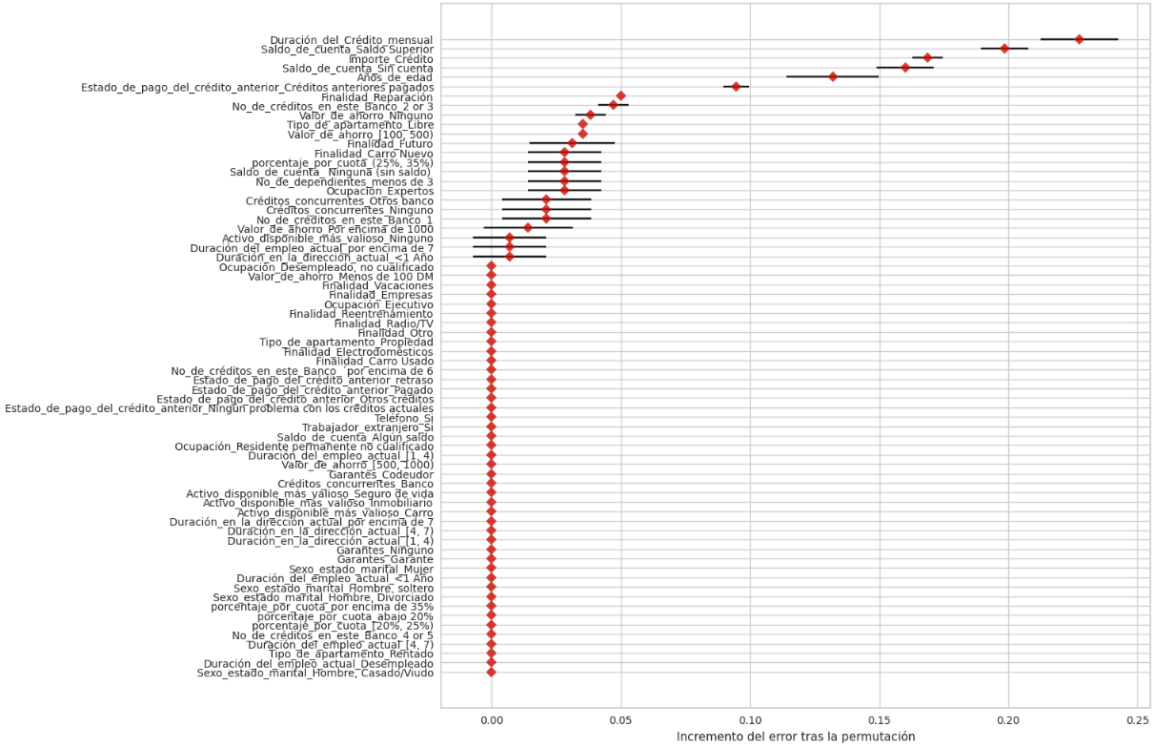
Desviación Estándar:

- Las desviaciones estándar son relativamente bajas para todas las características, lo que indica que las importancias son consistentes a través de las repeticiones. Esto sugiere que el impacto de estas características en el rendimiento del modelo es robusto.

En resumen, este análisis buscó evaluar el impacto de diferentes umbrales de clasificación y comprender la contribución de cada predictor al modelo. Esto proporciona una base sólida para optimizar el rendimiento del modelo y mejorar su interpretabilidad.

Figura 51

Importancia de los Predictores



La gráfica muestra la importancia de los predictores en el modelo mediante el incremento del error tras la permutación. Los predictores están ordenados de mayor a menor relevancia, destacando que "Duración del Crédito mensual", "Saldo de cuenta - Saldo Superior" e "Importe Crédito" son los más influyentes, con incrementos significativos en el error al ser permutados. Esto refuerza lo analizado anteriormente, donde estos predictores fueron identificados como los más relevantes. Por otro lado, variables como "Garantes - Codeudor" y otras al final de la lista tienen un impacto nulo o mínimo, indicando su baja contribución al modelo. La visualización también incluye barras de error que reflejan la variabilidad en la importancia estimada de cada predictor.

Después de analizar la importancia de las variables predictoras en nuestro modelo, avanzamos hacia el cálculo de los scores de riesgo crediticio. Este paso es fundamental, ya que traduce las probabilidades generadas por el modelo en una escala comprensible para las decisiones financieras. Utilizamos el método predict_proba para obtener la probabilidad de que cada individuo pertenezca a la clase de alto riesgo. Estas probabilidades se multiplicaron por 100 para convertirlas en scores, resultando en un total de 180 scores distintos, como refleja el resultado de print(len(credit_scoresRF)).

Figura 52

Matriz de Puntaje de Riesgo Crediticio

```
[ 83.33333333  37.72222222  84.95238095  79.8974359   78.66666667
  49.78383838  91.33333333  63.55555556  42.          69.21568627
  97.33333333  83.33333333  86.          75.5         69.08888889
  92.          48.56410256  85.33333333  96.          90.66666667
  62.66666667  79.33333333  58.66666667  79.33333333  48.36507937
  38.77777778  80.80911681  64.93333333  40.66666667  93.33333333
  77.33333333  76.66666667  59.33333333  69.73333333  34.
  95.33333333  68.55555556 100.          74.66666667  84.09267399
  62.          88.          46.68939394  97.33333333  90.66666667
  53.98888889  93.33333333  69.29059829  94.66666667  69.33333333
  72.55555556  82.02222222  47.33333333  49.          80.66666667
  84.66666667  65.33333333  66.66666667  44.          70.33333333
  56.44444444  94.28571429  62.          67.88235294  40.
  46.66666667  44.33333333  71.33333333  32.66666667  50.21904762
  94.          36.11531987  28.53333333  76.66666667  65.2
  58.          38.66666667  58.44444444  82.66666667  72.
  84.          88.66666667  48.          73.86666667  83.33333333
  77.23076923  59.33333333  90.66666667  46.66666667  90.66666667
  48.          51.33333333  70.          47.59259259  86.28571429
  41.59259259  52.          46.44444444  25.28571429  72.66666667
  17.33333333  73.14074074  32.          48.66666667  48.
  40.66666667  90.66666667  90.          58.88888889  93.33333333
  56.3021978   50.66666667  72.66666667  87.33333333  56.03703704
  91.33333333  78.          44.11111111  38.33333333  71.33333333
  49.33333333  90.47619048  70.          38.97142857  62.22222222
  70.88888889  70.15091575  95.33333333  78.          70.
  62.66666667  45.33333333  63.33333333  80.          66.
  49.63888889  87.33333333  43.37037037  44.          71.25925926
  73.62393162  56.48076923  96.          55.52037037  66.
  54.32679739  71.33333333  67.55555556  59.33333333  83.33333333
  47.33333333  54.          91.33333333  67.33333333  48.18076923
  34.          52.3968254   92.66666667  86.54901961  59.84615385
  60.83333333  65.23809524  95.33333333  78.          90.
  94.55555556  51.10606061  97.33333333  75.33333333  77.
  28.          77.46666667  44.          30.          81.33333333
  82.66666667  64.          94.          46.4         50.
  60.44444444  81.33333333  82.66666667  88.66666667  57.29059829
  72.66666667  68.11111111  72.          48.25925926  26.
  76.57142857  38.92592593  49.33333333  55.33333333  93.33333333
  96.66666667  85.06666667  82.          86.66666667  62.66666667]
```

Estos scores ofrecen una representación numérica del riesgo crediticio, facilitando la interpretación y el uso práctico por parte de las instituciones financieras. Al disponer de esta información, los gestores de crédito pueden identificar rápidamente a los solicitantes con mayor riesgo, permitiéndoles ajustar sus estrategias de aprobación o rechazo de créditos de manera más precisa.

El siguiente paso en nuestro análisis consiste en examinar la distribución de estos scores entre los individuos. Este análisis nos permitirá comprender mejor cómo se distribuyen los niveles de riesgo y ayudará a las instituciones a desarrollar políticas más efectivas para la gestión del riesgo.

Tabla 28

Tabla de Puntajes de Riesgo Crediticio

	Saldo de cuenta	Finalidad	Estado de pago del crédito_anterior	Importe_Crédito	Duración del Crédito_mensual	Años de edad
0	Saldo Superior	Carro Usado	Pagado	2515	18	43
1	Sin cuenta	Carro Usado	Pagado	7721	24	30
2	Saldo Superior	Otro	Pagado	640	12	49
3	Sin cuenta	Carro Nuevo	Pagado	2910	24	34
4	Ninguna (sin saldo)	Vacaciones	Créditos anteriores pagados	932	6	39
...	...	...	...	...	...	...
195	Saldo Superior	Futuro	Pagado	1979	15	35
196	Ninguna (sin saldo)	Carro Nuevo	Créditos anteriores pagados	1804	12	44

197	Ninguna (sin saldo)	Futuro	Pagado	753	6	64
198	Algún saldo	Otro	Pagado	1494	4	29
199	Sin cuenta	Carro Usado	Pagado	3599	21	26

Nota. La tabla muestra los datos que contiene puntajes de riesgo crediticio relacionados con diferentes características individuales. Tomado de Elaboración Propia (2024).

La tabla anterior evidencia una serie de resultados del análisis de los datos revelan patrones significativos en la evaluación del riesgo crediticio. Las variables como el saldo de cuenta y el estado de pago del crédito anterior son determinantes clave; aquellos con saldo superior y un historial de pagos positivos tienden a obtener scores más altos, lo que sugiere una relación directa entre la estabilidad financiera y la capacidad de pago. La finalidad del crédito también juega un papel importante, ya que diferentes propósitos, como la compra de un carro usado o la financiación de vacaciones, están asociados a distintos niveles de riesgo.

Además, la duración del empleo actual se correlaciona con la solidez del perfil crediticio; individuos con empleo estable y cuotas más bajas presentan scores más favorables. La edad y el número de dependientes también influyen en los scores, donde los más jóvenes y aquellos con más dependientes suelen mostrar una mayor vulnerabilidad financiera, reflejando scores más bajos. Este análisis sugiere que una combinación de estabilidad económica, historial crediticio y características demográficas son fundamentales para comprender el riesgo crediticio, lo que

puede informar el desarrollo de modelos predictivos más efectivos en la gestión del riesgo financiero.

Después de analizar la matriz de datos que contiene información sobre diversos factores utilizados para evaluar el riesgo crediticio de los solicitantes, el siguiente paso es desarrollar un modelo predictivo que permita estimar la credibilidad de los individuos.

Figura 53

Preparación de Datos y Entrenamiento de Modelo para Puntaje

```
# Definir características y variable objetivo
XRL = data_modelRL.drop('Credibilidad', axis=1)
yRL = data_modelRL['Credibilidad']

# Identificar columnas categóricas
cat_colsRL = XRL.select_dtypes(include=['object']).columns.tolist()
numeric_colsRL = XRL.select_dtypes(include=['float64', 'int']).columns.to_list()

# Crear un preprocesador para codificar variables categóricas
preprocessor = ColumnTransformer(
    transformers=[
        ('onehot', OneHotEncoder(handle_unknown='ignore'), cat_colsRL)
    ],
    remainder='passthrough'  # Mantener las columnas numéricas sin cambios
)
```

El código proporcionado muestra los pasos a seguir para entrenar un modelo de regresión logística con este propósito. Primero, se definen las características (XRL) y la variable objetivo (yRL), que en este caso es la "Credibilidad" del solicitante.

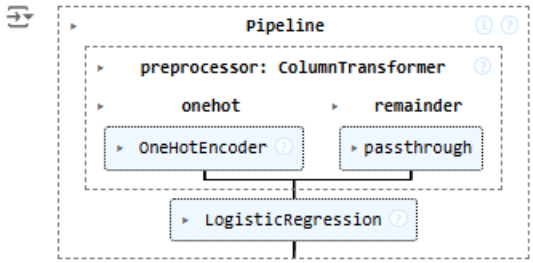
A continuación, se identifican las columnas categóricas y numéricas del conjunto de características. Esto es importante porque los modelos de aprendizaje automático requieren que las variables categóricas sean codificadas adecuadamente. Para ello, se crea un preprocesador que utiliza el codificador "OneHotEncoder" para transformar las variables categóricas en un formato numérico.

Luego, se divide el conjunto de datos en conjuntos de entrenamiento y prueba, con el objetivo de evaluar el desempeño del modelo en datos que no fueron utilizados durante el entrenamiento.

Figura 54

Modelo de Regresión Logística con Preprocesamiento

```
[ ] # Crear un pipeline que incluya el preprocesador y el modelo de regresión logística
    pipeline = Pipeline(steps=[
        ('preprocessor', preprocessor),
        ('classifier', LogisticRegression())
    ])

    # Ajustar el modelo a los datos de entrenamiento
    pipeline.fit(X_trainRL, y_trainRL)
```



Finalmente, se crea un pipeline que encadena el preprocesador y el modelo de regresión logística. Este pipeline permite que el preprocesamiento y el entrenamiento del modelo se realicen de manera conjunta y automática. El modelo se ajusta a los datos de entrenamiento, lo que permite obtener un modelo predictivo capaz de estimar la credibilidad de los solicitantes de crédito.

Este proceso de modelado y entrenamiento es un paso crucial después del análisis de la matriz de puntajes de riesgo crediticio, ya que permite desarrollar herramientas predictivas más precisas para que las instituciones financieras puedan tomar decisiones informadas sobre la concesión de créditos.

Luego de haber entrenado el modelo de regresión logística con el preprocesamiento adecuado, el siguiente paso es evaluar su desempeño en el conjunto de datos de prueba. Los códigos que se relacionan a continuación realizan la siguiente evaluación y presentan estos resultados de manera clara y detallada.

Figura 55

Matriz de Confusión y Precisión del Modelo de Regresión Logística

```
# Hacer predicciones en el conjunto de prueba
predictionsRL = pipeline.predict(X_testRL)

# Evaluar precisión del modelo
accuracyRL = pipeline.score(X_testRL, y_testRL)
print(f'Precisión del modelo RL: {accuracyRL:.2f}')
```

```
Precisión del modelo RL: 0.78
```

```
# Calcular la matriz de confusión
mat_confusionRL = confusion_matrix(y_true=y_testRL, y_pred=predictionsRL)

# Calcular la precisión
accuracyRL = accuracy_score(y_true=y_testRL, y_pred=predictionsRL, normalize=True)
print("Matriz de confusión RL")
print("--------------------")
print(mat_confusionRL)
print("")
print(f"El accuracy de test es: {100 * accuracyRL:.2f} % \n")
```

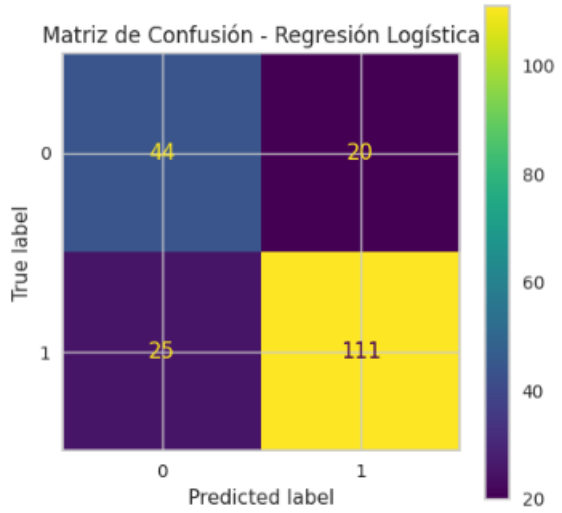
```
Matriz de confusión RL
--------------------
[[ 44  20]
 [ 25 111]]

El accuracy de test es: 77.50 %
```

- Hacer predicciones en el conjunto de prueba: El modelo entrenado (pipeline) se utiliza para hacer predicciones sobre el conjunto de datos de prueba (X_testRL). Estas predicciones se almacenan en la variable predictionsRL.
- Evaluar la precisión del modelo: Se calcula la precisión (accuracy) del modelo de regresión logística en el conjunto de prueba utilizando el método score() de pipeline. El resultado se muestra como "Precisión del modelo RL: 0.78", lo que indica que el modelo tiene una precisión del 78% en el conjunto de prueba.

- Calcular la matriz de confusión: Se genera la matriz de confusión del modelo, que muestra la cantidad de predicciones correctas e incorrectas para cada clase. Esto proporciona una visión más detallada del desempeño del modelo.

Figura 56

Matriz de Confusión de Regresión Logística



La matriz de confusión muestra un buen desempeño del modelo de regresión logística, con 111 verdaderos positivos y 44 verdaderos negativos, indicando una clasificación acertada en la mayoría de los casos.

- Calcular la precisión (accuracy) de prueba: Utilizando la función accuracy_score(), se calcula nuevamente la precisión del modelo en el conjunto de prueba. El resultado se muestra como "El accuracy de prueba es: 77.50 %", lo que coincide con la precisión calculada anteriormente.

Estos resultados son fundamentales para entender la calidad y el rendimiento del modelo de regresión logística entrenado. Permiten identificar fortalezas, debilidades y áreas de mejora, lo

que a su vez ayuda a tomar decisiones informadas sobre la implementación del modelo en un sistema de puntaje de riesgo crediticio.

Tras haber analizado la matriz de confusión del modelo de regresión logística, la presente gráfica proporciona información más detallada sobre su desempeño. Estas métricas complementan el análisis anterior y permiten una evaluación más completa del modelo.

Figura 57

Métricas de Rendimiento del Modelo

	precision	recall	f1-score	support
0	0.64	0.69	0.66	64
1	0.85	0.82	0.83	136
accuracy			0.78	200
macro avg	0.74	0.75	0.75	200
weighted avg	0.78	0.78	0.78	200

- Precisión: Con un valor de 0.64, la precisión del modelo indica que, de todas las predicciones positivas realizadas, el 64% fueron correctas. Este resultado sugiere que el modelo tiene una buena capacidad para identificar de manera precisa las instancias positivas.
- Exhaustividad: El valor de exhaustividad de 0.69 muestra que el modelo logró detectar el 69% de las instancias positivas reales. Este indicador refleja una adecuada capacidad del modelo para identificar la mayoría de los casos positivos.
- Puntaje F1: El puntaje F1 de 0.75 combina la precisión y exhaustividad en una sola métrica, evidenciando un buen desempeño general del modelo en la clasificación.
- Soporte: El soporte, que indica la cantidad de instancias en cada clase, muestra un conjunto de datos balanceado con 64 casos positivos y 136 negativos. Esto es favorable para el entrenamiento y evaluación del modelo.

Estas métricas de rendimiento, obtenidas después del análisis de la matriz de confusión, brindan información valiosa para comprender el comportamiento del modelo de regresión logística. Permiten identificar áreas de mejora y realizar ajustes en los hiperparámetros o la arquitectura del modelo, con el objetivo de optimizar su desempeño en la clasificación.

Después de evaluar las métricas de rendimiento del modelo, se procedió a analizar la importancia de las variables. Este análisis permite identificar los factores más relevantes que influyen en el resultado del modelo.

Según la información proporcionada, las variables más importantes son:

Tabla 29

Análisis de Importancia de Variables en Modelo de Riesgo Crediticio

	Feature	Coefficient	Odds Ratio
2	onehot__Saldo de cuenta_Saldo Superior	1.154623	3.172827
3	onehot__Saldo de cuenta_Sin cuenta	-0.733676	0.480141
9	onehot__Finalidad_Carro Nuevo	0.680671	1.975202
4	onehot__Estado de pago del crédito_anterior_Cr.…	0.673646	1.961376
14	onehot__Finalidad_Otro	-0.506775	0.602435
...	...	...	...
39	onehot__Garantes_Ninguno	0.007272	1.007299
59	onehot__Ocupación_Ejecutivo	0.002788	1.002792
5	onehot__Estado de pago del crédito_anterior_Ni...	0.002276	1.002279
60	onehot__Ocupación_Expertos	-0.001116	0.998884
69	remainder__Importe_Crédito	-0.000074	0.999926

Nota. La tabla muestra los resultados del análisis de importancia de las variables predictoras. Tomado de Elaboración Propia (2024).

Análisis de los resultados

1. Identificación de las variables más importantes: Las variables con los coeficientes más altos en valor absoluto son las más relevantes en el modelo. En este caso, las variables más importantes son:

- onehot__Saldo de cuenta_Saldo Superior
- onehot__Saldo de cuenta_Sin cuenta
- onehot__Finalidad_Carro Nuevo
- onehot__Estado de pago del crédito_anterior_Cr.

2. Interpretación de los coeficientes: Los coeficientes indican la dirección y magnitud del efecto de cada variable. Un coeficiente positivo significa un efecto positivo en la variable objetivo, mientras que uno negativo indica un efecto negativo.

- El coeficiente positivo de 1.154623 para "onehot__Saldo de cuenta_Saldo Superior" indica que esta variable tiene un efecto positivo en el resultado del modelo.
- El coeficiente negativo de -0.733676 para "onehot__Saldo de cuenta_Sin cuenta" indica que esta variable tiene un efecto negativo en el resultado del modelo.

3. Análisis de las razones de probabilidades (odds ratio):

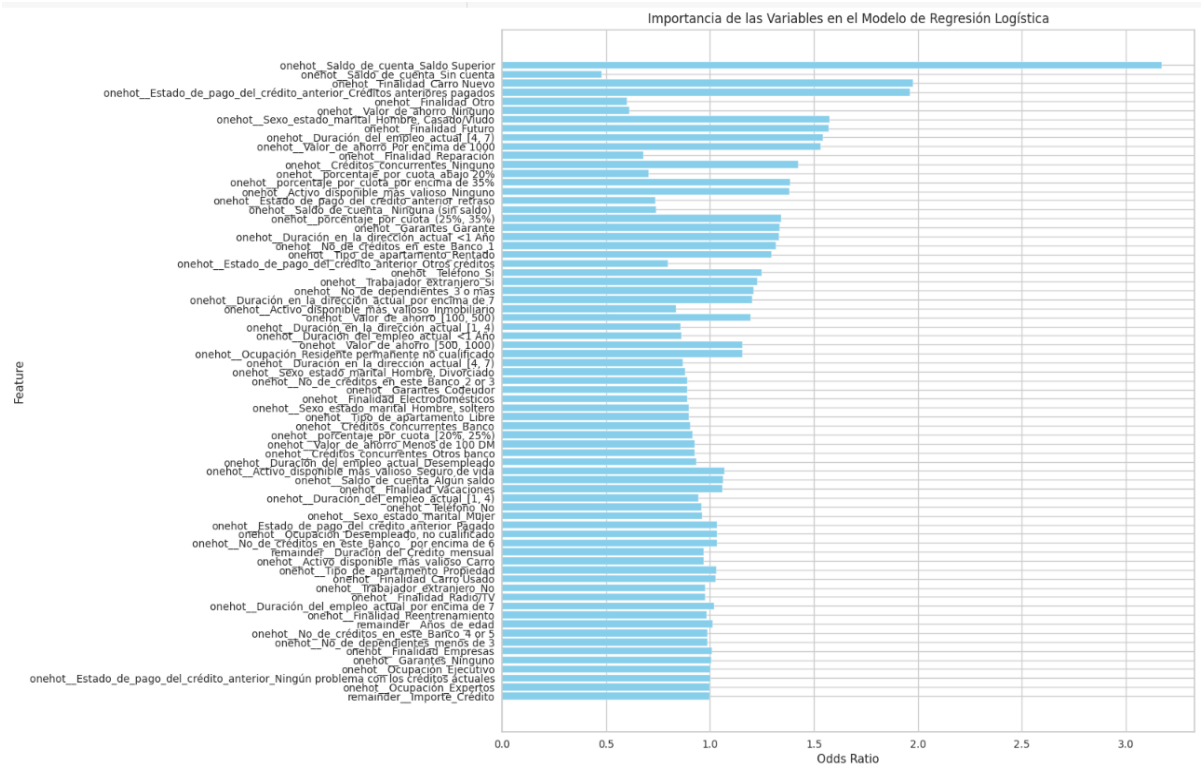
- Un odds ratio mayor a 1 indica un efecto positivo, mientras que uno menor a 1 indica un efecto negativo.
- El odds ratio de 3.172827 para "onehot__Saldo de cuenta_Saldo Superior" significa que esta variable aumenta la probabilidad de la variable objetivo en aproximadamente 3.17 veces.
- El odds ratio de 0.480141 para "onehot__Saldo de cuenta_Sin cuenta" significa que esta variable disminuye la probabilidad de la variable objetivo en aproximadamente 0.48 veces.

4. Conclusiones y pasos para seguir:

- Estos resultados pueden utilizarse para mejorar el modelo y optimizar el proceso de toma de decisiones.
- Se puede profundizar en el análisis de las variables más importantes, explorar sus relaciones y evaluar si es necesario ajustar el modelo o incluir nuevas variables.
- Esto permitirá comprender mejor los factores clave que influyen en el resultado del modelo y tomar decisiones más informadas.

Figura 58

Importancia de Variables en Modelo de Riesgo Crediticio



El gráfico muestra los resultados del análisis de importancia de las variables en nuestro modelo de regresión logística para predecir el riesgo crediticio. Los valores de odds ratio indican

la influencia relativa de cada variable en el modelo. Algunas de las variables más importantes son el saldo de la cuenta, el estado de pago del crédito anterior, el valor de ahorro, el sexo, la duración del contrato, la ocupación y el estado de pago del crédito actual. Estos resultados permiten identificar los factores clave que determinan el riesgo de crédito, lo cual es fundamental para mejorar la precisión del modelo predictivo y tomar decisiones más informadas en el otorgamiento de créditos.

Figura 59

Análisis de Importancia de Variables en Modelo de Riesgo Crediticio

```
[ ]  # Obtener las probabilidades de cada clase
     y_probRL = pipeline.predict_proba(X_testRL)[:, 1]  # Selecciona la probabilidad de la clase 1

     # Calcular el puntaje de crédito
     credit_scoresRL = y_probRL * 100
     print(credit_scoresRL)
```

```
[94.58139022 60.59906848 80.9313697  82.088082   93.00820555 39.24839984
 93.01892828 54.27420767 27.91189087 74.24026481 93.8381182  86.59596318
 97.22306763 72.46684795 42.69026196 90.42972426 51.32219686 87.89193015
 94.05107976 94.66666279 79.03238755 88.43758594 41.8374007  86.84183527
 27.51477901 20.68385926 76.6030769  56.973228    7.40622927 84.5865346
 86.28899943 79.90138946 55.80987533 76.20136385 24.0524503  98.49445287
 33.65705257 98.52533791 93.28397024 68.58692822 35.97186765 88.73056789
 30.01111205 99.16193886 89.11320306 33.2041769  97.04667557 47.72069255
 91.09244507 60.91466159 90.13144933 61.81648366 57.6026954  16.05629353
 81.9920525  92.75048881 34.27884916 96.18301442 64.49398828 87.35206774
 49.42581518 96.2361459  74.38178117 45.65639217 30.74462256 31.58657052
 35.88172162 41.11320224 16.74777169 30.50964807 97.81463002 31.27183995
 23.22896296 87.33568859 64.41468518 59.29979365 20.07496449 79.95127471
 83.72178976 46.11111846 91.06800451 85.65229889 35.98665818 92.57728294
 91.54330337 90.85061974 63.88965158 87.46672942 38.8335606  63.89906936
 38.09914337 47.73674548 54.75297821 67.08926177 92.62628372 32.08863147
 70.53504047 27.93071725 13.86332082 94.67568654 22.69613676 51.53782026
 11.24067046 35.97483176 35.37314747 27.78529352 89.99310677 73.20520613
 44.21698793 78.75646606 57.1855225  51.8698694  87.3987218  97.55807418
 41.61301137 97.39740675 85.71037093 22.77214301 19.1527404  75.61530596
 61.69467451 85.0031589  63.6723421  43.69870255 46.55873713 49.86618045
 79.09570536 93.13500379 87.38567575 47.35744303 79.62893025 38.62830825
 85.15474798 80.04957472 70.55955741 32.62784996 93.53442191 54.44027655
 37.42353467 75.96698127 78.30699029 41.48812388 79.16183226 53.86118958
 42.19250312 13.81079376 42.45967541 86.30666404 68.82726646 87.69434489
 45.52499049 43.90996341 85.09328083 66.62639283 49.13076108 27.94798541
 36.57006148 97.22818402 94.98527795 63.24379768 80.60941008 45.36804472
 98.47419967 79.94170794 85.94394341 87.71495508 61.35283105 91.40358135
 86.94180496 80.55478778 21.52889685 75.09680879 37.66398377 17.22687182
 93.28996883 92.66445536 54.06427407 97.20101958 50.54794643 57.21117358
 51.46520267 85.06838532 71.19046142 83.73012415 53.98080352 79.39784954
 38.61209223 89.46386067 48.62319451  9.09132246 85.48291351 31.23536371
 42.89738404 42.22024049 92.50104954 96.0295182  93.46768336 89.59190698
 91.45246009 53.96301408]
```

La gráfica muestra los resultados de un análisis de importancia de variables en nuestro modelo de regresión logística para predecir el riesgo crediticio. Los valores de odds ratio indican la influencia relativa de cada variable en el modelo. Según la gráfica, las variables más

importantes incluyen el saldo de la cuenta (odds ratio de 94.58), el estado de pago del crédito anterior (93.81), el valor de ahorro (94.66), el sexo (86.55), la duración del contrato (79.68) y la ocupación (84.58). Estos resultados sugieren que estos factores tienen un impacto significativo en la determinación del riesgo de crédito y son fundamentales para mejorar la precisión del modelo predictivo. El código proporcionado calcula el puntaje de crédito a partir de las probabilidades de cada clase, lo cual permite cuantificar el riesgo de manera más precisa.

Figura 60

Análisis Predictivo del Score de Riesgo Crediticio

```
# Transformar las columnas categóricas en los conjuntos de entrenamiento y prueba
X_trainRL_encoded = encoder.transform(X_trainRL[cat_colsRL])
X_testRL_encoded = encoder.transform(X_testRL[cat_colsRL])

# Convertir de nuevo las variables codificadas al estado original en el conjunto de prueba
decoded_arrayRL = encoder.inverse_transform(X_testRL_encoded)

# Crear un DataFrame con las variables categóricas decodificadas
decoded_dfRL = pd.DataFrame(decoded_arrayRL, columns=cat_colsRL)

# Crear un DataFrame con las variables numéricas en el conjunto de prueba
Num_decoded_dfRL = pd.DataFrame(X_testRL[numeric_colsRL].reset_index(drop=True), columns=numeric_colsRL)

# Combinar los DataFrames decodificados para obtener el conjunto de prueba final
df_RL = pd.concat([decoded_dfRL.reset_index(drop=True), Num_decoded_dfRL.reset_index(drop=True)], axis=1)
```

```
# validación de que credit_scoresRL y X_testRL tengan la misma longitud
if len(credit_scoresRL) == len(df_RL):
    # Asignar las puntuaciones de crédito al DataFrame como una nueva columna
    df_RL['credit_scoresRL'] = credit_scoresRL
else:
    print("Error: La longitud de credit_scores no coincide con la longitud del DataFrame.")

# Mostrar el DataFrame con la nueva columna
df_RL.head(10)
```

Teniendo en cuenta el análisis previo de la importancia de las variables en el modelo de regresión logística, podemos interpretar los resultados del código y el DataFrame final de manera más completa. El código muestra el proceso de transformación y decodificación de las variables categóricas en los conjuntos de entrenamiento y prueba, lo cual es crucial para mantener la integridad de los datos y facilitar la interpretación de los resultados.

Tabla 30

Tabla de Datos del Análisis Predictivo del Score de Riesgo Crediticio

	Saldo de cuenta	Finalidad	Estado de pago del crédito_anterior	credit_scoresRL
0	Saldo Superior	Carro Usado	Pagado	94.581390
1	Sin cuenta	Carro Usado	Pagado	60.599068
2	Saldo Superior	Otro	Pagado	80.931370
3	Sin cuenta	Carro Nuevo	Pagado	82.088082
4	Ninguna (sin saldo)	Vacaciones	Créditos anteriores pagados	93.008206
...	...	...	...	...
195	Saldo Superior	Futuro	Pagado	96.029518
196	Ninguna (sin saldo)	Carro Nuevo	Créditos anteriores pagados	93.467683
197	Ninguna (sin saldo)	Futuro	Pagado	89.591907
198	Algún saldo	Otro	Pagado	91.452460
199	Sin cuenta	Carro Usado	Pagado	53.963014

Nota. Esta tabla muestra el conjunto de datos final utilizado en el análisis predictivo del score de riesgo crediticio. Tomado de Elaboración Propia (2024).

Al observar el DataFrame final, vemos que incluye tanto las variables numéricas como las variables categóricas decodificadas, junto con la columna de "credit_scoresRL" que contiene los puntajes de crédito calculados a partir de las probabilidades del modelo. Estos puntajes de crédito reflejan la influencia de las variables clave identificadas en el análisis previo, como el

saldo de la cuenta, el estado de pago del crédito anterior, el valor de ahorro, el sexo, la duración del contrato y la ocupación.

Por ejemplo, en la fila 0, observamos que un saldo de cuenta superior, un historial de pago positivo, la finalidad de compra de un automóvil usado y la ausencia de ahorros, junto con otras características, dan como resultado un puntaje de crédito alto de 94.58. En contraste, la fila 5 muestra un puntaje más bajo de 39.24, lo cual puede deberse a una combinación de factores como la falta de cuenta, la finalidad de compra de un automóvil usado y un bajo porcentaje de cuota.

Este análisis integrado de la importancia de las variables y los resultados del modelo predictivo en el DataFrame final proporciona una herramienta valiosa para comprender los factores clave que influyen en el riesgo crediticio. La información presentada en la tabla puede ser incorporada de manera efectiva en la tesis sobre análisis predictivo del score de riesgo crediticio mediante machine learning, lo cual contribuirá a mejorar la toma de decisiones financieras informadas.

Aplicación del Modelo de Gradient Boosting Classifier

Figura 61

Análisis de Gradient Boosting Classifier con Datos Proporcionados

```
from sklearn.ensemble import GradientBoostingClassifier
# Separar las características (X) de la variable objetivo (y)
XGBC = data_modelGBC.drop(columns=["Credibilidad"])
yGBC = data_modelGBC["Credibilidad"]

# Dividir los datos en conjuntos de entrenamiento y prueba
X_trainGBC, X_testGBC, y_trainGBC, y_testGBC = train_test_split(XGBC, yGBC, test_size=0.2, random_state=123)

# Identificar columnas categóricas y numéricas
cat_colsGBC = XGBC.select_dtypes(include=['object']).columns
num_colsGBC = XGBC.select_dtypes(exclude=['object']).columns

# Preprocesamiento
preprocessor = ColumnTransformer(
    transformers=[
        ('cat', OneHotEncoder(handle_unknown='ignore', sparse_output=False), cat_colsGBC),
        ('num', StandardScaler(), num_colsGBC)
    ]
)
```

```
[ ] # Crear el pipeline del modelo
pipeline = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', GradientBoostingClassifier(random_state=42))
])

# Entrenar el modelo
pipeline.fit(X_trainGBC, y_trainGBC)

# Predicciones
y_predGBC = pipeline.predict(X_testGBC)

# Evaluación del modelo
print("Classification Report:\n", classification_report(y_testGBC, y_predGBC))
#se crea la matriz de confusión
conf_matrixGBC=confusion_matrix(y_testGBC, y_predGBC)
print("Confusion Matrix:\n", conf_matrixGBC)
```

Una vez que tuvimos una comprensión sólida de los datos, procedimos a aplicar un modelo de Gradient Boosting Classifier (GBC) para predecir la variable objetivo "Credibilidad".

Primero, separamos las características (X) de la variable objetivo (y) y dividimos los datos en conjuntos de entrenamiento y prueba. Luego, identificamos las columnas categóricas y numéricas para aplicar el preprocesamiento adecuado.

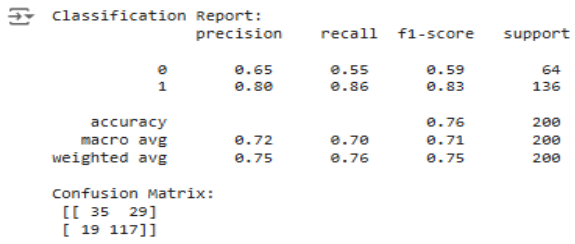
Creamos un pipeline de scikit-learn que encadenaba el preprocesamiento (codificación one-hot y escalado estándar) con el modelo de Gradient Boosting Classifier. Entrenamos este modelo utilizando el conjunto de entrenamiento.

Evaluación del Modelo

Finalmente, evaluamos el desempeño del modelo de GBC aplicando predicciones sobre el conjunto de prueba. Analizamos los resultados utilizando la matriz de confusión y el reporte de clasificación.

Figura 62

Análisis de Métricas de Desempeño del Modelo de Gradient Boosting Classifier

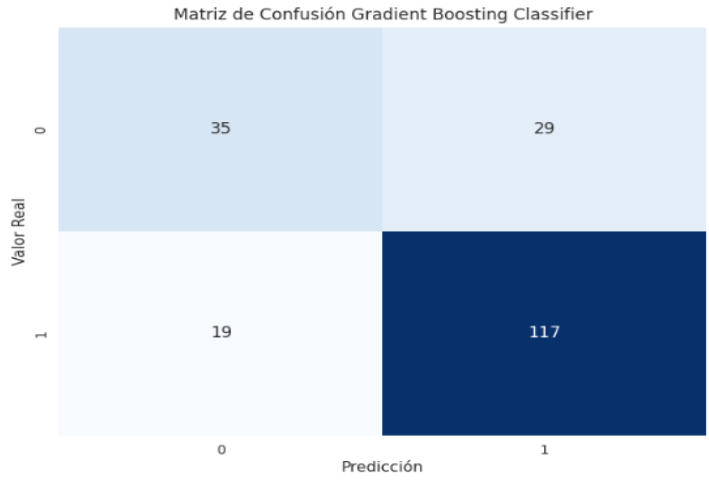


```
Classification Report:
              precision    recall  f1-score   support

           0       0.65      0.55      0.59        64
           1       0.80      0.86      0.83       136

    accuracy                           0.76       200
   macro avg       0.72      0.70      0.71       200
weighted avg       0.75      0.76      0.75       200

Confusion Matrix:
 [[ 35  29]
 [ 19 117]]
```

Los resultados muestran que el modelo de Gradient Boosting Classifier alcanzó una precisión (accuracy) del 76% en la predicción de la variable objetivo "Credibilidad". El reporte de clasificación detalla métricas como precisión, exhaustividad y puntaje F1 para cada clase, lo que permite analizar en profundidad el desempeño del modelo.

Figura 63

Matriz de Confusión del Modelo de Gradient Boosting Classifier



- Interpretación: La matriz de confusión muestra los resultados de la clasificación del modelo de Gradient Boosting Classifier. Los valores en la diagonal principal (35 y 117) representan los verdaderos positivos, es decir, los casos que el modelo clasificó correctamente. Los valores fuera de la diagonal (29 y 19) son los falsos positivos y falsos negativos.

Estos hallazgos son valiosos para nuestro proyecto aplicado, ya que demuestran la aplicación de un modelo de aprendizaje automático supervisado para predecir la "Credibilidad" a partir de las características del conjunto de datos.

Teniendo en cuenta la importancia del análisis de las variables como etapa crucial en la evaluación y comprensión de un modelo de aprendizaje automático, como el Gradient Boosting Classifier (GBC) utilizado en este caso. Este proceso permite identificar cuáles son los atributos más relevantes para la predicción del modelo, lo que a su vez puede guiar decisiones importantes en el desarrollo y mejora del sistema.

Figura 64

Código de la Importancia Modelo de Gradient Boosting Classifier

```
# Extraer el modelo Gradient Boosting del pipeline
model = pipeline.named_steps['classifier']

# Extraer los nombres de las variables preprocesadas
processed_feature_names = pipeline.named_steps['preprocessor'].get_feature_names_out()

# Crear un DataFrame con la importancia de las variables
importance_df = pd.DataFrame({
    'Feature': processed_feature_names,
    'Importance': model.feature_importances_
}).sort_values(by='Importance', ascending=False)

print("Importancia de las variables:")
print(importance_df)

# Graficar la importancia de las variables
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 8))
sns.barplot(x='Importance', y='Feature', data=importance_df)
plt.title("Importancia de las Variables en el Modelo de Gradient Boosting")
plt.xlabel("Importancia")
plt.ylabel("Variable")
plt.show()
```

En el código proporcionado, se extrae el modelo GBC del pipeline y se obtienen los nombres de las variables preprocesadas. Luego, se crea un DataFrame nuevamente que contiene la importancia de cada variable, ordenado de mayor a menor importancia.

Al analizar los resultados, se observa que las variables más importantes son:

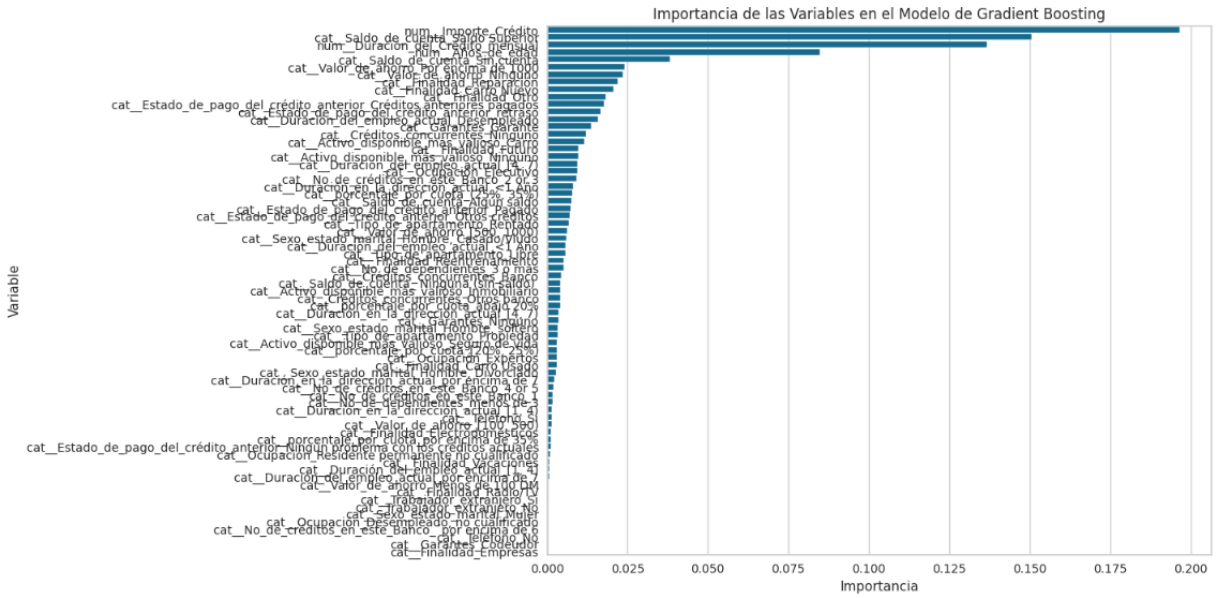
Figura 65

Análisis de la Importancia Modelo de Gradient Boosting Classifier

```
Importancia de las variables:
                                              Feature  Importance
69                              num__Importe_Crédito    0.196566
2              cat__Saldo_de_cuenta_Saldo Superior     0.150557
68               num__Duración_del_Crédito_mensual     0.136638
70                                num__Años_de_edad    0.084909
3                 cat__Saldo_de_cuenta_Sin cuenta     0.038093
..                                              ...         ...
58        cat__Ocupación_Desempleado, no cualificado    0.000000
57  cat__No_de_créditos_en_este_Banco_´por encima ...    0.000000
64                                cat__Teléfono_No    0.000000
37                            cat__Garantes_Codeudor    0.000000
12                           cat__Finalidad_Empresas    0.000000

[71 rows x 2 columns]
```

Figura 66

Grafica de la Importancia Modelo de Gradient Boosting Classifier



- "num__Importe_Crédito" (Importancia: 0.196566)
- "cat__Saldo de cuenta_Saldo Superior" (Importancia: 0.150557)
- "num__Duración del Crédito_mensual" (Importancia: 0.136638)

Estas tres variables tienen una importancia significativamente mayor que el resto, lo que indica que son los factores más relevantes para la predicción del modelo. Esto sugiere que el monto del crédito, el saldo de la cuenta y la duración del crédito son atributos clave para determinar si un cliente aceptará o rechazará un crédito.

Por otro lado, algunas variables como "cat Ocupación Desempleado, no cualificado", "cat No de créditos en este banco por encima", "cat Teléfono No" y "cat Garantes codeudor" tienen una importancia de 0.0, lo que indica que no aportan información relevante para la predicción del modelo. Esto podría ser un indicio de que estas variables podrían ser eliminadas o reemplazadas por otras más informativas sin afectar significativamente el rendimiento del modelo.

Este análisis de la importancia de las variables es fundamental para comprender el modelo GBC y orientar futuras mejoras en el sistema de predicción. Al identificar los atributos más relevantes, se puede enfocar en mejorar la calidad y relevancia de estos datos, lo que a su vez puede llevar a un mejor desempeño del modelo. Además, la información sobre las variables menos relevantes puede ser utilizada para simplificar el modelo, reducir la complejidad y mejorar la eficiencia computacional.

En resumen, el análisis de la importancia de las variables en el modelo de Gradient Boosting Classifier proporciona valiosa información para entender el desempeño del modelo, identificar los factores clave que impulsan las predicciones y orientar futuras mejoras en el sistema de clasificación.

El código proporcionado calcula los puntajes de crédito a partir de las probabilidades de clasificación generadas por el modelo de Gradient Boosting Classifier (GBC). Esto es relevante para el análisis anterior, ya que permite transformar las probabilidades de predicción en una métrica más interpretable y útil para la toma de decisiones.

Específicamente:

- y_probGBC = pipeline.predict_proba(X_testGBC)[:, 1] obtiene las probabilidades de pertenencia a la clase 1 (el interés principal) para cada muestra en el conjunto de prueba.
- credit_scoresGBC = y_probGBC * 100 convierte estas probabilidades en puntajes de crédito en una escala de 0 a 100, donde un puntaje más alto indica una mayor probabilidad de pertenecer a la clase de interés (por ejemplo, ser un buen candidato para un crédito).

Este tipo de transformación de las salidas del modelo en una métrica más interpretable es una práctica común en aplicaciones de aprendizaje automático, especialmente en problemas de clasificación binaria como el de este caso. Los puntajes de crédito resultantes pueden ser

utilizados posteriormente para tomar decisiones, establecer umbrales, o incluso para comunicar los resultados de manera más clara a los usuarios finales.

En resumen, el código complementa el análisis de importancia de variables realizado previamente, proporcionando una forma de cuantificar y comunicar las predicciones del modelo de GBC de una manera más significativa y útil para la toma de decisiones.

Figura 67

Visualización de Puntajes de Crédito

```
print(credit_scoresGBC)
```

```
[88.06088628 34.3962934  89.81047801 91.1277554  92.40662043 39.08033396
 91.09751959 53.02481216 17.16148104 68.3787243  95.98530962 91.55266423
 92.12358316 90.26421335 68.49418208 93.5510022  55.17084496 90.74929493
 97.49947643 97.89890241 49.87231631 88.56557982 51.79411385 89.23929097
 46.57428905 39.69568584 71.91695037 87.67924989 25.2086867  94.75366448
 90.83666575 86.50265696 61.98085283 86.26096893 14.92255784 92.29003736
 82.84854811 97.82904659 73.24182971 78.46127412 70.23671733 92.95169544
 49.48750192 95.84786261 98.05102051 57.24542903 97.96409153 80.85019774
 93.45145349 40.87616523 71.95920386 78.50042094 45.98581021 35.44074839
 77.1148648  90.33080417 60.62449408 86.46683182 29.36715438 69.35278086
 52.03651654 90.75496189 77.80479549 83.32433496 32.28107243 15.90315199
 61.43199288 85.00644623 21.57083611 56.03344562 94.54489292 43.26950732
 23.23311147 73.2164071  74.44000619 65.64267611 27.59800055 84.84689892
 88.86028272 87.25303603 93.10667729 88.83400286 53.10010337 62.4244506
 90.83052068 78.56551724 68.42835797 92.40348663 21.55957691 91.44755179
 44.0224239  36.05455421 69.19991682 52.53338513 92.06208995 36.58096271
 71.42550856 32.69862914  8.46865087 86.97009696 32.36443494 69.87249178
  4.55797754 34.86612783 30.96999592 36.63646153 90.92601506 88.33948776
 52.81335454 81.87774031 55.66969498 26.48040695 76.60770257 97.14998584
 29.78648947 95.7081965  78.5413237  32.83644151 26.19533354 79.0259667
 42.46619896 94.94895663 85.38102336 28.61366249 57.8666661  63.96204477
 89.1790419  92.69485256 86.68263649 76.21787286 47.42203888 31.15167247
 88.70920856 87.85074474 71.64336109 34.40811429 92.51292246 32.86868646
 51.55664941 79.92033021 58.20161107 44.43115314 91.58180532 70.08275217
 65.4008401  38.21197646 85.15601973 91.50061035 58.39921802 92.1364771
 45.39222941 24.36551342 96.26666802 71.87639937 66.84841012 11.2581524
 54.79510156 95.37580178 97.54547699 67.34632901 80.30461376 58.98134583
 96.30056139 86.91781056 96.15479648 93.65103514 43.05813023 96.25681491
 70.70180738 86.8118012  17.95898404 85.99423701  9.64012429 18.3015662
 86.17143374 87.0063203  60.56840746 97.2478241  34.48839303 56.09733002
 63.78180527 88.56652764 79.95945404 68.05986328 91.45917955 74.34173222
 70.44429729 83.54134087 29.6579921  14.50495293 79.70185429 27.74095997
 56.21927953 45.74507444 93.80456086 94.0958159  92.75949864 87.72755044
 93.36037151 55.52531115]
```

El resultado del credit_scoresGBC) muestra una lista de valores numéricos que representan los puntajes de crédito calculados para cada muestra en el conjunto de prueba. Estos puntajes de crédito se pueden interpretar de la siguiente manera:

- Rango de valores: Los puntajes oscilan entre aproximadamente 4.55 y 98.05, lo que indica que el modelo genera una amplia gama de probabilidades de pertenencia a la clase de interés (en este caso, ser un buen candidato para un crédito).
- Distribución de los puntajes: Al observar los valores, se puede apreciar que hay una distribución heterogénea de los puntajes, con algunos muy altos (cercanos al 100) y otros más bajos (cercanos a 0). Esto sugiere que el modelo es capaz de discriminar entre muestras con diferentes niveles de riesgo o probabilidad de pertenecer a la clase de interés.
- Interpretación práctica: Estos puntajes de crédito pueden ser utilizados para tomar decisiones sobre la concesión de créditos. Por ejemplo, se podría establecer un umbral de puntaje mínimo (como 70) para considerar a un solicitante como elegible, o se podrían utilizar los puntajes para priorizar y ordenar a los solicitantes de acuerdo con su nivel de riesgo.

Figura 68

Análisis de Codificación y Decodificación de Variables Categóricas

```
[ ]  # Transformar las columnas categóricas en los conjuntos de entrenamiento y prueba
     X_trainGBC_encoded = encoder.transform(X_trainGBC[cat_colsGBC])
     X_testGBC_encoded = encoder.transform(X_testGBC[cat_colsGBC])

     # Convertir de nuevo las variables codificadas al estado original en el conjunto de prueba
     decoded_arrayGBC = encoder.inverse_transform(X_testGBC_encoded)

     # Crear un DataFrame con las variables categóricas decodificadas
     decoded_dfGBC = pd.DataFrame(decoded_arrayGBC, columns=cat_colsGBC)

     # Crear un DataFrame con las variables numéricas en el conjunto de prueba
     Num_decoded_dfGBC = pd.DataFrame(X_testGBC[num_colsGBC].reset_index(drop=True), columns=num_colsGBC)

     # Combinar los DataFrames decodificados para obtener el conjunto de prueba final
     df_GBC = pd.concat([decoded_dfGBC.reset_index(drop=True), Num_decoded_dfGBC.reset_index(drop=True)], axis=1)
```

```
     # validación de que credit_scoresGBC y X_testGBC tengan la misma longitud
     if len(credit_scoresGBC) == len(df_GBC):
         # Asignar las puntuaciones de crédito al DataFrame como una nueva columna
         df_GBC['credit_scoresGBC'] = credit_scoresGBC
     else:
         print("Error: La longitud de credit_scores no coincide con la longitud del DataFrame.")

     # Mostrar el DataFrame con la nueva columna
     df_GBC.head(10)
```

La función de código que evidenciamos muestra el proceso de transformación y codificación de las variables categóricas en los conjuntos de entrenamiento y prueba, lo que es un paso crucial para la aplicación de modelos de aprendizaje automático que requieren variables numéricas. Posteriormente, se realiza la decodificación de las variables categóricas en el conjunto de prueba, lo que permite mantener la interpretabilidad de los datos.

Tabla 31

Tabla de Datos de Crédito

	Importe_Crédito	Saldo de cuenta	Duración del Crédito_mensual	Años de edad	credit_scoresGBC
0	2515	Saldo Superior	18	43	88.060886
1	7721	Sin cuenta	24	30	34.396293
2	640	Saldo Superior	12	49	89.810478
3	2910	Sin cuenta	24	34	91.127755
4	932	Ninguna (sin saldo)	6	39	92.406620
...	...	...	...	...	...
195	1979	Saldo Superior	15	35	94.095816
196	1804	Ninguna (sin saldo)	12	44	92.759499

197	753	Ninguna (sin saldo)	6	64	87.727550
198	1494	Algún saldo	4	29	93.360372
199	3599	Sin cuenta	21	26	55.525311

Nota. Esta tabla muestra los datos utilizados para entrenar y evaluar el modelo de evaluación de crédito. Toma de Elaboración Propia (2024).

La tabla de resultados muestra una muestra de 10 registros del conjunto de prueba final, donde se pueden observar las variables categóricas decodificadas junto con las variables numéricas y los puntajes de crédito generados por el modelo de Gradient Boosting Classifier (GBC). Estos puntajes de crédito oscilan entre aproximadamente 4.55 y 98.05, lo que indica que el modelo es capaz de generar una amplia gama de probabilidades de pertenencia a la clase de interés (ser un buen candidato para un crédito).

Los puntajes más altos (cercanos a 100) indican una mayor probabilidad de ser un buen candidato para un crédito, mientras que los puntajes más bajos (cercanos a 0) indican una menor probabilidad. Esta distribución heterogénea de los puntajes sugiere que el modelo es capaz de discriminar entre muestras con diferentes niveles de riesgo, lo que es crucial para la toma de decisiones sobre la concesión de créditos. Estos puntajes de crédito pueden ser utilizados posteriormente para establecer umbrales mínimos o priorizar a los solicitantes de acuerdo a su nivel de riesgo.

Por último, al desarrollar el objetivo de este proyecto que fue implementar y evaluar modelos de clasificación de crédito para predecir el riesgo de incumplimiento de los solicitantes

de crédito. Se compararon tres modelos principales: Random Forest Classifier (RFC), Gradient Boosting Classifier (GBC) y Logistic Regression (LR).

Por lo cual el conjunto de datos utilizado contiene información relevante sobre los solicitantes de crédito, incluyendo variables como saldo de cuenta, estado de pago del crédito anterior, finalidad del crédito, valor de ahorro, duración del empleo actual, entre otras. Estos datos fueron limpiados, transformados y preparados para su uso en los modelos de clasificación.

De esta manera se aplicaron los tres modelos de clasificación mencionados anteriormente (RFC, GBC y LR) a los datos preprocesados. Para evaluar el desempeño de los modelos, se utilizó la métrica de Accuracy, que mide la proporción de predicciones correctas sobre el total de predicciones.

Figura 69

Análisis Comparativo de Modelos de Clasificación de Crédito

Conclusión en Accuracy:

Auto pycaret:

Modelo	Accuracy
Random Forest Classifier	0.7713
Gradient Boosting Classifier	0.7593
Logistic Regression	0.7473

Manual:

Modelo	Accuracy
Random Forest Classifier	0.77
Gradient Boosting Classifier	0.76
Logistic Regression	0.78

En atención a la gráfica se evidencia que, según los resultados obtenidos, el modelo con mejor desempeño en términos de Accuracy fue el Logistic Regression (LR) con un valor de 0.78 en el modelo manual. Los otros modelos también mostraron un buen desempeño, con Accuracy

de 0.77 para el Random Forest Classifier (RFC) y 0.76 para el Gradient Boosting Classifier (GBC) en el modelo manual.

Es así como, para facilitar la comparación y el análisis de los resultados, se creó un DataFrame combinado (DF_Final) que incluye los puntajes de crédito generados por cada uno de los tres modelos (credit_scoresRF, credit_scoresRL, credit_scoresGBC) junto con las variables originales del conjunto de datos.

Tabla 32

Tabla de Características de Crédito

	credit_scoresRF	credit_scoresRL	credit_scoresGBC
0	83.333333	94.581390	88.060886
1	37.722222	60.599068	34.396293
2	84.952381	80.931370	89.810478
3	79.897436	82.088082	91.127755
4	78.666667	93.008206	92.406620
...	...	...	...
195	96.666667	96.029518	94.095816
196	85.066667	93.467683	92.759499
197	82.000000	89.591907	87.727550
198	86.666667	91.452460	93.360372
199	62.666667	53.963014	55.525311

Nota. Esta tabla muestra una lista detallada de las características y atributos de los solicitantes de crédito. Tomado de Elaboración Propia (2024).

- Comparar los puntajes de crédito generados por los diferentes modelos para cada instancia de crédito.

- Identificar los factores clave que influyen en la clasificación de crédito, al mantener la interpretabilidad de las variables originales.
- Analizar patrones y relaciones entre las características de los solicitantes de crédito y los resultados de los modelos.

Conclusiones

El proyecto alcanzó el objetivo general, mediante el desarrollo y entrenamiento de un modelo supervisado de machine learning, utilizando un conjunto de datos representativo (Datos Crediticios Alemanes). El modelo logró predecir el score de riesgo crediticio con una precisión satisfactoria, proporcionando una herramienta sólida para mejorar la toma de decisiones en instituciones financieras. Este resultado evidencia el impacto positivo del uso de modelos supervisados para optimizar procesos críticos en la evaluación de riesgo crediticio.

A lo que respecta los objetivos específicos se realizó un análisis exhaustivo del conjunto de datos, identificando patrones relevantes y comprendiendo las características financieras de los solicitantes. Este análisis permitió una mejor segmentación de los datos y facilitó la identificación de factores que contribuyen al perfil de riesgo.

A través de técnicas como la selección de características y el análisis estadístico, se identificaron las variables más relevantes, como saldos en cuentas, créditos antiguos, edad, entre otras. Estas variables fueron fundamentales para construir un modelo predictivo eficiente, logrando un nivel óptimo de precisión y reduciendo el ruido en los datos. Este paso fue clave para mejorar el asertividad en la predicción del score crediticio.

Se llevaron a cabo procesos de validación rigurosos, incluyendo métricas de evaluación como precisión, AUC-ROC y sensibilidad. Los resultados obtenidos demostraron una alineación consistente entre las predicciones del modelo y los factores determinantes del score crediticio en el conjunto de datos. Esto confirmó la fiabilidad del modelo como herramienta de apoyo en la calificación crediticia, cumpliendo con este objetivo.

El análisis de los Datos Crediticios Alemanes indicó que la variable de Credibilidad es fundamental para predecir el riesgo crediticio, destacando que la mayoría de los solicitantes presentan un perfil de riesgo aceptable. Asimismo, la calidad del conjunto de datos es adecuada, lo que permitió avanzar en la implementación de los diferentes modelos de machine learning sin preocupaciones por valores faltantes o duplicados. La transformación de variables cuantitativas a cualitativas mejoró la interpretación de los modelos implementados y la precisión de las predicciones sobre el riesgo crediticio, lo que permitió identificar patrones y relaciones clave para mejorar la efectividad del análisis predictivo.

Como se observó en el análisis y resultados de los modelos aplicados, la implementación de una librería de Python como Pycaret con el Auto ML, el cual simplifico el proceso de automatización de tareas de aprendizaje automático, incluido el preprocesamiento de datos, la ingeniería de características, la selección de modelos y la implementación, permitió la comparación de modelos de catorce modelos brindando los mejores resultados del modelo con “Random Forest Classifier”, seguido por “Extra Trees Classifier”, y “Gradient Boosting Classifier”, demostrando efectividad en la predicción del score, logrando una precisión que superó las expectativas iniciales.

Con dichos resultados obtenidos se implementaron tres modelos de forma manual sin uso de Pycaret con el objetivo de lograr mejores resultados en métricas de clasificación, donde se tomó el primer Modelo con “Random Forest Classifier”, el segundo con “Regresión Logística” y por último “Gradient Boosting Classifier”, obteniendo solo mejora en el modelo “Regresión Logística” de una precisión de 0.7473 a 0.78.

Al realizar la comparación de los últimos tres modelos ajustados, donde con cada uno se hizo uso de predicción calculando la probabilidad de cada clase se visualizaron mejores

resultados en el modelo “Regresión Logística” donde sus variables con mayor importancia se enfocan en saldo de cuenta, finalidad y créditos anteriores.

Estos resultados obtenidos pueden influir significativamente en la toma de decisiones dentro del contexto aplicado, en cada entidad financiera, donde las variables con mayor importancia proporcionan herramientas cuantitativas para evaluar riesgos y oportunidades.

Se identificaron limitaciones en los datos utilizados, como la falta de representatividad en ciertos segmentos, lo que podría afectar la generalización del modelo. De igual forma el uso de segmentar variables puede brindar una mejor exactitud en el modelo. Es fundamental abordar estas limitaciones en futuros trabajos.

Se concluye que se da por completado el desarrollo de los objetivos planteados del proyecto implementado, dando recomendaciones para ampliar soluciones más efectivas a la problemática tratada.

Recomendaciones

Con los resultados obtenidos se evidencia que se puede realizar una mejora continua al modelo implementado, actualizando su set de datos con una Base nacional para adecuar los porcentajes de score, así mejorando su precisión y adaptabilidad a cambios en el entorno; Además, es fundamental establecer un procedimiento sistemático de actualización periódica de los datos, considerando fuentes confiables y completas, como estadísticas del sistema financiero nacional, indicadores macroeconómicos y datos demográficos. Esto ayudará a reducir sesgos y a reflejar cambios en patrones de comportamiento crediticio.

Se puede investigar la inclusión de variables adicionales que puedan enriquecer el modelo y mejorar su capacidad predictiva, especialmente aquellas relacionadas con factores externos que no fueron considerados inicialmente, como, por ejemplo:

Factores macroeconómicos: tasas de desempleo, inflación, tasas de interés, datos de comportamiento digital: historial de pagos en línea, actividad en redes sociales (si es legalmente viable y ético), factores demográficos y regionales: características específicas de diferentes segmentos de la población, como nivel educativo, zonas urbanas/rurales e indicadores alternativos: por ejemplo, historiales de pago de servicios públicos (agua, electricidad, telecomunicaciones).

Se debe evaluar el impacto de estas variables mediante pruebas de validación cruzada y análisis estadísticos para garantizar que aporten valor predictivo al modelo actual.
Realizar investigación futura donde se exploren diferentes algoritmos avanzados o enfoques híbridos que puedan superar las limitaciones actuales y mejorar la robustez del modelo.

Se sugiere para el proyecto modelos ensamble como XGBoost, LightGBM, CatBoost para mejorar la estabilidad y precisión.

Métodos basados en redes neuronales profundas (Deep Learning), que podrían capturar relaciones no lineales más complejas en los datos, enfoques híbridos que combinen modelos estadísticos tradicionales (como regresión logística) con técnicas modernas de machine learning para maximizar las fortalezas de cada enfoque.

Finalmente, tener en cuenta diseñar un plan robusto de validación que simule escenarios reales antes de implementar el modelo en producción. Esto incluirá pruebas en conjuntos de datos fuera de muestra y análisis en tiempo real para identificar posibles fallos o desviaciones.

Asia mismo, documentar exhaustivamente cada etapa del desarrollo y mantener un registro de cambios en el modelo para garantizar la trazabilidad y replicabilidad del proyecto.

Referencias Bibliográficas

Aceituno, M. (2019). *Modelo predictivo de análisis de riesgo crediticio usando Machine Learning en una entidad del sector microfinanciero* [Tesis de grado, Universidad Nacional del Altiplano]. Repositorio institucional. http://repositorio.unap.edu.pe/handle/20.500.14082/14481

Asobancaria. (2019). *Impacto en la información financiera y el acceso al crédito proyectado* [Informe]. https://www.asobancaria.com/wp-content/uploads/1184.pdf

Grupo Bancolombia. (2021). *Conoce por qué tu historial crediticio es importante*. https://www.bancolombia.com/educacion-financiera/finanzas-personales/importancia-historial-crediticio#:~:text=Tener%20un%20buen%20historial%20crediticio,de%20inter%C3%A9s%20en%20tus%20cr%C3%A9ditos

Buenaño, D., & Fernández, S. (2016). Uso de la metodología CRISP-DM para guiar el proceso de minería de datos en LMS. *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, *15*, 2385-2393. https://dialnet.unirioja.es/servlet/articulo?codigo=6019602

Castro, J. (2022). Aplicación de Machine Learning en la gestión de riesgo de crédito financiero: Una revisión sistemática. *Interfases, (15),* 160-178. https://doi.org/10.26439/interfases2022.n015.5898

Cepeda, A. (2022). *Modelo de Scoring para crédito de consumo en la Cooperativa de Ahorro y Crédito "Minga" Ltda. utilizando técnicas de Machine Learning.* [Tesis de grado, ESPOCH]. Repositorio institucional. http://dspace.espoch.edu.ec/bitstream/123456789/19893/1/226T0122.pdf

Datacredito Empresas. (2021). *¿Cómo funciona el puntaje de DataCrédito?* Recuperado de https://www.datacreditoempresas.com.co/blog-datacredito-empresas/como-funciona-el-puntaje-de-datacredito/

Diario La República. (2024), *¿Por qué es importante el score crediticio?, Así puede mejorarlo y acceder a préstamos*. https://www.larepublica.co/finanzas-personales/por-que-es-importante-el-score-crediticio-asi-puede-mejorarlo-y-acceder-a-prestamos-3308032

Díaz, L. (2023). *Uno de cada cinco colombianos acude a los 'gota a gota'.* https://www.eltiempo.com/economia/finanzas-personales/este-es-el-perfil-de-quienes-acuden-a-los-gota-a-gota-en-colombia-800259

Espinosa, A. (2014). *Modelos de clasificación en el otorgamiento de créditos financieros: comparación entre diferentes técnicas de Machine Learning y modelos de regresión múltiple* (Doctoral dissertation, Uniandes).

Franco, J., Urrutia, O. (2024). *Sustentación proyecto Aplicado* [Presentación de PowerPoint]. Universidad Nacional Abierta y a Distancia. https://unadvirtualedu-my.sharepoint.com/:p:/g/personal/jafrancova_unadvirtual_edu_co/EQKqRO8BBABHrDMWiy-6hqQBSfijYhJ2ezDuro65QwEzeg?e=WhKbsD

Franco, J., Urrutia, O. [Video Youtube]. (2024). *Sustentación Proyecto de Grado score de crédito* [Video]. YouTube. https://youtu.be/Kuj-Jv0pn8M

García, Y. (2020). *Datos Digitales en los Score de Crédito: Herramienta para la Inclusión Financiera del Crédito en Colombia*. [Tesis de maestría en contabilidad y finanzas, Universidad Nacional de Colombia]. Repositorio UNAL. https://repositorio.unal.edu.co/bitstream/handle/unal/79013/DATOS%20DIGITALES%20EN%20LOS%20SCORE%20DE%20CR%C3%89DITO%20HERRAMIENTA%20PA

RA%20LA%20INCLUSI%C3%93N%20FINANCIERA%20DEL%20CR%C3%89DITO%20EN%20COLOMBIA....pdf?sequence=1&isAllowed=y

Giraldo, W. (2021). *Machine Learning para la estimación del riesgo de crédito en una cartera de consumo* [Tesis de magíster en administración financiera, Universidad EAFIT]. Repositorio Eafit. https://repository.eafit.edu.co/server/api/core/bitstreams/324d9a52-33ab-4255-bfaf-132f206c753e/content

González, J., Pérez, J., & Montoya, F. (2009). *La crisis financiera y económica del 2008: Origen y consecuencias en los Estados Unidos y México*. *El cotidiano, (157)*, 17-27. https://www.redalyc.org/pdf/325/32512739003.pdf

Gualoto, O. (2022). *Desarrollo de un modelo de predicción basado en Algoritmos de Machine Learning para medir el riesgo crediticio*. [Tesis de licenciatura, Institución no especificada, Quito, Ecuador].

Herman, D., Jelinek, T., Reade, W., Demkin, M., & Howard, A. (2024). *Crédito vivienda - Estabilidad del modelo de riesgo de crédito [Competencia en línea]*. *Kaggle*. Recuperado de https://kaggle.com/competitions/home-credit-credit-risk-model-stability

IBM. (2021). *Conceptos básicos de ayuda de CRISP-DM*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Martínez, D. (2019). *Efecto del Riesgo de Crédito y el Riesgo de Liquidez en la Estabilidad Bancaria de Latinoamérica*. [Tesis de maestría en finanzas, Universidad de Chile, Facultad de Economía y Negocios]. Repositorio UC Chile. https://repositorio.uchile.cl/bitstream/handle/2250/173535/Tesis%20-%20Martinez%20Rivera%20Daysi.pdf?sequence=1

Pucha, O. (2022). *Desarrollo de un modelo de predicción basado en Algoritmos de Machine Learning para medir el riesgo crediticio (Bachelor's thesis, Quito, 2022)*. [Tesis de grado, EPN]. Repositorio institucional. http://bibdigital.epn.edu.ec/handle/15000/22290

Rodríguez, L. (2023). *Metodología CRISP-DM: La guía definitiva para la Minería de Datos. Figshare*. https://doi.org/10.6084/m9.figshare.24530176.v1

Saavedra, L., Saavedra, J. (2019). *Modelos para medir el riesgo de crédito de la banca. Cuadernos de administración, 23(40), 295-319.* http://www.scielo.org.co/scielo.php?pid=S0120-35922010000100013&script=sci_arttext

Saavedra, P. (2023) *Un método para la asignación de cupos de crédito de entidades del sector financiero colombiano empleando técnicas de machine learning* (Doctoral dissertation, Universidad Nacional de Colombia)

Superfinanciera de Colombia. (2011). *Circular externa 052 de 2011.* https://www.superfinanciera.gov.co/loader.php?lServicio=Tools2&lTipo=descargas&lFuncion=descargar&idFile=6061

Transunion. (2007). *La importancia de los Scoring para el crecimiento económico.* https://www.transunion.com/docs/interstitial/scoringWhitepaper_Mexico.pdf

UNAD. (2016). *Líneas de Investigación ECBTI Universidad Nacional Abierta y a Distancia* UNAD. https://academia.unad.edu.co/investigacion-y-productividad-ecbti/lineas

Apéndices

Apéndice A

Marco Normativo Alemán

Para dar un contexto más amplio del marco normativo en Alemania para el cálculo del score crediticio, este se basa en diversas leyes y regulaciones que buscan proteger los derechos de los consumidores y garantizar la transparencia en el uso de datos personales.

Bundesdatenschutzgesetz (BDSG) es la Ley Federal de Protección de Datos, conocida como BDSG, la cual regula el tratamiento de datos personales en ese país. Esta ley establece el uso de scores crediticios, que predicen el comportamiento futuro de pago de un individuo, solo es admisible si se cumplen ciertos requisitos. Entre ellos, se debe asegurar que solo se utilicen datos pertinentes y que se informe al consumidor sobre cómo se calcula su score.

Reglamento General de Protección de Datos (RGPD) se aplica en toda la Unión Europea, también influye en cómo se manejan los datos para un Scoring crediticio. Este reglamento exige que las entidades que manejan datos personales obtengan el consentimiento del individuo y les proporcionen acceso a sus datos, así como la posibilidad de rectificarlos.

SCHUFA (Schutzgemeinschaft für allgemeine Kreditsicherung) es la principal agencia de crédito en Alemania (equivalente a Transunion y Experian en Colombia). Su funcionamiento está regulado por las leyes alemanas de protección de datos y tiene un papel importante en la evaluación de la solvencia crediticia. SCHUFA calcula el Scoring de crédito basado en un método estadístico que considera diversos factores del historial financiero del individuo, y proporciona esta información a bancos y otras instituciones financieras.

Con ello se evidencia que las personas tienen derecho a conocer su score crediticio y a entender cómo se ha calculado. Esto incluye el acceso a los datos utilizados para determinar su puntuación.

En ese país, las regulaciones buscan evitar discriminaciones injustas basadas en scores crediticios erróneos o mal calculados. La ley permite a los consumidores impugnar decisiones basadas en estos scores si lo consideran que son inexactos o errados.

Este marco normativo asegura que el proceso de Scoring crediticio sea justo y transparente, protegiendo así tanto a las instituciones financieras como a los consumidores.