



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Un método para la asignación de cupos de crédito de entidades del sector financiero colombiano empleando técnicas de machine learning

Edher Daniel Saavedra Porras

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión

Medellín, Colombia

2023

Un método para la asignación de cupos de crédito de entidades del sector financiero colombiano empleando técnicas de machine learning

Edher Daniel Saavedra Porras

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:

Magister en Ingeniería - Analítica

Director (a):

Albeiro Espinosa Bedoya, Ph.D.

Línea de Investigación:

Profundización

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión

Medellín, Colombia

2023

Agradecimientos

Con sincero agradecimiento deseo reconocer a aquellos que me han brindado su apoyo en este camino:

A mi amada esposa, cuyo apoyo incondicional, paciencia y comprensión han sido el faro que iluminó los momentos desafiantes de esta travesía. A mi querida familia, a quienes considero mi fortaleza, agradezco por su constante respaldo, donde cada palabra de aliento, gesto de apoyo y comprensión ha sido invaluable.

A mi tutor, agradezco profundamente por su orientación experta y su dedicación incansable. Sus conocimientos y disposición permitieron la culminación de este trabajo. A mi alma mater, la Universidad Nacional de Colombia, le expreso mi gratitud por proporcionarme el espacio y los recursos necesarios para llevar a cabo esta investigación.

Finalmente, a todos aquellos que, de alguna manera, contribuyeron a este logro, les doy las gracias. Este camino ha sido un esfuerzo colectivo, y cada uno de ustedes ha dejado una huella imborrable en mi viaje hacia la culminación de este sueño académico.

Resumen

Un método para la asignación de cupos de crédito de entidades del sector financiero colombiano empleando técnicas de machine learning

El análisis de riesgo de crédito desempeña un papel crucial en el sector financiero, ya que evalúa variables que podrían deteriorarse en circunstancias particulares, llevando a un incumplimiento de obligaciones y, aunque la metodología CAMEL es ampliamente utilizada, esta se basa en un sistema de calificaciones simple. Evaluar la influencia de factores macro y microeconómicos en la estabilidad financiera, especialmente en un contexto de preocupaciones globales, junto con la monitorización de indicadores es crucial para mitigar incumplimientos crediticios. A agosto de 2022, la cartera morosa en el sector financiero colombiano asciende a COP 23.4 billones, por lo cual, se busca proponer un método para la asignación de cupos de crédito basado en machine learning, analizar métodos existentes, desarrollar un método basado en indicadores financieros, y validar el modelo propuesto comparándolo con otros en la literatura.

El análisis de los antecedentes muestra que los métodos de machine learning superan a los estadísticos tradicionales en la estimación de riesgos crediticios, destacándose técnicas como Random Forest, Support Vector Machines, y redes neuronales. Además, se aplicaron criterios ponderados para evaluar la elección de dichos métodos, considerando la frecuencia de aplicación, resultados destacados en la literatura y opiniones de expertos. Random Forest y Árboles de Decisión obtuvieron las calificaciones más altas en el ranking debido a que se destacan su flexibilidad y capacidad para manejar diversos desafíos en diferentes aplicaciones.

El análisis se basó en más de 50 indicadores financieros recopilados de la Superintendencia Financiera de Colombia, abarcando diversas entidades del sector financiero, para luego implementar un modelo de clasificación de riesgo crediticio mediante Random Forest, logrando una precisión excepcional del 99.9% en datos de prueba y 99.79% en entrenamiento. La interpretación de la importancia de las características y la matriz de confusión respaldan la robustez del modelo.

Finalmente, se compararon los resultados con árboles de decisión y regresión logística, obteniendo un accuracy de 99.9% para Random Forest, 97.9% en la métrica de recall y 98.9% en F1 Score, resultados superiores a los modelos en comparación y destacando el notable rendimiento superior de Random Forest en la predicción de riesgo crediticio. Estos hallazgos respaldan su elección como una herramienta eficaz en la gestión de riesgos crediticios en el contexto colombiano.

Palabras clave: Riesgo crediticio, CAMEL, machine learning, Random Forest, indicadores financieros, sector financiero colombiano.

Abstract

A method for assigning credit quotas for entities in the Colombian financial sector using machine learning techniques

Credit risk analysis plays a crucial role in the financial sector, as it evaluates variables that could deteriorate under specific circumstances, leading to non-compliance with obligations. Although the CAMEL methodology is widely used, it relies on a simple rating system. Evaluating the influence of macro and microeconomic factors on financial stability, especially in a context of global concerns, and monitoring indicators are crucial to mitigate credit defaults. As of August 2022, the non-performing portfolio in the Colombian financial sector amounts to COP 23.4 trillion. Therefore, we aim to propose a method for assigning credit quotas based on machine learning, analyze existing methods, develop a method based on financial indicators, and validate the proposed model by comparing it with others in the literature.

The background analysis indicates that machine learning methods outperform traditional statistics in estimating credit risks. Techniques such as Random Forest, Support Vector Machines, and neural networks stand out. Weighted criteria were applied to evaluate the choice of these methods, considering the frequency of application, notable results in the literature, and expert opinions. Random Forest and Decision Trees obtained the highest scores in the ranking due to their flexibility and ability to handle various challenges in different applications.

The analysis was based on more than 50 financial indicators collected from the Financial Superintendency of Colombia, encompassing various entities in the financial sector. Subsequently, a credit risk classification model was implemented using Random Forest, achieving an exceptional accuracy of 99.9% in test data and 99.79% in training. The interpretation of the importance of the features and the confusion matrix supports the robustness of the model.

Finally, the results were compared with decision trees and logistic regression, yielding an accuracy of 99.9% for Random Forest, 97.9% in the recall metric, and 98.9% in F1 Score, results superior to the models in comparison, highlighting the remarkable superior performance of Random Forest in credit risk prediction. These findings support its selection as an effective tool in credit risk management in the Colombian context.

Keywords: Credit risk, CAMEL, machine learning, Random Forest, financial indicators, Colombian financial sector.

Contenido

	Pág.
Resumen	5
Lista de figuras.....	10
Lista de tablas	11
Introducción	12
1. Planteamiento del problema	14
1.1 Justificación	14
1.2 Objetivos.....	15
1.2.1 Objetivo general	15
1.2.2 Objetivos específicos	15
1.3 Antecedentes.....	16
1.4 Metodología a utilizar.....	21
1.5 Alcances del trabajo	22
2. Marco Teórico	23
3. Análisis de los principales métodos de machine learning identificados en la literatura para la asignación de cupos de crédito	27
4. Método propuesto para el análisis previo y selección de indicadores financieros.....	39
4.1 Insumos del modelo.....	39
4.2 Preprocesamiento de los datos	41
4.3 Modelo de Random Forest	43
5. Validación del modelo propuesto	48
6. Conclusiones y recomendaciones.....	50
6.1 Conclusiones	50
6.2 Recomendaciones	50
Bibliografía	52

Lista de figuras

	Pág.
Figura 3-1: Fortalezas y debilidades en métodos de machine learning.	37
Figura 4-2: Importancia de las características.	44
Figura 4-3: Gráfico de codo.	45
Figura 4-4: Matriz de confusión.....	46

Lista de tablas

	Pág.
Tabla 1-1: Metodología de trabajo de grado.....	22
Tabla 3-2: Ranking de calificación de métodos de machine learning.....	34
Tabla 4-1: Indicadores financieros y rubros como insumos del modelo.....	40
Tabla 5-1: Comparación de métricas de desempeño entre modelos.....	49

Introducción

El análisis de riesgo de crédito constituye un pilar fundamental en el funcionamiento de las entidades financieras, desempeñando un papel crucial en la toma de decisiones relacionadas con la asignación de créditos y la gestión de carteras. En este contexto, la metodología CAMEL (Capital Adequacy, Assets Quality, Management and Performance, Efficiency and Profitability, Liquidity) [1] ha sido una herramienta ampliamente utilizada a nivel mundial para evaluar la solidez de las instituciones financieras. Sin embargo, la simplicidad inherente a este modelo y las limitaciones de los métodos estadísticos tradicionales han motivado la búsqueda de enfoques más precisos y avanzados [2].

La estabilidad financiera de una entidad y su capacidad para hacer frente a obligaciones crediticias se ven influenciadas por factores macro y microeconómicos, por lo cual la importancia del análisis de riesgo de crédito cobra mayor relevancia en un contexto de incertidumbre global, donde se vislumbran temores de recesión, aumentos en las tasas de interés y volatilidad en los mercados, siendo el monitoreo y análisis de riesgo crediticio, elementos cruciales para anticipar y mitigar posibles situaciones de incumplimiento.

El presente estudio aborda la relevancia del análisis de riesgo de crédito en el sector financiero colombiano, destacando la importancia de identificar perfiles crediticios sólidos para minimizar la morosidad y optimizar la asignación de cupos de crédito. En agosto de 2022, el saldo de la cartera vencida en el sector financiero colombiano alcanza los COP 23.4 billones [44], subrayando la necesidad de estrategias más precisas y efectivas en la evaluación de riesgos.

Como bien se mencionaba, la metodología CAMEL, a pesar de su amplio uso, presenta limitaciones al basarse en un sistema de calificaciones simple. Además, los métodos estadísticos tradicionales muestran restricciones en su capacidad predictiva. Estos antecedentes teóricos y prácticos han motivado la incursión en la aplicación de modelos de machine learning para el análisis de riesgo crediticio, donde la literatura existente respalda la superioridad de modelos como Regresión Logística, Support Vector Machines, Redes Neuronales, K-Nearest Neighbors, Árboles de Decisión y Random Forest, en comparación con métodos estadísticos tradicionales. La adaptación de estas técnicas al

contexto colombiano implica un avance significativo en la capacidad predictiva y en la identificación temprana de situaciones de riesgo.

El objetivo general de este estudio es proponer un método basado en técnicas de machine learning para la asignación de cupos de crédito a entidades del sector financiero colombiano. Para lograr este propósito, se plantea analizar los principales métodos de machine learning identificados en la literatura, desarrollar un método basado en indicadores financieros y validar el modelo propuesto en comparación con los métodos existentes.

El alcance de la investigación abarca la recopilación y análisis de información de más de 50 indicadores financieros de entidades financieras colombianas, con un periodo de estudio que se extiende desde enero de 2015 hasta abril de 2023. La aplicación de técnicas de machine learning se centra en la predicción de riesgo crediticio, contribuyendo así a la toma de decisiones fundamentadas en la asignación de créditos. Cabe señalar que este estudio presenta ciertas limitaciones, ya que la utilización de datos históricos implica que el modelo se basa en condiciones económicas y financieras pasadas y su capacidad predictiva puede verse afectada por cambios abruptos en el entorno. Además, la disponibilidad y calidad de datos pueden variar entre entidades financieras, lo que podría influir en la robustez del modelo propuesto.

En cuanto a la metodología y la aplicación del modelo de machine learning, se emplea el algoritmo Random Forest para la clasificación de riesgo crediticio. Este enfoque se justifica por su eficacia demostrada en la literatura y su capacidad para manejar múltiples variables predictoras, mostrando ser una estrategia efectiva para mejorar la precisión y la capacidad predictiva en un contexto financiero dinámico y cambiante.

1. Planteamiento del problema

El análisis de riesgo de crédito juega un papel clave para entidades del sector financiero debido a la importancia del estudio de las variables que puedan deteriorarse ante circunstancias particulares y que lleven a un impago de las obligaciones [1], [6]. Una de las metodologías más usadas a nivel mundial para llevar a cabo esta labor es la CAMEL (C = capital adequacy (solventía), A = assets quality (calidad de los activos), M = management and performance (eficiencia administrativa), E = efficiency and profitability (rentabilidad), L = liquidity (liquidez)), la cual consiste en un grupo de indicadores que miden la solidez de una institución financiera, que cuentan con una calificación por cada criterio y con unas ponderaciones otorgadas a estos que dependen del interés y juicio de quien usa el modelo [1].

El modelo CAMEL, pese a ser un modelo ampliamente utilizado, tanto por entidades vigiladas como por los mismos reguladores y entidades calificadoras de riesgo, se basa en un sistema de calificaciones simple y de fácil interpretación que resume en un sólo indicador la situación general de una entidad financiera [1]. De igual manera, los métodos estadísticos tradicionales, en la mayoría de casos, proveen una capacidad de pronóstico limitada, por lo cual, se busca la aplicación de modelos de machine learning al análisis de riesgo crediticio con el fin de determinar con mayor exactitud, las condiciones financieras y la perspectiva de riesgo de crédito de las entidades del sector financiero colombiano con el fin de tomar una decisión acerca de la apertura de un cupo de crédito [2], de allí, que se mitiguen eventos de probables impagos hacia las entidades prestamistas y/o se disminuyan pérdidas de oportunidades de inversión debido a calificaciones crediticias erróneas.

1.1 Justificación

Factores macroeconómicos tanto como microeconómicos intervienen al momento de determinar la estabilidad financiera de una entidad y su capacidad para afrontar sus obligaciones crediticias, a la fecha en la cual se elabora este documento, la coyuntura mundial genera impactos en las compañías colombianas, acentuándose temores de recesión en las principales economías del planeta, continuando las expectativas de

incrementos en las tasas de interés por parte de los bancos centrales de dichos países, desvalorizándose los principales mercados accionarios y de renta fija y percibiendo los impactos de los altos niveles inflacionarios tanto a nivel global como local.

La importancia de monitorear los indicadores de las entidades dado los factores mencionados anteriormente, radica en la probabilidad de mitigar situaciones de incumplimientos crediticios que se puedan presentar. A agosto de 2022, el saldo de la cartera de los establecimientos de crédito del sector financiero colombiano que reporta mora mayor a 30 días se eleva a COP 23.4 billones, presentándose un aumento intermensual de COP 120 mil millones en la cartera vencida y presentándose un indicador de calidad de cartera, medida como la proporción entre la cartera vencida y la cartera bruta, de 3.7% [44]. Si bien este indicador se mantiene estable frente a meses anteriores, la posibilidad de aplicar técnicas de machine learning en la asignación de cupos de crédito a entidades del sector financiero colombiano, podría representar una disminución de este indicador al identificar mejores perfiles crediticios que cumplan oportunamente con las obligaciones pactadas.

1.2 Objetivos

1.2.1 Objetivo general

- Proponer un método para la asignación de cupos de crédito basado en técnicas de machine learning para entidades pertenecientes al sector financiero colombiano.

1.2.2 Objetivos específicos

- Analizar los principales métodos de machine learning identificados en la literatura para la asignación de cupos de crédito.
- Desarrollar un método basado en el análisis previo y en una selección de indicadores financieros.
- Validar el modelo propuesto en comparación con los principales métodos identificados en la literatura.

1.3 Antecedentes

El análisis del riesgo de crédito busca blindar a las entidades financieras ante las probabilidades de incumplimiento de las obligaciones de sus clientes o contrapartes, sin embargo, no existe una metodología estándar con la cual se administre este problema en cuanto al uso del machine learning, por lo cual, a continuación, se relacionan los casos presentados en la literatura sobre cómo puede ser abordado este tema:

El documento “Machine Learning para la estimación del riesgo de crédito en una cartera de consumo” [3], argumenta que actualmente es posible evolucionar de los métodos estadísticos tradicionalmente utilizados para la estimación del riesgo de crédito, como lo son la regresión lineal y la regresión logística, para lo cual compararon esta última técnica frente a Random Forest, Support Vector Machines y Multi-layer Perceptron, en cuanto a precisión y aplicado a una cartera de consumo. Los autores concluyen que el modelo más equilibrado al momento de la evaluación es el Random Forest, ya que presentó el mejor ajuste en las métricas de precisión evaluadas, adicionalmente, resaltaron la importancia de garantizar la calidad de la captura de las variables relevantes para un modelo de riesgo. Resultados similares se obtuvieron en las conclusiones de [11], [17], [34], [35] y [39], siendo Random forest, SVM y Gradient Boosting Decision Tree los principales métodos de conjunto para una calificación crediticia más precisa.

Asimismo, [4], [5] y [10], llevaron a cabo estudios para demostrar la superioridad predictiva de los modelos de machine learning sobre los métodos estadísticos tradicionales, esta vez, midiendo el riesgo de crédito hipotecario derivado de la crisis económica de 2007, estimando el riesgo crediticio de los clientes retail de los bancos colombianos y prediciendo el riesgo de quiebra de los bancos de la Unión Europea a partir de las experiencias de la última década, respectivamente. Para lo anterior e igual que [3], emplearon tanto técnicas tradicionales como regresión logística, como métodos más recientes de machine learning como árboles de decisión, Gradient Boosting, redes neuronales y Support Vector Machines. En el primero de los artículos se aborda la crisis financiera del 2007 a causa del gran número de defaults en los créditos hipotecarios y se analiza un conjunto de créditos para ver la aplicación de las técnicas de machine learning en el ámbito empresarial, en el segundo de los artículos se observa el efecto de estas técnicas en los bancos al aplicarlas para identificar los clientes que podrían incurrir en un estado de mora en Colombia, tratando de predecir el próximo pago de la cuota de un cliente a partir de datos básicos de

la operación, del cliente y de pagos de cuotas anteriores registrados, el último artículo se enfoca en el riesgo de quiebra de los bancos de la Unión Europea, obteniendo como variables más significativas y como predictores más fuertes a las ganancias, la solvencia y la capacidad de administración. Como conclusión final y común de los anteriores trabajos, se evidenció que los árboles de decisión eran la técnica más recomendable debido a sus buenos resultados de predicción y por el análisis de las variables explicativas. Estudios que llegaron a conclusiones similares donde a través de diferentes métricas para determinar la precisión de la predicción, se ratificó la efectividad de los árboles de decisión y Random Forest, fueron [13], [22], [26], [30] y [32].

Sin embargo, pese a que numerosos estudios indican que los árboles de decisión y la técnica Random Forest suelen arrojar los mayores indicadores de pronóstico y los mejores rendimientos, no puede generalizarse dicha conclusión, ya que como lo demuestran los autores del trabajo “Comparative study of support vector machines and random forests machine learning algorithms on credit operation” [14], aunque Random Forest posee ventajas sobre Support Vector Machines en cuanto a su velocidad de procesamiento y simplicidad operativa, Support Vector Machines tiene el beneficio de una mayor precisión de clasificación que Random Forest. De igual manera, [20] llevaron a cabo un estudio con el fin de predecir los puntajes crediticios de las empresas, con aprendizaje automático y métodos estadísticos modernos, tanto en datos sectoriales como agregados. Los autores realizaron un análisis de 1881 empresas que operan en tres sectores diferentes que solicitaron préstamos del banco público más grande de Turquía. Los resultados del experimento se compararon en términos de precisión de clasificación, sensibilidad, especificidad, precisión y coeficiente de correlación de Mathews. Se observó que el análisis de regresión logística, Support Vector Machines, Random Forest y XGBoost tienen un mejor rendimiento que el árbol de decisión y el vecino más cercano para todos los conjuntos de datos. La anterior conclusión fue similar a la arrojada por el estudio “Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods” [24], el cual se enfocó en el análisis de la estimación de riesgos en operaciones de factoring y la probabilidad de que se pague una factura en un plazo aceptable, obteniendo como resultado una favorabilidad para los modelos de regresión, los cuales conducen a mayores beneficios y distribuyen mejor el riesgo.

En línea con lo anterior, múltiples son los estudios que demuestran la efectividad de Support Vector Machines, tal y como lo demuestra el documento “A predictive intelligence system of credit scoring based on deep multiple kernel learning” [28], cuyos autores aplicaron un clasificador de kernel múltiple profundo en una evaluación del riesgo crediticio de las tarjetas de crédito en China, encontrando que dicho clasificador supera a los modelos convencionales permitiendo realizar una mejor gestión del riesgo y evitando posibles deudas incobrables. Un estudio similar que soporta la tesis anterior, realizado en el 2022 [18], concluyó que la evaluación de clasificación múltiple tiene una mayor precisión que los modelos tradicionales y que la aplicación de la investigación puede proporcionar referencias a los bancos para fortalecer sus capacidades de prevención y control de riesgos y evitar pérdidas financieras, lo cual, de igual manera, se encuentra en línea con los resultados obtenidos de los estudios [36] y [41].

A pesar de lo anterior, existen autores que declinan de la competencia entre las técnicas mencionadas y optan por modelos que fusionen las mejores características de estas, como lo realizado en [25], donde se investigó acerca del Random Survival Forest for Competing Risks (CR Rsf), el cual es un método de estimación y predicción basado en árboles. Los autores aplicaron CR Rsf a un conjunto de datos financieros que involucra dos riesgos crediticios en competencia: incumplimiento y reembolso anticipado, donde los hallazgos sugieren que CR Rsf puede ser una alternativa útil a los modelos de riesgos competitivos existentes. Así mismo, en el trabajo que se llevó a cabo en [27], los autores desarrollaron un modelo de fusión de votación suave, que incorpora regresión logística, máquina de vectores de soporte (SVM), bosque aleatorio (RF), eXtreme Gradient Boosting (XGBoost) y Light Gradient Boosting Machine (LightGBM), con el fin de mejorar la precisión predictiva del riesgo crediticio de las PYME. Para verificar la factibilidad y efectividad del modelo propuesto, utilizaron datos de 123 pymes a nivel nacional que trabajaron con un banco chino de 2016 a 2020, incluida información financiera y registros de morosidad. Los resultados mostraron que la precisión del modelo de fusión de votación suave es mayor que la de un solo algoritmo de machine learning, lo que proporciona una base teórica para el control del riesgo crediticio en el futuro y ofrece referencias importantes para que los bancos puedan otorgar créditos de una manera más precisa y confiable. De igual manera, otro trabajo que potencializa el rendimiento y la interpretabilidad de los métodos de calificación crediticia, es el enfocado en la regresión de árbol logístico penalizado (PLTR) [19], el cual utiliza información de árboles de decisión para mejorar el rendimiento de la

regresión logística y permite capturar los efectos no lineales que pueden surgir en los datos de calificación crediticia, mientras se preserva la interpretabilidad intrínseca del modelo de regresión logística. Los resultados sugieren que PLTR predice el riesgo crediticio con mucha más precisión que la regresión logística y se compara competitivamente con el método de bosque aleatorio, un resultado similar que ofrece la técnica Bolasso (Bootstrap-Lasso) [42], la cual, aplicada al algoritmo Random Forest, proporciona mejores resultados para la evaluación del riesgo de crédito, ya que tiene una buena precisión de clasificación y es mejor que otros métodos en términos de AUC y precisión, lo que resulta en una mejora efectiva del proceso de toma de decisiones de los prestamistas. Análisis y conclusiones semejantes a lo mencionado anteriormente se presentaron en los documentos [12] y [23].

Técnicas similares fueron aplicadas en los documentos “A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction” [16] y “A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique” [37], donde los autores emplearon random forest (RF), XGBoost, Adaptive Boosting (AdaBoost), Categorical Boosting (CatBoost) y Light Gradient Boosting Machine (LightGBM) como algoritmos base de métodos envolventes e integrados para seleccionar características y usar estos algoritmos como clasificadores para predecir el incumplimiento de préstamos personales. La conclusión a la cual llegaron los autores fue que es mejor usar diferentes algoritmos en la selección y clasificación de características, donde específicamente, AdaBoost y CatBoost funcionan mejor entre todos los clasificadores. Por otro lado, en el segundo artículo mencionado, los autores llevaron a cabo un estudio donde se desarrolló un nuevo modelo de evaluación de riesgo crediticio de conjunto de aprendizaje profundo para tratar con datos crediticios desequilibrados. Primero, se desarrolló un método mejorado de técnica de sobremuestreo de minorías sintéticas (SMOTE) para superar las deficiencias conocidas, después de lo cual se combinó un nuevo método de clasificación de conjuntos de aprendizaje profundo con la red de memoria a largo-corto plazo (LSTM) y el impulso adaptativo (AdaBoost). Luego, se desarrolló un algoritmo para entrenar y aprender los datos crediticios procesados. Posteriormente, se emplearon el área bajo la curva (AUC), Kolmogorov-Smirnov (KS) y la prueba no paramétrica de Wilcoxon para comparar el rendimiento del modelo propuesto y otros modelos de calificación crediticia ampliamente utilizados en dos

conjuntos de datos crediticios desequilibrados. Los resultados de las pruebas experimentales indicaron que el modelo de conjunto de aprendizaje profundo propuesto era generalmente más competitivo que otros modelos al abordar problemas de evaluación de riesgo crediticio desequilibrados.

Cabe mencionar que dentro de las técnicas de machine learning que también han incursionado en la predicción financiera, se encuentran las redes neuronales, cuyos resultados de pronóstico vale la pena resaltar y son sobresalientes, justo como lo detallan los documentos “An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data scarcity” [15], “Evolutionary extreme learning machine with novel activation function for credit scoring” [40] y “Extreme Learning Machine for Credit Risk Analysis” [43], los cuales encontraron en el Extreme Learning Machine (ELM) la herramienta adecuada para la evaluación del riesgo de crédito de diversos conjuntos de datos, que finalmente compararon contra modelos como el bayesiano ingenuo, los árboles de decisión y el perceptrón multicapa. Los resultados de la simulación de las medidas estadísticas de rendimiento corroboraron que ELM supera a los clasificadores ingenuos de Bayes, árboles de decisión y perceptrón multicapa en un 1,8248%, 16,6346% y 5,8934%, respectivamente.

Finalmente, la literatura analizada también abarca técnicas de machine learning alternativas a las convencionalmente utilizadas en la predicción financiera, tal y como lo es el uso del NLP (Natural Language Processing) en el documento [29], donde se utiliza un algoritmo de aprendizaje automático supervisado para extraer información relevante de los informes anuales de diferentes bancos de 19 países europeos entre 2005 y 2017 y, examina si dicha información puede determinar el riesgo crediticio. En general, los autores concluyeron que los resultados tienen implicaciones para las empresas, los reguladores y los participantes del mercado que buscan evidencia sobre la credibilidad de los informes anuales para transmitir información relevante que refleje el riesgo crediticio real. Otra técnica en la cual se basaron los autores del documento “A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment” [33], consiste en la combinación de métodos de computación con el criterio de opinión de expertos, donde un algoritmo de aprendizaje automático no supervisado permite a los expertos crear uno o más escenarios de agrupamiento con respecto a las funciones, definiendo una cantidad deseada de agrupamientos por escenario, luego, se

tienen en cuenta las opiniones de los expertos para resolver un problema de calificación crediticia en forma de un problema de optimización sujeto a restricciones a través de métodos de computación basados en aprendizaje automático supervisado y algoritmos de optimización evolutiva. La segregación por etapas también fue evidenciada dentro de la literatura [21], donde se propone un enfoque de calificación del riesgo crediticio en dos etapas: en la etapa 1, se utiliza el enfoque de generalización apilada (apilamiento) para entrenar el modelo y determinar la probabilidad de incumplimiento. En la etapa 2, se seleccionan los préstamos que se predice que no serán impagos en la etapa 1, se genera un nuevo conjunto de datos, se crea un modelo de predicción de ganancias usando el algoritmo de apilamiento y se introduce la tasa interna de retorno (TIR) como medida de rentabilidad. Incluso, el autor Breeden, J.L. [31], realizó un documento a manera de encuesta que examina la extensa gama de métodos de aprendizaje automático y áreas de aplicación para el riesgo crediticio. En todo momento, se destacan preguntas abiertas, ideas para mejoras y un marco para pensar sobre cómo elegir el mejor método de aprendizaje automático para un problema específico.

De los documentos mencionados se puede evidenciar que, si bien existe variedad de técnicas y resultados, es claro que en la mayoría de las ocasiones las capacidades de pronóstico son superiores en las técnicas de machine learning en comparación con los métodos estadísticos tradicionales, en cuanto a la estimación de riesgo de crédito, esta tesis fue demostrada en el trabajo [38] al afirmarse que los modelos de aprendizaje automático brindan ganancias sustanciales en el poder y la precisión discriminatorios, en relación con los modelos estadísticos, sin embargo, múltiples son las aplicaciones y ejemplos para una población de retail como clientes de las entidades financieras, más lo que se pretende analizar es el efecto de tales metodologías sobre las propias entidades bancarias del sector financiero colombiano, desde el punto de vista de un agente institucional como otorgador de cupos de crédito para dichas entidades.

1.4 Metodología a utilizar

Con el fin de llevar a cabo el trabajo propuesto inicialmente se realizará la recolección y análisis de datos, donde obtendremos diferentes indicadores financieros para medir la

solidez de las entidades del sector financiero colombiano durante diferentes periodos de tiempo.

Posteriormente, se realizará el preprocesamiento y limpieza de datos, con el fin de obtener unos indicadores depurados y consistentes en el tiempo, para luego, realizar la fase de entrenamiento de los modelos de machine learning elegidos de acuerdo con la experiencia y lo observado en el estado del arte. Finalmente, se elegirán métricas de precisión de pronóstico y error, con el fin de seleccionar el mejor modelo entre los testeados y poder realizar conclusiones sustentadas acerca de los resultados obtenidos.

Tabla 1-1: Metodología de trabajo de grado.

Objetivos	Actividades	Indicador
1. Analizar los principales métodos de machine learning identificados en la literatura para la asignación de cupos de crédito	Elección de bases de datos para la búsqueda	Elaboración de los antecedentes / Número de artículos analizados
	Creación de string de búsqueda	
	Depuración de string de búsqueda	
	Elección de artículos referentes a riesgo crediticio con técnicas de machine learning	
	Análisis de métodos empleados	
2. Desarrollar un método basado en el análisis previo y en indicadores financieros seleccionados	Recolección de indicadores financieros	Definición del modelo de machine learning a implementar
	Preprocesamiento y limpieza de datos	
	Elección del método a implementar con base en el análisis del primer objetivo	
	Entrenamiento del modelo	
3. Validar el modelo propuesto en comparación con los principales métodos identificados en la literatura	Obtención de las métricas resultantes de los métodos de la literatura	Métricas de precisión cuantificables
	Obtención de las métricas de precisión de pronóstico y error del modelo elaborado	
	Comparación de las métricas	

1.5 Alcances del trabajo

- El modelo de riesgo de crédito desarrollado en este documento se aplicará a los establecimientos de crédito del sector financiero colombiano.
- Esta investigación buscará elaborar un modelo de riesgo de crédito basado en técnicas de machine learning.
- El modelo de riesgo de crédito desarrollado será comparado con otros modelos consultados en la literatura.
- El modelo de riesgo de crédito no será aplicable a personas naturales, únicamente a entidades vigiladas por la Superintendencia Financiera de Colombia.
- El modelo de riesgo de crédito desarrollado aplicará para entidades que cuenten con track record suficiente para la alimentación del modelo.

2. Marco Teórico

Riesgo: Riesgo es una palabra antigua y de uso común en muchas lenguas. En su uso corriente denota incertidumbre asociada a un evento futuro o a un evento supuesto. Una descripción con sentido común del término riesgo debería incluir las circunstancias que amenacen con disminuir la seguridad, el bienestar social, la salud, el bienestar y la libertad de una entidad determinada. Esta descripción no apunta a definiciones técnicas o específicas del riesgo, pero ejemplifica el rango de aplicaciones que posee ese término y aclara que el concepto de riesgo está estrechamente ligado a valores humanos significativos [60].

El riesgo puede consistir en la mera posibilidad de un hecho adverso, en la causa de un evento, en la magnitud de la consecuencia, en alguien o algo considerado como peligroso y también en la conceptualización de un procedimiento para la estimación de una cantidad [60].

Riesgo de crédito: Posibilidad de que una entidad incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que un deudor o contraparte incumpla sus obligaciones [6].

Metodología CAMEL: La metodología CAMEL evalúa la solidez financiera de las empresas con base en indicadores cuantitativos, contemplando cinco áreas: capital adecuado (C), calidad del activo (A), capacidad de la gerencia (M), rentabilidad (E) y situación de liquidez (L) [59]. En el marco de la regulación y el sector financiero colombianos, las variables para la medición de las cinco áreas más utilizadas son:

- Capital

$$\text{Solvencia VaR} = \frac{\text{Patrimonio Técnico}}{\text{Activos Ponderados por Riesgo} + (\text{Riesgo Mercado} * 100/9)}$$

$$\text{Índice de capacidad} = \frac{\text{Patrimonio Básico}}{\text{Monto total de los créditos para el plazo Overnight}}$$

- Activos

$$\text{Indice de calidad de la cartera} = \frac{\text{Cartera Calificada en (BCDE)}}{\text{Cartera Bruta Total}}$$

$$\text{Indice de cartera vencida} = \frac{\text{Cartera vencida total (por altura de mora)}}{\text{Cartera Bruta Total}}$$

$$\text{Indice de cubrimiento de cartera} = \frac{\text{Provisiones totales de la cartera}}{\text{Cartera Vencida total (por altura de mora)}}$$

- Calidad de administración

$$\text{Calidad de la Administración} = \frac{\text{Costos Administrativos mensuales}}{\text{Margen Financiero Bruto mensuales}}$$

$$\text{Cubrimiento Financiero} = \frac{\text{Gastos por Intereses mensuales}}{\text{Ingresos por Intereses mensuales}}$$

- Rentabilidad

$$ROA = \left[\left(\frac{\text{Utilidad}}{\text{Activos}} + 1 \right)^{\frac{12}{m}} - 1 \right] * 100$$

$$ROE = \left[\left(\frac{\text{Utilidad}}{\text{Patrimonio}} + 1 \right)^{\frac{12}{m}} - 1 \right] * 100$$

- Liquidez

$$\text{Indicador de Liquidez} = \frac{\text{Indicador de Riesgo de Liquidez a siete días}}{\text{Activos Líquidos Netos}}$$

Solvencia: Capacidad de la empresa para cumplir con sus obligaciones de pago, sin importar cuándo tenga que asumir ese pago. Las empresas serán más o menos solventes cuando puedan mantener durante más tiempo los recursos suficientes para hacer frente a

sus costes. Una empresa puede tener un soporte para asumir sus obligaciones de pago en el corto plazo, lo que le permite ser estable. No obstante, cuanto más tiempo pueda asumir esas obligaciones de pago, mayor será su estabilidad [7].

$$\text{Ratio de solvencia} = \frac{\text{Activo}}{\text{Pasivo}}$$

De acuerdo con la normatividad colombiana y, aplicando para los establecimientos de crédito, estos deben cumplir con niveles de patrimonio adecuados y con relaciones mínimas de solvencia con el fin de proteger la confianza del público en el sistema y asegurar su desarrollo en condiciones de seguridad y competitividad [58]. El Decreto Único 2555 de 2010 define la relación de solvencia total como el valor del patrimonio técnico dividido por el valor de los activos ponderados por nivel de riesgo crediticio, de mercado y operacional, la cual debe ser mínimo del 9%, mientras que la relación de solvencia básica se define como el valor del patrimonio básico ordinario neto de deducciones dividido por el valor de los activos ponderados por nivel de riesgo crediticio, de mercado y operacional, la cual debe ser de mínimo el 4.5%.

$$\text{Relación de Solvencia Total} = \frac{\text{Patrimonio Técnico}}{\text{APNR}}$$

$$\text{Relación de Solvencia Básica} = \frac{\text{Patrimonio Básico Ordinario neto de deducciones}}{\text{APNR}}$$

Rentabilidad: Variación de la riqueza en un horizonte de inversión y la cual se mide como la diferencia entre el precio un activo en entre el precio de la acción en el momento t y el precio en el momento t - 1, dividido por el precio en el momento t – 1 [8].

$$R = \frac{P_t - P_{t-1}}{P_{t-1}}$$

Machine Learning: Es una forma de la IA que permite a un sistema aprender de los datos en lugar de aprender mediante la programación explícita. Sin embargo, machine learning no es un proceso sencillo. Conforme el algoritmo ingiere datos de entrenamiento, es

posible producir modelos más precisos basados en datos. Un modelo de machine learning es la salida de información que se genera cuando entrena su algoritmo de machine learning con datos. Después del entrenamiento, al proporcionar un modelo con una entrada, se le dará una salida. Por ejemplo, un algoritmo predictivo creará un modelo predictivo. A continuación, cuando proporcione el modelo predictivo con datos, recibirá un pronóstico basado en los datos que entrenaron al modelo [9].

Sobreajuste: Comportamiento de aprendizaje automático no deseado que se produce cuando el modelo de aprendizaje automático proporciona predicciones precisas para los datos de entrenamiento, pero no para los datos nuevos. Cuando los científicos de datos utilizan modelos de aprendizaje automático para hacer predicciones, primero entrenan el modelo en un conjunto de datos conocido. Luego, basándose en esta información, el modelo intenta predecir los resultados para los nuevos conjuntos de datos. Un modelo sobreajustado puede proporcionar predicciones inexactas y no puede funcionar bien para todos los tipos de datos nuevos [61].

3. Análisis de los principales métodos de machine learning identificados en la literatura para la asignación de cupos de crédito

De acuerdo con el análisis realizado sobre la literatura y desarrollado en el numeral 1.3, se observó que los métodos de machine learning más utilizados en el campo de la predicción del riesgo crediticio son: Regresión Logística, Support Vector Machines, Redes Neuronales, K-Nearest Neighbors, Árboles de Decisión y Random Forest. A continuación, se describe cada uno de los métodos mencionados y cómo se aplicaron en la literatura consultada:

Regresión Logística: es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor. Es importante tener en cuenta que, aunque la regresión logística permite clasificar, se trata de un modelo de regresión que modela el logaritmo de la probabilidad de pertenecer a cada grupo. La asignación final se hace en función de las probabilidades predichas.

Si una variable cualitativa con dos niveles se codifica como 1 y 0, matemáticamente es posible ajustar un modelo de regresión lineal por mínimos cuadrados $\beta_0 + \beta_1 x$. El problema de esta aproximación es que, al tratarse de una recta, para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, lo que entra en contradicción con el hecho de que las probabilidades siempre están dentro del rango [0,1], por lo cual, se emplea la regresión logística en lugar de la regresión lineal [49].

Específicamente en la literatura consultada, este método se aplicó en un caso de puntajes crediticios para empresas que operaban en tres sectores de la industria y que solicitaron préstamos al banco público más grande de Turquía [20]. En este estudio se analizaron

1881 empresas y los resultados del experimento se compararon en términos de precisión de clasificación, sensibilidad, especificidad, precisión y coeficiente de correlación de Mathews, donde se observó que la regresión logística obtuvo los mejores rendimientos, incluso por encima de los árboles de decisión y KNN para todos los conjuntos de datos.

Support Vector Machines: una máquina de vectores de soporte (SVM) es un algoritmo de aprendizaje supervisado que se puede emplear para clasificación binaria o regresión, construye un hiperplano óptimo en forma de superficie de decisión, de modo que el margen de separación entre las dos clases en los datos se amplía al máximo. Los vectores de soporte hacen referencia a un pequeño subconjunto de las observaciones de entrenamiento que se utilizan como soporte para la ubicación óptima de la superficie de decisión [50].

SVM tiene varias ventajas, como la capacidad de manejar eficientemente conjuntos de datos de alta dimensionalidad, la capacidad de manejar datos no lineales y la capacidad de controlar el ajuste del modelo mediante la elección del parámetro de regularización. Sin embargo, también tiene algunas limitaciones, como la sensibilidad a la escala de los datos y la necesidad de seleccionar cuidadosamente los parámetros del modelo [51].

Mediante la revisión de la literatura se observó que este método se aplicó en el desarrollo de un modelo de riesgo de crédito para títulos de deuda de la India calificados por las principales agencias de calificación crediticia utilizando la regresión logística ordinal y mejorando la precisión de este mediante el uso de Support Vector Machines. El resultado arrojó una precisión de 90% a través de esta metodología mientras que únicamente con la regresión logística ordinal dicha métrica alcanzó el 68% [34].

De igual manera, las máquinas de soporte vectorial obtuvieron resultados destacables en el estudio mencionado anteriormente y relacionado con el análisis de puntajes crediticios de un conjunto de empresas que solicitaron préstamos a un banco de Turquía [20].

Redes Neuronales: una red neuronal es un método de la inteligencia artificial que enseña a las computadoras a procesar datos de una manera que está inspirada en la forma en que lo hace el cerebro humano. Se trata de un tipo de proceso de machine learning llamado

aprendizaje profundo, que utiliza los nodos o las neuronas interconectados en una estructura de capas que se parece al cerebro humano. Crea un sistema adaptable que las computadoras utilizan para aprender de sus errores y mejorar continuamente [52].

Algunas de las ventajas que presenta este método son aprendizaje adaptativo, ya que poseen la capacidad de aprender a realizar tareas basadas en un entrenamiento o en una experiencia inicial, auto-organización, puesto que una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje, tolerancia a fallos, pues la destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño, operación en tiempo real, debido a que los cómputos neuronales pueden ser realizados en paralelo, para esto se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad y, fácil inserción dentro de la tecnología existente, ya que se pueden obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas, ello facilitará la integración modular en los sistemas existentes [53].

K-Nearest Neighbors: es un algoritmo basado en instancia de tipo supervisado de Machine Learning. Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos). Sirve esencialmente para clasificar valores buscando los puntos de datos “más similares” (por cercanía) aprendidos en la etapa de entrenamiento y haciendo conjeturas de nuevos puntos basado en esa clasificación. Es un método que busca en las observaciones más cercanas a la que se está tratando de predecir y clasifica el punto de interés basado en la mayoría de datos que le rodean. Es un algoritmo supervisado, es decir, se tiene etiquetado el conjunto de datos de entrenamiento, con la clase o resultado esperado dada “una fila” de datos, así mismo, es basado en instancia, lo cual significa que el algoritmo no aprende explícitamente un modelo (como por ejemplo en Regresión Logística o árboles de decisión), en cambio, memoriza las instancias de entrenamiento que son usadas como “base de conocimiento” para la fase de predicción.

La ventaja de este método es que es sencillo de aprender e implementar, sin embargo, como desventaja, utiliza todo el dataset para entrenar “cada punto”, por lo cual requiere de uso de mucha memoria y recursos de procesamiento (CPU). Por estas razones kNN tiende a funcionar mejor en datasets pequeños y sin una cantidad enorme de features [54].

Árboles de Decisión: un árbol de decisión es un algoritmo supervisado de aprendizaje automático que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Para dividir el espacio muestral en sub-regiones es preciso aplicar una serie de reglas o decisiones, para que cada sub-región contenga la mayor proporción posible de individuos de una de las poblaciones. Si una sub-región contiene datos de diferentes clases, se subdivide en regiones más pequeñas hasta fragmentar el espacio en sub-regiones menores que integran datos de la misma clase. El tipo de problema a resolver dependerá de la variable a predecir, si es una variable dependiente se estaría ante un problema de regresión, mientras que, si es una variable categórica, se afrontaría un problema de clasificación. Algunas ventajas que posee este método son las siguientes [55]:

- Son fáciles de construir, interpretar y visualizar.
- Selecciona las variables más importantes y en su creación no siempre se hace uso de todos los predictores.
- Si faltan datos no se puede recorrer el árbol hasta un nodo terminal, pero sí se pueden hacer predicciones promediando las hojas del sub-árbol que se alcance.
- No es preciso que se cumplan una serie de supuestos como en la regresión lineal (linealidad, normalidad de los residuos, homogeneidad de la varianza, etc.).
- Sirven tanto para variables dependientes cualitativas como cuantitativas, como para variables predictoras o independientes numéricas y categóricas. Además, no necesita variables dummies, aunque a veces mejoran el modelo.
- Permiten relaciones no lineales entre las variables explicativas y la variable dependiente.

De otro lado, algunas desventajas que poseen los árboles de decisiones son:

- Tienen al sobreajuste u overfitting de los datos, por lo que el modelo al predecir nuevos casos no estima con el mismo índice de acierto.
- Se ven influenciadas por los outliers, creando árboles con ramas muy profundas que no predicen bien para nuevos casos. Se deben eliminar dichos outliers.
- No suelen ser muy eficientes con modelos de regresión.
- Crear árboles demasiado complejos puede conllevar que no se adapten bien a los nuevos datos. La complejidad resta capacidad de interpretación.
- Se pueden crear árboles sesgados si una de las clases es más numerosa que otra.
- Se pierde información cuando se utilizan para categorizar una variable numérica continua.

En la literatura consultada se observó que este método se aplicó en un estudio sobre la predicción de quiebra en los bancos de la Unión Europea [10], obteniendo como resultado la importancia de las utilidades, la adecuación del capital y la capacidad de gestión en los modelos de predicción de quiebras bancarias. Adicionalmente, el árbol de decisión utilizado obtuvo los mejores indicadores de rendimiento. Así mismo, en un trabajo aplicado al riesgo crediticio en Colombia [5], donde se buscaba soportar los procesos de las áreas de riesgo de los bancos al identificar clientes que podrían incurrir en un estado de mora, prediciendo el próximo pago de la cuota a partir de datos básicos de la operación, del cliente y de pagos de cuotas anteriores registradas, se obtuvo que los árboles de decisión resultan ser más exactos que otras técnicas para la predicción del riesgo crediticio con un área bajo la curva ROC de 88.29%.

Random Forest: un bosque aleatorio es un conjunto de árboles de decisión combinados con bagging. Al usar bagging, distintos árboles de decisión ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y se tiene una predicción que generaliza mejor [56].

Al aplicar el algoritmo de random forest para problemas de clasificación o regresión, se presentan una serie de ventajas y desafíos clave [57]. Las ventajas son las siguientes:

1. Riesgo reducido de sobreajuste: Los árboles de decisión corren el riesgo de sobreajustarse, ya que tienden a ajustar todas las muestras dentro de los datos de entrenamiento. Sin embargo, cuando hay una gran cantidad de árboles de decisión en un random forest, el clasificador no se ajustará demasiado al modelo, ya que el promedio de árboles no correlacionados reduce la varianza general y el error de predicción.
2. Aporta flexibilidad: Dado que el random forest puede manejar tareas de regresión y clasificación con un alto grado de precisión, es un método popular entre los científicos de datos. El agrupamiento de características también convierte al clasificador de random forest en una herramienta eficaz para estimar los valores perdidos, ya que mantiene la precisión cuando falta una parte de los datos.
3. Importancia de la característica fácil de determinar: El random forest facilita la evaluación de la importancia o contribución de las variables al modelo. Hay algunas formas de evaluar la importancia de las características. La importancia de Gini y la disminución media de impurezas (MDI) se utilizan generalmente para medir cuánto disminuye la precisión del modelo cuando se excluye una variable determinada. Sin embargo, la importancia de la permutación, también conocida como precisión de disminución media (MDA), es otra medida de importancia. MDA identifica la disminución promedio en la precisión mediante la permutación aleatoria de los valores de las características en las muestras para pruebas.

Por su parte, algunos desafíos que enfrenta este método se listan a continuación:

1. Proceso que requiere mucho tiempo: Dado que los algoritmos de random forest pueden manejar grandes conjuntos de datos, pueden proporcionar predicciones más precisas, pero pueden ser lentos para procesar los datos, ya que están computando datos para cada árbol de decisión individual.

2. Requiere más recursos: Dado que los random forest procesan conjuntos de datos más grandes, requerirán más recursos para almacenar esos datos.
3. Más complejo: La predicción de un único árbol de decisiones es más fácil de interpretar en comparación con un bosque de ellos.

Al verificar el estado del arte investigado, se encontró que random forest fue aplicado en un modelo de riesgo de crédito para títulos de deuda de la India calificados por las principales agencias de calificación crediticia utilizando la regresión logística ordinal y mejorando la precisión de este mediante el uso de random forest. El resultado arrojó una precisión de 90% a través de esta metodología [34]. De igual manera, se observó que en un estudio que comparó el modelo de regresión logística frente a diversos modelos de machine learning, para la estimación del riesgo de crédito en una cartera de consumo con el fin de estimar los clientes en entrarían en mora, se obtuvo como resultado que el modelo más equilibrado es el random forest, presentando mejor ajuste de acuerdo con diversas métricas de exactitud evaluadas [3]. Finalmente, este método también obtuvo resultados destacables en el estudio relacionado con el análisis de puntajes crediticios de un conjunto de empresas que solicitaron préstamos a un banco de Turquía [20].

Adicionalmente, con el fin de contar con criterios más estructurados para la profundización en un método de machine learning, se aplicaron 4 criterios ponderados con el fin de obtener una calificación acerca de la elección del método:

- Criterio # 1: Número de veces en las cuales se aplicó un modelo en la literatura revisada.
- Criterio # 2: Número de veces en las que el modelo obtuvo el mejor resultado en términos de precisión en la literatura revisada.
- Criterio # 3: Mejor resultado obtenido con base en el artículo “A survey of machine learning in credit risk” de Breeden, J.L. de 2021 [31], el cual realiza un comparativo de diversos métodos de machine learning.
- Criterio # 4: Criterio de expertos.

A cada uno de los criterios se les otorgó un peso individual con el fin de obtener una calificación ponderada final, siendo de 15% para el criterio 1, 35% para el criterio 2, 20% para el criterio 3 y 30% para el criterio 4. Posteriormente, se evaluaron cada uno de los métodos bajo análisis según los criterios expuestos y se estandarizaron en una calificación de 1 a 10, siendo 1 la peor calificación y 10 la mejor calificación. En la tabla 3-1 se puede observar el ranking de calificaciones donde random forest y los árboles de decisión fueron los modelos mejor calificados con el objetivo de realizar predicciones de riesgo de crédito.

Tabla 3-1: Ranking de calificación de métodos de machine learning.

	Criterio # 1	Criterio # 2	Criterio # 3	Criterio # 4	Calificación	Ranking
Regresión Logística	5,0	5,0	1,7	8,3	5,3	3
Support Vector Machines	5,0	6,7	6,7	1,7	4,9	4
Redes Neuronales	3,3	1,7	3,3	6,7	3,8	5
K-Nearest Neighbors	1,7	1,7	8,3	1,7	3,0	6
Árboles de decisión	8,3	8,3	10,0	1,7	6,7	2
Random Forest	10,0	10,0	5,0	10,0	9,0	1

Con respecto al criterio 2, son diversas las métricas que se mencionan en la literatura consultada, las cuales buscan evaluar el rendimiento de los modelos de machine learning, a continuación, se detalla la explicación de algunas de dichas métricas:

- **Precisión (Precision):** Se refiere a la calidad del modelo de machine learning y mide la proporción de predicciones positivas que fueron realmente correctas.

$$Precision = \frac{VP}{VP + FP}$$

Donde:

VP = Verdaderos positivos (predicciones positivas correctas)

FP = Falsos Positivos (predicciones positivas incorrectas)

- **Exhaustividad (Recall):** Indica la cantidad que el modelo de machine learning es capaz de identificar y mide la proporción de predicciones positivas que fueron correctamente identificadas por el modelo.

$$Recall = \frac{VP}{VP + FN}$$

Donde:

VP = Verdaderos positivos (predicciones positivas correctas)

FN = Falsos negativos (predicciones positivas incorrectamente clasificadas como negativas)

- Valor F1 (F1 Score): Combina las medidas de precisión y exhaustividad en un solo valor, lo cual permite comparar más fácilmente y de forma equilibrada el rendimiento combinado del modelo.

$$F1 = 2 \frac{precision \times recall}{precision + recall}$$

- Exactitud (Accuracy): Se refiere al porcentaje de casos que el modelo ha acertado, es decir, la proporción de predicciones correctas realizadas por el modelo sobre el total de predicciones.

$$Accuracy = \frac{VP + VN}{VP + VN + FP + FN}$$

Donde:

VP = Verdaderos positivos (predicciones positivas correctas)

VN = Verdaderos negativos (predicciones negativas correctas)

FP = Falsos Positivos (predicciones positivas incorrectas)

FN = Falsos negativos (predicciones positivas incorrectamente clasificadas como negativas)

Sin embargo, existe un inconveniente con la métrica de exactitud en cuanto a que su funcionamiento puede no ser confiable cuando las clases se encuentran

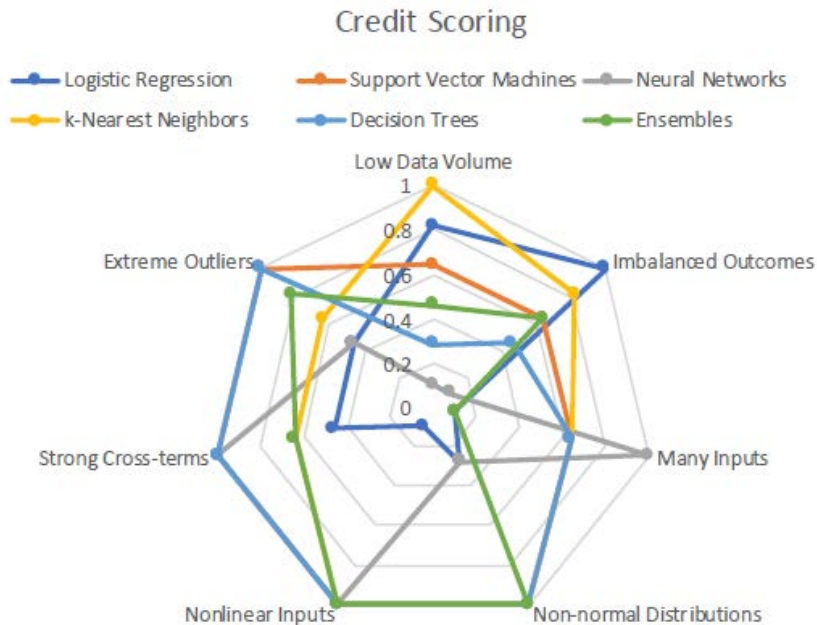
desbalanceadas, es decir, que el conjunto de datos no tiene una distribución equitativa de observaciones entre las diferentes clases, por lo cual se hace necesario tomar medidas para que este tipo de situaciones no afecte al modelo en términos de precisión [47].

- Curva ROC-AUC (Receiver Operating Characteristic - Area Under the Curve): La curva ROC permite evaluar el rendimiento de los algoritmos de clasificación binaria (dos clases o categorías), proporcionando una representación gráfica donde se plasma la tasa de verdaderos positivos (TVP) en el eje y, contra la tasa de falsos positivos (TFP) en el eje x, para un solo clasificador en una variedad de umbrales, con el fin de observar el rendimiento del clasificador. Cada punto en la curva ROC representa un par (TFP, TVP) correspondiente a un umbral de clasificación específico.

Complementando la curva ROC, se encuentra la métrica AUC que es el área bajo la curva ROC, la cual revela una medida cuantitativa del rendimiento global del modelo y cuanto sea mayor, mejor será el rendimiento de un clasificador binario para una tarea de clasificación dada y mejor será la capacidad de discriminación entre las clases. En otras palabras, se puede considerar como la probabilidad de que el modelo clasifique un ejemplo positivo de manera más alta que un ejemplo negativo elegido al azar [48].

En relación al criterio 3 y de acuerdo con [31], se realizó una comparación intuitiva de las fortalezas y debilidades potenciales de varios métodos de aprendizaje automático para la calificación crediticia, evaluando particularmente el bajo volumen de datos, resultados desequilibrados, la cantidad de inputs, distribuciones no normales, inputs no lineales, términos cruzados fuertes y valores atípicos extremos, obteniendo resultados entre 0 siendo el más débil y 1 el más fuerte en el gráfico 3-1.

Figura 3-1: Fortalezas y debilidades en métodos de machine learning.



Es necesario aclarar, de acuerdo con la literatura revisada, que catalogar un método de machine learning como mejor o peor que otro, puede resultar inapropiado e incluso inexacto, si no se realizan las evaluaciones bajo un conjunto de criterios estandarizados, tal y como se menciona en [31]: "...Los métodos tienen fortalezas y debilidades específicas que se alinean con diferentes aplicaciones. En una aplicación específica, el mejor método suele implicar la combinación de elementos de varios métodos, tanto estadísticos como de aprendizaje automático".

Con respecto al criterio 4, tal y como menciona [45], el criterio de expertos permite darle validez a una investigación a través de un juicio confiable, definiéndolo como "una opinión informada de personas con trayectoria en el tema, que son reconocidas por otros como expertos cualificados en este y que pueden dar información, evidencia, juicios y valoraciones". Es así como, se realizó una encuesta a diversos científicos de datos acerca de sus experiencias y opiniones trabajando con los métodos bajo análisis, donde, cómo se observó en la tabla 3-1, Random Forest fue el método de machine learning de mayor preferencia para los expertos en analítica frente a sus capacidades de predicción. Es

válido, entonces, argumentar que este criterio posea uno de los mayores pesos al reposar su fiabilidad en un método que enumera diversas ventajas, como el nivel de profundización de la valoración que ofrece, su facilidad de puesta en acción, la no exigencia de muchos requisitos técnicos y humanos para su ejecución, el poder utilizar diferentes estrategias para recoger la información, lo cual es de gran utilidad para determinar el conocimiento sobre contenidos y temáticas difíciles, complejas y novedosas o poco estudiadas, y la posibilidad de obtener información pormenorizada sobre el tema sometido a estudio [46].

4. Método propuesto para el análisis previo y selección de indicadores financieros

4.1 Insumos del modelo

La presente investigación se basa en la recopilación de información directamente extraída de la sección de indicadores gerenciales de la Superintendencia Financiera de Colombia. El análisis se centra en diversas entidades financieras, abarcando establecimientos de crédito, corporaciones financieras, compañías de financiamiento comercial e Instituciones Oficiales Especiales (IOE's). El periodo considerado para este estudio abarca desde enero de 2015 hasta abril de 2023, con el objetivo de evaluar la salud y el desempeño financiero de cada entidad.

El proceso de análisis de datos incluyó más de 50 indicadores financieros, permitiendo una evaluación integral de cada institución. Para garantizar la validez y consistencia de los resultados, se llevaron a cabo tratamientos específicos frente a la imputación de datos. Estos tratamientos incluyeron:

- Eliminación de filas correspondientes a bancos que desaparecieron, con el propósito de eliminar valores nulos y datos en cero.
- Exclusión de filas correspondientes a años iniciales con información en cero para los bancos más recientes que aún no existían en esos periodos.
- Eliminación de filas en las fechas en las que no se disponía de datos para las IOE's.
- Para el indicador de Plusvalía, donde no se contaba con datos desde enero hasta mayo de 2015, se asignó el primer dato existente de cada banco.

- Asignación de cero al indicador de activos líquidos/pasivos con costo, para aquellas entidades con pasivos con costo en cero, con el fin de evitar resultados indeterminados.
- Establecimiento de cero en el indicador de eficiencia financiera (gastos financieros/ingresos financieros de cartera) para todas las entidades con ingresos financieros de cartera igual a cero, evitando así resultados indeterminados.
- Para entidades sin registros de solvencia y patrimonio técnico, se asignó un valor de cero a la solvencia y el valor del patrimonio al patrimonio técnico. En casos de datos puntuales faltantes, se incluyó la información del mes más cercano.
- Con el objetivo de mejorar la integridad y confiabilidad del modelo, se excluyeron entidades con demasiada poca información con el fin de no obstaculizar el aprendizaje del modelo.

A continuación, se presentan los indicadores y rubros que se tuvieron en cuenta en el modelo de riesgo de crédito elaborado:

Tabla 4-1: Indicadores financieros y rubros como insumos del modelo.

CALIDAD	OPERACIONES DE TRANSFERENCIA TEMPORAL DE VALORES
ACTIVOS/PATRIMONIO	INSTRUMENTOS FINANCIEROS A VALOR RAZONABLE
CUBRIMIENTO	CRÉDITOS DE BANCOS Y OTRAS OBLIGACIONES FINANCIERAS
UTILIDAD/ACTIVO	TÍTULOS DE INVERSIÓN EN CIRCULACIÓN
UTILIDAD/PATRIMONIO	PLUSVALÍA
UTILIDAD / INGRESO FINANCIERO (DUPONT)	OTROS ACTIVOS ACTIVOS
ACTIVOS	INGRESOS FINANCIEROS CARTERA
PASIVOS	INTERESES DEPÓSITOS Y EXIGIBILIDADES
PATRIMONIO	% COMERCIAL DE LA CARTERA
GANANCIAS (EXCEDENTES) Y PÉRDIDAS	% CONSUMO DE LA CARTERA
CARTERA DE CREDITOS Y OPERACIONES DE LEASING	% VIVIENDA DE LA CARTERA
MARGEN NETO DE INTERESES = IFI-GI	% MICROCRÉDITO DE LA CARTERA
MARGEN FINANCIERO BRUTO-MFB = IFDI-GFDI	ACTIVOS LÍQUIDOS/PASIVO CON COSTO

MARGEN OPERACIONAL ANTES DE DEPR Y AMORT-MOADA =MFB-CA-DNR	ACTIVOS LÍQUIDOS/ACTIVOS TOTALES
MARGEN OPERACIONAL NETO DESPUES DE DEPR Y AMORT-MODDA= MOADA-DA	FONDEO DE MERCADO/ACTIVOS TANGIBLES
BENEFICIOS A EMPLEADOS	GASTOS FCROS DEPÓSITOS/INGRESOS FCROS CARTERA
GASTOS DE OPERACIONES	GASTOS LABORALES/MARGEN OPERACIONAL NETO DESPUÉS DE DEPR Y AMORT
INGRESOS DE OPERACIONES	SOLVENCIA BÁSICA
DISPONIBLE	PATRIMONIO TÉCNICO
POSICIONES ACTIVAS EN OPERACIONES DE MERCADO MONETARIO Y RELACIONADAS	INDICADOR DE CALIDAD DE CARTERA (CON CASTIGOS)
INVERSIONES Y OPERACIONES CON DERIVADOS	INDICADOR DE CUBRIMIENTO (CON CASTIGOS)
CUENTAS POR COBRAR	ACTIVOS/ INGRESO FINANCIERO
PASIVOS CON COSTO	QUEBRANTO PATRIMONIAL (PATRIMONIO/(CAPITAL SOCIAL + CAPITAL GARANTIA))
OPERACIONES DE REPORTO O REPO	FINANCIAMIENTO CON PASIVOS DE LARGO PLAZO (PasCP-ActCP / ActLP)
OPERACIONES SIMULTÁNEAS	RENDIMIENTO ACUMULADO DE LA CARTERA

4.2 Preprocesamiento de los datos

Con el fin de llevar a cabo un análisis de datos y predictivo acerca de un modelo de riesgo de crédito, inicialmente se da la importación de las bibliotecas respectivas para este fin, incorporando bibliotecas esenciales como Pandas para manipulación de datos, NumPy para operaciones matriciales y estadísticas, Matplotlib para visualización, Scikit-Learn para aprendizaje automático, y Statsmodels para análisis estadísticos avanzados. Posteriormente, se define la función 'series_to_supervised', la cual desempeña un papel crucial al convertir series temporales en un formato apto para el aprendizaje supervisado. Esta transformación es esencial para entrenar modelos de aprendizaje automático con datos secuenciales, proporcionando al modelo información contextual del pasado para prever eventos futuros.

Con respecto a la carga y preprocesamiento de datos, primeramente, se cargan estos desde un archivo Excel, empleando Pandas para estructurar y explorar la información. Se destaca la conversión de la columna de fechas a un formato temporal, un paso crucial para el análisis temporal. Además, se realiza la codificación de etiquetas y la normalización de ciertas variables numéricas para estandarizar la magnitud de los datos. La normalización de datos se ejecuta con Min-Max Scaling, una técnica que escala los valores entre 0 y 1, preservando las relaciones proporcionales entre ellos. Este procedimiento garantiza que las diferencias en la escala no afecten el rendimiento del modelo.

Luego, frente a la preparación de datos para modelos de aprendizaje automático, la función 'series_to_supervised' es nuevamente empleada para la preparación específica de datos destinados a un modelo de clasificación de Bosques Aleatorios (Random Forest). Esta preparación es esencial para crear un conjunto de datos supervisado que vincula las observaciones pasadas con las futuras, permitiendo al modelo aprender patrones temporales y realizar predicciones.

Seguidamente, los datos preparados se dividen en conjuntos de entrenamiento y prueba, un proceso crítico para evaluar la capacidad del modelo para generalizar a datos no vistos. La estrategia utilizada es la división clásica del 80-20, donde el 80% se destina al entrenamiento y el 20% a la evaluación. De igual manera, se separan las características (X) de las etiquetas (y). Esto establece claramente qué variables serán utilizadas para predecir y cuáles serán las predicciones, en el caso particular de este proyecto, se emplean como características los 50 indicadores financieros detallados previamente en la tabla 3-2. Estos indicadores abarcan diversas dimensiones que proporcionan una visión integral de la salud y el rendimiento de las entidades financieras. La variable predictora clave en este análisis es una nueva columna binaria, donde el valor 0 denota que la entidad ha caído en default y el valor 1 indica que no ha experimentado tal situación. Es crucial destacar que el concepto de default se define en el contexto colombiano y de este trabajo como el estado en el cual las entidades financieras han dejado de operar históricamente, ya sea a través de fusiones, adquisiciones por otras entidades, o cese de operaciones. Específicamente, este estado de default se evalúa en momentos en los cuales los indicadores financieros de las entidades se encuentran en un estado deteriorado. De esta manera, la binarización de la variable predictora refleja con precisión la capacidad del

modelo para prever eventos críticos en el panorama financiero, proporcionando insights valiosos para la gestión de riesgos crediticios en el contexto colombiano.

Adicionalmente, se implementó una estrategia integral destinada a mejorar la calidad y eficiencia del conjunto de datos. En este sentido, se llevó a cabo una cuidadosa reducción de dimensionalidad y regularización de la información inicial. Este procedimiento tuvo como objetivo principal la eliminación de features redundantes o que no aportaban información sustancial al modelo, contribuyendo así a la simplificación y optimización de la estructura del conjunto de características.

Durante la reducción de dimensionalidad, se emplearon técnicas específicas para identificar y retener únicamente aquellas variables que mejor capturaban la variabilidad presente en los datos. Este paso permitió conservar las características más relevantes, descartando aquellas que no añadían valor significativo al análisis. Además, se implementó un proceso de regularización destinado a mitigar el riesgo de overfitting, una problemática común en modelos predictivos complejos.

La limpieza de características no esenciales se llevó a cabo mediante la eliminación selectiva de columnas del conjunto de características, basándose en criterios de relevancia y contribución al rendimiento del modelo. Este enfoque no solo mejoró la eficacia del algoritmo de Random Forest al reducir la complejidad del conjunto de datos, sino que también favoreció la interpretación de los resultados al enfocarse en las variables más influyentes.

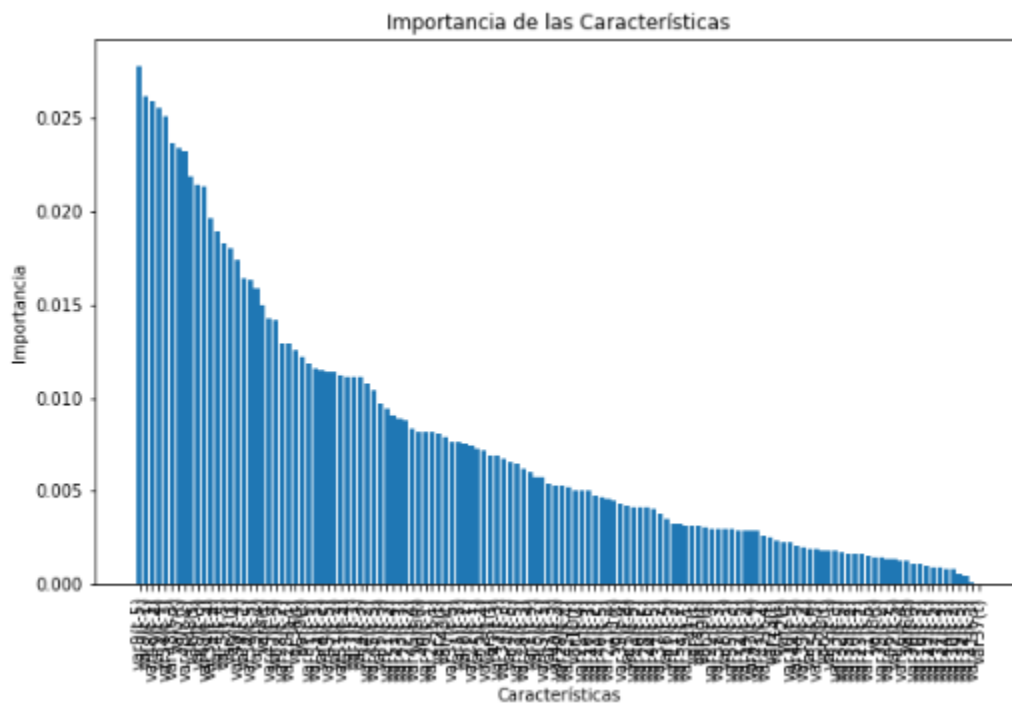
4.3 Modelo de Random Forest

Para continuar con la implementación del modelo de clasificación de riesgo crediticio mediante Random Forest, se emplea el conjunto de datos de entrenamiento para capacitar al modelo, seguido por la evaluación de su desempeño en el conjunto de prueba. La precisión del modelo obtenida fue de 99.9% en el conjunto de datos de testeo y de 99.79%

en el conjunto de datos de entrenamiento y, siendo la precisión la proporción de predicciones correctas entre el total de predicciones, el resultado indica que el modelo tiene una capacidad excepcional para predecir correctamente las etiquetas de las instancias en ambos conjuntos de datos, además, la precisión en los datos de entrenamiento indica que el modelo ha aprendido de manera efectiva los patrones y relaciones en los datos de entrenamiento, mostrando un rendimiento muy alto incluso en el conjunto con el cual fue entrenado.

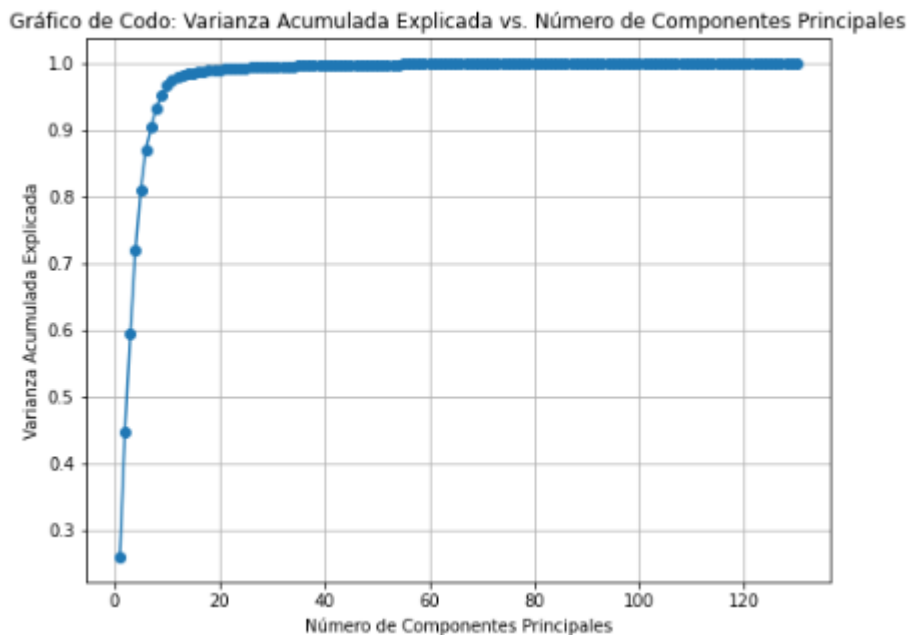
Una faceta fundamental de este análisis radica en la interpretación de la importancia de las características. Se utiliza el atributo `feature_importances_` del modelo para clasificar las variables según su contribución al proceso de toma de decisiones del modelo. Este análisis, crucial para comprender qué aspectos son más influyentes, se visualiza de manera intuitiva mediante un gráfico de barras, proporcionando una visión clara de las variables que más afectan las predicciones del modelo.

Figura 4-1: Importancia de las características.



Además, se incorpora un análisis de componentes principales (PCA), una técnica de reducción de dimensionalidad que ayuda a entender la estructura subyacente de los datos. El gráfico de codo generado revela la cantidad óptima de componentes principales a retener para explicar la mayor cantidad posible de la variabilidad en los datos, un paso esencial para optimizar la eficiencia del modelo.

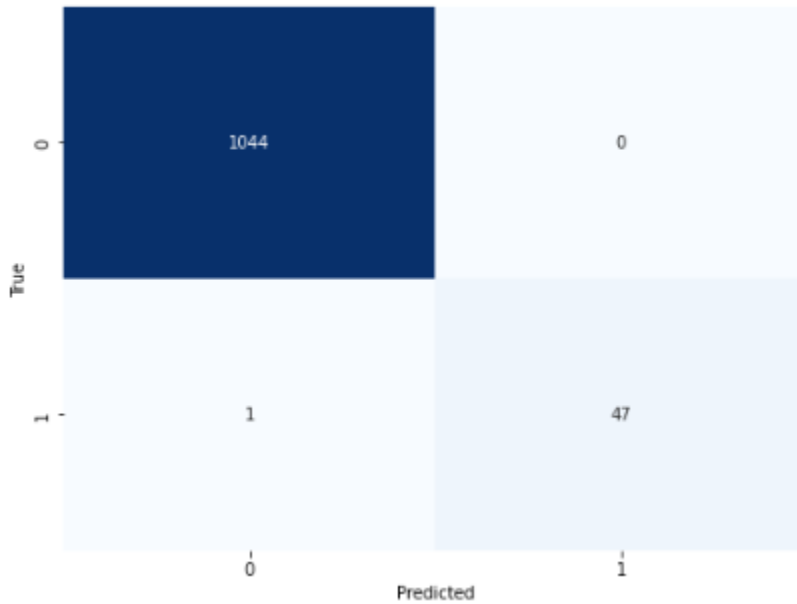
Figura 4-2: Gráfico de codo.



La exploración se amplía hacia una evaluación más detallada del rendimiento del modelo mediante la construcción y visualización de la matriz de confusión. La cantidad de verdaderos positivos (1044) indica que el modelo está acertando en la predicción de la clase positiva en la mayoría de los casos, el valor bajo de falsos negativos (0) sugiere que el modelo rara vez omite la clase positiva cuando debería predecirla, el valor bajo de falsos positivos (1) indica que el modelo comete errores al predecir la clase positiva solo en unos pocos casos, finalmente, el valor de verdaderos negativos (47) muestra que el modelo predice correctamente la clase negativa en la mayoría de los casos. Este análisis proporciona información valiosa sobre la capacidad del modelo para realizar predicciones

correctas y gestionar errores, aspectos cruciales en el contexto de la gestión de riesgos financieros.

Figura 4-3: Matriz de confusión.



Las métricas específicas, como el recall y el F1 Score, se calculan para ofrecer una comprensión más profunda de la capacidad del modelo para identificar instancias positivas, subrayando la importancia de equilibrar la precisión y la exhaustividad en el contexto del riesgo crediticio. El recall obtenido de 0.9792 (aproximadamente) significa que el modelo ha identificado correctamente la gran mayoría de los casos positivos reales en el conjunto de datos, en otras palabras, el 97.92% de los casos positivos reales fueron detectados por el modelo, lo que indica un buen rendimiento en términos de identificación de los casos positivos. Por su parte, el F1 Score arroja un valor de 0.98947 (aproximadamente) y, siendo una métrica que combina la precisión y el recall, este valor proporciona una medida equilibrada entre la capacidad del modelo para clasificar correctamente las instancias positivas (verdaderos positivos) y la capacidad de evitar falsos negativos. Un F1 Score alto es especialmente valioso en problemas de clasificación desequilibrada, como el riesgo crediticio, donde las instancias de una clase (por ejemplo, incumplimiento de pago) son significativamente menos frecuentes que las de la otra clase (por ejemplo, no incumplimiento), tal y como sucede en la investigación en cuestión. Este

resultado indica que el modelo es robusto y tiene un buen desempeño en la identificación de casos de riesgo crediticio sin comprometer la precisión en la clasificación de casos no riesgosos.

Finalmente, se realiza un reentrenamiento del modelo con nuevas particiones de los datos, utilizando diferentes semillas aleatorias, donde los resultados obtenidos en las métricas de desempeño del modelo fueron similares a los mencionados anteriormente. Este enfoque busca evaluar la consistencia del rendimiento del modelo en múltiples escenarios y resaltar su capacidad para generalizar efectivamente a conjuntos de datos no vistos.

5. Validación del modelo propuesto

Como se ha corroborado en la exhaustiva revisión de la literatura en la sección correspondiente, los árboles de decisiones y la regresión logística son dos de las técnicas de machine learning más ampliamente empleadas en la predicción de riesgo crediticio y las que obtuvieron un mejor puntaje frente a los criterios evaluados en el capítulo 3, además de Random Forest. En concordancia con estas conclusiones, en el presente estudio se llevó a cabo una comparación detallada de los resultados obtenidos mediante el modelo de Random Forest con aquellos derivados de modelos basados en árboles de decisión y regresión logística.

Dicha comparación se realizó en términos de las métricas de desempeño de accuracy, recall y F1 Score, teniendo en cuenta las ventajas de estas medidas y sus definiciones mencionadas en el capítulo 3. Adicionalmente, con el fin de evaluar la capacidad del modelo para generalizar patrones con datos no observados y ver cómo se comporta con diferentes conjuntos de datos, se realizaron dos particiones de la información al cambiar la semilla con la que se desempeña el modelo. Además, la práctica anterior permite mitigar riesgos de sobreajuste, ya que al cambiar la semilla y, por lo tanto, la composición de los conjuntos de entrenamiento y prueba, se puede poner a prueba la variabilidad del modelo en diferentes situaciones.

Los hallazgos de esta comparación se presentan de manera detallada en la Tabla 3-3, junto con cada una de las particiones realizadas al conjunto de datos con el fin de eliminar sesgos en la distribución de estos, donde el color verde refleja indicadores superiores y favorables para el modelo, el color amarillo refleja indicadores aceptables para el modelo y el color rojo refleja indicadores inferiores y de baja calidad para el modelo. En el conjunto de datos original, se observa como las métricas de desempeño de Random Forest son mayores a las de los demás modelos, siendo más similares a las de los árboles de decisión, pero significativamente superiores a las de regresión logística.

Posterior a esto, se realiza la primera partición de datos con el fin de evaluar las métricas con un conjunto de datos diferente, obteniendo, de nuevo, un resultado más favorable para Random Forest, al superar las métricas de árboles de decisión y regresión logística.

Finalmente, se realiza una segunda partición de los datos, poniendo de nuevo a prueba la variabilidad del modelo, obteniendo un resultado consistente en la fortaleza de Random Forest, ya que nuevamente el desempeño de este modelo supera al de los demás en los términos de las métricas empleadas.

De manera general, entonces, es significativo destacar que, al evaluar las métricas de desempeño de cada modelo, se observa claramente que los resultados del modelo de Random Forest superan de manera destacada a los obtenidos mediante los modelos basados en árboles de decisión y regresión logística. Esta disparidad en el rendimiento sugiere que el enfoque de Random Forest demuestra ser particularmente efectivo en la tarea de predicción de riesgo crediticio, proporcionando resultados notoriamente mejores en comparación con las técnicas más tradicionales. Estos resultados respaldan la elección y la eficacia del modelo de Random Forest como una herramienta robusta y precisa en la evaluación y mitigación de riesgos crediticios.

Tabla 5-1: Comparación de métricas de desempeño entre modelos.

Partición Original			
	Random Forest	Árboles de Decisión	Regresión Logística
Accuracy	0,99908	0,99084	0,95696
Recall	0,97917	0,91667	0,10417
F1 Score	0,98947	0,89796	0,17544

Partición 1			
	Random Forest	Árboles de decisión	Regresión Logística
Accuracy	0,99634	0,98718	0,96062
Recall	0,91667	0,81250	0,18750
F1 Score	0,95652	0,84783	0,29508

Partición 2			
	Random Forest	Árboles de decisión	Regresión Logística
Accuracy	0,99725	0,99451	0,96703
Recall	0,93750	0,95833	0,31250
F1 Score	0,96774	0,93878	0,45455

6. Conclusiones y recomendaciones

6.1 Conclusiones

El presente estudio aborda de manera integral el análisis de riesgo de crédito en el sector financiero colombiano, proponiendo un enfoque innovador basado en técnicas de machine learning, específicamente el algoritmo Random Forest, cuya aplicación del modelo demostró una precisión excepcional del 99.9% en el conjunto de datos de prueba y del 99.79% en el conjunto de entrenamiento, lo cual confirma la eficacia del método propuesto en la predicción de riesgo crediticio en el contexto colombiano.

Adicionalmente, se comparó el modelo de Random Forest contra técnicas de árboles de decisión y regresión logística en términos de las métricas accuracy, recall y F1 Score, obteniendo resultados superiores para Random Forest y evidenciando la consistencia de su variabilidad ante diferentes conjuntos de datos.

6.2 Recomendaciones

Para fortalecer y ampliar el alcance de la investigación realizada, se proponen las siguientes recomendaciones:

Incorporación de Nuevas Variables: Explorar la inclusión de variables adicionales en el modelo, como indicadores macroeconómicos y políticas gubernamentales, para evaluar su impacto en la predicción de riesgo crediticio. La consideración de factores externos puede enriquecer la capacidad del modelo para anticipar cambios en el entorno financiero.

Refinamiento de Indicadores Financieros: Realizar un análisis detallado de los indicadores financieros utilizados, considerando su relevancia y peso en la predicción. La identificación de variables clave puede contribuir a la simplificación del modelo y mejorar su interpretabilidad.

Ampliación del Conjunto de Datos: La obtención de más datos históricos permitiría entrenar el modelo en condiciones económicas y financieras más diversas, mejorando su capacidad

para generalizar a situaciones no vistas previamente. Esto podría reducir la posibilidad de overfitting al ajustar el modelo a patrones más representativos.

Validación Cruzada Rigurosa: Implementar técnicas de validación cruzada más rigurosas, como la validación cruzada estratificada, para evaluar la robustez del modelo en diferentes particiones de los datos. Esto ayudaría a identificar posibles sesgos y mejorar la generalización del modelo.

Dada la dinámica del entorno financiero, se sugiere la implementación de mecanismos de monitoreo continuo para evaluar la eficacia del modelo en condiciones cambiantes. Además, la colaboración con expertos del sector financiero puede proporcionar perspectivas valiosas sobre la interpretación de los resultados y la relevancia práctica del

modelo propuesto. En conclusión, la implementación de estas recomendaciones puede contribuir al perfeccionamiento del modelo de asignación de cupos de crédito basado en machine learning, asegurando su vigencia, precisión y capacidad de adaptación a un entorno financiero dinámico.

Bibliografía

[1] BRC Standard & Poor's, "Metodología de calificación para instituciones financieras Emisores de deuda", chrome-extension://efaidnbmnnnibpcajpcgiclfndmkaj/https://www.brc.com.co/archivos/3_Tipos_Metodologias_calificacion/3_3_Metodologias_calificaciones/3_3_1_sector_financiero/3_3_1_1_Establec_credito/cal-met-005%20Met%20InsFinanDeuda%20V3.pdf.

[2] J.Y. Crespo, "CAMEL vs. discriminante, un análisis de riesgo al sistema financiero venezolano" Scielo, vol.15 no.33, Dec. 2011, http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1657-42062011000200002.

[3] W. Ossa y V. Jaramillo, "Machine Learning para la estimación del riesgo de crédito en una cartera de consumo", Universidad EAFIT, 2021, chrome-extension://efaidnbmnnnibpcajpcgiclfndmkaj/https://repository.eafit.edu.co/bitstream/handle/10784/29589/Wbeimar_OssaGiraldo_Veronica_JaramilloMarin_2021.pdf?sequence=2&isAllowed=y.

[4] J. Grau, "Machine learning y riesgo de crédito", Universidad Pontificia Comillas, 2020, chrome-extension://efaidnbmnnnibpcajpcgiclfndmkaj/https://repositorio.comillas.edu/xmlui/bitstream/handle/11531/39062/TFG-%20Grau%20Alvarez%2C%20Jaime.pdf?sequence=1&isAllowed=y.

[5] D. Borrero y O. Bedoya, "Predicción de riesgo crediticio en Colombia usando técnicas de inteligencia artificial", UIS Ingenierías, vol. 19, Universidad Industrial de Santander, jul. 2020, doi: <https://doi.org/10.18273/revuin.v19n4-2020004>.

[6] Superintendencia Financiera de Colombia, "Capítulo II: Reglas relativas a la gestión de riesgo crediticio", Circular Externa 010, 2008, chrome-extension://efaidnbmnnnibpcajpcgiclfndmkaj/https://fasecolda.com/cms/wp-content/uploads/2019/08/ce100-1995-cap-ii.pdf.

[7] Cámara de Comercio de Oviedo, "Solvencia Financiera", 2021, <https://www.mba-asturias.com/economia/que-es-solvencia-finanzas/>.

- [8] F. D. Freitas, A. F. de Souza, and A. R. de Almeida, "Prediction-based portfolio optimization model using neural networks," *Neurocomputing*, vol. 72, no. 10–12, pp. 2155–2170, Jun. 2009, doi: 10.1016/j.neucom.2008.08.019.
- [9] IBM, "¿Qué es Machine Learning?", <https://www.ibm.com/co-es/analytics/machine-learning>.
- [10] K. Tamás & V. Miklós, "EU-27 bank failure prediction with C5.0 decision trees and deep learning neural networks", *Research in International Business and Finance*, vol. 61, 2022, doi: 10.1016/j.ribaf.2022.101644.
- [11] Z. Yao & G. Changchun, "Extreme Learning Machine Enhanced Gradient Boosting for Credit Scoring", *Algorithms*, vol. 15, 2022, doi: 10.3390/a15050149.
- [12] A. Samar & A. Marco, "Developing An Intelligent System For Predicting Bankruptcy", *Journal of Theoretical and Applied Information Technology*, vol. 100, 2022. https://www.scopus.com/record/display.uri?eid=2-s2.0-85128612727&origin=resultslist&sort=plf-f&src=s&st1=credit+risk&st2=machine+learning&sid=f9bb9a8ef9189a277ef72e4bfedc0578&sot=b&sdt=b&sl=64&s=%28TITLE-ABS-KEY%28credit+risk%29+AND+TITLE-ABS-KEY%28machine+learning%29%29&relpos=10&citeCnt=0&searchTerm=&featureToggles=FEATURE_NEW_DOC_DETAILS_EXPORT:1.
- [13] A. Srinivas & R. Somula, "Loan Default Prediction Using Machine Learning Techniques", *Lecture Notes in Networks and Systems*, vol. 385, 2022, doi: 10.1007/978-981-16-8987-1_56.
- [14] T. Germanno, J. Rodrigues, R. Ricardo y K. Sergei, "Comparative study of support vector machines and random forests machine learning algorithms on credit operation", *Software - Practice and Experience*, vol. 51, 2021, doi: 10.1002/spe.2842.
- [15] L. Yun, X. Zhang y H. Yin, "An extreme learning machine based virtual sample generation method with feature engineering for credit risk assessment with data scarcity", *Expert Systems with Applications*, vol.202, 2022, doi: 10.1016/j.eswa.2022.117363.

- [16] W. Guo y Z. Zhou, "A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction", *Journal of Forecasting*, vol. 41, 2022, doi: 10.1002/for.2856.
- [17] R. García y M. Moreno, "The generalized Vasicek credit risk model: A Machine Learning approach", *Finance Research Letters*, vol. 47, 2022, doi: 10.1016/j.frl.2021.102669.
- [18] T. Wang, R. Liu y G. Qi, "Multi-classification assessment of bank personal credit risk based on multi-source information fusion", *Expert Systems with Applications*, vol. 191, 2022, doi: 10.1016/j.eswa.2021.116236.
- [19] E. Dumitrescu, S. Hué, C. Hurlin y S. Tokpavi, "Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects", *European Journal of Operational Research*, vol. 297, 2022, doi: 10.1016/j.ejor.2021.06.053.
- [20] S. Dogan, Y. Buyukkor y M. Atan, "A COMPARATIVE STUDY OF CORPORATE CREDIT RATING PREDICTION WITH MACHINE LEARNING", *Operations Research and Decisions*, vol. 32, 2022, doi: 10.37190/ord220102.
- [21] Wang, C., Liu, Q., Li, S., "A two-stage credit risk scoring method with stacked-generalisation ensemble learning in peer-to-peer lending", *International Journal of Embedded Systems*, vol. 15, 2022, doi: 10.1504/IJES.2022.123312.
- [22] Jadwal, P.K., Pathak, S., Jain, S., "Analysis of clustering algorithms for credit risk evaluation using multiple correspondence analysis", *Microsystem Technologies*, 2022, doi: 10.1007/s00542-022-05310-y.
- [23] Tian, J., Li, L., "Digital Universal Financial Credit Risk Analysis Using Particle Swarm Optimization Algorithm with Structure Decision Tree Learning-Based Evaluation Model", *Wireless Communications and Mobile Computing*, vol. 2022, 2022, doi: 10.1155/2022/4060256.
- [24] Coenen, L., Verbeke, W., Guns, T., "Machine learning methods for short-term probability of default: A comparison of classification, regression and ranking methods", *Journal of the Operational Research Society*, vol. 73, 2022, doi: 10.1080/01605682.2020.1865847.

- [25] Frydman, H., Matuszyk, A., “Random survival forest for competing credit risks”, *Journal of the Operational Research Society*, vol. 73, 2022, doi: 10.1080/01605682.2020.1759385.
- [26] Hu, L., Chen, J., Vaughan, J., (...), Sudjianto, A., Nair, V.N., “Supervised Machine Learning Techniques: An Overview with Applications to Banking”, *International Statistical Review*, vol. 89, 2021, doi: 10.1111/insr.12448.
- [27] Gao, G., Wang, H., Gao, P., “Establishing a credit risk evaluation system for smes using the soft voting fusion model”, *Risks*, vol. 9, 2021, doi: 10.3390/risks9110202.
- [28] Wu, C.-F., Huang, S.-C., Chiou, C.-C., Wang, Y.-M., “A predictive intelligence system of credit scoring based on deep multiple kernel learning”, *Applied Soft Computing*, vol. 111, 2021, doi: 10.1016/j.asoc.2021.107668.
- [29] Acheampong, A., Elshandidy, T., “Does soft information determine credit risk? Text-based evidence from European Banks”, *Journal of International Financial Markets, Institutions and Money*, vol. 75, 2021, doi: 10.1016/j.intfin.2021.101303.
- [30] Kokate, S., Chetty, M.S.R., “Credit risk assessment of loan defaulters in commercial banks using voting classifier ensemble learner machine learning model”, *International Journal of Safety and Security Engineering*, vol. 11, 2021, doi: 10.18280/IJSSE.110508.
- [31] Breeden, J.L., “A survey of machine learning in credit risk”, *Journal of Credit Risk*, vol. 17, 2021, doi: 10.21314/JCR.2021.008.
- [32] Liu, R., Yang, X., Dong, X., Sun, B., “Credit risk assessment of banks' loan enterprise customer based on state-constraint”, *Computing and Informatics*, vol. 40, 2021, doi: 10.31577/cai_2021_1_145.
- [33] Lappas, P.Z., Yannacopoulos, A.N., “A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment”, *Applied Soft Computing*, vol. 107, 2021, doi: 10.1016/j.asoc.2021.107391.

- [34] Balakrishnan, C., Thiagarajan, M., “Credit risk modelling for Indian debt securities using machine learning”, *Buletin Ekonomi Moneter dan Perbankan*, vol. 24, 2021, doi: 10.21098/BEMP.V24I0.1401.
- [35] Ampountolas, A., Nde, T.N., Date, P., Constantinescu, C., “A machine learning approach for micro-credit scoring”, *Risks*, vol. 9, 2021, doi: 10.3390/risks9030050.
- [36] Moscato, V., Picariello, A., Sperlí, G., “A benchmark of machine learning approaches for credit score prediction”, *Expert Systems with Applications*, vol. 165, 2021, doi: 10.1016/j.eswa.2020.113986.
- [37] Shen, F., Zhao, X., Kou, G., & Alsaadi, F. E., “A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique”, *Applied Soft Computing*, vol. 98, 2021, 106852.
- [38] Moscatelli, M., Parlapiano, F., Narizzano, S., & Viggiano, G., “Corporate default forecasting with machine learning”, *Expert Systems with Applications*, vol. 161, 2020, 113567.
- [39] Li, J. P., Mirza, N., Rahat, B., & Xiong, D., “Machine learning and credit ratings prediction in the age of fourth industrial revolution”, *Technological Forecasting and Social Change*, vol. 161, 2020, 120309.
- [40] Tripathi, D., Edla, D. R., Kuppili, V., & Bablani, A., “Evolutionary extreme learning machine with novel activation function for credit scoring”, *Engineering Applications of Artificial Intelligence*, vol. 96, 2020, 103980.
- [41] Luo, C., “A comprehensive decision support approach for credit scoring”, *Industrial Management & Data Systems*, 2019, Vol. 120 No. 2, pp. 280-290.
<https://doi.org/10.1108/IMDS-03-2019-0182>.
- [42] Arora, N., & Kaur, P. D., “A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment”, *Applied Soft Computing*, vol. 86, 2020, 105936.
- [43] Qasem, M. H., & Nemer, L., “Extreme Learning Machine for Credit Risk Analysis”, *Journal of Intelligent Systems*, vol. 29(1), 2020, 640-652.

- [44] Superintendencia Financiera de Colombia, “Actualidad del Sistema Financiero”, 2022. <https://www.superfinanciera.gov.co/inicio/informes-y-cifras/informes/informe-actualidad-del-sistema-financiero-colombiano/resultados-del-sistema-financiero-colombiano-agosto-de--10112528>.
- [45] Escobar-Pérez, J. y Cuervo-Martínez, A., “Validez de contenido y juicio de expertos: una aproximación a su utilización”, Avances en Medición, vol. 6, pp. 27-36, 2008, http://www.humanas.unal.edu.co/psicometria/files/7113/8574/5708/Articulo3_Juicio_de_expertos_27-36.pdf.
- [46] Cabero Almenara, J. y Llorente Cejudo, M. C., “La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información (TIC)”. Eduweb. Revista de Tecnología de Información y Comunicación en Educación, vol. 7 (2) pp.11-22, 2013, <http://tecnologiaedu.us.es/tecnoedu/images/stories/jca107.pdf>
- [47] Martínez, J. (2020). Precision, Recall, F1, Accuracy en clasificación. IArtificial.net. <https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/>
- [48] Sanahuja, P. (2021). Entendiendo la curva ROC y el AUC: Dos medidas del rendimiento de un clasificador binario que van de la mano. Pol Martí Sanahuja. <https://polmartisanahuja.com/entendiendo-la-curva-roc-y-el-auc-dos-medidas-del-rendimiento-de-un-clasificador-binario-que-van-de-la-mano/>.
- [49] Amat, J. (2016). Regresión logística simple y múltiple. Cienciadedatos.net. https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple.
- [50] Cuenca, D. & León, D. (2022). Support Vector Machine. Amazon AWS. https://rstudio-pubs-static.s3.amazonaws.com/570352_e34015b16f1a47e883e04c6195d4711f.html.
- [51] Amat, J. (2017). Máquinas de Vector Soporte (Support Vector Machines, SVMs). Cienciadedatos.net. https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines.

- [52] Amazon AWS (2023). ¿Qué es una red neuronal?. Amazon AWS.
<https://aws.amazon.com/es/what-is/neural-network/>.
- [53] Matich, D. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. Universidad Tecnológica Nacional. chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monografias/matich-redesneuronales.pdf.
- [54] Aprendemachinelearning.com (2018). Clasificar con K-Nearest-Neighbor ejemplo en Python. Aprendemachinelearning.com.
<https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/>.
- [55] Ferrero, R. (2020). Qué son los árboles de decisión y para qué sirven. Máximaformacion.es. <https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/>.
- [56] Martínez, J. (2020). Random Forest (Bosque Aleatorio): combinando árboles. Iartificial.net. <https://www.iartificial.net/random-forest-bosque-aleatorio/>.
- [57] IBM (2023). ¿Qué es un bosque aleatorio? IBM. <https://www.ibm.com/mx-es/topics/random-forest>.
- [58] Unidad de Regulación Financiera (2023). Decreto Único 2555 de 2010. chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www.urf.gov.co/webcenter/ShowProperty?nodeId=/ConexionContent/WCC_CLUSTER-107284.
- [59] Asobancaria (2019). Metodología de selección de las entidades financieras que participarán en el esquema del indicador bancario de referencia. chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/<https://www.asobancaria.com/wp-content/uploads/2019-08-metodologia-de-seleccion-entidades-actualizada-VF.pdf>.
- [60] ARL SURA (2023). Tomado del artículo publicado en la revista Gerencia de Riesgos y Seguros de la Fundación MAPFRE ESTUDIOS. <https://www.arlsura.com/index.php/66-centro-de-documentacion-anterior/prevencion-de-riesgos-/280--sp-28739>.
- [61] Amazon Web Services (2023). ¿Qué es el sobreajuste?.
<https://aws.amazon.com/es/what-is/overfitting/>.