REyF Revista de Economía y Finanzas

www.reveyf.es

2024 – Vol. 2 – n.º 5 - Artículo 4

Construcción de un score crediticio mediante el análisis de datos alternativos y métodos de inteligencia artificial para impulsar la inclusión financiera

Paulina Milenka González Manrique[a] y María Coveñas de León[b]

Seleccionado en la segunda convocatoria de trabajos de fin de máster de la Revista de Economía y Finanzas

Tutor: Iván Blanco Sánchez

[a] *Graduada en Ingeniería Financiera - ITESO Universidad y Máster en Instituciones y Mercados Financieros - CUNEF.*

[b] *Máster en Instituciones y Mercados Financieros - CUNEF.*

JEL CODES:
C38; G21; O16

KEYWORDS:
Financial inclusion;
Artificial intelligence;
Alternative data;
XGBoost;
Credit risk assessment

Abstract: This study develops a credit scoring model using alternative data and artificial intelligence to promote financial inclusion. Techniques such as XGBoost, logistic regression, and Random Forest were implemented, demonstrating the effectiveness of unconventional data in credit risk assessment and its impact on financial inclusion.

CÓDIGOS JEL:
C38; G21; O16

PALABRAS CLAVE:
Inclusión financiera;
Inteligencia artificial;
Datos alternativos;
XGBoost; Evaluación del riesgo crediticio

Resumen: Este estudio desarrolla un modelo de score crediticio utilizando datos alternativos e inteligencia artificial, con el objetivo de promover la inclusión financiera. Se implementaron técnicas como XGBoost, regresión logística y Random Forest, demostrando la efectividad de los datos no convencionales en la evaluación del riesgo crediticio e impactando la inclusión financiera.

E-mail: paulina.gonzalez@cunef.ea
https://doi.org/10.32826/reyf.v2i5.375

1. Introducción

En la actualidad, el acceso al crédito sigue siendo un desafío crítico para amplios segmentos de la población en diferentes países, especialmente para aquellos que son emergentes y luchan con problemas de desigualdad y pobreza. Frente a este panorama, surge la necesidad de democratizar el acceso al crédito, utilizando enfoques innovadores que aprovechen la tecnología y datos no convencionales. Este Trabajo de Fin de Máster (TFM) se titula "Construcción de un Score Crediticio mediante el Análisis de Datos Alternativos y métodos de Inteligencia Artificial para impulsar la Inclusión Financiera", y tiene como objetivo fundamental la creación de un modelo de riesgo crediticio que aproveche el potencial de los datos alternativos y la inteligencia artificial para transformar la evaluación del riesgo crediticio y brindar mayores oportunidades para un espectro más amplio de la población.

La capacidad de predecir el comportamiento crediticio no debería depender únicamente del historial financiero de una persona; en cambio, puede enriquecerse o partir del uso de datos alternativos. Estos datos pueden incluir información de encuestas detalladas, comportamientos en redes sociales, hábitos de consumo, patrones de pago de servicios, entre otros. Al combinar estos datos con técnicas avanzadas de inteligencia artificial, es posible crear modelos predictivos más robustos, inclusivos y precisos en la identificación de riesgos.

Este TFM investigará cómo los datos de fuentes alternativas pueden ser analizados mediante inteligencia artificial para desarrollar un score crediticio que sea tanto predictivo como inclusivo. Con este enfoque, el estudio busca enfrentar directamente los desafíos de la inclusión financiera y el acceso al crédito, ofreciendo soluciones que permitan a los otorgantes de crédito comprender mejor a sus clientes y expandir sus servicios a poblaciones tradicionalmente no bancarizadas, mejorando así su calidad de vida.

2. Contexto actual del crédito en Latinoamérica y España

2.1. Panorama mundial

Según el Global Findex Database 2023 del Banco Mundial, aproximadamente 1.4 mil millones de adultos en todo el mundo aún no utilizan servicios financieros formales. A pesar de las mejoras recientes en la inclusión financiera, todavía existen desafíos significativos, especialmente para segmentos de la población con bajos ingresos, menor nivel educativo, residentes en zonas rurales, mujeres y adultos mayores (Banco Mundial 2023).

La falta de acceso a una cuenta bancaria impide que estos grupos puedan acceder a mecanismos de financiación convencionales, ya que a menudo carecen de garantías y de un historial crediticio necesario para demostrar su capacidad de pago. Como resultado, se les niegan oportunidades de crédito que podrían ayudarles a cubrir necesidades financieras, reducir la vulnerabilidad económica y mejorar su capacidad para enfrentar emergencias financieras o cubrir necesidades de liquidez durante un ciclo de ingresos (Deloitte 2023).

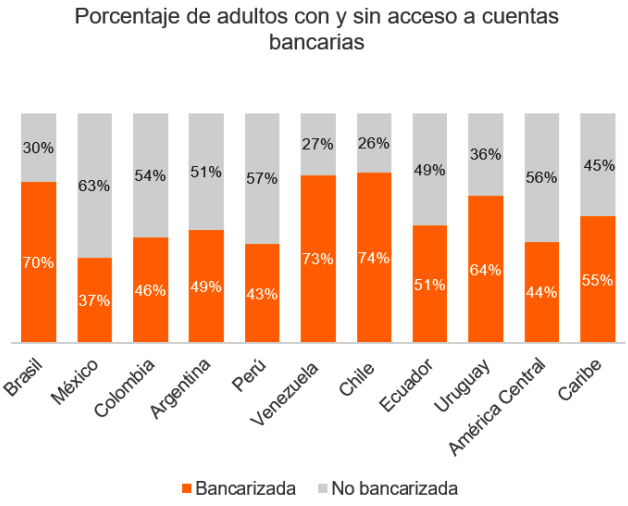
Además, estas personas suelen tener que ajustar sus patrones de consumo para hacer frente a necesidades inesperadas, lo que implica prescindir de productos y servicios básicos necesarios para mantener una calidad de vida digna. Operan con presupuestos muy limitados, lo que les dificulta planificar objetivos a largo plazo. Esta situación evidencia la brecha que los servicios financieros formales aún no han logrado cubrir para ciertos segmentos de la población (Banco Mundial 2023).

La falta de acceso al crédito a menudo lleva a las personas a recurrir a mecanismos de financiación informales, como casas de empeño o prestamistas informales, que imponen tasas de interés excesivamente altas. Esto dificulta la recuperación financiera y perjudica la estabilidad financiera a largo plazo de las familias (MasterCard 2023).

2.2. Contexto en Latinoamérica

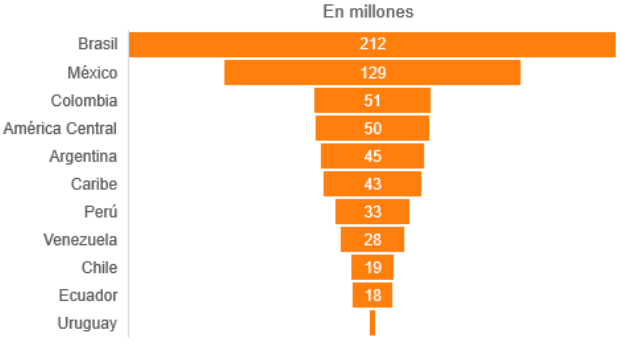
La situación en América Latina es más compleja debido a la diversidad de realidades económicas y sociales en la región. Un porcentaje significativo de la población en países como México, Colombia, y Perú, entre otros, sigue siendo no bancarizada. Las principales razones incluyen altos niveles de pobreza, disparidades regionales en el acceso a servicios, y a menudo, una mayor inestabilidad económica.

Gráfico 1: Distribución de la población adulta en 2020



Fuente: Informe de BPC Latin America Digital Banking

Gráfico 2: Tabla de población en Sudamérica en 2020



Fuente: Informe de BPC Latin America Digital Banking

Una encuesta que realizó MasterCard en diferentes países de Latinoamérica en el 2023 muestra que el 58 % de las personas tienen acceso a una tarjeta de crédito, mientras que el 38 % tienen acceso a un préstamo o línea de crédito tal como se muestra en la imagen 1. Sin embargo, en ciertos casos, los índices de acceso al crédito pueden ser mucho más bajos. Por ejemplo, en El Salvador, México y Perú, los porcentajes de penetración del crédito pueden ser inferiores al 50%.

Imagen 1:

	Argentina	Brasil	Colombia	El Salvador	Guatemala	México	Perú
Tarjeta de crédito	74 %	75 %	55 %	50 %	64 %	39 %	47 %
Préstamo o línea de crédito	19 %	28 %	30 %	47 %	43 %	18 %	29 %

Fuente: MasterCard

2.3. Contexto en España

En España, aunque el sistema bancario es robusto y ampliamente desarrollado, existe un segmento mínimo de la población que sigue siendo no bancarizado. Este grupo incluye a menudo a inmigrantes, personas mayores, y residentes en comunidades rurales, quienes no tienen acceso fácil a servicios financieros tradicionales como cuentas bancarias, préstamos y créditos. Las razones de esta exclusión son diversas e incluyen la falta de documentación necesaria, bajos niveles de ingresos, falta de historial crediticio, y en algunos casos, una profunda desconfianza hacia las instituciones financieras.

El Informe del Banco de España sobre la Exclusión Financiera (2023) destaca que, a pesar de los avances en la digitalización y la expansión de los servicios financieros, estos desafíos de exclusión persisten, especialmente para estos grupos vulnerables. Esta situación subraya la necesidad de soluciones innovadoras que consideren a todos los segmentos de la población, incluyendo aquellos tradicionalmente excluidos del sistema financiero formal (Banco de España 2023).

Estos desafíos refuerzan la importancia de desarrollar modelos de inteligencia artificial que utilicen datos alternativos para evaluar el riesgo crediticio y promover una mayor inclusión financiera. Los datos alternativos, pueden proporcionar una imagen más completa del comportamiento financiero de los individuos. Al integrar estos datos con técnicas avanzadas de inteligencia artificial, se pueden crear modelos más precisos y justos que permitan a los otorgantes de crédito expandir sus servicios a personas que tradicionalmente han sido excluidas del sistema financiero formal.

2.4. Inclusión financiera

Por todos estos motivos, es indispensable que a la población se le brinden oportunidades y soluciones que mejoren su calidad de vida, lo cual es posible gracias a la inclusión financiera. El Banco Mundial define la inclusión financiera como un facilitador clave para reducir la pobreza y fomentar la prosperidad compartida. La inclusión financiera se refiere al acceso que tienen las personas y las empresas a productos y servicios financieros útiles y asequibles —como transacciones, pagos, ahorro, crédito y seguros— que se ofrecen de manera responsable y sostenible (Banco Mundial 2023).

Además, la inclusión financiera es una prioridad global, ya que contribuye al logro de varios de los Objetivos de Desarrollo Sostenible de la Agenda 2030 de la ONU. El Grupo de los 20 (G20) también ha reafirmado su compromiso de promover la inclusión financiera en todo el mundo, subrayando su importancia para mejorar el acceso a servicios financieros y reducir las desigualdades económicas y sociales (Banco Mundial 2023).

3. Evolución de scores crediticios: impacto de datos alternativos e inteligencia artificial en el riesgo crediticio

La capacidad de las instituciones financieras para ofrecer productos crediticios de calidad está esencialmente relacionada con su capacidad para evaluar el riesgo de forma adecuada, lo cual depende de su acceso a la información crediticia. Margarita Henao, directora ejecutiva de Daviplata en Colombia, explica que “la falta de información es una de las dos barreras más importantes para la inclusión financiera, principalmente para el crédito”. Henao enfatiza la necesidad de contar con formas innovadoras de calcular la calificación crediticia que se basen en datos de comportamiento, para así ayudar a los clientes a mejorarla (MasterCard 2023).

Este tipo de análisis está en línea con las tendencias globales en la transformación digital de los servicios financieros, donde la inteligencia artificial y el análisis de datos están jugando un papel crucial para mejorar la evaluación del riesgo y la inclusión financiera. Por ejemplo, informes de instituciones como McKinsey y el Banco Interamericano de Desarrollo destacan la importancia de integrar tecnologías avanzadas y nuevas metodologías para enfrentar los

desafíos actuales en el sector financiero (McKinsey & Company 2023 y IDB Publications 2023).

En la era digital actual, esto representa una oportunidad para impactar en la inclusión financiera, ya que con el desarrollo de la Inteligencia Artificial y los novedosos modelos de Machine Learning, se puede predecir el riesgo crediticio de una persona, utilizando datos alternativos que no necesariamente provengan de fuentes tradicionales como un historial crediticio. De hecho, los datos alternativos representan una oportunidad para proporcionar a las instituciones financieras una herramienta efectiva para evaluar a personas que no tienen historial crediticio tradicional. Al incluir datos alternativos en los modelos de evaluación, los otorgantes de crédito pueden obtener una visión más completa del comportamiento de un individuo, lo que mejora la precisión en la evaluación del riesgo crediticio y reducir la probabilidad de impago.

Los datos alternativos abarcan información como transacciones digitales, actividad en redes sociales, pagos de servicios, uso de telefonía móvil y datos de geolocalización, que reflejan comportamientos financieros y socioeconómicos. Este mercado está creciendo rápidamente: se espera que alcance un valor de $159 mil millones para 2030, con un aumento del 36% en su uso en servicios financieros en los últimos dos años. Empresas y startups están adoptando inteligencia artificial para desarrollar modelos crediticios innovadores, aprovechando la digitalización y estas fuentes no convencionales para evaluar riesgos de forma más inclusiva y precisa.

4. Recopilación de datos alternativos relevantes de la población de interés

En el capítulo anterior, se exploró cómo los datos alternativos están revolucionando la construcción de modelos de riesgo crediticio más inclusivos, especialmente en contextos donde las fuentes de datos tradicionales, como los historiales crediticios de los Burós de Crédito o las instituciones financieras, no están disponibles. Este desafío es particularmente relevante para las personas no bancarizadas, cuyo comportamiento financiero no queda registrado en los sistemas tradicionales. Con el fin de superar estas limitaciones y desarrollar un modelo predictivo que refleje de manera más precisa y equitativa el riesgo crediticio, se diseñó y llevó a cabo un proceso de recopilación de datos exhaustivo utilizando métodos innovadores.

En primer lugar, se intentó establecer contacto con diversas Fintechs con el objetivo de obtener datos anonimizados del comportamiento crediticio de sus clientes para fines académicos y de investigación. Sin embargo, debido a diversas restricciones y la falta de respuesta favorable, no se logró acceder a estos datos.

Posteriormente, se procedió a la construcción de una encuesta tras una revisión exhaustiva de la literatura existente y discusiones con expertos en el campo de la evaluación crediticia y la inclusión financiera. Las preguntas fueron cuidadosamente seleccionadas para captar información relevante y útil para la evaluación del riesgo crediticio. Las preguntas de la encuesta abarcaron varias categorías importantes:

- Datos demográficos: Incluyeron la fecha de nacimiento, sexo, estado civil, ciudad de residencia y tipo de residencia. Estas preguntas se formularon para comprender mejor el contexto social y económico de los encuestados, proporcionando una base para analizar cómo diferentes factores demográficos podrían influir en la solvencia crediticia de una persona.

- Situación familiar y educativa: Se preguntó a los encuestados sobre la presencia de hijos a su cargo y su nivel de estudios alcanzado. Esta información es crucial para entender las responsabilidades financieras y el potencial de ingresos futuros de los individuos, factores que impactan significativamente en su perfil de riesgo.

- Situación laboral y profesional: Preguntas sobre la situación laboral actual, tipo de contrato, profesión y tiempo en el cargo actual ayudaron a evaluar la estabilidad laboral y la capacidad de generar ingresos de los encuestados. Estos factores son esenciales para predecir la capacidad de los individuos para cumplir con sus obligaciones crediticias.

- Habilidades cognitivas y financieras: La encuesta incluyó preguntas de razonamiento lógico y matemático, como secuencias numéricas y problemas hipotéticos, diseñadas para evaluar habilidades cognitivas que pueden correlacionarse con la gestión financiera responsable. Este enfoque permite una comprensión más profunda de cómo las habilidades cognitivas pueden influir en el comportamiento financiero.

- Comportamiento financiero: Se preguntó sobre la posesión de cuentas bancarias, contratación de préstamos, ingresos netos mensuales, ahorros disponibles, historial de pagos de facturas, y posesión de bienes como coche o smartphone. Estas preguntas proporcionaron una visión directa de la salud financiera de los encuestados, su acceso a servicios financieros y su comportamiento financiero general, aspectos críticos para la construcción de un perfil de riesgo crediticio.

- Historial de cumplimiento y responsabilidad: También se formularon preguntas sobre posibles retrasos en el pago de facturas de servicios en los últimos 12 meses y cualquier problema legal en los últimos cinco años. Este tipo de información es vital para evaluar la responsabilidad financiera y la estabilidad de los encuestados, proporcionando un indicador adicional del riesgo crediticio.

La encuesta fue distribuida ampliamente a través de plataformas digitales y redes sociales para alcanzar una amplia variedad de individuos con diferentes perfiles socioeconómicos. Esta estrategia de difusión se implementó para minimizar los sesgos de selección y asegurar una muestra representativa de la población objetivo, compuesta por personas con diversos niveles de ingresos, ocupaciones, niveles educativos, y acceso a servicios financieros. El proceso de difusión aleatoria permitió la inclusión de participantes de diferentes contextos geográficos y económicos, asegurando que los datos recopilados fueran diversos y reflejaran una amplia gama de comportamientos financieros.

Una vez recopilados los datos, estos fueron almacenados en una base de datos segura y se sometieron a un riguroso proceso de limpieza para garantizar su integridad y precisión. Esto incluyó la eliminación de entradas duplicadas o inconsistentes, asegurando que solo se utilizaran datos de alta calidad para el análisis.

5. Selección de variables clave que puedan predecir el comportamiento crediticio en estas poblaciones y uso de algoritmos de inteligencia artificial para desarrollar un modelo predictivo

Tras la recopilación de datos alternativos pertinentes, el siguiente paso en este estudio es el desarrollo de modelos predictivos capaces de evaluar el comportamiento crediticio de los individuos de manera precisa. Para alcanzar este objetivo, se implementarán y evaluarán cuatro técnicas de Machine Learning ampliamente reconocidas: Regresión Logística, Random Forest, XGBoost, y Support Vector Machines (SVM). Cada uno de estos algoritmos se selecciona por su capacidad para manejar conjuntos de datos complejos y por su eficacia en la identificación de patrones no evidentes que puedan influir en el riesgo crediticio.

El proceso de selección del modelo más adecuado se basará en la comparación del rendimiento de estos algoritmos utilizando técnicas de validación cruzada. Esta metodología permite evaluar la robustez de los modelos y reducir el riesgo de sobreajuste, garantizando que los modelos no solo funcionen bien con los datos de entrenamiento, sino que también mantengan su precisión en datos no vistos anteriormente. Las métricas de evaluación clave que se utilizarán incluyen precisión, recall, F1-score, y el área bajo la curva ROC (Receiver Operating Characteristic), que ofrecen una visión completa de la capacidad predictiva y la sensibilidad de los modelos a diferentes tipos de errores.

En el ámbito de la evaluación del riesgo crediticio, los modelos de Machine Learning han demostrado ser herramientas potentes para predecir el comportamiento financiero de individuos y empresas. Los algoritmos seleccionados para este estudio —Regresión Logística con Softmax, Random Forest, XGBoost y SVM— son especialmente valorados por su capacidad para manejar grandes volúmenes de datos y capturar relaciones complejas entre variables. Cada uno de estos modelos ofrece ventajas únicas en términos de flexibilidad, interpretabilidad y capacidad de manejo de datos desequilibrados, características esenciales en la predicción del riesgo crediticio.

5.1. Regresión logística con Softmax

La regresión logística es un modelo estadístico utilizado para la clasificación binaria, es decir, para predecir una de dos posibles resultados. La regresión logística multinomial, también conocida como regresión logística con Softmax, extiende este concepto para manejar múltiples clases. El modelo utiliza la función de activación Softmax para asignar probabilidades a cada clase, asegurando que la suma de las probabilidades sea igual a uno.

Estudios previos han demostrado la eficacia de la regresión logística en la predicción de riesgos financieros (Hosmer et al. 2013 y Lee & Chen 2005) subrayan su importancia en la modelización de credit default. Además, (Mays 2004) destaca cómo la regresión logística puede ser utilizada para mejorar los modelos de scoring crediticio, proporcionando una base sólida para decisiones crediticias.

La regresión logística con Softmax es una herramienta poderosa y versátil para la clasificación multiclase, especialmente útil en la evaluación del riesgo crediticio. Su capacidad para manejar múltiples clases y proporcionar probabilidades detalladas permite a los analistas financieros tomar decisiones informadas y gestionar mejor el riesgo. Con su fundamento en técnicas estadísticas robustas y su aplicabilidad en numerosos campos, sigue siendo una técnica valiosa en el análisis de datos y la ciencia de la decisión.

5.2. Random Forest

Random Forest es un algoritmo de ensemble que consiste en la combinación de múltiples árboles de decisión. Cada árbol en el bosque se construye utilizando una muestra aleatoria del conjunto de datos y selecciona de manera aleatoria un subconjunto de características para realizar las divisiones en cada nodo. La predicción final del Random Forest se obtiene promediando las predicciones de todos los árboles (en el caso de regresión) o por votación mayoritaria (en el caso de clasificación).

El proceso de construcción de un árbol individual en un Random Forest se puede resumir en los siguientes pasos:

1. Seleccionar aleatoriamente una muestra del conjunto de datos con reemplazo (muestra de bootstrap).

2. En cada nodo de decisión, seleccionar aleatoriamente un subconjunto de características y elegir la mejor característica para dividir el nodo.

3. Repetir el proceso hasta que se cumpla un criterio de parada (por ejemplo, una profundidad máxima del árbol o un número mínimo de observaciones en un nodo).

Esta combinación de árboles de decisión crea un "bosque" robusto que es capaz de manejar datos de alta dimensionalidad y capturar interacciones no lineales entre las variables. Random Forest es conocido por su capacidad para reducir el sobreajuste y mejorar la precisión general del modelo (Breiman 2001).

Random Forest se ha utilizado ampliamente en la evaluación del riesgo crediticio debido a su capacidad para manejar datos de alta dimensionalidad y capturar interacciones no lineales entre las variables. Este modelo es particularmente útil para predecir la probabilidad de incumplimiento crediticio, ya que puede integrar una gran cantidad de factores y reducir el riesgo de sobreajuste mediante el uso de múltiples árboles de decisión. Un estudio realizado por (Malley et al. 2012) destaca cómo Random Forest puede ser utilizado para mejorar la precisión de las predicciones en el ámbito financiero.

5.3. XGBoost

XGBoost (eXtreme Gradient Boosting) es un potente algoritmo de boosting basado en árboles de decisión que ha demostrado ser altamente eficiente y preciso, especialmente en grandes conjuntos de datos y en competiciones de ciencia de datos. Introducido por (Tianqi Chen y Carlos Guestrin en 2016) XGBoost se ha convertido en uno de los algoritmos más populares debido a su capacidad para manejar datos

complejos y grandes volúmenes de información de manera eficiente.

El principio fundamental de XGBoost es combinar varios modelos débiles (árboles de decisión) para formar un modelo fuerte. Cada árbol se construye secuencialmente, donde cada árbol intenta corregir los errores del árbol anterior. La técnica de regularización que emplea XGBoost previene el sobreajuste, mejorando así la precisión y la capacidad de generalización del modelo.

5.4. Support Vector Machines (SVM)

Support Vector Machines (SVM) es un potente algoritmo de clasificación que busca encontrar el hiperplano que mejor separa las clases en el espacio de características. Introducido por (Cortes y Vapnik 1995), SVM se ha convertido en una técnica popular debido a su capacidad para manejar datos de alta dimensionalidad y su eficacia en problemas de clasificación complejos.

En su forma más simple, SVM encuentra el hiperplano que maximiza la distancia (margen) entre las muestras de diferentes clases. Este margen máximo asegura que las muestras de cada clase se encuentren lo más alejadas posible del hiperplano de separación, lo que mejora la capacidad del modelo para generalizar a datos no vistos.

SVM es especialmente útil para problemas de clasificación en los que las clases son altamente separables en el espacio de características. En la evaluación del riesgo crediticio, SVM puede ser utilizado para clasificar a los solicitantes de crédito en categorías como "buen crédito" y "mal crédito" basándose en sus características financieras y demográficas. Su capacidad para manejar datos no lineales y de alta dimensionalidad lo hace adecuado para capturar las complejidades inherentes en los datos de riesgo crediticio (Baesens et al. 2003).

6. Selección de variables predictoras y objetivo

En el desarrollo de un modelo de riesgo crediticio utilizando datos alternativos, es fundamental preparar y transformar adecuadamente los datos de entrada para garantizar que el modelo sea robusto y preciso. A partir de 200 encuestas recabadas después de la difusión a cientos de personas, se extrajeron múltiples variables que proporcionan información sobre los encuestados. A continuación, se explican detalladamente los pasos realizados para seleccionar, transformar y preparar las variables predictoras y la variable objetivo para el entrenamiento de los cuatro modelos de machine learning.

Dado que muchas preguntas de la encuesta eran de tipo categórico (por ejemplo, "Sexo", "Estado civil", "Nivel de estudios"), fue necesario convertir estas variables en un formato numérico que pudiera ser utilizado por los algoritmos de Machine Learning. Este proceso, conocido como codificación de variables categóricas, se realizó mediante la técnica de One-Hot Encoding. Esta técnica crea variables binarias (0 o 1) para cada categoría única dentro de una variable, permitiendo que los modelos interpreten las diferentes categorías como independientes entre sí.

La transformación de variables categóricas es crucial porque los modelos de Machine Learning requieren entradas numéricas. Además, esta transformación permite que los algoritmos capten las relaciones no lineales entre las variables categóricas y la variable objetivo, mejorando la capacidad predictiva del modelo.

Para construir el modelo de riesgo crediticio, se utilizó como variables predictoras (X) todas las preguntas de la encuesta que proporcionan información demográfica, socioeconómica, laboral, y comportamiento financiero de los encuestados. Estas variables se consideran alternativas en comparación con los datos tradicionales utilizados en la evaluación crediticia, como los historiales crediticios. La variable objetivo (y) seleccionada para este estudio fue la respuesta a la pregunta "¿Cuáles son sus ingresos netos mensuales?", una variable crítica que refleja la capacidad financiera de los individuos para cumplir con sus obligaciones crediticias.

Las variables predictoras (X) como se mencionaba antes, incluyen tanto datos demográficos como características financieras y de comportamiento:

Datos Demográficos:

Fecha de Nacimiento: Convertida a edad para análisis. La edad es una variable numérica continua que puede correlacionarse con la estabilidad financiera y el comportamiento crediticio.

Sexo: Incluye tres categorías: "Hombre", "Mujer" y "Prefiero no decirlo". La inclusión de la opción "Prefiero no decirlo" requiere la creación de tres variables dummy: Sexo_Hombre, Sexo_Mujer, y Sexo_Prefiero_no_decirlo, para representar cada opción de manera binaria.

Estado Civil: Contiene varias categorías como "Casado", "Soltero", "Divorciado", "Unión libre". Cada estado civil se convierte en una columna binaria para capturar las diferentes situaciones personales que pueden influir en las decisiones financieras.

Características Socioeconómicas y Laborales:

Su Residencia es: Incluye opciones como "Casa propia", "Con padres", "Piso propio", "Piso de alquiler", "Habitación", "Otro". Estas categorías reflejan la estabilidad de la vivienda y las posibles responsabilidades financieras, convirtiéndose en variables dummy para el análisis.

¿Tiene hijos a su cargo?: Respuestas numéricas que indican el número de dependientes. Esto es un indicador de las responsabilidades financieras de una persona.

Nivel de Estudios: Las opciones como "ESO", "Bachillerato", "Formación Profesional", "Grado", "Posgrado" se transforman en variables dummy para capturar el impacto del nivel educativo en el comportamiento crediticio.

Situación Laboral: Respuestas como "Autónomo", "Desempleado", "Funcionario", etc., se codifican como variables dummy para reflejar las diferentes situaciones laborales que pueden afectar la estabilidad financiera.

Profesión: Dado que esta es una respuesta abierta, las profesiones se categorizan según sectores (por ejemplo, sector público, sector privado, freelance) y se convierten en variables dummy.

Comportamientos Financieros y Activos:

¿Tiene alguna cuenta bancaria?: Esta variable binaria ("Sí" o "No") se convierte directamente en una variable dummy.

¿Tiene contratado algún préstamo?: Las respuestas incluyen opciones como "Sí", "No", "Hipotecario", "Automotriz", etc., y se transforman en variables dummy para capturar los diferentes tipos de préstamos.

¿Dispone de coche? y ¿Tiene un smartphone?: Estas son variables binarias simples ("Sí" o "No") y se utilizan directamente como variables dummy.

¿Ha tenido algún retraso en los pagos de sus facturas de servicio en los últimos 12 meses? y ¿Ha tenido algún problema con la ley en los últimos 5 años?: Estas preguntas se utilizan para crear variables binarias que pueden indicar un mayor riesgo crediticio.

Preguntas Cognitivas y Numéricas:

Preguntas como "¿Qué número debería ser el siguiente? 9 16 23 30 37 44 51 ..." y "¿La suma de tres números consecutivos es 48? ¿Cuál es mayor?" se transforman en variables categóricas. Estas preguntas miden habilidades cognitivas y se convierten en variables dummy para que los modelos puedan utilizar esta información.

6.1. Creación de la variable objetivo y análisis de relevancia

La variable objetivo (Y) seleccionada para este estudio es "¿Cuáles son sus ingresos netos mensuales?". Esta variable se trató inicialmente como una variable categórica ordinal, donde cada categoría representa un rango diferente de ingresos mensuales. Para simplificar el problema de clasificación y mejorar la capacidad del modelo para predecir el riesgo crediticio, las categorías originales se agruparon en dos clases principales:

Ingresos Bajos (0): Incluye respuestas en los rangos "Menor o igual a 850 dólares" y "Entre 850 y 1500 dólares".

Ingresos Altos (1): Comprende los rangos "Entre 1500 y 2500 dólares", "Entre 2500 y 4000 dólares" y "5000 dólares o más".

El objetivo de esta categorización es simplificar la clasificación y maximizar la capacidad del modelo para distinguir entre diferentes niveles de riesgo financiero basados en los ingresos. Al agrupar los ingresos medios con los ingresos altos bajo la misma categoría (1), se prioriza la identificación de aquellos individuos con ingresos bajos, ya que estos podrían tener un mayor riesgo de incumplimiento.

Para realizar el análisis de relevancia de esta variable objetivo, se ponderaron las diferentes categorías en función de su frecuencia en los datos y su impacto potencial en la evaluación del riesgo crediticio. Esta estrategia asegura que el modelo pueda aprender a distinguir correctamente entre los individuos con mayores y menores riesgos financieros.

Adicionalmente, se asignó un mayor peso a la categoría de ingresos bajos (0) durante el entrenamiento del modelo. Esta ponderación refleja la importancia de identificar con precisión a los individuos con menores ingresos, ya que estos son los que tienen más probabilidades de incumplir sus obligaciones crediticias. El ajuste de las proporciones de las clases y la aplicación de técnicas de balanceo cuando fue necesario garantizó que el modelo no estuviera sesgado hacia la clase mayoritaria y que pudiera aprender de manera efectiva los patrones relevantes en los datos.

Este enfoque en la construcción de la variable objetivo es fundamental para desarrollar un modelo robusto y fiable que pueda proporcionar predicciones precisas sobre el riesgo crediticio, especialmente en el contexto de la inclusión financiera.

7. División del conjunto de datos en conjuntos de entrenamiento y prueba

Una vez que se transformaron las variables y se estableció la variable objetivo, se procedió a dividir los datos en conjuntos de entrenamiento y prueba. Para evaluar cómo diferentes proporciones de datos afectan el rendimiento del modelo, se realizaron múltiples divisiones de los datos con diferentes tamaños de prueba: 10%, 20%, 30%, y 50%. Este proceso, conocido como split de datos, permite evaluar la capacidad del modelo para generalizar a nuevos datos que no se han visto durante el entrenamiento.

El uso de diferentes tamaños de conjuntos de prueba es crucial para entender la estabilidad del modelo y su capacidad de generalización. Por ejemplo, con un 10% de datos para prueba, se está utilizando la mayor parte de los datos para el entrenamiento, lo cual puede resultar en un sobreajuste. Por otro lado, con un 50% de datos para prueba, se evalúa el modelo en condiciones más desafiantes, lo que puede revelar su verdadera capacidad de generalización.

7.1. Justificación del escalado de variables e imputación de valores faltantes

Antes de entrenar los modelos, se aplicaron técnicas de escalado e imputación de datos para asegurar la calidad y la robustez del modelo. Se realizó un escalado de las variables numéricas para normalizar los datos, lo que significa que todas las variables se ajustaron a la misma escala. Esto es importante porque los algoritmos de Machine Learning como la Regresión Logística o SVM son sensibles a la magnitud de las variables, y las características con valores más grandes pueden dominar el modelo si no se escalan correctamente.

Durante el proceso de limpieza de datos, se detectaron algunos valores faltantes. Estos valores fueron imputados utilizando la técnica de imputación por la mediana, que reemplaza los valores faltantes con la mediana de la variable correspondiente. Esta técnica es robusta frente a valores atípicos y asegura que el modelo no sea sesgado por la falta de datos.

El escalado y la imputación son pasos esenciales para preparar los datos de manera efectiva para el entrenamiento del modelo. Sin estas transformaciones, los modelos podrían producir resultados menos precisos o inestables debido a la variabilidad y la falta de consistencia en los datos de entrada.

8. Construcción y entrenamiento de los modelos de Machine Learning

Con los datos preparados y transformados, se procedió a entrenar cuatro modelos de Machine Learning: Regresión Logística, Random Forest, XGBoost y Support Vector

Machines (SVM). Cada modelo fue evaluado en términos de su capacidad predictiva utilizando métricas de evaluación como precisión, recall, F1-score y el área bajo la curva ROC (Receiver Operating Characteristic). Estas métricas proporcionan una visión integral del rendimiento del modelo, especialmente en la identificación precisa de individuos con diferentes niveles de riesgo crediticio.

8.1. Matriz de confusión

La matriz de confusión se utiliza para visualizar el rendimiento de un modelo clasificando correcta o incorrectamente las instancias en distintas clases. Contiene cuatro elementos básicos:

- Verdaderos Positivos (VP): Número de instancias de la clase positiva (ingresos altos) correctamente predichas como positivas.
- Verdaderos Negativos (VN): Número de instancias de la clase negativa (ingresos bajos) correctamente predichas como negativas.
- Falsos Positivos (FP): Número de instancias de la clase negativa (ingresos bajos) incorrectamente predichas como positivas.
- Falsos Negativos (FN): Número de instancias de la clase positiva (ingresos altos) incorrectamente predichas como negativas.

La matriz de confusión proporciona una visión detallada de cómo un modelo maneja tanto las predicciones correctas como los errores. Es crucial para entender las implicaciones prácticas de los resultados del modelo. Por ejemplo, en el contexto del riesgo crediticio, un alto número de falsos negativos (FN) podría indicar que el modelo está clasificando incorrectamente a individuos con ingresos altos como de ingresos bajos, lo que podría llevar a una empresa a rechazar créditos a clientes potencialmente valiosos.

8.2. Precisión (PRECISON)

La precisión mide la proporción de instancias predichas como positivas que realmente son positivas.

Una alta precisión significa que cuando el modelo predice que un individuo tiene ingresos altos, generalmente está en lo correcto. En el contexto del riesgo crediticio, una alta precisión implica que el modelo es confiable para identificar correctamente a los individuos de ingresos altos y evitar falsos positivos, lo que minimiza el riesgo de otorgar créditos a clientes que no pueden pagar.

8.3. Sensibilidad o Recall (RECALL)

El recall mide la proporción de instancias positivas que el modelo identificó correctamente.

Un alto recall indica que el modelo captura la mayoría de las instancias positivas (ingresos altos). En términos prácticos, esto significa que el modelo tiene menos probabilidades de perder clientes potenciales de ingresos altos. Sin embargo, un alto recall puede venir a expensas de un mayor número de falsos positivos (FP), lo cual puede no ser ideal en ciertos escenarios.

8.4. F1-Score

El F1-score es la media armónica de precisión y recall, proporcionando un balance entre ambas métricas.

El F1-score es particularmente útil cuando se necesita un equilibrio entre precisión y recall, especialmente en situaciones donde hay una distribución de clase desequilibrada. En el contexto de la evaluación del riesgo crediticio, un F1-score alto para la clase de ingresos bajos (mayor riesgo) sería crucial para minimizar tanto los errores de falsos positivos como los falsos negativos.

8.5. Área bajo la curva ROC (AUC ROC)

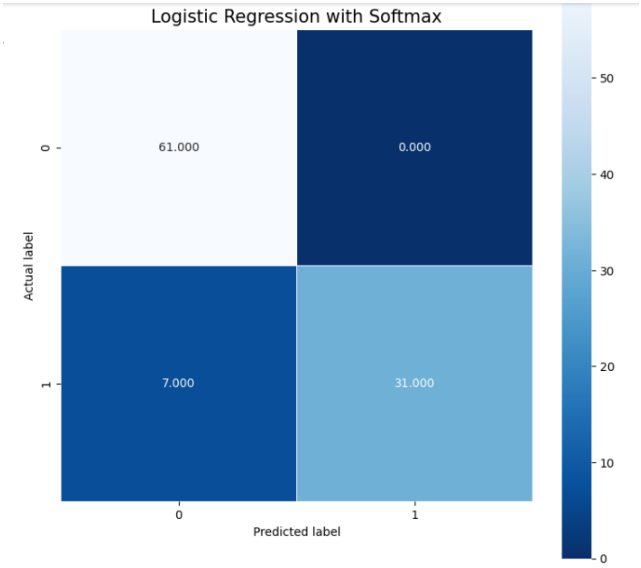
El AUC ROC mide la capacidad del modelo para distinguir entre clases. Un AUC cercano a 1 indica que el modelo tiene una excelente capacidad para clasificar correctamente las instancias de ingresos bajos y altos. Un AUC de 0.5 sugiere que el modelo no es mejor que un clasificador aleatorio.

9. Resultados del entrenamiento de cada modelo

9.1. Regresión logística con SOFTMAX

Matriz de Confusión:

Gráfico 3:



Accuracy Score for Logistic Regression with Softmax: 0.9292929292929293

	precision	recall	f1-score	support
0.0	0.90	1.00	0.95	61
1.0	1.00	0.82	0.90	38
accuracy			0.93	99
macro avg	0.95	0.91	0.92	99
weighted avg	0.94	0.93	0.93	99

Verdaderos Negativos (VN): 61

Falsos Positivos (FP): 0

Falsos Negativos (FN): 7

Verdaderos Positivos (VP): 31

Gráfico 4:

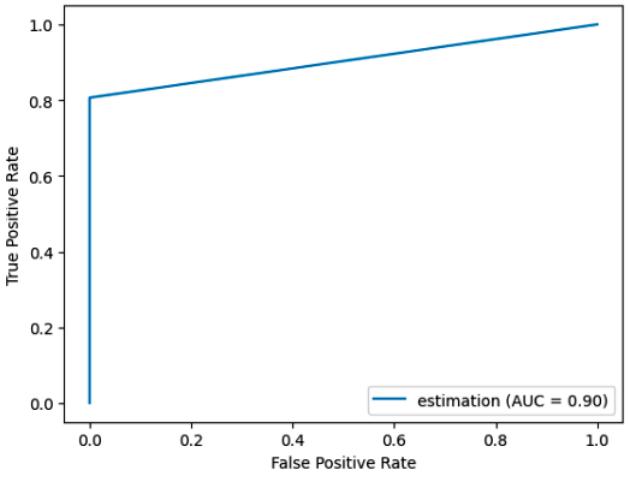
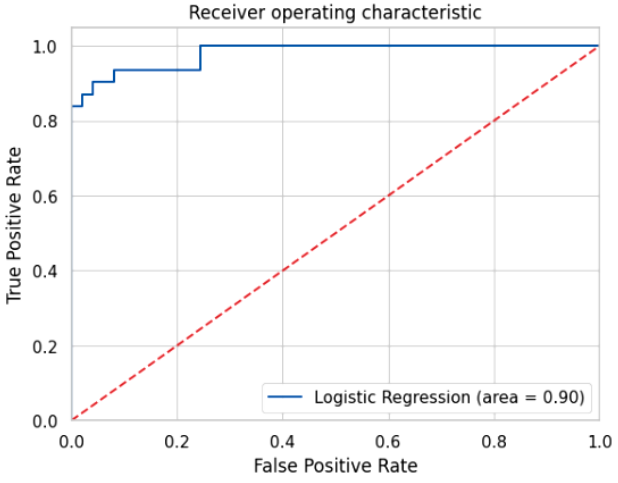


Gráfico 5:



La matriz de confusión muestra que el modelo de regresión logística predijo correctamente 61 casos de ingresos bajos (VN) y 31 casos de ingresos altos (VP). Sin embargo, hay 7 casos de ingresos altos que fueron clasificados incorrectamente como ingresos bajos (FN), y 0 casos de ingresos bajos fueron incorrectamente clasificados como ingresos altos (FP).

9.1.2. Métricas de Evaluación

Precisión para ingresos bajos (0): 0.90

Recall para ingresos bajos (0): 1.00

F1-Score para ingresos bajos (0): 0.95

Precisión para ingresos altos (1): 1.00

Recall para ingresos altos (1): 0.82

F1-Score para ingresos altos (1): 0.90

Accuracy: 0.93

AUC ROC: 0.90

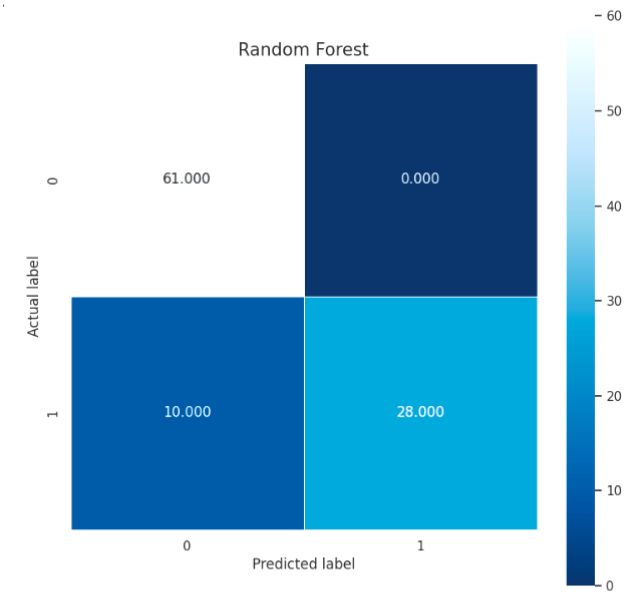
La regresión logística con Softmax mostró un rendimiento sólido con una alta precisión general (0.93) y un AUC ROC de 0.90, lo que indica una buena capacidad para distinguir entre individuos de ingresos bajos y altos. El modelo tiene

una excelente precisión para ambas clases, pero su recall para la clase de ingresos altos (1) es ligeramente menor (0.82), lo que implica que algunos individuos con ingresos altos no fueron correctamente identificados.

9.2. Random Forest

Matriz de Confusión:

Gráfico 6:



Accuracy Score for Random Forest: 0.898989898989899

	precision	recall	f1-score	support
0.0	0.86	1.00	0.92	61
1.0	1.00	0.74	0.85	38
accuracy			0.90	99
macro avg	0.93	0.87	0.89	99
weighted avg	0.91	0.90	0.90	99

VN: 61

FP: 0

FN: 10

VP: 28

La matriz de confusión para el modelo de Random Forest muestra que predijo correctamente 61 casos de ingresos bajos (VN) y 28 casos de ingresos altos (VP). Hay 10 casos de ingresos altos clasificados incorrectamente como ingresos bajos (FN).

Gráfico 7:

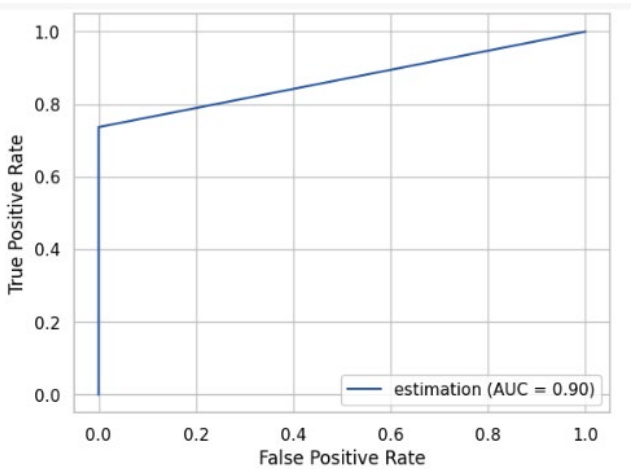
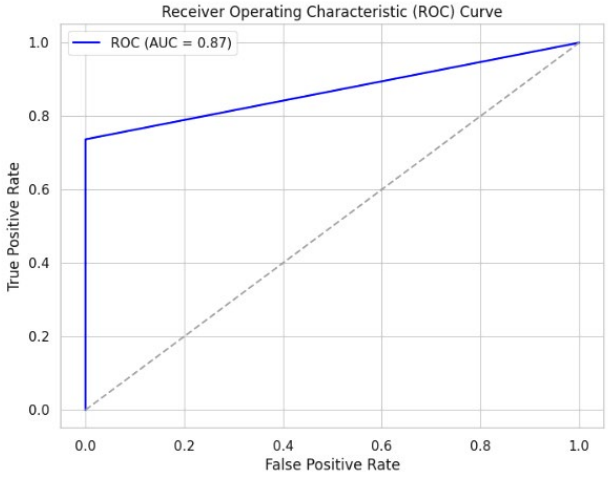


Gráfico 8:



Métricas de Evaluación:

Precisión para ingresos bajos (0): 0.86

Recall para ingresos bajos (0): 1.00

F1-Score para ingresos bajos (0): 0.92

Precisión para ingresos altos (1): 1.00

Recall para ingresos altos (1): 0.74

F1-Score para ingresos altos (1): 0.85

Accuracy: 0.90

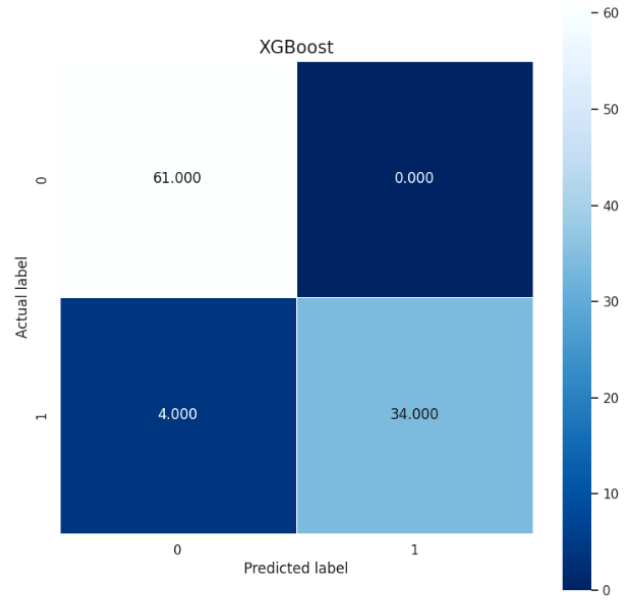
AUC ROC: 0.87

Random Forest presenta un desempeño robusto con una precisión de 0.90 y un AUC ROC de 0.87. El modelo muestra una precisión perfecta (1.00) para la clase de ingresos altos, pero el recall para esta clase es solo 0.74, lo que indica que es menos efectivo que la regresión logística en identificar todos los casos de ingresos altos.

9.3. XGBoost

9.3.1. Matriz de Confusión

Gráfico 9:



Accuracy Score for XGBoost: 0.9595959595959596

	precision	recall	f1-score	support
0.0	0.94	1.00	0.97	61
1.0	1.00	0.89	0.94	38
accuracy			0.96	99
macro avg	0.97	0.95	0.96	99
weighted avg	0.96	0.96	0.96	99

VN: 61

FP: 0

FN: 4

VP: 34

La matriz de confusión muestra que el modelo XGBoost predijo correctamente 61 casos de ingresos bajos (VN) y 34 casos de ingresos altos (VP). Hay solo 4 casos de ingresos altos clasificados incorrectamente como ingresos bajos (FN).

Gráfico 10:

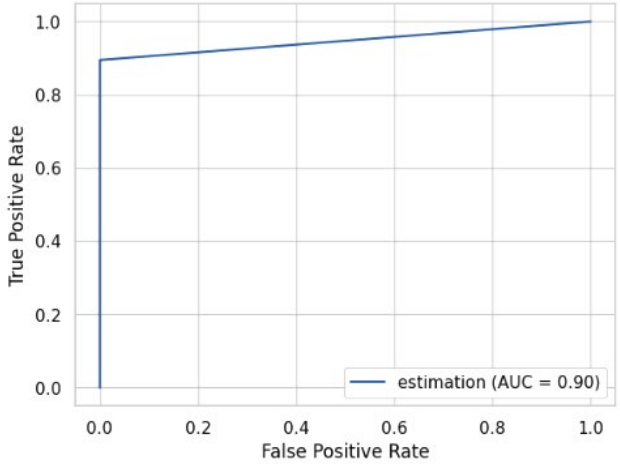
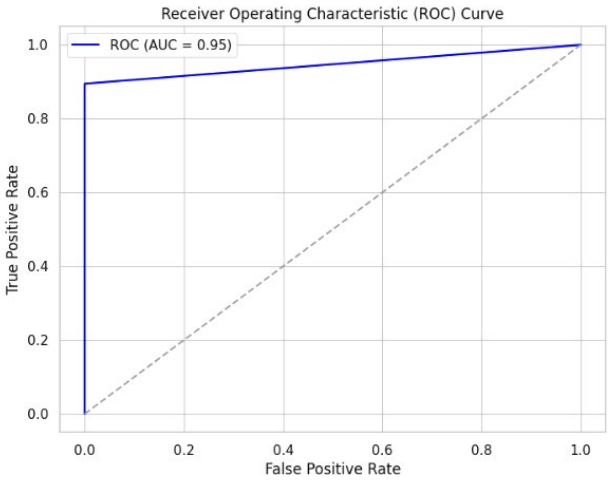


Gráfico 11:



Métricas de Evaluación:

Precisión para ingresos bajos (0): 0.94

Recall para ingresos bajos (0): 1.00

F1-Score para ingresos bajos (0): 0.97

Precisión para ingresos altos (1): 1.00

Recall para ingresos altos (1): 0.89

F1-Score para ingresos altos (1): 0.94

Accuracy: 0.96

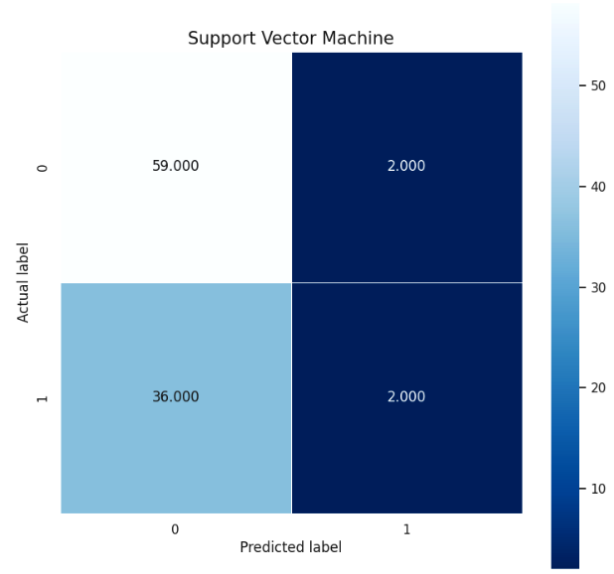
AUC ROC: 0.95

Análisis: XGBoost es el modelo con el mejor desempeño en términos de precisión (0.96) y AUC ROC (0.95). El modelo logra un alto recall para ambas clases, lo que indica que es muy efectivo para identificar correctamente tanto ingresos bajos como ingresos altos. El bajo número de falsos negativos (4) y la alta precisión sugieren que XGBoost es el modelo más confiable para predecir el riesgo crediticio.

9.4. Support Vector Machines (SVM)

9.4.1. Matriz de Confusión

Gráfico 12:



Accuracy Score for Support Vector Machine: 0.6161616161616161

	precision	recall	f1-score	support
0.0	0.62	0.97	0.76	61
1.0	0.50	0.05	0.10	38
accuracy			0.62	99
macro avg	0.56	0.51	0.43	99
weighted avg	0.57	0.62	0.50	99

VN: 59

FP: 2

FN: 36

VP: 2

La matriz de confusión para SVM muestra un desempeño significativamente peor. El modelo predijo correctamente 59 casos de ingresos bajos (VN) y solo 2 casos de ingresos altos (VP). Hay 36 casos de ingresos altos clasificados incorrectamente como ingresos bajos (FN).

Gráfico 13:

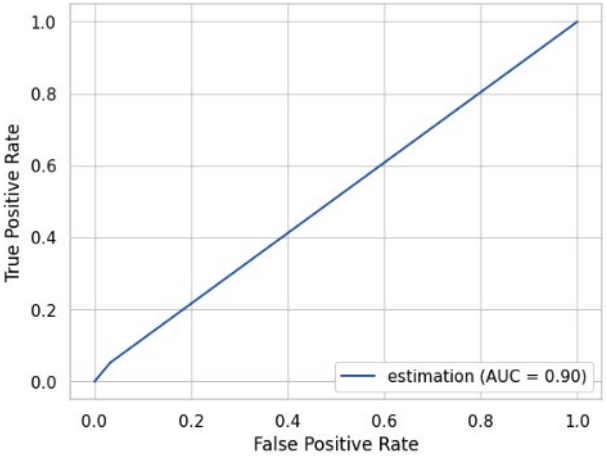
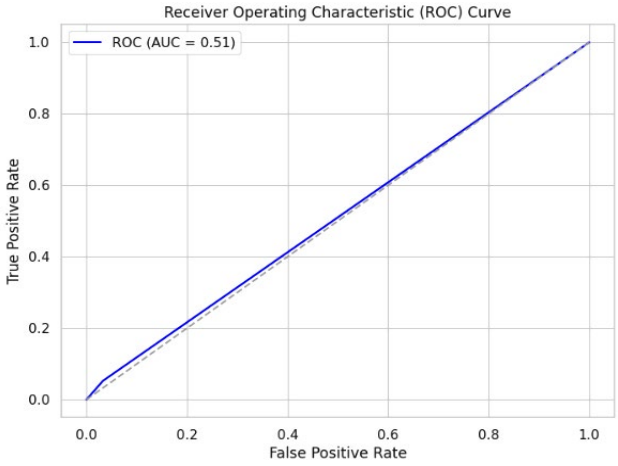


Gráfico 14:



Métricas de Evaluación:

Precisión para ingresos bajos (0): 0.62

Recall para ingresos bajos (0): 0.97

F1-Score para ingresos bajos (0): 0.76

Precisión para ingresos altos (1): 0.50

Recall para ingresos altos (1): 0.05

F1-Score para ingresos altos (1): 0.10

Accuracy: 0.62

AUC ROC: 0.51

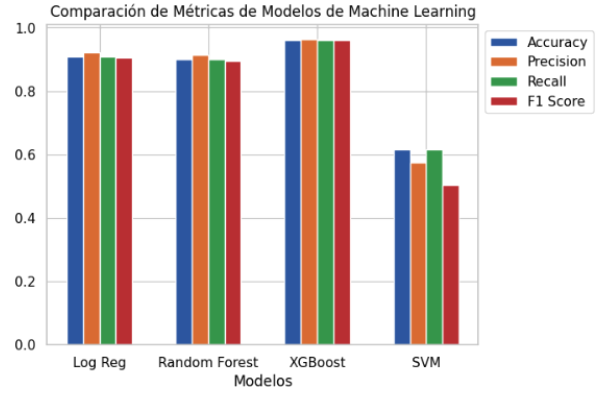
SVM muestra el peor rendimiento entre todos los modelos evaluados, con una precisión general de solo 0.62 y un AUC ROC de 0.51, lo cual es apenas mejor que un clasificador aleatorio. La baja precisión y recall para la clase de ingresos altos indican que el modelo no es adecuado para distinguir entre individuos de ingresos altos y bajos.

10. Comparación y conclusiones de los resultados obtenidos

Para evaluar de manera efectiva el rendimiento de los distintos modelos de machine learning aplicados a la predicción del riesgo crediticio, se presenta a continuación un gráfico comparativo que muestra las métricas clave —precisión (accuracy), precisión (precision), recall y F1-score— para cada uno de los modelos utilizados: Regresión Logística con Softmax, Random Forest, XGBoost, y Support Vector Machines (SVM).

El gráfico mostrado a continuación, proporciona una representación visual clara del rendimiento de cada modelo, permitiendo comparar directamente su eficacia en diferentes dimensiones del análisis predictivo. Además, refuerza las conclusiones previas basadas en las matrices de confusión y las métricas de evaluación individuales, proporcionando una base sólida para determinar el modelo más adecuado para la tarea de predicción del riesgo crediticio utilizando datos alternativos.

Gráfico 15:



	Model	Accuracy	Precision	Recall	F1 Score
0	Log Reg	0.909091	0.920779	0.909091	0.906108
1	Random Forest	0.898990	0.913217	0.898990	0.895164
2	XGBoost	0.959596	0.962082	0.959596	0.959115
3	SVM	0.616162	0.574588	0.616162	0.502627

El gráfico comparativo de las métricas confirma que XGBoost es el modelo más robusto y efectivo para la predicción del riesgo crediticio en este conjunto de datos. A continuación, se detalla un análisis comparativo de cada modelo basado en el gráfico y las métricas obtenidas:

Análisis Comparativo de los Modelos Basado en el Gráfico

XGBoost:

Rendimiento General: XGBoost se destaca con los valores más altos en todas las métricas clave: accuracy (0.959596), precision (0.962082), recall (0.959596), y F1-score (0.959115). Esto demuestra que es el modelo más equilibrado y eficaz para predecir tanto ingresos bajos como altos, minimizando tanto los falsos positivos como los falsos negativos.

Justificación de Resultados: La alta precisión y recall indican que XGBoost es capaz de identificar correctamente tanto las clases positivas como negativas. Esto es crucial en el contexto del riesgo crediticio, ya que asegura que los clientes de alto riesgo sean correctamente identificados mientras se minimizan los errores.

Regresión Logística con Softmax:

Rendimiento General: Este modelo muestra un rendimiento sólido, especialmente en términos de accuracy (0.909091) y precision (0.920779). Sin embargo, su recall (0.909091) es ligeramente menor que el de XGBoost, lo que sugiere una pequeña reducción en su capacidad para identificar todos los casos de ingresos altos.

Implicaciones Prácticas: Aunque la regresión logística es algo menos precisa que XGBoost en términos de recall, sigue siendo una opción viable para la predicción del riesgo crediticio debido a su simplicidad y eficiencia computacional. Sin embargo, la presencia de algunos falsos negativos puede implicar un riesgo de perder clientes de alto valor.

Random Forest:

Rendimiento General: El modelo de Random Forest presenta un desempeño razonablemente bueno, con un accuracy de 0.898990 y un F1-score de 0.895164. Sin embargo, su recall (0.898990) es menor que el de XGBoost y la regresión logística, lo que sugiere que es menos efectivo en identificar todos los casos de ingresos altos.

Consideraciones: Aunque Random Forest es robusto frente al sobreajuste, su rendimiento inferior en recall puede limitar su aplicabilidad en escenarios donde es crucial identificar todos los casos de alto riesgo.

Support Vector Machines (SVM):

Rendimiento General: SVM muestra el rendimiento más bajo en todas las métricas, con un accuracy de 0.616162, precision de 0.574588, recall de 0.616162, y un F1-score de 0.502627. Estos resultados indican que SVM no es un modelo adecuado para este problema específico.

Limitaciones del Modelo: El bajo rendimiento de SVM puede ser atribuido a varios factores, incluyendo el pequeño tamaño del conjunto de datos (200 encuestas), que puede no ser suficiente para que SVM aprenda patrones significativos. Además, el kernel lineal utilizado podría no ser el más adecuado para la naturaleza del conjunto de datos.

El tamaño del conjunto de datos (200 encuestas) es un factor crítico en la interpretación de estos resultados:

Sesgo y Varianza: Un conjunto de datos pequeño aumenta la probabilidad de sesgo de generalización, donde los modelos pueden aprender patrones específicos del conjunto de datos de entrenamiento que no generalizan bien a nuevos datos. Esto podría explicar el bajo rendimiento de SVM y el desempeño más equilibrado de XGBoost, que maneja bien datos con diferentes distribuciones y variaciones.

Impacto en las Métricas de Evaluación: Las métricas de rendimiento, como el F1-score y el AUC ROC, pueden verse afectadas por la falta de representatividad de un conjunto de datos pequeño. En aplicaciones prácticas, esto sugiere que los resultados obtenidos en este estudio deben ser validados con conjuntos de datos más grandes y diversos para asegurar su robustez y aplicabilidad general.

Basándonos en el análisis detallado de las matrices de confusión, las métricas de evaluación, y el gráfico comparativo, XGBoost emerge como el modelo más efectivo para la predicción del riesgo crediticio utilizando datos alternativos. Su capacidad para manejar la complejidad del

problema y su alto rendimiento en todas las métricas clave lo hacen ideal para aplicaciones en el mundo real. No obstante, es fundamental considerar que un conjunto de datos más grande y representativo podría proporcionar más confianza en estos resultados.

El gráfico presentado refuerza esta conclusión al mostrar claramente que XGBoost supera consistentemente a los otros modelos en todas las métricas evaluadas, haciendo de él la mejor opción para la tarea específica de predicción del riesgo crediticio en este contexto.

11. Conclusión del trabajo de fin de máster: potencial e impacto en la inclusión financiera

El presente Trabajo de Fin de Máster (TFM) se propuso desarrollar un modelo de score crediticio utilizando datos alternativos y algoritmos de inteligencia artificial para fomentar la inclusión financiera, especialmente en segmentos de la población vulnerables y no bancarizados. Este enfoque innovador se centra en romper las barreras tradicionales que limitan el acceso al crédito, ofreciendo una herramienta que evalúa el riesgo crediticio de manera más inclusiva y justa.

En el contexto actual, tanto en América Latina como en algunos casos específicos en España, una gran parte de la población permanece excluida del sistema financiero formal. Factores como la falta de historial crediticio, bajos ingresos, o la ausencia de garantías suelen impedir que estas personas accedan a servicios financieros básicos. Esta exclusión financiera afecta negativamente su capacidad para mejorar su calidad de vida, enfrentar emergencias financieras o invertir en oportunidades que puedan sacarles de la pobreza.

11.1. Potencial del modelo de score crediticio basado en datos alternativos

El desarrollo del modelo de score crediticio basado en datos alternativos y machine learning ha demostrado que es posible evaluar el riesgo crediticio de manera efectiva sin depender únicamente de los datos financieros tradicionales. Los datos alternativos utilizados en este estudio, provenientes de encuestas detalladas, han permitido capturar un panorama más completo del comportamiento financiero de los individuos, incluyendo factores como estabilidad laboral, educación, hábitos de consumo y acceso a servicios básicos. Esto es fundamental para ofrecer una evaluación más justa y precisa del riesgo crediticio, especialmente para aquellos que han sido excluidos por métodos de evaluación tradicionales.

11.2. Éxito del estudio y resultados obtenidos

El análisis comparativo de los modelos de machine learning, que incluyó algoritmos como Regresión Logística con Softmax, Random Forest, XGBoost y Support Vector Machines (SVM), reveló que es posible construir modelos robustos y eficientes para la evaluación del riesgo crediticio. XGBoost se destacó como el modelo más robusto, con la capacidad de manejar datos complejos y proporcionar resultados precisos en todas las métricas clave (accuracy, precision, recall y F1-score). Este modelo demostró ser particularmente eficaz en minimizar tanto los falsos positivos como los falsos negativos, lo cual es crucial para asegurar que los individuos de alto riesgo sean correctamente identificados y se minimicen los errores en la evaluación crediticia.

La conclusión de este TFM refuerza la idea de que el uso de datos alternativos y técnicas avanzadas de inteligencia artificial puede mejorar significativamente la precisión y equidad en la evaluación del riesgo crediticio. Además, subraya el potencial transformador de estas herramientas para impactar positivamente en la vida de miles de personas, especialmente en regiones con altas tasas de exclusión financiera. Al proporcionar una alternativa más inclusiva a los modelos tradicionales de evaluación crediticia, este enfoque podría facilitar el acceso al crédito para segmentos de la población que históricamente han sido excluidos, contribuyendo así a mejorar su estabilidad financiera y calidad de vida.

11.3. Implicaciones para el futuro de la inclusión financiera

El éxito de este estudio abre la puerta a una serie de oportunidades para expandir la inclusión financiera a través de soluciones basadas en tecnología e inteligencia artificial. Implementar estos modelos en la práctica podría permitir a las instituciones financieras expandir sus servicios a un grupo demográfico más amplio, mejorando la accesibilidad y equidad en el acceso al crédito. Esto es especialmente relevante en el contexto global actual, donde la inclusión financiera se reconoce como un motor clave para reducir la pobreza y fomentar la prosperidad económica.

Al aprovechar datos no tradicionales y técnicas avanzadas de machine learning, este TFM demuestra que es posible desarrollar soluciones innovadoras que respondan a las necesidades de los no bancarizados, proporcionando herramientas más justas y precisas para evaluar el riesgo crediticio. Este enfoque no solo beneficia a los individuos al facilitar el acceso a servicios financieros básicos, sino que también representa una oportunidad significativa para las instituciones financieras de ampliar su base de clientes y fomentar un crecimiento económico más inclusivo.

En conclusión, el estudio no solo ha sido un éxito en términos de demostrar la viabilidad de utilizar datos alternativos para la evaluación del riesgo crediticio, sino que también subraya la importancia de seguir innovando en el uso de tecnologías emergentes para promover la inclusión financiera a nivel global. Se recomienda seguir investigando y expandiendo este enfoque con conjuntos de datos más grandes y diversos, lo que podría proporcionar una base más sólida para la implementación práctica de estos modelos en diversos contextos socioeconómicos y geográficos

12. Apéndice

Encuesta utilizada para la recolección de datos:

Pregunta 1: ¿Fecha de nacimiento (DDMMYYYY)?

Pregunta 2: ¿Sexo?

- Hombre
- Mujer
- Prefiero no decirlo

Pregunta 3: ¿Estado civil?

- Casado
- Soltero
- Divorciado
- Unión libre

Pregunta 4: ¿Ciudad de residencia?

Pregunta 5: Su residencia es

- Casa propia
- Con padres
- Piso propio
- Piso de alquiler
- Habitación
- Otro

Pregunta 6: ¿Hijos a su cargo?

- 0
- 1
- 2
- 3
- 4
- 5 o más

Pregunta 7: Si al llegar a la esquina Jim dobla a la derecha o a la izquierda puede quedarse sin gasolina antes de encontrar una estación de servicio. Ha dejado una atrás, pero sabe que, si vuelve, se le acabará la gasolina antes de llegar. En la dirección que lleva no ve ningún surtidor. Por tanto:

- Puede que se quede sin gasolina
- Se quedará sin gasolina
- No debió seguir
- Se ha perdido
- Debería seguir a la derecha
- Debería girar a la izquierda

Pregunta 8: ¿Nivel de estudios?

- ESO
- Bachillerato
- Formación Profesional
- Grado
- Posgrado

Pregunta 9: ¿Situación laboral?

- Funcionario
- Contrato fijo
- Contrato temporal
- Autónomo
- Pensionista
- Contrato fijo discontinuo
- Desempleado
- Becario
- Otros

Pregunta 10: ¿Profesión?

Pregunta 11: ¿Tiempo en el cargo actual?

- Menos de un año
- De 1 a 5 años
- De 5 a 10 años
- De 10 a 15 años
- Más de 20 años

Pregunta 12: La nota media conseguida en una clase de 20 alumnos ha sido de 6. Ocho alumnos han suspendido con un 3 y el resto superó el 5. ¿Cuál es la nota media de los alumnos aprobados?

- 6
- 5
- 8
- 3

Pregunta 13: ¿Cuáles son sus ingresos netos mensuales?

- Menor o igual a 850 dólares
- Entre 850 y 1500 dólares
- Entre 1500 y 2500 dólares
- Entre 2500 y 4000 dólares
- 5000 dólares o más

Pregunta 14: ¿De cuánto dinero ahorrado dispone actualmente?

- Menos de 150 dólares
- Entre 150 y 300 dólares
- Entre 300 y 500 dólares
- Entre 500 y 1000 dólares
- Mas de 1000 dólares

Pregunta 15: ¿Qué número debería ser el siguiente? 9 16 23 30 37 44 51...

- 55 66
- 56 62
- 58 66
- 58 65

Pregunta 17: ¿Tiene alguna cuenta bancaria?

- Sí
- No

Pregunta 18: ¿Tiene contratado algún préstamo?

- No
- Hipotecario
- Automotriz
- Tributario
- Personal
- Tarjeta de crédito
- Otro

Pregunta 19: ¿Ha tenido algún retraso en los pagos de sus facturas de servicio en los últimos 12 meses?

- Si
- No

Pregunta 20: ¿Dispone de coche?

- Sí
- No
- Otro

Pregunta 21: ¿Tiene un smartphone?

- Sí
- No

Pregunta 22: ¿Ha tenido algún problema con la ley en los últimos 5 años?

- Sí
- No

Pregunta 23: La suma de tres números consecutivos es 48. ¿Cuál es mayor?

- 16
- 14
- 17
- 15

13. Bibliografía

Banco de España (2023) Un repaso de las diversas iniciativas desplegadas a nivel nacional e internacional para hacer frente a los riesgos de exclusión financiera. Documentos Ocasionales, N.º 2305. Disponible en: Banco de España.

Banco Mundial (2023) Global Findex Database 2023. Disponible en: https://www.worldbank.org/en/publication/globalfindex [11 junio 2024].

BPC Latin America (2023) Digital Banking Report. Disponible en: https://www.bpcbt.com/report/digital-banking-in-latin-america [11 junio 2024].

Deloitte (2023) Alternative Data: The Fuel for AI-Powered Credit Scoring. Disponible en: https://www2.deloitte.com/xe/en/insights.html?icid=disidenav_insights [11 junio 2024].

Ericsson (2023) Mobile Data Traffic Forecasts. Disponible en: https://www.ericsson.com/en/mobility-report/reports [11 junio 2024].

ICEMD (2023) Inteligencia Artificial y Datos Alternativos en la Evaluación del Riesgo Crediticio. Disponible en: https://www.icemd.com/digital-knowledge/articles/inteligencia-artificial-y-datos-alternativos-en-la-evaluacion-del-riesgo-crediticio/ [11 junio 2024].

IDB Publications (2023) Innovations in Credit Scoring for Financial Inclusion. Disponible en: https://www.iadb.org/en [11 junio 2024].

MasterCard (2023) Latinoamérica: Informe de Inclusión Financiera. Disponible en: https://www.mastercard.com/news/ [11 junio 2024].

McKinsey & Company (2023) The Digital Future of Credit Scoring. Disponible en: https://www.mckinsey.com/ [11 junio 2024].

Provenir (2023) Fintech Solutions for Financial Inclusion. Disponible en: https://www.provenir.com/solutions/financial-inclusion/ [11 junio 2024].

Baesens, B., Setiono, R., Mues, C., y Vanthienen, J. (2003) 'Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation', Management Science, 49(3), pp. 312-329.

Breiman, L. (2001) 'Random Forests', Machine Learning, 45(1), pp. 5-32.

Chen, T., y Guestrin, C. (2016) 'XGBoost: A Scalable Tree Boosting System', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

Cortes, C., y Vapnik, V. (1995) 'Support-Vector Networks', Machine Learning, 20(3), pp. 273-297.