TRABAJO INVESTIGATIVO:

MODELO DE *CREDIT SCORE* ALTERNATIVO PARA PERSONAS CON INGRESOS INDETERMINADOS EN COLOMBIA: BASADO EN *MACHINE LEARNING*

JOHAN SEBASTIAN MAYOR CORTÉS

JUAN PABLO PORRAS CASANOVA

PONTIFICIA UNIVERSIDAD JAVERIANA CALI

FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS

MAESTRÍA EN FINANZAS

SANTIAGO DE CALI

2023

TRABAJO INVESTIGATIVO:

MODELO DE *CREDIT SCORE* ALTERNATIVO PARA PERSONAS CON INGRESOS INDETERMINADOS EN COLOMBIA: BASADO EN *MACHINE LEARNING*

JOHAN SEBASTIAN MAYOR CORTÉS

JUAN PABLO PORRAS CASANOVA

Trabajo de grado presentado como requisito parcial para optar por el título de Magíster en Finanzas

Director:

DAVID ARANGO

Estadístico / Magister en Economía Aplicada

PONTIFICIA UNIVERSIDAD JAVERIANA CALI

FACULTAD DE CIENCIAS ECONÓMICAS Y ADMINISTRATIVAS

MAESTRÍA EN FINANZAS

SANTIAGO DE CALI

2023

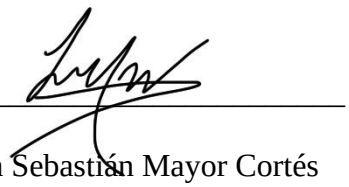
Santiago de Cali, 3 de julio de 2023

Doctor (a)
FABIÁN FERNANDO OSORIO TINOCO
Decano
Facultad de Ciencias Económicas y Administrativas
Pontificia Universidad Javeriana
La Ciudad

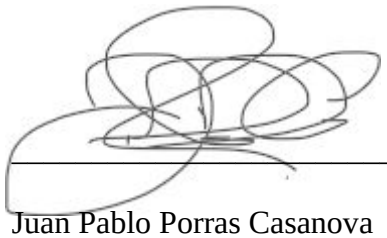
Por medio de la presente estamos entregando a usted el Trabajo de Grado cuyo título es MODELO DE *CREDIT SCORE* ALTERNATIVO PARA PERSONAS CON INGRESOS INDETERMINADOS EN COLOMBIA: BASADO EN *MACHINE LEARNING*.

Esperamos que este Trabajo cumpla con los requisitos académicos exigidos y que alcance el propósito para el cual fue elaborado.

Atentamente



Johan Sebastián Mayor Cortés
Cédula 1.107.103.440



Juan Pablo Porras Casanova
Cédula 1.095.824.111

Santiago de Cali, 3 de julio de 2023

Doctor (a)
Fabián Fernando Osorio Tinoco
Facultad de Ciencias Económicas y Administrativas
Pontificia Universidad Javeriana
La Ciudad

Por medio de la presente me permito comunicarle, que en mi calidad de director de trabajo de grado he leído detenidamente el informe final del estudio titulado "MODELO DE *CREDIT SCORE* ALTERNATIVO PARA PERSONAS CON INGRESOS INDETERMINADOS EN COLOMBIA: BASADO EN *MACHINE LEARNING"*, realizado por los estudiantes de la Facultad de Ciencias Económicas y Administrativas de la Universidad Javeriana Johan Sebastian Mayor Cortés con c.c. 1.107.103.440 y Juan Pablo Porras Casanova con c.c. 1.095.824.111, y considero que cumple con todos los requisitos requeridos para ser presentada a evaluación.

Atentamente

David Arango Londoño

DAVID ARANGO
Director del Trabajo de Grado

ARTÍCULO 23 de la resolución No. 13 de julio 6 de 1946.

“La Universidad no se hace responsable por los conceptos emitidos por sus alumnos en sus trabajos de Tesis. Sólo velará porque no se publique nada contrario al dogma y a la moral católica y porque la Tesis no contenga ataques o polémicas puramente personales; antes bien, se vea en ellas al anhelo de buscar la Verdad y la Justicia”.

"**MODELO DE CREDIT SCORE ALTERNATIVO PARA PERSONAS CON INGRESOS INDETERMINADOS EN COLOMBIA: BASADO EN MACHINE LEARNING**" Aprobado por el Comité de Trabajos de Grado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana para optar por el título de Magíster en Finanzas.

Fabian Fernando Osorio Tinoco
Decano
Facultad de Ciencias Económicas y Administrativas

Gladis Rodríguez Muñoz
Directora de Posgrados

Orlando Joaqui Barandica
Jurado

David Arango Londoño

David Arango Londoño
Director del Trabajo de Grado

Santiago de Cali, 23 de agosto de 2023

Contenido

Pág.

Lista de tablas

Pág.

Lista de figuras

Pág.

Lista de anexos

Pág.

Resumen

El acceso al crédito en Colombia sigue siendo un desafío para una gran parte de la población desbancarizada. A pesar de los esfuerzos realizados para aumentar la inclusión financiera, muchas personas no pueden obtener préstamos debido a la falta de historial crediticio o a los requisitos exigidos por la banca tradicional. En este estudio, se plantea la pregunta de investigación de cómo generar un modelo alternativo de *credit score* utilizando *machine learning* para analizar perfiles de personas con ingresos indeterminados. Los objetivos del estudio son identificar variables representativas, construir un modelo utilizando el algoritmo *RandomForest*, comparar este modelo con los modelos tradicionales de regresión GLM, regresión logística y *support vector machine*, y evaluar la viabilidad de las colocaciones de crédito mediante simulaciones. La metodología incluye el uso de datos de la encuesta de demanda de inclusión financiera y el tratamiento de los datos utilizando en R. Los resultados muestran que se pueden obtener modelos alternativos de *credit score* utilizando variables categóricas y *machine learning*.

Palabras clave: crédito, inclusión financiera, modelo alternativo, *machine learning*, variables categóricas, Colombia.

Abstract

Access to credit in Colombia continues to be a challenge for a large part of the unbanked population. Despite efforts to increase financial inclusion, many people are unable to obtain loans due to lack of credit history or traditional banking requirements. In this study, the research question of how to generate an alternative credit score model using machine learning to analyze profiles of people with indeterminate income is posed. The objectives of the study are to identify representative variables, build a model using the RandomForest algorithm, compare this model with traditional GLM regression, logistic regression, and support vector machine models, and evaluate the feasibility of credit placements through simulations. The methodology includes the use of data from the financial inclusion demand survey and the treatment of the data using R. The results show that alternative credit score models can be obtained using categorical variables and machine learning.

Keywords*: credit, financial inclusion, alternative model, machine learning, categorical variables, Colombia.*

1. Introducción

La inclusión financiera se refiere a la capacidad de las personas de tener acceso a productos y servicios financieros básicos, como cuentas de ahorro, créditos, seguros y otros servicios que pueden ayudar a mejorar su bienestar económico y social. En Colombia, la inclusión financiera ha sido una prioridad en la agenda gubernamental y empresarial, con el objetivo de reducir la brecha entre los que tienen acceso a los servicios financieros y los que no (Banco Mundial, 2022). Sin embargo, para muchas personas especialmente aquellas que no tienen un historial crediticio establecido, es difícil acceder a créditos y otros productos financieros. En estos casos, un modelo de *credit score* alternativo podría ser una solución efectiva para fomentar la inclusión (ASOBANCARIA, 2022).

Estos modelos alternativos utilizan información no tradicional para evaluar el riesgo crediticio de una persona, como los pagos de facturas de servicios públicos, el uso de tarjetas de crédito prepagadas y la actividad en redes sociales. Estos modelos pueden ser especialmente útiles para evaluar el riesgo crediticio de personas que no tienen un historial crediticio establecido, lo que les permite acceder a servicios financieros y mejorar su situación económica. Por otra parte, enfoques de la inteligencia artificial como el *machine learning* han permitido desarrollar modelos no lineales con alto grado de predictibilidad (Maisueche, 2019); así mismo sus bondades les permiten reducir el sobreajuste de las variables permitiendo que el mismo modelo seleccione las variables de mayor relevancia. Con lo anterior se logra obtener resultados razonables con poca información sin importar si sus variables son numéricas o categóricas.

En este marco, el objetivo de este trabajo se basó en estimar un modelo de *credit score* alternativo para colocaciones de crédito de bajo monto, utilizando como innovación herramientas de machine learning y la sustitución de la variable tradicional del ingreso. Buscando por medio de

factores categóricos calcular y evaluar el riesgo de crédito. Para ello se hizo uso de la información disponible de la encuesta demanda de inclusión financiera en Colombia publicada por la banca de las oportunidades (Pérez, 2022), donde por medio de un desarrollo empírico se preparó la información categórica recopilada en la base mencionada para seleccionar y crear variables que posteriormente permitieron la inferencia de la probabilidad de incumplimiento del sujeto de crédito. Con este fin, se planteó el desarrollo de un modelo *machine learning* en donde se comparan los algoritmos *RandomForest, Suport Vector Machine* y la regresión logística. La asertividad de estos fue medida por medio de la variable *Accuracy* la cual fue un referente para la elección del método *machine learning* que permitió una mejor aproximación con la información disponible. Adicionalmente, se analizó la viabilidad del modelo con validación cruzada para proyectar la eficiencia del modelo elegido en la práctica con simulaciones de préstamos de bajo monto.

Con lo anterior se logró llegar a un modelo de variables categóricas que permitió inferir el riesgo de crédito sustituyendo la variable tradicional del ingreso, que en muchos casos para la población financiera excluida es de difícil acceso o verificación ocasionando su situación de exclusión. Así mismo, dado que el modelo que se planteó se basa en aspectos cualitativos del solicitante se excluye el análisis de burós de crédito con el que tradicionalmente se mide la experiencia crediticia lo cual es una limitante para acceder al crédito tradicional. Por todo lo anterior se constituyó el resultado del desarrollo académico en un aporte al proceso de otorgamiento de créditos de bajo monto en Colombia.

El acceso al crédito en Colombia pese a que ha aumentado en los últimos años sigue teniendo una fuerte barrera de crecimiento en la población des bancarizada la cual representa gran parte de la población. Según un informe de la Superintendencia Financiera de Colombia- SFC (2018), el 32% de la población adulta en Colombia no tenía ningún producto financiero. Además,

muchas personas que tienen acceso a servicios financieros formales no pueden obtener préstamos debido a la falta de historial crediticio o el incumplimiento de requisitos exigidos por la banca tradicional. Teniendo presente el efecto apalancador que tiene la deuda financiera, esta exclusión generada por el modelo tradicional pone en desventaja a esta población con lo que se limita su capacidad para iniciar un negocio, invertir en su educación o vivienda.

En Colombia se han realizado esfuerzos para incrementar el acceso a productos financieros, según el más reciente estudio de inclusión financiera la población adulta colombiana (personas de 60 años o más) que tiene al menos un producto financiero pasó del 51,6% en 2017 al 65,3% en 2022 (Colombia Fintech, 2023).

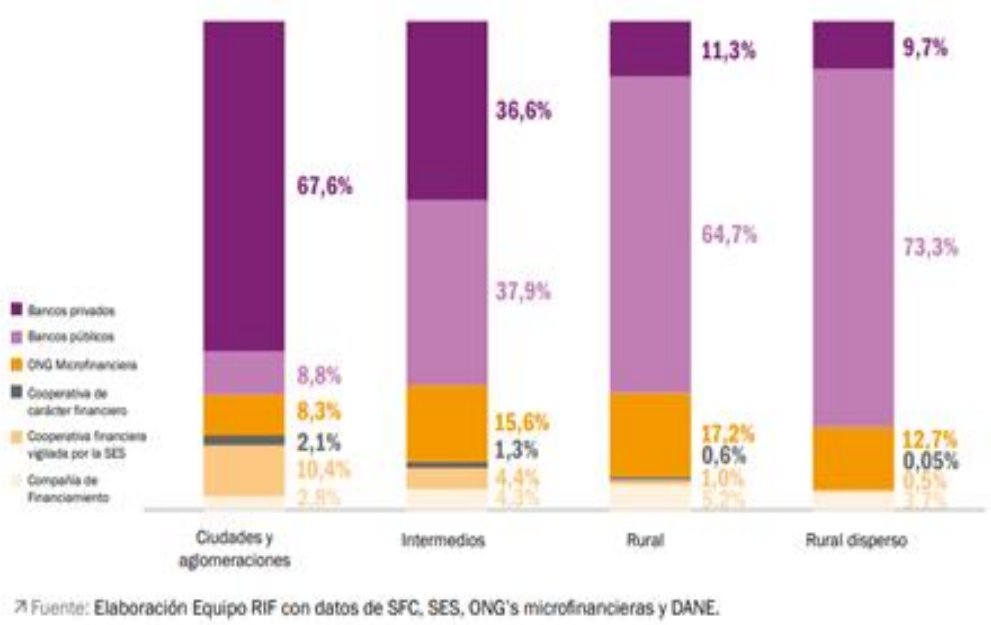
Lo cual si bien presenta un avance también demuestra que aún queda camino por recorrer. El camino a seguir es lastrado por la escasa información disponible sobre la población excluida para correr los modelos de riesgo tradicional, los cuales contemplan variables como el salario promedio, comportamiento de pago registrado en centrales de riesgo, entre otras. La exigencia de esta información conlleva a que en áreas rurales o en comunidades marginadas sean especialmente vulnerables a la exclusión financiera, ya que a menudo son lugares donde prima el trabajo informal. Dado lo anterior, es relevante indicar que si bien esta población recoge gran parte de las características de la población con exclusión financiera comparten el poco deseado grupo con los estudiantes que carecen de experiencia crediticia los cuales son excluidos de la banca tradicional ocasionando que recurran a mecanismos de financiación irregular. Como se observa en la figura 1 la banca tradicional obtiene mayor presencia en las principales ciudades con un 67,6% y en el sector rural esta participación decae hasta un 11,3% siendo más crítico en el rural disperso llegando al 9,7%. Esto ocasiona que la única fuente de financiación dependa de la banca pública y planes

estatales para mejorar la inclusión financiera en estas zonas (Superintendencia Financiera de Colombia-SFC, 2021).

La figura 1 explica el comportamiento de la población por locación (urbana/intermedios/rural) donde son atendidas para desembolso de microcréditos por entidades financieras.

Figura 1

Comportamiento de solicitud de microcrédito en Colombia 2021



Nota. Tomado de *Reporte de inclusión financiera-RIF*, por Superintendencia Financiera de Colombia-SFC (2021)

Entendiendo que las brechas de la inclusión financiera habitualmente son causadas por la escasa información cuantitativa verificable entorno a variables sensibles como el historial crediticio o ingresos, se ve necesario ir de la mano con mecanismos disruptivos para seguir evolucionando de la mano de la innovación con las tecnologías de inteligencia artificial enfocadas

en el *machine learning*, las cuales permiten inferir modelos con alto grado de asertividad haciendo uso de variables categóricas y poca información. Se origina la pregunta de investigación a razón de los hallazgos.

De lo anterior surge el siguiente interrogante: ¿Cómo generar un modelo alternativo de *credit score*, que permita analizar perfiles de personas con ingresos indeterminados haciendo uso del *machine learning*?

2. Marco teórico

Se precede de varios enfoques los cuales buscan entender y solucionar factores que aporten a la inclusión financiera, algunos direccionados a analizar determinantes en acceso de crédito formal e informal, otros que buscan comprender modelos cualitativos para asignar un *credit score* a préstamos de consumo ordinario, así como estudios que simulan préstamos masivos para entender la viabilidad de llevar el servicio de microcrédito de la mano con el desarrollo de un país.

2.1 Inclusión financiera

La inclusión financiera es fundamental para el desarrollo económico y social de un país, ya que permite que personas y empresas accedan a servicios financieros que les permitan ahorrar, invertir y obtener crédito para impulsar sus actividades productivas (Gamba, et al., 2016).

La inclusión financiera también es un elemento importante para reducir la pobreza y la desigualdad, ya que permite que las personas accedan a herramientas para mejorar su bienestar económico (Demirgue et al., 2014).

En Colombia, la inclusión financiera ha sido promovida por diferentes instituciones, tanto públicas como privadas. En el marco del Plan Nacional de Desarrollo 2018-2022, se han implementado diferentes estrategias para mejorar la inclusión financiera en el país, incluyendo el fortalecimiento de la oferta de servicios financieros en zonas rurales y la promoción de la educación financiera entre la población (Ministerio de Vivienda, Ciudad y Territorio, 2021).

Una de las herramientas utilizadas para promover la inclusión financiera en Colombia ha sido el uso de tecnologías financieras, como la banca móvil y las plataformas de pagos digitales. Estas tecnologías permiten que personas y empresas accedan a servicios financieros desde

cualquier lugar y en cualquier momento, lo que facilita el acceso a servicios financieros en zonas donde la oferta tradicional de servicios financieros es limitada (Chaparro, 2021).

2.2 Teoría de inclusión financiera

La teoría de la inclusión financiera se refiere a la idea de que todas las personas deben tener acceso a los servicios financieros (Banco Mundial, 2022), los cuales son necesarios para llevar a cabo sus actividades económicas y mejorar su calidad de vida. La inclusión financiera implica que las personas tengan acceso a una amplia gama de productos y servicios financieros, incluyendo cuentas bancarias, crédito, seguros y pagos electrónicos, entre otros. A continuación, se presentan algunas referencias que ejemplifican la teoría de la inclusión financiera:

El Banco Mundial publica regularmente el Informe *Global Findex*, que proporciona datos sobre la inclusión financiera en todo el mundo. En el informe más reciente, se analiza el acceso a los servicios financieros y se destaca la importancia de la inclusión financiera para el desarrollo económico (Demirgüç et al., 2018).

Este informe del Banco Mundial presenta un análisis más detallado de los resultados del Informe Global Findex 2017. El informe destaca la necesidad de mejorar la inclusión financiera en todo el mundo, especialmente en los países en desarrollo (Demirgüç et al., 2018).

Los autores Jappelli y Pagano (1994) presentan un modelo teórico de cómo los obstáculos a la inclusión financiera pueden limitar el ahorro y el crecimiento económico. Los autores argumentan que la inclusión financiera puede mejorar el bienestar económico al permitir que las personas ahorren y participen en la economía de manera más efectiva.

En resumen, la teoría de la inclusión financiera destaca la importancia de garantizar que todas las personas tengan acceso a los servicios financieros que necesitan para mejorar su calidad

de vida. La inclusión financiera puede tener implicaciones importantes para el crecimiento económico y el bienestar de las personas en todo el mundo.

2.3 Economía popular

La financiación de la economía popular ha sido una preocupación en Colombia durante muchos años. El acceso a crédito y otros servicios financieros es una de las principales barreras para el crecimiento de estas empresas (Castro, 2018). Además, muchos de estos trabajadores informales y pequeños empresarios tienen dificultades para acceder a los sistemas financieros tradicionales debido a su falta de historial crediticio o garantías.

En el PND de Colombia se han establecido estrategias para mejorar el acceso de la economía popular al financiamiento. Por ejemplo, el Plan Nacional de Desarrollo 2018-2022 incluye programas de financiamiento para las pequeñas y medianas empresas, que incluyen el acceso a crédito y otros servicios financieros (Ministerio de Hacienda y Crédito Público, 2019). Además, el PND promueve la creación de instituciones financieras especializadas que atiendan las necesidades de los trabajadores informales y pequeños empresarios (Ministerio de Vivienda, Ciudad y Territorio, 2021). En conclusión, la financiación de la economía popular es un tema relevante en el contexto del desarrollo económico de Colombia y ha sido abordado en el Plan Nacional de Desarrollo del país. Es necesario seguir trabajando en el desarrollo de estrategias para mejorar el acceso de la economía popular al financiamiento y promover su crecimiento.

2.4 Factores que se pueden utilizar para un *credit score* alternativo

El sistema de puntuación crediticia es una herramienta clave que utilizan los prestamistas para evaluar la solvencia crediticia de un individuo. Sin embargo, los sistemas de puntuación crediticia tradicionales se basan principalmente en factores como el historial de pagos y el uso de

crédito. En la actualidad, existen diferentes factores que pueden ser utilizados para crear un sistema de puntuación crediticia alternativo, entre ellos se encuentran:

• Historial de pagos de servicios públicos. El historial de pagos de servicios públicos puede ser un indicador de la capacidad de una persona para pagar sus deudas. Algunos prestamistas han comenzado a utilizar la información de pagos de servicios públicos para evaluar la solvencia crediticia de un individuo. Este factor es especialmente útil para aquellos que tienen poco o ningún historial de crédito (National Consumer Law Center, 2019).

• Historial de transacciones bancarias. El historial de transacciones bancarias puede proporcionar información valiosa sobre los hábitos de gasto de una persona y su capacidad para administrar sus finanzas. Los prestamistas pueden revisar los estados de cuenta bancarios para evaluar la frecuencia y el monto de las transacciones realizadas. Este factor puede ser útil para aquellos que tienen un historial de crédito limitado o para aquellos que no tienen acceso a crédito (National Consumer Law Center, 2019).

• Historial de empleo. La estabilidad laboral y los ingresos regulares son factores críticos para determinar la solvencia crediticia de una persona. El historial de empleo puede ayudar a los prestamistas a determinar la capacidad de un individuo para pagar sus deudas a largo plazo. Este factor es especialmente útil para aquellos que tienen un historial de crédito limitado o para aquellos que no tienen acceso a crédito (National Consumer Law Center, 2019).

• Uso de redes sociales y otras fuentes de datos. Los prestamistas pueden utilizar diferentes fuentes de datos para evaluar el riesgo crediticio de un individuo. Algunos prestamistas han comenzado a utilizar las redes sociales para evaluar la solvencia crediticia de un individuo. Este factor es especialmente útil para aquellos que tienen un historial de crédito limitado o para aquellos que no tienen acceso a crédito (SEON, 2022).

2.5 Selección adversa

La teoría de la selección adversa es una teoría económica que sostiene que la presencia de información asimétrica entre los prestamistas y los prestatarios puede resultar en una situación en la que los prestatarios menos solventes tengan más incentivos para solicitar préstamos que los prestatarios más solventes. La puntuación crediticia es una herramienta que se utiliza para evaluar el riesgo crediticio de los prestatarios y, por lo tanto, puede desempeñar un papel importante en la reducción de la selección adversa en el mercado crediticio. La literatura existente ha abordado la relación entre la teoría de la selección adversa y la puntuación crediticia de diversas maneras. A continuación, se presentan algunas referencias que ejemplifican esta relación:

En este artículo clásico, Stiglitz y Weiss (1981) presentan una teoría de cómo la información asimétrica puede dar lugar a la selección adversa en el mercado crediticio y cómo esto puede llevar a la exclusión de los prestatarios más solventes. Los autores argumentan que la puntuación crediticia puede ayudar a reducir la selección adversa al proporcionar información adicional sobre el riesgo crediticio de los prestatarios.

En el informe del Banco Interamericano de Desarrollo, los autores abordan el papel que desempeña la selección adversa en el mercado crediticio informal y cómo esto puede limitar el acceso de los prestatarios menos solventes a los préstamos. Los autores argumentan que la puntuación crediticia puede ayudar a reducir la selección adversa en el mercado crediticio informal y mejorar el acceso al crédito para los prestatarios menos solventes (Didier & Schmukler, 2014).

Laibson (1997) presenta un modelo teórico de cómo la selección adversa puede afectar la elección de los prestatarios entre distintos tipos de préstamos. El autor muestra cómo la puntuación crediticia puede desempeñar un papel importante en la reducción de la selección adversa al proporcionar información adicional sobre el riesgo crediticio de los prestatarios.

2.6 Artículos de información no crediticia

El uso de la información no crediticia se refiere a la utilización de factores adicionales a los datos financieros tradicionales para evaluar el riesgo crediticio. En el contexto de los microcréditos, esto puede incluir información sobre el carácter y la historia laboral del solicitante, así como su historial de pago de servicios públicos y teléfono celular. El uso de la información no crediticia puede ser especialmente útil para las personas que no tienen un historial crediticio establecido o para aquellos que viven en comunidades donde los datos crediticios no están ampliamente disponibles.

A continuación, se presenta una referencia de un estudio sobre el uso de la información no crediticia en microcréditos:

Kiva (2013) es una plataforma de préstamos en línea que presenta un protocolo para el uso de la información no crediticia en la evaluación del riesgo crediticio. El protocolo incluye una serie de factores que pueden ser utilizados para evaluar el carácter y la capacidad de pago del solicitante, incluyendo la historia laboral, la experiencia empresarial y el historial de pago de servicios públicos y teléfono celular.

En el mercado crediticio español, el Banco de España ha identificado varios tipos de modelos de *scoring* alternativos, incluyendo aquellos que utilizan información sobre el comportamiento de pago de servicios públicos, información sobre transacciones financieras y datos obtenidos de las redes sociales (Banco de España, 2017).

Para este estudio, el Banco de España (2017) examina los modelos de *scoring* alternativos que se utilizan en el mercado crediticio español, incluyendo sus ventajas y desventajas. Además, presenta una descripción detallada de los distintos tipos de modelos de *scoring* alternativos y su aplicación en la evaluación del riesgo crediticio.

En este artículo Khandani, et al. (2010) examinan el uso de algoritmos de aprendizaje automático para la evaluación del riesgo crediticio en Estados Unidos. Los autores demuestran que estos algoritmos pueden mejorar la precisión del *scoring* crediticio al incorporar datos no tradicionales.

En conclusión, el uso de la información no crediticia puede ser una herramienta útil para evaluar el riesgo crediticio en microcréditos, especialmente en entornos donde la información crediticia es limitada o no está disponible. Es importante tener en cuenta los factores adicionales que se utilizan para evaluar el riesgo crediticio y cómo se integran en el proceso de toma de decisiones de préstamos.

2.7 Viabilidad de otorgamiento de créditos

La viabilidad del otorgamiento de un crédito va siempre ligada a: capacidad de pago del deudor, moralidad comercial, solvencia, garantías y calidad de información brindada. Esto representa una dependencia importante en variables cuantitativas que permiten soportar un modelo de credit score para el análisis del otorgamiento de un préstamo, se presentan avances enfocados en correlación de variables cualitativas como: ubicación demográfica del cliente, estrato, género, nivel de estudios, etc. con variables cuantitativas como lo es el nivel de ingreso, este estudio ayudó a agilizar la respuesta de otorgamiento de créditos de consumo ordinario al momento de que el algoritmo daba peso a las variables previamente estudiadas (Peña et al., 2011).

Se presentan otro tipo de estudios donde se analiza la asignación de puntajes crediticios aleatorios para lograr evaluar el impacto, utilizando premisas de asignación de micro préstamos a corto plazo (no mayores a un año) utilizando un software de calificación crediticia para de forma aleatoria afectar la decisión de aprobación de créditos enfocados en mejorar la prestación de intermediación financiera. Arrojó que los resultados llevaron a determinar una conducta más

sólida, menos negocios y un menor bienestar subjetivo. Estos resultados no avalan si el microcrédito funciona o no efectivamente (Karlan & Zinman, 2021).

2.8 Riesgo financiero de crédito

El riesgo financiero hace referencia a las situaciones que pueden generar una disminución en el valor de un activo determinado, situaciones ampliamente catalogadas por Duffie y Singleton (2003) como: riesgo de mercado, riesgo liquidez, riesgo operacional y riesgo de crédito; este último de importancia para el presente estudio. Como lo definen los autores mencionados, dicha clasificación hace referencia a la probabilidad asociada al incumplimiento de una persona jurídica o natural en el pago de una obligación adquirida previamente asociada a diversos factores como disminución en su capacidad de pago y condiciones no previstas.

El cálculo de este riesgo se ha profundizado a través del tiempo por medio de modelos *credit scoring* cada vez más sofisticados, según Puertas y Marti (2011) y afirman Crook et al. (2017) que el *credit scoring* es todo sistema automático que permite asignar un puntaje de clasificación para aprobar o negar una solicitud en función de la capacidad de pago analizada. El aporte de esta herramienta ha sido sumamente relevante para decidir aquellas solicitudes que por sus cualidades superan los requisitos mínimos de crédito, pero no cumplen de manera holgada las capacidades de pago entrando en una zona gris, que tradicionalmente es revisada por un comité o analista para concluir su proceso.

Ahora bien, tradicionalmente estos modelos utilizan técnicas de regresión ordinarias que se ajustan debido a la información disponible (ingreso salarial, consulta en burós crediticios y otras variables cualitativas ya mencionadas) los cuales son inflexibles al momento que se limitan las variables de entrada. Estos modelos pueden ser afectados por multicolinealidad o sobre ajustes en sus *inputs* al tener un desbalance en la información como ocurre en los estudios de crédito en

población con ingreso indeterminado, siendo más pertinentes modelos basados en *machine learning* (Yu, 2012).

2.9 *AI* y *Machine learning*

La inteligencia artificial (AI por sus siglas en inglés) es un término que se ha utilizado ampliamente para describir sistemas capaces de replicar el nivel cognitivo del humano, en su primera aplicación en 1956, John McCarthy menciona el termino de *AI* para referirse a la creación de robots con funcionalidades humanas y en la actualidad se acuña el término para referirse a algoritmos y metodologías basadas en las formas en las que el cerebro humano resuelve problemas permitiendo llegar a soluciones complejas (Saeed & Ondracek, 2012).

Por ello ha tomado fuerza la investigación de modelos de *scoring* basados en *machine learning*; Ampountolas, et al. (2021) recopilaron en su investigación titulada *"A machine learning approach for micro-credit scoring"*. Las más recientes investigaciones enfocadas en modelos de *credit score* las cuales catalogadas como modelos de caja negra, se caracterizan por la flexibilidad de sus modelos al no tener en cuenta relaciones de correlación ni comportamiento de los datos para realizar las proyecciones logrando un porcentaje de acierto predictivo más eficiente que el modelo tradicional con variables cualitativas.

3. Objetivos

3.1 Objetivo principal

Estimar un modelo de *credit score* alternativo para colocaciones de crédito de bajo monto utilizando herramientas de *machine learning* a partir de variables categóricas sin contemplar los ingresos.

3.2 Objetivos específicos

• Identificar y seleccionar las variables representativas con base en la información disponible de superintendencia financiera de Colombia para desarrollar el modelo *credit score*.

• Construir modelo usando el algoritmo *RandomForest* para determinar la PI.

• Comparar el modelo de *RandomForest* construido, contra el modelo de regresión GLM de *credit score* y modelo *support vector machine.*

• Correr los modelos elaborados con simulaciones de préstamo con monto y tasa fija sobre capital limitado (*ceteris paribus*) para determinar la viabilidad de las colocaciones y su rotación.

4. Metodología

El desarrollo de modelos *machine learning* para determinar capacidades de pago o riesgos de incumplimiento asociados a créditos de consumo no es algo nuevo, de hecho, existen numerosos estudios donde la relevancia de variables como el ingreso han sido determinantes para el éxito del modelo. Sin embargo, debido a la escasa información verificable que existe sobre la población no bancarizada y su importancia en los modelos tradicionales se planteó una opción que sustituya dicha variable por medio de la creación de un factor calculado con base a datos categóricos, bajo un desarrollo empírico.

Para efectos académicos el modelo a desarrollar tuvo como insumo la información disponible del más reciente informe de Demanda de Inclusión Financiera realizada por Bancoldex - Banca de las Oportunidades en Colombia (Pérez, 2022). Este informe cuya base de datos es abierta fue desarrollada con las siguientes metodologías:

Metodología de muestreo no probabilístico por cuotas: con el objetivo de garantizar la representatividad de la muestra. Se establecieron cuotas basadas en el perfil sociodemográfico de los participantes para asegurar una muestra diversa y representativa de la población, se encuestó 5.500 personas; la mayor de las muestras de la historia de la encuesta que realiza la entidad.

Metodología de recolección de datos: las encuestas fueron realizadas utilizando diferentes métodos de recolección de datos, como entrevistas telefónicas, encuestas autoadministradas en línea y encuestas cara a cara. La elección del método de recolección de datos varió según la población objetivo y los objetivos específicos de cada encuesta.

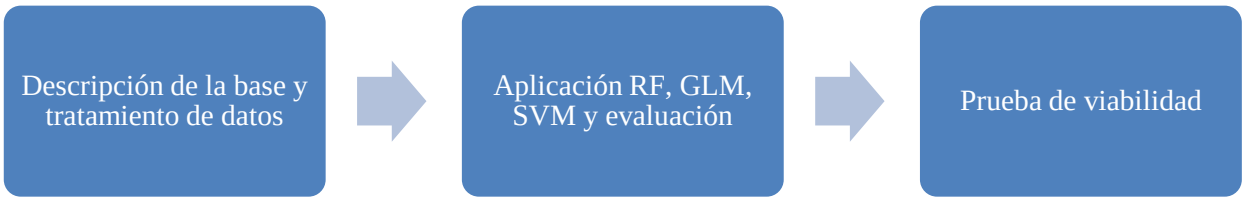
Instrumentos utilizados para la medición: los cuestionarios de las encuestas incluyeron preguntas sobre el perfil socio-demográfico de los participantes, su experiencia con los servicios financieros y su percepción sobre los factores relevantes para un sistema de puntuación crediticia

alternativo en Colombia. Los cuestionarios también incluyeron escalas de medición para evaluar la importancia de diferentes variables en la evaluación del riesgo crediticio y la disposición de los participantes a utilizar un sistema de puntuación crediticia alternativo.

De lo anterior se aseguró tener una base confiable para el desarrollo académico, con lo que el enfoque metodológico corresponde a tres etapas para la construcción del *credit score*.

Figura 2

Desarrollo del modelo



Nota. Elaboración propia, 2023

Pasando el tratamiento de datos se procedió a identificar el funcionamiento y utilidad de los algoritmos de inteligencia artificial *Generalized Linear Model-GLM*, *Support Vector Machine-SVM* y *Random Forest-RF*, para su comparación y evaluación con la finalidad de comprender sus similitudes y diferencias.

GLM (*Generalized Linear Model*). Representa un enfoque estadístico que generaliza el modelo lineal clásico al permitir diferentes distribuciones de probabilidad y funciones de enlace. Es utilizado con el objetivo de modelar la relación entre una variable de respuesta y una o más variables predictoras en diversas disciplinas científicas (Nelder & Wedderburn, 1972).

La fórmula general de un GLM es:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \quad \text{(Fórmula 1)}$$

Donde:

E(Y) es la media de la variable de respuesta Y.

β_0, β_1, β_2, ..., β_p son los coeficientes del modelo que se estiman a partir de los datos.

X_1, X_2, ..., X_p son las variables predictoras.

El GLM permite realizar predicciones utilizando la fórmula del modelo, donde la función de enlace convierte la media de la variable de respuesta en una combinación lineal de las variables predictoras y los coeficientes del modelo (McCullagh & Nelder, 1983; Dobson & Barnett, 2018).

Random Forest. Es un algoritmo de aprendizaje automático que es utilizado tanto para problemas de clasificación como de regresión. La técnica es basada en árboles de decisión que combina múltiples árboles y utiliza el promedio de sus predicciones para obtener resultados más precisos y estables (Breiman, 2001).

A diferencia de un solo árbol de decisión, que puede ser propenso a sobreajustar los datos de entrenamiento, Random Forest utiliza un enfoque de conjunto (ensemble) donde se crean múltiples árboles independientes y se promedian sus resultados. Esto se conoce como el método de *bagging (bootstrap aggregating)*. Cada árbol se entrena con una muestra aleatoria de los datos de entrenamiento y utiliza solo un subconjunto aleatorio de las características (variables predictoras) disponibles (Liaw & Wiener, 2002).

La fórmula general para la predicción en un *Random Forest* no se representa de manera explícita como en los modelos lineales o los GLM. En lugar de eso, se basa en la votación o promedio de las predicciones de los árboles individuales.

Según El funcionamiento básico de un Random Forest se puede resumir en los siguientes pasos:

1. Seleccionar un número definido de árboles a construir en el bosque.

2. Para cada árbol:

• Tomar una muestra aleatoria con reemplazo de los datos de entrenamiento.

• Tomar un subconjunto aleatorio de características (variables predictoras).

• Construir el árbol de decisión utilizando la muestra y las características seleccionadas.

3. Para realizar una predicción:

• Para un problema de clasificación, se realiza una votación entre los árboles para determinar la clase más común.

• Para un problema de regresión, se promedian las predicciones de los árboles para obtener un valor numérico (Hastie et al, 2009).

Support Vector Machine (SVM). Este algoritmo de aprendizaje automático es utilizado tanto para problemas de clasificación como de regresión. Es un método que busca encontrar el hiperplano óptimo que mejor separa las muestras de diferentes clases en un espacio de alta dimensión. La idea principal detrás de SVM es encontrar un hiperplano que maximice la separación entre las clases. Un hiperplano es una superficie de decisión que divide el espacio en regiones que se asignan a diferentes clases. Para problemas de clasificación linealmente separables, SVM busca el hiperplano que tiene la mayor distancia (margen) a los puntos de ambas clases (Cortes & Vapnik, 1995; Vapnik, 2000).

La fórmula general del hiperplano en SVM es:

$f(x) = w \cdot x + b$ (Fórmula 2)

Donde:

f(x) es la función de decisión que asigna un punto x a una clase.

w es el vector de pesos que define la orientación y dirección del hiperplano.

x es el vector de características (variables predictoras).

b es el sesgo (bias) o término de intersección que permite mover el hiperplano fuera del origen.

En SVM, la tarea consiste en encontrar el hiperplano que maximice la separación entre las clases, lo que se traduce en resolver un problema de optimización. La función de decisión f(x) asignará la clase de acuerdo con el lado del hiperplano en el que caiga el punto x (Hastie, et al, 2008).

En la tabla 1 se relacionan las similitudes y diferencias entre los tres modelos, concluyendo que los tres son utilizados para construir modelos predictivos utilizando la base de regresión , siendo esta indexada a cualidades especificas de cada uno para lograr modelar escenarios complejos que permitan identificar relaciones entre variables cualitativas y/o cuantitativas.

Tabla 1

Similitudes y diferencias de los modelos

Similitudes / Diferencias	GLM	RandomForest	SVM
Construir modelos predictivos a partir de datos de entrenamiento	Si	Si	Si
Buscan aprender patrones y relaciones en los datos para realizar predicciones precisas sobre nuevos datos	Si	Si	Si
Enfocados en problemas de regresión	Si	Si	Si
Enfocados en problemas de clasificación binaria	Si	No	Si
Enfocados en problemas de multiclase	No	No	Si
Enfoque basado en modelos lineales generalizados que asume una distribución de probabilidad y una función de enlace específicas para el problema	Si	No	No
Ensamblado de árboles de decisión que utiliza promedios o votaciones para realizar predicciones	No	Si	No
Maximización de márgenes que busca encontrar un hiperplano óptimo de separación	No	No	Si

Nota. Elaboración propia, 2023

5. Resultados

A continuación, se presenta los resultados de acuerdo con las tres fases establecidas en la metodología (figura 2). Cabe resaltar que las fases están asociadas a los objetivos específicos, el primer objetivo que es identificar y seleccionar las variables representativas con base en la información disponible de superintendencia financiera de Colombia para desarrollar el modelo *credit score*, comprende la primera fase del modelo (descripción y tratamiento de datos); los objetivos 2 y 3 se refieren a construir el modelo usando el algoritmo *RandomForest* para determinar la PI y comparar el modelo de *RandomForest* construido, contra el modelo de regresión GLM de *credit score,* y modelo *support vector machine,* se cumplen en la segunda fase que se denomina creación y evaluación; finalmente el cuarto objetivo que es correr los modelos elaborados con simulaciones de préstamo con monto y tasa fija sobre capital limitado (*ceteris paribus*) para determinar la viabilidad de las colocaciones y su rotación, se cumple en la tercera fase que es la prueba de viabilidad.

5.1 Fase 1. Descripción de la base y tratamiento de datos

La base sobre la cual se desarrolló el presente estudio corresponde al resultado de la encuesta de demanda de inclusión financiera publicada por la Banca de las Oportunidades (2022) en colaboración con la Superintendencia Financiera de Colombia y el Banco de la República en su tercera toma. Los resultados son producto de un trabajo de campo realizado entre el 5 de abril del 2022 y el 27 de mayo del 2022 bajo la modalidad presencial dirigida por el Centro Nacional de Consultoría, la cual se aplicó a personas de nacionalidad colombiana con un rango de edad entre 18 y 70 años, donde el tiempo promedio de cada encuesta fue de 60 minutos. Como se detalla en la presentación de los resultados de la encuesta:

> Se encuestaron a más de 5.500 adultos colombianos ubicados a lo largo del territorio nacional, muestra significativamente mayor a la observada en las mediciones anteriores (cercana a los 1.400 adultos). Los resultados encontrados tienen representatividad estadística por sexo, grupos etarios, nivel educativo, ingresos, regiones y nivel de ruralidad (Banca de Oportunidades, 2022)

Es de resaltar que dada la fecha en que se originan los datos de la base, estos incluyen los efectos desarrollados tras los efectos de la pandemia (Covid19) por lo cual la información carece de algún sesgo originado por dicho evento.

La base consta de 5.610 registros que respondieron las 72 preguntas del cuestionario, estas respuestas fueron codificadas en 432 opciones, de acuerdo al diccionario creado por la entidad para mejorar su interpretación (Banca de Oportunidades, 2022), estos datos fueron superiores a la muestra objetiva de 5.513 establecida en la ficha técnica de la encuesta, la cual se encuentra en el siguiente link:

https://www.bancadelasoportunidades.gov.co/es/publicaciones/encuestas-de-demanda

Estos resultados le permiten obtener un margen de error del 1,8% con un nivel de confianza del 95%.

Cabe resaltar que, dada la estructura de la encuesta, la gran mayoría de preguntas son de múltiple respuesta, por lo cual se observan en su base 432 opciones, de las cuales no todas tienen registro, dejando campos en vacío, por lo que se requirió de realizar un tratamiento a base.

Para el tratamiento de los datos se utilizó el software R, en donde se realizó un pre-filtro de las preguntas que se presumen tienen coherencia para inferir la probabilidad de incumplimiento, por lo cual se creó una base inicial “Datos” la cual incluyó 37 preguntas con sus múltiples opciones de respuesta que para efectos del modelo se consideran variables con un total de 185. Sobre estas

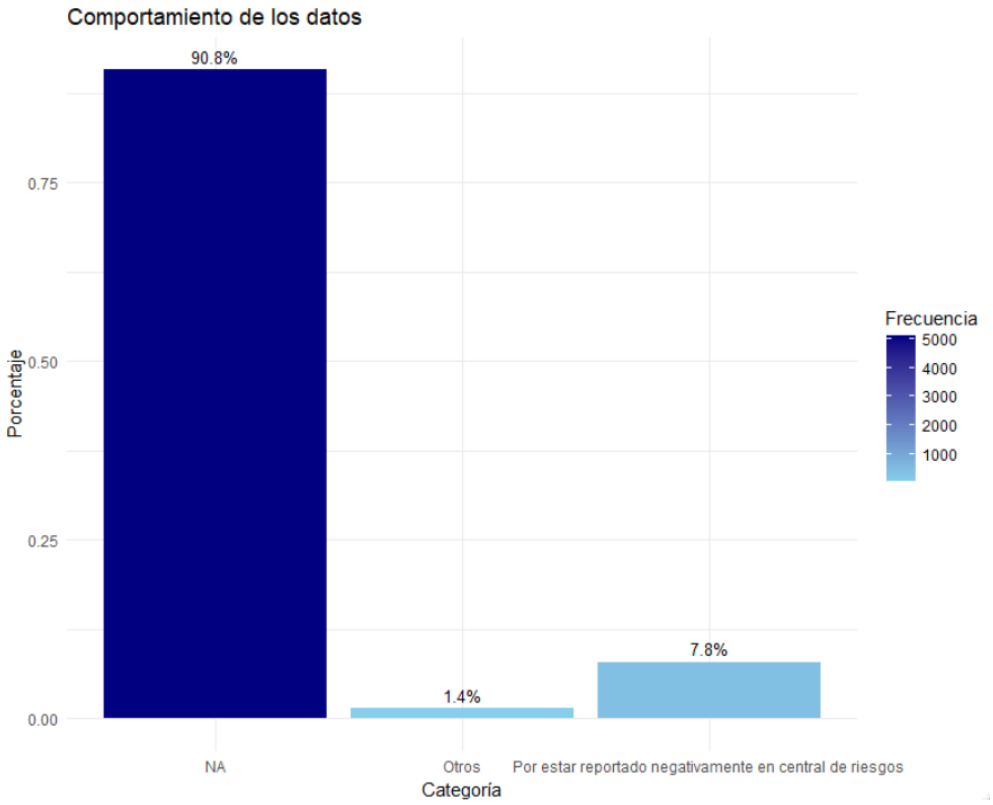
se hizo el conteo de los campos vacíos, encontrando registros con más de 5.000 campos vacíos sobre los cuales se realizó la interpretación, encontrando que son variables que dependen de respuestas anteriores, así el campo vacío se interpreta con relación de la respuesta de su dependencia, por lo cual se normalizan rellenando los campos en vacío con la lógica que aplique según la variable.

Como resultado se modificaron las variables asociadas a las preguntas P103, T110, P203, P210, T307. Así mismo, fue necesario consolidar las preguntas que formarían la variable a predecir, tomando como elección las variables asociadas a la pregunta P306 (¿Por qué le negaron el (los) crédito(s) que solicitó?) la cual indagaba sobre el motivo asociado en diferentes productos que solicitó el individuo tales como tarjeta de crédito, crédito libre inversión, crédito vivienda, entre otros.

Consolidando estos resultados en una variable nueva titulada P306, la cual hace referencia a comportamientos negativos en calidad crediticia sobre personas que solicitaron créditos y les fue negado; razón por la cual fue necesario complementar la información con las personas que no solicitaron créditos y presentaban los mismos comportamientos, esto se complementó con la variable P302 (¿Cuál(es) es(son) la(s) razón(es) por la(s) que no solicitó crédito en el último año?) de la cual se obtuvieron respuestas como: “estoy reportado en centrales de riesgo”, “estoy sobre endeudado”, “falta de garantías”, “ingresos insuficientes”. Al unificar los resultados de las variables P302 y P306 en una nueva titulada Variable Y se evidenció que cerca del 90,5% de los encuestados no comentan tener causales de negación o reportes negativos en el sector financiero dejando un 9,5% de los encuestados como personas con comportamientos de difícil recaudo con los que se establece el modelo de estudio (figura3). Lo anterior se resumió en 535 registros que cumplen el criterio de incumplimiento y 5.075 que no los cumplen.

Figura 3

Causal de negación o no consumo de productos de crédito

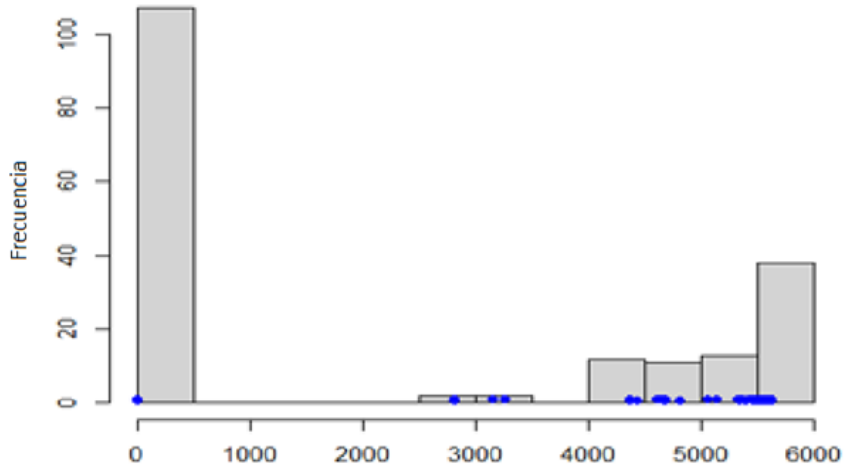


Nota. Elaboración propia en R, 2023.

De lo anterior se procedió a realizar el conteo de variables que continúan con campos vacíos encontrando 78 variables que siguen teniendo campos vacíos y que gran parte de estas variables tienen más de 3.000 campos vacíos, al profundizar este comportamiento se encontró que corresponden a variables complementarias de la encuesta como “otros” o variables con muy pocos campos diligenciados por lo cual se excluyen del presente análisis.

Figura 4

Comportamiento de las variables según campos vacíos



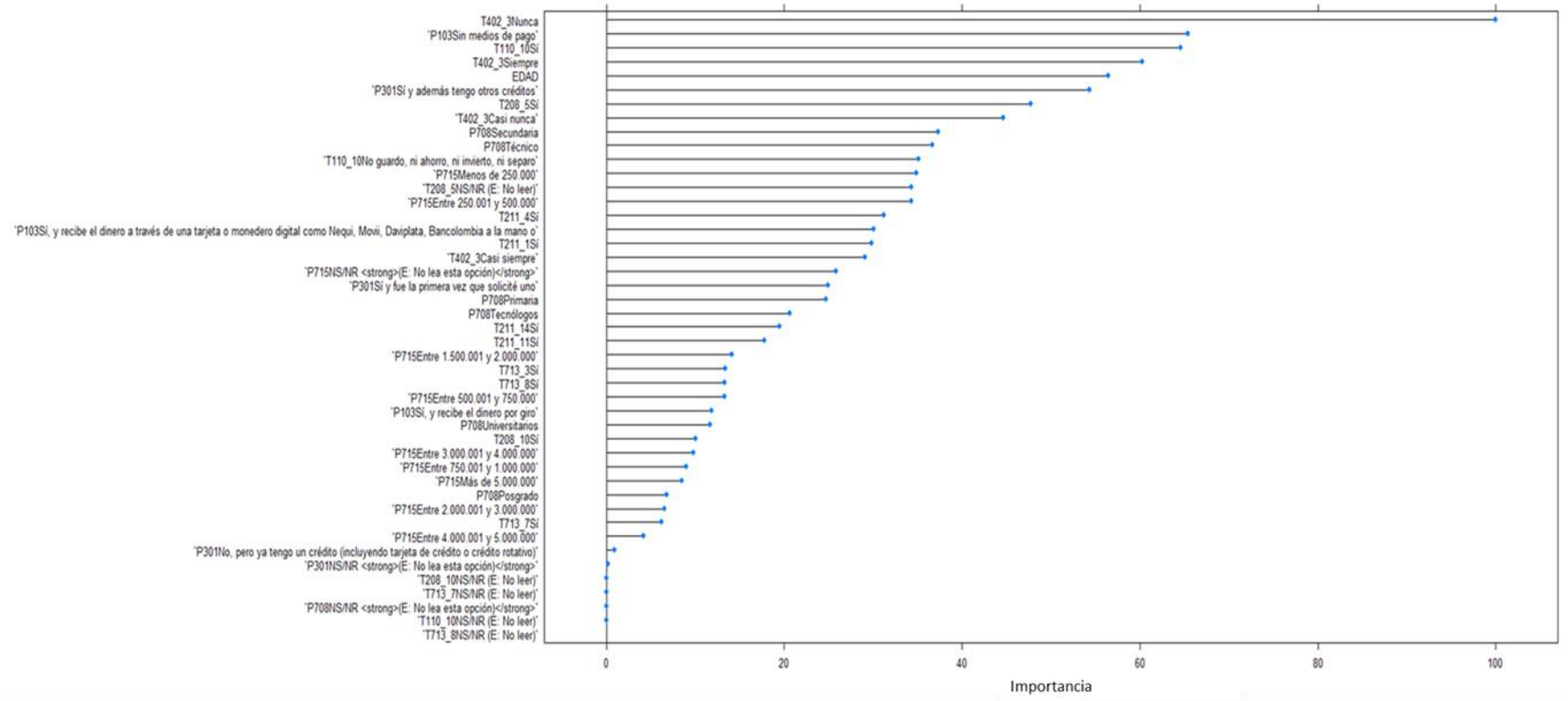
Nota. Elaboración propia en R, 2023

Con esto se llega a una base de 105 variables con la cual se empieza el análisis. Dado que las variables a tratar son de carácter cualitativo para comprender la correlación de estos con la variable a proyectar Y se realizó una prueba Chi Cuadrado donde se guardaron los valores P en el arreglo valoresp, sobre las variables que tienen correlación se realizó un análisis bivariado entre cada variable seleccionada para evitar sobre ajustes en un modelo tradicional de GLM.

Producto de este análisis se llegan a 43 variables que incluyen la variable Y. Sobre esta base se corre el modelo inicial haciendo uso de la librería *Caret en R* la cual permite correr modelos GLM, SVM y *RandomForest*. Con esto se obtiene el primer acercamiento a la importancia de las variables seleccionadas sobre el modelo, por lo cual haciendo uso de la variable importancia se establecen como relevantes para el modelo todas aquellas que superen el 80%. Resultado de dicho proceso se obtiene como relevante las variables "T402_3", "EDAD", "T110_10", "P301", "P103", "P301", "P709", "T402_3”,"T405_102", T405_105","T402_3", "T307_6", "Variable Y".

Figura 5

Importancia con la variable Y dentro del modelo



Nota. Elaboración propia en R, 2023

Los porcentajes de importancia con respecto a la variable Y “posibilidad de incumplimiento”, demuestra que tan correlacionada está la variable dentro de la ecuación lineal que se genera en el modelo para su funcionamiento.

Con estas se procede a un análisis sobre los resultados de cada variable y su interpretación para posterior uso en los modelos que siguen.

5.2 Fase 2. Creación y evaluación

En esta fase se construyó el modelo usando el algoritmo *Random Forest* para determinar la PI y se comparó el modelo de *Random Forest* construido, contra el modelo de regresión GLM de *credit score*, y modelo *support vector machine.*

Se detalló primero una breve explicación de que significa y como funciona cada modelo que se procedió a comparar:

5.2.1 *Generalized Linear Model* (GLM)

Representa un enfoque estadístico que generaliza el modelo lineal clásico al permitir diferentes distribuciones de probabilidad y funciones de enlace. Es utilizado con el objetivo de modelar la relación entre una variable de respuesta y una o más variables predictoras en diversas disciplinas científicas (Nelder & Wedderburn, 1972).

5.2.2 RandomForest

Es un algoritmo de aprendizaje automático que es utilizado tanto para problemas de clasificación como de regresión. La técnica es basada en árboles de decisión que combina múltiples árboles y utiliza el promedio de sus predicciones para obtener resultados más precisos y estables (Breiman, 2001).

5.2.3 *Support Vector Machine* (SVM)

Este algoritmo de aprendizaje automático es utilizado tanto para problemas de clasificación como de regresión. Es un método que busca encontrar el hiperplano óptimo que mejor separa las muestras de diferentes clases en un espacio de alta dimensión. Un hiperplano es una superficie de decisión que divide el espacio en regiones que se asignan a diferentes clases. Para problemas de clasificación linealmente separables, SVM busca el hiperplano que tiene la mayor distancia (margen) a los puntos de ambas clases (Cortes & Vapnik, 1995; Vapnik, 2000).

5.2.4 Correlaciones lineales del modelo GLM en R

Estas son las correlaciones de la ecuación lineal generada, para entender cómo afecta las respuestas a las preguntas dentro del modelo para calcular un resultado de posibilidad de incumplimiento de una persona según sus respuestas. Los hallazgos encontrados en esta parte principalmente fueron generados con respecto a la pregunta T402_3 , la habla sobre con qué frecuencia se retrasa en el cumplimiento de sus pagos comprometidos, al momento de que las personas respondieron que nunca se atrasan, el modelo las castiga con mayor probabilidad de incumplimiento, esto debido a que revisando toda la base de datos, se logra identificar que las misma personas que respondían que nunca se atrasaban en sus pagos, estaban reportadas en centrales de riesgo, por tal motivo el modelo identifica esa correlación como anti intuitiva y ese es el motivo que al responder positivamente reduce en la ecuación la posibilidad de incumplimiento.

El otro hallazgo importante se encontró en la pregunta P715, la cual trata sobre el nivel de gasto mensual en diferentes tipos de escalas, se logró identificar que no hubo una correlación directamente proporcional según el aumento del gasto por cada escala con relación a que este factor ayuda a disminuir el nivel de incumplimiento, se presentó todo lo contrario, los rangos de gastos

que fueron menores a $750.000 COP por mes, el modelo incrementó la posibilidad de incumplimiento, así como los rangos de gastos mayores de $3.000.000 COP también identificó que la posibilidad de incumplimiento es más alta; mientras que los gastos que están en un rango mayor a 750.000 COP y menores a 3.000.000 COP, se logró identificar que el modelo redujo la probabilidad de incumplimiento. Ver Anexo A.

Ahora bien, es de resaltar que dada la estructura de regresión que realizaron los algoritmos de IA de SVM y RF no fue posible obtener la ecuación con la que las variables se relacionan, debido a que trabajan como "cajas negras", razón por la cual se realizó el presente análisis con la variable *Accuracy* que se repite en los modelos seleccionados.

Tabla 2

Accuracy de los modelos

	GLM	**SVM**	**RF**
Accuracy	0.9037432	0.9076651	0.5944784

Nota. Elaboración Propia

Estos resultados confirmaron que es posible desarrollar un modelo de regresión, que excluya la variable de ingreso, sin generar grandes afectaciones en la eficiencia de los estos mismos. Lo anterior se validó con el desempeño favorable, superior al 90% obtenido, en dos de los tres modelos de algoritmos seleccionados en el análisis. Siendo el de mayor eficiencia el algoritmo de inteligencia artificial.

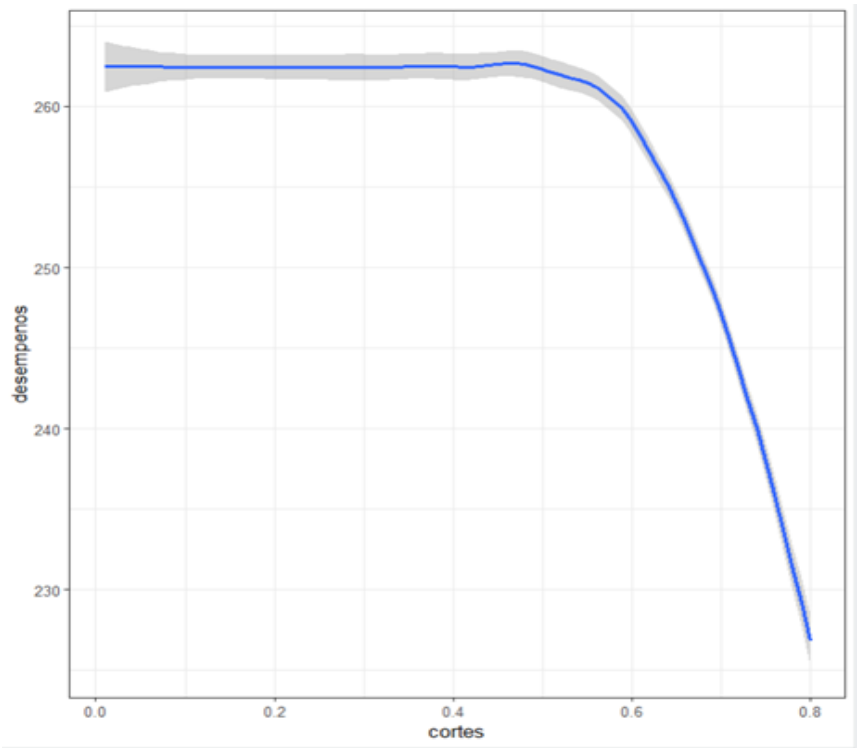
5.3 Fase 3. Prueba de viabilidad

En esta fase se cumplió el último objetivo del estudio que es correr los modelos elaborados con simulaciones de préstamo con monto y tasa fija sobre capital limitado (*ceteris paribus*) para determinar la viabilidad de las colocaciones y su rotación.

Simulación de préstamos aleatorios: se crea una base aleatoria en *R* de cincuenta mil registros utilizando la función *oversampled* en el programa, con el fin de simular préstamos bajo la siguiente característica, préstamos de $50.000 COP por persona, con una tasa de interés del 44% e.a, a un plazo de préstamo de un mes, con el fin de analizar el desempeño del modelo creado, el cual arrojó los siguientes resultados.

Figura 6

Corte óptimo del modelo de préstamo simulado



Nota. Elaboración propia en R, 2023

Se realizó un corte óptimo del modelo del 48% como rango de probabilidad de incumplimiento y analizamos los resultados del modelo.

Tabla 3

Resultados de la población según el corte optimo del modelo

Viabilidad	Cumple	Incumple
False	154	130
True	3.917	45.799

Nota. Elaboración propia en R

Se logró identificar que el modelo identifica factible prestar a 45.799 personas (91,5%) con una probabilidad de cumplimiento asertiva, 3.9017 personas (7,8%) el modelo recomienda no prestarles porque no cumplirían con su obligación del crédito, 130 (0,26%) personas que el modelo recomendaría no prestar las cuales si pudieran llegar a cumplir con la obligación financiera y por último 154 (0,30%) personas a las cuales se les prestaría, pero incumplirían con la devolución del préstamo. Si estos datos numéricos se llevan a valores monetarios, tendría el siguiente resultado.

Tabla 4

Resultados de la población según el corte optimo del modelo valorizado

Viabilidad	Cumple	Incumple
False capital	-$ 7.700	
False int.	$ -	
True capital		$ 2.289.950
True int.		$ 70.653

Nota. Elaboración propia

Se pudo determinar que los intereses ganados de las personas que cumplirán con la obligación financiera correspondientes a los $70MCOP cubrirán la pérdida de las personas que

simplemente no devolverán el capital prestado en este caso $7,7MCOP, por tal motivo se logró identificar un beneficio bruto de $62,3MCOP para esta simulación.

6. Conclusiones

Como conclusiones se encontró que durante la construcción del modelo de GLM dentro del tratamiento de datos se identificó un patrón que a primera vista que no es intuitivo sobre la variable T402_3 de la cual se indaga la frecuencia con la que ocurren determinadas situaciones; para el tercer literal se preguntó con qué periodicidad la persona se encuentra atrasada con algunos de los pagos comprometidos obteniendo como posibles respuestas: siempre, casi siempre, a veces, casi nunca y nunca. El análisis de correlación e importancia de los datos y la variable Y indicó que la respuesta “nunca” tiene una correlación positiva en el riesgo de incumplimiento con una importancia del 100%, la cual se verificó sobre los datos evidenciando que estas mismas personas se encontraban reportadas en centrales de riesgo o les fue negado una solicitud por sobre endeudamiento, demostrando la validez de la correlación identificada. En menor grado de importancia (44,65%) se ubicó la respuesta “casi nunca” con correlación positiva.

Al ejecutar los modelos GML, SVM y RF incluidos en la librería Caret, se evidencia que los modelos de mejor desempeño fueron los GML y SVM donde ambos llegaron a una *Accuracy* cercano al 90% la cual fue superior a la obtenida en el modelo generado por el *RandomForest*, donde esta última no superó el 60% de asertividad. Es de resaltar que este resultado se obtiene excluyendo de las variables predictoras el nivel de ingresos de los encuestados con los cual se identifica que dicho factor no es excluyente para el estudio de préstamos, con lo que se permite abrir a futuras investigaciones definir hasta qué punto el ingreso es relevante para proyectar incumplimientos sobre el monto solicitado.

Adicionalmente se puede determinar que esta iniciativa podría llegar a tener un potencial significativo según los resultados arrojados del préstamo aleatorio, donde se logra identificar que

las ganancias que se tendrían por el préstamo del capital, llamado intereses lograran cubrir las pérdidas relacionadas a las personas que no cumplan con su obligación financiera.

Referencias

Ampountolas, A., Nyarko, T., Date, P., & Constantinescu, C. (2021). A Machine Learning Approach for Micro-Credit Scoring. *Risks., 9*(3), 1-20. doi:10.3390/risks9030050

ASOBANCARIA. (19 de Septiembre de 2022). Uso de información alternativa para fortalecer los modelos de scoring. *Banca & Economia*(1346). https://www.asobancaria.com/wp-content/uploads/2022/09/1346_BE.pdf

Banca de Oportunidades. (2022). *Encuesta de demanda de inclusión financiera.* https://www.bancadelasoportunidades.gov.co/es/publicaciones/encuestas-de-demanda

Banco de España. (2017). *Informe de estabilidad financiera.* https://www.bde.es/f/webbde/Secciones/Publicaciones/InformesBoletinesRevistas/InformesEstabilidadFinancera/17/ficheros/IEFMayo2017.pdf

Banco Mundial. (29 de Marzo de 2022). *La inclusión financiera es un elemento facilitador clave para reducir la pobreza y promover la prosperidad.* https://www.bancomundial.org/es/topic/financialinclusion/overview

Breiman, L. (2001). Random Forests. *Machine learning*(43), 5-32. doi:10.1023/A:1010933404324

Castro, A. (2018). *Marco constitucional sobre la economía popular solidaria y el sector financiero popular solidario en, Economía popular y solidaria: ¿realidad o utopía? Caracterización de las entidades de fomento.* Quito, Ecuador: Editorial Universitaria Abya-Yala. doi:10.7476/9789978104903

Chaparro, A. M. (2021). *Fintech, una apuesta de la tecnología para la inclusión financiera en Colombia. Alianza EFI economía formal e inclusiva.* https://alianzaefi.com/wp-content/uploads/2023/01/WP3-2021-003.pdf

Colombia Fintech. (10 de Enero de 2023). *Demanda de inclusión Financiera.* https://colombiafintech.co/lineaDeTiempo/articulo/encuesta-de-demanda-de-inclusion-financiera

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine learning*(20), 273-297. doi:10.1007/bf00994018

Crook, J., Edelman, D., & Thomas, L. (2017). *Credit Scoring and its Applications. The University of Edinburgh.* https://www.research.ed.ac.uk/en/publications/credit-scoring-and-its-applications-4

Demirgüç, A., Klapper, L., Singer, D., Ansar, S., & Hess, J. (2018). *The Global Findex Database 2017 Measuring Financial Inclusion an the Fintech Revolution.* World Bank Group, Washington D. C., Estados Unidos. http://hdl.handle.net/10986/29510

Demirgue, A., Klapper, L., Singer, D., & Van Oudheusden, P. (2014). *Global Findex Database 2014: Measuring Financial Inclusion around the World.* https://thedocs.worldbank.org/en/doc/681361466184854434-0050022016/original/2014GlobalFindexReportDKSV.pdf

Didier, T., & Schmukler, S. (2014). *Emerging Issues in Financial Development Lessons from Latin America.* The World Bank: https://openknowledge.worldbank.org/server/api/core/bitstreams/ad9d5f87-8616-5a4d-a60b-df3bf0be93f1/content

Dobson, A. J., & Barnett, A. (2018). *An Introduction to Generalized Linear Models* (Cuarta ed.). London: Chapman and Hall.

Duffie, D., & Singleton, K. (2003). *Credit Risk: Pricing, Measurement, and Management.* Princeton: Princeton University Press. doi:10.1515/9781400829170

Gamba, S., Pacheco, D., & Yaruro, A. (2016). *Informe especial de inclusión financiera. Informes especiales de estabilidad financiera.* Bogotá: Banco de la Republica. https://www.banrep.gov.co/sites/default/files/publicaciones/archivos/iepref_sep_5_201

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning.* New York, NY: Publisher Name Springer. doi:10.1007/b94608_15

Jappelli, J., & Pagano, M. (1994). Saving, Growth, and Liquidity Constraints. *The Quarterly Journal of Economics, 109*(1), 83-109. doi:10.2307/2118429

Karlan, D., & Zinman, J. (2021). Microcredit in Theory and Practice: Using Randomized Credit Scoring for Impact Evaluation. *Science, 332*(6035), 1278-1284. doi:10.1126/science.1200138

Khandani, A. E., Kim, A., & Lo, A. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767-2787. 10.1016/j.jbankfin.2010.06.001

Kiva. (2013). *Anual Report.* kiva.org: https://www.kiva.org/cms/kiva_annual_report_2013_0.pdf

Laibson, D. (1997). Golden Eggs and Hyperbolic Discounting. *The Quarterly Journal of Economics, 112*(2), 443-478. 10.1162/003355397555253

Liaw, A., & Wiener, M. (December de 2002). Classification and Regression by Random Forest. *R News, 2/3*(18).

Maisueche, A. (2019). *Utilización del Machine Learning en la Industria 4.0.* Tesis de Maestría, Universidad de Valladolid de España, Máster en Ingeniería Industrial. Escuela de Ingenierías Industriales Universidad de Valladolid, Valladolid, España. https://core.ac.uk/download/pdf/228074134.pdf

McCullagh, P., & Nelder, J. A. (1983). *Generalized Linear Models* (Segunda ed.). London: Chapman and Hall.

Ministerio de Hacienda y Crédito Público. (2019). *Artículo 53 de la Ley 1955 de 2019 - Plan Nacional de Desarrollo 2018-2022.* https://www.minhacienda.gov.co/webcenter/portal/AtencionPublico/pages_atencinalciudadano/sentenciasconciliaciones/art53ley1955pnd2018-2022

Ministerio de Vivienda, Ciudad y Territorio. (17 de Agosto de 2021). *Plan Nacional de Desarrollo 2018-2022.* https://www.minvivienda.gov.co/ministerio/planeacion-gestion-y-control/planeacion-y-seguimiento/plan-nacional-de-desarrollo-2018-2022#:~:text=El%20Plan%20Nacional%20de%20Desarrollo%202018%2D%202022%20es%20un%20pacto,construir%20el%20futuro%20de%20Colombia

National Consumer Law Center. (2019). *Credit Invisibility and Alternative Data: Promises and Perils.* https://www.nclc.org/wp-content/uploads/2022/08/ib-credit-invisib-alt-data-july19-1.pdf

Nelder, J. A., & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, 135*(3), 370-384. doi:10.2307/2344614

Peña, A., Lochmüller, C., Murillo, J., Pérez, M., & Vélez, C. (2011). Modelo cualitativo para la asignación de créditos de consumo y ordinario. El caso de una cooperativa de crédito. *Revista Ingenierías, 10*(19), 101-111. http://www.redalyc.org/articulo.oa?id=75022317009

Pérez, J. D. (2022). *Encuesta de demanda de inclusión financiera.* https://www.bancadelasoportunidades.gov.co/sites/default/files/2022-08/Formulario%20Encuesta%20de%20Demanda%202021_1.pdf

Puertas, R., & Marti, M. (2011). Análisis del Credit Scoring. *RAE, 53*(3), 303-315. https://riunet.upv.es/bitstream/handle/10251/59864/Credit%20Scoring%20RAE%20%28303-315%29.pdf?sequence=2&isAllowed=y

Saeed, M., & Ondracek, J. (2012). *Doing business in india: international perspectives (with particular reference to business process outsourcing (BOP) industry*. India: Excel India Publishers New Delhi.

SEON. (Julio de 2022). *Social Media Credit Scoring: Pros, Cons, and How to Do It.* https://seon.io/resources/social-media-credit-scoring/

Stiglitz, J. E., & Weiss, A. (Jun. de 1981). Credit Rationing in Markets with Imperfect Information. *The American Economic Review, 71*(3), 393-410. https://www.jstor.org/stable/1802787

Superintendencia Financiera de Colombia-SFC. (2018). *Reporte de inclusión financiera 2018.* https://imgcdn.larepublica.co/cms/2019/06/20164130/Banca_RIF2018_FINAL.pdf

Superintendencia Financiera de Colombia-SFC. (2021). *Reporte de inclusión financiera 2021*. https://www.bancadelasoportunidades.gov.co/sites/default/files/2022-07/Reporte%20de%20inclusi%C3%B3n%20financiera%202021.pdf

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory.* Springer. https://statisticalsupportandresearch.files.wordpress.com/2017/05/vladimir-vapnik-the-nature-of-statistical-learning-springer-2010.pdf

Yu, B. (2012). Analysis of a Random Forests Model. *Journal of Machine Learning Research*(13), 1064-1095. https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

Anexos

Anexo A. Correlación e importancia de ecuación lineal modelación R modelo GLM

Código de pregunta & respuesta	Importancia vs variable Y	Resultado	Analisis
`T402_3Casi nunca`	44.65	0.46491073	Rta. hace sentido de que haya más posibilidad de incumplimiento, debido a que, en la base de datos a pesar de responder casi nunca en retraso de pago, estaban reportados en centrales de riesgo.
`T402_3Casi siempre`	29.06	-0.29420885	Rta. que hace sentido con respecto a su respuesta, reduce posibilidad de incumplimiento.
T402_3Nunca	100.00	0.83625236	Rta. hace sentido de que haya más posibilidad de incumplimiento, debido a que, en la base de datos a pesar de responder casi nunca en retraso de pago, estaban reportados en centrales de riesgo.
T402_3Siempre	60.24	-0.57135339	Rta. que hace sentido con respecto a su respuesta, reduce posibilidad de incumplimiento, a pesar de estar atrasado, no estan reportados en centrales de riesgo.
EDAD	56.40	-0.01330859	Rta. que hace sentido en la correlacion a mayor edad menos posibilidad de incumplimiento.
`T110_10No guardo, ni ahorro, ni invierto, ni separo`	35.09	-0.37586688	Rta. que hace sentido, genera menos posibilidad de incumplimiento.
`T110_10NS/NR (E: No leer) `	-	16,89891255	Rta. que se identifica como mayor posibilidad de incumplimiento no responder la pregunta.
T110_10Sí	64.57	-0.72585532	Rta. que hace sentido.
`P103Sí, y recibe el dinero a través de una tarjeta o monedero digital como Nequi, Móvil, Daviplata, Bancolombia a la mano o`	30.06	-0.36655188	Rta. que hace sentido por ser beneficiario de subsidio, existe menos posibilidad de incumplimiento.
`P103Sí, y recibe el dinero por giro`	-	-0.16570160	Rta. que hace sentido por ser beneficiario de subsidio, existe menos posibilidad de incumplimiento.
`P103Sin medios de pago`	65.36	0.51171878	Rta. con sentido, por no tener ningún tipo de producto electrónico, presenta una probabilidad de incumplimiento mayor.
`P301No, pero ya tengo un crédito (incluyendo tarjeta de crédito o crédito rotativo) `	-	30,32761336	Rta. con sentido, po tener un crédito ya existente, aumenta su posibilidad de incumplimiento.

Código de pregunta & respuesta	Importancia vs variable Y	Resultado	Analisis
`P301NS/NR <strong> (E: No lea esta opción) </strong>`	-	17,33543215	Rta. con sentido por no responder la pregunta, aumenta la posibilidad de incumplimiento.
`P301Sí y además tengo otros créditos`	54.28	0.73387258	Rta. con sentido, por tener varios créditos existentes, aumenta su posibilidad de incumplimiento.
`P301Sí y fue la primera vez que solicité uno`	24.94	0.26975347	Rta. con sentido, por ser su primer crédito.
`T402_3.1Casi nunca`	-	NA	
`T402_3.1Casi siempre`	-	NA	
T402_3.1Nunca	-	NA	
T402_3.1Siempre	-	NA	
`T402_3.2Casi nunca`	-	NA	
`T402_3.2Casi siempre`	-	NA	
T402_3.2Nunca	-	NA	
T402_3.2Siempre	-	NA	
`T208_5NS/NR (E: No leer) `	34.30	-2,84536008	Rta. por no responder la pregunta, referente a ingresos por honorarios la posibilidad de incumplimiento disminuye.
T208_5Sí	47.69	-0.40089148	Rta. por recibir pensión la posibilidad de incumplimiento disminuye.
`T208_10NS/NR (E: No leer) `	-	16,91474525	Rta. con sentido, por no responder si recibe un subsidio del gobierno su posibilidad de incumplimiento incrementa.
T208_10Sí	-	-0.10106048	Rta. con sentido, por responder si recibe un subsidio del gobierno su posibilidad de incumplimiento disminuye.
T211_1Sí	29.81	-0.19067822	Rta. con sentido, por pagar el arriendo la posibilidad de incumplimiento disminuye.
T211_4Sí	31.22	-0.20636776	Rta. con sentido, por pagar minutos a celular la posibilidad de incumplimiento disminuye.
T211_11Sí	-	0.19913459	Rta. con sentido, por pagar productos financieros la posibilidad de incumplimiento disminuye.
T211_14Sí	-	0.15047182	Rta. con sentido, por pagar impuestos la posibilidad de incumplimiento disminuye.
`T713_7NS/NR (E: No leer) `	-	-32,90294	Rta. con sentido, por no responder la pregunta de tener computador en casa, mayor la posibilidad de incumplimiento.

Código de pregunta & respuesta	Importancia vs variable Y	Resultado	Analisis
T713_7Sí	-	0.04726634	Rta. con sentido, por tener computador en casa, menor la posibilidad de incumplimiento.
`T713_8NS/NR (E: No leer) `	-	16,75417356	Rta. con sentido, por no responder la pregunta de tener computador en casa, menor la posibilidad de incumplimiento.
T713_8Sí	-	0.16167993	Rta. con sentido, por tener carro, mayor la posibilidad de incumplimiento.
T713_3Sí	-	-0.15630856	Rta. con sentido, por tener un celular inteligente, menor la posibilidad de incumplimiento.
`P715Entre 1.500.001 y 2.000.000`	-	-0.17700301	Rta. con sentido, tener gastos sobre ese rango presenta una posibilidad de incumplimiento menor.
`P715Entre 2.000.001 y 3.000.000`	-	0.11927738	Rta. con sentido, tener gastos elevados sobre ese rango presenta una posibilidad de incumplimiento mayor.
`P715Entre 250.001 y 500.000`	34.29	0.37819543	Rta. con sentido, tener gastos tan bajos sobre ese rango presenta una posibilidad de incumplimiento mayor.
`P715Entre 3.000.001 y 4.000.000`	-	0.31379486	Rta. con sentido, tener gastos elevados sobre ese rango presenta una posibilidad de incumplimiento mayor.
`P715Entre 4.000.001 y 5.000.000`	-	-0.17062558	Rta. con sentido, tener gastos sobre ese rango presenta una posibilidad de incumplimiento menor.
`P715Entre 500.001 y 750.000`	-	0.13455641	Rta. con sentido, tener gastos tan bajos sobre ese rango presenta una posibilidad de incumplimiento mayor.
`P715Entre 750.001 y 1.000.000`	-	-0.08759925	Rta. con sentido, tener gastos sobre ese rango presenta una posibilidad de incumplimiento menor.
`P715Más de 5.000.000`	-	0.41566833	Rta. con sentido, tener gastos elevados sobre ese rango presenta una posibilidad de incumplimiento mayor.
`P715Menos de 250.000`	34.87	0.49452441	Rta. con sentido, tener gastos tan bajos sobre ese rango presenta una posibilidad de incumplimiento mayor.
`P715NS/NR <strong> (E: No lea esta opción) </strong>`	25.81	0.63840370	Rta. con sentido no responder la pregunta genera posibilidad de incumplimiento mayor.

Código de pregunta & respuesta	Importancia vs variable Y	Resultado	Analisis
`P708NS/NR <strong> (E: No lea esta opción) </strong>`	-	16,10412348	Rta. con sentido no responder la pregunta genera posibilidad de incumplimiento mayor.
P708Posgrado	-	-0.29769901	Rta. con sentido tener un grado de estudio presenta un nivel de incumplimiento menor.
P708Primaria	-	-0.68721015	Rta. con sentido tener un grado de estudio presenta un nivel de incumplimiento menor.
P708Secundaria	37.29	-1,02719353	Rta. con sentido tener un grado de estudio presenta un nivel de incumplimiento menor.
P708Técnico	36.63	-1,05165704	Rta. con sentido tener un grado de estudio presenta un nivel de incumplimiento menor.
P708Tecnólogos	-	-0.64552232	Rta. con sentido tener un grado de estudio presenta un nivel de incumplimiento menor.
P708Universitarios	-	-0.35593406	Rta. con sentido tener un grado de estudio presenta un nivel de incumplimiento menor.