
**UNIVERSIDAD POLITÉCNICA SALESIANA**

**SEDE GUAYAQUIL**

**CARRERA DE COMPUTACIÓN**

# ANÁLISIS DE SESGO DE MODELOS DE APRENDIZAJE AUTOMÁTICO EN LA PREDICCIÓN DE RIESGO DE CRÉDITO EN LA BANCA EN EL ECUADOR

Trabajo de titulación previo a la obtención del
Título de Ingeniero en Ciencias de la Computación

AUTORES: ANGIE LISSETTE GÓMEZ GUTIÉRREZ

EDUARDO IVAN SOSA BRACCO

TUTOR: ING. GALO ENRIQUE VALVERDE LANDIVAR

Guayaquil – Ecuador

2024

**CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN**

Nosotros, Angie Lissette Gómez Gutiérrez con documento de identificación N° 0955493051 y Eduardo Ivan Sosa Bracco con documento de identificación N° 0929010700; manifestamos que:

Somos los autores y responsables del presente trabajo; y, autorizamos a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Guayaquil, 04 de febrero del año 2024

Atentamente,

Angie Lissette Gómez Gutiérrez
0955493051

Eduardo Ivan Sosa Bracco
0929010700

**CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA**

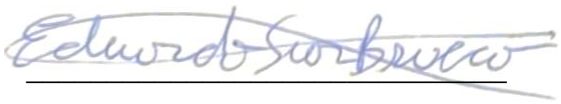
Nosotros, Angie Lissette Gómez Gutiérrez con documento de identificación N° 0955493051 y Eduardo Ivan Sosa Bracco con documento de identificación N° 0929010700, expresamos nuestra voluntad y por medio del presente documento cedemos a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que somos autores del Artículo Académico: "Análisis de sesgo de modelos de Aprendizaje Automático en la predicción de riesgo de crédito en la banca en el Ecuador", el cual ha sido desarrollado para optar por el título de: Ingeniero en Ciencias de la Computación, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribimos este documento en el momento que hacemos la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Guayaquil, 04 de febrero del año 2024

Atentamente,

Angie Lissette Gómez Gutiérrez
0955493051



Eduardo Ivan Sosa Bracco
0929010700

## CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Galo Enrique Valverde Landivar con documento de identificación N° 0912511532, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: Análisis de sesgo de modelos de Aprendizaje Automático en la predicción de riesgo de crédito en la banca en el Ecuador, realizado por Angie Lissette Gómez Gutiérrez con documento de identificación N° 0955493051 y Eduardo Ivan Sosa Bracco con documento de identificación N° 0929010700, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Guayaquil, 04 de febrero del año 2024

Atentamente,

Ing. Galo Enrique Valverde Landivar
0912511532

## DEDICATORIA

Con mucho aprecio quiero dedicar este artículo de titulación primeramente a Dios, por darme la sabiduría y salud para realizar este proyecto. A mi mamá Juana Gómez, que es la persona más valiosa para mí por la confianza y amor que me transmitió mediante su apoyo incondicional durante todos mis años de estudio. En cada tropiezo y desafío que enfrenté, sus palabras de aliento fueron mi motivación para que pudiera seguir adelante. Ha sido mi mayor defensora y mi guía, animándome a perseguir mis sueños y a superar los obstáculos que se presentaban. A mi abuelita Cristina Gutiérrez, a quien quiero expresarle mi profundo agradecimiento por su amor hacia mí. Su presencia constante y sus oraciones han sido una bendición en cada paso que he dado. A mi ángel guardián, mi papá Félix Gómez por guiarme desde el cielo en cada logro de mi vida. A mis tías, Gladys, Erika, Magali y Regina, cada una de ustedes ha dejado una marca única en mi vida. Han sido modelos a seguir, mostrándome el significado de la fortaleza y determinación para realizar las cosas. A mis tíos, Félix, Julio y Marcos por que mediante sus logros y experiencias me han inspirado a alcanzar mis propias metas y a nuca rendirme frente a los desafíos. A mis primos, Andrés y Brianna por llenarme de alegrías y amor a lo largo de estos años. A mi compañero de proyecto, Eduardo por ser un gran amigo y apoyo durante estos años. Y a mi novio Ebert, por ser mi compañero y mi apoyo incondicional en esta hermosa etapa universitaria, que cuando no creía en mí misma, siempre estaba ahí para recordarme cuán capaz y valiosa soy. A medida que finalizo mi artículo científico, no puedo dejar de reflexionar sobre el arduo camino que he recorrido y la importancia que han tenido en él. Sin su presencia constante, sin su aliento incondicional, sin su fe en mí, no hubiera sido posible llegar hasta aquí. Me siento verdaderamente agradecida y bendecida de contar con personas tan maravillosas en mi vida.

**Angie Lissette Gómez Gutiérrez.**

## DEDICATORIA

Este artículo lo dedico principalmente a Dios y a mis padres por el apoyo en este paso hacia un mundo profesional lleno de retos y metas , también le dedico esto a mis tíos maternos que me apoyaron a lo largo de mi preparación en esta carrera, también le dedico esto a mi compañera y amiga de carrera Angie Gómez la cual demostró ser una chica inteligente, increíble y sobre todo responsable también esto va principalmente dedicado a mi persona, él cual siempre logra demostrar que todo es posible y que todo lo que yo me propongo lo puedo cumplir sin importar lo difícil o complicado que pueda llegar a ser.

**Eduardo Ivan Sosa Bracco**

## AGRADECIMIENTO

Querida Universidad Politécnica Salesiana, ingenieros de la carrera de Ciencias de la Computación y nuestro apreciado tutor, el Ing. Galo Valverde. Nos dirigimos a ustedes con profundo agradecimiento y gratitud por haber sido parte de este articulo científico. Su apoyo, orientación y dedicación han sido fundamentales para nuestro crecimiento y éxito. A la Universidad Politécnica Salesiana, por brindarnos un buen entorno académico gracias a la calidad de educación y los recursos disponibles, hemos podido desarrollar nuestras habilidades y conocimientos en nuestra área de estudio A nuestros profesores durante estos cuatro años de estudio por su dedicación, compromiso en la impartición de conocimiento y su pasión por enseñar. En especial, queremos agradecer a nuestro tutor, Ing. Galo Valverde, por su orientación constante y su valioso asesoramiento a lo largo de este proceso. Su experiencia y sabiduría han sido fundamentales para la estructuración y el desarrollo de nuestra investigación.

**Angie Lissette Gómez Gutiérrez.**

**Eduardo Iván Sosa Bracco**

## RESUMEN

Evaluar el sesgo en los modelos de aprendizaje automático diseñados para predecir el riesgo crediticio en la banca es fundamental para garantizar decisiones justas y transparentes. El propósito de este análisis es investigar el potencial de sesgo en los algoritmos utilizados para evaluar la confiabilidad y sensibilidad de los datos necesarios en el proceso de calificación crediticia.

Uno de los aspectos fundamentales cubiertos en este estudio es la necesidad de corregir los errores identificados y tomar acciones específicas para reducir la presencia de datos erróneos en los modelos de pronóstico. Esto incluye la introducción de métricas de capital al evaluar el desempeño del modelo para mejorar la transparencia y la rendición de cuentas en el uso de algoritmos en las decisiones crediticias.

El análisis detallado tiene como objetivo identificar posibles fuentes de sesgo, como conjuntos de datos desequilibrados o variables discriminatorias, que podrían afectar la precisión y confiabilidad de los pronósticos. Tener datos desequilibrados puede generar resultados sesgados porque el modelo puede favorecer a ciertos grupos sobre otros debido a la falta de representación equitativa en el conjunto de datos de entrenamiento.

Para lograr el objetivo de una predicción de resultados correcta y precisa, se propone implementar técnicas de reducción de sesgos, como la ponderación de muestras o la incorporación de características de equidad en el proceso de capacitación del modelo. Los modelos también deben monitorearse y actualizarse continuamente para garantizar la confiabilidad y precisión a medida que cambian los conjuntos de datos y las condiciones ambientales.

**Palabras clave:** aprendizaje automático, análisis, riesgo de crédito, transparencia, decisiones crediticias.

## ABSTRACT

Assessing bias in machine learning models designed to predict credit risk in banking is critical to ensuring fair and transparent decisions. The purpose of this analysis is to investigate the potential for bias in the algorithms used to evaluate the reliability and sensitivity of the data needed in the credit rating process.

One of the fundamental aspects covered in this study is the need to correct identified errors and take specific actions to reduce the presence of erroneous data in forecast models. This includes the introduction of capital metrics when evaluating model performance to improve transparency and accountability in the use of algorithms in credit decisions.

The detailed analysis aims to identify potential sources of bias, such as unbalanced data sets or discriminating variables, that could affect the accuracy and reliability of the forecasts. Having unbalanced data can lead to biased results because the model may favor certain groups over others due to a lack of equal representation in the training data set.

It also emphasizes the importance of eliminating discriminatory variables that may introduce biases into the credit rating process. These variables can include sensitive information such as age, gender, or race, which if not used appropriately can negatively affect the model's decisions.

To achieve the goal of correct and accurate outcome prediction, it is proposed to implement bias reduction techniques, such as sample weighting or incorporating fairness characteristics in the model training process. Models must also be continually monitored and updated to ensure reliability and accuracy as data sets and environmental conditions change.

In summary, the article emphasizes the importance of analyzing and eliminating errors in machine learning-based prediction models for credit risk.


**Key words**: Bias, machine learning, credit risk, transparency, bias correction, credit decisions.

## ÍNDICE DE CONTENIDO



## ÍNDICE DE FIGURAS

# 1. INTRODUCCIÓN

El uso de los modelos de aprendizaje automático en la predicción de riesgo de crédito en la banca en el Ecuador causa preocupaciones referentes a la existencia de posibles sesgos en las decisiones crediticias. Debido a que podrían conducir a una exclusión injusta de ciertos clientes o grupos de clientes en el proceso de solicitud y otorgamiento de crédito, de manera que contribuye a mantener desigualdades socioeconómicas preexistentes.

Esta situación se produce debido a los sesgos en la predicción de crédito que se miden en diferentes factores, como lo son el nivel de desempleo y la falta de acceso a servicios básicos según (Gonzalo, 2019). Provocando de esta forma que se incremente el riesgo de incumplimiento de pagos por parte de los clientes y su vez creando un impacto negativo en la rentabilidad y solvencia de las entidades financieras, puesto que mediante la pérdida de crédito en estas áreas reduce los ingresos generados por intereses y tarifas, de tal forma que se ve afectado la rentabilidad de las instituciones al no poder recuperar los fondos prestados.

La presencia de préstamos incobrables o en mora puede incrementar el nivel de activos problemáticos, lo que afecta su capacidad de otorgar nuevos préstamos. A pesar de que, mediante el uso de un modelo de aprendizaje automático se puede mejorar significativamente la predicción para los casos en mora al considerar los factores relevantes y utilizar algoritmos avanzados para el análisis de datos.

Sin embargo, estos modelos pueden verse influenciados por sesgos inherentes a los datos utilizados en su entrenamiento, lo que puede resultar en discriminación o falta de equidad en el proceso de otorgamiento de créditos. Para (RUIZ, 2022) reducir el sesgo en estos modelos es crucial al momento de evitar la discriminación y para promover la equidad. Para lograrlo, se pueden implementar técnicas como el uso de datos más diversos, ajustes de decisiones y evaluación regular en términos de equidad. Es esencial abordar este desafío para garantizar que los modelos sean justos, imparciales y promuevan una inclusión financiera.

En la actualidad, las empresas se encuentran inmersas en un entorno empresarial altamente competitivo y en constante evolución. En este contexto, el acceso y la utilización efectiva de los datos se han convertido en un factor determinante para el éxito de las empresas, por eso las organizaciones capaces de aprovechar la gran cantidad de datos disponibles provenientes de sistemas de información, dispositivos electrónicos y redes sociales, tiene una clara ventaja

competitiva en el mercado, si reconocen que estos datos representan uno de sus activos más valiosos.

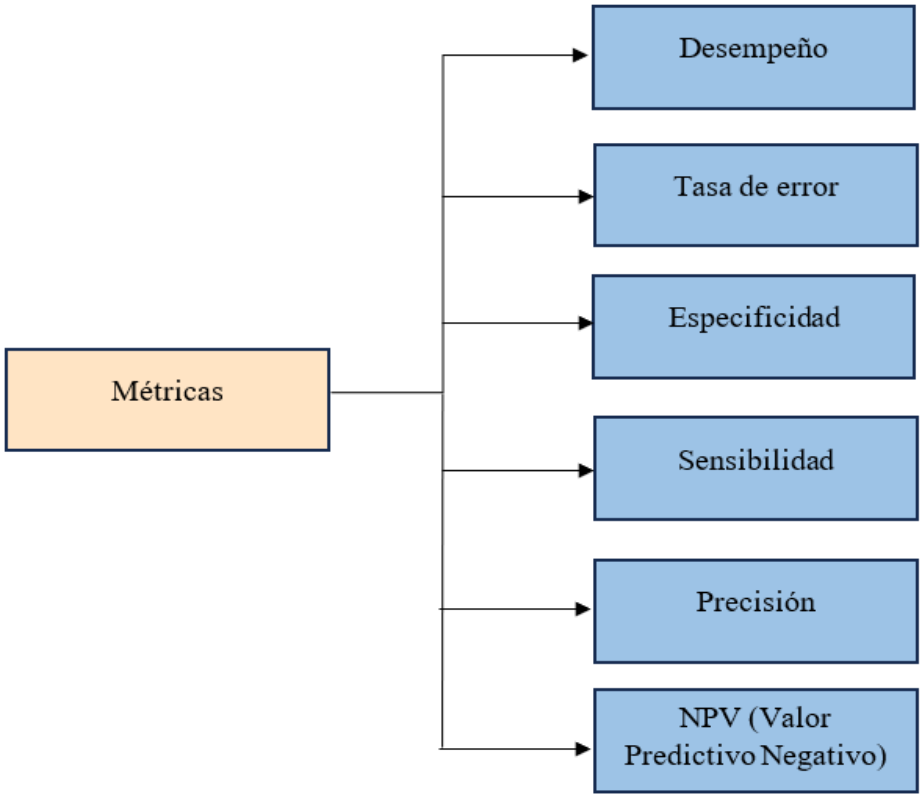
La gestión adecuada de estos datos se ha vuelto esencial para generar valor en las organizaciones y tomar decisiones estratégicas fundamentadas. En este sentido, han surgido herramientas y técnicas avanzadas de análisis de datos que permiten a las entidades financieras obtener información valiosa, identificar patrones y tendencias del mercado, y así anticiparse a las necesidades y preferencias de los clientes. Sin embargo, también es importante destacar que el manejo de grandes cantidades de datos conlleva varios desafíos. Debido a que las organizaciones deben asegurarse de que los datos se almacenen de manera segura y a su vez cumplan con las regulaciones de privacidad y protección de datos dispuestos por la ley. Además, deben garantizar la calidad e integridad de dichos datos para evitar decisiones basadas en información errónea o incompleta y así evitar afectar a los clientes o futuros clientes.

En este panorama, la banca se encuentra en una posición única, ya que maneja grades volúmenes de datos relacionados con clientes, transacciones financieras, historiales crediticios, el buró de crédito y otros aspectos relevantes para la toma de decisiones crediticias. Por esto la predicción del riesgo de crédito es una de las áreas más críticas en el sector financiero, y el uso de modelos de aprendizaje automático ha demostrado ser una herramienta efectiva para mejorar la precisión de dichas predicciones.

El aprendizaje automático, o machine learning, se refiere a la capacidad de los sistemas informáticos para aprender y mejorar automáticamente a partir de la experiencia de los datos. Al aplicar algoritmos de aprendizaje automáticos a conjuntos de datos de crédito, es posible desarrollar modelos predictivos que ayuden a evaluar la probabilidad de incumplimiento de los clientes, y reducir el número de clientes en mora.

En este artículo, la investigación se centrará en examinar las posibles implicaciones éticas y legales asociadas con los sesgos en los modelos de predicción de riesgo de crédito en la banca en el Ecuador. Además, se realizará una evaluación efectiva e interpretación de los modelos de predicción mediate la utilización de métricas relevantes presentadas en la figura 1.

*Figura 1*. Métricas para la predicción de riesgo de crédito.




*Nota*. Métricas relevantes para el algoritmo de predicción de riesgo de crédito. Fuente: Elaboración Propia.

El resultado de esta investigación busca proporcionar una visión más completa del impacto del sesgo en la predicción del riesgo de crédito en la banca ecuatoriana, permitiendo que las instituciones financieras tomen medidas correctivas y mejoren la equidad en sus procesos de otorgamiento de créditos.

Durante el periodo comprendido entre noviembre de 2023 y diciembre de 2023, se evaluaron las limitaciones abordadas en este estudio. La información recopilada para la base de datos fue obtenida en la ciudad de Guayaquil, abarcando diferentes zonas. Durante dicho periodo de estudio, se llevaron a cabo encuestas y entrevistas a una muestra aleatoria de personas que habían solicitado crédito, obteniendo un total de 114 encuestas. Además, se recopilaron variables socioeconómicas relevantes, como nivel de ingresos, historial crediticio, buró de crédito y ocupación de los individuos.

## 2. REVISIÓN DE LITERATURA

Para (Belén, 2020) las instituciones bancarias son consideradas actores clave en el sistema financiero en Ecuador, dado que desempeñan un papel fundamental en la economía al ser los principales proveedores de crédito en el país.

Existen diferentes tipos de entidades financieras, como los bancos privados, bancos públicos y las cooperativas de ahorro y crédito, entre otras. Estas instituciones ofrecen diferentes tipos de productos y servicios financieros para satisfacer las necesidades de diferentes grupos de personas y además se encuentran reguladas y supervisadas por la Superintendencia de Bancos, que es el responsable de garantizar la estabilidad y solidez del sistema financiero ecuatoriano. El objetivo del análisis es investigar el desempeñado de las instituciones bancarias ecuatorianas en el acceso de crédito, así como identificar los principales desafíos que enfrentan dichas instituciones.

Los créditos bancarios se clasifican en las siguientes categorías presentadas en la tabla 1.

**Tabla 1**. *Categorías de créditos bancarios*

| Calificación | Descripción |
|---|---|
| A-1 | Créditos que no presentan morosidad alguna y tienen un riesgo de pérdida esperada de 1%. |
| A-2 | Créditos que representan una morosidad de 1 a 15 días y tienen un riesgo de pérdida esperada de 2%. |
| A-3 | Créditos que representan una morosidad de 16 a 30 días y tienen un riesgo de pérdida esperada de 3 a 5%. |
| B | Créditos de riesgo potencial |
| C | Créditos deficientes. |
| D | Créditos de dudoso recaudo. |
| E | Pérdidas. |

*Nota.* La tabla muestra las diferentes categorías de créditos bancarios. Fuente: SB -Estructuras del Sistema de Operaciones Activas y Contingentes – SOAC.

Según (Noriega, 2023) la aplicación de modelos de aprendizaje automático en la predicción del riesgo de crédito ha ganado relevancia en el sector financiero debido a que se centra en entrenar a los modelos para que adquieran conocimiento a partir de los datos, en lugar de ser programadas de manera explícita para ello, debido a que son capacitados para identificar patrones y correlaciones en grandes data sets, con el objetivo de tomar decisiones optimas y realizar proyecciones basadas en los resultados. Además, también se destacan por su capacidad para mejorar la precisión y efectividad de las evaluaciones crediticias, y tienen un rendimiento significativo en comparación con métodos tradicionales como el análisis discriminante lineal.

Los algoritmos de clasificación, como las redes neuronales y los árboles de decisión, son muy precisos en la predicción de riesgo y la calificación de crédito, ya que ayudan a identificar grupos grandes de solicitantes de préstamos y ofrecer modelos de puntuación dinámicos basados en conglomerados para mejorar la precisión de la evaluación de riesgos. Al combinar estos métodos con los métodos tradicionales, es posible obtener una mayor capacidad de predicción y una mejor protección contra pérdidas crediticias.

Por ello, resulta importante que las entidades financieras reconozcan y consideren el impacto del sesgo en la evaluación del riesgo de crédito, debido a que si existe quiere decir que hay una inclinación desproporcionada a favor o en contra de un grupo en comparación a otro, al momento de realizar la evaluación de crédito.

En el ámbito de la ciencia e ingeniería, el sesgo se define como un error sistemático, que puede ocurrir cuando se realiza un muestreo injusto de una población o cuando un proceso de estimación produce resultados imprecisos en promedio. Dentro de los tipos más comunes de sesgos en los modelos está el sesgo de selección que se define como un error sistemático que surgen del procedimiento utilizado para seleccionar los sujetos y de los factores que afectan su participación en un estudio, esto produce una distorsión de los resultados y lleva a conclusiones incorrectas.

Por ejemplo, si un modelo es entrenado utilizando un conjunto de datos que únicamente incluye información de clientes de raza blanca, es probable que el modelo presente un sesgo en contra de los clientes de otras razas.

Asimismo, está el sesgo de representación que se refiere a la tendencia de los algoritmos y sistemas de búsqueda en línea, que al mostrar resultados sesgados o discriminatorios se basan

en características como la raza, el género o la clase social, lo que implica que los resultados de búsqueda pueden estar segados hacia ciertos grupos, mostrando una representación negativa de ciertos segmentos de la población

Por ejemplos, si un modelo utiliza una variable binaria que representar el género, es probable que presente un sesgo en contra de aquellas personas que se identifican como no binarias.

Por otro lado, también está el sesgo de aprendizaje que surge cuando un modelo no logra capturar de manera adecuada la verdadera relación subyacente entre las variables de entrada y las variables del objetivo que se busca predecir, esto ocurre cuando el modelo presenta limitaciones o suposiciones que afectan su capacidad para representar de manera precisa la complejidad del problema en cuestión.

Por ejemplo, si un modelo es entrenado utilizando un conjunto de datos que tiene una mayor proporción de préstamos otorgados a hombres que a mujeres, es probable que el modelo presente un sesgo en contra de las mujeres.

En consecuencia, evaluar y minimizar el sesgo en los modelos de predicción del riesgo crediticio se ha convertido en un tema esencial en el sector financiero, como señala (RUIZ, 2022). A medida que las instituciones financieras utilizan con mayor frecuencia modelos de aprendizaje automático para la toma de decisiones crediticias, aumentan las preocupaciones sobre posibles sesgos y resultados injustos o discriminatorios.

Debido a esto, la evaluación del sesgo se centra en identificar sesgos sistemáticos en los datos de entrenamiento y en los resultados, mientras que la mitigación de sesgos se centra en la creación de métodos y planes para minimizar o erradicar por completo los sesgos, para garantizar la equidad y la transparencia en la toma de decisiones crediticias, promoviendo así la confianza de los clientes, la integridad del sistema financiero y la protección de los derechos individuales.

## 3. METODOLOGÍA

Se utilizaron dos enfoques complementarios para obtener una base de datos amplia. En primer lugar, se procedió a realizar encuestas en la ciudad de Guayaquil con el fin de obtener datos reales de personas que solicitaron crédito. La elección de Guayaquil como ubicación se basó en su relevancia en el ámbito financiero del Ecuador y la disponibilidad de una muestra representativa de individuos.

Una vez recopilados los datos de las encuestas, se procedió a utilizar la base de datos real como punto de partida para generar datos ficticios adicionales. Estos datos ficticios se crearon de para simular la información histórica de instituciones financieras en Ecuador. La combinación de estos enfoques de recopilación de datos permitió obtener un dataset "datosCredito" más amplio y diverso.

Para determinar las variables más relevantes del modelo de evaluación de riesgo de crédito, se han identificado los siguientes campos, presentados en la tabla 2.

**Tabla 2.** *Variables de Evaluación de Riesgo*

| Campo | Detalle |
|---|---|
| Estado Civil | Estado civil de la persona |
| Nivel de Educación | Tipo de educación de la persona |
| Profesión | Profesión de la persona |
| Actividad Económica | Actividad económica de la persona |
| Tipo de Préstamo | Tipo de préstamo que está solicitando |
| Monto Solicitado | Monto que solicito la persona |
| Tasa de Interés | Tasa de interés según el tipo de préstamo |
| Edad | Edad de la persona |
| Nivel de ingresos | Ingresos de la persona |
| Cargas Familiares | Número de personas dependientes económicamente de la persona |
| Otros ingresos | Ingresos adicionales de la persona |
| Activos | Bienes y propiedades de la persona |
| Préstamos Instituciones | Prestamos o deudas con instituciones que tenga la persona |

| | |
|---|---|
| Riesgo | Si la persona ha tenido antecedentes de incumplimiento de pagos como cliente |

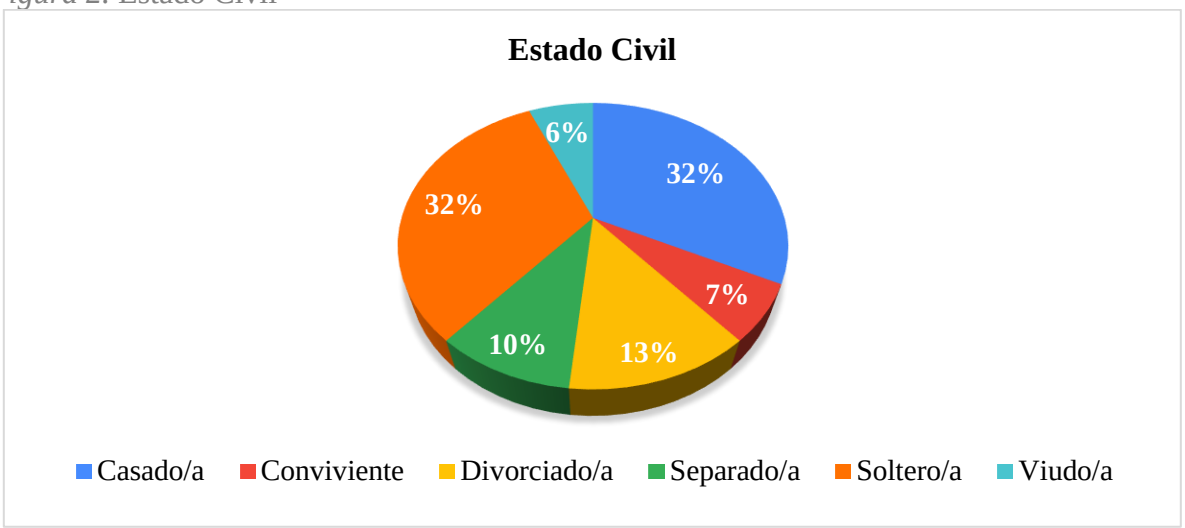
*Nota.* Contiene las variables fundamentales para la evaluación de riesgo de crédito. Fue recopilado de. Lcda. Merced Cuenca, Cela Gonzalo & Juan Cuenca (2019) "Propuesta de modelo de machine learning para la evaluación de crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito".

El campo "Riesgo" desempeña un papel fundamental en el modelo de evaluación de riesgo de crédito, debido a que actúa como la variable predictora principal. Se trata de una variable binaria que indica si el cliente presenta o no riesgo crediticio. Un valor de 0 indica que el cliente no presenta riesgo, mientras que un valor de 1 indica que el cliente tiene riesgo. La determinación de esta variable se basa en la clasificación de los clientes en buenos pagadores y malos pagadores.

La consideración de datos como el nivel de ingresos, ocupación, historial crediticio, cargas familiares y morosidad de los solicitantes de crédito permite realizar una evaluación de riesgo más completa y precisa. Estos datos se tienen en cuenta para evaluar las características individuales de cada cliente y ayudar a prevenir posibles sesgos en la evaluación crediticia.

En la figura 2 se ilustra el resultado de la encuesta en relación con la pregunta sobre el estado civil.

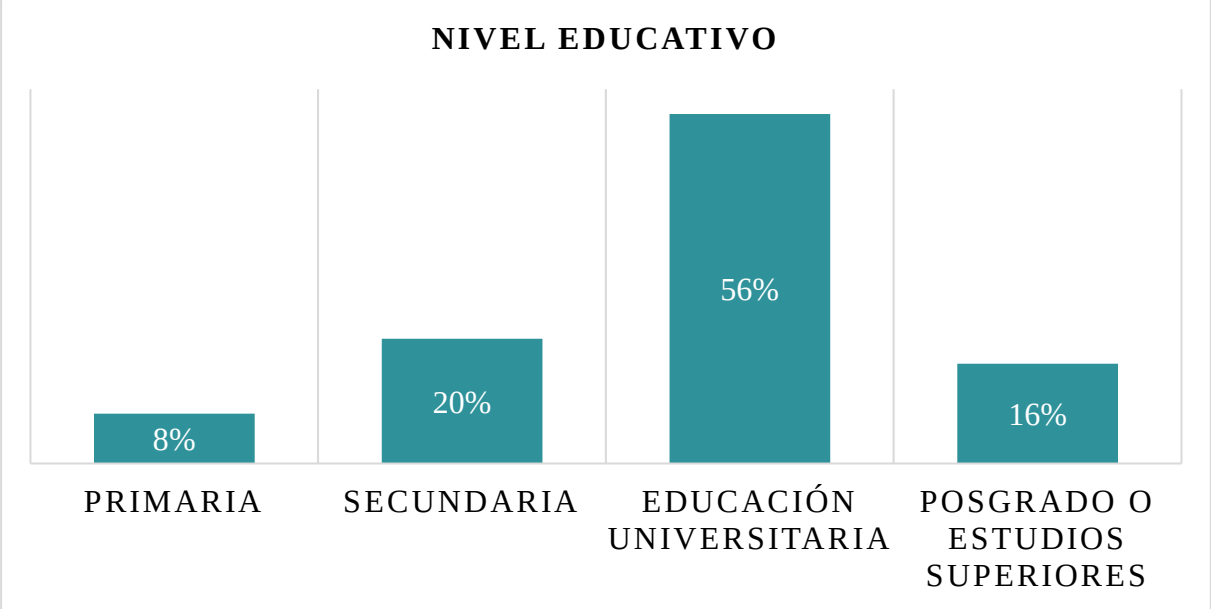
*Figura 2.* Estado Civil




*Nota.* Estadística de la encuesta sobre el Estado Civil. Fuente: Elaboración Propia.

La figura 3 presenta los resultados obtenidos en la encuesta respecto a la pregunta acerca del nivel educativo.

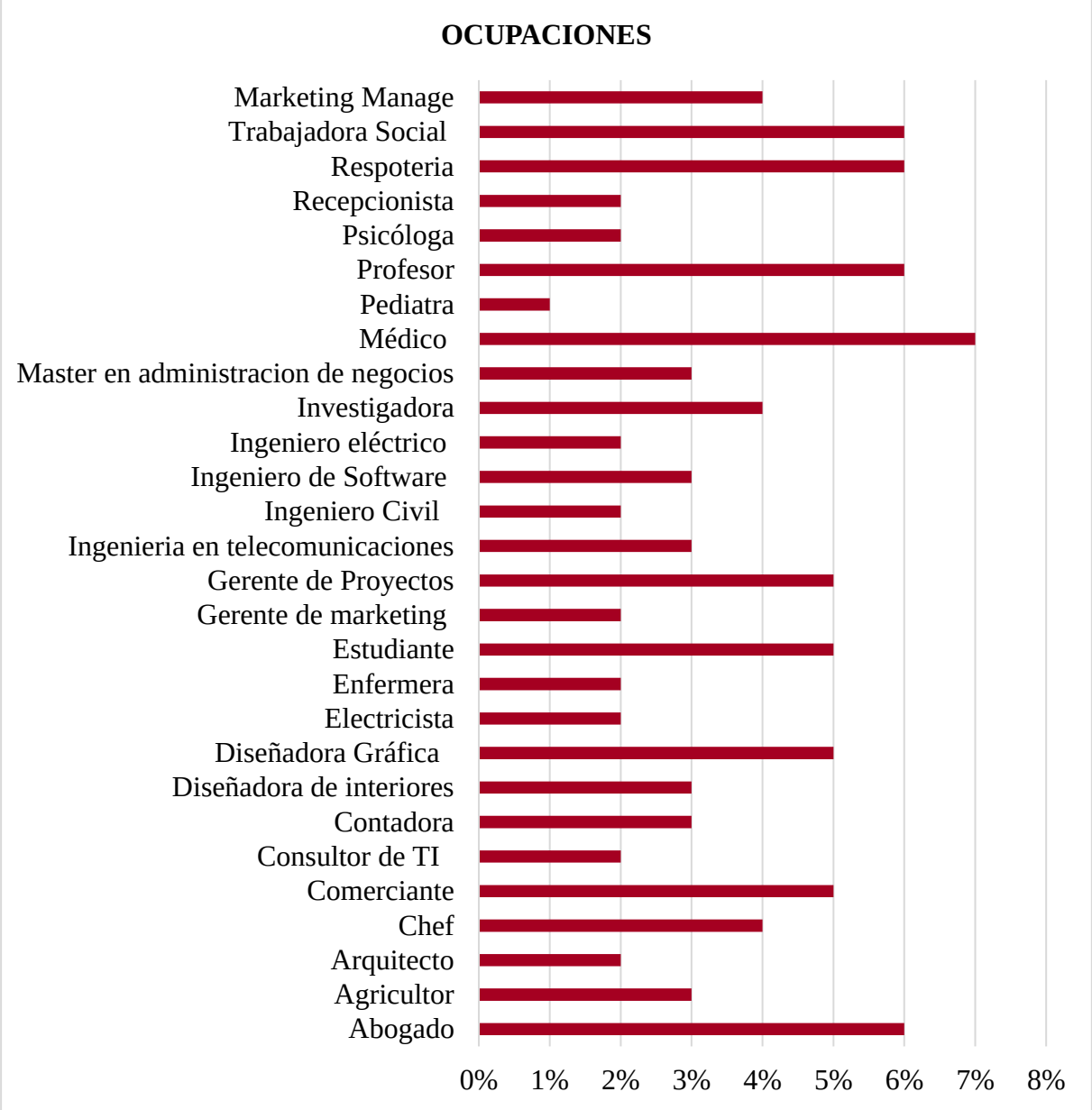
*Figura 3.* Nivel Educativo




*Nota.* Estadística de la encuesta sobre el nivel de estudios. Fuente: Elaboración Propia.

La figura 4 muestra el resultado de la encuesta en respuesta a la pregunta sobre la ocupación de los encuestados.

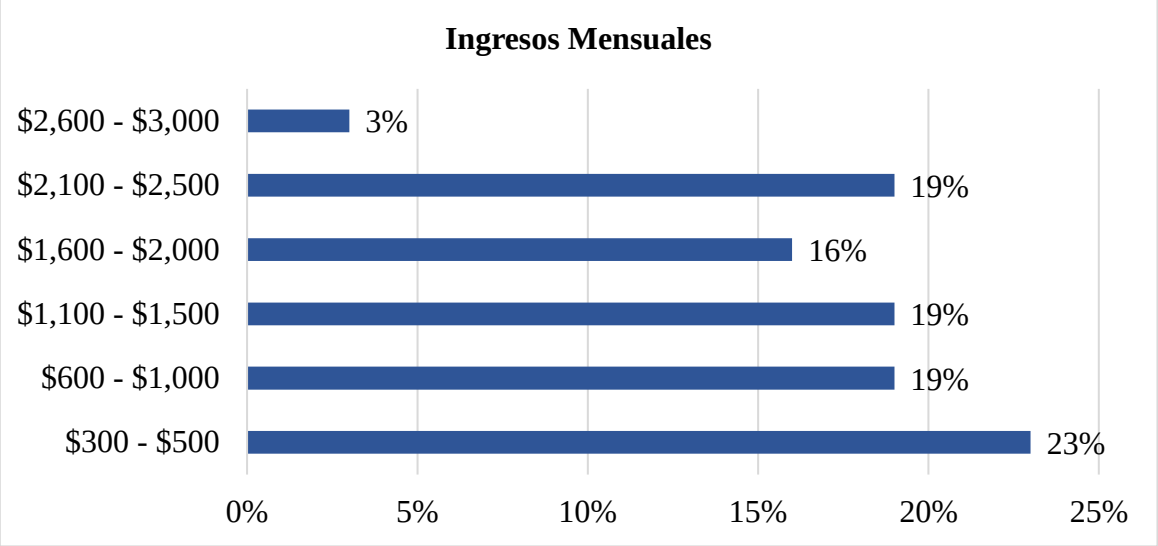
*Figura 4.* Ocupaciones




*Nota.* Estadística de la encuesta sobre las ocupaciones de los encuestados. Fuente: Elaboración Propia.

La figura 5 muestra el resultado en relación con la pregunta sobre los ingresos mensuales.

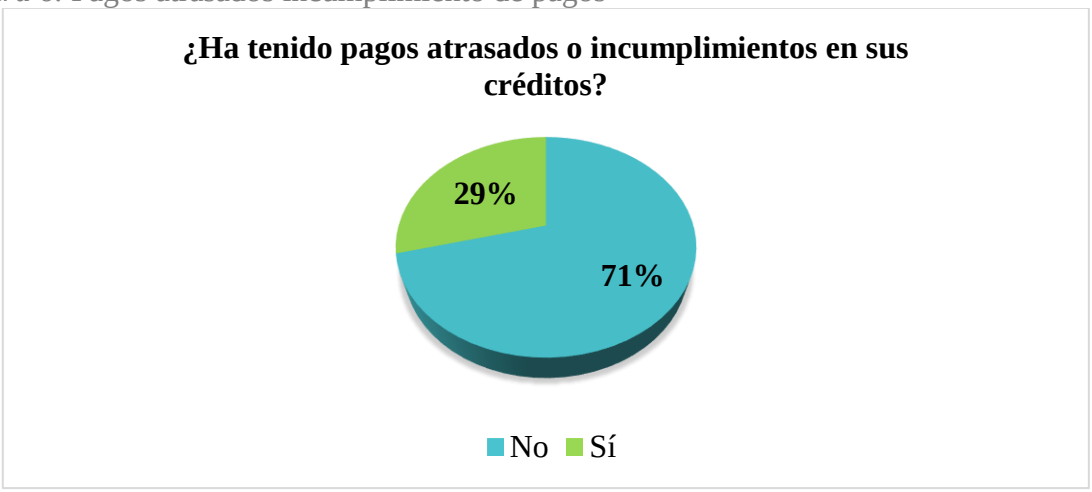
*Figura 5.* Ingresos Mensuales




*Nota.* Estadística de la encuesta sobre los ingresos mensuales. Fuente: Elaboración Propia.

En la figura 6 se representa el resultado de la encuesta en respuesta a la pregunta "¿Ha tenido pagos atrasados o incumplimientos en sus créditos?"

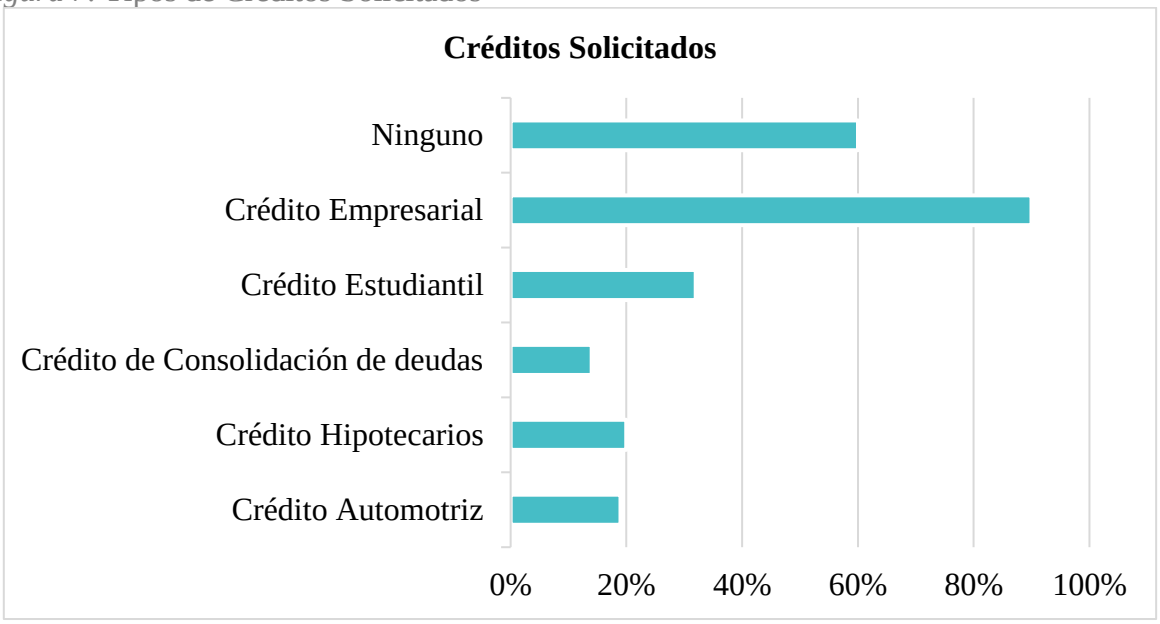
*Figura 6.* Pagos atrasados incumplimiento de pagos




*Nota.* Estadística de la encuesta sobre incumplimiento de pagos o pagos atrasados de los encuestados. Fuente: Elaboración Propia.

La Figura 7 se muestra el resultado en relaciona la pregunta sobre el tipo de créditos solicitados anteriormente por los encuestados.

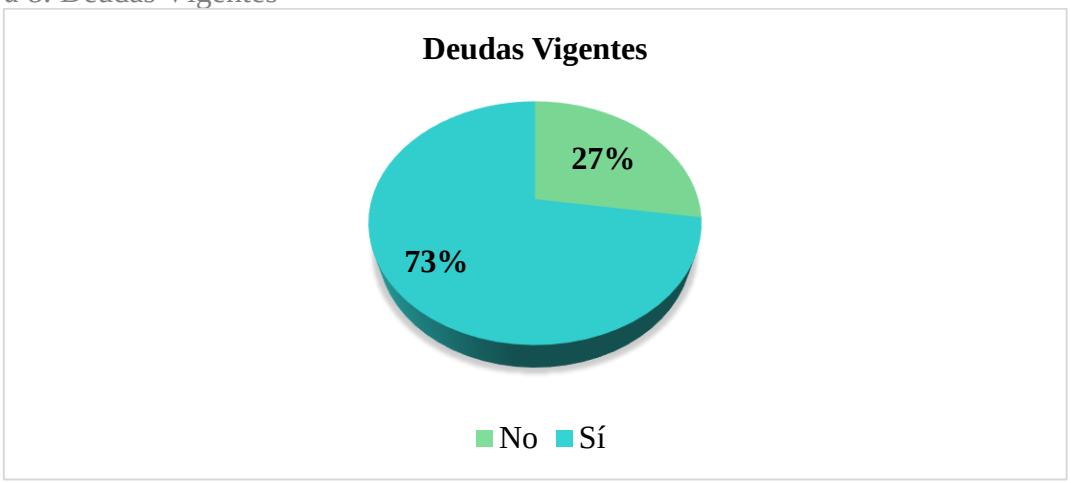
*Figura 7.* Tipos de Créditos Solicitados




*Nota.* Estadística sobre los tipos de créditos solicitados por los encuestados. Fuente: Elaboración Propia.

En la figura 8 se muestra el resultado de la encuesta acerca de la pregunta "¿Tiene deudas existentes en la actualidad?"

*Figura 8.* Deudas Vigentes




*Nota.* Estadística sobre deudas vigentes de los encuestados en la actualidad. Fuente: Elaboración Propia.

En la figura 9 se presenta la carga de datos provenientes de un archivo .csv con datos etiquetados. Del mismo modo se muestra la cabecera de la base de datos.

*Figura 9.* Encabezado de la base de datos.

| estado_civil | nivel_educacion | profesion | historial_crediticio | nivel_ingresos | ocupacion | buro_credito | nivel | monto_solicitado | cargas_familiares | tasa_interes | riesgo_moroso |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Viudo | Post Grado | Ingeniero civil | 0 | 4506 | Desempleado | 0 | 1 | 6292 | 3 | 2 | 0 |
| Soltero | Secundaria | Medico | 1 | 2511 | Empleado | 0 | 1 | 4536 | 4 | 5 | 1 |
| Viudo | Universitario | Gerente | 1 | 7874 | Desempleado | 0 | 1 | 4862 | 4 | 12 | 0 |
| Soltero | Universitario | Arquitecto | 1 | 4781 | Empleado | 0 | 1 | 5592 | 5 | 3 | 0 |

*Nota.* Muestra de los datos utilizados. Fuente: Elaboración Propia.

Se llevó a cabo un análisis descriptivo de los datos, centrándose en los campos presentados en la figura 9. Este análisis proporcionó una visión general de las características de los datos y permitió identificar posibles patrones o tendencias.

Tras finalizar el análisis descriptivo, se llevaron a cabo diversas técnicas de limpieza en los datos, incluyendo la eliminación de valores faltantes y la normalización de los datos. Después de procesar los datos y culminar con la limpieza de datos, se procedió a realizar pruebas en dos modelos de aprendizaje automático.

El primer modelo por evaluar fue el algoritmo de regresión logística. Este algoritmo de aprendizaje automático es utilizado para resolver problemas de clasificación binaria, donde se busca obtener una función lineal que relacione variables independientes, ya sean cuantitativas o cualitativas. De esta forma se utilizarán las variables encontradas para clasificar a los clientes si tendrá los medios económicos necesarios para afrontar el crédito otorgado o si el cliente pudiese presentar dificultades económicas que le impidan cumplir con sus obligaciones referentes a los pagos del crédito otorgado.

En el proceso de análisis, se realizó la lectura del archivo que contiene los datos utilizados para el entrenamiento. Uno de los pasos iniciales fue la conversión de la variable de ocupación en valores numéricos, donde se asignó el valor 1 para indicar una ocupación presente y 0 para una ocupación ausente. Al considerar la ocupación como variable independiente, se busca determinar si existe alguna relación entre la ocupación de los clientes y su capacidad para realizar pagos puntuales.

Posteriormente, se separaron las variables independientes y la variable objetiva del conjunto de datos. Esto se hace para evitar dificultades con los clientes de alto riesgo al momento de predecir el sesgo en la evaluación de riesgo de crédito.

Después de la separación de las variables, se dividió el conjunto de datos en dos conjuntos, uno de entrenamiento y otro de prueba. Finalmente, se inició la etapa de entrenamiento del modelo para evaluar el modelo utilizando las métricas planteadas para así verificar si pudo predecir correctamente los datos proporcionados.

En el segundo modelo de evaluación, se utilizó el algoritmo de redes neuronales, que se caracteriza por su capacidad de aprender y evolucionar a medida que se entrena con los datos. En este modelo, se utilizaron las mismas métricas y funciones de análisis de sesgo que en el primero.

La implementación del modelo se realizó con las mismas herramientas que el modelo anterior, con la única excepción de ***TensorFlow***, que se utiliza para crear redes neuronales junto con ***Keras***.

La red neuronal se declaró con una secuencia de tres capas. La primera capa, que es la capa de entrada, consta con 32 nodos y la función de activación ***'ReLu (Rectified Linear Unit)'***. La función ReLu se utiliza para la capa de entrada y capa oculta debido a que ayuda a que las redes neuronales aprendan patrones complejos en los datos. Mientras, la función softmax en la capa de salida es utilizada para obtener probabilidades normalizadas para cada clase.

## 4. RESULTADOS

El objetivo del modelo de predicción es detectar si un cliente está en riesgo de mora, es decir, si es probable que incumpla con el pago de las cuotas del crédito otorgado, o si es un cliente solvente. Para ello, se puede verificar la precisión de ambos modelos a través de las métricas presentadas en la tabla 3.

**Tabla 3.** Comparación de métricas entre ambos algoritmos.

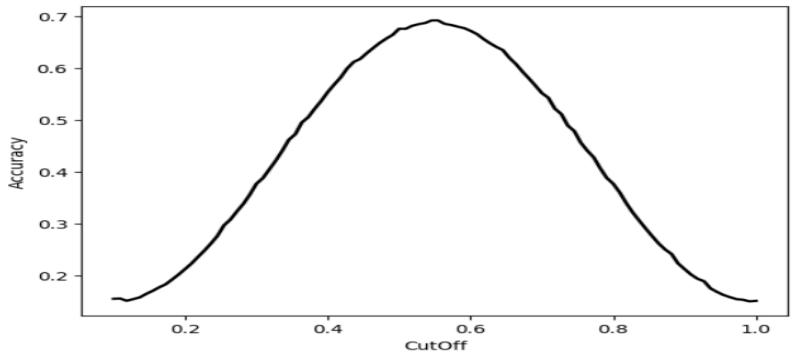
| Algoritmo | Desempeño | Error | Sensibilidad | Especificidad | Precisión | NPV |
|---|---|---|---|---|---|---|
| Regresión logística | 0.6870152 | 0.31278 | 0.6548182 | 0.6919541 | 0.68923 | 0.665756 |
| Redes neuronales | 0.7236 | 0.2764 | 0.91561 | 0.05363 | 0.73891 | 0.32547 |

*Nota.* Resultados obtenidos de las métricas del algoritmo de regresión logística y redes neuronales. Fuente: Elaboración Propia.

Los datos obtenidos del dataset “datosCredito” mediante el algoritmo de regresión logística indican que el modelo de predicción presenta un rendimiento moderado, con un desempeño del 68.70% y una tasa de error del 31.28%. La sensibilidad del modelo es del 65.48% y su especificidad es del 69.19%, lo que indica que también son valores moderados. Mientras que la precisión y el valor predictivo negativo del modelo son similares, con un 68.92% y un 66.58% respectivamente.

En base a los resultados obtenidos, se presenta también el historial de predicciones en la figura 10, en donde se muestra el punto de corte (CutOff), que es crucial para establecer el nivel de riesgo necesario que permite clasificar a un cliente como de riesgo o no. Además, el gráfico también proporciona información sobre el desempeño del modelo, lo cual es fundamental para evaluar el desempeño y la exactitud de las clasificaciones realizadas por el modelo.

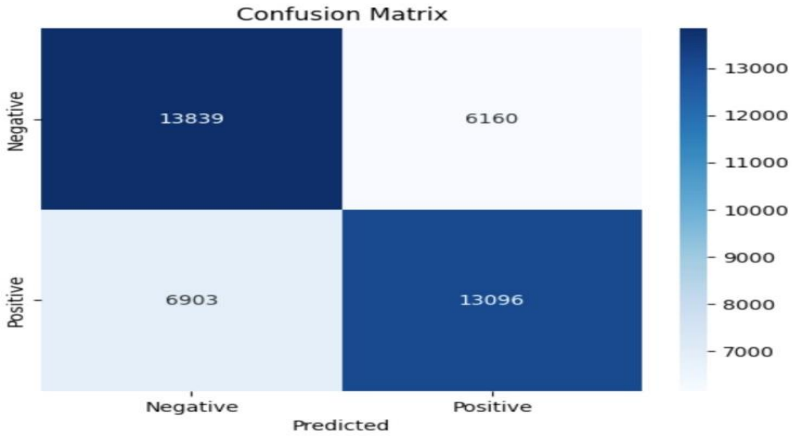
*Figura 10.* Histograma de regresión logística




*Nota.* Histograma de predicción de riesgo de crédito mediante el algoritmo de regresión logística. Fuente: Elaboración propia.

A la vez, se realizó una matriz de confusión para clasificar el modelo entre negativos, positivos, falsos positivos y falsos negativos como se muestra en la figura 11.

*Figura 11.* Matriz de confusión del modelo regresión logística



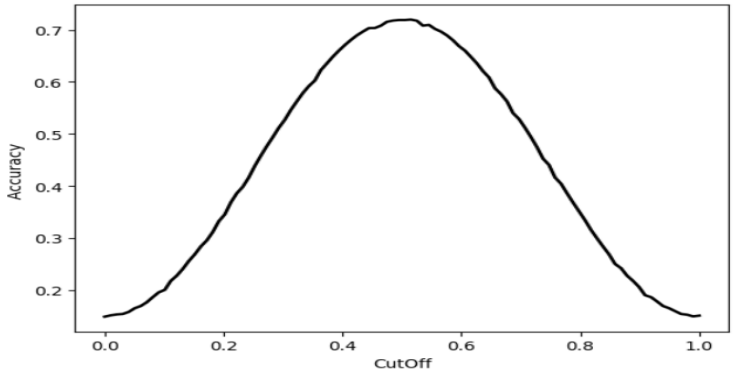

*Nota.* Resultados de valores en la matriz de confusión. Fuente: Elaboración propia.

En el esquema se puede apreciar que el algoritmo clasificó los datos del dataset “datosCredito” en 13096 registros como verdaderos positivos, 13839 registros como verdaderos negativos, del mismo modo, 6903 registros como falsos positivos y 6160 registros como falsos negativos.

Mientras que los datos obtenidos del dataset “datosCredito” mediante el algoritmo de redes neuronales indican que el modelo tiene un desempeño del 72.36% y su tasa de error es del 27.64%. En cuanto a la sensibilidad del modelo, se encuentra en un 91.56%, lo que indica que tiene la capacidad de identificar la gran mayoría de los casos positivos. Sin embargo, la especificidad del modelo es baja, situada en un 5.36%. En cuanto a la precisión, se encuentra en un 73.89%, lo que indica que el modelo tiene una alta proporción de predicciones positivas.

Por otro lado, el valor predictivo negativo del modelo se sitúa en un 32.54%. Asimismo, se generó una gráfica sobre los datos obtenidos del historial de predicciones que se muestra en la figura 12.

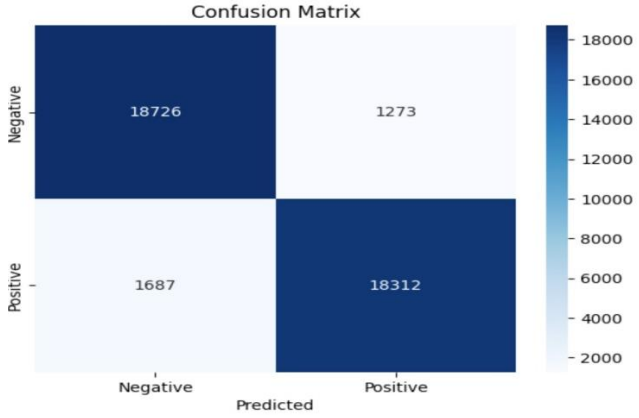
*Figura 12.* Histograma de redes neuronales




*Nota.* Histograma de predicción de riesgo de crédito mediante el algoritmo de redes neuronales. Fuente: Elaboración propia.

Mediante la figura 13, se muestra la matriz de confusión obtenida de los resultados.

*Figura 13.* Matriz de confusión redes neuronales




*Nota.* Resultados de valores en la matriz de confusión. Fuente: Elaboración propia.

Mediante el gráfico se resalta que el algoritmo catalogó del dataset “datosCredito”, 18312 registros como verdaderos positivos, 18726 registros como verdaderos negativos, del mismo modo, 1687 registros como falsos positivos y 1273 registros como falsos negativos.

## 5. CONCLUSIÓN

Los modelos de aprendizaje automático han demostrado ser una herramienta importante al predecir el riesgo crediticio en el sector financiero. Tanto el algoritmo de regresión logística como el algoritmo de redes neuronales destacan en el campo aplicado a la predicción de riesgo de crédito en el sector financiero. Debido a sus capacidades para identificar patrones y correlaciones en grandes conjuntos de data sets, demostrando significativamente mejoras en cuanto a la precisión y eficacia de las evaluaciones crediticias en comparación a los métodos tradicionales.

La recopilación de datos mediante encuesta a personas con experiencia al solicitar un crédito y entrevistas con analistas y expertos de la industria permitió analizar y comparar las características relevantes que fueron señaladas mediante sus opiniones. Debido a que es una forma valiosa de aumentar la precisión de los modelos de aprendizaje automático al estimar el riesgo de mora de los clientes y así contribuir a una mejor toma de decisiones al momento de analizar y tratar los datos recopilados, lo que a su vez permite aumentar la confianza en la precisión de los resultados obtenidos.

Al comparar el modelo de regresión logística y el de redes neuronales, se obtuvo un mayor porcentaje de sensibilidad del 91.56 % para el algoritmo de red neuronal, lo que lo hace que sea una opción más factible en escenarios donde la detección precisa de los casos positivos es importante, como lo es al momento de evaluar el sesgo para la predicción de crédito. Sin embargo, es importante tener en cuenta que cada modelo tiene fortalezas y limitaciones únicas que deben evaluarse en función de los requisitos específicos de cada necesidad financiera.

La aplicación de estos modelos proporciona un impacto positivo en la automatización de procesos, la eficiencia operativa y la capacidad de respuesta a patrones complejos del mercado. Sin embargo, es importante que las instituciones financieras consideren las implicaciones éticas asociadas con el uso de algoritmos predictivos, garantizando la transparencia y la equidad en su implementación. Por ello, los resultados obtenidos ayudan a que el sector financiero pueda ver por el bienestar económico no solo de la institución sino también del cliente y así generar una confianza pública.

# REFERENCIAS

AL-Najjar Dana, Al-Rousan Nadia, & AL-Najjar Hazem. (2022). Machine Learning to Develop Credit Card Customer Churn Prediction. *Theoretical and Applied Electronic Commerce Research*. https://doi.org/10.3390/jtaer17040077

Álvarez Jaime Grau. (2020). MACHINE LEARNING Y RIESGO DE CRÉDITO. *Universidad Pontificia Comillas Facultad de Ciencias Económicas y Empresariales*. https://repositorio.comillas.edu/rest/bitstreams/411260/retrieve

Banco Central del Ecuador. (2022, June). *NOTA METODOLÓGICA SOBRE LAS ESTADÍSTICAS MONETARIAS Y FINANCIERAS: NUEVA SEGMENTACIÓN DE CRÉDITO*. https://contenido.bce.fin.ec/documentos/PublicacionesNotas/Catalogo/IEMensual/Indices/m2043052022.htm

Benites Elizalde, Consuelo de Jesús, Arteaga Gordón, & Edgar Andrés. (2022). *Aplicación de modelos de machine learning para clasificación de clientes en una empresa auxiliar de servicios financieros ecuatoriana especializada en crédito vehicular para el período junio 2019- octubre 2021* [Universidad Central del Ecuador]. http://www.dspace.uce.edu.ec/handle/25000/27960

Cela Gonzalo, & Cuenca Juan Pablo. (2019). Propuesta de modelo de machine learning para la evaluación de riesgo de crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito La Merced. *Universidad Católica de Cuenca (UCACUE)*. https://www.researchgate.net/publication/337480778

Dans Enrique. (n.d.). *El machine learning y sus sesgos*. Retrieved February 2, 2024, from https://www.enriquedans.com/2019/11/el-machine-learning-y-sus-sesgos.html

Espinoza Rosero, & Gloria Belén. (2020). El Crédito Bancario y las Pymes en Ecuador. *Yachana Revista Científica*, *9*(2). http://repositorio.ulvr.edu.ec/handle/44000/4048

Ferrante Enzo. (2021). Inteligencia artificial y sesgos algorítmicos ¿Por qué deberían importarnos? *Universidad Nacional Del Litoral*. https://biblat.unam.mx/hevila/Nuevasociedad/2021/no294/3.pdf

Figueroa Jorge, Ferreira Guillermo, González Reinaldo, Martínez Fernández, & Tamahí Constanza. (2022). *Comparación de modelos machine learning aplicados al riesgo de crédito.* [Universidad de Concepción]. http://repositorio.udec.cl/jspui/handle/11594/9846

Granda Iván. (2017). MATRICES DE TRANSICIÓN DEL SISTEMA DE BANCOS. *Superintendencia de Bancos Del Ecuador  Dirección Nacional de Estudios e Información*. https://www.superbancos.gob.ec/bancos/wp-content/uploads/downloads/2017/08/MT1_mar_17.pdf

Hermitaño  Juler. (2022). Aplicación de Machine Learning en la Gestión de Riesgo de Crédito Financiero: Una revisión sistemática. *Universidad de Lima*. https://doi.org/10.26439/interfases2022.n015.5898

Liming Brotcke. (2022). Time to Assess Bias in Machine Learning Models for Credit Decisions. *Riask and Financial Management*. https://doi.org/10.3390/jrfm15040165

López Luis. (2022). EXPLICACIÓN Y PREDICCIÓN DEL DEFAULT EN CRÉDITOS, CON LA IMPLEMENTACIÓN DE MODELOS DE MACHINE LEARNING. *Universidad Potificia*

*Comillas*. https://repositorio.comillas.edu/jspui/bitstream/11531/57070/4/TFG%20-%20Lopez%20Blanco%2C%20Luis%20Ramiro.pdf

Mayank Anand, Arun Velu, & PawanWhig. (2022). Prediction of Loan Behaviour with Machine Learning Models for Secure Banking. *Computer Science and Engineering (JCSE), 3*. https://icsejournal.com/index.php/JCSE/article/view/237/117

Noriega Jomark Pablo, Rivera Luis Antonio, & Herrera Jose Alfredo. (2023). *Machine Learning for Credit Risk Prediction: A Systematic Literature Review*. https://doi.org/10.20944/preprints202308.0947.v1

Poveda Myriam. (2019). Riesgo de crédito: Evidencia en el sistema bancario ecuatorianoCredit risk: Evidence in Ecuadorian banking system. *Universidad Técnica de Ambato*. https://revistas.uta.edu.ec/erevista/index.php/bcoyu/article/view/842/811

Pucha Gualoto, & Oscar Iván. (2022). *Desarrollo de un modelo de predicción basado en Algoritmos de Machine Learning para medir el riesgo crediticio.* [q]. http://bibdigital.epn.edu.ec/handle/15000/22290

Ruiz Georges. (2022). *Predicción del Default en Carteras de Crédito: Un Enfoque de Machine Learning (ML).*

Småros Johanna, & Kaleva Henri. (2020, August 6). *La Guía Completa sobre Machine Learning en la Previsión de la Demanda en Retail*. https://www.relexsolutions.com/es/publicaciones/la-guia-completa-sobre-machine-learning-en-la-prevision-de-la-demanda-en-retail/

Vasquez Leyva Oliver. (2019). *SISTEMA PREDICTIVO BASADO EN UN MODELO CREDIT SCORING DE APRENDIZAJE AUTOMÁTICO PARA LA MEDICIÓN DEL RIESGO CREDITICIO EN LOS CRÉDITOS PYME DE LA EDPYME ALTERNATIVA S.A.* https://repositorio.uss.edu.pe/handle/20.500.12802/6357

ANEXO

# Encuesta sobre Perfil Socioeconómico y Experiencia Crediticia

1. Nombres

_____

2. Apellidos

_____

3. Edad

_____

4. Género

*Marca solo un óvalo.*

- ◯ Masculino
- ◯ Femenino

5. Estado Civil

*Marca solo un óvalo.*

- ◯ Soltero/a
- ◯ Casado/a
- ◯ Conviviente
- ◯ Separado/a
- ◯ Divorciado/a
- ◯ Viudo/a

6. Nivel educativo

*Marca solo un óvalo.*

- ◯ Sin educación formal
- ◯ Primaria
- ◯ Secundaria
- ◯ Educación universitaria
- ◯ Posgrado o estudios superiores

7. Ocupación

_____

8. Ingresos mensuales

_____

9. ¿Ha solicitado crédito anteriormente?

*Marca solo un óvalo.*

- ◯ Sí
- ◯ No

10. ¿Ha tenido pagos atrasados o incumplimientos en sus créditos?

*Marca solo un óvalo.*

- ◯ Sí
- ◯ No

11. Tipo de créditos solicitados anteriormente

*Marca solo un óvalo.*

- ◯ Crédito hipotecario
- ◯ Crédito automotriz
- ◯ Crédito estudiantil
- ◯ Crédito empresarial
- ◯ Crédito de consolidación de deudas
- ◯ Otro: _____

12. ¿Tiene deudas existentes en la actualidad?

*Marca solo un óvalo.*

- ◯ Sí
- ◯ No