Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería Mecánica Industrial

DISEÑO DE INVESTIGACIÓN DEL MODELO ESTADÍSTICO PARA ESTIMAR LA PROBABILIDAD DE IMPAGO DE CRÉDITOS DE UN BANCO GUATEMALTECO MEDIANTE REGRESIÓN LOGÍSTICA MULTIVARIADA.

Dale Jim Meléndez Herrera
Asesorado por el Mtra. Claudia Lorena García Bran

Guatemala, marzo de 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DISEÑO DE INVESTIGACIÓN MODELO ESTADÍSTICO PARA ESTIMAR LA PROBABILIDAD DE IMPAGO DE CRÉDITOS DE UN BANCO GUATEMALTECO MEDIANTE REGRESIÓN LOGÍSTICA MULTIVARIADA.

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

DALE JIM MELÉNDEZ HERRERA
ASESORADO POR LA MTRA. CLAUDIA LORENA GARCÍA BRAN

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO INDUSTRIAL

GUATEMALA, MARZO DE 2023

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANA	Ing. Aurelia Anabela Cordova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANA	Ing. Aurelia Anabela Cordova Estrada
EXAMINADOR	Ing. Selvin Estuardo Joachin Juarez
EXAMINADORA	Inga. Mayra Saadeth Arreaza Martínez
EXAMINADORA	Inga. María Martha Wolford Estrada
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

DISEÑO DE INVESTIGACIÓN MODELO ESTADÍSTICO PARA ESTIMAR LA PROBABILIDAD DE IMPAGO DE CRÉDITOS DE UN BANCO GUATEMALTECO MEDIANTE REGRESIÓN LOGÍSTICA MULTIVARIADA.

Tema que me fuera asignado por la Dirección de la Escuela de Estudios de Postgrado, con fecha 25 de septiembre de 2022

Dale Jim Meléndez Herrera

EEPFI-PP-1783-2022
Guatemala, 10 de noviembre de 2022

Director
César Ernesto Urquizú Rodas
Escuela Ingenieria Mecanica Industrial
Presente.

Estimado Ing. Urquizú

Reciba un cordial saludo de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería.

El propósito de la presente es para informarle que se ha revisado y aprobado el Diseño de Investigación titulado: **MODELO ESTADÍSTICO PARA ESTIMAR LA PROBABILIDAD DE IMPAGO DE CRÉDITOS DE UN BANCO GUATEMALTECO MEDIANTE REGRESIÓN LOGÍSTICA MULTIVARIADA**, el cual se enmarca en la línea de investigación: **Todas las áreas - Estadística multivariada**, presentado por el estudiante **Dale Jim Meléndez Herrera** carné número **200715321**, quien optó por la modalidad del "PROCESO DE GRADUACIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA OPCIÓN ESTUDIOS DE POSTGRADO". Previo a culminar sus estudios en la Maestría en ARTES en Estadistica Aplicada.

Y habiendo cumplido y aprobado con los requisitos establecidos en el normativo de este Proceso de Graduación en el Punto 6.2, aprobado por la Junta Directiva de la Facultad de Ingeniería en el Punto Décimo, Inciso 10.2 del Acta 28-2011 de fecha 19 de septiembre de 2011, firmo y sello la presente para el trámite correspondiente de graduación de Pregrado.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS
ESCUELA DE POSTGRADO
FACULTAD DE INGENIERIA
DE GUATEMALA

Mtra. Claudia Lorena García Bran
Asesor(a)

Lcda. Claudia Lorena García Bran
Contadora Pública y Auditora
No. De Colegiado 8829

Mtro. Edwin Adalberto Bracamonte Orozco
Coordinador(a) de Maestría

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERIA
DIRECCIÓN
ESCUELA DE POSTGRADO

Mtro. Edgar Darío Alvaréz Cotí
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÍA

EEP-EIMI-1437-2022

El Director de la Escuela Ingenieria Mecanica Industrial de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del Asesor, el visto bueno del Coordinador y Director de la Escuela de Estudios de Postgrado, del Diseño de Investigación en la modalidad Estudios de Pregrado y Postgrado titulado: **MODELO ESTADÍSTICO PARA ESTIMAR LA PROBABILIDAD DE IMPAGO DE CRÉDITOS DE UN BANCO GUATEMALTECO MEDIANTE REGRESIÓN LOGÍSTICA MULTIVARIADA**, presentado por el estudiante universitario **Dale Jim Meléndez Herrera**, procedo con el Aval del mismo, ya que cumple con los requisitos normados por la Facultad de Ingeniería en esta modalidad.

ID Y ENSEÑAD A TODOS

Ing. César Ernesto Urquizú Rodas
Director
Escuela Ingenieria Mecanica Industrial

Guatemala, noviembre de 2022

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.262.2023

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería Mecánica Industrial, al Trabajo de Graduación titulado: **DISEÑO DE INVESTIGACIÓN DEL MODELO ESTADÍSTICO PARA ESTIMAR LA PROBABILIDAD DE IMPAGO DE CRÉDITOS DE UN BANCO GUATEMALTECO MEDIANTE REGRESIÓN LOGÍSTICA MULTIVARIADA,** presentado por: **Dale Jim Meléndez Herrera,** después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, marzo de 2023

AACE/gaoc

ACTO QUE DEDICO A:

Dios	Por ser el centro de mi vida y ser la fuerza que me impulsa a lograr mis metas.
Mis padres	Por darme siempre su apoyo y creer en mí.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por ser la casa de estudios que me instruyó y formó mi forma de pensar.
Facultad de Ingeniería	Por ser formadora de pensamiento critico y haberme enseñado a ver la vida con ojos diferentes.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
Gal.	Galón
Hrs.	Horas
KWH.	Kilo Vatio Hora

GLOSARIO

Calidad crediticia Es el rasgo característico que posee un crédito o instrumento de deuda y está determinada por la probabilidad de incumplimiento. La calidad crediticia es más alta en la medida que el incumplimiento es más bajo.

Calificación de riesgo Proceso de categorización de un deudor o emisor y de valores emitidos por los mismos, que consiste en la asignación de un nivel de riesgo crediticio específico pretende categorizar la probabilidad de que un evento adverso suceda.

Cancelación anticipada Pago del préstamo o cancelación parcial de una obligación antes de la fecha de vencimiento programada.

Capacidad de pago Es la posibilidad de que un prestatario actual o potencial pueda generar los beneficios económicos necesarios para honrar sus obligaciones y mantener en el tiempo un nivel de solvencia.

Créditos en mora Cartera en incumplimiento de capital y/o intereses, que se encuentra con acciones de cobranza o no, que ha dejado de generar ingresos por intereses, y con riesgos potenciales en su recuperación.

Créditos vigentes	Préstamos que presentan cumplimiento tanto en el pago de capital como de los intereses conforme al plan de pagos establecido en el contrato de crédito.
Cartera de créditos	Comprende los créditos otorgados por una institución financiera a terceros, que se originan en la actividad principal de intermediación financiera, sin considerar el estado actual de recuperación.
Garantía o colateral	Activo tangible negociable (mueble o inmueble, real o financiero), mientras también se denota de carácter intangible como el ingreso de la persona solicitante.
Incumplimiento o *default*	Falta de pago de una obligación o cualquier otro tipo de violación de las condiciones de un contrato de préstamo. Se considera comúnmente se utiliza este terminado cuando un cliente cuenta con una mora por más de 90 días con respecto
Intermediación financiera	Proceso mediante el cual, un agente, se encargade trasladar los excesos de liquidez de los ahorradores e inversores a las personas que necesitan capital, de tal forma que el inversor genere rendimientos, mientras que el gestor atrae un mayor capital

Liquidez	Es la capacidad de una persona o entidad de hacer frente a sus deudas de corto plazo por poseer activos fácilmente convertibles en efectivo.
Obligación	Es un deber impuesto por ley o por un contrato. El término también se utiliza para describir un valor u otro instrumento financiero, como un bono o un pagaré, que contiene la promesa del emisor de pagar al propietario.
Relación cuota ingreso	Denota la porción del ingreso que representa la cuota del crédito para el deudor

INTRODUCCIÓN

El presente estudio de investigación es una sistematización de ponderación de los solicitantes de crédito para identificar, medir y pronosticar la probabilidad de impago de operaciones de crédito de un banco del sistema guatemalteco.

El riesgo crediticio es la probabilidad de perdida de una entidad financiera dada la materialización de impago de los agentes económicos que es inherente al negocio bancario y es el principal giro de negocio.

Un banco del sistema busca evolucionar, ampliar su alcance y otorgar nuevos canales de atención al cliente, para realizar esto es necesario implementar sistemas de control para reducir el riesgo de crédito por lo que es necesario establecer un modelo estadístico basado en regresión lineal multivariada que logre una efectividad igual o superior al análisis cualitativo que se realiza actualmente. Este cuenta con medios de otorgamiento de créditos cualitativos que requieren de una analista de créditos experto que califique de forma cualitativa y cuantitativa a un cliente, además cuenta con mecanismos de recuperación efectivos que logran mantener la mora menor al 2% del total de la cartera.

La metodología de la investigación es de tipo cuantitativo, con diseño no experimental y el alcance es correlacional multivariado. Las variables que se analizarán corresponden a los datos que los solicitantes entregan al solicitar un crédito, por lo que son demográficas y de comportamiento de pagos para un periodo de tiempo de los años 2010 a 2022.

Se espera que el resultado brinde un banco un parámetro matemático para decidir si se le otorga el crédito a cada solicitante, con el fin de otorgar únicamente a los clientes que tengan probabilidad de pago completo de sus créditos y se pueda reducir las pérdidas.

En la primera fase, se revisará la información documental y se buscarán todas las fuentes para sustentar los temas a tratar para análisis. En la segunda fase, se recolectará la información del cliente y el comportamiento de pago anterior se recopilan al momento de la solicitud del préstamo, la empresa proporciona esta base de datos. En la tercera fase se utiliza los modelos y variables relevantes para estimar la probabilidad de incumplimiento de pagos, además utilizando la métrica de error estimado se evalúa la precisión de cada modelo. En la cuarta fase, se analizarán los resultados y se elaborará un informe final.

El trabajo de investigación es factible ya que todos los recursos necesarios están disponibles para todas las fases del estudio.

El informe final se dividirá en los siguientes 4 capítulos:

- Primer capítulo: El marco referencial incluirá los temas que sustentarán y formarán la base para esta investigación.

- Segundo capítulo: Se desarrollará el marco teórico el cual se dividirá en dos partes, la primera parte es la estadística, en la cual se describen los conceptos básicos, teorías y ecuaciones que fundamentan el estudio; en la segunda parte se detallará la teoría de gestión de riesgo.

- Tercer capítulo: Se presentarán los resultados.

- Cuarto capítulo: Se discutirán los resultados.

1. ANTECEDENTES

Las entidades financieras están expuestas a diversos riesgos, el riesgo más importante es el crediticio ya que es inherente al esquema de negocio, por tal razón existen diversos estudios orientados hacia la minimización de este.

De acuerdo con Hajiyev (2021) quien realizó una investigación descriptiva de encuestas y censos con el fin de estudiar la administración de riesgos en el mercado en Azerbaiyán, como primera parte investigó los modelos de medición de riesgo de crédito que enfrentará el banco de forma sectorial y por firma. En la segunda parte aborda las diferencias encontradas con otros países europeos, en la tercera parte explica los modelos y teorías de medición del riesgo de crédito tradicionales y nuevos utilizados en el sector bancario para minimizar los efectos del riesgo de crédito. Por último, se presenta la estructura general del sector bancario de Azerbaiyán y se explican las normas legales relativas al sector. En los resultados encontró que los bancos de Azerbaiyán utilizan los métodos de probabilidad lineal*, logit y probit*, Z-score de Altman, rendimiento del capital ajustado al riesgo y la matriz de crédito, el 78.9% de los bancos investigados cuenta con por lo menos un método de los mencionados para analizar los créditos y su cartera se desempeña mejor que aquellos bancos que utilizan métodos cualitativos. En la comparación entre bancos y países se determina que es necesario medir el riesgo de crédito en función del país y del banco ya que los riesgos son sustancialmente diferentes.

De esta forma es posible ver que los métodos estadísticos cuantitativos son ampliamente utilizados en los bancos europeos mostrando resultados consistentemente mejores que los bancos que no los utilizan para mejorar la

rentabilidad, de la misma forma existen diversos estudios buscan incrementar la rentabilidad mediante análisis estadísticos numéricos como en la investigación.

Según Navarrete (2017) quien aborda en su investigación de estadística multivariada titulada "*Swaps* de Incumplimiento Crediticio (CDS)" identificó y analizó diferentes tipos de riesgos asociados con derivados financieros, no sólo riesgo de crédito genérico. También se aborda su valoración, a través de modelos econométricos, y tipos de actividades que se formalizan con CDS. Finalmente, se analizó la relación entre CDS y Índices bursátiles y subíndices de los principales países europeos. El uso de modelos matemáticos sugiere que los CDS son una buena herramienta para la gestión del riesgo de crédito, pero su uso implica asumir nuevos riesgos. Al tratarse de productos negociados en mercados no organizados, dependiendo de las necesidades de las partes, presentan riesgos operativos, legales, de liquidez y de contraparte. Lo anterior es típico de mercado bursátil o extrabursátil. Los distintos modelos de rentabilidad presentados en este estudio indican que es necesario cuantificar el riesgo de crédito para asegurar la rentabilidad, incluyendo la prima de riesgo en los productos ofertados.

En ese orden de ides Torre (2018) apegado la norma de riesgo Basilea II busca optimizar las ganancias mediante el enfoque IRB de esa misma norma la cual asume que los factores de riesgo involucrados en la probabilidad de incumplimiento y la ratio de pérdida en caso de incumplimiento son independientes. Mediante una investigación de estadística multivariada basado en regresión logística analiza y cuantifica los efectos sobre el requerimiento de capital mínimo ante la presencia de correlación entre los factores que afectan la Probabilidad de Incumplimiento y la pérdida en caso de incumplimiento. La misma cartera se simula con diferente requerimiento de capital mínimo en el método IRB y se compara con dos conjuntos el Probabilidad de Incumplimiento

y pérdida en caso de incumplimiento reales y correlacionados; La principal conclusión es que a medida que crece la dependencia entre Probabilidad de Incumplimiento y pérdida en caso de incumplimiento, se subestima más el capital mínimo, por lo que los bancos tendrán medidas de mayor protección ante contingencias de incumplimiento de pagos si utilizan el modelo de reservas correlacional propuesto en la investigación.

Estas investigaciones tienen un enfoque principal en mejorar la rentabilidad basado en la gestión de riesgo, el riesgo de mayor implicación es el crediticio por lo que algunos autores como Prasad (2020) quien realizó una investigación de estadística multivariada correlacional con el objetivo de examinar el impacto de la gestión del riesgo de crédito en la rentabilidad de los bancos irlandeses. Tomo una muestra de 4 bancos irlandeses, para el estudio se calcularon métricas a partir de los informes financieros anuales de los bancos durante 11 años. Estas métricas han sido analizadas mediante análisis de regresión para comprender la relación entre la gestión del riesgo de crédito, la morosidad y el coeficiente de solvencia (variables independientes) y rentabilidad para el accionista (variable dependiente). Los modelos se desarrollan haciendo tres variaciones en el conjunto de datos: eliminando un valor atípico, retrasando la variable dependiente por un período y retrasando la variable dependiente con un valor atípico eliminado. También se examina una línea de tendencia no lineal para modelos relevantes. La calidad del conjunto de datos también se verifica mediante pruebas de diagnóstico. La prueba incluye autocorrelación y heterocedasticidad; también se realiza una prueba de multicolinealidad y para finalizar, se realizaron pruebas de significancia. Los resultados del estudio indican que existe una asociación negativa entre el préstamo moroso y la rentabilidad para el accionista, y una relación positiva entre el índice de adecuación de capital y la rentabilidad para el accionista. También se observa que también existe una relación positiva entre la situación económica y la

rentabilidad de los bancos irlandeses. Los datos no muestran evidencia de multicolinealidad y autocorrelación, pero hay alguna evidencia de la presencia de heterocedasticidad en los datos. Esta investigación amplia los diferentes métodos que se pueden emplear para calibrar el modelo.

Por su parte Jaramillo (2021) realiza una investigación de estadística multivariada correlacional en la cual creó y comparó la precisión del modelo de regresión logística frente a algunos modelos de *Machine Learning* para la estimación del riesgo de crédito en una cartera de consumo; los modelos contrastados son Regresión logística, *Random Forest, Support Vector Machine y Multi-layer Perceptron* que miden la capacidad de predecir la eficiencia de la estimación de los clientes que van a entrar en mora, el autor citado concluyó que el modelo con resultados más equilibrados al momento de la evaluación es el *Random Forest*, dado que fue el que presentó el mejor ajuste de acuerdo con las métricas de exactitud evaluadas. El modelo de regresión logística tiene un desempeño similar y que se perfilan como competidores de la metodología tradicional, con el valor agregado de que fácilmente aplicables y tienen un potencial de refinamiento importante. También concluye que a pesar de la capacidad de predicción cada entidad debe realizar su propia medición y seleccionar el modelo, cuyo desempeño esté más acorde al apetito de riesgo de la entidad y que sea mejor para sus objetivos de negocio y niveles de exposición de riesgo deseados.

Continuando con la idea Devia (2015) quien realizó una investigación que pretendía realizar modelos estadísticos del riesgo de crédito, el estudio aborda desde tres técnicas distintas basado en los enfoques: paramétricos, semiparamétricos y no paramétricos. El primero fue un análisis de supervivencia. Este modelo de probabilidad de *default* se deriva de la función de distribución condicional del tiempo y se calcula usando tres clases de estimadores.

posteriormente se construyó un segundo modelo de probabilidad de incumplimiento, basado en la reincidencia en el impago de créditos, que proveyó una fórmula que describe la probabilidad de impago para un mismo sujeto, conociendo su historial de crediticio, solvencia y el tiempo del impago anterior. El tercer modelo evaluado fue de calificación crediticia construido mediante técnicas de regresión logística el cual calcula una estimación de la propensión a la probabilidad de default de los clientes dado un conjunto de variables dependientes sobre las que se realizó una regresión logística, Todos los modelos realizados fueron hechos a la medida y pueden ser utilizados únicamente por la empresa en estudio como una herramienta para monitorear y evaluar el perfil de pagos de los clientes con probabilidad de default. El modelo elegido el segundo puesto que fue el que mejor demostró predictibilidad en la probabilidad de default.

Según Andrade (2020) quien realizó una investigación de estadística multivariada con el objetivo de crear formas novedosas de calificación crediticia que cierren la brecha entre las redes neuronales simples y las metodologías avanzadas en aprendizaje profundo aplicado a la calificación crediticia. Propone una nueva metodología para aprender representaciones de datos útiles de clientes bancarios introduciendo una etapa de supervisión, esta propuesta propone agrupar los datos de entrada utilizando la transformación de acuerdo con el peso de la evidencia y compararlo con un modelo sin agrupación. El método propuesto aprende representaciones de datos que pueden capturar la solvencia de los clientes en una estructura de agrupamiento bien definida. Además, conservan la coherencia espacial de la solvencia de los clientes. Los modelos propuestos utilizan la teoría probabilística para inferir la solvencia de los clientes desconocidos, lo cual es una clara ventaja sobre los enfoques tradicionales cualitativos y se parametrizó un modelo de mezcla gaussiana con redes neuronales para mejorar. Finalmente, se abordó la calificación crediticia como un problema de aprendizaje multimodal. Como resultado, se creó un

modelo que genera datos crediticios futuros, basados en datos de aplicaciones capaces de reducir la dimensionalidad de los datos de entrada, sino que también es capaz de aprender una representación de datos útil que captura la solvencia de los clientes. Los grupos identificados son adecuados para un enfoque de calificación crediticia basada en segmentos, que logra un mayor rendimiento en comparación con el enfoque tradicional de calificación crediticia, en el que solo se ajusta un clasificador a todo el conjunto de datos; también se encontró que agregar solicitudes rechazadas mejora la precisión de clasificación de nuestros modelos propuestos y potencialmente resuelve el problema del sesgo de selección.

Continuando con la misma línea de investigación Argomaniz (2019) llevó a cabo estadística multivariada para determinar los factores que influyen en la probabilidad de Incumplimiento de una base de datos crediticia americana. Se utilizó el modelo *logit* para determinar el umbral de Probabilidad de Incumplimiento que presenta mayores utilidades. La base de datos contemplaba préstamos rescindidos otorgados a particulares, orientados al consumo con 139 variables. A partir de los datos brutos se redujo el número de variables a las más significativas, se consideraron dos modelos y el modelo elegido fue el que presentó mayores utilidades para la Institución Financiera. Se evaluó la capacidad del modelo por medio de *Backtesting*. Los dos modelos presentan equivalencia en cuanto a poder discriminatorio y de calibración, por lo que la elección del modelo se basó en rentabilidad potencial del segundo modelo el cual estima se 31,199,636.64 de dólares anuales que el modelo 1. También concluyeron que los resultados arrojados por las pruebas estadísticas parecen muy consistentes precisos del modelo, lo que sugiere que las variables elegidas son una buena combinación para modelar la probabilidad de incumplimiento.

Una vez teniendo claro que los autores anteriores mencionan sobre los modelos estadísticos resultan en mejores resultados Montalván (2019) realizó una investigación de estadística multivariada con el fin de contrastar la hipótesis de sí el modelado de calificación crediticia de carteras utilizando redes neuronales funciona mejor que los métodos de regresión logística. Para ello se utilizó información de las carteras de instituciones financieras ecuatorianas, analizando las solicitudes al momento inicial del crédito. La capacidad predictiva y la capacidad de discriminar entre buenos y malos clientes se basan en el estadístico KS, el coeficiente de Gini, la matriz de confusión, la curva ROC y el criterio de información de Akaike. Como resultado, se encontró que el modelo de red neuronal resultó con un mejor ajuste a los datos reales, ya que el criterio de información de Akaike fue 8029.2 puntos más bajo que el criterio de regresión logística. Además, las estadísticas de KS, los coeficientes de Gini y la curva ROC muestran que el uso de la red neuronal proporciona una mejor clasificación de clientes que las estadísticas del modelo de regresión logística en 5,19, 5,84 y 2,92 puntos porcentuales, respectivamente. Finalmente, la matriz de confusión en el modelo de red neuronal resultó menos propensa a errores en comparación con el mismo umbral óptimo.

De los autores anteriores se puede observar que la gestión del riesgo crediticio por medio de métodos estadísticos ha tenido mejores resultados que los métodos clásicos y se usa ampliamente en bancos europeos, por lo que se denota la necesidad de implementar esta clase de modelos en la banca guatemalteca.

2. PLANTEAMIENTO DEL PROBLEMA

2.1. Contexto general

Los bancos en Guatemala tienen como giro principal de negocio otorgar créditos. Esta actividad conlleva un riesgo inherente de pérdidas para la entidad, en caso de que los clientes a los que se les otorgaron los recursos no los devuelvan en las condiciones pactadas, es decir, cuando los clientes incumplen sus compromisos.

El sistema bancario en Guatemala es tradicional, está en la fase inicial de la banca digital, actualmente las sucursales físicas son las preferidas por la población para adquirir prestamos, con el objetivo de prepararse para la era digital es necesario agilizar los procesos a la vez que se incrementan las medidas de mitigación de riesgo; para mantener el riesgo de impago en valores mínimos se hace necesario profundizar en análisis de los clientes, un banco del sistema se encuentra en el proceso de otorgamiento, realiza un análisis cualitativo por medio de un analista de créditos y de forma cuantitativa únicamente considera los ingresos y las deudas de los clientes.

2.2. Descripción del problema

La entidad bancaria antes mencionada cuenta con mecanismos de recuperación efectivos, su mora ha permanecido en los últimos años menor al 2% del total de la cartera, actualmente el proceso de análisis requiere un analista de créditos experto que califique de forma cualitativa y cuantitativa al cliente. Para incrementar el volumen de colocación, otorgar créditos por nuevos canales de

atención y controlar la mora, se necesitan de modelos matemáticos que permitan determinar la probabilidad de pago a clientes de forma automatizada y eficaz, con estos métodos no se cuenta. Así que de esto se deriva la necesidad de crear e implementar un modelo para prevenir el riesgo de crédito y sea igual o más efectivo que el análisis cualitativo que se realiza actualmente.

2.3. Formulación del problema

2.3.1. Pregunta central

¿Cuál es el modelo estadístico que describe la probabilidad de impago de crédito con el mejor ajuste?

2.3.2. Preguntas auxiliares

- ¿Cuáles son las características del cliente que pueden ayudar a identificar su probabilidad de impago?

- ¿Qué cantidad de créditos históricos es necesario analizar para poder crear un modelo estadístico?

- ¿Cuál es el potencial de mejora que pueda representar la implementación de un modelo de probabilidad de *default* crediticio?

2.4. Delimitación del problema

El problema está relacionado con un banco guatemalteco que tiene operaciones a nivel nacional y cuenta con datos de clientes de los últimos 10

años de los que es útil tomar en cuenta los clientes del banco que han tenido por lo menos un crédito aprobado.

3. JUSTIFICACIÓN

El problema planteado está enmarcado en el sector financiero guatemalteco y en la línea de investigación de estadística multivariada regresiva para brindar un parámetro que determine la probabilidad de *default* de los solicitantes de crédito de una entidad bancaria. Para esto, se utilizarán los datos históricos del banco desde 2012 al 2022 lo cual servirá de base para proponer un modelo estadístico personalizado para el banco mencionado anteriormente.

Esto servirá para la gestión del riesgo crediticio del banco, lo cual ayudará a reducir la cantidad de créditos impagos por lo que disminuirán las pérdidas operativas.

La motivación para realizar esta investigación es el facilitar el análisis de créditos y tecnificarlo ya que actualmente el análisis de créditos se hace de acuerdo con criterios establecidos en un manual interno y las decisiones cuentan con un grado de subjetividad del analista de créditos.

Este informe brindará un modelo hecho a la medida para un banco en particular que servirá de base para poder utilizarlo en los sistemas de análisis del banco, con lo cual se pretende obtener la probabilidad de default del solicitante

El beneficio que representa este modelo servirá para los accionistas quienes verán reflejado un incremento en la rentabilidad y personal gerencial del banco, ya que podrán gestionar el riesgo de una forma más técnica. También, los analistas de créditos que contarán con una herramienta que facilitará su trabajo

y por último, se beneficiará la comunidad académica al contar con un documento de referencia para futuras investigaciones sobre el riesgo crediticio.

El presente estudio generará un modelo personalizado para un banco en particular, pero servirá de referencia futura para otras investigaciones relacionadas el riesgo crediticio de las entidades bancarias.

4. OBJETIVOS

4.1. General

Proponer un modelo estadístico para estimar la probabilidad de impago de créditos de un banco del sistema guatemalteco mediante regresión logística multivariada.

4.2. Específicos

- Identificar las variables cuantitativas y cualitativas de mayor incidencia sobre la probabilidad de impago mediante un análisis de la significancia para determinar las características que se necesita evaluar del cliente.

- Calcular la muestra necesaria mediante muestreo aleatorio para elaborar el modelo estadístico.

- Estimar mediante el análisis de la rentabilidad y certeza, los posibles riesgos y potencial de mejora que pueda implicar el modelo propuesto para su implementación.

5. NECESIDADES QUE CUBRIR Y ESQUEMA DE LA SOLUCIÓN

El estudio propuesto contribuirá con un modelo estadístico que pretende automatizar la evaluación de créditos estimando la probabilidad de que los clientes incumplan con el pago de sus créditos para un banco en particular. De igual forma se desconocen las características que inciden en una mayor probabilidad que no realicen sus pagos, por lo que el banco necesita estimar la rentabilidad de implementar el modelo propuesto, para conocer las implicaciones que puedan causar evaluar a clientes y medir la capacidad de incrementar o reducir la rentabilidad.

Para realizar el estudio el banco particular proporcionará la base de datos histórica de los clientes que han adquirido un crédito por lo menos un crédito en los años 2012 al 2022.

Con los datos proporcionados se construirá un modelo llamado Regresión logística multivariada, el cual brinda como resultado, la probabilidad de impago de un cliente la efectividad se medirá por medio del Error medio cuadrado (RMSE) el cual servirá para determinar la rentabilidad y el potencial de mejora que tiene el modelo respecto a el actual método de otorgamiento de créditos. También se buscará encontrar cuales son las variables con mayor influencia en que una persona sea propensa a incumplir el pago de sus deudas, la cual se comprobará mediante la Prueba de Wald.

6. MARCO TEÓRICO

6.1. Métodos de análisis estadístico de datos

Existen diferentes métodos que aplican diferentes modelos matemáticos y estadísticos para poder determinar la causalidad de las variables, Oroes (2014) describe el análisis estadístico por medio de dos tipos.

- Estadística descriptiva: Pretende describir las características esenciales de sus datos sin hacer inferencias ni predicciones. El análisis descriptivo es un requisito antes de realizar cualquier otro análisis, ya que ayuda a seleccionar el método matemático o estadístico apropiado para aplicar al conjunto de datos.

- Estadística inferencial: Busca analizar conjuntos de datos o conjuntos de datos, ya sea encontrando relaciones entre dos o más variables de uno o más conjuntos de datos relacionados o probando hipótesis sobre el conjunto de datos. Para llevar esto acabo existen diferentes métodos tales como regresiones lineales.

6.1.1. Regresión lineal múltiple

La regresión múltiple según Díaz y Morales (2012) se centra en la dependencia de una variable de respuesta de un conjunto de regresores o predictores. Usando el modelo de regresión, se mide el impacto de cada regresor en la respuesta. Uno de los objetivos es una estimación para predecir la media

de la variable dependiente, basada en el conocimiento de las variables independientes o predictoras.

Fernandez, Córdova y Cordero (2002) definen que cuando hay más de una variable independiente en una distribución, el análisis de regresión se denomina regresión múltiple. El procedimiento para investigar qué función se ajusta mejor a la distribución observada es descomponer los valores observados en valores sumados en una ecuación Y= $y_i + e_i$, el primer resultado corresponde a la parte explicada por la ecuación de regresión, y la segunda parte corresponde a la parte no explicada, comúnmente conocida como residual o error. El método más común para seguir estudios de ajuste funcional es el método de mínimos cuadrados, aplicable a funciones lineales, parabólicas, exponenciales, potenciales e hiperbólicas. Otra dificultad con la regresión múltiple es la incapacidad de graficar las distribuciones y el ajuste correspondiente. Si Y es la variable dependiente y se calcula a partir de las variables independientes, estamos encontrando la ecuación de regresión de Y respecto a X. La variable dependiente también se denomina variable endógena y las variables dependientes son variables exógenas o explicaciones, en este tema, y el comienzo de la regresión múltiple, se usarán funciones lineales.

6.1.2. Regresión logística

La regresión logística es de acuerdo con Alvarez (1995) una técnica de análisis multivariada donde la variable dependiente u objetivo es una variable binaria y la(s) variable(s) independiente(s) puede(n) ser cualitativa(s) o cuantitativa(s). Si la variable independiente en el modelo es cualitativa con categorías H, será necesario hacer H + 1 variables ficticias o *Dummy* para que todas las posibles respuestas de esa variable estén representadas en el modelo. Las variables binarias solo pueden tener dos valores. La regresión logística se

divide en dos tipos básicos: regresión simple, cuando el modelo tiene solo una variable independiente, y regresión múltiple, cuando el modelo tiene múltiples variables independientes. En la regresión logística, la variable dependiente es binaria correspondiente al valor nominal (sí o no) con frecuencia. Para construir un modelo matemático, necesitas números por lo que se asignan valores numéricos (0 o 1). Esto se obtiene considerando la probabilidad de ocurrencia de algún valor de la variable dependiente. Los modelos ayudan a saber qué factores aumentan o disminuyen más la probabilidad de un evento.

6.1.3. Multicolinealidad

El hecho de que una o más de las variables independientes iniciales en la ecuación de regresión múltiple realmente dependan entre sí se denomina multicolinealidad de acuerdo con Fernandez, Córdova y Cordero (2002), Cuando dos de las variables independientes son completamente dependientes entre sí, es decir, su coeficiente de correlación lineal simple $r_{xx} = 1$, entonces se dice que la multicolinealidad es perfecta. Debido a la dependencia completa entre dos variables independientes, el sistema de ecuaciones de mínimos cuadrados se convierte en un sistema indeterminadamente compatible porque una o más ecuaciones son combinaciones lineales de las otras variables, creando un sistema con número infinitesimal de raíces y hay multicolinealidad imperfecta cuando el determinante está formado por coeficientes de correlación lineal simples distintos de 0. Si el determinante de las correlaciones (R_x) es 0, entonces hay multicolinealidad. La multicolinealidad perfecta parece difícil. Sin embargo, no es difícil que dos o más variables independientes aparezcan con coeficientes de correlación elevados. En estos casos, la ecuación de regresión se puede calcular matemáticamente, pero los cálculos estadísticos serán poco confiables porque es imposible distinguir el efecto de cada variable independiente sobre la variable dependiente.

6.1.4. Análisis discriminante

Para identificar variables según Díaz y Morales (2012) en base a su pertenencia a uno de varios grupos (poblaciones) predefinidos, dicho individuo debe ser asignado a una de estas variables, dependiendo de la información que posea. La técnica del análisis discriminante proporciona los requisitos y criterios para tomar esta decisión.

6.1.5. Análisis de correlación de conformidad

Díaz y Morales (2012) mencionan que este análisis busca una relación lineal entre un conjunto de variables predictivas y un conjunto de criterios medidos u observados. Se prueban dos combinaciones lineales, una para variables predictoras y otra para variables de criterio (dependientes). El análisis normativo puede extenderse a más de dos grupos.

6.1.6. Testing

Para poder determinar qué modelo se adapta mejor a la situación partícula del banco, es necesario realizar pruebas para tal motivo existen métodos de poder comprobar el modelo que podrá predecir con mayor éxito la probabilidad de default.

6.1.7. A/B Testing

El objetivo del A/B *testing*, es conducir el experimento de manera controlada, redireccionando de acuerdo con pros y contras en la valoración de la información

Según explica Dong y Liu (2018) su nombre se debe a que el método consiste en dividir en dos secciones (A y B) la información que recolecta aleatoriamente, y la estructura proporcionalmente para poder determinar si es mejor mantener el rumbo que se lleva en cierta situación, o bien cambiar de curso, o de elección. Es decir, este algoritmo nos ayuda a comprender si los cambios que estamos pensando hacer, serán beneficiosos o no, por lo tanto, ayuda a determinar que modelos de *Machine Learning* resultan más eficientes para el tipo de análisis requerido, Esta es la manera en que se lleva a cabo la prueba A/B:

El primer paso al ejecutar una prueba A / B es determinar el resultado que se desea lograr y elegir la métrica con la que se medirá el progreso con respecto a ese resultado. En la experimentación, esta métrica a veces se denomina Criterio de evaluación general u OEC. Frecuentemente, la OEC es una métrica aproximada del resultado deseado, en lugar de una medición directa del mismo. Se trabaja con aproximación ya que se consigue ese resultado en menor tiempo requerido. Un OEC que podría medirse en el orden de horas o días nos permite integrar rápidamente la retroalimentación experimental.

El segundo paso es determinar los parámetros del experimento en sí. Los dos parámetros que debe determinar son los tamaños de las muestras (cómo se dividen los usuarios entre los grupos de control y de tratamiento) y la duración del experimento. La forma en que los usuarios se dividen nos dirá qué usuarios verán el nuevo modelo de aprendizaje automático y qué usuarios seguirán viendo el modelo implementado actualmente.

La duración del experimento depende de un análisis de potencia (probabilidad de que obtenga un falso negativo) y el nivel de significación

(probabilidad de no rechazar la hipótesis nula cuando es verdadera, es decir, probabilidad de falsos positivos).

El seguir estos pasos, nos asegurará llevar a cabo correctamente el análisis A/B, para determinar el rumbo que convenga seguir, teniendo un mayor grado de certeza del éxito del experimento.

6.1.8. Matriz de confusión

Prosiguiendo con el análisis de datos, y la clasificación de estos, se llega a la Matriz de Confusión, herramienta que según Perner (2013) “mide la calidad de los trabajos de clasificación, y lo realiza mediante una visión general de las asignaciones correctas (en la diagonal) y las incorrectas (errores de omisión y comisión en los valores fuera de la diagonal”. Como lo menciona De los Santos (2018) Es una herramienta que mide:

6.1.9. Exactitud

La exactitud expresa qué tan cerca está la medición del valor real. Estadísticamente, la exactitud está relacionada con el sesgo de predicción. Estos se conocen verdaderos positivos. Se expresa como la proporción de verdaderos positivos predichos por el algoritmo para todos los casos positivos. En la práctica, la precisión es el número de predicciones positivas correctas.

$$\frac{(VP + VN)}{(VP + FP + FN + VN)}$$

(Ec.1)

Donde:

- VP: Verdadero positivo

- FP: Falso Positivo
- VN: Verdadero negativo
- FN: Falso Negativo

La precisión se refiere a la distribución de un conjunto de valores obtenidos a partir de mediciones de amplitud repetidas. Cuantas menos discordancias, mejor será la precisión. Está representado por la relación entre el número de predicciones correctas (positivas y negativas) y el número total de predicciones.

En forma práctica es el porcentaje de casos positivos detectados.

Se calcula como:

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

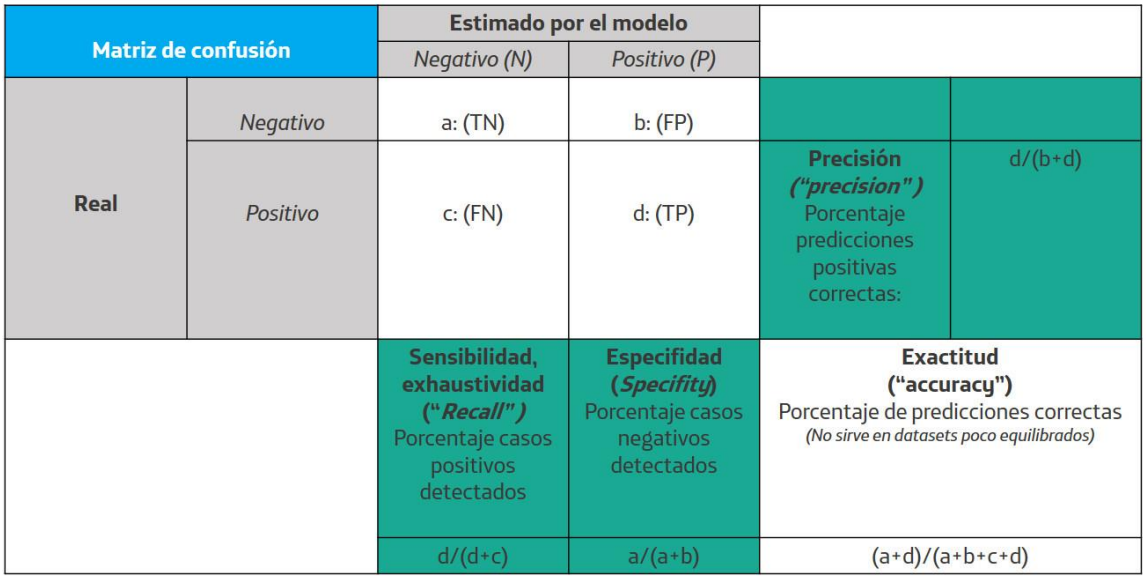
(Ec.2)

Se calcula como:

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

(Ec.3)

Figura 1. **Matriz de confusión**



Matriz de confusión		Estimado por el modelo			
		Negativo (N)	*Positivo (P)*		
Real	*Negativo*	a: (TN)	b: (FP)		
	Positivo	c: (FN)	d: (TP)	**Precisión** ***("precision")*** Porcentaje predicciones positivas correctas:	d/(b+d)
		Sensibilidad, exhaustividad ***("Recall")*** Porcentaje casos positivos detectados	**Especificidad** (***Specifity***) Porcentaje casos negativos detectados	**Exactitud ("accuracy")** Porcentaje de predicciones correctas *(No sirve en datasets poco equilibrados)*	
		d/(d+c)	a/(a+b)	(a+d)/(a+b+c+d)	

Fuente: *de los Santos (2018), Matriz de confusión y métricas asociadas .* Consultado el 27 de agosto de 2022. Recuperado de https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/.

Tal y como se puede observar, la Matriz de confusión, se puede emplear en el campo de la inteligencia artificial y en el problema de la clasificación estadística, para visualizar el desempeño de un algoritmo que se emplea en aprendizaje supervisado.

6.2. Gestión de riesgos

La gestión de crédito es una actividad imprescindible en los bancos y las instituciones financieras, los préstamos constituyen una gran parte de los activos estas instituciones indican el volumen de actividad crediticia. dado el impacto de una cartera crediticia morosa, administrar y mantener el crecimiento sostenido es un desafío para los gerentes de instituciones financieras. En términos de liquidez,

capacidad de otorgar créditos, utilidad y rentabilidad, se deben utilizar estrategias para aumentar la rentabilidad y reducir el riesgo. Un sistema bancario fuerte es esencial para una economía saludable del país, según Baldwin y Mason (1983), la contingencia en un banco es cuando el negocio cuenta con problemas para cumplir con sus obligaciones financieras. Por su parte Arko, Samuel Kofi (2012) indican que el riesgo crediticio es la fuente más grande de riesgo para las instituciones financieras, por lo que una gestión crediticia eficaz es fundamental para garantizar el desarrollo de operaciones de préstamo bancario eficaces; Poudel (2012) indica que la gestión del riesgo crediticio sigue siendo un predictor importante del desempeño de la economía del país y de los bancos del sistema. Además de ser un elemento imprescindible para el rendimiento esperado por los accionistas. Los bancos otorgan crédito a las personas con mayor probabilidad de pago, pero existe siempre una cantidad de sujetos de crédito que incumplen con sus obligaciones, lo que genera deudas incobrables que afectan el desempeño general de estos préstamos.

6.2.1. Tipos de riesgo en la banca

Según la Junta Monetaria de Guatemala (2006), el concepto de riesgo suele ser un evento o situación incierta que, de ocurrir, afecta las operaciones, rentabilidad o integridad de la institución. En otras palabras, el riesgo es la posibilidad o amenaza de daño, lesión, responsabilidad, pérdida u otro evento indeseable causado por una debilidad externa o interna que se puede evitar o mitigar tomando precauciones. Los siguientes tipos de riesgos bancarios han sido destacados por el regulador bancario guatemalteco:

- Riesgo de crédito: Se refiere a la posibilidad de sufrir pérdidas si el deudor o la contraparte no cumple con sus obligaciones en los términos pactados.

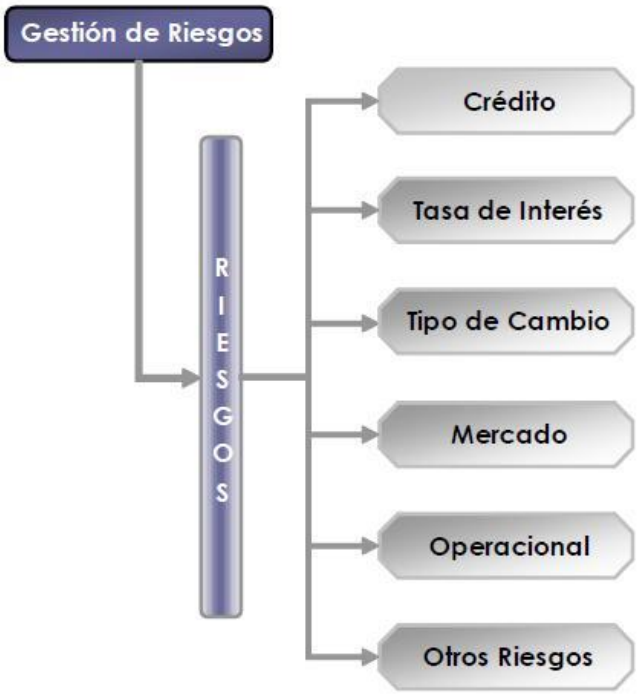
- Riesgo de tasa de interés: la posibilidad de sufrir pérdidas en el futuro como consecuencia de diferencias en términos de actividades de repreciación y cambios adversos en los tipos de interés.

- Riesgo de tasa de cambio: la existencia de posiciones en moneda extranjera y la posibilidad de sufrir pérdidas en el futuro por tipos de cambio desfavorables de determinadas divisas.

- Riesgo de mercado: se refiere a la posibilidad de perder posiciones en balance y fuera de balance por cambios en los precios de mercado.

- Riesgo operacional: pérdidas potenciales por inadecuación o falla de procesos, personas, sistemas internos o eventos inducidos por fuerzas externas.

Figura 2. **Tipos de riesgos bancarios**



Fuente: Super Intendencia de Bancos de Guatemala (2014). Gestión de Riesgos. Consultado el 13 de octubre de 2022. Recuperado de https://www.sib.gob.gt/web/sib/sbr/enfoque/riesgos/.

6.2.2. Crédito

Los bancos juegan un papel muy importante en todos los países del mundo. De acuerdo con la Junta Monetaria de Guatemala (2005) La principal función del sistema bancario es transferir el excedente creado de las personas individuales o jurídicas y transferirlo a las personas que necesiten de esta liquidez para desarrollar sus actividades comerciales o de consumo, a esta actividad se le denomina Intermediación. El sistema financiero incluye mercados e intermediarios financieros. Los mercados financieros actúan como "intermediarios" que conectan unidades de pérdidas y ganancias para beneficio

mutuo. En este proceso, el riesgo de crédito lo asumen los intermediarios, no las personas que colocaron en el mercado sus excedentes de liquidez, de acuerdo con Suresh y Paul, (2018). Los intermediarios financieros también garantizan la liquidez de las unidades sobrantes recibidas. Dejan sus ahorros y nuevamente reducen el riesgo con menores costos de información.

6.2.3. Clasificación de créditos según garantías

Los bancos en busca de cubrir sus pérdidas acuden a la solicitud de garantías en caso de incumplimiento de pago, ya que estos cubren una parte del riesgo los bancos deciden otorgar una tasa de interés de acuerdo con rapidez de la liquidez del bien puesto en garantía, En Guatemala, según la Junta Monetaria (2005), los préstamos se clasifican por tipo de garantía, las cuales son:

- Fiduciaria: Es el que presta el propio deudor o varios deudores. La obligación de pago se establece formalmente mediante una firma, como un endoso o fianza.

- Hipotecaria: Implica la pignoración de bienes inmuebles como terrenos, fincas o casas para que los deudores puedan pagar sus deudas.

- Prendaria: Incluye bienes muebles que están disponibles para pagar el crédito, por ejemplo: automóviles, menaje, joyas, entre otros.

- Inmobiliario: Incluye muebles, tales como: alquileres, deudores, etc.

- Garantía de fondos propios: Este préstamo se otorga cuando el prestatario tiene sus propios valores o bienes como garantía contra el incumplimiento del préstamo solicitado.

6.2.4. Análisis de riesgo crediticio

Para poder incrementar la rentabilidad en los bancos hay que tener un control de los créditos que se otorgan, con esto en mente (Samaniego, 2008) precisa que existen tres factores que permiten medir el riesgo de crédito: probabilidad de incumplimiento, exposición y severidad o tasa de recuperación.

- La probabilidad de incumplimiento se interpreta como la de que el prestamista eluda sus responsabilidades contractuales.

- La exposición es el valor de pérdida en el que se incurre en el momento del incumplimiento de la contraparte.

- Severidad es el porcentaje de pérdida resultante luego del incumplimiento y la recuperación, para lo cual se puede tener en cuenta las garantías que pueda presentar el cliente al momento de solicitar el crédito, el cual puede mitigar el impacto de pérdidas.

Mecanismos de contención de riesgo contemplan crear parámetros y modelos de evaluación estándar, Estos modelos se basan en comparar las probabilidades de incumplimiento con los solicitantes en función de las ponderaciones de las características cualitativas y cuantitativas del cliente. Tales como historial de pagos en créditos propios o del sistema bancario. Perfilar de forma exitosa puede reducir el riesgo de incumplimiento debido a la evaluación antes mencionada. Esto permite que las instituciones financieras eviten daños proporcionando más que una serie de créditos. Para poder mejorar la predicción en los mecanismos de autorización de crédito del banco, existen diferentes métodos, mismos que son indicados por el ente regulador de bancos en Guatemala, Superintendencia de Bancos define el riesgo de crédito como la

probabilidad de pérdidas como consecuencia de que un prestatario o contraparte incumpla sus obligaciones en los términos acordados, para poder mitigar este riesgo el ente regulador indica evaluar como mínimo los siguientes aspectos para conceder un crédito:

Tabla I. **Análisis de requisitos para concesión de créditos**

Análisis financiero	Empresariales mayores	Otros solicitantes
Comportamiento financiero histórico.	X	X
Capacidad de generar flujos de fondos suficientes	X	
Capacidad de pago		X
Experiencia de pago en la institución y en otras instituciones	X	X
Relación entre el servicio de deuda y los flujos de fondos proyectados del solicitante	X	X
Nivel de endeudamiento	X	X
Relación entre el monto del activo crediticio y el valor de las garantías	X	X

Fuente: Junta Monetaria de Guatemala (2005). *Reglamento para la Administración Integral de Riesgo*.

Dada la sencillez de los requisitos mínimos se vuelve necesario diseñar modelos predictivos que mejoren el desempeño de la cartera de créditos. El presente estudio aborda los créditos de consumo ya que son el giro principal del banco, para tal motivo se puede definir los Créditos de consumo según la Junta Monetaria de Guatemala (2005) como aquellos “activos crediticios que en su conjunto no sean mayores de tres millones de quetzales (Q3,000,000.00), si fuera en moneda nacional, o no sean mayores al equivalente de trescientos noventa mil dólares de los Estados Unidos de América (US$390,000.00), o su equivalente, si se trata de moneda extranjera, otorgados a una sola persona

individual destinados a financiar la adquisición de bienes de consumo o atender el pago de servicios o de gastos no relacionados con una actividad empresarial.”

Complementario a esta información requerida por el ente regulador se puede aplicar un modelo de calificación crediticia el cual puede expresarse conceptualmente como “métodos estadísticos para clasificar a los solicitantes de crédito, cuantificando el riesgo de los buenos y de los malos créditos” de acuerdo con (Hand y Henley, 1997).

6.2.5. Ponderación de créditos

Para poder implementar un modelo de calificación crediticia existen diferentes métodos, el presente estudio se enfocará en utilizar herramientas de minería de datos y *machine learning* para poder implementar un modelo de calificación crediticia.

La analítica de datos predictivos se ha utilizado para diversas funciones, de acuerdo con (Kelleher, MacNamee y D'Arcy, 2015) se describe como:

“El arte de construir y utilizar modelos que realizan predicciones basadas en la extracción de patrones de datos históricos.” Sus aplicaciones incluyen:

- Predicción de precios
- Predicción de dosis de medicamentos
- Evaluación de riesgos, entre otros

Cada una de las aplicaciones tienen en común, que en cada caso se utiliza un modelo para realizar predicciones que ayudan para tomar una decisión, y en algunos casos la predicción podría mostrar un aspecto en términos de tiempo,

pero no sucede en todos los casos. Otra aplicación que se denota en común es que las predicciones se basan en datos históricos.

6.3. Ciencia de datos

Tal como lo menciona Hand, Mannila y Smyth (2001) la minería de datos se trata del “proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos.” Además, toma como base los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El propósito general de la minería de datos se enfoca en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.

Consiste en un conjunto de técnicas matemáticas, estadísticas y computacionales que permiten realizar diferentes análisis de datos y el desarrollo de modelos descriptivos, combinados con métodos de las ciencias del comportamiento.

6.3.1. Machine learning

Continuando con el tema de analítica de datos, *machine learning* representa un gran avance, ya que es un proceso automatizado de extracción de patrones de datos; Tal como lo menciona (Kelleher, MacNamee, & D'Arcy, 2015, p.134):

“Machine learning supervisado, se utiliza para la realización de modelos, mediante técnicas en las que el modelo automáticamente detecta la relación existente entre las características descriptivas de un conjunto de datos y los datos objetivo, basándose en datos históricos o instancias.”

Es decir, se trata de un sistema de respuesta automatizado que puede emplear estrategias de aprendizaje para desarrollar o mejorar las capacidades de respuesta automatizada. Esto lo hace mediante un proceso en el que se toman las cualidades descriptivas de datos históricos, a las cuales se les clasifica mediante un orden (*Query*) según el criterio que se tenga, para luego procesar estas cualidades mediante un modelo de predicción, que mostró las predicciones o resultados que sirvieron para tomar una decisión.

Dicha información sugiere que, las estrategias de aprendizaje automático toman en cuenta la selección de comunicaciones como oportunidades de aprendizaje para mejorar e incrementar las capacidades de respuesta automatizada se fundamenta en criterios de selección (el propósito de esto es asegurar que el sistema no aprenda de ejemplos poco confiables o insignificantes).

6.3.2. Transformación de datos

Para poder modelar es necesario contar con datos creados aptos para cada uno de los modelos, en este sentido existen cierto tipo de transformaciones que pueden mejorar el desempeño de los modelos, tales como lo pueden ser:

6.3.2.1. Oversampling

En lo que respecta a los conjuntos de datos, estos frecuentemente se encuentran desequilibrados, lo cual representa amplios dominios de aplicación en la minería de datos.

El o*versampling* es “un método visual de comparación por puntos. En este método se plotea una cuadricula comparativa por pares, de un grupo

seleccionado de atributos de interés, para luego, ambos en el grupo de datos y los datos mostrados” (Dong y Liu, 2018, p. 215)

Si el *oversampling* se realiza apropiadamente, se generan grupos minoritarios más grandes en el grupo de datos resultante, que muestran el resultado esperado. Es decir, toma el grupo de datos y lo subdivide en un mayor número de muestras.

6.3.2.2. Downsampling

Al contrario del *oversampling*, este tipo de muestreo se encarga de tomar el grupo de datos y reducirlo en muestras más pequeñas que grupo de datos original.

A continuación, se muestra gráficamente la diferencia entre ambos tipos de muestreo.

Figura 3. **Nivelación de muestreo**

Fuente: ResearchGate (2008). Differences between undersampling and oversampling. Consultado el 15 de octubre de 2022. Recuperado de https://www.researchgate.net/figure/Differences-between-undersampling-and-oversampling_fig1_341164819.

6.3.2.3. Selección de variables

Para poder evaluar un modelo se deben seleccionar las variables que se van a utilizar para que el modelo muestre la variable de interés, con sus respectivos resultados. El análisis de datos de alta dimensión es un desafío para los investigadores e ingenieros en los campos del aprendizaje automático y la minería de datos.

El propósito de seleccionar variables o *Feature Selection*, es “reducir el paquete de datos a su esencia, a lo que realmente interesa evaluar, sus características esenciales y cualidades” (Dong y Liu 2018, p. 216), Se divide en dos tipos principalmente:

- Filtro: Selecciona solo los atributos que se encuentran en la cima y cumplen ciertos criterios.

- Envolvente: Selecciona datos iterativamente, retroalimentándose cíclicamente, solo con los atributos que mejoran el desempeño del algoritmo.

Posteriormente al verificar la comprobación de los datos, y aplicar los diferentes métodos mencionados se puede aplicar métodos de aprendizaje de datos.

7. PROPUESTA DEL ÍNDICE DE CONTENIDOS

8. METODOLOGÍA

El estudio no experimental se llevará a cabo mediante diferentes métodos estadísticos para lo cual el banco en particular brindará las bases de datos necesarias. A continuación, se describen los métodos a utilizar y las variables que intervienen, así como los pasos necesarios para realizar la investigación.

8.1. Características del estudio

El **enfoque** del estudio propuesto es cuantitativo, ya que el estudio se enfocará en crear un modelo matemático-numérico que modele la probabilidad de impago de los clientes.

El **alcance** es correlacional, multivariado, dado que el estudio pretende crear un modelo de regresión logística basado en múltiples variables para determinar una probabilidad cuantitativa de impago.

El **diseño** adoptado será observacional, pues la información de riesgo crediticio se analizará en su estado original sin ninguna manipulación; además será transversal pues se estudiará la base general de clientes del banco, pues se analizará el comportamiento de pago o impago de los créditos.

8.2. Unidades de análisis

La población en estudio será estará constituida por clientes de la entidad bancaria sujeto de estudio, la cual se encuentra dividida en subpoblaciones dadas por depositantes y prestatarios, para este estudio únicamente se analizará a los prestatarios históricos de los años 2012-2022 de la cual se extraerán muestras de forma aleatoria simple, que serán estudiadas en su totalidad.

8.3. Variables

Tabla II. **Variables del estudio**

Variable	Definición teórica	Definición operativa
Identificación	Identificador del cliente. No tiene afectación en el modelo.	Obtenido de base de datos, variable numérica los valores se encuentran entre: 0 - ∞
Agencia	Consiste en el lugar en el que se realizó la solicitud del crédito por parte del cliente. La sucursal a la que pertenezca el cliente.	Obtenido de base de datos, los valores representan los códigos de la sucursal determinado por el banco, variable numérica de 10 dígitos los valores se encuentran entre: 0 - ∞
Tipo de crédito	Se refiere al tipo de crédito que solicita el cliente si es una línea de crédito revolvente o es un préstamo total.	Obtenido de base de datos, los valores son crédito o línea revolvente, cadena de valores.
Re-crédito	Esta variable indica si el crédito es obtenido para cancelar un crédito anterior.	Obtenido de base de datos, variable dicotómica puede tomar los valores de Sí o No y se representará por medio de 1 para sí y 0 para No.
Monto	Consiste en el valor inicial de dinero otorgado al cliente en calidad de préstamo.	Obtenido de base de datos, valor numérico monetario representado en moneda Q., los datos son de escala de razón
Saldo	Es el saldo de otros prestamos del cliente al día de la solicitud.	Obtenido de base de datos, representan en las dimensionales de: Q., los datos son de escala de razón.
Plazo	Es el tiempo que el cliente requiere para pagar el préstamo.	Obtenido de base de datos, variable numérica los datos son de escala de razón.

Continuación de la tabla II

Tasa de interés	Tasa otorgada de interés al cliente.	Obtenido de base de datos, variable numérica los datos son de escala de razón.
Cuota	El monto que el cliente pagará manualmente por su crédito.	Obtenido de base de datos, variable numérica los datos son de escala de razón.
Garantía	El valor que el cliente entrega como garantía de préstamo para mitigar el impacto en caso de impago.	Obtenido de base de datos, los valores representan el tipo de garantía que respalda el crédito.
Acciones	Detalla la cantidad de acciones del banco que el cliente tiene	Obtenido de base de datos, variable dicotómica puede tomar los valores de Sí o No y se representará por medio de 1 para sí y 0 para No.
Días de mora histórica anterior	Esta variable marca el máximo de días mora que el cliente a alcanzado en créditos anteriores.	Obtenido de base de datos, variable numérica los datos son de escala de razón, se mide en días.
Forma de pago	El banco cuenta con 2 formas de pago, descuento automático en nómina y pago voluntario en la agencia bancaría.	Obtenido de base de datos, cadena de valores con las siguientes los valores pueden ser: Voluntario o descuento automático.
Edad	Es el número de años que tiene el deudor. El rango esta entre 18 y 100.	Obtenido de base de datos, variable numérica los datos son de escala de razón y se representan en años.
Ocupación	Describe la actividad a la que se dedica el cliente.	Obtenido de base de datos, los valores son códigos numéricos que representan las diferentes profesiones.
Nivel educativo	Es el grado de escolaridad del deudor.	Obtenido de base de datos, los valores pueden tomar los valores de: sin educación, primaria, básica, técnica, grado y posgrado.

Continuación de la tabla II

Salario	Equivale a los ingresos mensuales recibidos por el cliente.	Obtenido de base de datos, variable numérica los datos son de escala de razón y se representan en Q.
Antigüedad laboral	Consiste en el tiempo que lleva el deudor vinculado como empleado de una empresa. Se mide en años y varía entre 0 y 46.	Obtenido de base de datos, variable numérica los datos son de escala de razón y se representan en años.
Estado civil	Describe si el cliente es soltero o casado.	Obtenido de base de datos, los valores pueden tomar los valores de soltero y casado, y se representará por medio de 1 para casado y 0 para soltero
Género	Se refiere al género del deudor, es decir, masculino o femenino.	Obtenido de base de datos, puede tomar los valores de hombre o mujer y se representará por medio de 1 para hombre y 0 para mujer.
Dependientes	Es el número de personas que dependen económicamente del deudor. Oscila entre 0 y 25	Obtenido de base de datos, variable numérica los datos son de escala de razón y se representan en cantidad de cargas (personas).
Tipo de vivienda	Determina si el cliente tiene propiedad inmueble o eroga una cantidad de dinero manualmente donde vive.	Obtenido de base de datos, los valores pueden ser: propia, alquilada, familiar y otro.
Tipo de contrato	Indica la relación laboral del cliente	Obtenido de base de datos, es una cadena de valores, los valores pueden ser: Término definido, término indefinido, de servicios, jubilado, desempleado.
Región	Determina el departamento en el cual el cliente reside	Obtenido de base de datos, los valores son códigos numéricos asignados por el banco a cada región.

Continuación de la tabla II

Municipio	Este indica la municipalidad donde el cliente reside	Obtenido de base de datos, los valores son códigos numéricos asignados por el banco a cada región.
Impago	Estado final del crédito	Obtenido de base de datos, variable dicotómica puede tomar los valores de Sí o No y se representará por medio de 1 para sí y 0 para No

Fuente: elaboración propia.

8.4. Fases del estudio

El siguiente estudio se desarrollará en 4 fases las cuales servirán para crear el modelo matemático propuesto.

8.4.1. Fase 1: Revisión de literatura

Se reunirá toda la información referente a riesgo crediticio, definiciones, gráficas, normas aplicables para enmarcar la dinámica de los clientes respecto a su comportamiento de pagos y los factores que influyen para que el cliente cese sus pagos en un determinado tiempo de la madurez de los créditos. También se reunirá información sobre los diferentes métodos para poder calificar a los clientes de forma cuantitativa de acuerdo con el riesgo que representan según sus características.

8.4.2. Fase 2: Gestión o recolección de la información

El banco proporcionará la base de datos históricos de los años 2012-2022, de los cuales se calculará una muestra aleatoria simple, para poder seleccionar la muestra se utilizará el programa estadística R en su versión 4.2.1 para Windows; al contar con la muestra seleccionada se aplicarán *querys* a la base de datos proporcionada para poder obtener la información de los sujetos seleccionados.

8.4.3. Fase 3: Análisis de información

El análisis de datos se realizará por medio del programa estadístico R en su versión 4.2.1 para Windows, la base de datos se dividirá en 2 subgrupos, un set de prueba y un set de entrenamiento mediante muestreo aleatorio estratificado de acuerdo con la variable dependiente. Para que el modelo cuente con suficientes datos de clientes que pagaron y no pagaron para entrenar y probar el modelo, al set de datos de entrenamiento se realizará la prueba de Wald, que determina la relación entre las variables respecto a la variable objetivo, así se creará un nuevo set de datos con las variables significativas, posteriormente se realizará un modelo de regresión logística multivariada con el nuevo set de datos. Para determinar la validez del modelo se aplicará el modelo generado al set de datos de prueba, con el objetivo de determinar la efectividad con datos que no se utilizaron para modelación, por último, se calculará el RMSE el cual servirá de igual forma para determinar el impacto económico por medio del monto de los créditos.

8.4.4. Fase 4: Interpretación de información

El modelo final será seleccionado solo con las variables significantes para el modelo por lo que se analizará la influencia de estas variables en la contingencia de impago de créditos, además de realizar un contraste respecto a las investigaciones anteriores y el marco teórico, de igual forma se emitirán recomendaciones para elaboración de políticas de riesgo crediticio basado en los resultados del modelo.

8.4.5. Fase 5: Informe final

Se redactará los resultados y la interpretación de los resultados propuesto en la fase anterior.

9. TÉCNICAS DE ANÁLISIS DE INFORMACIÓN

Las técnicas utilizadas para realizar el estudio se especificarán a continuación: Los datos serán proporcionados por la empresa y serán datos históricos de clientes que han obtenido por lo menos un crédito entre los años 2012 al 2022.

9.1. Regresión logística:

La regresión logística estima la probabilidad de que ocurra un evento (variable dependiente), en función de un conjunto de datos determinado (variables independientes). El resultado es una probabilidad de que el evento ocurra. La regresión logística se utiliza únicamente cuando la variable dependiente es binaria.

9.2. Muestreo aleatorio estratificado:

El muestreo estratificado es una técnica de muestreo aleatorio en la que los investigadores primero dividen el universo en subgrupos o estratos más pequeños según las características comunes de los participantes y luego seleccionan al azar de esos grupos para formar la muestra final.

9.3. Prueba de Wald:

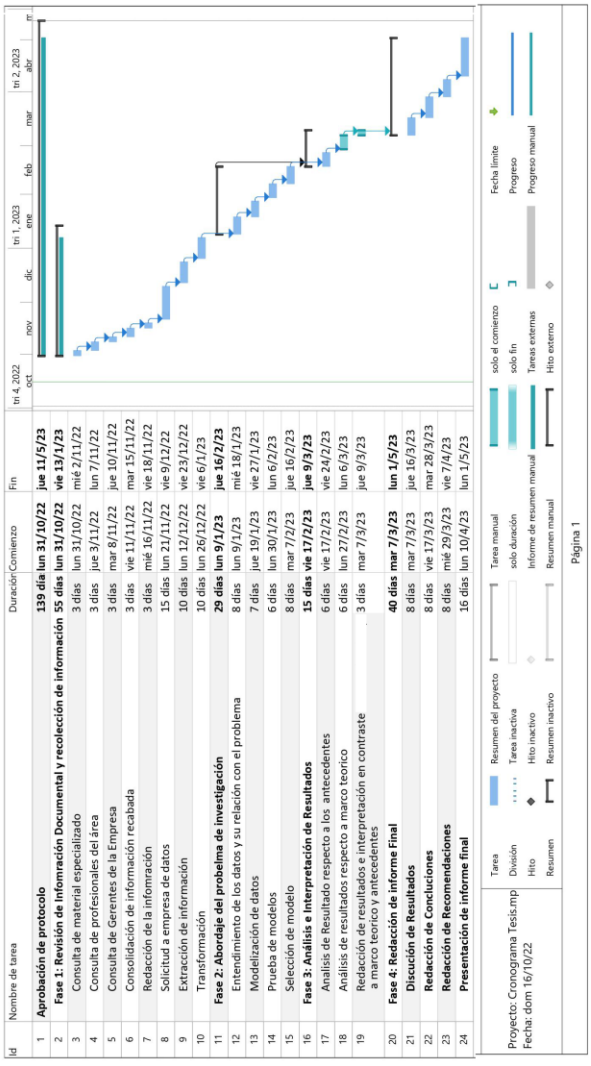
La prueba de Wald es una forma de determinar si una variable independiente en un modelo es significativa para explicar la variable dependiente, esta prueba permite prescindir de las variables no significativas.

9.4. Error cuadrático medio:

El error cuadrático medio conocido comúnmente como RMSE es una medida que determina la diferencia de un valor creado por un modelo y el valor real para los mismos parámetros. Este valor determina qué tan lejos están los puntos de datos de la línea de regresión; es decir indica qué tan concentrados están los datos alrededor de la línea óptima.

10. CRONOGRAMA

Figura 4. **Cronograma**



Id	Nombre de tarea	Duración	Comienzo	Fin
1	**Aprobación de protocolo**	**139 día**	**lun 31/10/22**	**jue 11/5/23**
2	**Fase 1: Revisión de Infomración Documental y recolección de información**	**55 días**	**lun 31/10/22**	**vie 13/1/23**
3	Consulta de material especializado	3 días	lun 31/10/22	mié 2/11/22
4	Consulta de profesionales del área	3 días	jue 3/11/22	lun 7/11/22
5	Consulta de Gerentes de la Empresa	3 días	mar 8/11/22	jue 10/11/22
6	Consolidación de información recabada	3 días	vie 11/11/22	mar 15/11/22
7	Redacción de la infomración	3 días	mié 16/11/22	vie 18/11/22
8	Solicitud a empresa de datos	15 días	lun 21/11/22	vie 9/12/22
9	Extracción de información	10 días	lun 12/12/22	vie 23/12/22
10	Transformación	10 días	lun 26/12/22	vie 6/1/23
11	**Fase 2: Abordaje del probelma de investigación**	**29 días**	**lun 9/1/23**	**jue 16/2/23**
12	Entendimiento de los datos y su relación con el problema	8 días	lun 9/1/23	mié 18/1/23
13	Modelización de datos	7 días	jue 19/1/23	vie 27/1/23
14	Prueba de modelos	6 días	lun 30/1/23	lun 6/2/23
15	Selección de modelo	8 días	mar 7/2/23	jue 16/2/23
16	**Fase 3: Análisis e Interpretación de Resultados**	**15 días**	**vie 17/2/23**	**jue 9/3/23**
17	Analisis de Resultado respecto a los antecedentes	6 días	vie 17/2/23	vie 24/2/23
18	Análisis de resultados respecto a marco teorico	6 días	lun 27/2/23	lun 6/3/23
19	Redacción de resultados e interpretación en contraste a marco teorico y antecedentes	3 días	mar 7/3/23	jue 9/3/23
20	**Fase 4: Redacción de informe Final**	**40 días**	**mar 7/3/23**	**lun 1/5/23**
21	**Discución de Resultados**	8 días	mar 7/3/23	jue 16/3/23
22	**Redacción de Concluciones**	8 días	vie 17/3/23	mar 28/3/23
23	**Redacción de Recomendaciones**	8 días	mié 29/3/23	vie 7/4/23
24	**Presentación de informe final**	16 días	lun 10/4/23	lun 1/5/23

Fuente: elaboración propia, realizado con MS Project.

11. FACTIBILIDAD DEL ESTUDIO

Para realizar el estudio se cuenta con los recursos humanos, financieros, tecnológicos, acceso a información, permisos, equipo, infraestructuras disponibles son suficientes para llevar a cabo la investigación por lo tanto es factible.

11.1. Recurso humano

Para este estudio se requerirá del investigador, un asesor y un revisor de tesis

11.2. Recursos financieros

A continuación, se detallan los recursos financieros necesarios para poder llevar a cabo el estudio, mismos que serán cubiertos en su totalidad por parte del investigador.

Tabla III. **Presupuesto de gastos**

Elemento	Unidad	Costo Unitario/(Q.)	Cantidad necesaria	Costo /(Q.)
FASE: Recolección de datos				
Transporte	Gal	40	3	120
Almacén seguro de información	Unitario	300	1	300
Tiempo del investigador	Hr	40	16	640
FASE: Análisis y redacción				
Internet	Mes	400	4	1600
Electricidad	KWH	1	300	300
Uso de equipo de computo	Mes	50	4	200
Tiempo del investigador	Hr	40	300	12000
Contingencias	Unitario	1000	1	1000
Total				**16160**

Fuente: elaboración propia.

11.3. Recursos tecnológicos

Para la elaboración del estudio se utilizará los siguientes softwares:

- Microsoft Windows 10
- SQL Server Management Studio versión 18.12.1
- R versión 4.2.1

11.4. Acceso a información y permisos

La información que se utilizará en este estudio será proporcionada por la empresa sujeto de estudio, por lo que cuenta con autorización, acuerdo de

confidencialidad, publicación, divulgación y permisos necesarios para este estudio.

11.5. Equipo e infraestructura

Se utilizará un equipo de cómputo y un almacenamiento seguro para la información de la empresa.

REFERENCIAS

1. Alvarez, R., 1995. Estadística multivariante y no paramétrica con SPSS. AbeBooks. Recuperado el 27 de agosto de 2022, de: https://www.abebooks.com/9788479781804/Estad%C3%ADstica-multivariante-param%C3%A9dica-SPSS-ALVAREZ-8479781807/plp [Accessed September 18, 2022].

2. Andrade, R. (2020) Deep Generative Models in Credit Scoring.The Arctic University of Nowrway. Oslo, Noruega.

3. Argomaniz, L. (2019) On Credit Score Models. Universidade de Lisboa. Lisboa, Portugal.

4. Arko, S. & Kofi, (2012a). Determining the causes and impact of non-performing loans on the operations of microfinance institutions: A case of Sinapi Aba Trust. Recuperado el 13 de octubre de 2022, desde: http://hdl.handle.net/123456789/4958/.

5. Baldwin, C. Y., & Mason, S. P. (1983). The resolution of claims in financial distress the case of Massey Ferguson. The Journal of Finance, Estados Unidos.

6. De los Santos, P. (2018), Machine Learning a tu alcance: La matriz de confusión, Think Big, España, Recuperado el 27 de agosto de 2022, de: https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/.

7. Devia, A. (2015). Contribuciones al Análisis Estadístico. Universidade Da Coruña. Coruña, España.

8. Díaz L. & Morales M. (2012), Análisis Estadístico de Datos Multivariados, Universidad Nacional de Colombia.

9. Dong, G., & Liu, H. (2018). Feature Engineering for Machine Learning and Data Analytics. Arizona: CRC Press.

10. Fernández, S., Córdoba A. & Cordero J. (2002). Estadística Descriptiva, Madrid: ESIC.

11. Hajiyev R. (2021). Risks in The Banking Sector and Credit Risk Measurement Methods: Research On Azerbaijan Banking Sector. Università Ca' Foscari. Venezia, Italia.

12. Hand, D. y Henley, W., (1997). Statistical classification methods in consumer calificación crediticia: A review. Royal Statistical Society.

13. Jaramillo M. (2021), Machine Learning para la Estimación del Riesgo de Crédito en una Cartera de Consumo. Universidad EAFIT.Medellín, Colombia.

14. Junta Monetaria (2005), Reglamento para la Administración del Riesgo de Crédito, Anexo a la Resolución JM 93-2005, Guatemala.

15. Junta Monetaria (2006). Reglamento para la Administración Integral del Riesgo. Resolución JM-54-2006 y sus reformas. Guatemala.

16. Kelleher, J., MacNamee, B. & D'Arcy, A., (2015). Fundamentals of Machine Learning For Predictive Data Analytics. Inglaterra: The MIT Press.

17. Montalván, C. (2019) Credit Scoring, Aplicando Técnicas de Regresión Logística y Redes.

18. Navarrete, A. (2017). Riesgo De Crédito Y Credit Default Swaps. Universidad de Sevilla. Sevilla, España.

19. Neuronales, para una Cartera de Microcrédito, Universidad Andina Simón Bolívar. Quito, Ecuador.

20. Oroes, M. (2014). Descriptive and Inferential Statistics. Lulu.

21. Padmalatha Suresh and Justin Paul (2018). Management of Banking and Financial Services. Pearson Education. India.

22. Perner, P. (2013), Machine Learning and Data Mining in Pattern Recognition, Springer.

23. Poudel, R.P. (2012). The impact of credit risk management on the financial performance of commercial banks in Nepal, International Journal of Arts and Commerce, Recuperado el 13 de octubre de 2022, desde: https://ijac.org.uk/images/frontImages/gallery/Vol._1_No._5/2.pdf.

24. Prasad N. (2020), Impact of Credit Risk Management on The Profitability of The Irish Banks. National College of Ireland. Dublín, Irlanda.

25. Samaniego Medina, R. (2008). El riesgo de crédito en el marco del Acuerdo de Basilea II. Madrid: Delta Publicaciones Universitarias.

26. Super Intendencia de Bancos. (2014) Gestión de Riesgos. Recuperado el 13 de octubre de 2022, desde https://www.sib.gob.gt/web/sib/sbr/enfoque/riesgos/.

27. Torre, V. (2018) Analysis of PD-LGD Correlation Effects on The Minimum Capital Requirement. Universitat de Barcelona. Barcelona, España.

ANEXO

Tabla IV. **Matriz de coherencia**

ELEMENTOS	PROBLEMA DE INVESTIGACIÓN (Vacíos de conocimiento) problema estadístico	PREGUNTAS DE INVESTIGACIÓN
GENERAL O CENTRAL	Se desconoce si un modelo estadístico puede ayudar a reducir la cantidad de créditos otorgados a clientes que pagarán sus créditos.	¿Cuál es el modelo estadístico describe la probabilidad de impago de crédito?
ESPECÍFICOS O AUXILIARES	01.Se desconoce las características de los clientes que tienen una mayor incidencia en la probabilidad de que no realicen sus pagos	01. ¿Cuáles son las características del cliente que pueden ayudar a identificar su probabilidad de impago?
ESPECÍFICOS O AUXILIARES	02. Se desconoce si la cantidad de datos históricos almacenados en el banco es suficiente para poder elaborar un modelo estadístico de probabilidad	02. ¿Qué cantidad de créditos históricos es necesario para poder crear un modelo estadístico?
ESPECÍFICOS O AUXILIARES	03. Se desconocen las implicaciones que puedan causar evaluar a clientes por medio de un modelo estadístico de probabilidad de impago de créditos.	03. ¿Qué medidas preventivas hay que tomar para ejecutar un modelo que minimice el riesgo y genere confiabilidad?

Continuación de la tabla IV

ELEMENTOS	OBJETIVOS	PROCEDIMIENTO Y TÉCNICAS ESTADÍSTICAS	METODOLOGÍA
GENERAL O CENTRAL	Proponer un modelo estadístico que permita identificar, medir y pronosticar la probabilidad de impago de operaciones de crédito de un banco del sistema guatemalteco.	Regresión logística multivariada	El enfoque es de tipo cuantitativo regresivo multivariable que buscará determinar la probabilidad de impago de un cliente.
ESPECÍFICOS O AUXILIARES	01. Determinar las variables cuantitativas y cualitativas de mayor incidencia sobre la probabilidad de impago.	Prueba de Wald	Listar de las variables que influyen en que una persona sea propensa a pagar sus deudas.
ESPECÍFICOS O AUXILIARES	02. Diseñar una modelo que capture adecuadamente la probabilidad de impago de los clientes del banco.	Muestreo aleatorio estratificado	Determinar la cantidad de créditos históricos necesario para llevar a cabo el estudio, mediante técnicas de muestreo

Continuación de la tabla IV

ESPECÍFICOS O AUXILIARES	03. Determinar los posibles riesgos y potencial de mejora que pueda implicar la implementación del modelo.	Error medio cuadrado (RMSE)	Determinar la rentabilidad del modelo mediante el error medio cuadrado y el potencial de mejora que tiene el modelo respecto a el actual método de otorgamiento de créditos.

Fuente: elaboración propia.