UNIVERSIDAD TECNOLÓGICA DE LA MIXTECA

ANÁLISIS DE RIESGO CREDITICIO PARA LAS SOCAP UTILIZANDO APRENDIZAJE AUTOMÁTICO

TESIS
PARA OBTENER EL GRADO DE MAESTRÍA EN TECNOLOGÍAS DE CÓMPUTO APLICADO

PRESENTA:
ERWIS MELCHOR PÉREZ

DIRECTOR DE TESIS:
DR. AGUSTIN SANTIAGO ALVARADO

CO-DIRECTOR DE TESIS:
MTCA. MOISÉS EMMANUEL RAMÍREZ GUZMÁN

Huajuapan de León, Oaxaca, Diciembre 2023

Dedicado a mi esposa Karina,
a mis padres Rodrigo y Maria
por todo el apoyo que me han brindado.

Dedicado a mi hermana Citlali
y mis sobrinos Gonzalo Rodrigo y
Gianna Maricruz

Agradecimientos

Agradezco infinitamente a la mujer, amiga y compañera más importante de mi vida, mi esposa Karina, siendo mi apoyo y motivación constante.

De igual manera agradezco a mis padres, hermana y mis sobrinos que han estado apoyándome en cada etapa de mi carrera profesional.

Agradezco a la Universidad Tecnológica de la Mixteca, por brindarme la oportunidad y el espacio para realizar mis estudios de posgrado. Así mismo, agradezco al Consejo Nacional de Humanidades, Ciencia y Tecnología (CONAHCyT) por la beca otorgada para la realización de mis estudios de Maestría bajo el número de CVU 1150059. A cada uno de los profesores que me impartieron clases y esos amigos que formé durante mi paso por la maestría. En especial al Dr. Iván, quien fue mi tutor y constantemente se mantuvo pendiente de mi progreso durante la maestría y animarme a realizar una estancia en el INAOE.

Este trabajo no habría sido posible sin la aceptación del Dr. Agustín como su tesista, el apoyo constante del profesor Moi quien dispuso de tiempo, consejos, paciencia y conocimientos para ser una base en mi formación académica. A la Dra. Araceli, por aceptar ser mi asesora en el área de finanzas y motivarme a participar en congresos nacionales e internacionales exponiendo mi trabajo de tesis.

Finalmente, agradezco a mis sinodales, Dr. Eduardo Sánchez Soto, Dr. Christian Eduardo Millán Hernández, Dr. José Anibal Arias Aguilar y M.A. María Sánchez Zárate por el tiempo invertido en la revisión de este trabajo.

Acrónimos

AG Algorítmo genético.

CNBV Comisión Nacional Bancaria de Valores.

CONCAMEX Confederación de Cooperativas de Ahorro y Préstamo de México.

CVC Cartera Vencida.

CVG Cartera Vigente.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

DT Decision Trees.

GD Gradiente Descendente.

IA Inteligencia Artificial.

IMOR Índice de morosidad.

IV Valor de la Información.

LASSO Least absolute shrinkage and selection operator.

LRASCAP Ley para Regular las Actividades de las Sociedades Cooperativas de Ahorro y Préstamo.

MARS Multivariate adaptive regression splines.

MCC Coeficiente de Correlación de Matthews.

ML Machine Learning.

PE Pérdidas Esperadas.

PI Probabilidad de Incumplimiento.

RF Random Forest.

RL Regresión Logística.

RN Red Neuronal.

RNA Redes Neuronales Artificiales.

SFM Sistema Financiero Mexicano.

SOCAP Sociedades Cooperativas de Ahorro y Préstamo.

SP Severidad de la Pérdida.

SVM Máquinas de Vector de Soporte.

WoE Peso de la Información.

Índice general

Índice de figuras

Índice de tablas

Resumen

En la actualidad, la principal actividad económica de las Sociedades Cooperativas de Ahorro y Préstamo (SOCAP) son los otorgamientos de créditos a sus socios. Al momento de analizar una solicitud de crédito se calcula la probabilidad de incumplimiento o *score* crediticio, siendo éste el principal factor que determina el otorgamiento del crédito.

La correcta aprobación o negación de las solicitudes de créditos es importante para la salud financiera de las SOCAP. Este proceso requiere de un análisis e interpretación minucioso de los resultados. La identificación de los atributos y su relación para generar un modelo matemático se realiza utilizando datos históricos almacenados sobre los socios que han solicitado créditos.

El presente trabajo de investigación se enfoca en la construcción de una base de datos y el desarrollo de una biblioteca que contiene funcionalidades para el preprocesamiento de los datos usando técnicas del peso de evidencia y la ganancia de la información (*WoE IV)*, selección de características y *clustering* y la utilización de clasificadores para predecir el incumplimiento de las solicitudes de crédito en la entidad financiera EIZ con presencia en el Estado de Oaxaca.

La metodología aplicada a los conjuntos de datos presenta el mejor rendimiento en cuanto a las métricas de error tipo I y II, *accuracy* y coeficiente de correlación de Matthews (*MCC*). Los clasificadores propuestos en el presente trabajo tienen como prioridad minimizar el error tipo II, ya que este representa al segmento de personas que potencialmente se les podría aprobar un crédito y no pagarían en tiempo y forma. La minimización de este error permite a las microfinancieras tener finanzas más sanas para así poder llevar los servicios de financiamiento a segmentos de población a los que no llegan grandes instituciones financieras como los bancos.

Capítulo 1

Introducción

1.1. Introducción

Las Sociedades Cooperativas de Ahorro y Préstamo (SOCAP) son las instituciones dirigidas al sector social, sin ánimo especulativo y sin fines de lucro que, conforme a la Ley General de Sociedades Cooperativas y de la Ley de Ahorro y Crédito Popular, tienen como objetivo principal el realizar operaciones de ahorro y préstamo con sus socios (Condusef, 2021).

Una de las actividades económicas principales de las SOCAP es el otorgamiento de créditos. Se entiende por crédito a la adquisición de un monto de dinero a disposición de una persona física o moral, el cual puede ser adquirido a través de las instituciones bancarias y financieras en México, este trabajo de investigación se centra exclusivamente en el análisis del riesgo crediticio de personas físicas presentes en las SOCAP en el Estado de Oaxaca, las cuales se encuentran autorizadas por la Comisión Nacional Bancaria de Valores (CNBV) (CNBV, 2021b). Para el trabajo de investigación se realiza el convenio con la institución financiera Esperanza Indígena Zapoteca SC de AP de RL de CV (denominada EIZ), para la construcción de una base de datos que pueda funcionar para predecir el riesgo crediticio.

En nuestro país, los créditos se encuentran clasificados en: consumo, comercial y vivienda[1]. La asignación de un crédito comprende en la solicitud de un nuevo socio o uno con historial en la entidad. La autorización deberá permitir y asegurar el cumplimiento en tiempo y forma el pago del crédito. Dicha aprobación se basa en la tarea de analizar las características o parámetros definidos por cada institución financiera a partir de las entidades reguladoras y que el solicitante cumpla con ellas.

La evaluación del riesgo de cualquier préstamo monetario se realiza calculando la probabilidad de incumplimiento, esta es una estimación para indicar si el contratante podrá o no terminar el contrato en tiempo y forma (García et al., 2016). Este problema ha sido tratado por diferentes investigaciones como es el caso de Solarte y Cerezo (2018), donde se analizan los resultados del uso de los modelos de análisis discriminante, regresión logística y redes neuronales para la clasificación de las solicitudes de créditos.

En el presente trabajo de investigación se implementó una biblioteca para clasificar solicitudes de crédito considerando los datos generales, económicos, historial crediticio y de actividades económicas de los solicitantes. El objetivo de esta biblioteca es que pueda ser una herramienta para apoyar en la toma de decisiones a las SOCAP (EIZ) al momento de deliberar si una persona es susceptible a un préstamo.

[1] Anexo C SOCAP `https://www.cnbv.gob.mx/Anexos/Anexo%20C%20SOCAP.pdf`

1.2. Planteamiento del problema

Las entidades financieras han ido incorporando el uso de las tecnologías de la información y servicios para brindar un mejor servicio a sus usuarios. El servicio de créditos no es la excepción, donde se han incorporado los sistemas inteligentes para evaluar si un cliente califica o no para la aprobación de un préstamo.

Franco y Chang (2017) plantean que la morosidad es un riesgo que cualquier institución financiera enfrenta. Una de las principales razones de la insolvencia y la permanencia de las instituciones financieras en el mercado es el número elevado de créditos que presentan días de mora e incumplimiento de pago (Cabrera-Cruz, 2014).

En el trabajo de investigación desarrollado por Rimarachín y Sánchez (2018) concluyen que la morosidad de un crédito se encuentra íntimamente relacionada con el riesgo crediticio de las instituciones financieras. Mientras que Oblitas et al. (2021) describen el inconveniente presentado al no estar realizando la calificación ni evaluación de manera adecuada de los créditos, presentando un enfoque directo en los indicadores de morosidad.

La importancia en la identificación del riesgo crediticio ha aumentado considerablemente con el paso del tiempo, permitiendo reducir las pérdidas y el aumento de gastos dentro del sector financiero, ya que es la causante de afectar la salud financiera de las entidades que emiten los créditos. Cabrera-Cruz (2014) define al *credit scoring* como un sistema que evalúa automáticamente el riesgo asociado con cada solicitud de crédito. Rayo Cantón et al. (2010) lo definen como un sistema que determina la probabilidad del incumplimiento del pago de un crédito que es otorgada a una persona. En las instituciones financieras, el procedimiento de identificación se realiza normalmente de manera manual, ya que la adquisición de sistemas para solventar esta necesidad tiene costos muy elevados, por lo que resultan de difícil acceso para algunas pequeñas entidades financieras del país.

La morosidad es un tema de gran importancia para mantener una buena salud financiera Delgado et al. (2020), se considera a la morosidad como la deuda que tienen los socios de las entidades financieras, pues a través de ella se permite medir su desempeño y salud financiera de las instituciones. Al cierre del segundo trimestre del año 2021, el Estado de Oaxaca resultó ser una de las entidades federativas con mayor Índice de Morosidad[2] (IMOR) con un 8.41 %, siendo el segundo Estado con mayor índice de morosidad del país (CNBV, 2021a).

La importancia de este trabajo se aprecia en la Tabla 1.1 y en la Figura 1.1, donde se puede observar el comportamiento del IMOR en el cuarto trimestre de los años 2017, 2018, 2019, 2020 y tercer trimestre del 2021 (Apéndice A para ver relación de Cartera Vigente CVG vs Cartera Vencida CVC), en donde se puede observar que la mayoría de las SOCAP con presencia de operación en el Estado de Oaxaca muestran un incremento en su IMOR reportado ante la CNBV (CNBV, 2021a).

[2]Calculado como: Cartera vencida / Cartera total, es el porcentaje de operaciones que han sido morosas en relación con la cantidad de préstamos concedidos por la institución financiera (CNBV, 2021a)

Tabla 1.1: Instituciones con IMOR en el Estado de Oaxaca. Fuente: (CNBV, 2021a)

Nombre de la entidad	**IMOR %**				
	dic 2017	dic 2018	dic 2019	dic 2020	sep 2021
Caja Popular Mexicana, S.C. de A.P. de R.L. de C.V.	2.69	2.68	3.24	4.54	5.59
Finagam, S.C. DE A.P. DE R.L. DE C.V.	7.70	7.06	4.44	13.62	18.16
Cooperativa Acreimex, S.C. de A.P. de R.L. de C.V.	5.49	4.74	8.10	3.57	3.30
Cooperativa Yolomecatl, S.C. de A.P. de R.L. de C.V.	4.55	4.52	3.85	3.59	3.46
Cooperativa Lachao, S.C. de A.P. de R.L. de C.V.	0.92	1.48	2.27	1.94	2.33
Caja Solidaria San Dionisio Ocotepec, S.C. de A.P. de R.L. de C.V.	12.23	1.00	1.47	3.23	5.48
Esperanza Indígena Zapoteca, S.C. de A.P. de R.L. de C.V.	4.39	6.04	3.62	4.04	5.10

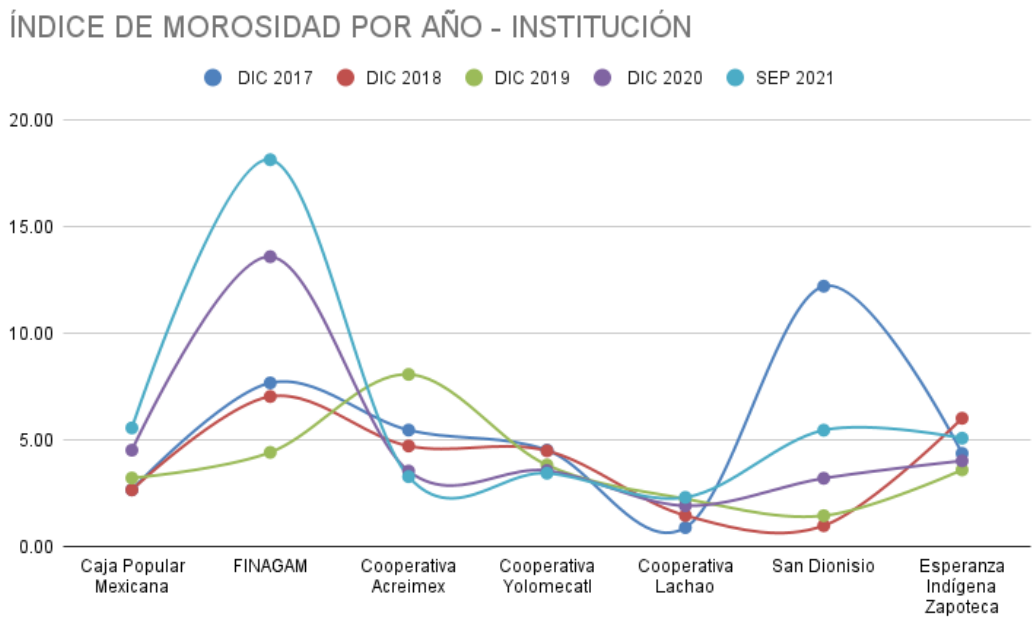


Figura 1.1: Índice de morosidad por Institución 2017-2021. CNBV. (2021a).

Todo lo anterior motiva el desarrollo de esta investigación, cuyo objetivo principal es proponer la implementación de un modelo basado en métodos de aprendizaje automático que permita minimizar el riesgo crediticio al identificar aquellos solicitantes que tengan una liquidez potencial dentro de las entidades financieras. Por lo que este proyecto pretende abrir las puertas a la integración de sistemas expertos utilizando Inteligencia Artificial (IA) a las SOCAP del Estado de Oaxaca.

1.3. Trabajos relacionados

Las SOCAP presentan un problema al momento de realizar la identificación de las solicitantes que son aptos para el otorgamiento de un crédito a través del análisis e interpretación de sus datos personales, bancarios y económicos, aplicando los lineamientos establecidos por la regulación vigente.

Las Redes Neuronales Artificiales (RNA) han mostrado resultados prometedores con el paso del tiempo, como principal objetivo es el simular el comportamiento del cerebro humano, siendo capaz de adquirir conocimiento a través de la experiencia. Angelini et al. (2008) proponen el uso de una *RNA* compuesta por dos capas ocultas y una capa de salida, para la evaluación del riesgo crediticio enfocado a 76 pequeñas empresas Italianas, con datos entre los años de 2001 y 2003, para cada registro se tiene 15 campos de los cuales 8 provienen de sus estados financieros y los otros 7 son datos históricos con el banco proveedor. Durante el preprocesamiento del conjunto de datos se identificaron algunos datos faltantes o en formato erróneo sobre las empresas solicitantes. Por lo tanto, fue necesario eliminar aquellos datos incorrectos y normalizar los datos al máximo valor de entrada.

En el trabajo realizad por Ghodselahi y Amirmadhi (2011) proponen un modelo híbrido para el análisis del riesgo crediticio a partir de la base de datos de origen alemán *Statlog (German Credit Data)* (*GCD*) (Hofmann, 1994), el conjunto de datos se encuentra conformado por 20 atributos, de los cuales 7 son numéricos y 13 no numéricos. Dentro de los atributos se encuentran el historial crediticio, saldo de cuenta, propósito del préstamo, situación laboral, edad, vivienda, trabajo, entre otros. Los clasificadores utilizados fueron las Máquinas de Soporte Vectorial (SVM, por sus siglas en inglés), RNA y Árboles de Decisión (DT, por sus siglas en inglés).

Marcano-Cedeno et al. (2011) recomiendan el uso de diferentes RNAs para abordar el problema; sin embargo, las tasas de error obtenidas suelen ser elevadas, con esto proponen la aplicación de un algoritmo de entrenamiento de RNAs inspirado en la propiedad biológica de metaplasticidad de las neuronas. Este algoritmo es especialmente eficiente cuando se dispone de pocos patrones de una clase, o cuando la información se encuentra incompleta. Utilizando dos conjuntos de datos, una Australiana con 15 atributos y una base de datos *GCD*.

Amézquita Reyes et al. (2012) proponen el desarrollo de un modelo de *credit scoring* para la toma de decisiones al momento de realizar el otorgamiento de microcréditos basado en un sistema difuso-genético, que permita predecir la probabilidad del riesgo de incumplimiento de los clientes, teniendo en cuenta las características específicas de cada solicitante. Concluyendo que el modelo propuesto logra ser más preciso y se acerca al comportamiento histórico crediticio de los clientes. Utilizando una base de datos con un total de 60,749 registros, con 59 atributos, como lo son ingresos, edad, gasto, actividad económica, estado civil, género, escolaridad, estados financieros y ventas.

Oreski et al. (2012) en Croacia, proponen utilizar los datos históricos de los bancos para predecir la capacidad del cliente para pagar el préstamo en tiempo y forma. Utilizando *AG* para seleccionar las características más significativas al momento de realizar la clasificación del riesgo crediticio mediante una *RNA*. Con la necesidad que surge para predecir aquellas personas que son susceptibles a créditos, García et al. (2017) proponen realizar el procedimiento por iteraciones al modelo predictivo de incumplimiento usado por las entidades regulatorias como la CNBV. Con una base de datos de 43,323 registros y 6 atributos, los cuales son: variables de cumplimiento/incumplimiento, número de impagos, historial crediticio, meses del crédito, solvencia y límite del crédito.

En la tesis de Ramos Martinez (2017) se propone un método para generar modelos de clasificación de créditos, utilizando AG para determinar el modelo de puntuaciones y *clustering* jerárquico aglomerativo para la segmentación de grupos de riesgo. Utilizando un total de 1459 pequeñas empresas, considerando los datos del estado financiero como lo son: pasivos, activos, ventas y utilidades. La desventaja de esta propuesta es que la aplicación permite al usuario manipular los datos de tal manera que pueda obtener una salida satisfactoria.

Solarte y Cerezo (2018) muestran la aplicación y la efectividad de tres clasificadores para la evaluación de las solicitudes de créditos. Los clasificadores evaluados fueron el modelo de análisis discriminante, Regresión Logística (RL) y las RN. Los resultados obtenidos muestran la superioridad de las RN obteniendo un rendimiento del 86.9 % durante la tarea de clasificación. Para el experimento fue utilizado una base de datos conformada por 673 registros de una entidad financiera, utilizando los atributos de estado civil, edad, género, actividad económica, tipo vivienda, dependientes económicos, plazo del préstamo, ingresos, línea de crédito y garantía.

Barajas-Juárez et al. (2019) desarrollan un *software* aplicando un AG para identificar si un individuo puede ser factible o no de un préstamo bancario, considerando los atributos propuestos por el banco; entre las variables analizadas destacan: edad, ocupación, nivel educativo, estado civil, género y tipo de vivienda. Estas variables coinciden con las analizadas por la mayoría de las entidades financieras en México. Cabe mencionar que cada entidad realiza su análisis de riesgo crediticio con parámetros que ellos definen y en relación con la información solicitada por las entidades reguladoras.

Ossa Giraldo et al. (2021) muestran el análisis de desempeño entre los clasificadores: *RL, Random Forest* (*RF*, por sus siglas en inglés), *SVM* y *Multi-Layer Perceptron* (*MLP*, por sus siglas en inglés) para la evaluación del riesgo crediticio. De los clasificadores utilizados, *RF* es el clasificador que presenta el mayor *accuracy* como medida de rendimiento. Para el desarrollo fue utilizado una base de datos correspondiente a 15,060 registros con 18 variables, entre los cuales se encuentran: edad, monto del crédito, plazo, ingresos, gastos, residencia, tiempo en trabajo, etc.

Freire-López (2021) propone desarrollar un modelo de clasificación de clientes en *Insotec* que permita disminuir el riesgo crediticio y mejore los tiempos de respuesta a las solicitudes. Obteniendo resultados de precisión de los algoritmos, evidenciando el éxito de *RF* con una precisión de efectividad de un 97.2 % y una tasa de error de un 2.8 %. Utilizando una base de datos clasificada en buenos y malos pagadores, compuesto por 18 variables y 63,896 registros. Dentro los cuales se encuentran: tipo de préstamo, año y mes de última transacción, género, estado civil, grado de estudio, edad, patrimonio, ingreso y egresos.

Una particularidad de estos trabajos relacionados es que han realizado la evaluación del riesgo crediticio en bancos y para la implementación procesan y analizan las bases de datos de instituciones bancarias alemanas, italianas y croatas, haciendo el uso de técnicas de aprendizaje automático con datos de empresas pequeñas. Es importante mencionar que la mayoría de los trabajos citados realizan la implementación de RNA, AG y RL, enfocadas y llevadas a implementar en Bancos y las solicitudes de créditos por parte de las pequeñas empresas, dichas entidades bancarias realizan el análisis en relación con los parámetros que ellos mismos tienen definidos, dentro los más significativos se pueden encontrar: edad, género, estado civil, ocupación, vivienda, dependientes económicos y gastos. Siendo esta la razón principal por el cual se pretende trabajar con entidades financieras del Estado de Oaxaca, ayudando de esta manera a mejorar la salud financiera de las entidades involucradas en el desarrollo del trabajo de investigación.

En el presente proyecto de investigación se pretende implementar técnicas de aprendizaje automático para el análisis de riesgo crediticio en las SOCAP del Estado de Oaxaca, teniendo como beneficio ayudar a la toma de una mejor decisión al momento de deliberar un préstamo, teniendo en cuenta las variables y características consideradas para la evaluación de los solicitantes de créditos en la población del Estado.

1.4. Justificación

La elaboración de un modelo basado en técnicas de aprendizaje automático, el cual permitirá ayudar a las entidades financieras del Estado de Oaxaca a reducir él IMOR en los productos de préstamos otorgados y reducir el monto de cartera vencida en las instituciones. El presente trabajo de investigación se enfoca en las SOCAP que tienen presencia de operación en el Estado de Oaxaca, ayudando a la evaluación de las solicitudes de crédito y la prevención del riesgo de incumplimiento de pago.

El desarrollo de esta investigación se efectúa con el propósito de ayudar a reducir el número de solicitudes con riesgo de incumplimiento, utilizando la información proporcionada por la entidad; dichos datos fueron utilizados para crear un conjunto de datos y entrenar un modelo de aprendizaje automático.

En la actualidad, las SOCAP del Estado de Oaxaca, carecen de sistemas inteligentes que permitan la identificación de socios candidatos a la asignación de un crédito. Esta carencia impide que el personal humano tenga un fundamento y respaldo del que se encuentra realizando una buena selección de los socios, lo que ha incrementado el aumento del IMOR, afectando la salud financiera de la entidad. Una ventaja adicional es que el tiempo de respuesta para el análisis de una solicitud de crédito podría reducirse a la captura de la misma y la ejecución del modelo para obtener una respuesta de manera inmediata a diferencia del proceso tradicional que lleva días u horas.

El propósito de las SOCAP es promover la Inclusión Financiera de las personas en las comunidades donde operan, brindando productos y servicios financieros de calidad que ayuden a mejorar su situación económica y cooperar con el gobierno federal en su difusión, implementación y gestión de los programas de apoyo que promueven CNBV (2021b).

Cabe aclarar que existen muchas propuestas que abarcan este tema de investigación, sin embargo, estos se concentran en analizar datos de bancos extremadamente grandes y de carácter internacional encargadas de proporcionar créditos a pequeñas empresas.

1.5. Hipótesis

Al aplicar técnicas de aprendizaje automático se podrá generar un modelo de clasificación eficaz para predecir el incumplimiento de las solicitudes de crédito para las SOCAP del Estado de Oaxaca.

1.6. Objetivos

1.6.1. Objetivo general

Implementar una biblioteca con modelos de clasificación basados en técnicas de aprendizaje automático que permitan predecir el incumplimiento de pago a las solicitudes de crédito en las

SOCAP en el Estado de Oaxaca.

1.6.2. Objetivos específicos

1. Revisar el estado del arte relacionado con modelos para el análisis y predicción del riesgo crediticio.
2. Examinar información referente a técnicas de aprendizaje automático aplicadas a problemas de clasificación.
3. Identificar bibliotecas para desarrollo de técnicas de aprendizaje automático.
4. Generar una base de datos de socios para su análisis.
5. Realizar la propuesta de un modelo que permita la implementación de una técnica de aprendizaje automático.
6. Identificar aquellas variables que permitan determinar el riesgo de incumplimiento de una persona al solicitar un crédito.
7. Establecer las métricas que serán usadas para validar los resultados obtenidos del modelo propuesto.
8. Implementar un modelo clasificación para la identificación del riesgo crediticio en las solicitudes de crédito en institución que es objeto de estudio.
9. Aplicar y analizar los resultados de la implementación de un modelo de aprendizaje automático para clasificación utilizando los datos facilitados por la SOCAP.

1.7. Metas

1. Reporte de trabajos relacionados con modelos generados para calcular el riesgo crediticio en el sector financiero aplicando técnicas de aprendizaje automático.
2. Investigación relacionada con el análisis de riesgo crediticio, detección de atributos y métricas de rendimiento.
3. Reporte de herramientas de análisis de datos y bibliotecas de aprendizaje automático para la implementación de los algoritmos propuestos.
4. Creación de una biblioteca utilizando un modelo de clasificación basada en técnicas de aprendizaje automático para la identificación del riesgo crediticio en las SOCAP del Estado de Oaxaca.
5. Reporte comparativo del desempeño de los modelos de aprendizaje automático implementados.
6. Elaboración del documento de tesis.
7. Publicación de un artículo arbitrado.

1.8. Metodología

La metodología ocupada en el desarrollo de este proyecto contiene los pasos que se muestran en la **Figura 1.2**:

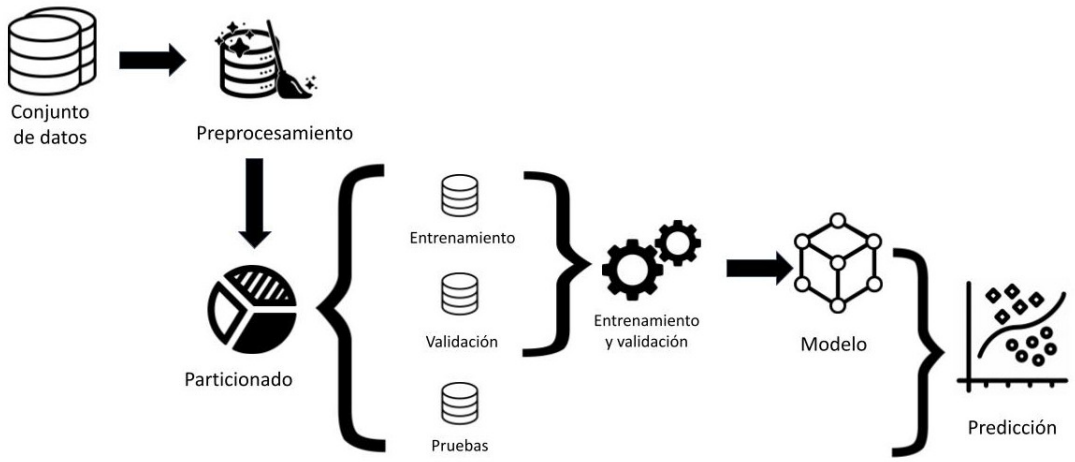


Figura 1.2: Representación de la metodología a usar durante el desarrollo de la tesis

En el primer paso se obtiene información relevante de la base de datos, proporcionada por la SOCAP a través de un convenio de colaboración y confidencialidad.

Como segundo paso, se contempla el **Preprocesamiento** de los datos, en donde se normaliza, selecciona, limpia y destacan las características que pueden ser útiles para los algoritmos a implementar.

El tercer paso llamado **Particionado** consiste en la división de los datos en conjuntos de entrenamiento, validación y pruebas de manera aleatoria; los conjuntos de entrenamiento y validación son utilizados para el ajuste de los hiperparámetros de los clasificadores y el conjunto de pruebas es para obtener las métricas de rendimiento de los clasificadores.

En el cuarto paso consiste en realizar el **Entrenamiento** de los clasificadores utilizando el conjunto de datos de entrenamiento, mediante la validación cruzada y la búsqueda en malla para identificar los hiperparámetros que puedan generar el mejor rendimiento. El quinto paso denominado **Validación y comparación** de los resultados obtenidos por los clasificadores, en caso de que el rendimiento no sea satisfactorio, el proceso se vuelve a ejecutar para realizar ajustes de los hiperparámetros.

Como paso final se realiza la evaluación del rendimiento de los clasificadores con el conjunto de pruebas. La metodología planteada se aborda con mayor detalle en el Capítulo 3.

1.9. Alcances y limitaciones

Esta tesis contempla la utilización de técnicas de aprendizaje automático, que permitan predecir el incumplimiento de una solicitud de crédito a partir del uso de un modelo creado utilizando la base de datos proporcionada por la SOCAP. El conjunto de datos creado para desarrollar los modelos está limitado por los registros e información (políticas y reglas aplicadas durante la evaluación de solicitudes) proporcionados por la entidad financiera a través de un convenio de colaboración.

Capítulo 2

Marco teórico

2.1. Introducción

Las aplicaciones que hacen uso del aprendizaje computacional han incrementado en diversas áreas de la ciencia como son: medicina, computación, finanzas, entre otras. Diversos autores han implementado algoritmos sofisticados de aprendizaje automático que permiten evaluar con mayor precisión el perfil crediticio de los solicitantes en las entidades financieras, permitiendo identificar señales de riesgo que podrían pasar desapercibidas con enfoques tradicionales para tomar decisiones más informadas, y así poder otorgar créditos a aquellos individuos con menor riesgo de morosidad. Los beneficiarios del uso de esta tecnología son las instituciones financieras, al mitigar pérdidas, y los solicitantes al promover prácticas crediticias más responsables.

En este capítulo se realiza una breve descripción de las instituciones de crédito, riesgo crediticio y modelos de aprendizaje automático. Se revisa la importancia de las SOCAP en la economía mexicana, así como la evaluación del riesgo crediticio.

2.2. La importancia de las SOCAP en la economía mexicana

Las SOCAP, también son conocidas como cajas populares o cajas de ahorro. Surgen en México a principios de la década de los 50's y promovidas por la Iglesia Católica, teniendo como principal objetivo impactar de forma positiva en las condiciones de vida las familias mexicanas, para obtener préstamos con tasas bajas de interés.

En México, existe un rezago de la inclusión financiera, especialmente en algunas poblaciones rurales y semiurbanas (Rovirosa et al., 2015). Demirgüç-Kunt y Singer (2017) consideran que la inclusión financiera significa que los adultos tengan acceso a los servicios financieros y puedan utilizarlos de manera efectiva. Estos servicios deben de ofrecerse al consumidor de manera responsable, segura y en un ambiente regulado. Por lo tanto, las SOCAP juegan un papel muy importante, estas se encuentran comprometidas a llevar a cabo la inclusión financiera, que puede definirse como el acceso de las empresas y los hogares a servicios financieros apropiados que satisfagan sus necesidades (Lázaro y Sosa, 2020).

2.2.1. Inclusión Financiera en México

La CNBV define la inclusión financiera como el acceso y uso de servicios financieros formales con una regulación adecuada que garantice la protección de los datos del consumidor y promueva la educación financiera para mejorar las oportunidades en la mayoría de la población (CNBV, 2016). Esta definición describe cuatro componentes fundamentales:

- **Acceso:** Describe la infraestructura disponible para brindar servicios y productos, es decir, los canales de acceso entre las instituciones financieras y los ciudadanos.
- **Uso:** Se refiere a la adquisición de productos o servicios financieros por parte del público y la frecuencia con la que son utilizados. Esto se refiere a la demanda, comportamiento y las necesidades de la población.
- **Protección y defensa al consumidor:** Se refiere a que los productos y servicios financieros, ya sean nuevos o existentes, dentro de un marco que al menos garantice la transparencia de la información, manejo justo y mecanismos efectivos para resolver quejas y asesorar a los clientes sobre prácticas desleales y fraudulentas. Además, el objetivo del marco regulatorio es promover la inclusión financiera y la protección de los datos personales de los usuarios.
- **Educación financiera:** Consiste en que la población adquiera aptitudes, habilidades y conocimientos que permitan a la población a gestionar y planificar adecuadamente sus finanzas personales, evaluar diferentes productos y servicios financieros para la toma de decisiones de acuerdo a sus intereses y la selección de productos que satisfagan sus necesidades y la responsabilidad que implica el uso de estos servicios.

2.2.2. El papel del crédito en la economía

En México, como en otras partes del mundo, las cooperativas de ahorro y préstamos se han convertido en empresas productivas y una alternativa financiera para que la población en situación económica más desfavorable, pueda incrementar sus ingresos, la creación de empleos, promoviendo la inclusión social, protegiendo a los socios y además de ofrecer servicios al resto de la sociedad, de esta manera promueven sistemas financieros que beneficien al cooperativismo, la cual consiste en la captación del ahorro y en el pago puntual de los préstamos (Lázaro y Sosa, 2020).

Los créditos otorgados por las entidades cooperativas permiten a las personas realizar pagos, adquirir productos, invertir en bienes o iniciar un negocio. De la misma manera, pueden solventar una emergencia o cubrir gastos de imprevistos (Credimejora, 2022). La adquisición del primer crédito inicia a generar un historial crediticio, reflejado en las centrales de información financiera, como el Buró de Crédito. Con el paso del tiempo, esto se vuelve la "carta de presentación" de una persona ante las entidades financieras, ya que todos los datos sobre su comportamiento financiero se registran para crear informes y evaluar la generación de puntajes. Mientras mejor sea el cumplimiento de los pagos, estas entidades financieras estarán dispuestas a continuar con la otorgación de algún crédito.

2.2.3. Instituciones de crédito

Una institución de crédito se define como una organización que se especializa en adquirir recursos de sus clientes y el otorgamiento de crédito. Este grupo incluye los bancos, cajas de ahorro, microfinancieras y cooperativas de ahorro y préstamos BBVA (2022). Así mismo, son partícipes del mercado financiero al captar recurso público para luego invertirlos en activos como depósitos bancarios, valores y títulos.

Tipos de instituciones de crédito

Westreicher (2018) clasifica a las entidades de crédito con base en la forma de captar recursos y entregar créditos, entre los cuales se encuentran:

- **Bancos:** Son instituciones cuya actividad principal es la administración del capital y ofrecer distintas clases de financiamiento para el uso de las personas físicas y de las empresas.
- **Cajas de ahorro y préstamos:** Dentro esta clasificación se encuentran las SOCAP que realizan las mismas actividades que los bancos. Las personas, pequeñas y medianas empresasn son el público objetivo de las SOCAP. Además, su organización y forma de operación varía en función de la legislación vigente de cada país.
- **Entidades de dinero electrónico:** Se trata de instituciones que emiten dinero que se encuentra almacenado en un medio virtual y es aceptado por otras empresas distintas al emisor como medio de pago.

2.2.4. SOCAP

Una cooperativa es una asociación autónoma de personas que se unen voluntariamente para satisfacer sus necesidades y aspiraciones económicas, sociales y culturales comunes debidamente controlada (CONCAMEX, 2021).

Las SOCAP son aquellas sociedades creadas y organizadas con el principal propósito de realizar operaciones de ahorro y préstamo, formando parte del Sistema Financiero Mexicano (SFM) e integrantes del sector popular, sin ánimo especulativo, sin fines de lucro y se encuentran organizadas conforme a la Ley General de Sociedades Cooperativas (Condusef, 2021).

De acuerdo con la Confederación de Cooperativas de Ahorro y Préstamo de México (CONCAMEX), las SOCAP son instituciones que tienen como objetivo principal contribuir con la Inclusión Financiera de la población donde operan. Su meta es proporcionar productos y servicios financieros de calidad; como créditos, ahorros e inversiones, que ayuden a mejorar la situación económica de las personas. Además, estas instituciones también colaboran con el Gobierno Federal en la difusión, entrega y administración de los programas de apoyo que este promueva (CONCAMEX, 2021).

La Ley para Regular las Actividades de las Sociedades Cooperativas de Ahorro y Préstamo (LRASCAP) es el marco legal que normaliza las SOCAP. Esta ley fue publicada en el Diario Oficial de la Federación el 13 de agosto de 2009 y reconoce a las SOCAP como parte del SFM con la característica de pertenecer al sector popular y sin ánimo especulativo. Además, se reconocen por no ser intermediarios financieros con fines de lucro (Condusef, 2021)

2.3. Análisis de riesgo crediticio

Brown y Moles (2014) definen el riesgo crediticio como la probabilidad que una parte contractual incumpla sus obligaciones al vencimiento del crédito, también es conocido como el riesgo de impago, refiriéndose al impago de un crédito en parte o en su totalidad. Saavedra García y Saavedra García (2010) consideran que la evaluación del riesgo de crédito consiste en calcular la probabilidad de que una persona incumpla con sus obligaciones del pago de un crédito en tiempo y forma.

J.-P. Li et al. (2020) mencionan que la evaluación del riesgo crediticio funciona como ayuda a las instituciones financieras a definir las políticas bancarias y estrategias comerciales. En particular, su objetivo es apoyar a los profesionales en el proceso de toma de decisiones sobre la asignación de préstamos a un solicitante basándose en diferentes parámetros.

2.3.1. Crédito y riesgo

El crédito se define como la transferencia de un bien actual a cambio de recibir un bien futuro. Además, también se puede definir como un acuerdo entre un acreedor y un deudor, donde el acreedor confía en las características del deudor para permitirle utilizar bienes y riqueza durante un periodo de tiempo determinado, con la posibilidad de recuperarlos al finalizar dicho plazo (Valdés, 2005).

Por otra parte, Hand y Henley (1997) definen al crédito como una cantidad de dinero que una institución financiera presta a un consumidor y que debe ser reembolsado, con intereses, en cuotas que generalmente se encuentran en intervalos regulares de tiempo. Siendo estos de plazo fijo (donde el monto solicitado y los intereses generados se pagarán al vencimiento del contrato) o pueden ser con pagos periódicos, así mismo puede incrementar el monto solicitado y la duración dependerá de las condiciones del préstamo.

La palabra riesgo tiene su origen en el latín *risicare*, que significa tener el valor de explorar por un sendero peligroso. De acuerdo con de Lara (2008), el riesgo es considerado una parte de los procesos involucrados en la toma de decisiones. En el ámbito de las finanzas, el concepto de riesgo se encuentra vinculado a las pérdidas que pueden experimentarse en los portafolios de inversión, haciendo referencia a la probabilidad de presentar una pérdida en el futuro. Además, en el sector financiero, se pueden encontrar distintos tipos de riesgos que se clasifican en las siguientes categorías:

- **Riesgo de mercado:** La pérdida que puede presentar un inversionista debido a la variación de los precios en el mercado o en factores de riesgo, como lo son las tasas de interés, tipos de cambio, etc.).
- **Riesgo de liquidez:** Se trata de las pérdidas que una institución puede enfrentar al necesitar más recursos para financiar sus activos. Mientras que Banxico (2022) define al riesgo de liquidez como la incapacidad de uno o varios participantes de un contrato de crédito para pagar su obligación en el momento, pero sí en una fecha posterior.
- **Riesgo operativo:** Se asocia con los errores en los procedimientos, en los modelos o en las personas responsables de ejecutar dichos sistemas. Así mismo, algunos problemas en la organización pueden estar relacionados con fraudes o por la falta de capacitación de los empleados.
- **Riesgo de reputación:** Se refiere a las posibles pérdidas que podrían ocurrir si no se logran concretar oportunidades de negocio debido a un daño a la reputación de la institución.
- **Riesgo de crédito:** El incumplimiento de una contraparte en una operación que implica un compromiso de pago. Banxico (2022) define al riesgo de crédito como la situación en que una de las partes involucradas de un contrato financiero no puede cumplir con sus obligaciones financieras, lo que ocasiona pérdidas para la otra parte del contrato.

Vargas Sánchez y Mostajo Castelú (2014) plantean que la evaluación del riesgo de crédito se enfoca en medición de las pérdidas relacionadas con las operaciones crediticias. El riesgo de crédito se divide en dos partes: la primera denominada como riesgo de impago, que es la probabilidad de que un prestatario incumpla con su responsabilidad de hacer los pagos establecidos en el contrato. El segundo componente es la severidad de la pérdida en caso de incumplimiento, es decir, la parte que el inversionista pierde.

El riesgo de crédito se define como la posibilidad de que una parte no cumpla con sus obligaciones según los términos acordados. El riesgo de crédito también se conoce como riesgo de rendimiento o riesgo de contraparte (Brown y Morales, 2008). Todos estos definen lo mismo: el impacto de los efectos del crédito en las operaciones de una empresa.

García et al. (2016) definen a la propiedad de incumplimiento como una medida de la probabilidad de que un acreditado deje de cumplir con sus obligaciones establecidas en el contrato de crédito. En relación con los lineamientos establecidos por la CNBV, algunas instituciones financieras en el país desarrollan sus propios modelos de *scoring*. Los factores que se consideran importantes al momento de realizar el monitoreo del riesgo en las instituciones financieras son (Vargas Sánchez y Mostajo Castelú, 2014) :

- **Severidad de la Pérdida (SP):** Es la pérdida que sufre un banco cuando el deudor no cumple con sus obligaciones, teniendo en cuenta todos los gastos implicados para recuperar el dinero.
- **Probabilidad de Incumplimiento (PI):** Es la probabilidad que un acreditado deje de cumplir con sus obligaciones estipuladas en el contrato crediticio. De acuerdo con la normativa vigente, es necesario realizar el cálculo de la estimación de la PI por crédito.
- **Pérdidas Esperadas (PE):** Se utiliza para determinar la cantidad promedio que se puede perder, y está asociado a la política de provisiones preventivas, la institución debe tomar medidas preventivas para evitar riesgos de créditos. Las provisiones que se presentan de acuerdo a los mandatos de la CNBV, a través del saldo insoluto y la calificación obtenida.

2.3.2. Modelos de *credit scoring*

El proceso de identificación del riesgo crediticio es el reconocimiento de todos los factores que, al presentar comportamientos adversos, originan un incremento de este. La manera más común de identificar el riesgo de crédito es mediante el uso de tecnologías que son llamadas *scoring*, las cuales complementan el análisis y sirven como herramienta de apoyo en la toma de decisiones (Vargas Sánchez y Mostajo Castelú, 2014).

Hand y Henley (1997) definen al *credit scoring* como los métodos estadísticos para categorizar a los solicitantes de créditos, incluyendo a aquellos con un historial crediticio en la institución financiera. Por otra parte, Medina y Selva (2013) denomina al *credit scoring* a todo sistema de evaluación que permite valorar de forma automática el riesgo asociado a cada solicitud de crédito. El riesgo de cada solicitud se encuentra en función de la solvencia del deudor, tipo de crédito, plazo y características del cliente. Además, se considera como un sistema automático, el cual permite reducir los costos y tiempo de evaluación de las solicitudes de crédito. Mientras que Fica et al. (2018) lo definen como aquellos procesos que proponen automatizar la gestión de créditos en cuanto a conceder o no una determinada operación crediticia sujeta a un conjunto de variables relevantes en la toma de decisiones.

2.3.3. Enfoques tradicionales de la medición del riesgo de crédito

Los métodos convencionales para evaluar el riesgo de crédito se basan en las variables económicas y financieras que afectan el rendimiento de la empresa durante el tiempo (Saavedra García y Saavedra García, 2010). Estos modelos implican la perspectiva que cada analista utiliza para evaluar según su experiencia en la asignación de créditos. Los modelos convencionales son: Basilea y las 5 C's del crédito.

2.3.3.1. Basilea I

En julio de 1988 se realizó la primera publicación del convenio de capital del Comité de Supervisión Bancaria de Basilea (BCBS, por sus siglas en inglés), también llamada Basilea I, con la finalidad de realizar la especificación de las bases del cálculo del capital necesario para cumplir los riesgos que enfrentaba el sistema bancario (Carrascal y María, 2015).

El BCBS reune a las autoridades regulatorias bancarias del mundo, con el fin de mejorar la regulación, supervisión y estándares bancarios para fortalecer la estabilidad financiera. En marzo del 2009, la CNBV se integró al BCBS para supervisar entidades financieras, lo que representó un reconocimiento internacional a la labor de modernización del marco legal, prácticas de regulatorias y supervisión bancaria en México (CNBV, 2022).

2.3.3.2. Las 5 Cs

Para la evaluación de una solicitud de crédito, el experto toma en consideración las normas, políticas y manuales de la institución, con base en ello otorga o rechaza la solicitud. Para otorgar un crédito se debe de realizar un análisis profundo de las características y condiciones de cada solicitante. Las 5 C's del crédito es una herramienta que debe ser utilizada en el análisis crediticio, permitiendo tomar la decisión de otorgar o no el crédito, así mismo las condiciones con las cuales deberá de otorgarse (Madrigal et al., 2017).Saavedra García y Saavedra García (2010) definen estos factores:

- **Capacidad:** Es la capacidad de pago que tiene el solicitante, teniendo como resultado de la evaluación la manera en que este terminará de pagar el préstamo, flujo de efectivo e historial crediticio.
- **Capital:** Consiste en analizar la situación financiera de quien recibe el crédito, lo que brindará información acerca de su capacidad de pago, fuente de ingresos y gastos, capacidad de asumir deudas y el tiempo promedio de pago.
- **Colateral:** Se refiere a los recursos que tiene el acreditado para garantizar la liquidación del crédito, es decir, las garantías. Mientras mayor sea la facilidad de conversión a efectivo, menor será el riesgo asociado al crédito.
- **Carácter:** Se refiere a los datos relacionados con los hábitos de pago y comportamiento en operaciones crediticias, considerando los siguientes elementos:
  - Historial crediticio.
  - Reporte de buró de crédito.
  - Antecedentes penales.
  - Referencias bancarias.
- **Condiciones:** Situación económica de la región y la organización, y pueda tener un impacto directo en la generación de ingresos para la empresa. Las condiciones se refieren a las de la entidad que emite el crédito y la persona que lo solicita. Dentro de los principales atributos con las que debe de cumplir el acreditado son:
  - Edad.
  - Ocupación.
  - Lugar de residencia.
  - Tiempo de residencia.

- Escolaridad.
- Estado civil.
- Género.
- Tipo de vivienda.

2.4. Preprocesamiento

El preprocesamiento de datos es fundamental al entrenar un modelo de ML porque garantiza que los datos de entrada estén limpios, consistentes y listos para ser utilizados de manera efectiva por el modelo, promoviendo la mejora del rendimiento; así mismo, facilitar la interpretación y capacidad de generalización ante nuevas entradas. Hernández et al. (2008) mencionan que el propósito principal del preprocesamiento de datos es corregir y estructurar los datos que serán la base de análisis o descubrimiento de conocimiento, en este caso los que serán utilizados por los modelos de aprendizaje automático.

La justificación de este proceso radica en que es común que los datos contengan errores o no tengan las estructuras adecuadas para facilitar su procesamiento, es común que existan valores inusuales e incoherencias por diferentes razones, entre las cuales se encuentran (Han et al., 2012):

- **Datos incompletos:** los datos carecen de valores de atributos y no se cuenta con el detalle de la información.
- **Ruido:** la información contiene errores o valores atípicos que se desvían de lo esperado.
- **Inconsistencias:** contienen discrepancias en los datos.

Aplicar algunas técnicas de preprocesamiento permite que los algoritmos de aprendizaje automático sean más eficientes. Por ejemplo, al aplicar técnicas como la reducción de la dimensionalidad, los algoritmos de aprendizaje podrían consumir menos recursos de cómputo en su entrenamiento y su efectividad podría mejorar (Kotsiantis et al., 2006).

Las técnicas de preprocesamiento de datos tienen por objetivo mejorar la calidad de los datos, ayudando de esta manera a incrementar la precisión y eficiencia de los procesos de análisis (Han et al., 2012). Dentro de las acciones más comunes están:

- **Limpieza de datos:** Esta actividad consiste en llenar los valores faltantes, identificar o remover los datos inconsistentes.
- **Integración de datos:** Se realiza la combinación de los datos desde múltiples fuentes.
- **Transformación de datos:** Los datos se transforman o combinan, implicando: la normalización, suavizado, agregación y obtener un conjunto de datos que sea de calidad.
- **Reducción de datos:** Se aplican técnicas para reducir el volumen de datos, para obtener una representación reducida de los datos, manteniendo la integridad de los datos originales.

Normalización de los datos

C. Li (2019) define a la normalización como la transformación de los datos, de tal forma que todos los atributos estén aproximadamente en la misma escala, de lo contrario las variables con mayor desviación estándar dominan sobre las que tengan una desviación estándar más baja durante el entrenamiento del modelo. Los métodos de normalización usados se describen a continuación:

- **Normalización *Min-Max*** La transformación de los datos se realiza aplicando el método *Min-Max* que se describe en la ecuación 2.1. Este método escala todos los atributos del conjunto de datos entre el rango de 0 y 1.

$$X_i = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{2.1}$$

  Donde X_i, X_{min}, X_{max} representan cada uno de los valores de entrada, los valores mínimos y máximos de los atributos del conjunto de datos, respectivamente.

- **Normalización *Z-score*** La estandarización de los datos se realiza mediante la ecuación 2.2.

$$Z = \frac{X - X_{mean}}{X_{stddev}} \tag{2.2}$$

  Donde X_{mean} y X_{stddev} son el promedio y la desviación estándar de cada uno de los atributos en el conjunto de datos. Siendo este tipo de normalización con la cual se obtiene el mejor rendimiento de los modelos utilizados.

2.4.1. Extracción y selección de características

En los conjuntos de datos analizados, uno de los problemas son la gran cantidad de características y una cantidad limitada de registros. Para abordar este problema, en los métodos de extracción de características es necesario descubrir una relación entre el espacio d-dimensional y un nuevo espacio k-dimensional, donde $k<=d$, con la menor pérdida de información. Es decir, el conjunto de datos en el espacio de características d y el k mantenga la misma información y minimizar el error de clasificación (Dougherty, 2012). En otras palabras, este proceso busca determinar un subconjunto de características que permitan mejorar el rendimiento del modelo y potencialmente reducir la complejidad del mismo.

K. Wang et al. (2022) definen la selección de características como un proceso de preprocesamiento, cuyo objetivo es encontrar un subconjunto de estas características dentro del conjunto original utilizando criterios de evaluación especificos. Dicha selección tiene por objetivo que el subconjunto obtenido sea lo más pequeño, manteniendo la misma calidad de los datos del conjunto original con la finalidad de mejorar la predicción del modelo clasificador implementado.

Zhou et al. (2021) definen a la selección de características como uno de los pasos esenciales para construir modelos de manera eficiente, con la finalidad de mejorar la fiabilidad, la generalización y reducir el sobreajuste. Es el procedimiento de elegir un subconjunto de características más significativas para utilizarlas en la evaluación del modelo.

Usualmente, para realizar la tarea de selección de características, se dividen en 3 categorías (Jović et al., 2015).

- **Empaquetado (*wrapper*)**: Para cada combinación de vector de características, la probabilidad de error de clasificación del clasificador tiene que ser estimada, y después se selecciona la combinación resultante con la probabilidad de error mínima.

- **Filtrado (*filter*)**: Se seleccionan características independientemente del algoritmo de modelado, después de encontrar las mejores características, los clasificadores pueden utilizarlas. Los métodos de filtrado pueden clasificar características individuales o evaluar subconjuntos enteros de características.

- **Embebido (*embedded*)**: La selección de características se realiza durante la ejecución del algoritmo de modelización. Entre los más comunes se encuentran algoritmos de *DT, random forest, XGBoost* pero también otros como los de regresión logística y sus variantes.

Los métodos para extracción y selección de características explorados para esta tesis se listan a continuación:

Least absolute shrinkage and selection operator (LASSO)

El método del Operador de Selección y Contracción Mínina Absoluta (*LASSO*, por sus siglas en inglés) recomendado por Tibshirani (1996), se puede representar mediante la reducción de la función de log-verosimilitud negativa sujeta a restricciones ponderadas, definida por la ecuación 2.3:

$$\sum_{i=1}^{n}[-Y_{i,y+12}(\beta_0 + \beta^{'} X_{i,t}) + \log(1 + \exp(\beta_0 + \beta^{'} X_{i,t}))] \tag{2.3}$$

sujeto a $\sum_{k=1}^{p} |\beta_k| \leq s$, en la ecuación 2.3 n representa el número de instancias y p es la cantidad de características predictoras usadas en el modelo. La frecuencia de la contracción puede ser controlada por el parámetro de sintonía de la penalización de la rigurosidad de s. LASSO elige a las características con mayor capacidad de predicción mediante la reducción hacia cero de algunos de sus coeficientes y la minimización de otros.

Multivariate adaptive regression splines (*MARS*)

J. H. Friedman (1991) propone un procedimiento flexible para organizar las relaciones existentes entre un conjunto de variables de entrada y el objetivo dependiente donde se involucran las interacciones con menos variables. MARS es un método estadístico no paramétrico basado en una estrategia de divide y vencerás en la que los conjuntos de datos de entrenamiento se dividen en segmentos lineales separados por partes (*splines*) de diferentes gradientes.

$$y = c_0 + \sum_{i=1}^{k} c_i B_i(X) \tag{2.4}$$

donde c_0 es un coeficiente constante, $B_i(X)$ es la función base y c_i es un coeficiente de la función base. En la función base se obtienen varias formas de conexiones entre las variables independientes. La función tiene la siguiente forma:

$$max(0, X - c) \tag{2.5}$$

o

$$max(0, c - X) \tag{2.6}$$

donde c es una constante, X representa las variables independientes. El propósito de la función base es convertir las variables independientes X en nuevas variables (por ejemplo $X^{'}$). Donde $X^{'}$ tomará el valor de X si X es mayor que c, y tomará el valor de cero si el valor de X es menor que c (ZhangAnthony y GohSmith, 2016).

The Gini criterion

Borland et al. (1998) definen que el criterio de *Gini* es utilizado para asignar la probabilidad $p(j|t)$ a un objeto de la clase j seleccionado al azar. La probabilidad estimada de que el objeto esté realmente en la clase j es $p(j|t)$, definida por la siguiente ecuación 2.7:

$$S = -\sum_{j} p_j log(p_j) \tag{2.7}$$

donde p_j es la probabilidad de pertenencia a la clase j, la cual se calcula como la proporción de la clase j en el conjunto.

La ganancia de información sirve para determinar qué atributo de un conjunto de vectores de características de entrenamiento es más útil para discriminar entre las clases. Esta ganancia indica la importancia de un atributo del vector de características, y es utilizada para decidir el orden de los atributos en los nodos de un árbol de decisión.

WoE and IV

L. Chen (2022) define al Valor de la Información (*IV*, por sus siglas en inglés), como el cálculo derivado del Peso de Evidencia (*WoE*, por sus siglas en inglés), la cual es una forma de codificación de las variables independientes. Para la codificación de una variable con WoE el primer paso es agruparlas, después de esto, la fórmula de cálculo de la WoE se realiza por medio de la ecuación 2.8.

$$WOE = ln(\frac{py_q}{pn_q}) \tag{2.8}$$

en donde py_q es la proporción de datos buenos (es decir, sin riesgo) de un grupo formado con respecto a todos los datos buenos en todas las muestras, pn_q es la proporción de datos malos (es decir, datos con riesgo) de un grupo, con respecto a los datos malos en la muestra. y_q es la cantidad de los datos buenos en este conjunto, n_q es la cantidad de los malos pagadores.

IV es la variable basada en WoE, y se calcula con la fórmula 2.9:

$$IV = (py_q - pn_q) * WOE = (py_q - pn_q) * ln(\frac{py_q}{pn_q}) \tag{2.9}$$

Después de calcular el IV de cada grupo de variables a partir de la fórmula 2.9, se calcula la suma de los valores IV de cada variable integrando los valores de IV del grupo de características por medio de la fórmula 2.10:

$$IV = \sum q^n IV_q \tag{2.10}$$

donde n es el número de grupos en las variables.

Análisis de componentes principales

El principal objetivo del *Principal Components Analysis* (*PCA*, por sus siglas en inglés) es representar los datos en un espacio de características de menor dimensión. Cada componente principal (*PC1, PC2,...*) es una combinación lineal de los atributos originales, y hay tantos componentes principales como atributos contenga el conjunto de datos. El primer componente principal (*PC1*) es la combinación lineal normalizada de atributos que tiene la mayor varianza. El uso de los componentes principales reduce la dimensionalidad de los datos, explicando de esta manera la estructura interna de los datos y no considera las etiquetas de los datos. Dichos componentes se obtienen diagonalizando la matriz de covarianza de los datos, sus direcciones y magnitudes vienen dadas por los vectores y valores propios de la matriz de covarianza (Dougherty, 2012).

Análisis discriminante lineal

El análisis discriminante lineal (*LDA*, por sus siglas en inglés) es un método supervisado (es decir, reconoce que los datos comprenden varias clases etiquetadas) que resulta útil para reducir la dimensionalidad. Trata explícitamente de optimizar la separabilidad de las clases. Los vectores base de esta transformación, conocidos como canónicos (que son combinaciones lineales de las características originales), se encuentran maximizando la relación discriminante de Fisher (*FDR*, por sus siglas en inglés). LDA se generaliza para problemas de multiclase, donde para C clases, se busca las proyecciones en *C-1*. Es decir, para dos clases estas deben estar lo más

separados posible y sus varianzas deben ser lo más pequeñas posible. La dimensionalidad puede reducirse de dos a una, preservando (la mayor parte de) la información discriminatoria de los datos. La dirección que produce la mejor discriminación es la que maximiza la distancia entre las medias de las clases de datos proyectadas, normalizada por una medida de la dispersión dentro de la clase (Dougherty, 2012).

2.5. Aprendizaje computacional

Russell y Norvig (2004) definen que el campo de la IA no se limita a la comprensión y a la realización de tareas de clasificación, sino que se centra en el desarrollo de sistemas inteligentes, abarcando diversas áreas como el aprendizaje y la percepción. La IA busca sintetizar y automatizar tareas para una gran cantidad de actividades intelectuales humanas, siendo un campo genuinamente universal.

Una de las áreas de la IA que han tenido mayor desarrollo en los últimos años es el ML, su desarrollo se ha visto beneficiado por el incremento constante de las fuentes de datos que puede ser ocupadas para construir estos modelos. El ML se define como un conjunto de métodos que son capaces de detectar y clasificar automáticamente patrones en los datos y usarlos para predecir datos futuros, o para la toma de decisiones en condiciones de incertidumbre Murphy (2012). El ML puede dividirse en dos categorías:

- Aprendizaje supervisado.
- Aprendizaje no-supervisado.

Chapelle et al. (2006) definen otras categorías como:

- Aprendizaje por refuerzo.
- Aprendizaje semi-supervisado.

Aprendizaje supervisado

El propósito del aprendizaje supervisado tiene como objetivo aprender a relacionar un conjunto de entradas x con un conjunto de salidas y, partiendo de un conjunto de pares de entradas y salidas conocidas, $D = (x_i, y_i)_{i=1}^N$ donde D es llamado el conjunto de entradas de entrenamiento y N es el número de ejemplos de entrenamiento (Murphy, 2012). En una configuración simple, cada entrada x_i es un vector de dimensión d que representan las variables, que son llamadas características o atributos. Mientras que la forma de la variable de salida o respuesta puede ser una variable y_i con un valor escalar de tipo categórica o nominal de algún conjunto finito.

El aprendizaje supervisado pretende proporcionar directamente un poder discriminatorio para la clasificación de patrones (Deng y Yu, 2014). Así mismo, González (2015) define al aprendizaje supervisado como aquel que intenta desarrollar modelos que sean capaces de predecir el valor de las variables dependientes usando las variables independientes.

Aprendizaje no supervisado

Este tipo de aprendizaje solo tienen entradas $D = (x_i)_{i=1}^N$ y el objetivo es encontrar regularidades en los datos. Para este tipo de aprendizaje el problema es menos definido, ya que se desconocen los patrones que se buscan y no existe una métrica de error que se puede utilizar

para comparar los resultados, ya que no se cuenta con una salida deseada para cada entrada (Murphy, 2012).

El aprendizaje no supervisado se refiere a no utilizar información de supervisión específica de la tarea (por ejemplo, etiquetas de clases objetivas) en el proceso de aprendizaje Deng y Yu (2014).

El aprendizaje no supervisado es aquel donde no existe una distinción clara entre variables que se han dependientes e independientes, en este caso se pretende encontrar la estructura que explique la estructura de los datos. Uno de los ejemplos más representativos de tipo de aprendizaje es el análisis de conglomerados (*clustering*) en el cual el objetivo es encontrar grupos de datos que compartan características similares (González, 2015).

En el aprendizaje no supervisado, una unidad de procesamiento de salida se entrena para responder a grupos de patrones dentro de la entrada. En este paradigma, el sistema identifica las características sobresalientes desde un punto de vista estadístico dentro del conjunto de datos de entrada (Abraham, 2005).

Aprendizaje por refuerzo

El aprendizaje por refuerzo consiste en aprender a actuar o comportarse cuando se le dan señales ocasionales al sistema de recompensa o de castigo. Las acciones pueden afectar no solo a la recompensa inmediata, sino también a la siguiente situación y, a través de ella, a todas las recompensas posteriores. La exploración continua y la recompensa son dos características más importantes del aprendizaje por refuerzo. El objetivo de la modelo es adquirir la habilidad de tomar acciones de una manera que optimicen las futuras recompensas que obtiene durante su funcionamiento (Sutton y Barto, 1998).

Aprendizaje semi-supervisado

Este tipo de aprendizaje se encuentra en un punto intermedio entre lo que se realiza en el aprendizaje supervisado y el no supervisado. La idea es utilizar algunos datos etiquetados para la clasificación e ir aprendiendo junto con los datos no etiquetados. Para este tipo de aprendizaje, se tienen datos de entrenamientos etiquetados $(x_i, y_i)_{i=1}^{N}$ y los datos muestras para las cuales no se conocen sus etiquetas $(x_i)_{i=N+1}^{N+u}$ (Chapelle et al., 2006).

Adicionalmente, en los últimos años se han desarrollado avances importantes en diversas ramas del aprendizaje computacional gracias al uso del aprendizaje profundo. La aplicación de modelos más complejos se dieron a partir de factores como el surgimiento de equipos de cómputo de mayor capacidad, la disponibilidad de grandes volúmenes de datos, y mejoras importantes en los algoritmos de entrenamiento de los algoritmos, por citar algunas. El aprendizaje profundo permite que modelos computacionales compuestos por múltiples capas de procesamiento aprendan representaciones de datos con diversos niveles de abstracción para aprender y hacer representaciones de manera jerárquica (LeCun, Bengio, y Hinton, 2015).

2.5.1. Métodos de agrupamiento

K-means

El algoritmo *K-means* es un algoritmo capaz de agrupar un conjunto de datos de forma rápida y eficaz, en solo unas pocas iteraciones Géron (2019). La implementación de este método consta de los siguientes pasos (Cambronero y Moreno, 2006):

1. Colocar centroides aleatoriamente.
2. Etiquetar las muestras.
3. Actualizar los centroides.
4. Reubicar cada punto de datos.
5. Repetir los puntos 3 y 4, hasta que ninguno de los puntos cambie de posición.

El parámetro principal del algoritmo de *k-means* recibe el número de *clustering* a identificar, esta entrada puede elegirse por intuición o utilizando el método *Elbow* (Thorndike, 1953), el cual puede ayudar a determinar el número adecuado de *k*. En este trabajo, se utiliza *k-means* para segmentar los datos en un número de *clustering* definido por el método *Elbow*, a partir de los cuales se construyeron los modelos supervisados de aprendizaje computacional.

DBSCAN

La *Density-Based Spatial Clustering of Applications with Noise* (*DBSCAN*, por sus siglas en inglés), este algoritmo agrupa las muestras que se encuentran cercanas a los centroides y los lejanos son considerados como los datos atípicos, esta implementación consiste en los siguientes pasos Géron (2019):

1. Para cada muestra, el algoritmo realiza el conteo de cuántas instancias se encuentran a una pequeña distancia ϵ. Esta región es llamada vecindad ϵ de la muestra.
2. Si se obtiene un mínimo de muestras en la vecindad ϵ, ésta es considerada una instancia central.
3. Todas las muestras cercanas a la vecindad de un centroide pertenecen al mismo *cluster*.
4. Cualquier muestra que se encuentra lejos de la vecindad es considerada como un dato atípico.

Para la implementación de este algoritmo es necesario definir el número mínimo de muestras por cada *clúster* y la longitud de conexión de éstas, es definida por la distancia Euclidiana.

2.5.2. Redes neuronales

Una neurona es una célula del cerebro encargada de recibir, procesar y enviar señales eléctricas. Estudios sugieren que la capacidad de procesamiento de información en el cerebro se origina a partir de redes de este tipo de células (Russell y Norvig, 2004).

Graupe (2007) define a las RNA como redes computacionales que intentan simular las redes de neuronas del sistema nervioso central humano o animal.

Callejas et al. (2018) consideran que las RNA forman parte de la IA, mientras que Fausett (1993) la define como herramientas útiles para la resolución de distintos problemas, los cuales pueden caracterizarse como de mapeo, de agrupación y de optimización, existiendo distintas RN para cada tipo de problema.

La estructura de las neuronas de una red neuronal tienen una estrecha relación con los algoritmos utilizados empleados para entrenar el aprendizaje de la red. La clasificación de las redes neuronales se basada en su estructura que permite identificar tres arquitecturas definidas a continuación por Haykin (2009):

Redes neuronales de capa simple *feed-forward*

Es la forma más simple de las capas de redes neuronales, se tiene una capa de entrada que se proyecta sobre una capa de salida, es decir, esta red es estrictamente del tipo hacia adelante. La denominación se refiere a la única capa de salida sin considerar la capa de entrada, ya que no realiza ningún cálculo durante el procedimiento (*feedforward*). En la Figura 2.1 se muestra este tipo de red.

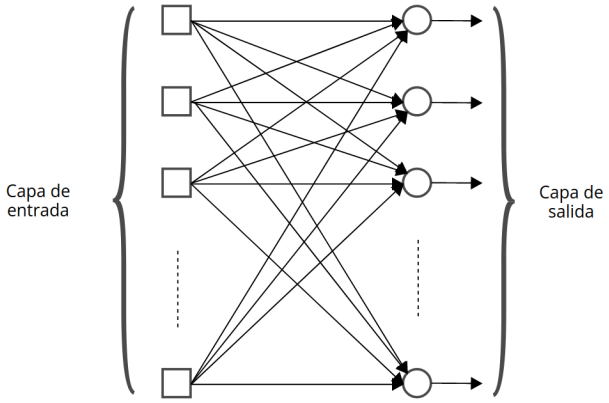


Figura 2.1: Red neuronal con una sola capa de neuronas. Fuente y acoplado de Haykin (2009)

Redes neuronales multicapa *feed-forward*

La arquitectura de este tipo de redes tiene este nombre debido a que tiene más de una capa oculta. El término oculta se refiere a las capas de neuronas que están ubicadas entre las capas de entrada y salida.

La capa de entrada suministra los elementos a la segunda capa, la cual se considera la primera capa oculta, así mismo las salidas de la primera capa oculta funcionan como entrada para la tercera capa, comportándose de esta manera sucesivamente para las demás capas contenidas en la red. Mientras que la salida de la última capa constituye la respuesta de la red neuronal. En la Figura 2.2 se describe este tipo de red multicapa.

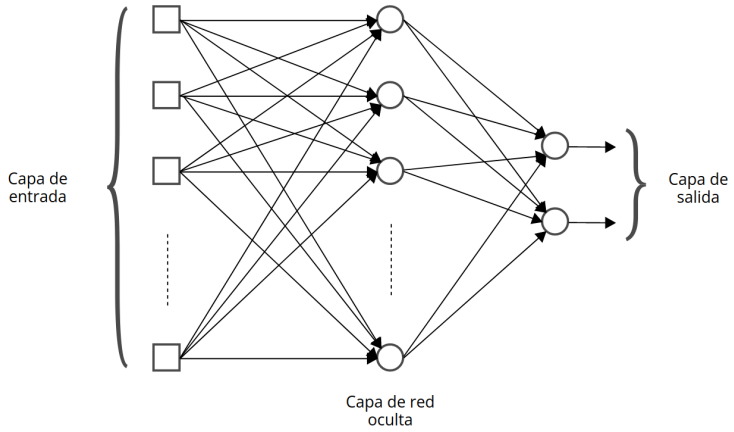


Figura 2.2: Red totalmente conectada con alimentación de una capa oculta y una capa de salida. Fuente y acoplado de: Haykin (2009)

Funciones de activación

El cálculo del valor de ajuste para los pesos de la red neuronal (δ) requiere el cálculo de la derivada de la función de activación asociada a cada neurona $\varphi(\cdot)$ asociada a esa neurona. Para que esta derivada exista, se requiere que la función $\varphi(\cdot)$ sea continua. En términos básicos, el único requisito que debe satisfacer una función de activación es que sea derivable Haykin (2009). A continuación se describen las funciones de activación más utilizadas en las RNA.

La **función logística** es la función más común en redes neuronales. En su forma general está definida por la ecuación 2.11, la gráfica de la función se observa en la **Figura 2.3**:

$$\varphi(v) = \frac{1}{1 + exp(-av)} \quad , \qquad \text{a} > 0 \tag{2.11}$$

donde $v_j(n)$ es el campo local inducido de la neurona j, a es un parámetro positivo ajustable, y la amplitud de su salida se encuentra dentro del rango $0 \leq y_j \leq 1$.

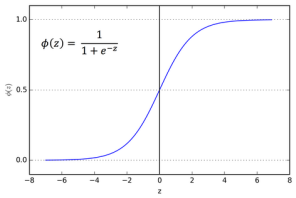


Figura 2.3: Función de activación *Simoide*.
Fuente: Haykin (2009)

Otra de las funciones no lineales ampliamente utilizadas es la **tangente hiperbólica** definida por la ecuación 2.12 y su gráfica de la función tangente hiperbólica se observa en la **Figura 2.4** :

$$\varphi(v) = a * tanh(b * v_j(n)) \tag{2.12}$$

donde a y b son constantes positivas. En realidad, la función tangente hiperbólica no es más que la función logística reescalada y sesgada.

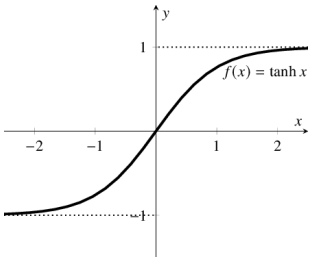


Figura 2.4: Función de activación tangente hiperbólica.
Fuente: Haykin (2009)

La **función ReLU** definida en la ecuación 2.13 significa que si la salida es positiva, saldrá el mismo valor, en caso contrario saldrá 0, cuando z es negativo, desactiva la neurona. La gráfica de la función puede observarse en la **Figura 2.5**.

$$\varphi(v) = max(0, z) \tag{2.13}$$

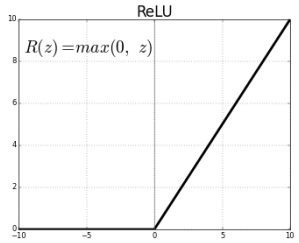


Figura 2.5: Función de activación ReLU.
Fuente: Haykin (2009)

Parámetros de optimización

Los algoritmos de optimización más populares son el Descenso de Gradiente Estocástico (*SGD*, por sus siglas en inglés), *Adagrad* y *Adam*. En general, el *SGD* es más rápido que los demás, existiendo una relación entre la velocidad y la precisión (Ghayoumi, 2022).

Los algoritmos de optimización actualizan los pesos y los sesgos de la red, además las tasas de aprendizaje pueden ser constantes o adaptativas. La tasa de aprendizaje es el valor del tamaño del paso que los algoritmos de optimización definen la razón de cambio en el ajuste de los pesos del modelo durante su entrenamiento. Su valor es pequeño y generalmente está entre 0.0 y 1.0. La selección de un valor inadecuado puede afectar a la convergencia del entrenamiento.

Ghayoumi (2022) define al Gradiente Descendente (GD) como uno de los métodos de optimización más populares y proporciona el valor óptimo a lo largo de la dirección de descenso gradual. Este método es la solución más general y popular cuando la función objetivo es convexa. Uno de sus problemas es el proceso de actualización en cada paso. Cuando se calcula el gradiente de todos los datos, el coste de cálculo del GD es elevado, especialmente cuando el tamaño de los datos es muy alto. También tarda más tiempo en converger cuando los datos incluyen más ruido o datos sesgados.

El GD calcula el gradiente para encontrar mínimos locales, mientras que el Gradiente Descendente Estocástico (SGD, por sus siglas en inglés) es un método que permite optimizar una función objetivo utilizando un elemento por cada entrenamiento. El costo del cálculo es eficiente y no depende necesariamente de la totalidad de los datos. El inconveniente es la selección de la tasa de aprendizaje adecuada (Ghayoumi, 2022).

En este método, la tasa de aprendizaje se ajusta de forma adaptativa a los cuadrados de todos los valores de gradientes anteriores, entre mayor sea ésta, la velocidad de aprendizaje es más rápida. Existe la posibilidad que, a través del tiempo, todos los valores de gradiente sean mayores y hagan que la tasa de aprendizaje se acerque a cero, y entonces los parámetros no se actualicen correctamente (Ghayoumi, 2022).

Adam es la abreviatura de Estimación Adaptativa del Momento y calcula diferentes tasas de aprendizaje. Este método es una combinación del método del momento y el adaptativo, siendo el más adecuado para problemas no convexos con una gran cantidad de datos y alta dimensión del espacio de características, siendo uno de los más adecuados para proyectos de aprendizaje profundo (Ghayoumi, 2022).

2.5.3. Árboles de decisión

En el ámbito de la minería de datos, un DT es un modelo predictivo que se puede utilizar para representar como modelos clasificaciones como a los modelos de regresión. Los DT representan un modelo jerárquico de decisiones y sus ramificaciones. Aquellos que son responsables para la toma de decisiones emplean los DT para identificar la estrategia con mayor probabilidad de alcanzar sus metas. Para la resolución de problemas, existen dos tipos de árboles: los de clasificación y los de regresión (Rokach y Maimon, 2008).

Ville (2013) describen a los DT como un modelo de aprendizaje supervisado donde los datos son particionados continuamente de acuerdo a ciertos parámetros evaluados hasta formar el árbol. Los datos obtenidos del árbol pueden ser explicados basándose en los nodos de decisión y sus hojas; las hojas son las decisiones finales y los nodos de decisión son los puntos donde los datos son separados. La principal característica de los DT son las divisiones de manera recursiva de un campo objetivo de datos en función a los valores de entrada para crear particiones y subconjuntos de datos en cualquier nivel del árbol. La calidad de división del árbol es la función para maximizar la ganancia del modelo por medio de Tangirala (2020):

- ***Entropy:*** La ganancia de la información se basa en la entropia, midiendo el grado de impureza y aleatoriedad de un conjunto de datos. Si todas las observaciones de los subconjuntos pertenecen a una clase, la entropía de este conjunto de datos es 0. Dicha entropía se define como la suma de la probabilidad de cada etiqueta multiplicada por la probabilidad logarítmica de la misma etiqueta.

$$Entropy(L) = \langle C_1|L\rangle log_2\langle C_1|L\rangle + \langle C_2|L\rangle log_2\langle C_2|L\rangle + ... + \langle C_j|L\rangle log_2\langle C_j|L\rangle \tag{2.14}$$

$$Entropy(L) = p_1 log_2 p_1 + p_2 log_2 p_2 + ... + p_j log_2 p_j \tag{2.15}$$

$$Entropy(L) = -\sum_{i=1}^{j} p_i log_2(p_i) \tag{2.16}$$

- ***Information Gain:*** La ganancia de la información se basa en la entropía, siendo la diferencia entre la entropía de una clase y la entropía condicional de la clase y la característica seleccionada. Mide la reducción de la incertidumbre tras dividir el conjunto en una característica, si el valor de la ganancia de la información aumenta, dicha característica f es más útil para la clasificación, siendo esta la mejor para la división del árbol, definida por la ecuación 2.17.

$$IG(L, f) = Entropy(L) - \sum_{v=1}^{V} \frac{|L^V|}{|L|}(Entropy(L^V)) \tag{2.17}$$

- ***GINI Index:*** El índice GINI determina la pureza de una clase específica después de dividirla según un atributo en particular. Si L es un conjunto de datos con j etiquetas de clase diferentes, definida por 2.18:

$$GINI(L) = 1 - \sum_{i=1}^{j} p_i^2 \tag{2.18}$$

2.5.3.1. Árboles de clasificación

Los árboles de clasificación se emplean para asignar un objeto o una instancia en un conjunto previamente definido de clases según los valores de sus atributos. Siendo comunes en áreas como las finanzas, el *marketing*, la ingeniería y la medicina. Este tipo de árbol resulta ser útil como una técnica de exploración y siendo un modelo predictivo para representar modelos de clasificación y de regresión (Rodriguez Molina, 2022). Los DT consisten en nodos que forma una estructura de un árbol, donde se inicia desde un nodo "raíz", caracterizado por no contar con nodos entrantes. Todos los demás nodos poseen exactamente una conexión de entrada. Se nombra como un nodo "interno" o de "pruebas" a aquel que tiene conexiones de salida, mientras que a los restantes se les conoce como "hojas", nodos "terminales" o nodos de "decisión". En un DT cada nodo interno divide el espacio de instancia en dos o más subespacios, dependiendo de los valores de los atributos de entrada. En el caso más simple, en cada prueba se considera un solo atributo, lo que lleva a realizar la división del espacio de instancias dependiendo del valor de cada atributo. Para el caso de los atributos numéricos, esta condición se refiere a un rango (Rokach y Maimon, 2008).

Las hojas tienen la capacidad de contener un conjunto de probabilidades que indican la probabilidad de que el atributo objetivo tenga un valor específico. En la representación gráfica, los nodos internos son representados como círculos, mientras que las hojas son denotadas con cuadrados (ver Figura 2.6). De cada nodo interno, es decir, es aquel que no es una hoja, de él pueden surgir dos o más ramas. Cada nodo corresponde a un atributo específico, y las ramas a un rango de valores. Estos rangos deben de dividir el conjunto de valores del atributo proporcionado. El proceso de clasificación de las instancias se realiza siguiendo un recorrido desde el nodo raíz del árbol hasta una hoja, con base en los resultados de las pruebas realizadas durante todo el recorrido. Se inicia desde la raíz del árbol, se considera el atributo relacionado con dicha raíz y se determina la rama que se relaciona con el valor observado de ese atributo. Posteriormente, se procede al nodo al que conduce esa rama. Estas operaciones se repiten hasta llegar a una hoja. El árbol de decisión integra tanto atributos numéricos como no numéricos (Luna Santander et al., 2023).

Según Breiman et al. (1984) la precisión del árbol se ve influenciado por su complejidad. Usualmente, la complejidad del árbol se evalúa utilizando métricas como: el total de nodos, el número de hojas, la profundidad del árbol y la cantidad de atributos utilizados. El control de la complejidad se realiza explícitamente a través de los criterios de parada y los métodos de poda que se apliquen.

La inducción en los árboles de derivación está estrechamente relacionada con la inducción de reglas. Cada camino desde la raíz hasta una de sus hojas puede transformarse en una regla, simplemente uniendo las pruebas a lo largo del camino para formar la parte del antecedente, y tomando la predicción de la clase de la hoja como el valor de la clase. El conjunto de reglas resultantes puede simplificarse para mejorar su comprensión por parte de un usuario y su precisión (Quinlan, 1987).

2.5.3.2. Representación gráfica y terminología

Myles et al. (2004) define a un árbol de decisión es un modelo jerárquico compuesto por reglas de decisión, que se aplican recursivamente para dividir el espacio de características de un conjunto de datos en subespacios puros de una sola clase. El resultado de los programas de los DT se enfoca en el propio DT y no en la representación del espacio de características.

Los DT presentan dos tipos de nodos: nodos rama y nodos hoja (representados por círculos

y cuadrados, respectivamente, en la Figura 2.6). El nodo X se denomina nodo raíz y representa todo el espacio de características. Los demás nodos (nodos X1-X5 en este ejemplo) representan cada uno un subespacio del espacio de características original, los nodos X31, X32, X41, X21 y X51 son las hojas del árbol. Así mismo, una ruta de decisión se encuentra representada mediante una secuencia de nodos que conecta un nodo padre con uno de sus hijos. Las reglas de decisión son las expresiones de desigualdad que describen los límites de la decisión.

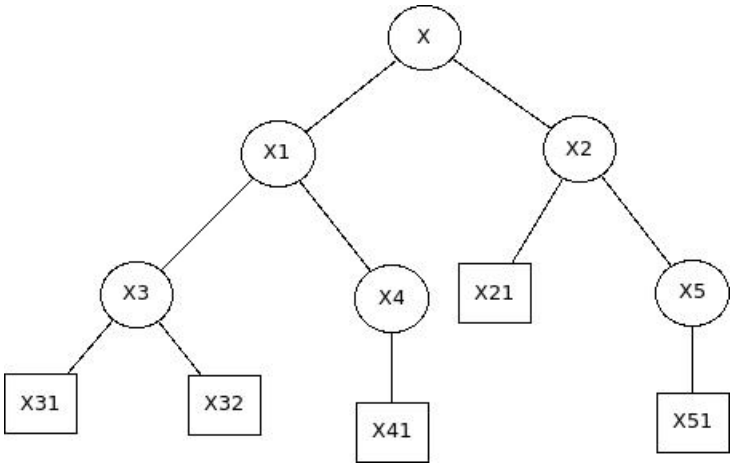


Figura 2.6: Estructura de un árbol de decisión

.

Fuente: Myles et al. (2004)

Toda la construcción de un árbol, gira en torno a tres elementos:

- La selección de las divisiones.
- Las decisiones de cuándo declarar un nodo como terminal o seguir dividiéndolo.
- La asignación de cada nodo terminal a una clase.

Breiman et al. (1984) consideran un problema la construcción del árbol al momento de determinar las divisiones binarias de un conjunto de datos llamado L en trozos cada vez más pequeños con la finalidad de identificar los nodos terminales y sus asignaciones. El procedimiento se trata de encontrar buenas divisiones y saber el momento indicado cuándo se dejará de dividir.

Grabczewski (2014) declara que existen numerosos algoritmos de construcción que dan lugar a modelos en donde las divisiones se realizan sobre la base de condiciones simples relativas a características individuales, las cuales se declaran a continuación.

Por lo general, los criterios para dividir o ramificar se basan en evaluaciones de la impureza de un nodo (López Maldonado et al., 2014). La impuresa del nodo se refiere al nivel en el que un nodo contiene diferentes clases dentro del problema de clasificación que se busca resolver. Un nodo se considera puro cuando solo contiene instancias de una sola clase del problema. Los algoritmos más conocidos son el ID3, CART, C4.5 y C5.0 (Galiano, 2002).

ID3

El Dicotomizador Iterativo 3 (ID3, por sus siglas en inglés) es una de las primeras ideas de la inducción usando árboles de decisión. Su criterio de división se basa en la teoría de la información, donde un nodo se divide en tantos subnodos como el número de valores posibles

de la característica utilizada para la división. El inconveniente más grave de ID3 es el requisito de que la descripción de los datos solo puede incluir características discretas. Cuando en la tabla de datos los atributos son numéricos, primero hay que hacerlos discretos (Grabczewski, 2014).

Rokach y Maimon (2008) describe al método ID3 como un algoritmo de árbol de decisión bastante básico. Basado en el concepto de la ganancia de información como criterio para dividir, este método finaliza su crecimiento cuando todas las instancias coinciden con un único valor de una característica objetivo o cuando la máxima ganancia de información no es mayor que cero. ID3 no implementa técnicas de poda, ni maneja atributos numéricos o valores faltantes (SÁNCHEZ-CERVANTES et al., 2017; Muñoz, 2018; Aguilar Castillo, 2021).

El algoritmo ID3 utiliza el concepto de la ganancia de información para determinar qué atributo colocar en cada nodo de decisión. Identificando la capacidad de un atributo para separar efectivamente los ejemplos de entrenamiento en las distintas clases, eligiendo aquel que proporcione mayor información útil en la separación. De manera general, la información proporcionada de una distribución de probabilidad $P = (p1, p2, p3, ..., pn)$ y un ejemplo S, se define como la entropía de P, y se calcula mediante la fórmula 2.19 (Rodríguez, 2021):

$$EntropiaP = -\sum_{i=1}^{n} p_i * log(p_i) \tag{2.19}$$

Si se poseen métricas que evalúen el grado de la combinación de clases para todos los ejemplos y para cualquier ubicación en el desarrollo del árbol. Se debe contar con una nueva métrica para elegir el atributo que debe etiquetar el nodo actual. Esto define la ganancia para una prueba T y una posición p de la siguiente manera (Rodríguez, 2021) :

$$Ganancia(p, T) = EntropiaP - \sum_{i=1}^{n} p_j * Entropia(p_j) \tag{2.20}$$

Donde los valores (p_j) representan el conjunto de todos los posibles valores para el atributo T. De esta manera, esta medida se emplea para identificar que atributo es más adecuado y construir el árbol de decisión, donde cada nodo contiene el atributo con la mayor ganancia de información entre todos los atributos que aún no se han sido considerados en la ruta desde el nodo raíz (Hssina et al., 2014).

CART

Grabczewski (2014) define a los Árboles de Clasificación y Regresión (CART, por sus siglas en inglés) son uno de los métodos de inducción de árboles de decisión más populares y de mayor éxito. El algoritmo no es paramétrico y está enfocado a la construcción de árboles binarios, es decir, cada nodo interno tiene exactamente dos aristas salientes (Rokach y Maimon, 2008).

Una característica importante de CART es su capacidad para generar árboles de regresión. En estos árboles, las hojas predicen un número real y no una clase. Buscando las divisiones que minimicen el error cuadrático de la predicción en cada hoja, la cual se basa en la media ponderada del nodo.

Sea $A = x_1, x_2, x_3, ..., x_n$ el conjunto de atributos o variables predictoras. Si A es una de n número de categorías, entonces hay $2^{n-1} - 1$ posibles divisiones para este nodo predictor (Hssina et al., 2014).

Rutkowski et al. (2014) describen el método CART con el siguiente procedimiento:

- Se crea el nodo raíz al que llaman L_0.
- Se procesa un subconjunto Sq del conjunto de entrenamiento X para inducir los nodos del árbol Lq incluido el nodo raíz.
- Si todos los elementos del conjunto Sq son de la misma clase, entonces el nodo es etiquetado como nodo hoja y la división no es realizada.
- Para cada atributo disponible x_i, el conjunto de los valores que los atributos pueden tomar A_i es dividido en dos subconjuntos disjunto A_I^i y A_D^i.
- Los conjuntos A_I^i y A_D^i dividen al conjunto de entrenamiento S_q en dos conjuntos izquierdo y derecho.

C4.5

Surge a partir de ID3 y comparte muchas soluciones con su antecesor. Las principales diferencias introducidas en C4.5 son (Grabczewski, 2014):

- Modificación de la medida de impureza de los nodos.
- Soporte para el manejo directo de atributos continuos.
- Introducción de un método de poda.
- Métodos precisos para tratar los datos con valores perdidos.

La división cesa cuando el número de instancias a dividir es inferior a un determinado umbral. La poda se encuentra basada en errores que se realizan después de la fase de crecimiento. Para solucionar este tipo de inconvenientes, se utiliza un nuevo cálculo que permite medir una razón de ganancia (Hssina et al., 2014):

$$RelGan(p,T) = \frac{Gan(p,T)}{infDiv(p,T)} \tag{2.21}$$

Donde:

$$infDiv(p,test) = -\sum_{j=1}^{k} p(\frac{i}{j}) * log(p^{'}(\frac{i}{j})) \tag{2.22}$$

$p^{'}(\frac{i}{j})$ es la proporción de elementos presentes en la posición p, tomando el valor de la j-ésima prueba. Para evaluar la ganancia o la relación de ganancia, el algoritmo estima las probabilidades de salidas diferentes, entonces el nuevo criterio de ganancia toma la forma:

$$Gan(p) = F(Info(T) - Info(p,T)) \tag{2.23}$$

2.5.4. Máquinas de Soporte Vectorial

El modelo SVM se entrena basándose en el ajuste de los parámetros, cuyo objetivo es crear un límite de decisión entre dos clases que pueda permitir la predicción de etiquetas a partir de uno o más vectores de características H. Wang y Hu (2005). La función *kernel* permite a las SVM realizar una clasificación bidimensional de un conjunto de datos, proyectando los datos de un espacio de baja dimensión a un espacio de dimensión superior Patle y Chouhan (2013). Los tipos de *kernel* utilizados en las SVM se en listan a continuación:

- ***kernel* lineal:** se describe mediante la siguiente ecuación 2.24:

$$K(x, x_i) = x \cdot x^T \tag{2.24}$$

- ***kernel* polinomial:** es una función direccional, esto quiere decir que la salida depende de la dirección de los dos vectores en el espacio de baja dimensión, la magnitud de la salida depende de la magnitud del vector x_i, definido por la función 2.25.

$$K(x, x_i) = (1 + x \cdot x_i^T)^d \tag{2.25}$$

  donde d es el grado de la función del *kernel*.

- ***Radial* básico:** es uno de los *kernel* más utilizados y definida por 2.26:

$$K(x, x_i) = e^{-\gamma \|x - x_i\|^2} \tag{2.26}$$

  donde el parámetro $\gamma > 0$

- ***Sigmoid*:** 2.27

$$K(x, x_i) = tanh(\gamma x_i^T x_j + r) \tag{2.27}$$

donde γ, r y d son los parámetros del *kernel*. La selección del *kernel* depende de la aplicación y no es fija.

2.5.5. *XGBoost*

Se trata de una implementación eficiente y escalable del marco de trabajo de *gradient boosting* de J. Friedman et al. (2000). El *eXtreme Gradient Boosting* (*XGBoost*) es un algoritmo basado en el algoritmo *gradient boosting tree* (T. Chen y He, 2017). El algoritmo *XGBoost* se encuentra basado en la teoría de la clasificación y el árbol de regresión, siendo un método eficaz para los problemas de regresión y clasificación (Qiu et al., 2021). Además, la función objetivo evita que el modelo presente sobre entrenamiento. $D = (x_i, y_i)$ representa un conjunto de datos que contiene n ejemplos y m características, y los resultados de predicción están conformados por las siguientes:

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in \varphi \tag{2.28}$$

$$\varphi = \{f(x) = w_s(x)\}(s : R^m \longrightarrow T, w_s \in R^T) \tag{2.29}$$

donde $\hat{y}$, representa la etiqueta de predicción, x_i representa una de las muestras y $f_k(x_i)$ es la muestra dada, φ simboliza el árbol de regresión, $f(x)$ y w representa el peso de las hojas y número de hojas, respectivamente.

2.5.6. Optimización de hiperparámetros

La validación cruzada es un método estadístico utilizado para evaluar y comparar algoritmos de aprendizaje computacional mediante la división del conjunto de entrenamiento en dos segmentos: uno que se usa para aprender o entrenar y un segundo para validar el modelo Refaeilzadeh et al. (2009). La validación se realiza de manera cruzada, es decir, los conjuntos de entrenamiento y validación deben de cruzarse en secuencias sucesivas de modo que cada dato pueda ser validado. En la validación cruzada en *k-fold*, los datos se dividen primero en k segmentos de igual tamaño. A continuación, se realizan k iteraciones de entrenamiento y validación, de forma que en cada iteración se reserva un segmento de los datos para la validación, mientras que los *k-1* segmentos restantes se utilizan para el entrenamiento. En la **Figura 2.7** se puede observar el funcionamiento con un $k = 3$.

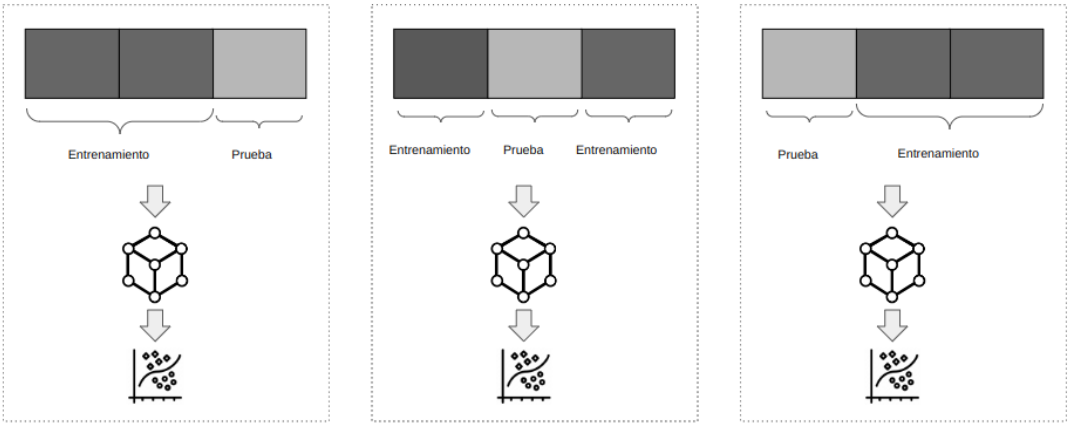


Figura 2.7: Procedimiento para $k = 3$ para la validación cruzada Refaeilzadeh et al. (2009).

2.6. Medidas de rendimiento

Las medidas de rendimiento juegan un papel muy importante en problemas de clasificación donde evalúan el desempeño de los clasificadores aplicados sobre los datos y permiten identificar el mejor clasificador dependiendo del objetivo de investigación.

La precisión es una de las métricas más utilizadas, realizando la evaluación con base en la precisión de predicción, lo que puede ser inadecuado en el caso de contar con bases de datos no balanceadas y los costos de errores varían considerablemente (Danjuma, 2015).

Selvik y Abrahamsen (2017) definen a la precisión como la relación entre los casos positivos predichos correctamente con respecto al número total de casos positivos predichos.

Matriz de confusión

La matriz de confusión, según la definición de Kohavi y Provost (1998), es una herramienta que facilita la evaluación del rendimiento de un algoritmo de clasificación. Cada columna de esta matriz representa la cantidad de predicciones realizadas por cada clase, y cada fila representa a las instancias que pertenecen a la clase real. En otras palabras, esta matriz permite identificar los tipos de aciertos y errores que el modelo presenta al aprender de los datos (Barrios, 2019).

Para la clasificación binaria se encuentran dos posibles resultados predichos para cada instancia, donde T (True, Positivo, Verdadero) y F (False, Negativo), como se muestra en la **Tabla 2.1**.

Tabla 2.1: Matriz confusión (Mahbobi et al., 2021)

	Predicción Positiva	Predicción Negativa
Positiva	Verdadero Positivo (TP)	Falso Negativo (FN)
Negativo	Falso Positivo(FP)	Verdadero-Negativo(TN)

- **Verdadero-Positivo (TP, por sus siglas en inglés):** El modelo clasifica una solicitud no moroso correctamente.

- **Falso-Negativo (FN, por sus siglas en inglés):** El modelo clasifica una solicitud no moroso como un moroso.
- **Falso-Positivo (FP, por sus siglas en inglés:** El modelo clasifica una solicitud morosa como no moroso.
- **Verdadero-Negativo (TN, por sus siglas en inglés):** El modelo clasifica una solicitud morosa correctamente.

Debido a la distribución desbalanceada del conjunto de datos analizado para medir el rendimiento de los clasificadores, se utilizan las medidas que a continuación se describen (Luna Santander et al., 2023).

La exactitud (en inglés, *Accuracy*) hace referencia a la cercanía del resultado de una medición del valor verdadero. Se presenta como la relación entre los resultados correctos (tanto los TP y los TN) y el total de casos analizados. La fórmula que se utiliza para calcular esta medida de rendimiento se muestra en la ecuación 2.30 (Borja-Robalino et al., 2020):

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2.30}$$

La precisión (en inglés, *Precision*) se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción de verdaderos positivos dividida entre todos los resultados positivos (tanto verdaderos positivos, como falsos positivos). Se refiere al porcentaje de casos positivos detectados. Se calcula con la ecuación 2.31 (Borja-Robalino et al., 2020):

$$Precision = \frac{TP}{TP + TN} \tag{2.31}$$

El error tipo I es la proporción de los casos positivos clasificados incorrectamente (Akobeng, 2016). En el contexto del problema de riesgo crediticio se refiere a aquellas solicitudes reconocidas como malos pagadores. Se calcula con la ecuación .

$$Error\ tipo\ I = \frac{FN}{FN + TP} \tag{2.32}$$

El error tipo II es la proporción de los casos negativos identificados incorrectamente como positivos (Akobeng, 2016). En el contexto del problema en estudio se refiere a aquellas solicitudes reconocidas como buenos pagadores; sin embargo, estos resultan ser malos pagadores y presentar posibles pérdidas al incumplimiento de sus pagos, por lo tanto, este error el de mayor interés para la institución financiera. Se calcula con la ecuación .

$$Error\ tipo\ II = \frac{FP}{FP + TN} \tag{2.33}$$

La sensibilidad (en inglés *Recall* o *Sensitivity*) es conocida como la tasa de verdaderos positivos, siendo la porción de casos positivos que fueron identificados correctamente, calculado con la ecuación 2.34 (Borja-Robalino et al., 2020).

$$Recall = \frac{TP}{TP + FN} \tag{2.34}$$

La especificidad (en inglés *Specificity*) es conocida como la tasa de verdaderos negativos, siendo la porción de casos negativos que fueron identificados correctamente por el clasificador y la fórmula para calcularla es 2.35 (Borja-Robalino et al., 2020).

$$Specificity = \frac{TN}{TN + FN} \tag{2.35}$$

El *F1 Score* se define como la media armónica de la precisión y sensibilidad. Es de gran utilidad cuando la distribución de las clases es desbalanceada. Se calcula con la ecuación 2.36 (Chicco y Jurman, 2020):

$$F1 - Score = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{2.36}$$

Chicco y Jurman (2020) definen al Coeficiente de Correlación de *Matthews* (*MCC*, por sus siglas en inglés) como un índice estadístico fiable para modelos de clasificación binaria, produciendo una puntuación alta solo si la predicción obtuvo buenos resultados en todas las cuatro categorías de la matriz de confusión, proporcionalmente tanto al tamaño de los elementos positivos como al tamaño de los elementos negativos en el conjunto de datos. El *MCC* es una alternativa para resolver conjuntos de datos desbalanceados, generando una puntuación alta solo si el modelo es capaz de predecir correctamente la mayoría de los datos positivos y negativos, calculados mediante la ecuación 2.37:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}} \tag{2.37}$$

Balanced Error Rate (BER, por sus siglas en inglés) es la media aritmética de la suma de la sensibilidad (ecuación 2.34) y la especificidad (ecuación 2.35), mediante la ecuación 2.38 cuando este valor se encuentra más cercano a 1 significa que el modelo se encuentra clasificando correctamente los patrones de entrada (Chicco et al., 2021)

$$BER = 1 - \frac{Recall + Specificity}{2} \tag{2.38}$$

Además, durante el análisis y la construcción de la base de datos de la institución financiera, nos dimos cuenta de que el error tipo II (2.33) permite determinar aquellos casos a los que ya no se les debe otorgar el crédito; sin embargo, el clasificador sugiere otorgar el crédito, y en la práctica, es indispensable reducir el número de estos casos. Por esta razón, en nuestros experimentos, es de suma importancia monitorear el rendimiento de los clasificadores utilizando el error de tipo II.

2.7. Bibliotecas de desarrollo

Existen distintas bibliotecas en *Python* que son una serie de colecciones de módulos relacionados entre ellos. Dichas librerías hacen que la programación resulte ser más sencilla.

Python

Python es un lenguaje de programación, el cual fue diseñado para ser fácil de aprender y se encuentra disponible para distintas multiplataformas. Es ideal para el desarrollo rápido de aplicaciones en diversas áreas debido a la gran cantidad de bibliotecas en diversas áreas como lo son: la carga de datos, visualización, estadística, procesamiento de lenguaje natural, procesamiento de imágenes y en especial para el propósito de este trabajo de investigación, para desarrollar aplicaciones de aprendizaje automático (Muller y Guido, 2017).

- **Numpy:** Se trata de una librería en *python* que simplifica las tareas de computación científica, incluye características para trabajar con matrices multidimensionales, funciones matemáticas de alto nivel y generadores de números pseudoaleatorios.
- **Matplotlib:** Es una biblioteca enfocada a la visualización de datos a partir de datos contenidos en listas o estructuras de datos definidas principalmente por la biblioteca o pandas.

- **SciPy:** Es una librería compuesta por diversas herramientas que ofrecen funcionalidad para abordar el cálculo de tareas científicas y analíticas. Dentro de la colección de herramientas se incluyen soluciones de álgebra lineal, optimización, interpolación, procesamiento de señales e imágenes, etc.
- **Pandas:** Es una biblioteca que permite la manipulación y análisis de datos sobre tablas numéricas y series temporales. Tiene opciones para importar y exportar tablas desde y hacia diferentes tipos de archivos como hojas de cálculo y bases de datos.
- **Scikit-learn:** Es una biblioteca (muy extensa) con algoritmos para procesamiento y análisis de datos. Además, presenta la compatibilidad con otras librerías de *Python* como *NumPy, SciPy y matplotlib*, siendo muy popular y destacada para el aprendizaje automático, con las siguientes características:
  - Herramientas simples y eficientes para el análisis predictivo de datos.
  - Accesible y reutilizable.
  - Basado en las librerías *Numpy*, *SciPy* y *matplotlib*.
  - Código abierto, utilizable comercialmente.

TensorFlow

TensorFlow es una biblioteca de software de aprendizaje profundo de código abierto para definir, entrenar y desplegar modelos de aprendizaje automático (Muller y Guido, 2017).

Keras (2022) define a *TensorFlow* como una plataforma de código abierto para el aprendizaje automático. Tiene un ecosistema integral y flexible de herramientas, bibliotecas y recursos que permiten a los investigadores crear y desarrollar fácilmente aplicaciones basadas en *ML*. Combina cuatro habilidades clave:

- Ejecución eficiente de operaciones de tensor de bajo nivel en Unidad Central de Procesamiento (*CPU*, por sus siglas en inglés), Unidades de Procesamiento Gráfico (*GPU*, por sus siglas en inglés) y Unidad de Procesamiento Tensorial (*TPU*, por sus siglas en inglés).
- Cálculo del gradiente de expresiones diferenciales arbitrarias.
- Escalamiento del cómputo a muchos dispositivos, como *clusters* de cientos de *GPU*.
- Exportación de programas a tiempos de ejecución externos, como servidores, navegadores, dispositivos móviles e integrados.

Keras

Nandy y Biswas (2018) definen a *Keras* como una biblioteca *frontend* de código abierto para RN. Funcionando como la columna vertebral para las RN, ya que cuenta con buenas capacidades para formar funciones de activación. *Keras* puede ejecutar diferentes marcos de aprendizaje profundo como el *backend*, funcionando con muchos marcos de aprendizaje profundo. Mientras que (Muller y Guido, 2017) la describen como una de las librerías de *python* más fáciles de usar para crear modelos de aprendizaje profundo.

Keras es una Interfaz de Programación de Aplicaciones (*API*, por sus siglas en inglés) para el aprendizaje profundo, creada en *Python* y ejecutada en la plataforma de aprendizaje automático *TensorFlow*. Su desarrollo se centr en permitir la experimentación rápida, con la finalidad de pasar de la idea al resultado lo más rápido posible para realizar investigaciones efectivas. Esta *API* tiene las siguientes características Keras (2022):

- **Simple:** Uso de librerías, las cuales facilitan el procesamiento de datos y construcción de modelos de clasificación.
- **Flexible:** Flujos de trabajo simples, rápidos y fáciles.
- **Potencia:** Proporciona un rendimiento y una escalabilidad en la industria: lo que permite que é ste pueda ser utilizado por organizaciones y empresas, como son la *NASA*, *YouTube* o *Waymo*.

Keras es la API de alto nivel de *TensorFlow*: presentando una interfaz accesible y altamente productiva para resolver problemas de aprendizaje computacional.

2.8. Estado del arte

En esta sección de la tesis se presentan aquellos trabajos más recientes relacionados con el tema de identificación del riesgo crediticio, donde trabajan con técnicas de sobre muestreo de información y modelos de híbridos.

2.8.1. Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk

Abedin, Guotai, Hajek, y Zhang (2022) proponen en este estudio un clasificador *Random Forest* (*RF*, por sus siglas en inglés) utilizando la técnica de sobre muestreo sintético ponderado de minorías (WSMOTE), proporcionando un equilibrio entre el rendimiento de la clase de incumplimiento y la clase de cumplimiento. Los métodos de muestreo híbridos utilizados fueron SMOTE, WSMOTE, WSMOTE-ensemble, RUS, MChanUS USOS y RUSSMOTE. De los algoritmos propuestos WSMOTE-ensemble y WSMOTE fueron los que tienen mayor precisión de la clase minoritaria en los clasificadores, como lo son: C4.5, k-NN, SVM y los clasificadores de conjuntos Bagging, Boosting, LB, RC, RTF y RF.

2.8.2. Making Deep Learning-Based Predictions for Credit Scoring Explainable

Dastile y Celik (2021) proponen un modelo de aprendizaje profundo el cual convierte los conjuntos de datos tabulares en imágenes, permitiendo de esta manera la aplicación de Redes Neuronales Convolucionales (CNNs, por sus siglas en inglés) 2D para la identificación y clasificación crediticia. Cada uno de los píxeles de la imagen corresponde a una casilla de las características del conjunto de datos. Los conjuntos de datos utilizados son German, Australian and Home Loan Equity (HMEQ). Estas tres bases de datos se encuentran públicas en los repositorios de *UCI* y *Kaggle*.

2.8.3. A new deep learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique

Shen, Zhao, Kou, y Alsaadi (2020) proponen el desarrollo de un nuevo modelo de aprendizaje profundo combinando la Long Short Term Memory (LSTM) y la Adaptative Boosting (AdaBoost) para la identificación del riesgo de crédito en los conjuntos de datos desbalanceados. Por lo tanto, utilizan la técnica de Synthetic Minority Oversampling Technique (SMOTE) para el sobremuestreo de la clase minoritaria de los datos. La fórmula matemática para la generación de una nueva muestra de la clase minoritaria se encuentra dada por 2.39:

$$x_{new} = x_i + (x_i^k - x_i) * \delta \tag{2.39}$$

donde x_i^k es uno de los vecinos cercanos a x_i y δ es un valor aleatorio entre (0,1). Por lo tanto, el nuevo elemento de la clase minoritaria x_{new} es un punto intermedio entre x_i y su vecino más cercano x_i^k.

Demostrando el método de aprendizaje propuesto excelentes resultados de *scoring* crediticio, permitiendo ayudar a desarrollar un sistema avanzado de *scoring* crediticio interno para bancos y otras instituciones financieras, que podría reducir los riesgos y aumentar los beneficios.

2.8.4. Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data

Zhang, Yan, Li, Tian, y Yoshida (2022) proponen un nuevo enfoque llamado DeepRisk que fusiona los datos demográficos de la empresa y los datos de comportamiento de financiación para predecir el riesgo de crédito de las PYME en Financiación de la Cadena de Suministro (SCF, por sus siglas en inglés). Dicho enfoque incluye datos demográficos de la empresa y datos de comportamiento de la financiación, se encuentra compuesta por 8 capas: la primera contiene los datos demográficos y datos de comportamiento del financiamiento, la segunda capa es una capa de incrustación que mapea los vectores de entrada a una dimensión superior, la tercera capa está diseñada para evitar el sobre ajuste mediante la activación de neuronas parciales en el proceso de aprendizaje. La capa 4 y 5 son utilizadas para extraer las representaciones abstractas de los vectores de entrada. La capa 6 concatena la derivación de los datos de la capa 5. La capa 7 es una capa oculta del modelo, y la capa 8 es la capa de salida para predecir la capacidad de pago.

2.8.5. A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network

Liu et al. (2022) proponen un modelo híbrido de dos etapas para mejorar el rendimiento de predicción del riesgo de crédito. En primer lugar, para hacer un uso completo de la información clasificada oculta en los datos crediticios, emplean XGBoost para transformar las características originales en una matriz de características dispersas de alta dimensión. En segundo lugar, para procesar de manera efectiva los datos de alta dimensión transformados y descubrir las relaciones entre las características, un modelo de red neuronal basado en gráficos (forgeNet) fue propuesto recientemente, éste permite procesar datos de alta dimensionalidad para predecir el riesgo crediticio.

2.8.6. Tabla comparativa

A continuación se muestra un cuadro comparativo de los modelos presentados.

Item	**Modelo**	**Sobremuestreo**	Base de datos
Abedin et al. (2022)	Bosques aleatorios	✓	Público
Dastile y Celik (2021)	CNN		Público
Shen et al. (2020)	Híbrido	✓	Privado
Zhang et al. (2022)	DeepRisk		Privado
Liu et al. (2022)	XGBoost		Privado
Metodología propuesta	Híbrido	✓	Privado y público

A partir del análisis y revisión de trabajos más recientes se han considerado algunas observaciones respecto al diseño de un modelo que permita identificar el riesgo crediticio en entidades financieras. La mayoría de los trabajos presentados describen el uso de métodos de sobremuestreo y selección de características que juegan un papel importante durante la evaluación del

riesgo crediticio. Tomando como referencia estos trabajos se propone la metodología utilizando los métodos de selección de características y agrupamiento que se describen en la siguiente sección.

Capítulo 3

Desarrollo del proyecto

Es este capítulo se describen las características de software, los entornos de desarrollo y las bibliotecas utilizadas. También, se describe el esquema general de los experimentos y la metodología usada para evaluar el rendimiento de los clasificadores utilizados en esta tesis para identificar el riesgo crediticio.

3.1. Especificaciones de software

En cuanto al software se refiere, se ha utilizado el lenguaje de programación Python 3.10.12 y las siguientes bibliotecas:

- keras 2.11.0
- numpy 1.24.1
- dtreeplt 0.1.43
- xgboost 1.7.3
- hyperopt 0.2.7
- jupyter 1.0.0
- matplotlib-inline 0.1.6
- mlxtend 0.21.0
- pandas 1.5.3

3.2. Construcción de la base de datos EIZ

La descripción detallada del conjunto de datos utilizado en experimentos de investigación en el campo del aprendizaje computacional es de vital importancia, ya que establece las bases sobre las cuales se construyen y evalúan los modelos y algoritmos. Una visión completa y transparente del conjunto de datos permite a otros investigadores comprender y reproducir los resultados de manera precisa, lo que fomenta la validación y comparación de enfoques. Por otro lado, el análisis de esta información permite identificar posibles sesgos o limitaciones inherentes al conjunto de datos, lo que es crucial para evitar generalizaciones erróneas o resultados engañosos.

La base de datos proporcionada por la institución financiera contó originalmente con un total de 15,833 registros y 42 características entre las fechas de 1 de enero del 2014 al 31 de julio de 2022, de los cuales se realizó el proceso de limpieza con el gestor de base de datos *PostgreSQL*, por la facilidad de uso y la implementación de sus funciones. Al finalizar, el proceso de limpieza, el conjunto de datos utilizado para la identificación del riesgo crediticio quedó compuesta por un total de 5,510 registros. Dicha reducción fue resultado de eliminar y unificar registros duplicados y registros con campos con información faltante. La mayoría de los registros duplicados tienen como diferencia el campo de la garantía dejada para la adquisición del crédito. Para

la unificación de estos registros se utilizó información proporcionada por la entidad financiera para realizar un etiquetado que contempla todas las garantías relacionadas con el mismo registro.

La anonimización de los datos es un proceso que busca preservar la privacidad de los datos sensibles de las personas pertenecientes a la institución financiera. Esta es una medida ética que además busca el cumplimiento legal, entre otros temas (Agrawal y Srikant, 2000). Bajo esta razón, se realizó el proceso de anonimización, para esto se eliminaron campos relacionados con los datos personales como el nombre, domicilio, CURP, RFC, datos bancarios, correos electrónicos, números de teléfono, entre otros, por lo que de un total de 42 características iniciales, se redujeron a 23. El resultado final de este proceso se observa en la **Tabla A.1**.

La selección de la variable dependiente para el diseño del conjunto de datos a analizar es de suma importancia en el desarrollo de esta investigación. Cauas (2015) define a una variable dependiente como aquella que interesa explicar en función de las otras características. El parámetro principal para la creación de esta variable son los días de mora, donde se define por la entidad financiera como a aquellos clientes morosos, siendo aquellos que tienen más de dos días de atraso en sus créditos. Además, a este grupo se agregan aquellas solicitudes que han sido negadas por las políticas de la institución, dicha codificación se encuentran en la **Tabla 3.1**.

Tabla 3.1: Codificación de la variable dependiente.

Clasificación	**Código**	**Descripción**
Moroso	0	Aquellos préstamos que tienen más de dos días de atraso en sus pagos. Además, aquellas solicitudes que han sido rechazadas por la institución.
No moroso	1	Aquellos préstamos que tiene como máximo 2 días de atraso en sus pagos.

Fuente: Elaboración propia.

Derivado a la anterior condición, se determina que el conjunto de datos final queda compuesto con 5,510 registros, de los cuales 3,215 se encuentran al corriente con sus pagos (o por lo menos no tienen más de 2 días de morosidad) y 2,295 son clasificados como morosos; esto se resume en la **Tabla 3.2**.

Tabla 3.2: Descripción del conjunto de datos.

Clasificación	**Cantidad**	**Porcentaje**
Moroso	2,295	41.65 %
No moroso	3,215	58.35 %
Total	**5,510**	**100.00 %**

Fuente: Elaboración propia.

Un preprocesamiento adicional fue realizado al conjunto de datos usando el *WoE* a las variables independientes: tasa normal, tasa moratoria, clave actividad, créditos trabajados, monto de la garantía, monto, código postal, edad, egreso, ingreso, plazo y dependientes (ver **Tablas A.4, A.5 y A.6**). El resto de características ya se encuentran categorizadas con sus respectivos valores.

3.3. Metodología del proyecto

Machado y Karray (2022) proponen utilizar de manera conjunta el aprendizaje no supervisado para la generación de grupos y modelos supervisados para la evaluación de cada uno de ellos. Tomando como referencia este modelo, nuestra propuesta busca la mejora de éste al agregar las etapas de selección de características y el preprocesamiento de los datos para el caso particular del conjunto de EIZ. La **Figura 3.1** muestra la metodología propuesta para el diseño y evaluación de los clasificadores. La primera tarea de desarrollo de un sistema inteligente es la adquisición de los datos con los que se realizó el entrenamiento del modelo.

Para el desarrollo de esta tesis, el conjunto de datos utilizados fue adquirido por medio de un convenio de confidencialidad con la institución financiera EIZ, donde fue necesario realizar la tarea de análisis de datos, limpieza, categorización y unificación de registros explicados en la sección anterior. Además, para la validación de la metodología utilizada fue necesario el uso de bases de datos libres disponibles en el repositorio de la *UC Irvine Machine Learning Repository*, estos conjuntos de datos han sido utilizados ampliamente para la predicción del riesgo crediticio por diversos autores (Gicić et al., 2023; Bhattacharya et al., 2023; Hayashi, 2022; Yan, 2023; Singh et al., 2021).

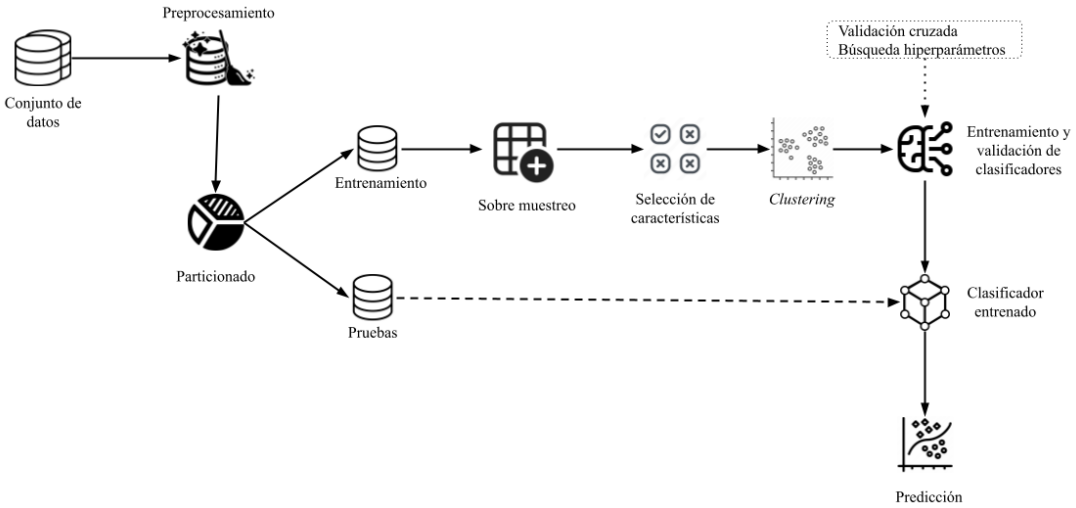


Figura 3.1: Representación de la metodología utilizada.

La fase del preprocesamiento consiste en normalizar el conjunto de datos entre los valores de 0 y 1. El particionado consiste en dividir el conjunto de datos de manera aleatoria, en los experimentos realizados se hizo la división de 80 % para el conjunto de entrenamiento y 20 % para las pruebas, esto se aprecia en la parte inicial de la metodología propuesta en la **Figura 3.1** cuyos pasos son revisados con mayor detalle a lo largo de este capítulo.

Debido a que los conjuntos de datos se encuentran desbalanceados, y después de un análisis del rendimiento de distintos métodos de sobre muestreo (SMOTE, SMOTENC, ROS, ADASYN), se aplica *SMOTE* para la creación de nuevas instancias de la clase minoritaria hasta mantener el mismo número de registros para ambas clases.

La selección de características consiste en aplicar distintos métodos reportados en el estado del arte para identificar las más significativas del conjunto de datos. El método de regresión lineal *LASSO* fue con el que se obtuvo un mayor rendimiento. La siguiente fase consiste en aplicar técnicas de *clustering* para determinados patrones en cada uno de los grupos creados. El

entrenamiento y validación de los clasificadores consiste en realizar el entrenamiento de los clasificadores utilizados al conjunto de entrenamiento, aplicando la búsqueda de hiperparámetros mediante el método en malla y aplicando la validación cruzada en 10 pliegues. Al finalizar esta fase, se obtiene el modelo entrenado con sus hiperparámetros optimizados.

La fase de evaluación del modelo consiste en aplicar el clasificador entrenado con el conjunto de pruebas en 50 iteraciones, donde se reporta el mayor rendimiento obtenido mediante las medidas del *accuracy* y error tipo II.

3.3.1. Metodología para la segmentación de datos

Las variables numéricas presentes en el conjunto de datos tienen rangos de valores muy amplios y dispersos, por lo que los algoritmos de clasificación pueden presentar problemas que la normalización en muchos casos no puede reducir. Por ejemplo, el monto solicitado presenta valores muy dispersos, por lo que al crearse grupos de estos pueden convertirse a variables categóricas (discretización) que permiten simplificar y mejorar el rendimiento de los modelos de ML (Aggarwal et al., 2015). En este proceso los valores numéricos pueden ser difíciles de interpretar, por lo que la segmentación facilita el análisis e identificación de patrones por parte de los clasificadores.

Este proceso recorre todas las variables numéricas de la base de datos hasta llegar a la variable dependiente (*target*), como se muestra en la **Figura 3.2**. Para cada columna numérica se inicia con 2 particiones, si al aplicar la ecuación 2.8 genera valores entre 0.3 y 0.5 se termina el análisis para dicha columna, o en caso contrario se incrementa el número de particiones. Este proceso se repite hasta llegar a un valor entre 0.3 y 0.5 ó hasta tener una cantidad de particiones cercana al total de valores diferentes en esa columna (L. Chen, 2022).

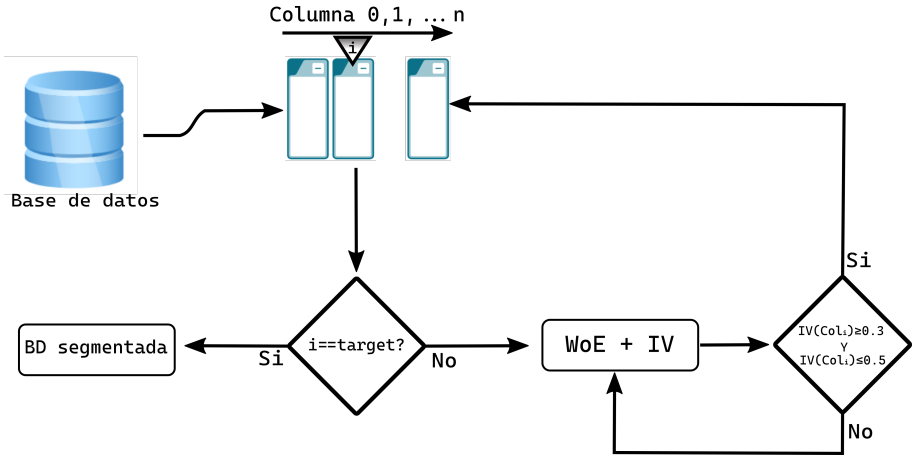


Figura 3.2: Proceso de segmentación de datos.

Al terminar este proceso, las variables numéricas independientes analizadas se convierten en categóricas, maximizando su capacidad de predicción en los clasificadores. En la **Figura 3.3** se observa el método que realiza la segmentación de los datos. De manera interna se encuentran las validaciones mencionadas con anterioridad.

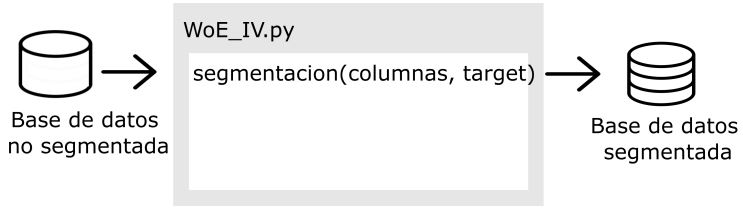


Figura 3.3: Clase para realizar la segmentación de datos.

3.3.2. Proceso para entrenamiento del modelo para detección de riesgo crediticio.

Después de realizar el proceso de segmentación de las variables numéricas (discretización), se tiene una base de datos con los atributos que tienen mayor capacidad discriminatoria para entrenar el modelo (ver **Figura 3.4**). El siguiente paso es normalizar las variables con el objetivo de que todos los atributos tengan el mismo peso al momento del entrenamiento de los modelos. Enseguida se realiza la separación de la base de datos en un conjunto de entrenamiento y otro de pruebas.

Después de aplicar normalización a todo el conjunto, éste es divido de manera aleatoria en un conjunto de entrenamiento y pruebas en una proporción 80:20. El conjunto de entrenamiento es utilizado para realizar de la selección de características con las cuales los clasificadores obtienen el mejor rendimiento al momento de realizar la predicción del riesgo de crédito. El método utilizado es el llamado *LASSO*, es un método de regresión lineal y regularización de características, donde la selección de características se realiza de manera automática, reduciendo a cero los coeficientes de las características menos significativas mediante la fórmula 2.3 descrita en el Capítulo 2. Este método se encuentra dentro de las librerías de *sklearn* en *python*, por lo que se utiliza en el proceso de solo se realiza la ejecución de dicha biblioteca.

A las características más significativas se aplica el método de *clustering* a cada instancia del conjunto de entrenamiento y de pruebas para asignarle un grupo. Además, se utiliza *PCA* para la visualización de los datos. Para cada grupo obtenido se entrena y obtiene un clasificador, y se realiza el ajuste de los hiperparámetros mediante la búsqueda en malla y la validación cruzada en 10 pliegues.

Para la evaluación del modelo se filtran las características seleccionadas durante el entrenamiento, enseguida se aplican los modelos de *clustering* para identificar el grupo al que pertenece cada instancia del conjunto de pruebas, y finalmente, se realiza una predicción aplicando el modelo generado para dicho grupo. La evaluación de las predicciones resultantes para cada grupo son utilizadas para la obtención del rendimiento sobre este conjunto.

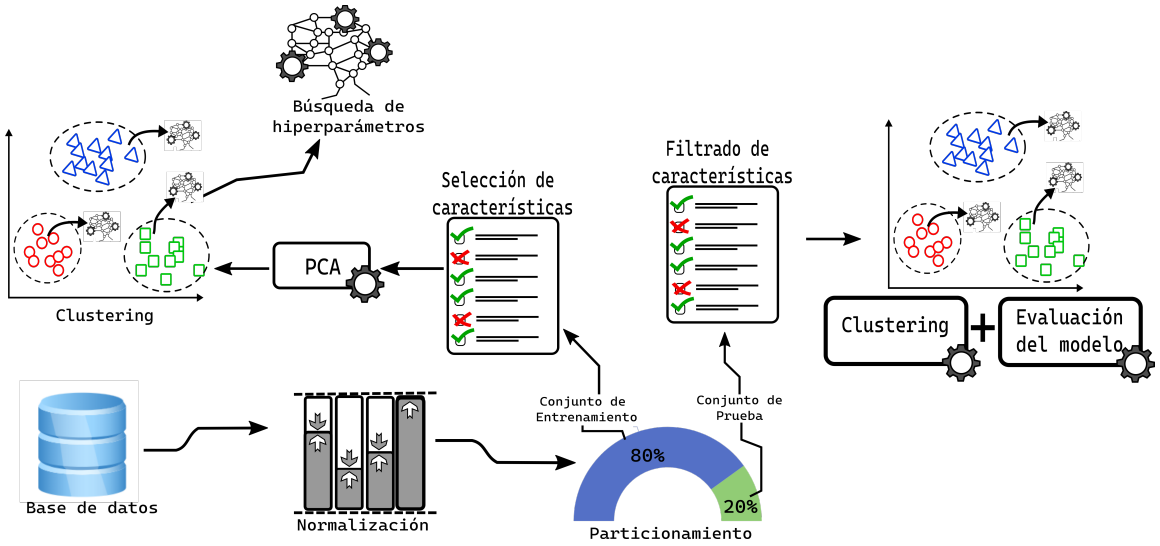


Figura 3.4: Proceso de entrenamiento del modelo.

En la **Figura 3.5** se observa los métodos definidos en la biblioteca desarrollada para la construcción del modelo de identificación del riesgo crediticio utilizando la metodología propuesta.

Los datos segmentados son la entrada que recibe este archivo, estos pueden leerse con el método `lectura_datos` que de manera interna realiza la normalización de los datos. Los datos pueden ser separados usando el método `split_train_test`, por default estos son separados en una proporción 80 : 20 de manera aleatoria y se guardan en variables internas.

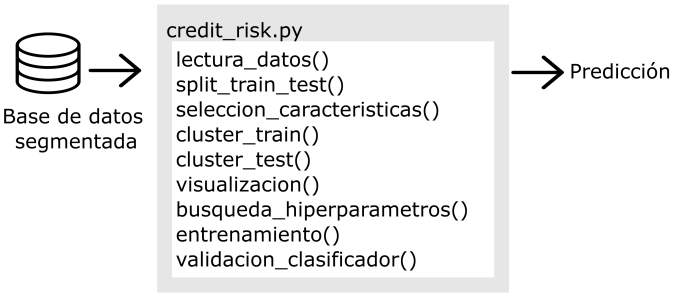


Figura 3.5: Métodos utilizados para la identificación del riesgo crediticio.

El método de `cluster_train` hace el entrenamiento de un modelo para cada clúster para poderlo posteriormente evaluar usando `cluster_test`. El método de *visualización* genera gráficas para mostrar la distribución de los datos en los *clústers* o grupos generados usando el algoritmo PCA con las 2 componentes principales de mayor importancia (ver **Figura 4.7**).

El método de `busqueda_hiperparametros` obtiene los hiperparámetros óptimos para cada grupo y clasificador utilizado. El método `entrenamiento` realizará el entrenamiento del clasificador con los hiperparámetros obtenidos.

El método de `validacion_clasificador` obtiene las métricas de rendimiento *Accuracy*, error tipo I y II y la *MCC* obtenidas por los clasificadores.

Capítulo 4

Pruebas y Resultados

En este cápitulo se muestran los resultados obtenidos del entrenamiento de los clasificadores *DT, RNN, XGBoost* y *SVM* utilizando 3 bases de datos públicas y otra de una institución financiera privada. Así mismo, se presentan los hiperparámetros obtenidos por los modelos de ML al realizar la búsqueda en malla y seleccionar aquellas que presentan el mejor comportamiento en cuanto a la medida de rendimiento (*accuracy*) con el conjunto de entrenamiento y validación. Se describen los resultados de implementar la selección de características mediante el método de regresión lineal *LASSO*, a este conjunto de características se aplica la reducción de dimensionalidad (PCA) para la aplicación de los algoritmos de *clustering k-means* y *DBSCAN* como se observa en la **Figura 3.4**.

Se han tomado como referencia trabajos relacionados que tratan el problema del riesgo crediticio, que han sido publicados recientemente para las bases de datos Alemana, Australiana y Japonesa. Con motivos de validación de la metodología propuesta, se utilizaron los conjuntos de datos que han sido ampliamente utilizados en el estado del arte. Los conjuntos de datos se obtuvieron del *UCI Machine Learning Repository* (Dua y Graff, 2017). Estos conjuntos de datos contienen información de instituciones financieras con presencia en Alemania, Japón y Australia. El conjunto de datos de Australia contiene 6 características numéricas y 8 no numéricas, todas ellas anonimizadas. El conjunto de datos de Japón contiene 6 características numéricas y 9 no numéricas, todas ellas anonimizadas. El conjunto de datos de Alemania contiene 7 características numéricas y 13 no numéricas, estado de la cuenta corriente existente, duración, historial crediticio, finalidad, importe del crédito, cuentas de ahorro, empleo actual desde entonces, tasa de pago a plazos en porcentaje de la renta disponible, estado personal y sexo, otros deudores, residencia actual desde entonces, propiedad, edad en años, otros planes de pago a plazos, vivienda, número de créditos existentes en este banco, trabajo, número de dependientes económicos, teléfono y trabajador extranjero.

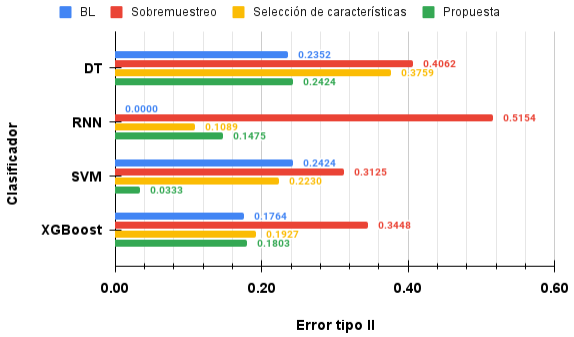
Diversos autores han experimentado con distintos métodos para mejorar el rendimiento de los modelos entrenados, como selección de características y de sobre muestreo, para obtener un conjunto de entrenamiento equilibrado (Zhou et al., 2021; Jemai y Zarrad, 2023; Dahiya et al., 2017; Hajek y Michalak, 2013; Hassani et al., 2020). Siendo ésta la razón por la cual se propone la metodología descrita (ver **Figura 3.1**) de la cual se analizan los resultados en los siguientes apartados.

Para evaluar el rendimiento obtenido por cada clasificador en cada conjunto de datos utilizado, se tomaron como referencia las medidas de rendimiento: *accuracy*, error tipo I y II y la *MCC*. El error tipo II representa los falsos positivos que son aquellas personas que los modelos han clasificado como no morosos; sin embargo, son personas que llegarán a presentar retraso en el cumplimiento de sus obligaciones al solicitar un préstamo. Además, la medida de rendimiento

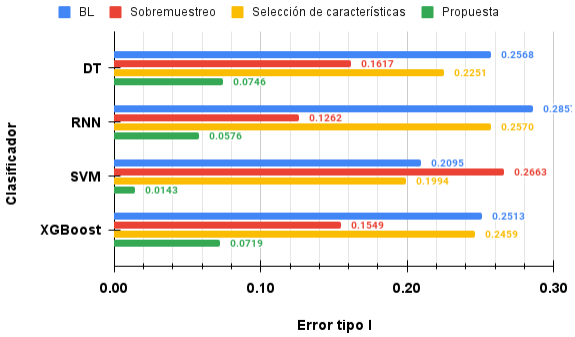
MCC con los valores cercanos a 1 indican que el modelo construido se encuentra realizando una buena generalización de los registros, separando correctamente los morosos de los no morosos. A continuación se describen los resultados obtenidos con las diferentes bases empleadas.

4.1. Base de datos *German Credit Data* (GCD)

En la **Figura 4.1** se observa el rendimiento obtenido por los clasificadores en las medidas de rendimiento del error tipo I y II, aplicados al conjunto de datos de crédito *GCD*, el experimento base consiste en la evaluación de los clasificadores sin utilizar preprocesamiento al conjunto de entrenamiento y se encuentra representado con las barras de color azul y las siglas *BL*, la barra de color rojo muestra el rendimiento de los clasificadores al aplicar el método de sobre muestreo, la barra de color amarillo muestra el rendimiento obtenido por el método de selección de características y la barra de color verde muestra el rendimiento obtenido con la metodología propuesta. En el eje X de la **Figura 4.1a** y **4.1b** se muestra el error tipo II y I obtenido por cada uno de los clasificadores que se observan en el eje Y. Como se observa en la **Figura 4.1**, los resultados obtenidos por la evaluación de los clasificadores demuestran que la implementación de la metodología propuesta representada por la barra de color verde (ver **Figura 3.1**) obtiene el mejor rendimiento en cuanto al error tipo II y I, para realizar la predicción del riesgo crediticio.



(a) Rendimiento comparativo del error tipo II para la base de datos *GCD*.

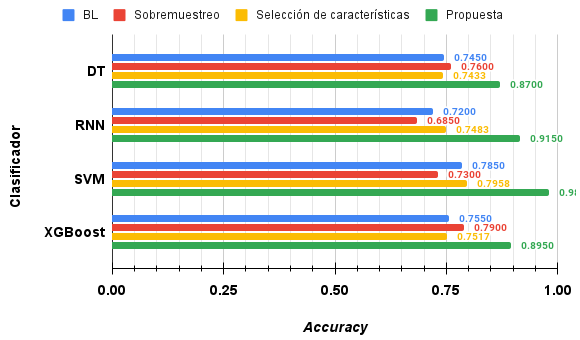


(b) Rendimiento comparativo del error tipo I para la base de datos *GCD*.

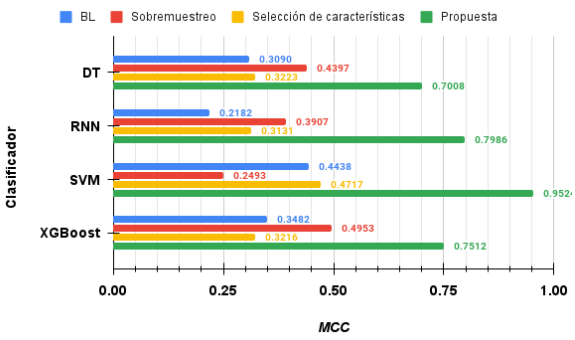
Figura 4.1: Rendimiento obtenido mediante el error tipo I y II para el conjunto *GCD*.

Para el conjunto de datos *GCD*, el uso de la metodología propuesta obtiene el mejor rendimiento en cuanto al error tipo II y I comparado con los obtenidos por los otros experimentos (excepto para *RNN* en el error tipo II sin emplear ningún método). Además, se observa que en general el utilizar la metodología propuesta, reduce el error tipo I para la mayoría de los clasificadores. A pesar de que las *RNN* obtienen el mejor rendimiento en el error tipo II para el experimento base, el error tipo I se incrementa, por lo tanto, este no es el más indicado para la predicción del riesgo crediticio para el conjunto de datos. Además, se confirma la superioridad de la metodología propuesta mostrada en color verde mediante el clasificador de las *SVM* para las medidas del error.

La **Figura 4.2** muestra las medidas de rendimiento del *accuracy* y la *MCC*, como se puede observar, en la **Figura 4.2a** el utilizar la metodología propuesta obtiene el mayor *accuracy* al igual que en la medida del *MCC* mostrada en la **Figura 4.2b**. Por lo tanto, se podría confirmar que para el problema de identificación del riesgo crediticio en el conjunto de datos *GCD*, las *SVM* son las que tienen el mayor rendimiento en las métricas analizadas.



(a) Rendimiento comparativo del *accuracy* para la base de datos *GCD*.

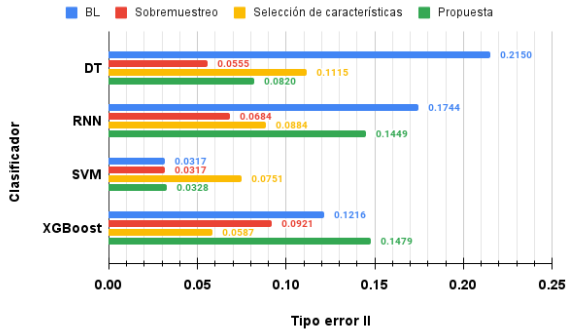


(b) Rendimiento comparativo de la *MCC* para la base de datos *GCD*.

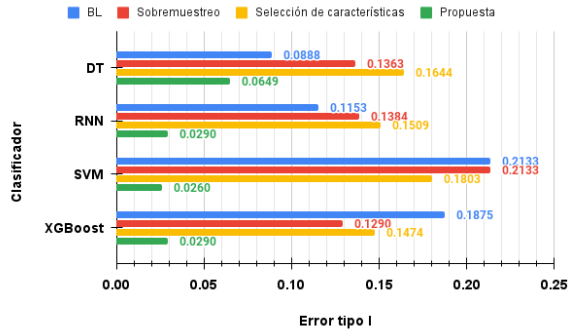
Figura 4.2: Rendimiento obtenido mediante el *accuracy* y la *MCC* para el conjunto *GCD*.

4.2. Base de datos Japonesa

La **Figura 4.3** muestra el rendimiento obtenido con la medida del *accuracy* y la *MCC*, como se observa, mediante la metodología propuesta las *SVM* obtienen el mejor rendimiento. Además, aparentemente para este conjunto de datos resulta conveniente utilizar los métodos de sobre muestreo y selección de características de manera individual para el caso de reportar el error tipo II; sin embargo, en la medida de la *MCC* resulta contradictoria, ya que tienen un rendimiento menor al obtenido utilizando con la metodología propuesta donde se logra realizar de mejor manera la generalización de los clasificadores.



(a) Rendimiento comparativo del error tipo II para la base de datos Japonesa.

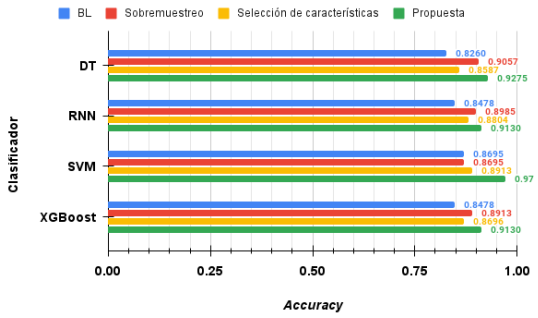


(b) Rendimiento comparativo del error tipo I para la base de datos Japonesa.

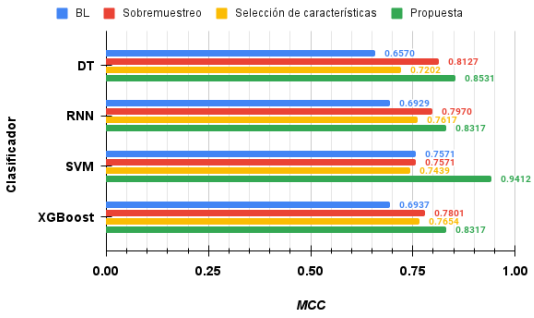
Figura 4.3: Rendimiento obtenido mediante el error tipo I y II para la base de datos Japonesa.

Para el conjunto de datos japones, el uso de la metodología propuesta obtiene el mejor rendimiento en cuanto al error tipo II y I comparado con los obtenidos por los otros experimentos. Además, se observa que en general el utilizar la metodología propuesta, reduce el error tipo I para la mayoría de los clasificadores. Además, se confirma la superioridad de la metodología propuesta mostrada en color verde mediante el clasificador de las SVM para las medidas del error.

La **Figura 4.4** muestra las medidas de rendimiento del accuracy y la MCC, como se puede observar, en la **Figura 4.4a** el utilizar la metodología propuesta obtiene el mayor *accuracy* al igual que en la medida del *MCC* mostrada en la **Figura 4.4b**. Por lo tanto, se podría confirmar que para el problema de identificación del riesgo crediticio en el conjunto de datos japonesa, las SVM son las que tienen el mayor rendimiento en las métricas analizadas



(a) Rendimiento comparativo del *accuracy* para la base de datos Japonesa.

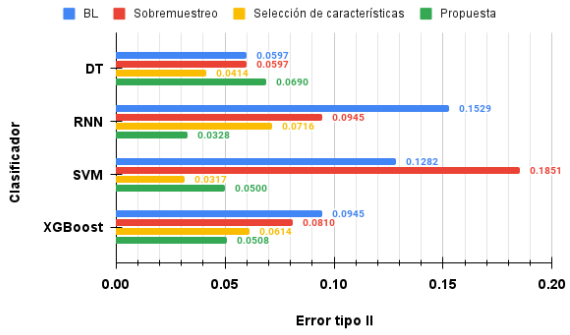


(b) Rendimiento comparativo de la *MCC* para la base de datos Japonesa.

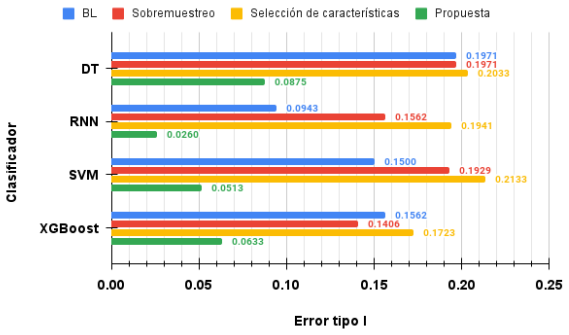
Figura 4.4: Rendimiento obtenido mediante *accuracy* y la *MCC* para la base de datos Japonesa.

4.3. Base de datos Australiana

En la **Figura 4.5** se muestra el rendimiento de los clasificadores al evaluar la base de datos australiana, el experimento base (evaluación de los clasificadores sin utilizar ninguno de los métodos utilizados y se encuentra representado con las barras de color azul (BL)), sobre muestreo, selección de características y la implementación de manera conjunta de estas dos últimas. En el eje X, se muestra el error tipo II y I obtenido por cada uno de los clasificadores en el eje Y. Como se observa, la evaluación realizada con la metodología propuesta tiene un rendimiento superior a la mayoría de las otras evaluaciones en la medida de rendimiento del error II y I. Cabe destacar que el clasificador *RNN* presenta al mayor rendimiento con la metodología propuesta en la medida del rendimiento del error tipo I y II. También se observa que la metodología propuesta es la que presenta mejor desempeño en los clasificadores para ambos tipos de error y se ve reflejado en el *accuracy* y *MCC*.



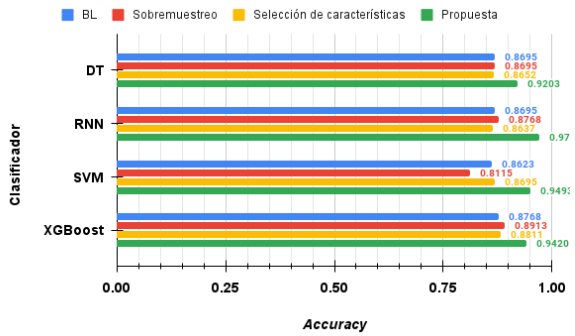
(a) Rendimiento comparativo del error tipo II para la base de datos Australiana.



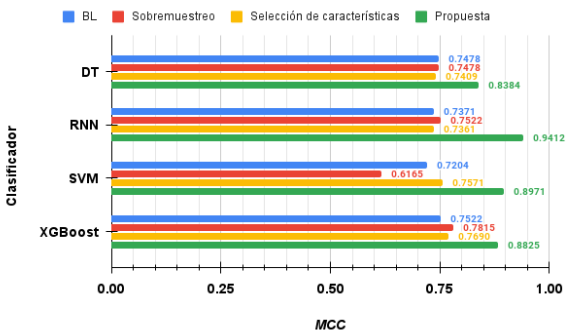
(b) Rendimiento comparativo del error tipo I para la base de datos Australiana.

Figura 4.5: Rendimiento obtenido mediante el error tipo II y I para la base de datos Australiana.

La **Figura 4.5** muestra el rendimiento obtenido con la medida del *accuracy* y la *MCC*. Puede observarse que el uso de la metodología propuesta, las *RNN* obtienen el mejor rendimiento. Como puede observarse, el uso de la metodología propuesta ayudar a minimizar el error tipo II y I, y mediante la *MCC* se puede confirmar que efectivamente el clasificador se encuentra generalizando con el mínimo margen de error ambas clases.



(a) Rendimiento comparativo del *accuracy* para la base de datos Australiana.



(b) Rendimiento comparativo de la *MCC* para la base de datos Australiana.

Figura 4.6: Rendimiento obtenido mediante *accuracy* y la *MCC* para la base de datos Australiana.

4.4. Base de datos de la entidad financiera EIZ

En este apartado se analizan los resultados del experimento utilizando la metodología propuesta con la base de datos de la entidad financiera. La identificación de características permitió identificar aquellas más significativas para predecir el riesgo crediticio. Utilizando la selección de características y aplicando los métodos descritos en el estado del arte, de un total de 23 utilizadas para la construcción de la base de datos, las características más significativas y que representan un mayor impacto al predecir el riesgo crediticio son las siguientes:

- Edad.
- Código postal.
- Tipo de vivienda.
- Dependientes.
- Estado civil.
- Género.
- Ingreso.
- Egreso.
- Tipo de préstamo.
- Tasa interés normal.
- Tasa interés moratoria.
- Monto solicitado.
- Avales.
- Créditos trabajados.
- Plazo.

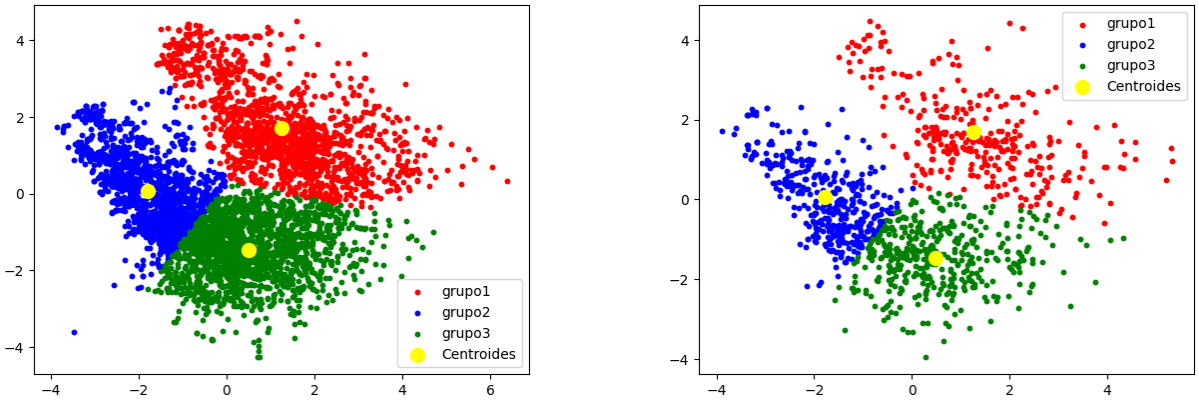
Utilizando el preprocesamiento de datos mediante la *WoE IV*, normalización de los datos, la selección de características y *clustering* utilizando la metodología mostrada en la **Figura 1.2**, se obtiene la segmentación de los datos en los grupos mostrados en la **Tabla 4.1**:

Tabla 4.1: Separación de grupos utilizando el algoritmo *k-means*.

Grupo	**Clasificación**	**Número de instancias**
Grupo 0	Moroso	1,176
	No moroso	984
Grupo 1	Moroso	846
	No moroso	580
Grupo 2	Moroso	550
	No moroso	1,008

En la **Figura 4.7** se puede apreciar la separación de los datos en los 3 grupos formados y

descritos en la **Tabla 4.1**.



(a) Representación gráfica de los grupos formados en el conjunto de entrenamiento.

(b) Representación gráfica de los grupos formados en el conjunto de pruebas.

Figura 4.7: Representación gráfica de los grupos formados

Además, considerando la importancia de obtener aquellos patrones que permitan identificar con mayor precisión los créditos que presentan un riesgo para la institución financiera, se identificó en la segmentación de los datos de los *montos* que mayormente son solicitados por las personas en la institución están clasificados en los tipos 13, 1, 6 y 2 (ver **Tabla A.5**), como se observa en la **Tabla 4.2**:

Tabla 4.2: Montos mayormente solicitados.

Segmentación	Clasificación	Números
13	Moroso	306
	No moroso	583

Segmentación	Clasificación	Números
1	Moroso	327
	No moroso	526

Segmentación	Clasificación	Números
6	Moroso	359
	No moroso	365

Segmentación	Clasificación	Números
2	Moroso	343
	No moroso	172

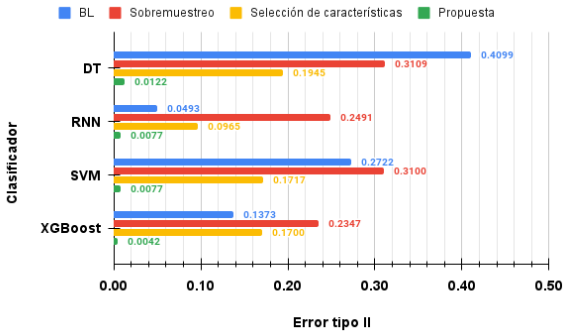
Así mismo, se identificó la segmentación del código postal 2, 0, 7 y 5 (ver **Tabla A.5**), este representa aquellas personas que pertenecen a las localidades con códigos postales que solicitan un mayor número de solicitudes de crédito, como se muestra en la **Tabla 4.3**.

Tabla 4.3: Segmentación de los datos mediante el código postal.

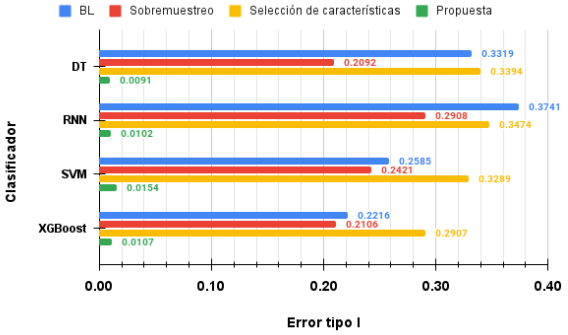
Segmentación	**Clasificación**	**Conteo**
2	**Moroso**	879
	No moroso	702
0	**Moroso**	570
	No moroso	527
7	**Moroso**	187
	No moroso	518
5	**Moroso**	249
	No moroso	429

En la **Figura 4.8** se muestra el rendimiento de los clasificadores en el conjunto de datos de la entidad financiera EIZ, utilizando el experimento base (evaluación de los clasificadores sin utilizar ninguno de los métodos utilizados y se encuentra representado con las barras de color azul (BL)), sobre muestreo, selección de características y la implementación de manera conjunta de estas dos últimas mediante la metodología propuesta.

En el eje X se muestra el error tipo II y I obtenido por cada uno de los clasificadores que se observan en el eje Y. Como se observa, la evaluación realizada con el método propuesto tiene un rendimiento superior en comparación con las otras evaluaciones en la medida de rendimiento del error II y I. Además, al utilizar la metodología, todos los clasificadores logran minimizar las medidas del error tipo I y II, sin embargo, el clasificador *XGBoost* presenta el mejor rendimiento en cuanto a estas medidas.



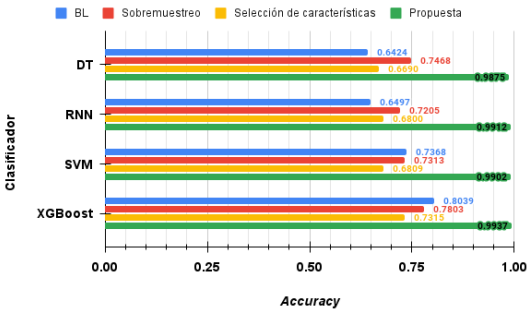
(a) Rendimiento comparativo del error tipo II para la base de datos EIZ.



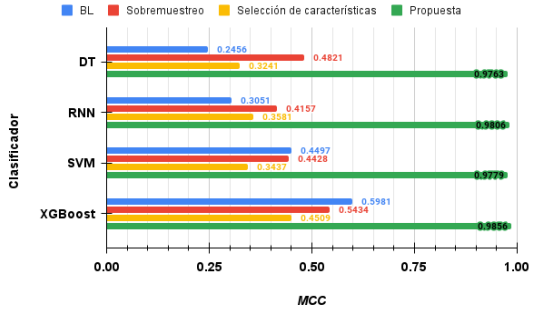
(b) Rendimiento comparativo del error tipo I para la base de datos EIZ.

Figura 4.8: Rendimiento obtenido mediante el error tipo I y II para la base de datos EIZ.

La **Figura 4.9a** muestra el rendimiento obtenido con la medida del *accuracy*. Como se observa, mediante la metodología propuesta los cuatro clasificadores propuestos obtienen un mejor rendimiento en comparación con el obtenido con los otros experimentos. La **4.9b** muestra el rendimiento obtenido con la medida de la *MCC*. Como se observa, mediante la metodología propuesta los cuatro clasificadores propuestos obtienen un mejor rendimiento en comparación con el obtenido con los otros experimentos. Además, el uso de la metodología ayuda a minimizar el error tipo II y I, y mediante la *MCC* podemos confirmar que efectivamente el clasificador se encuentra generalizando con el mínimo margen de error.



(a) Rendimiento obtenido del *accuracy* para la base de datos EIZ.



(b) Rendimiento obtenido de la *MCC* para la base de datos EIZ.

Figura 4.9: Rendimiento obtenido mediante *accuracy* y la *MCC* para la base de datos EIZ.

Capítulo 5

Conclusiones

La alta dimensionalidad y el desbalance de los conjuntos de datos para la evaluación del riesgo crediticio son factores que influyen en el rendimiento de los clasificadores. Por tanto, en esta tesis se propuso una metodología en la cual se combinan los métodos de selección de características, métodos de sobre muestreo y agrupamiento para la evaluación del riesgo crediticio. Así mismo se presenta el rendimiento obtenido de los clasificadores en 3 bases de datos que han sido utilizados ampliamente por diversos autores para la predicción del riesgo crediticio.

A partir de un convenio firmado con la institución financiera EIZ con presencia en el estado de Oaxaca, se creó un nuevo conjunto de datos. Con la aplicación de la metodología propuesta en los conjuntos de datos antes citados, se obtuvo un modelo de clasificación con métricas de rendimiento comparables o superiores a las obtenidas en los otros 3 conjuntos de dominio público analizados y reportados por otros autores en el estado del arte.

Las variables que tienen mayor capacidad de clasificación para el conjunto de datos EIZ son consistentes con las presentes en el conjunto Alemán que fueron la referencia para creación del conjunto de datos EIZ. Los conjuntos de datos japonés y australiano no tienen información sobre las variables, por lo que no puede establecerse una relación.

La precisión del clasificador para la gestión de riesgos en las instituciones de crédito es importante para su salud financiera, porque puede generar ganancias importantes y minimizar pérdidas. Los resultados obtenidos indican que la combinación de las características definidas por el experto y reforzado por las técnicas de selección permitieron construir un modelo de ML con alta capacidad de generalización para el proceso de predicción del incumplimiento de las solicitudes de crédito para las SOCAP del Estado de Oaxaca. El mismo método fue aplicado a bases de datos públicas, y el rendimiento obtenido por el modelo propuesto sobrepasa a los presentados por otros autores en el estado del arte.

Con los métodos de selección de características, sobre muestreo de información y agrupamiento, permiten una mejora significativa al momento de realizar la tarea de la evaluación del riesgo crediticio, cumpliendo de esta manera con el objetivo planteado en esta tesis. Además, en futuras investigaciones, se pretende mejorar la interpretabilidad de los resultados conforme al éxito y fracaso de los clasificadores. Donde se pretende desarrollar un algoritmo de selector de características basado en *clustering* con la finalidad de mejorar la interpretabilidad de los datos y facilitar la toma de decisiones en las instituciones financieras.

El análisis de las características más importantes obtenidas por el método de selección de características presentan una concordancia con aquellas analizadas por los expertos de las instituciones financieras. Entre estas características se encuentran la edad, monto de préstamo,

plazo, ocupación, historial crediticio, finalidad del crédito, ingresos y egresos y garantías.

Las aportaciones del desarrollo de este trabajo son el conjunto de datos EIZ creado a partir de la información proporcionada por la institución financiera, la metodología para el preprocesamiento y los modelos de clasificación que permiten predecir el incumplimiento de las solicitudes de crédito para el caso de estudio, cumpliendo así con la hipótesis planteada. Además, estos modelos reducen sustancialmente el error tipo II que permite detectar con mayor exactitud las solicitudes de créditos que al ser aprobados corresponden a potenciales deudores.

Trabajo futuro

Los resultados obtenidos en esta tesis nos motivan a:

1. Los métodos de selección de características y de sobre muestreo de la información fueron utilizados para incrementar el rendimiento de los clasificadores; sin embargo, al análisis de las características y métodos de sobre muestreo se puede agregar la distinción de características numéricas y categóricas.

2. Así mismo, se pretende analizar el rendimiento y propuesta de métodos de selección de instancias para realizar la limpieza de los conjuntos de datos, eliminando aquellas instancias con ruido presentes en los conjuntos de datos para la evaluación del riesgo crediticio.

Por lo tanto, como trabajo futuro se pretende desarrollar un modelo de selección de características basada en modelos de *clustering* basado en la densidad de la información para la predicción del riesgo crediticio y sector financiero.

Participaciones

Derivado a esta tesis y el análisis de distintos métodos, se presentaron las siguientes presentaciones:

1. Erwis Melchor-Pérez, Moises Emmanuel Ramírez-Guzmán, Araceli Hernández-Jímenez. **Análisis de riesgo crediticio para las SOCAP utilizando aprendizaje automático**. EL CONGRESO INTERNACIONAL DE ECONOMÍA FINANCIERA Y ADMINISTRACIÓN DE RIESGOS 2022.

2. Erwis Melchor-Pérez, Moises Emmanuel Ramírez-Guzmán, Araceli Hernández-Jímenez. **Aprendizaje Automático y SMOTE para identificar el riesgo crediticio en una Microfinanciera**. XIII Congreso Internacional de Contaduría, Administración, Mercadotecnia e Informática Administrativa 2022.

3. Erwis Melchor-Pérez, Moises Emmanuel Ramírez-Guzmán, Araceli Hernández-Jímenez. **Análisis de riesgo crediticio para las SOCAP usando técnicas de selección de características**. EL CONGRESO INTERNACIONAL DE ECONOMÍA FINANCIERA Y ADMINISTRACIÓN DE RIESGOS 2023.

4. Erwis Melchor-Pérez, Moises Emmanuel Ramírez-Guzmán, Araceli Hernández-Jímenez, Agustin Santiago-Alvarado. **Predicción del riesgo crediticio a microfinanciera usando aprendizaje computacional** Revista Mexicana de Economía y Finanzas Nueva Época (REMEF) (*The Mexican Journal of Economics and Finance*)

5. (En revisión) Erwis Melchor-Pérez, Moises Emmanuel Ramírez-Guzmán, Araceli Hernández-Jímenez. **Artificial neural network algorithm applied to classification of SOCAP credit applications**. Revista del Centro de Investigación de la Universidad la Salle, (Revista indexada el CONACyT).

6. Erwis Melchor-Pérez, Itahí Leticia Pérez-Feria. **Importancia del preprocesamiento y aprendizaje computacional en las finanzas**. 5^{to} Encuentro de Ciencias Empresariales. Universidad del Istmo, campus Ixtepec.

7. (En revisión) Erwis Melchor-Pérez, José Francisco Martínez-Trinidad, Jesús Ariel Carrasco-Ochoa, Agustín Santiago-Alvarado. *Oversampling and instance selection: an experimental analysis to predict credit risk.*

Referencias

Abedin, M. Z., Guotai, C., Hajek, P., y Zhang, T. (2022). Combining weighted SMOTE with ensemble learning for the class-imbalanced prediction of small business credit risk. *Complex & Intelligent Systems*, 1–21. doi: https://doi.org/10.1007/s40747-021-00614-4

Abraham, A. (2005). Nature and scope of AI techniques. *Handbook of measuring system design*.

Aggarwal, C. C., y cols. (2015). *Data mining: the textbook* (Vol. 1). Springer.

Agrawal, R., y Srikant, R. (2000). Privacy-preserving data mining. En *Proceedings of the 2000 ACM SIGMOD international conference on management of data* (pp. 439–450).

Aguilar Castillo, E. C. (2021). *Auditoría informática para detección de valores anómalos, caso de estudio: créditos concedidos en la asociación de empleados banco pichincha.* (B.S. thesis). Quito: UCE.

Akobeng, A. K. (2016). Understanding type i and type ii errors, statistical power and sample size. *Acta Paediatrica*, *105*(6), 605–609.

Amézquita Reyes, J. A., León Castro, J. G., y cols. (2012). *Una aproximación a un modelo de credit scoring aplicado a la etapa de otorgamiento en una entidad financiera colombiana. un enfoque desde la lógica difusa y los algoritmos genéticos.* (B.S. thesis). Universidad Piloto de Colombia.

Angelini, E., Di Tollo, G., y Roli, A. (2008). A neural network approach for credit risk evaluation. *The quarterly review of economics and finance*, *48*(4), 733–755. doi: https://doi.org/10.1016/j.qref.2007.04.001

Banxico. (2022). *Riesgos en las infraestructuras de los mercados financieros.* Descargado 22 de Marzo de 2022, de `http://educa.banxico.org.mx/banco_mexico_banca_central/sistema-pago-riesgo-infaestru.html` (Banco de México)

Barajas-Juárez, A. U., Gutierrez-Cruz, D., Rodríguez-Paéz, C. L., y Durán López, V. (2019). Indicadores para la aprobación de créditos bancarios con algoritmos genéticos. *Revista Aristas: Investigación Básica y Aplicada*, *7*(14).

Barrios, J. I. (2019). La matriz de confusión y sus métricas. *BigData*.

BBVA. (2022). *Institución de crédito.* Descargado 04 de Marzo de 2022, de `https://www.bbva.mx/educacion-financiera/i/institucion_de_credito.html` (Institución de crédito)

Bhattacharya, A., Biswas, S. K., y Mandal, A. (2023). Credit risk evaluation: a comprehensive study. *Multimedia Tools and Applications*, *82*(12), 18217–18267. doi: https://doi.org/10.1007/s11042-022-13952-3

Borja-Robalino, R., Monleon-Getino, A., y Rodellar, J. (2020). Estandarización de métricas de rendimiento para clasificadores machine y deep learning. *Revista Ibérica de Sistemas e Tecnologias de Informação*(E30), 184–196.

Borland, L., Plastino, A. R., y Tsallis, C. (1998). Information gain within nonextensive thermostatistics. *Journal of Mathematical Physics*, *39*(12), 6490–6501. doi: https://doi.org/10.1063/1.532660

Breiman, L., Friedman, J. H., Olshen, R. A., y Stone, C. J. (1984). *Classification and regression trees.* Chapman & Hall.

Brown, K., y Moles, P. (2014). Credit risk management. *K. Brown & P. Moles, Credit Risk Management*, *16*.

Brown, K., y Morales, P. (2008). *Credit Risk and Management.* Edinburgh Business School.

Cabrera-Cruz, A. M. (2014). *DiseÑo de credit scoring para evaluar el riesgo crediticio en una entidad de ahorro y crÉdito popular* (Tesis). Universidad Tecnológica de la Mixteca.

Callejas, I., Piñeros, J., Rocha, J., y Ferney Hernández, F. D. (2018). Implementación de una Red Neuronal Artificial tipo SOM en una FPGA para la resolución de trayectorias tipo laberinto. *Universidad INCCA de Colombia*.

Cambronero, C. G., y Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. *Intelgencia en Redes de Comunicación, Universidad Carlos III de Madrid*, *23*.

Carrascal, J. M. V., y María, J. (2015). *Modelos de medición del riesgo de crédito* (Tesis doctoral, Memoria para optar el grado de Doctor). Universidad Complutense de Madrid.

Cauas, D. (2015). Definición de las variables, enfoque y tipo de investigación. *Bogotá: biblioteca electrónica de la universidad Nacional de Colombia*, *2*, 1–11.

Chapelle, O., Schölkopf, B., y Zien, A. (2006). *Semi-supervised learning*. The MIT Press.

Chen, L. (2022). Internet financial risk model evaluation and control decision based on big data. *Wireless Communications and Mobile Computing*, *2022*. doi: https://doi.org/10.1155/2022/8606624

Chen, T., y He, T. (2017). XGBoost: extreme gradient boosting.

Chicco, D., y Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1). doi: https://doi.org/10.1186/s12864-019-6413-7

Chicco, D., Tötsch, N., y Jurman, G. (2021). The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*(13). doi: https://doi.org/10.1186/s13040-021-00244-z

CNBV. (2016). *¿Qué es la inclusión financiera?* Descargado 22 de Marzo de 2022, de `https://www.cnbv.gob.mx/Inclusi%C3%B3n/Paginas/Descripci%C3%B3n.aspx` (Comisión Nacional Bancaria y de Valores)

CNBV. (2021a). *Comisión nacional bancaria y de valores.* Descargado 03 de Noviembre de 2021, de `https://www.banxico.org.mx/SieInternet/consultarDirectorioInternetAction.do?sector=19&accion=consultarCuadro&idCuadro=CF829&locale=es` (Banca comercial, Crédito por entidad federativa, sector económico y situación de la cartera)

CNBV. (2021b). *Sociedades cooperativas de ahorro y préstamo (socap).* Descargado 08 de Noviembre de 2021, de `https://www.gob.mx/cnbv/acciones-y-programas/sociedades-cooperativas-de-ahorro-y-prestamo-socap` (Sociedades Cooperativas de Ahorro y Préstamo (Socap))

CNBV. (2022). *Reglas de capitalización.* Descargado 22 de Marzo de 2022, de `https://www.gob.mx/cnbv/acciones-y-programas/basilea-compliance` (Comisión Nacional Bancaria y de Valores)

CONCAMEX. (2021). *Confederación de Cooperativas de Ahorro y Préstamo de México.* Descargado 20 de Marzo de 2022, de `https://www.concamex.coop/es/` (Banca comercial, Crédito por entidad federativa, sector económico y situación de la cartera)

Condusef. (2021). *Sociedades cooperativas de ahorro y préstamo (SOCAP).* Descargado 20 de Octubre de 2021, de `https://www.condusef.gob.mx/?p=mapa-socap` (Sociedades Cooperativas de Ahorro y Préstamo (SOCAP))

Credimejora. (2022). *Importancia del crédito en la vida financiera.* Descargado 04 de Marzo de 2022, de `https://www.credimejora.com/informacion-hipotecaria/importancia-tener-credito`

Dahiya, S., Handa, S., y Singh, N. (2017). A feature selection enabled hybrid-bagging algorithm for credit risk evaluation. *Expert Systems*, *34*(6), e12217. doi: https://doi.org/10.1111/exsy.12217

Danjuma, K. J. (2015). Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. *arXiv preprint arXiv:1504.04646*. doi: https://doi.org/10.48550/arXiv.1504.04646

Dastile, X., y Celik, T. (2021). Making deep learning-based predictions for credit scoring explainable. *IEEEAccess*, *9*, 50426-50440. doi: 10.1109/ACCESS.2021.3068854

de Lara, A. (2008). *Medición y control de riesgos financieros* (3ra. Edición ed.). Limusa.

Delgado, G. S. P., Farroñan, E. V. R., y Falcon, A. W. C. (2020). La morosidad ante un confinamiento del Covid-19 en la caja rural de ahorro y crédito raíz, Perú. *Investigación Valdizana*, *14*(4), 206–212.

Demirgüç-Kunt, A., y Singer, D. (2017). Financial inclusion and inclusive growth: A review of recent empirical evidence. *World Bank Policy Research Working Paper*(8040).

Deng, L., y Yu, D. (2014). *Deep learning methods and applications.* Foundations and Trends in Signal Processing.

Dougherty, G. (2012). *Pattern recognition and classification: an introduction.* Springer Science & Business Media.

Dua, D., y Graff, C. (2017). *UCI machine learning repository.* Descargado de `http://archive.ics.uci.edu/ml`

Fausett, L. V. (1993). *Fundamentals of neural networks: Architectures, algorithms and applications* (1ra Edición ed.). Prentice Hall PTR.

Fica, A. L. L., Casanova, M. A. A., y Mardones, J. G. (2018). Análisis de riesgo crediticio, propuesta del modelo credit scoring. *Revista Facultad de Ciencias Económicas*, *26*(1), 181–207. doi: https://doi.org/10.18359/rfce.2666

Franco, E. G., y Chang, L. P. (2017). Modelos predictor de la morosidad con variables macroeconómicas. *Revista Ciencia Unemi*, *11*(26). doi: https://doi.org/10.1016/j.eswa.2015.06.001

Freire-López, J. (2021). *Modelo de clasificación de riesgo crediticio utilizando random forest en financiera del ecuador* (Tesis, para optar el Título de Master en Sistemas de Información con Mención en Data Science).

Friedman, J., Hastie, T., y Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337–407. doi: 10.1214/aos/1016218223

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.*, *19*(1). doi: 10.1214/aos/1176347963

Galiano, F. B. (2002). *Un método alternativo para la construcción de árboles de decisión* (TESIS DOCTORAL). UNIVERSIDAD DE GRANADA E.T.S. INGENIERÍA INFORMÁTICA.

García, J. C. T., García, M. Á. M., y Martínez, F. V. (2017). Administración del riesgo crediticio al menudeo en México: una mejora econométrica en la selección de variables y cambios en sus características. *Contaduría y administración*, *62*(2), 377–398. doi: https://doi.org/10.1016/j.cya.2017.01.003

García, J. C. T., Bolívar, H. R., y Vázquez, F. A. (2016). Actualización del modelo de riesgo crediticio, una necesidad para la banca revolvente en México. *Revista Finanza y Política Económica*, *8*(1). doi: http://dx.doi.org/10.14718/revfinanzpolitecon.2016.8.1.2

Ghayoumi, M. (2022). *Deep learning in practice.* CRC Press.

Ghodselahi, A., y Amirmadhi, A. (2011). Application of artificial intelligence techniques for credit risk evaluation. *International Journal of Modeling and Optimization*, *1*(3), 243.

Gicić, A., Đonko, D., y Subasi, A. (2023). Intelligent credit scoring using deep learning methods. *Concurrency and Computation: Practice and Experience*, *35*(9), e7637. doi: https://doi.org/10.1002/cpe.7637

González, F. A. (2015). Machine learning models in rheumatology. *ELSEVIER*, *22*(2).

Grabczewski, K. (2014). *Meta-learning in decision tree induction.* Springer.

Graupe, D. (2007). *Principles of artificial neural networks* (Vol. 6). Advanced Series on Circuits and System.

Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and tensorflow* (Second Edition ed.; O. Media, Ed.).

Hajek, P., y Michalak, K. (2013). Feature selection in corporate credit rating prediction. *Knowledge-Based Systems*, *51*, 72–84. doi: https://doi.org/10.1016/j.knosys.2013.07.008

Han, J., Kamber, M., y Pei, J. (2012). *Data mining: Concepts and techniques* (2.ª ed.). Morgan Kaufmann.

Hand, D. J., y Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the royal statistical society: series a (statistics in society)*, *160*(3), 523–541. doi: https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hassani, Z., Alambardar Meybodi, M., y Hajihashemi, V. (2020). Credit risk assessment using learning algorithms for feature selection. *Fuzzy Information and Engineering*, *12*(4), 529–544. doi: https://doi.org/10.1080/16168658.2021.1925021

Hayashi, Y. (2022). Emerging trends in deep learning for credit scoring: A review. *Electronics*, *11*(19), 3181. doi: https://doi.org/10.3390/electronics11193181

Haykin, S. (2009). *Neural networks and learning machines* (Third Edition ed.). Pearson Prentice Hall.

Hernández, C., Rodríguez, J. E. R., y cols. (2008). Preprocesamiento de datos estructurados. *Revista vínculos*, *4*(2), 27–48. doi: https://doi.org/10.14483/2322939X.4123

Hofmann, H. (1994). *Statlog (German Credit Data).* UCI Machine Learning Repository. (DOI: https://doi.org/10.24432/C5NC77)

Hssina, B., Merbouha, A., Ezzikouri, H., y Erritali, M. (2014). A comparative study of decision tree ID3 and C4.5. *International Journal of Advanced Computer Science and Applications*, *4*(2), 13–19.

Jemai, J., y Zarrad, A. (2023). Feature selection engineering for credit risk assessment in retail banking. *Information*, *14*(3), 200. doi: https://doi.org/10.3390/info14030200

Jović, A., Brkić, K., y Bogunović, N. (2015). A review of feature selection methods with applications. En *2015 38th international convention on information and communication technology, electronics and microelectronics (mipro)* (pp. 1200–1205). doi: 10.1109/MIPRO.2015.7160458

Keras. (2022). *Acerca de keras.* Descargado 10 de Junio de 2022, de `https://keras.io/about/` (Keras)

Kohavi, R., y Provost, F. (1998). Confusion matrix. *Machine learning*, *30*(2-3), 271–274.

Kotsiantis, S. B., Kanellopoulos, D., y Pintelas, P. E. (2006). Data preprocessing for supervised leaning. *International journal of computer science*, *1*(2), 111–117.

Lázaro, L. M. C., y Sosa, F. A. P. (2020). Análisis del impacto de las reformas financieras de 2014 en las sociedades cooperativas de ahorro y préstamo de méxico. *REVESCO: revista de estudios cooperativos*(135), 121–136. doi: https://dx.doi.org/10.5209/REVE.69190

LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Li, C. (2019). Preprocessing methods and pipelines of data mining: An overview. *arXiv preprint arXiv:1906.08510*. doi: https://doi.org/10.48550/arXiv.1906.08510

Li, J.-P., Mirza, N., Rahat, B., y Xiong, D. (2020). Machine learning and credit ratings prediction in the age of fourth industrial revolution. *Technological Forecasting and Social Change*, *161*, 120309. Descargado de `https://www.sciencedirect.com/science/article/pii/S0040162520311355` doi: https://doi.org/10.1016/j.techfore.2020.120309

Liu, J., Zhang, S., y Fan, H. (2022). A two-stage hybrid credit risk prediction model based on XGBoost and graph-based deep neural network. *Expert Systems with Applications*, *195*, 116624. Descargado de `https://www.sciencedirect.com/science/article/pii/S0957417422001142` doi: https://doi.org/10.1016/j.eswa.2022.116624

López Maldonado, G., y cols. (2014). *Análisis de la severidad de los accidentes de tráfico utilizando técnicas de minería de datos* (Tesis Doctoral no publicada). Universidad de Granada.

Luna Santander, F. A., y cols. (2023). Visión robótica de baja resolución, con recursos limitados.

Machado, M. R., y Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems With Applications*, *200*. doi: https://doi.org/10.1016/j.eswa.2022.116889

Madrigal, F., Chávez, L., y Díaz, A. (2017). Evaluación de las 5 cs de crédito en condiciones de incertidumbre. *de Estudios Organizacionales en las Ciencias Administrativas ante los Retos del Siglo XXI, Primera ed., México, Universidad Michoacana de San Nicolás de Hidalgo*, 2438.

Mahbobi, M., Kimiagari, S., y Vasudevan, M. (2021). Credit risk classification: an integrated predictive accuracy algorithm using artificial and deep neural networks. *Annals of Operations Research*.

Marcano-Cedeno, A., Marin-De-La-Barcena, A., Jiménez-Trillo, J., Piñuela, J., y Andina, D. (2011). Artificial metaplasticity neural network applied to credit scoring. *International journal of neural systems*, *21*(04), 311–317. doi: https://doi.org/10.1142/S0129065711002857

Medina, R. P., y Selva, M. L. M. (2013). Análisis del credit scoring. *Revista de Administração de Empresas*, *53*, 303–315. doi: https://doi.org/10.1590/S0034-75902013000300007

Muller, A. C., y Guido, S. (2017). *Introduction to machine learning with python*. O' Reilly.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. The MIT Press.

Muñoz, G. O. (2018). *Predicción de la tasa de éxito en las asignaturas de primer año para los alumnos de la universidad del bío - bío* (B.S. thesis). Universidad del Bío-Bío, Facultad de ciencias empresariales departamento de sistemas de información.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., y Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, *18*(6), 275–285. doi: https://doi.org/10.1002/cem.873

Nandy, A., y Biswas, M. (2018). *Reinforcement learning with open AI, Tensorflow and Keras using Python*. Apress. doi: https://doi.org/10.1007/978-1-4842-3285-9

Oblitas, M. M. R., Ramirez, E. T., García, W. E. V., Cárdenas, M. F. U., y Ramírez, E. C. (2021). Gestión de riesgo crediticio para afrontar la morosidad bancaria. *Revista científica Tzhoecoen*, *13*(1).

Oreski, S., Oreski, D., y Oreski, G. (2012). Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. *Expert systems with applications*, *39*(16), 12605–12617. doi: https://doi.org/10.1016/j.eswa.2012.05.023

Ossa Giraldo, W., Jaramillo Marin, V., y cols. (2021). *Machine learning para la estimación del riesgo de crédito en una cartera de consumo* (Tesis, para optar el Título de Magíster en Administración Financiera). Universidad EAFIT.

Patle, A., y Chouhan, D. S. (2013). Svm kernel functions for classification. , 1–9. doi: 10.1109/ICAdTE.2013.6524743

Qiu, Y., Zhou, J., Khandelwal, M., Yang, H., Yang, P., y Li, C. (2021). Performance evaluation of hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost models to predict blast-induced ground vibration. *Engineering with Computers*, 1–18. doi: https://doi.org/10.1007/s00366-021-01393-9

Quinlan, J. R. (1987). Simplifying decision trees. *International journal of man-machine studies*, *27*(3), 221–234. doi: https://doi.org/10.1016/S0020-7373(87)80053-6

Ramos Martinez, H. M. (2017). *Implementación de una herramienta de análisis de riesgo de crédito basado en el modelo de rating de crédito, algoritmos genéticos y clustering jerárquico aglomerativo* (Tesis, para optar el Título Profesional de Ingeniero de Sistemas).

Rayo Cantón, S., Lara Rubio, J., y Camino Blasco, D. (2010). Un modelo de credit scoring para instituciones de microfinanzas en el marco de Basilea II. *Journal of Economics, Finance and Administrative Science*, *15*(28), 89–124.

Refaeilzadeh, P., Tang, L., y Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, 532–538. doi: https://doi.org/10.1007/978-0-387-39940-9_565

Rimarachín, G. B. D., y Sánchez, Y. I. J. (2018). *Estrategias de riesgo crediticio para disminuir el Índice de morosidad de la cooperativa de ahorro y crédito Tumán* (Tesis, para optar el Título Profesional de licenciado en contabilidad). Facultad de Ciencias Empresariales escuela Académico Profesional de Contabilidad.

Rodriguez Molina, N. D. (2022). *Creación de una aplicación web para la gestión de conocimiento de procesos contables y soluciones dadas a sus incidentes* (B.S. thesis). Universidad Tecnológica ECOTEC.

Rodríguez, C. F. (2021). *Generación de árboles de decisión usando un algoritmo inspirado en la física* (Maestría en Computación Aplicada). Laboratorio Nacional de Informática Avanzada, A.C.

Rokach, L., y Maimon, O. (2008). *Data mining with decision trees: Theory and applications.* World Scientific Publishing Co. Pte. Ltd.

Rovirosa, J. E. C., Sosa, F. A. P., y Santana, M. A. E. (2015). La relación entre la inclusión financiera y el rezago social en México. *Cimexus*, *10*(1), 13–31.

Russell, S. J., y Norvig, P. (2004). *Inteligencia artificial: Un enfoque moderno* (2da Edición ed.). Pearson Educación.

Rutkowski, L., Jaworski, M., Pietruczuk, L., y Duda, P. (2014). The cart decision tree for mining data streams. *Information Sciences*, *266*, 1–15. doi: https://doi.org/10.1016/j.ins.2013.12.060

Saavedra García, M. L., y Saavedra García, M. J. (2010). Modelos para medir el riesgo de crédito de la banca. *Cuadernos de administración*, *23*(40), 295–319.

Selvik, J. T., y Abrahamsen, E. B. (2017). On the meaning of accuracy and precision in a risk analysis context. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, *231*(2), 91–100. doi: https://doi.org/10.1177/1748006X16686897

Shen, F., Zhao, X., Kou, G., y Alsaadi, F. E. (2020). A new deep learning ensemble credit risk evaluation model with and improved synthetic minority oversampling technique. *Applied Soft Computing Journal*.

Singh, I., Kumar, N., Srinivasa, K., Maini, S., Ahuja, U., y Jain, S. (2021). A multi-level classification and modified pso clustering based ensemble approach for credit scoring. *Applied Soft Computing*, *111*, 107687. doi: https://doi.org/10.1016/j.asoc.2021.107687

Solarte, J. C. M., y Cerezo, E. C. (2018). Modelos para otorgamiento y seguimiento en la gestión de riesgo de crédito. *Revista de Métodos Cuantitativos para la Economía y la Empresa*, *25*.

Sutton, R. S., y Barto, A. G. (1998). *Reinforcement learning: An introduction.* The MIT Press.

SÁNCHEZ-CERVANTES, M. G., FAJARDO-DELGADO, D., y OCHOA-ORNELAS, R. (2017). Estudio comparativo de árboles de decisión para la clasificación de densidad mamográfica. *nature*, *4*(10), 45–53.

Tangirala, S. (2020). Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612–619.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, *18*(4), 267–276. doi: https://doi.org/10.1007/BF02289263

Tibshirani, R. (1996). Regressión shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B(Methodological)*, *58*(11), 267-288. doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Valdés, D. I. (2005). *El buen uso del dinero.* Editorial Limusa.

Vargas Sánchez, A., y Mostajo Castelú, S. (2014). Medición del riesgo crediticio mediante la aplicación de métodos basados en calificaciones internas. *Investigación & Desarrollo*, *2*(14), 5–25.

Ville, B. D. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448-455. doi: https://doi.org/10.1002/wics.1278

Wang, H., y Hu, D. (2005). Comparison of svm and ls-svm for regression. , *1*, 279–283. doi: 10.1109/ICNNB.2005.1614615

Wang, K., Li, M., Cheng, J., Zhou, X., y Li, G. (2022). Research on personal credit risk evaluation based on XGBoost. *Procedia computer science*, *199*, 1128–1135. doi: https://doi.org/10.1016/j.procs.2022.01.143

Westreicher, G. (2018). *Entidades de crédito.* Descargado 04 de marzo 2022, de `https://economipedia.com/definiciones/`

Yan, G. (2023). Autoencoder based generator for credit information recovery of rural banks. *International Journal of Industrial Engineering: Theory, Applications and Practice*, *30*(2). doi: 10.23055/ijietap.2023.30.2.8697

Zhang, W., Yan, S., Li, J., Tian, X., y Yoshida, T. (2022). Credit risk prediction of SMEs in supply chain finance by fusing demographic and behavioral data. *Transportation Research Part E 158*, *158*, 102611. doi: https://doi.org/10.1016/j.tre.2022.102611

ZhangAnthony, W., y GohSmith, T. (2016). Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Ann. Statist.*(7), 45-52. doi: https://doi.org/10.1016/j.gsf.2014.10.003

Zhou, Y., Uddin, M. S., Habib, T., Chi, G., y Yuan, K. (2021). Feature selection in credit risk modeling: an international evidence. *Economic Research-Ekonomska Istraživanja*, *34*(1), 3064-3091. doi: https://doi.org/10.1080/1331677X.2020.1867213

Apéndice A

Conjuntos de datos

A.1. Descripción de conjunto de datos

La construcción y preprocesamiento apropiado del conjunto de datos sirve como base para el desarrollo de modelos de aprendizaje computacional más eficientes, además permite a otros investigadores replicar los resultados obtenidos.

La descripción del conjunto de datos creado para esta tesis busca facilitar a futuro la creación de modelos más robustos y confiables para la predicción del riesgo crediticio. Las **Tablas A.1 y A.2** contienen la descripción detallada de las variables que conforman el conjunto de datos EIZ.

Tabla A.1: Descripción de características del conjunto de datos de la institución financiera EIZ.

Característica	Descripción
Edad	Edad de la persona que realiza la solicitud de crédito.
Código postal	Código postal del domicilio del solicitante.
Tipo de vivienda	Tipo de vivienda en la que habita el solicitante (rentada, familiar, propia, compartida).
Dependientes	Número de dependientes económicos que tiene el solicitante.
Estado civil	El estado civil del solicitante (soltero, casado, divorciado, viudo, unión libre).
Género	Género del solicitante.
Clave actividad	Es la clave de la actividad productiva de la persona, está basada en el catálogo siti (proporcionado por la CNBV).
Nivel académico	Nivel académico reportado por el solicitante del crédito (sin estudios, primaria, secundaria, preparatoria, licenciatura, maestría, doctorado).
Teléfono	Indica si el solicitante tiene teléfono o no.
Ingreso	Ingreso económico que tiene el solicitante del crédito.
Egreso	Egreso económico que tiene el solicitante del crédito.
Tipo préstamo	Tipo de préstamo al que pertenece la solicitud.
Tasa normal	Tasa normal que tiene el préstamo.
Tasa moratoria	Tasa moratoria que tiene el préstamo en caso de ser incumplido.

Tabla A.2: Descripción de características del conjunto de datos de la institución financiera EIZ (cont).

Característica	Descripción
Monto	Monto económico solicitado.
Avales	Número de avales que deja el solicitante del crédito.
Créditos trabajados	Número de créditos que ha tenido el solicitante dentro de la institución financiera.
Bien	Tipo de garantía dejado por el solicitante.
Monto garantía	Valor monetario de la garantía dejada por el solicitante.
Finalidad	Indica la finalidad para la cual será utilizado el crédito.
Remesas	Indica si el solicitante recibe remesas del extranjero.
Plazo	Plazo por el cual será liberado el préstamo.
Target	Etiqueta de la solicitud, si la solicitud presenta riesgo de impago o no.

A.2. Categorización de variables

La correcta identificación de las variables categóricas en el proceso de entrenamiento de un modelo de aprendizaje computacional permite representar características con valores discretos que no siguen una relación numérica, requieren un tratamiento especial para evitar interpretaciones erróneas y sesgos en el modelo resultante. Una identificación precisa de estas variables permiten que el modelo capture adecuadamente las relaciones intrínsecas entre las categorías y sus efectos en la variable objetiva. Por otro lado, Una manipulación incorrecta de estas variables puede llevar a una disminución en la calidad del modelo y, como consecuencia, a conclusiones inexactas.

Debido al amplio rango de valores para las variables que se listan en las **Tablas A.4, A.5** y **A.6**, utilizando los criterios aplicados en L. Chen (2022) aplicados a problemas del ámbito crediticio, se sugiere agrupar las variables independientes a partir del cálculo de la WoE se realiza por medio de la ecuación 2.8. Para cada variable se hizo una búsqueda exhaustiva para determinar la mejor agrupación de las características, de tal forma que se buscó maximizar la capacidad de predicción (entre 0.3 a 0.5) según la **Tabla A.3**, esta tabla muestra los valores que maximizan la capacidad de predicción de las variables.

Tabla A.3: Reglas relacionadas con el Valor de la Información

Valor de la información	Predicción de la variable Bueno/Malo
Menos de 0.02	No es útil para la predicción.
0.02 a 0.1	Poder predictivo débil.
0.1 a 0.3	Poder predictivo medio.
0.3 a 0.5	Fuerte poder predictivo.
>0.5	Poder predictivo sospechoso.

Tabla A.4: Categorización de las variables. Fuente propia.

Variable	**Categoría**	**Valores**
Tasa normal	0	0 >& <= 18
	1	18 >& <= 20
	2	20 >& <= 24
	3	24 >& <= 42
	4	42 >& <= 48
	5	48 >& <= 72
Tasa moratoria	0	0 >& <= 18
	1	18 >& <= 20
	2	20 >& <= 24
	3	24 >& <= 42
	4	42 >& <= 48
	5	48 >& <= 72
Clave actividad	0	10007 >& <= 2071017
	1	2071017 >& <= 6131023
	2	6131023 >& <= 6999900
	3	6999900 >& <= 8719017
	4	8719017 >& <= 9506009
	5	9506009 >& <= 9999999
Créditos trabajados	0	0 >& <= 1
	1	1 >& <= 2
	2	10 >& <= 11
	3	11 >& <= 12
	4	12 >& <= 14
	5	14 >& <= 16
	6	16 >& <= 18
	7	18 >& <= 22
	8	2 >& <= 3
	9	22 >& <= 42
	10	3 >& <= 4
	11	4 >& <= 5
	12	5 >& <= 66
	13	6 >& <= 7
	14	7 >& <= 8
	15	8 >& <= 10
Edad	0	19 >& <= 31
	1	31 >& <= 36
	2	36 >& <= 40
	3	40 >& <= 45
	4	45 >& <= 49
	5	49 >& <= 53
	6	53 >& <= 57
	7	57 >& <= 62
	8	62 >& <= 67
	9	67 >& <= 92

Tabla A.5: Categorización de las variables (cont). Fuente propia.

Variable	**Categoría**	**Valores**
Monto garantía	0	0 >& <= 450
	1	1000 >& <= 1200
	2	1200 >& <= 1500
	3	1500 >& <= 1800
	4	15584 >& <= 800000
	5	1800 >& <= 2100
	6	2100 >& <= 2250
	7	2250 >& <= 3000
	8	3000 >& <= 3500
	9	3500 >& <= 5000
	10	450 >& <= 600
	11	5000 >& <= 7500
	12	600 >& <= 750
	13	750 >& <= 900
	14	7500 >& <= 15584
	15	900 >& <= 1000
Monto	0	10000 >& <= 10500
	1	10500 >& <= 15000
	2	15000 >& <= 20000
	3	20000 >& <= 25000
	4	25000 >& <= 30000
	5	30000 >& <= 50000
	6	3500 >& <= 5000
	7	499 >& <= 3500
	8	5000 >& <= 6000
	9	50000 >& <= 80000
	10	6000 >& <= 7000
	11	7000 >& <= 9000
	12	80000 >& <=1119682
	13	9000 >& <= 10000
Código postal	0	69999 >& <= 70110
	1	70110 >& <= 70140
	2	70140 >& <= 70330
	3	70330 >& <= 70341
	4	70341 >& <= 70705
	5	70705 >& <= 70725
	6	70725 >& <= 70730
	7	70730 >& <= 70735
	8	70735 >& <= 70746
	9	70746 >& <= 71720
Dependientes	0	0 >& <= 1
	1	1 >& <= 2
	2	2 >& <= 4
	3	3 >& <= 5
	5	4 >& <= 14

Tabla A.6: Categorización de las variables (cont). Fuente propia.

Variable	**Categoría**	**Valores**
Egreso	0	19 >& <= 31
	1	31 >& <= 36
	2	36 >& <= 40
	3	40 >& <= 45
	4	45 >& <= 49
	5	49 >& <= 53
	6	53 >& <= 57
	7	57 >& <= 62
	8	62 >& <= 67
	9	67 >& <= 92
Ingreso	0	0 >& <= 1800
	1	11693 >& <= 20510
	2	1800 >& <= 5692
	3	20510 >& <= 3206500
	4	5692 >& <= 6000
	5	6000 >& <= 7955
	6	7955 >& <= 11693
Plazo	0	0 >& <= 3
	1	10 >& <= 11
	2	3 >& <= 5
	3	5 >& <= 6
	4	6 >& <= 7
	5	7 >& <= 10