Sistema de Predicción de Riesgo Crediticio mediante el uso de Técnicas y Algoritmos de Minería de Datos Caso Estudio: Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE

Junia Cando, Mauricio Iván y Sampedro Giler, Francisco Gonzalo

Departamento de Ciencias de la Computación

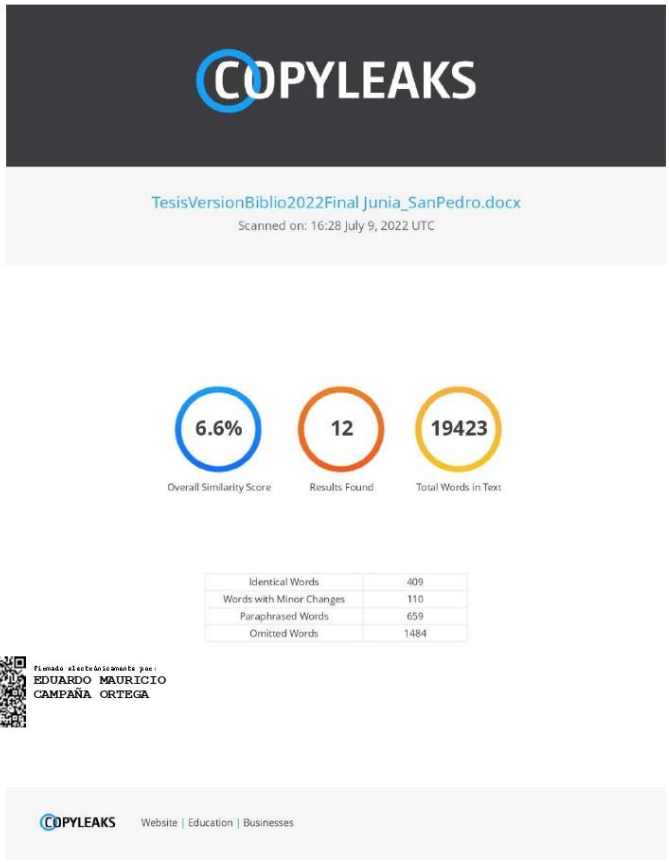
Carrera de Ingeniería de Sistemas e Informática

Trabajo de titulación, previo a la obtención del título de Ingeniero en Sistemas e Informática

MSc. Campaña Ortega, Eduardo Mauricio

03 de agosto de 2022

Análisis de similitud de contenidos.



.....

Campaña Ortega Eduardo Mauricio

C. C 1708856701

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Sistemas e Informática

Certificación

Certifico que el trabajo de titulación: "**Sistema de Predicción de Riesgo Crediticio mediante el uso de Técnicas y Algoritmos de Minería de Datos Caso Estudio: Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE**" fue realizado por los señores **Junia Cando Mauricio Iván** y **Sampedro Giler Francisco Gonzalo**; el mismo que cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, además fue revisado y analizado en su totalidad por la herramienta de prevención y/o verificación de similitud de contenidos; razón por la cual me permito acreditar y autorizar para que se lo sustente públicamente.

Sangolquí,03 de agosto 2022

Firma:

Firmado electrónicamente por:
EDUARDO MAURICIO
CAMPAÑA ORTEGA

.....

Campaña Ortega Eduardo Mauricio

C.C: 1708856701

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Sistemas e Informática

Responsabilidad de Autoría

Nosotros, **Junia Cando Mauricio Iván**, con cédula de ciudadanía n°1722469887 y **Sampedro Giler Francisco Gonzalo**, con cédula de ciudadanía n°2300504947, declaramos que el contenido, ideas y criterios del trabajo de titulación**: Sistema de Predicción de Riesgo Crediticio mediante el uso de Técnicas y Algoritmos de Minería de Datos Caso Estudio: Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE** es de nuestra autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos, y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 03 de agosto 2022

Firma

.....

Junia Cando Mauricio Iván

C.C.: 1722469887

.....

Sampedro Giler Francisco Gonzalo

C.C:2300504947

Departamento de Ciencias de la Computación

Carrera de Ingeniería de Sistemas e Informática

Autorización de Publicación

Nosotros, **Junia Cando Mauricio Iván**, con cédula de ciudadanía n°1722469887 y **Sampedro Giler Francisco Gonzalo**, con cédula de ciudadanía n°2300504947, autorizamos a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Sistema de Predicción de Riesgo Crediticio mediante el uso de Técnicas y Algoritmos de Minería de Datos Caso Estudio: Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de nuestra responsabilidad.

Sangolquí, 03 de agosto 2022

Firma

.....

Junia Cando Mauricio Iván

C.C.: 1722469887

.....

Sampedro Giler Francisco Gonzalo

C.C:2300504947

Dedicatoria

Dedico la tesis a mis padres, quienes siempre han estado ahí para apoyarme, me han dado todo para que pudiese salir adelante y terminar mi carrera, cuando me he caído han estado ahí para levantarme.

A mi hermana, mi mejor amiga, quien me ha sabido escuchar en momentos difíciles, siempre ha estado ahí para aconsejarme y no desviarme de la meta.

A mi compañero de tesis, con quien hemos pasado duros momentos pero que eso ha logrado que nos volvamos más unidos y siempre apoyarnos pase lo que pase.

A todos mis, amigos y personas que siempre han estado brindándome su apoyo para llevar a cabo este logro.

Junia Cando Mauricio Iván

A Dios, que me ha dado conocimiento y fuerza y ha guiado cada uno de mis pasos, y a la Virgen María, que siempre me ha acompañado y me ha intercedido.

Dedico mi tesis a mi madre, padre y a mi familia, que me han dado lo mejor de sí, incondicionalmente apoyándome, dándome excelentes enseñanzas de vida, son las que me motivan a crecer y me han ayudado a realizar mis objetivos a nivel personal y profesional, conquistando penurias juntas.

A mis amigos y a los que me han inspirado y apoyado para no renunciar y lograr mis objetivos.

Sampedro Giler Francisco Gonzalo

Agradecimiento

Agradezco a Dios que me ha bendecido, y me ha dado la fortaleza para salir adelante a pesar de las adversidades, a mis padres Sandra Cando y Jorge Junia a hermana Grace Junia y a Miguel Cedeño que sin el apoyo de ellos no hubiese sido posible este logro, a mi jefa María Augusta Escandón quien me ha apoyado y confiado en mi brindándome un trabajo estable y sobre todo un excelente ambiente laboral además de todas las facilidades para poder graduarme.

Por último, pero no menos importante agradezco al mgs. Ing. Mauricio Campaña y al personal que forma parte del Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, por confiar en nosotros y abrirnos las puertas para poder desarrollar la presente investigación.

Junia Cando Mauricio Iván

Quiero agradecer en primer lugar a Dios por permitirme tener salud y fuerzas para desarrollar el presente trabajo, por bendecir a mi familia y allegados, agradecer a mis padres Kleber Sampedro y Leida Giler, a mis hermanos Leonardo, Yaritza, Susana y Paola por ser un pilar y motivo de impulso de crecer día a día.

Expreso mi agradecimiento al mgs. Ing. Mauricio Campaña y al personal que forma parte del Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, por dar seguimiento y ser un apoyo en el desarrollo del presente trabajo.

Sampedro Giler Francisco Gonzalo

Índice de contenido

Índice de tablas

Índice de figuras

Resumen

Las inversiones y las líneas de crédito son fundamentales para el desarrollo de la economía y están alineados y son base para alcanzar el objetivo del Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, siendo una organización creada para crear ahorros entre sus participantes y brindarles beneficios y créditos.

Las expectativas y aspiraciones financieras de todos los participantes, brinda a los docentes, personal administrativo y de servicios, afiliados antiguos y nuevos las facilidades para iniciar y manejar sus ahorros, que se convertirán en un Fondo para un futuro mejor, Identificar la relación entre los créditos y los requisitos a cumplir por parte de los partícipes permitirá mediante la predicción de riesgo de crédito, mejorar los procesos de otorgamiento de créditos, asegurando que el riesgo crediticio sea lo menor posible.

Esta investigación de minería de datos se ha llevado a cabo para identificar las tendencias de bajo riesgo y alto riesgo. riesgo o PNL (préstamos con mora) a partir de los datos históricos y construir un modelo predictivo para ayudar a la gestión del Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, Para llevar a cabo el experimento se utiliza un modelo de Proceso de Descubrimiento de Conocimiento híbrido de seis pasos, aplicando la metodología SEMMA , se ha utilizado la herramienta Rapidminer, Los datos Los datos requeridos se recogieron de los repositorios de base de datos.

Se preprocesaron los datos para para la minería utilizando el software Rapidminer, utilizó tres algoritmos de minería de datos (Gradient Boosted Trees, Deep Learning y Naïve Bayes) para desarrollar el modelo predictivo. Los resultados indicaron que Naive Bayes es el mejor predictor con un 96,0167%.

Palabras claves: minería de datos, riesgo de crédito, predicción del riesgo de crédito, metodología de minería de datos, herramienta de minería de datos.

Abstract

Investments and credit lines are fundamental for the development of the economy and are aligned and are the basis for achieving the objective of the Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, being an organization created to create savings among its participants and provide them with benefits and credits.

The expectations and financial aspirations of all participants deposited in the Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, provide teachers, administrative and service personnel, old and new members with the facilities to start and manage their savings, which will become a Fund for a better future. Identifying the relationship between credits and the requirements to be met by the participants will allow through the prediction of credit risk, improve the processes of granting credits, ensuring that the credit risk is as low as possible.

This data mining research has been carried out to identify low risk and high-risk trends. To carry out the experiment, a six-step hybrid Knowledge Discovery Process model was used, applying the SEMMA methodology, the Rapidminer tool was used. The required data was collected from database repositories.

The data were preprocessed for mining using Rapidminer software, which used three data mining algorithms (Gradient Boosted Trees, Deep Learning y Naïve Bayes) to develop the predictive model. The results indicated that the Naïve Bayes is the best predictor with 96.0167%.

Key words: data mining, credit risk, prediction credit risk, methodology data mining, tool data mining.

Capítulo I

Introducción

Algunas de las áreas en las que se puede utilizar la minería de datos en la industria financiera son: segmentación y rentabilidad de clientes, solicitantes de préstamos que pueden conllevar riesgos, previsiones de morosidad, marketing, análisis crediticio, inversiones, transacciones fraudulentas, optimización de cartera de valores, gestión financiera y previsiones operativas, clientes rentables de tarjetas de crédito y venta cruzada, hay muchos tipos de créditos a considerar al solicitar créditos, y es importante conocer sus opciones (Hamid & Ahmed, 2016).

La clasificación de préstamos se refiere al proceso de calificar un préstamo y asignarlo a un grupo o grado en función del riesgo percibido y otras características relevantes del préstamo, el proceso continuo de revisión y clasificación de préstamos permite controlar la calidad de la cartera de préstamos y tomar medidas para contrarrestar la disminución de la calidad crediticia (Patel, Patil, Hembram, & Jaswal, 2020).

Los bancos necesitan utilizar un sistema de calificación interno más complejo que los más estandarizados. Un esquema de clasificación interna que es más complejo tiene como objetivo facilitar el seguimiento y la evaluación interbancaria (Patel, Patil, Hembram, & Jaswal, 2020). Los principales objetivos de la minería de datos son la predicción y la descripción; la predicción consiste en usar algunas variables en el conjunto de datos para predecir valores desconocidos de otras variables, y la descripción se enfoca en explicar los datos y encontrar patrones que los humanos puedan interpretar (Hamid & Ahmed, 2016).

La minería de datos es el proceso de extraer patrones ocultos de grandes cantidades de datos y se utiliza para tomar decisiones por escrito, el conocimiento derivado debe ser nuevo, no trivial,

relevante y aplicable al campo en el que se adquirió este conocimiento y también es el proceso de extraer información útil de datos sin procesar (Samsir, Suparno, & Giatman, 2020).

En el presente trabajo pretende encontrar un esquema de calificación que sea de utilidad para los agentes crediticios del “Fondo Complementario Previsional Cerrado de las Fuerzas Armadas ESPE”

Antecedentes

Muchos estudios han discutido temas relacionados dentro del marco de la minería de datos en el sector de la Banca y el Análisis de seguros. Por ejemplo, Jin et al. se utilizó un enfoque basado en datos y un enfoque de minería de datos para predecir el riesgo del préstamo y se compararon los modelos de minería de datos: árboles de decisión, máquina vectorial de soporte y redes neuronales, utiliza un enfoque de validación cruzada de 10 veces junto con el gran valor del porcentaje promedio de acierto para mostrar la mejor predicción. El análisis de la curva de elevación acumulativa se realiza para evaluar la calidad.

Sudhakar se centró en especificar la utilidad de las aplicaciones de minería de datos, estas aplicaciones utilizan varias técnicas de minería de datos, como árboles de decisión y redes neuronales de base radial. Este estudio incluyó la forma de implementar estas aplicaciones en un campo de evaluación del riesgo crediticio. McLeod presenta las propiedades de las redes neuronales y su idoneidad para el proceso de concesión de créditos (Jafar Hamid & Ahmed, 2016).

Según Chamatkar (Fosu et al., 2020), la minería de datos es la extracción de patrones y relaciones útiles de fuentes de datos, como bases de datos, textos y la web. Utiliza técnicas estadísticas y de coincidencia de patrones. La preocupación en la minería de datos son datos ruidosos, valores faltantes, datos estáticos, datos dispersos, datos dinámicos, relevancia, interés, heterogeneidad, eficiencia algorítmica y el tamaño y complejidad de los datos (Fosu et al., 2020).

Los datos que tenemos a menudo son vastos y ruidosos, lo que significa que son imprecisos y la estructura de datos es compleja. Aquí es donde una técnica puramente estadística no tendría éxito, debido a su inmensidad, por lo que la minería de datos es una solución. La minería de datos se ha convertido en una herramienta popular para analizar grandes conjuntos de datos. Los sistemas eficientes de gestión de bases de datos han sido activos muy importantes para la gestión de un gran corpus de datos, especialmente para la recuperación eficaz y eficiente de determinada información de una gran colección cuando ha sido necesario.

La proliferación de sistemas de gestión de bases de datos también ha contribuido a la reciente reunión masiva de todo tipo de información. La recuperación de información simplemente ya no es suficiente para la toma de decisiones porque la información debe extraerse y estar disponible (Weka 3 - Data Mining with Open Source Machine Learning Software in Java, s. f.).

La minería de datos se aplica en la industria bancaria y se acumulan enormes ventajas competitivas para aquellas industrias que la han implementado con éxito. Algunas de las áreas en las que la minería de datos se puede aplicar a la industria bancaria son en la identificación de factores de riesgo que predicen ganancias, reclamos y pérdidas, análisis de nivel de acreedores, análisis de marketing y ventas, desarrollo de nuevas líneas de productos, análisis financiero, estimación de provisión de reclamos pendientes y detección de fraude (Sadgali et al., 2019)].

La minería de datos ha destacado en los últimos años como una sofisticada metodología para buscar conocimiento que está escondido en las bases de datos de las organizaciones. El proceso de concesión de créditos es una de las funciones centrales de una entidad financiera; por lo tanto, el uso de instrumentos que de apoyo a ese proceso es deseable y puede convertirse en un factor clave en la gestión del crédito (Sousa, M. de M., & Figueiredo, R. S, 2014).

Planteamiento del problema

Actualmente en el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE cuenta con un sistema desarrollado en PowerBuilder que no ha sido actualizado desde el año 2013 y únicamente genera créditos para sus aportantes bajo las limitantes de sus servicios actuales mediante el otorgamiento de las prestaciones de cesantía establecidas en el Estatuto en forma complementaria e independiente a la prevista por el Seguro Social Obligatorio administrado por el Instituto Ecuatoriano de Seguridad Social; y, la inversión de los recursos del Fondo bajo criterios de seguridad, solvencia, diversificación del riesgo, rentabilidad y liquidez, en este sentido el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE requiere de un sistema inteligente que permita predecir el riesgo de acuerdo a las características de los solicitantes.

Los clientes solicitan créditos a diferentes bancos, algunos con trámites engañosos en diferentes momentos. Este entorno ha creado una importante competencia con otros bancos y los mantiene en apuros en la recaudación de efectivo y en la concesión de facilidades de crédito sin riesgo. La gestión y la medición del riesgo son el núcleo de todo servicio financiero. Hoy en día, el mayor reto en el mundo relacionado a los Servicios Financieros como los de la banca, seguros y Fondos de Cesantía es la implantación de sistemas de gestión de riesgos para identificar, medir y controlar la exposición del negocio.

En este sentido, el riesgo de crédito y de mercado presentan el reto central. Se puede observar un cambio importante en el ámbito de la forma de medirlos y tratarlos, a partir de la llegada de la tecnología avanzada de bases de datos y minería de datos. La evaluación del riesgo crediticio de los clientes implica elementos de decisión de gestión estructurados y no estructurados. Las decisiones estructuradas son aquellas donde se conocen de antemano los procesos necesarios para el

otorgamiento del préstamo y se dispone de varias herramientas computacionales para sustentar las decisiones.

Para decisiones no estructuradas, solo se utilizan la intuición y la experiencia de los gerentes. Los especialistas pueden apoyar a estos gerentes, pero las decisiones finales involucran una cantidad sustancial de elementos subjetivos. La minería de datos entra aquí para ayudar en la decisión no estructurada de la gerencia en la predicción y ejecución de los procedimientos de seguimiento necesarios.

Muchas de las organizaciones o incluso una sola persona, almacenan datos necesarios para devolver algún tipo de información de valor al propietario. Por lo tanto, los datos se almacenan de una manera que puede relacionarse con todos y cada uno de los demás datos y proporcionar algún hecho o información oculta necesaria para la evolución a más datos. Esta información, de vital importancia no es visible para el usuario, sino que está presente en la gran cantidad de datos obtenidos.

Cuanto mayor sea el tamaño de los datos, mayor será el tamaño de la información oculta y mayor será la posibilidad de derivar patrones y reglas a partir de ella. La minería de datos, como se ha mencionado en varios artículos, trabajos y libros, es la técnica para extraer este hecho o información del repositorio de datos almacenados. Este es el proceso para trabajar los datos utilizando algunos métodos de extracción.

Teniendo todo esto en cuenta se presentan las siguientes interrogantes:

Tabla 1

Preguntas de investigación

N.º	PREGUNTAS DE INVESTIGACIÓN
I	¿Cuáles son los indicadores adecuados para aplicar Minería de Datos para la evaluación del riesgo de crédito?
II	¿Qué metodología para el proceso de Minería de Datos es la más aplicable al perfil del proyecto?
III	¿Cuáles son los Indicadores y relaciones de interés en los contratos de créditos riesgosos y los contratos de créditos libre de riesgo para el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE?
IV	¿En qué medida el modelo permitirá la identificación del riesgo de crédito?

Nota: Preguntas que orientan a la investigación del presente trabajo.

Objetivos

General

Diseñar un modelo predictivo para la evaluación del riesgo en el otorgamiento de créditos en el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, utilizando técnicas y algoritmos de Minería de datos con soporte fuentes públicas de la Central de Crédito del Ecuador.

Específicos

- Analizar una pequeña porción de un conjunto de datos grande referente a los indicadores del riesgo de crédito en el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE.
- Explorar la información disponible mediante Herramientas de Visualización o Técnicas Estadísticas que permitan identificar las relaciones entre variables relacionadas al estudio de riesgo de crédito.
- Manipular los datos para crear, seleccionar y transformar las variables para el estudio de riesgo de crédito.
- Crear un Modelo válido utilizando las herramientas de software seleccionada.
- Diseñar una interfaz de usuario que permita evaluar la utilidad y confiabilidad de los resultados obtenidos mediante el modelo seleccionado.

Justificación

La concesión de créditos es una de las principales preocupaciones Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, ya que incluye los riesgos de impago. Según las directrices de la Superintendencia de Bancos y Seguros de Ecuador para los

fondos complementarios, las entidades financieras encargadas de brindar créditos, deben desarrollar sus propios sistemas de evaluación del riesgo crediticio.

Ciertas entidades cuentan con tales sistemas; sin embargo, han perdido una gran cantidad de dinero simplemente porque los modelos que utilizaron no pudieron predecir con precisión los incumplimientos de los clientes. Tradicionalmente, los bancos han utilizado modelos estáticos con factores demográficos para modelar patrones de riesgo crediticio.

Sin embargo, no se ha desarrollado un modelo dinámico que pueda adaptarse a factores político-económicos fluctuantes. En este artículo, proponemos un modelo que puede acomodar los factores asociados con las crisis político-económicas. El juicio humano prácticamente se elimina del proceso de evaluación del cliente. Usamos un sistema de inferencia difusa para crear una base de reglas usando un conjunto de predictores de incertidumbre.

Por lo tanto, presentamos un modelo que es más flexible a los factores político-económicos y puede producir resultados que son compatibles al máximo con situaciones de la vida real. La comparación entre la predicción realizada por el modelo propuesto y un préstamo en mora real indica poca diferencia entre ellos.

La principal innovación de esta investigación es producir una tabla de malos clientes mensualmente y crear un modelo dinámico basado en la tabla. Este modelo es un buen sustituto de los modelos estáticos actualmente en uso, ya que puede superar a los modelos tradicionales, especialmente frente a la crisis económica.

Alcance

De acuerdo con el objetivo y la finalidad fijados por el investigador, la cobertura de esta investigación se centraría en el Fondo Complementario Previsional Cerrado de Cesantía de la

Universidad de las Fuerzas Armadas ESPE, con una base de datos Sybase AnyWhere 8 subyacente durante los últimos 4 años. También ha estado utilizando un de contable Prometeo Debido a las limitaciones de recursos y tiempo para fusionar la información de los dos sistemas en un almacén de datos, esta investigación se llevará a cabo con los datos del Core Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE de los últimos 4 años.

En este estudio se aplican tareas de minería de datos tanto descriptivas como de modelización predictiva. En la modelización descriptiva, los grupos de acreedores se agrupan según los datos demográficos, el comportamiento crediticio, intereses expresados y otros factores descriptivos. Las estadísticas pueden identificar dónde los grupos de acreedores comparten similitudes y en qué se diferencian. Los acreedores más activos reciben una atención especial porque ofrecen el mayor (rendimiento del préstamo).

El objetivo de la modelización predictiva de la minería de datos es encontrar una descripción de cómo se comportará en el futuro determinados atributos de los datos. Por ejemplo, en las aplicaciones de riesgo crediticio, el análisis de los acreedores historial para predecir la probabilidad de la tasa de devolución del préstamo en un periodo de tiempo determinado.

Hipótesis

Un modelo predictivo permite identificar el riesgo crediticio en la generación de créditos en el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE.

Factibilidad del proyecto

Factibilidad tecnológica

Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE se encarga de afrontar todos los costos de sus grupos de investigación y de proveer todas

las herramientas que necesiten para continuar trabajando con sus investigaciones, en el caso del presente proyecto se utilizarán en su mayoría herramientas de código abierto y los equipos de infraestructura que se encuentran asignados al grupo de investigación, Sin embargo, se hará uso de los equipos personales de los estudiantes a cargo del proyecto para el desarrollo del modelo y la aplicación para validar el modelo.

Se requiere una computadora con las siguientes características:

Tabla 2

C*aracterísticas técnicas*

Cantidad de núcleos:	4
Cantidad de subprocesos:	8
Frecuencia turbo máxima:	Hz
Caché:	12 MB Intel® Smart Cache
Velocidad del bus:	4 GT/s
Frecuencia de incremento de TDP configurable:	2.90 GHz
Incremento de TDP configurable:	28 W
Frecuencia de descenso de TDP configurable:	1.30 GHz
Descenso de TDP configurable:	12 W
Memoria RAM:	16GB

Nota: La tabla describe los requisitos mínimos para los gestores de base de datos

Generador de consultas de predicción se proporciona en SQL Server Management Studio y SQL Server Data Tools

Factibilidad técnica

De acuerdo a las reuniones realizadas con el personal del Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, se identificó diferentes competencias que serán de beneficio para el desarrollo de este trabajo de titulación, más específicamente la empresa facilitará una descripción y el dominio del conocimiento respecto de la

consideración de un crédito y sus diferentes resoluciones , tanto si aprobado o rechazado, los requisitos para que un aportante en calidad de solicitante de crédito pueda realizar dicha solicitud.

Para el caso se ha identificado los profesionales que serán un apoyo vital para el desarrollo de este proyecto de tesis:

Tabla 3

Personal capacitado FCPCESPE

NOMBRE	CARGO	CONOCIMIENTOS
SR. GUILLERMO CHILUISA.	Oficial de Crédito	Productos financieros, préstamos y líneas de crédito.
ING. DIANA SÁNCHEZ.	Analista Técnica	Conocimiento en los requisitos de la empresa para aprobar o rechazar un crédito.

Nota: Tabla que describe al personal y cargo dentro del fondo complementario previsional cerrado de cesantía de la universidad de las fuerzas armadas ESPE.

Factibilidad operativa

Existe factibilidad para realizar la investigación porque se dispone del conocimiento suficiente en el campo de Gestión de Base de Datos, de los recursos económicos, bibliográficos y tecnológicos necesarios, así como el apoyo logístico y profesional de los especialistas. Fundamentalmente la facilidad

para acceder a la información. Los beneficiarios serán los funcionarios (Administradores) de las facultades y dependencias del fondo de Cesantías de la ESPE, conjuntamente con los funcionarios, para llevar un mejor control sobre los riesgos crediticios que se pueden o podrían presentar.

El proyecto cuenta con el apoyo del grupo de investigación conformado por las personas incluidas en este trabajo y con el personal designado en cada una de las áreas que conforman el Fondo de Cesantía de la ESPE, dicho grupo cuenta con la experticia que permita el correcto desarrollo del modelo de minería de datos.

A continuación, se detalla los miembros del fondo con sus respectivas funciones:

Tabla 4

Cargos y funciones FCPCESPE

NOMBRE	CARGO	FUNCIÓN
ING. BYRON BERMEO OLIVEROS.	Representante Legal	Representar legal, judicial y extrajudicial al Fondo. Revisar y aprobar los procesos administrativos internos, tales como: préstamos, cesantías y desafiliaciones.
SR. GUILLERMO CHILUISA.	Oficial de Crédito	Atención al cliente, información sobre cuenta individual para acceder a créditos quirografarios y prendarios.
ING. ADRIANA ORBEA, MSC.	Contadora	Revisar y registrar las operaciones diarias del Fondo,

		control financiero y emisión de Estados Financieros mensuales
ING. DIANA ROBALINO	Asistente Financiera y Administrativa	Atención al cliente, contacto a partícipes para acceder a la prestación de liquidación de cesantía o rendimientos, apoyo para procesos internos administrativos y financieros.
ING. DIANA SÁNCHEZ	Analista Técnica	Responsable de dar respuesta a las solicitudes o requerimientos del ente de control y administración del Fondo, encargada del desarrollo del Programa de Educación Financiera.

Nota: Tabla que describe las funciones de los encargados del fondo complementario previsional cerrado de cesantía de la universidad de las fuerzas armadas ESPE.

Factibilidad económica

Todo el proyecto será autofinanciado por los estudiantes y el grupo de investigación, De acuerdo a lo analizado, los computadores con los que contamos cumplen con los requisitos por lo tanto no se necesita realizar una inversión, además que se utilizará software libre.

Tabla 5

Costos del desarrollo

COMPONENTE	COSTO
Computadora 1	850$
Computadora 2	1100$
Internet	40$
Programador 1	500$
Programador 2	500$
Herramientas OpenSource: RStudio, Visual Studio y VisualCode	0,00
Total:	2990$

Nota: Tabla con la descripción de los costos de desarrollo

Tabla 6

Factibilidad mensual

	MAY	JUN	JUL	AGO	SEP
Computadora 1	850$	850$	850$	850$	850$
Computadora 2	1100$	1100$	1100$	1100$	1100$
Internet	40$	40$	40$	40$	40$
Programador 1	500$	500$	500$	500$	500$
Programador 2	500$	500$	500$	500$	500$
Herramientas OpenSource: RStudio, Visual Studio y VisualCode	0,00	0,00	0,00	0,00	0,00
	2990$	2990$	2990$	2990$	2990$

Nota: Tabla con la descripción de la factibilidad mensual.

Capítulo II

Marco Referencial

Estado del arte

Dada la hipótesis, se plantea un estudio de literatura, el cual permitirá identificar el estado actual sobre un problema. Para esto, se ha empleado varias fuentes digitales, entre las más importantes Google Académico, IEEExplore, en las cuáles se puede encontrar trabajos investigativos realizados a nivel mundial, ayudando a la población investigativa a tener un enfoque de la actualidad.

Definición del objetivo

El estado del arte, tiene una orientación sobre la hipótesis y el problema planteado, cuyo objetivo es identificar el nivel investigativo realizado hasta la actualidad.

Criterios de inclusión y exclusión

Debido a la gran cantidad de artículos relacionados que se obtiene de las diferentes fuentes de búsqueda, sería imposible lograr analizarlos todos, para reducir y acercarse al tema identificado, se debe aplicar reglas para minimizar y elegir los documentos de mayor relevancia que apoyarán en el desarrollo del objeto de estudio.

Criterios de inclusión

Para el criterio de inclusión de artículos de investigación seleccionados para realizar la búsqueda bibliográfica preliminar y que cumplieron con los criterios de búsquedas recopilados dentro del grupo de investigación, que recopilo los mejores artículos. Seleccionados por estar conformados siendo posible encontrar palabras claves utilizadas para formular la cadena de búsqueda.

Los artículos seleccionados cumplen con los siguientes criterios de inclusión:

- Artículos publicados a partir del 2016.
- Artículos que hablen sobre la minería de datos
- Artículos que hablen sobre riesgos crediticios.
- Artículos que hablen sobre modelos predictivos enfocados a finanzas y entidades prestamistas.
- Artículos que hablen sobre metodologías de minería de datos.
- Documentos, libros o revistas que tengan carácter de artículo científico.

Criterios de exclusión

Los otros artículos de investigación no fueron elegidos porque no cumplían los requisitos en términos de niveles de información y no proporcionaban palabras clave para la revisión preliminar de la literatura.

Se han utilizado los siguientes criterios de exclusión para rechazar los artículos:

- Artículos publicados a antes del 2016.
- Artículos con metodologías distintas a predicción en minería de datos.
- Artículos con temas de entornos financieros sin relación al riesgo de otorgar créditos.

Definición de la estrategia de búsqueda

La estrategia de búsqueda utilizada en esta revisión inicial de literatura comprende de lo siguientes etapas:

- Revisión inicial de literatura en las bases de datos académicas digitales que tienen documentos relacionados con las preguntas de investigación.

- Validación de estudios que permitan descartar todos los artículos que caen dentro de los criterios de exclusión y que permiten el desarrollo de las siguientes fases; se añaden estos artículos al listado de control.
- Integración del grupo de control de todos los artículos que han superado los criterios de inclusión y exclusión forman el grupo de control, de los cuales se realizó un estudio de ítems como son el título, introducción y conclusiones cuyo análisis está en relación con las preguntas de investigación.

Revisión de literatura

A continuación, se detallan todos los estudios que han cumplido con las características de la investigación, mediante Método general para Cochrane adaptado para la Ingeniería de Sistemas de Información (ISE) (Diaz et al., 2021).

Tabla 7

Documentos seleccionados

ID	Título	Palabras clave
D1	Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técnicas de minería de datos	**Minería de datos, KDD, árboles de decisión, reglas de decisión, ID3 y J48.**
D2	Predicting credit risk on the basis of financial and non-financial variables and data mining	**Artificial neural network (ANN), Decision trees, financial variables, credit risk assessment, non-financial variables, unbalanced data.**
D3	Exploration of credit risk of P2P platform based on data mining technology	**Credit risk assessment, Data mining, Peer-to-peer lending, Blockchain platform**

D4	Modelo de medición de riesgo crediticio en entidades financieras basado en minería de datos. Caso práctico: Cacpeco Ltda.	**Minería de datos crédito, riesgo financiero, instituciones financieras, cooperativas de ahorro y crédito estudios de casos**
D5	Evaluación del Riesgo Crediticio con Minería de Datos y Sistemas Expertos	**Árboles de Decisión, Minería de Datos, Sistemas Expertos.**

Nota: Tabla con los documentos seleccionados para realizar la investigación del presente trabajo.

Revisión bibliográfica: procesamiento

Para definir el proceso de búsqueda, se ha especificado la ventana temporal de enero / 2013 a Julio / 2022, el área de investigación se establece en tecnologías de la información, la subzona a ISE y la subzona de investigación a ingeniería de software. Por lo tanto, se estructuraron las siguientes cadenas de búsqueda.

Tabla 8

Cadenas de búsqueda

SC1: Prediction of Credit Risk Data Mining Methodology
SC2: Prediction of Credit Risk Data Mining Algorithm
SC3: Prediction of Credit Risk Data Mining Model
SC4: Prediction of Credit Risk Data Mining Tool
SC5: Prediction of Credit Risk Data Mining Methodology
SC6: Prediction of Credit Risk Data Mining Algorithm
SC7: Prediction of Credit Risk Data Mining Model
SC8: Prediction of Credit Risk Data Mining Tool

Nota: Tabla con las cadenas de búsquedas para selección de documentos.

A continuación, se describe una tabla con extracción del conocimiento, con los documentos que han sido útiles para el desarrollo del proyecto tesis

Tabla 9

Cantidad de documentos extraídos

Cadenas de búsqueda	ACM	IEEE XPLORER	SCOPUS	SPRINGER	WEB OF SCIENCE	Total
SC1	3	10	0	3	5	21
SC2	0	3	0	1	4	8
SC3	1	5	0	1	3	10
SC4	1	4	0	2	1	8
SC5	1	10	0	1	1	13
SC6	1	4	0	1	5	11
SC7	1	6	0	2	3	12
SC8	1	3	0	2	2	8
Total	9	45	0	13	24	91

Nota: Tabla comparativa para la selección de los repositorios de documentación de guía para la investigación.

Evaluación de la calidad de los documentos seleccionados

Se establece un sistema de documentos de rango para cada uno de los después de las preguntas de evaluación (EQ). Como resultado de esta clasificación, se obtienen un conjunto de documentos relevantes.

EQ1. ¿Aborda el documento los temas de investigación que son objetivos de esta revisión bibliográfica?

1 punto, aborda un tema

2 puntos, aborda dos temas

3 puntos, aborda tres o más temas

EQ2. ¿Se están aplicando los resultados esperados en estudios de casos o en soluciones desarrolladas?

1 punto, si se aplica un resultado esperado

2 puntos, si se aplican dos resultados esperados

3 puntos, si se aplican tres o más resultados esperados

EQ3. ¿El documento especifica obras relacionadas?

1 punto por cada trabajo relacionado, máximo 3 puntos.

EQ4. ¿Se utiliza su propia metodología en el desarrollo del trabajo?

1 punto, si la metodología es ajena

2 puntos, si la metodología es propia

EQ5. ¿Presenta el documento un análisis de la

resultados?

1 punto por cada resultado esperado que se analiza, máximo

tres puntos

EQ6. ¿El documento presenta una aplicación al caso

¿estudios?

1 punto por cada estudio de caso, máximo tres puntos.

EQ7. ¿Presenta el documento una evaluación del trabajo desarrollado?

1 punto, si se evalúa.

EQ8. ¿Sugiere el documento desarrollar trabajos futuros?

1 punto por cada trabajo futuro sugerido, máximo tres puntos

En la siguiente tabla se describen la calidad de los documentos según las preguntas y calificaciones planteadas anteriormente.

Tabla 10

Comparación ponderaciones de clasificación

ID	EQ1	EQ2	EQ3	EQ4	EQ5	EQ6	EQ7	EQ8	TOTAL
D1	2	1	1	1	3	2	1	1	12
D2	2	1	2	2	3	2	2	1	15
D3	1	2	1	1	1	2	2	1	11
D4	2	1	1	2	1	2	1	1	11
D5	2	1	2	2	1	1	2	1	12
TOTAL	9	6	7	8	9	9	8	5	61

Nota: Descripción de las ponderaciones según las preguntas de control.

Grupo de control

(Tello, Eslava Blanco, Tobias, 2013) Análisis y evaluación del nivel de riesgo en el otorgamiento de créditos financieros utilizando técnicas de minería de datos.

En este artículo se presenta la aplicación de la minería de datos en el sector financiero, para evaluar el nivel de riesgo en el otorgamiento de créditos. Se tomó una muestra de datos de 1000 registros, correspondientes a una cartera comercial de una entidad bancaria. Se utilizó la metodología Knowledge Discovery in Databases (KDD) y se desarrolló un software que permitió discretizar los datos, para poder utilizarlos como entradas en la herramienta de minería de datos WEKA.

Se comparan los resultados obtenidos al aplicar las técnicas de minería de datos, árboles de clasificación. Finalmente se obtiene como resultado las características que deben tener los clientes para recibir un crédito bancario.

(Sihem Khemakhem, 2018) Predicting credit risk on the basis of financial and non-financial variables and data mining

El documento explica que la minería de datos para predecir el riesgo crediticio es una herramienta beneficiosa para que las instituciones evalúen la salud financiera de las empresas. Este estudio tiene como objetivo proporcionar un nuevo método para evaluar el riesgo de crédito, teniendo en cuenta no solo las variables financieras.

Se determinaron las variables financieras más significativas para construir un modelo de calificación crediticia e identificar la calidad crediticia de las empresas. Además, se utilizó la técnica de sobre muestreo de minorías sintéticas para resolver el problema del desequilibrio de clases y mejorar el rendimiento del clasificador.

Los resultados mostraron que los índices de rentabilidad, la capacidad de reembolso, la solvencia, la duración de un informe crediticio, las garantías, el tamaño de la empresa, el número de préstamos, la estructura de propiedad y la duración de la relación bancaria corporativa resultaron ser los factores clave para predecir el incumplimiento.

(Cai, Zhang. 2020) Exploration of credit risk of P2P platform based on data mining technology

El propósito de esta investigación es ayudar a los bancos de inversión a seleccionar clientes con buen crédito, excluir a los clientes con mayor riesgo, minimizar el riesgo de los inversores y mantener los intereses máximos de los inversores. La tecnología de minería de datos se utilizó para establecer un modelo de evaluación del riesgo crediticio de préstamos de red de persona a persona de igual a igual (P2P), y luego se evaluó con precisión la situación crediticia del prestamista para reducir el riesgo de la

plataforma. En primer lugar, se recopilaron y clasificaron los datos de préstamos de la plataforma LendingClub (LC) en 2018, y luego se obtuvo el conjunto de datos. En segundo lugar, el conjunto de datos se muestreó por capas y se obtuvieron 10 conjuntos de datos equilibrados y los cuatro índices de evaluación se obtuvieron a través de la clasificación de datos, es decir, el índice de evaluación de la calificación crediticia de la plataforma P2P. Finalmente, los datos reales de la plataforma LC fueron evaluados por árbol de decisión y algoritmo de regresión lógica binomial.

Los resultados de la investigación mostraron que el algoritmo del árbol de decisiones podría mejorar la precisión de la selección preliminar y predecir la probabilidad de incumplimiento de los prestatarios con mayor precisión, a fin de filtrar a los prestatarios con una tasa de incumplimiento más alta y reducir el riesgo de préstamo de la plataforma.

La combinación de los dos algoritmos puede realmente estimar el estado crediticio de los prestamistas y mejorar la eficiencia de la transacción. Por lo tanto, la investigación sobre el riesgo crediticio de la plataforma blockchain basada en tecnología de minería de datos es de gran importancia para mejorar el nivel crediticio de los inversores, mejorar sus ingresos por inversiones, ahorrar costos de transacción, optimizar la asignación de recursos crediticios y lograr una supervisión eficaz por crédito.

(Salinas, 2020) Modelo de medición de riesgo crediticio en entidades financieras basado en minería de datos. Caso práctico: Cacpeco Ltda.

En este articulo nos detalla la investigación realizada en La Cooperativa de Ahorro y Crédito CACPECO LTDA. La institución posee deficiencias en cuanto a su proceso de selección de potenciales sujetos de crédito, ya que la misma es realizado a través de análisis manuales en herramientas ofimáticas convencionales, sin embargo, no se logran analizar todos los factores deseados que permitan determinar con exactitud cuando un socio es un posible candidato; generando como resultado una deficiente promoción de productos crediticios y, por ende, un bajo porcentaje de colocación de crédito.

El objetivo del artículo es proveer al área de crédito información estructurada, fácil de comprender y un método útil para el análisis de riesgo del perfil del socio. Con el desarrollo de este trabajo se pretende obtener por parte de los autores un modelo de medición de riesgo crediticio (predictivo) utilizando técnicas de minería de datos que permita promocionar diferentes productos de una cartera en forma más eficientemente a los clientes e incrementar el número de créditos concedidos.

(Soto, 2014) Evaluación del Riesgo Crediticio con Minería de Datos y Sistemas Expertos

El presente documento evaluó el uso de algoritmos de Arboles de Decisión sobre una minería de Datos para la creación de un Sistema Experto que permitiera evaluar el riesgo crediticio en base la información de la minería.

Se empleó una data de 23 mil créditos evaluados. Por las reglas encontradas el algoritmo ha utilizado 4 atributos y no se ha tomado en cuenta los atributos con cantidad. La aplicación diseñada, y que valida si el crédito se aprueba o no, es muy sencilla. Luego se debe extender a una que permita generar un histórico de las evaluaciones de crédito y que permita emitir reportes. El sistema deberá en un futuro cercano permitir usar la minería u almacenar en la misma.

Metodología

Esta investigación está diseñada para aplicar la tecnología de minería de datos para la evaluación del riesgo crediticio en el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas ESPE, La metodología es el proceso utilizado para recopilar información y datos con el fin de tomar decisiones empresariales.

Diseño de la investigación

Esta investigación se ajusta a la investigación experimental. Estudio que se ciñe estrictamente a un diseño de investigación científica (Hassan et al., 2018). Incluye una hipótesis, una variable que puede

ser manipulada por el investigador, y variables que pueden medirse, calcularse y compararse. Y lo que es más importante, la investigación experimental se realiza en un entorno controlado.

El investigador recoge datos y resultados que apoyan o rechazan la hipótesis. Este método de investigación se denomina prueba de hipótesis o método de investigación deductivo (Hassan et al., 2018).

Para llevar a cabo el experimento en esta investigación, el investigador ha seguido el proceso de minería de datos híbrida Modelos. Según Swiniarski y Kurgan (Pazmiño-Maji et al., s. f.), el modelo híbrido mejora el proceso de descubrimiento del conocimiento mediante la combinación de los modelos académicos e industriales en los proyectos de minería de datos. El desarrollo del modelo híbrido se adoptó a partir del modelo de proceso SEMMA, ya que se puede utilizar para la investigación académica.

Por lo tanto, estos modelos están orientados a la investigación, que presentan el paso de minería de datos en la modelización. Los seis pasos de los modelos híbridos permiten una serie de mecanismos de retroalimentación. Además, el conocimiento descubierto en el último paso para un dominio específico puede aplicarse en otros dominios

Herramientas

SQL Anywhere

SQL Anywhere es una base de datos relacional fabricada por SAP. Proporciona múltiples funciones avanzadas como: referencias integradas, procedimientos almacenados, tablas de proxy, bloqueo de nivel de línea, alta disponibilidad, gestión de usuarios y eventos del sistema, modo en memoria, paralelismo entre consultas, duplicación de bases de datos, integración con directorios LDAP,

cifrado sólido, soporte de múltiples interfaces, etc. (Get SQL Anywhere - Manage, Synchronize and Exchange Data, s. f.).

Weka 3

Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Contiene herramientas para la preparación de datos, clasificación, regresión, agrupación, minería de reglas de asociación y visualización.

Weka es un software de código abierto publicado bajo la Licencia Pública General de GNU (Weka 3 - Data Mining with Open Source Machine Learning Software in Java, s. f.).

Rapidminer

RapidMiner está disponible como aplicación independiente para el análisis de datos y también puede integrarse como motor de DM en otras aplicaciones, es una plataforma gratuita, flexible y de código abierto implementada en Java, por lo que se ejecuta en todas las plataformas y sistemas operativos importantes, y es muy fácil utilizar su código en ella (Graczyk, Lasota, & Trawiński, 2009). Representa un enfoque completamente nuevo para el diseño de una aplicación en la que incluso los problemas más complicados y complejos se convierten en algo sencillo (Burger, Karasek, Smékal, Uher, & Dostal, 2010).

Es una de las soluciones de análisis predictivo y minería de datos más utilizadas en el mundo actual, tiene muchas ventajas. Algunos de ellos son: Se proporcionará de forma gratuita como Software de código abierto, pero con licencias comerciales adecuadas para el desarrollo de aplicaciones comerciales de código cerrado, una comunidad gratuita experimentada con soporte patentado avanzado y una comunidad de usuarios y desarrolladores madura y en crecimiento relativamente económica de desarrollar (Burger, Karasek, Smékal, Uher, & Dostal, 2010). Otras soluciones de minería

de datos que también se ofrecen, lo que lo hace particularmente adecuado para fines de investigación. Y una interfaz gráfica de usuario intuitiva, clara y fácil de usar (Windarto & Wanto, 2018).

DBeaver

DBeaver es una herramienta de base de datos universal, gratuita y de código abierto para desarrolladores y administradores de bases de datos.

- La usabilidad es el objetivo principal de este proyecto, la interfaz de usuario del programa está cuidadosamente diseñada e implementada.
- Es gratuito y de código abierto (ASL) y multiplataforma
- Se basa en un marco de trabajo de código abierto y permite escribir varias extensiones (plugins).
- Soporta cualquier base de datos que tenga un driver JDBC.
- Puede manejar cualquier fuente de datos externa que puede o no tener un controlador JDBC.
- Existe un conjunto de plugins para diferentes bases de datos y diferentes utilidades de gestión de bases de datos (por ejemplo, ERD, transferencia de datos, comparación, exportación/importación de datos, generación de datos simulados, etc.).
- Tiene un gran número de características (About | DBeaver Community, s. f.).

Capítulo III

Marco teórico

Se llevó a cabo una revisión de la literatura conceptual y empírica con el fin de tener una comprensión conceptual sobre la minería de datos y de su aplicación en el área Financiera, también se revisaron trabajos de investigación previos sobre el área problemática para tener una idea de la posible aplicabilidad de la extracción de datos en la evaluación del riesgo crediticio. Se revisaron diversos libros, revistas, artículos, artículos Conceptuales y páginas web relacionados con el tema de la minería de datos y el descubrimiento de conocimiento en la base de datos.

¿Qué es Minería de Datos?

Los datos no se limitan a tuplas representadas solo por números o caracteres. Los avances en la tecnología de administración de bases de datos permiten la integración de diferentes tipos de datos, como imágenes, videos, texto y otros datos digitales en una sola base de datos, facilitan el procesamiento multimedia (Riquelme, Ruiz y Gilbert, 2006).

Por lo tanto, la combinación especial de técnicas estadísticas y herramientas tradicionales de gestión de datos ya no es suficiente para analizar esta gran colección desigual de datos.

La tecnología de Internet actual y sus crecientes necesidades requieren el desarrollo de tecnologías de minería de datos más avanzadas para interpretar la información y el conocimiento de los datos distribuidos en todo el mundo. En este siglo, la demanda seguirá creciendo y el acceso a grandes volúmenes de datos multimedia traerá la mayor transformación de la sociedad global. Por lo tanto, el desarrollo de tecnologías avanzadas de minería de datos seguirá siendo un área importante de investigación y, por lo tanto, se deben dedicar más recursos a este campo en evolución en los próximos años.

Existen diversos campos en los que se almacenan grandes volúmenes de información en bases de datos centralizadas y distribuidas, como bibliotecas digitales, archivos de imágenes, bioinformática, salud, finanzas e inversión, producción y producción, ventas y marketing, redes de telecomunicaciones, etc.

No es de extrañar que la minería de datos, como un tema verdaderamente interdisciplinario, pueda definirse de muchas formas diferentes. Incluso el término minería de datos no presenta realmente todos los componentes principales en la imagen. Para referirnos a la extracción de oro de rocas o arena, decimos extracción de oro en lugar de extracción de roca o arena.

Figura 1

Minería de datos

Nota: Minería de datos-búsqueda de conocimiento (patrones interesantes) en datos (Zaki y Wong, 2001).

De manera análoga, la minería de datos debería haberse denominado más apropiadamente "minería de conocimiento a partir de datos", que lamentablemente es algo largo. Sin embargo, a corto plazo, la minería del conocimiento puede no reflejar el énfasis en la minería a partir de grandes

cantidades de datos. Sin embargo, la minería es un término vívido que caracteriza el proceso que encuentra un pequeño conjunto de pepitas preciosas a partir de una gran cantidad de materia prima.

Por lo tanto, un nombre tan inapropiado que incluye tanto "datos" como "minería" se convirtió en una opción popular. Además, muchos otros términos tienen un significado similar a la minería de datos, por ejemplo, minería de conocimientos a partir de datos, extracción de conocimientos, análisis de datos / patrones, arqueología de datos y dragado de datos (Zaki y Wong, 2001).

Aplicaciones de la Minería de datos

Algunas de las tareas importantes de la minería de datos incluyen la identificación de aplicaciones para las técnicas existentes, y desarrollar nuevas técnicas para dominios tradicionales o de nueva aplicación, como el comercio electrónico y la bioinformática. Existen numerosas áreas donde la minería de datos se puede aplicar, prácticamente en todas las actividades humanas que generen datos:

- Comercio y banca: segmentación de clientes, previsión de ventas, análisis de riesgo.
- Medicina y Farmacia: diagnóstico de enfermedades y la efectividad de los tratamientos.
- Seguridad y detección de fraude: reconocimiento facial, identificaciones biométricas, accesos a redes no permitidos, etc.
- Recuperación de información no numérica: minería de texto, minería web, búsqueda e identificación de imagen, video, voz y texto de bases de datos multimedia. • Astronomía: identificación de nuevas estrellas y galaxias.
- Geología, minería, agricultura y pesca: identificación de áreas de uso para distintos cultivos o de pesca o de explotación minera en bases de datos de imágenes de satélites
- Ciencias Ambientales: identificación de modelos de funcionamiento de ecosistemas naturales y/o artificiales (p.e. plantas depuradoras de aguas residuales) para mejorar su observación, gestión y/o control.

- Ciencias Sociales: Estudio de los flujos de la opinión pública. Planificación de ciudades: identificar barrios con conflicto en función de valores sociodemográficos.

Descripción general de la Minería de datos

La minería de datos podría definirse como el proceso para derivar conocimiento y patrones interesantes de una gran colección de datos, La capacidad de extraer conocimiento útil de enormes conjuntos de datos y utilizar este conocimiento se está volviendo importante en el mundo competitivo de hoy. Por ejemplo, en la comunidad empresarial, la minería de datos se puede utilizar para descubrir nuevas tendencias de compra, planificar estrategias de inversión y detectar gastos no autorizados en los sistemas de contabilidad (DivaPortal,2021).

Metodología SEMMA

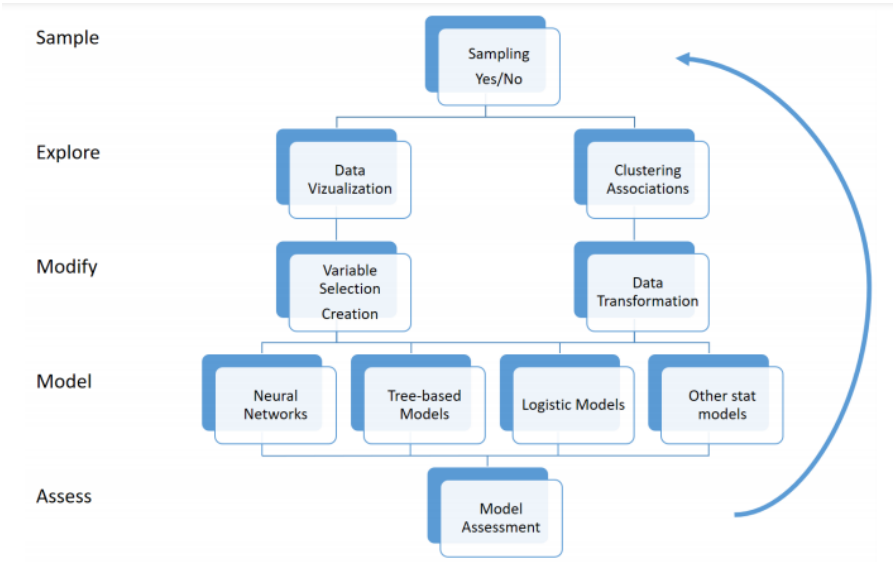
SEMMA es un acrónimo de Sample, Explore, Modify, Assess (Enterprise Miner). SAS Institute, que desarrolló el modelo, lo describe no como un método de minería de datos, sino como un conjunto de herramientas para llevar a cabo las tareas centrales de la minería de datos. SEMMA se centra principalmente en los aspectos de desarrollo de modelos de la minería de datos y se utiliza en el software SAS Enterprise Mine.

El movimiento entre los diferentes pasos no es estricto, durante el proyecto puede moverse hacia adelante y hacia atrás y repetir los pasos. Según SAS, más que una simple metodología para la minería de datos, SEMMA es un conjunto de herramientas funcionales que se enfocan en los aspectos de auto evolución de un modelo de minería de datos (Enterprise Miner, 2020).

SEMMA consta de cinco pasos diferentes, que se describen en general en las secciones a continuación y en la Figura 2.2, pero no es obligatorio incluir todos los pasos en el proyecto.

Figura 2

Etapas SEMMA



Nota: Figura que describe los pasos a seguir según la metodología SEMMA (Bulkley y Gayle, 2018).

VISTA DE LOS PASOS EN LAS DIFERENTES FASES DE SEMMA (BULKLEY Y GAYLE)

Sample (Muestra)

El primer paso se llama Muestra. Aquí comenzará el muestreo de los datos que luego se utilizará para modelar. Los datos recogidos deben ser grandes y suficientes para contener la información necesaria, pero lo suficientemente pequeño para ser fácil de procesar (P.C et al).

Esta fase también incluye la partición de los datos a crear formación, validación y muestras de ensayo, es posible facilitar los procesos de minería de datos, reduciendo costos y tiempo para la organización (P.C et al).

Explorer (Explorar)

En este paso, los datos serán explorados y buscados para cualquier tipo de patrones y relaciones interesantes. Esto se hace para obtener una comprensión de los datos y de eso sacar conclusiones y

obtener ideas. Esto se puede hacer con el uso de la visualización, pero si las visualizaciones no muestran ninguna tendencia clara, se puede utilizar un análisis estadístico en su lugar (P.C et al).

La minería de datos estadísticos permite que se rastree, detecte e identifique con éxito, y posteriormente descartar los datos que representan anomalías u omisiones en las etapas posteriores hacia el descubrimiento de información (P.C et al).

Modify (Modificar)

Este paso se basa en el paso anterior Explorar. En este paso, los datos comienzan a ser modificados y preparados para ser utilizados en un modelo específico, la selección y transformación de los datos se realiza de acuerdo con las variables retenidas para el proceso de minería, lo que permitirá, en función de estas variables, adecuar el enfoque de diseño y la selección del modelo (P.C et al).

Puede incluir segmentación adicional de la muestra y la creación de nuevas variables.

Modelo (Model)

En el cuarto paso, el modelo está empezando a ser creado. Aquí se aplicarán diferentes técnicas de modelado a los datos y variables ahora modificados y bien seleccionados (P.C et al). Esto se esforzará por lograr el objetivo de obtener un modelo confiable, que luego se puede utilizar para predecir un resultado o clasificar datos desconocidos.

Las herramientas de software existentes permiten el uso de técnicas y métodos de minería de datos, que tienden a descubrir asociaciones o combinaciones entre los datos, realizando predicciones de resultados con un alto grado de precisión (P.C et al).

Entre las técnicas más utilizadas para el modelado de datos se encuentran: métodos estadísticos, agrupamiento, redes neuronales, árboles de decisión, lógica difusa, reglas de asociación, etc. (P.C et al).

Evaluar (Assess)

En el paso final de SEMMA, una evaluación de los resultados de los modelos y el rendimiento se lleva a cabo contra las muestras que se utilizan para la validación y las pruebas. Con esta evaluación, se toma una decisión si el modelo es útil y confiable (P.C et al).

Sobre la base del modelo obtenido en la fase anterior, se realiza una evaluación de los resultados para verificar el éxito del proyecto. Una buena práctica para la validación del modelo es elegir una muestra de datos diferente y aplicarla para verificar los resultados. Si esto es óptimo, proceda al proceso de producción, de lo contrario se desarrollará otro modelo.

Préstamos bancarios

Son promesas contractuales de prestar a un solicitante específico en condiciones predeterminadas una determinada cantidad monetaria, esto se utiliza ampliamente en la economía, a medida que ha crecido el uso de los contratos de préstamo también ha crecido la información sobre ellos (Ergungor, 2001).

Dos características de los contratos de préstamo, son, las comisiones que deben pagarse a lo largo de la vida del contrato, y la cláusula de cambio material adverso, resultan ser particularmente importantes en los modelos.

La estructura de honorarios puede incluir un honorario de compromiso, que es una cuota inicial que se paga cuando se firma el contrato, mientras que la estructura de comisiones puede incluir una comisión de contrato, que es una comisión inicial que se paga cuando se adquiere el contrato, una comisión anual, que se paga sobre la cantidad prestada, y una de uso, que se cobra sobre el crédito disponible no utilizado, un contrato de préstamo rara vez incluye los tres tipos de comisiones juntas (Ergungor, 2001).

Riesgo crediticio

El riesgo de crédito o riesgo de impago implica la incapacidad o la falta de voluntad de un cliente o una contraparte para cumplir los compromisos en relación con las operaciones de préstamo, comercio, cobertura, liquidación y otras operaciones financieras, el riesgo de crédito se compone generalmente del riesgo de transacción o de impago y del riesgo de cartera (Bartosova, 2008).

El riesgo de cartera comprende a su vez el riesgo intrínseco y el de concentración, el riesgo de crédito de la cartera de un banco depende de factores externos e internos. Los factores externos son el estado de la economía, las grandes oscilaciones de los precios de las materias primas/acciones, los tipos de cambio y los tipos de interés, las restricciones comerciales, las sanciones económicas, las políticas gubernamentales, etc. Los factores internos son las deficiencias en las políticas/administración de préstamos, la ausencia de límites prudenciales de concentración de créditos, la definición inadecuada de los límites de préstamo para los agentes de crédito/comités de crédito, las deficiencias en la evaluación de la posición financiera de los prestatarios, la excesiva dependencia de las garantías y la inadecuada fijación de los precios del riesgo, la ausencia de un mecanismo de revisión de los préstamos y de vigilancia posterior a la sanción, etc. (Bartosova, 2008).

Modelo predictivo

Un modelo de predicción trata de estratificar al prestamista según su probabilidad de tener un determinado resultado. El modelo permite entonces identificar a los solicitantes de créditos que tienen una mayor probabilidad de un evento (Goodman, 2008).

La variable de resultado del modelo de predicción puede ser cualquier cosa, podemos distinguir las variables de resultado con variables continuas o variables categóricas. Las variables continuas se describen mediante valores numéricos y se utilizan modelos de regresión para predecirlas, por ejemplo, la regresión lineal. Las variables categóricas están restringidas a un número limitado de clases o

categorías y para su predicción se utilizan modelos de clasificación. Si el resultado tiene dos categorías, se denomina clasificación binaria y las técnicas típicas son árboles de decisión y regresión logística (Banerjee, 2009).

Tareas de Minería de Datos

Las tareas de minería de datos se pueden clasificar generalmente en dos tipos, en función de lo que intenta lograr una tarea específica. Estas dos categorías son tareas descriptivas y tareas predictivas. Las tareas descriptivas de minería de datos caracterizan las propiedades generales de los datos, mientras que las tareas predictivas de minería de datos realizan inferencia sobre el conjunto de datos disponible para predecir cómo se comportará un nuevo conjunto de datos.

La minería de datos se puede utilizar para realizar diferentes tareas. Pero, la tarea de la minería de datos depende del uso del resultado de la minería de datos que significa para qué propósito se utilizará el resultado (Gupta & Chandra, 2020)

Las tareas se clasifican de la siguiente manera:

- **Análisis exploratorio de datos:** es simplemente explorar los datos sin tener una idea clara de lo que estamos buscando. Estas técnicas son interactivas y visuales.
- **Modelado descriptivo:** Describe todos los datos, incluye modelos para la distribución de probabilidad general de los datos, la partición del espacio p-dimensional en grupos y modelos que describen las relaciones entre las variables. En resumen, se aplica para describir los datos existentes.
- **Modelado predictivo:** Para predecir el futuro teniendo o basándose en los datos o comportamientos existentes. El modelo permite predecir el valor de una variable a partir de los valores conocidos de otras variables

- **Descubrir patrones y Reglas:** Se ocupa de la detección de patrones; el objetivo es detectar comportamientos fraudulentos detectando regiones del espacio que definen los diferentes tipos de transacciones donde los datos apuntan significativamente diferentes del resto. A diferencia del modelado descriptivo, el descubrimiento de patrones se ocupa de generar patrones y reglas ocultas importantes
- **Recuperación por contenido:** Es encontrar un patrón similar al patrón de interés en el conjunto de datos. Esta tarea se usa más comúnmente para conjuntos de datos de texto e imágenes

JAVA

El lenguaje de programación Java ha sido muy bien recibido por la comunidad mundial de desarrolladores de software y proveedores de contenidos de Internet. Los usuarios de Internet y de la World Wide Web se benefician del acceso a aplicaciones seguras e independientes de la plataforma, que pueden provenir de cualquier lugar de Internet los desarrolladores de software que crean aplicaciones en el lenguaje de programación Java se benefician al desarrollar el código una sola vez, sin necesidad tener que adaptar sus aplicaciones a todas las plataformas de software y hardware (Gosling & Holmes,2005).

Para muchos, el lenguaje se conoció primero como herramienta para crear applets para la World Wide Web. Un applet es una miniaplicación que se ejecuta dentro de una página web. Un applet puede realizar tareas e interactuar con los usuarios en sus páginas del navegador sin utilizar recursos del servidor web después de ser descargado

Algunos applets pueden, por supuesto, hablar con el servidor para hacer su trabajo, pero eso es cosa suya, el lenguaje de programación Java es realmente valioso para entornos de red distribuidos como la Web. Sin embargo, va mucho más allá de este dominio para proporcionar un potente lenguaje de programación de propósito general adecuado para construir una variedad de aplicaciones que, o bien

no dependen de las características de la red, o las quieren por razones diferentes. La capacidad de ejecutar código descargado en hosts remotos de forma segura es un requisito crítico requisito crítico para muchas organizaciones (Gosling & Holmes,2005).

El lenguaje de programación Java está diseñado para lograr la máxima portabilidad con el menor número posible de dependencias de implementación como sea posible (Gosling & Holmes,2005). Un int, por ejemplo, es un entero con signo de 32 bits en todas las implementaciones, independientemente de la arquitectura de la CPU en la que se ejecute el programa. Definir todo lo posible sobre el lenguaje y su entorno de ejecución permite a los usuarios ejecutar el código compilado en cualquier lugar y compartir el código con cualquiera que tenga un entorno de ejecución Java.

ANGULAR

AngularJS fue uno de los primeros frameworks para el desarrollo de SPAs. Fue capaz de desbancar a jQuery al ofrecer a los desarrolladores características como la vinculación de datos bidireccional y la posibilidad de organizar módulos para importar scripts externos. Una de sus principales ventajas sobre la mayoría de los competidores es su carácter accesible. Con sólo insertar el enlace CDN en el documento HTML y añadir la directiva ng-app a la etiqueta <body>, la aplicación estaba lista. Además, la documentación y los tutoriales proporcionados por el equipo de desarrolladores eran muy completos y directos (Wohlgethan,2018).

En el verano de 2014, se anunció Angular 2. Angular 2 supuso una reescritura completa del framework. Junto con esta reescritura, también cambiaron muchos de los conceptos centrales del framework. Mientras que AngularJS se centraba en ámbitos y controladores como como patrón de arquitectura, Angular 2 se basa completamente en una jerarquía de componentes (Wohlgethan,2018).

HTML

El lenguaje de marcado de la World Wide Web siempre ha sido HTML. HTML se diseñó principalmente como lenguaje para describir semánticamente documentos científicos, aunque su diseño general y sus adaptaciones a lo largo de los años han permitido utilizarlo para describir otros tipos de documentos, el área principal que no ha sido abordada adecuadamente por HTML es un tema impreciso denominado Aplicaciones Web. Esta especificación trata de rectificar esta situación, al tiempo que actualiza las especificaciones de HTML para abordar las cuestiones planteadas en los últimos años (Hickson & Hyatt, 2011).

TYPESCRIPT

TypeScript es una extensión de JavaScript destinada a facilitar el desarrollo de aplicaciones JavaScript a gran escala. Mientras que cada programa JavaScript es un programa TypeScript, TypeScript ofrece un sistema de módulos, clases, interfaces y un rico sistema de tipos gradual. La intención es que TypeScript proporcione una transición suave para los programadores de JavaScript -los modismos de programación de JavaScript bien establecidos son compatibles sin ninguna reescritura o anotaciones importantes. Una consecuencia interesante es que el sistema de tipos de TypeScript no es estático por diseño. El objetivo de este trabajo es capturar la esencia de TypeScript dando una definición precisa de este sistema de tipos sobre un conjunto de construcciones centrales del lenguaje. Nuestra principal contribución, más allá de las conocidas ventajas de una formalización robusta y matemática, es una refactorización en un fragmento interno seguro y una capa adicional de reglas inseguras (Bierman, & Torgersen, 2014).

Capítulo IV

Desarrollo de la investigación

Introducción

El estudio de investigación es el caso de otorgamiento de créditos en el Fondo Complementario Cerrado de Cesantías de la Universidad de las Fuerzas Armadas ESPE, las características específicas y los diferentes escenarios o calidad de vida de los partícipes afecta en gran manera el proceso de otorgamiento de créditos. Estos factores disminuyen la posibilidad de otorgar préstamos para algunas personas debido a por su ubicación y avance social, financiero, tecnológico, etc., es de rápido crecimiento. Ver capítulo 1 para mayor entendimiento de la problemática.

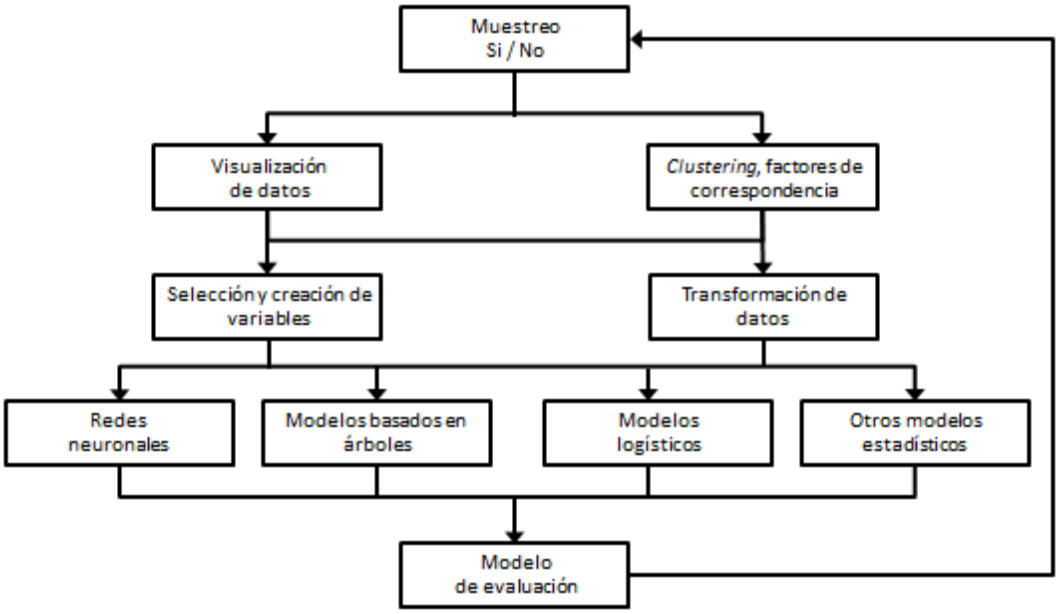
El resultado de esta investigación permitirá determinar un modelo que facilite al Fondo determinar las mejores características de una persona apta para recibir un crédito ya sea quirografario o prendario.

Planificación del proyecto de minería de datos

La planificación está relacionada con el enfoque del proyecto. Las metodologías dirigen el proyecto a la realidad. En este trabajo de investigación se a identificado la metodología SEMMA como la más adecuada de acuerdo a las características del negocio y de la muestra, que en este caso son los partícipes del Fondo Complementario Cerrado de Cesantías de la Universidad de las Fuerzas Armadas ESPE

Figura 3

Flujo Metodología Semma



Nota: Descripción del flujo de la metodología SEMMA (Bulkley y Gayle, 2018).

En la siguiente tabla se muestra las fases que se han tomado en cuenta para el desarrollo del presente trabajo de investigación

Tabla 11

planificación del proyecto

ACTIVIDAD	MES							
	1	2	3	4	5	6	7	8
Análisis de metodologías	x							
Elaboración planificación de proyecto de minería		x						
Comprensión del negocio		x						
Recopilación de datos de participes del fondo		x						
Diseño y creación del repositorio de datos		x	x	x				
Comprensión de datos			x	x				
Preparación de datos				x	x			
Modelamiento y evaluación					x	x	x	
Despliegue						x	x	x
Conclusiones y Recomendaciones								x

Nota: Descripción de la planificación del proyecto.

Fase 1. Muestra (Sample)

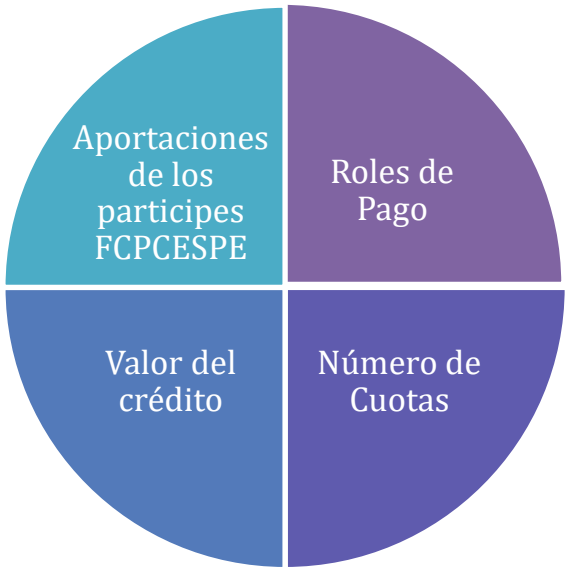
En este estudio de caso se analizaron datos, para identificar las variables relacionadas con características representativas de los diferentes participes. Se seleccionó la base de datos proporcionada por el Fondo Complementario Cerrado de cesantías de la Universidad de las Fuerzas Armadas ESPE, para realizar el mismo análisis en busca de un modelo que permita caracterizar cómo las características de cada participe afectan el otorgamiento de créditos.

Recolección de datos

Al identificar los datos se realizó la configuración de un DataSet en formato csv, con los datos más representativos que permitirán alcanzar el objetivo de la minería, siendo estos datos descritos a continuación:

Figura 4

Factores críticos que afectan en la evaluación del riesgo crediticio.



Nota: Factores que inciden en la evaluación del riesgo crediticio. Fuente (Elaboración propia, 2022)

El DataSet resultante considerando como origen la base de datos del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, base que cuenta con los datos que permiten identificar los diferentes tipos de créditos realizados en las diferentes sedes por los partícipes,

Datos que fueron obtenidos directamente desde Gestor de Base de Datos Sybase, base que cuenta con el registro de 2811 socios distribuidos en las diferentes sedes del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, y generados durante el lapso de 5 años.

El formato del DataSet es del tipo:

- Tipo Texto, la información contenida está en formato texto separado por comas.

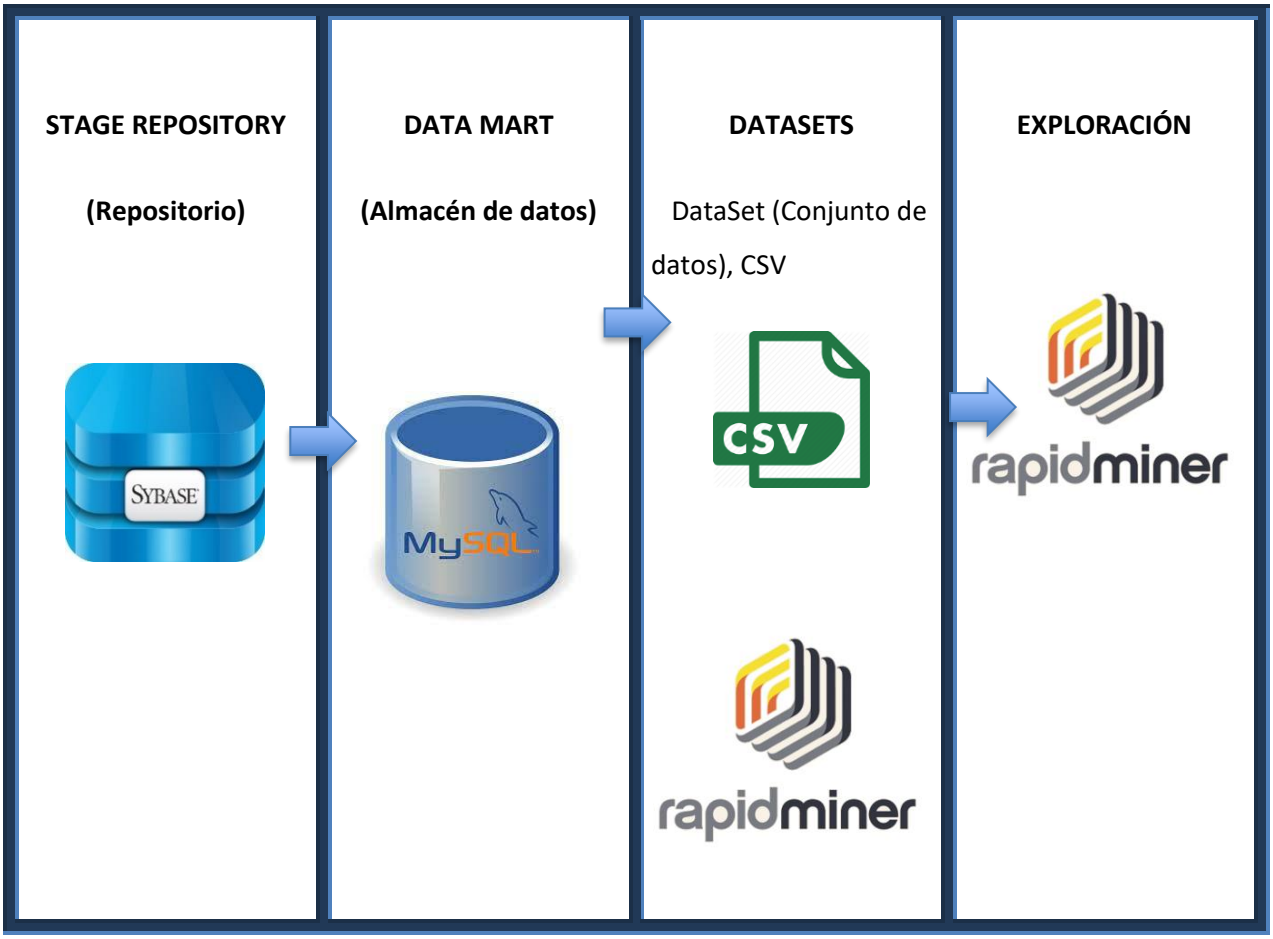
Arquitectura de la solución

Una vez recopilada la información, la arquitectura gestionará todo el proceso de minería de datos para el que fue diseñada. La especificación de la arquitectura se basa en 5 pasos que acompañan el proceso desde la carga de datos hasta la visualización de datos:

- Repositorio Stage (Sybase)
- Data mart (MySQL)
- Json Data mart
- Minería de datos
- Exploración y visualización

Figura 5

Arquitectura para la investigación



Nota: Descripción de la arquitectura para la investigación en cada una de sus etapas. fuente *(elaboración propia, 2022).*

Como siguiente paso, se detalla los 5 pasos de la arquitectura de la solución:

Stage repository (Repositorio): es una herramienta ETL capaz de extraer datos, transformarlos, aplicar reglas comerciales y luego cargarlos en el destino deseado. Puede integrar todo tipo de datos, incluido el big data. DataStage simplifica el análisis empresarial al proporcionar datos de alta calidad para ayudar a impulsar la inteligencia empresarial. Además, proporciona una interfaz gráfica para gestionar de forma intuitiva los procesos de integración de datos.

Data mart (Almacén de datos): Un data mart es una estructura de datos integrada en un repositorio o base de datos. Esta estructura almacena información para que la use una herramienta de análisis o visualización de datos.

Minería de datos: es el proceso de clasificar grandes conjuntos de datos para identificar patrones y relaciones que pueden ayudar a resolver problemas a través del análisis de datos. Las técnicas y herramientas de minería de datos permiten a las empresas predecir tendencias futuras y tomar decisiones comerciales más informadas.

Exploración: La información se extrae directamente del repositorio Data mart para explorar los datos y visualizar los resultados del proceso, así como los factores clave que alteran la aceptación o rechazo de créditos por parte de los agentes crediticios.

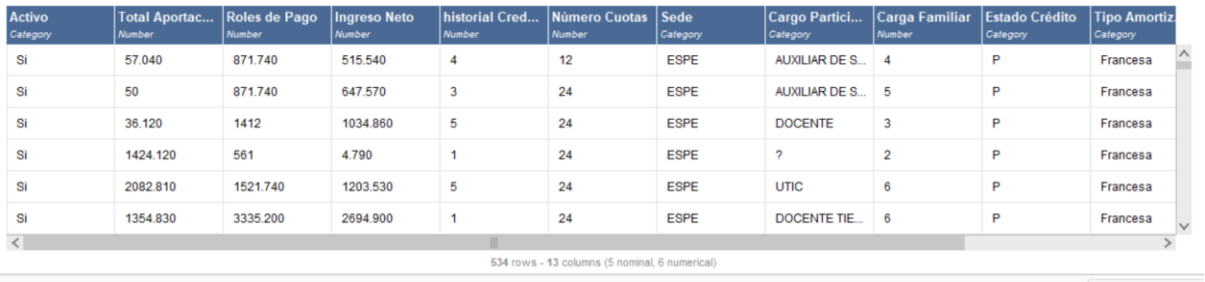
Fase 2. Exploración (Explore)

Establecido el conjunto de datos son créditos registrados en el Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE. En el que se puede identificar el tipo de crédito, el monto, el plazo y el estado del mismo desde su aprobación hasta su cancelación total, procesos que se repiten según el tipo de solicitud de cada socio que han sido realizadas hasta ahora. Se diferencian los créditos por sus características sean “Quirografarios” o “Prendarios”.

La cantidad de registros anuales que se registraron en el DataSet, permiten identificar los créditos que han sido cancelado en su totalidad y los que presentaron inconvenientes, realizando un conteo por número de créditos aprobados como se muestra a continuación, descritos y relacionados por 5 variables nominales y 6 numéricas.

Figura 6

Data set FCPCESPE



Activo Category	Total Aportac... Number	Roles de Pago Number	Ingreso Neto Number	historial Cred... Number	Número Cuotas Number	Sede Category	Cargo Partici... Category	Carga Familiar Number	Estado Crédito Category	Tipo Amortiz Category
Si	57.040	871.740	515.540	4	12	ESPE	AUXILIAR DE S...	4	P	Francesa
Si	50	871.740	647.570	3	24	ESPE	AUXILIAR DE S...	5	P	Francesa
Si	36.120	1412	1034.860	5	24	ESPE	DOCENTE	3	P	Francesa
Si	1424.120	561	4.790	1	24	ESPE	?	2	P	Francesa
Si	2082.810	1521.740	1203.530	5	24	ESPE	UTIC	6	P	Francesa
Si	1354.830	3335.200	2694.900	1	24	ESPE	DOCENTE TIE...	6	P	Francesa

534 rows - 13 columns (5 nominal, 6 numerical)

Nota: identificación de variables de los datos. Fuente (elaboración propia, 2022).

Exploración de datos

El proceso de familiarización con cada una de las variables inicia con la creación de los dataset a partir de los repositorios de datos del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, inicia después de que los datos no modificados se coloquen finalmente en los repositorios de etapa. El análisis se realiza según el tipo de datos, estructura de datos y las conexiones o relaciones con los diferentes origines de datos.

Comprender las relaciones entre las variables en el contexto del problema es el resultado del procedimiento mediante las variables que se han identificado pasaran en las siguientes etapas a convertirse en información valiosa que luego de ser comprendidas tomar un mayor aporte en la construcción del modelo multidimensional que incluirá a los datos seleccionados para el proceso de minería, a continuación se muestra una descripción con los datos del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE que intervendrán en el modelo.

Dimensión tipo de Crédito: contiene información específica de todos los tipos de créditos que se han realizado en el Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE.

Dimensión tiempo: contiene la información de fechas y desglose, sean estos en años o meses, etc.

Hechos de observación: contiene la información relacionada a los créditos concesionados especificando datos financieros del participe, la fecha y la sede.

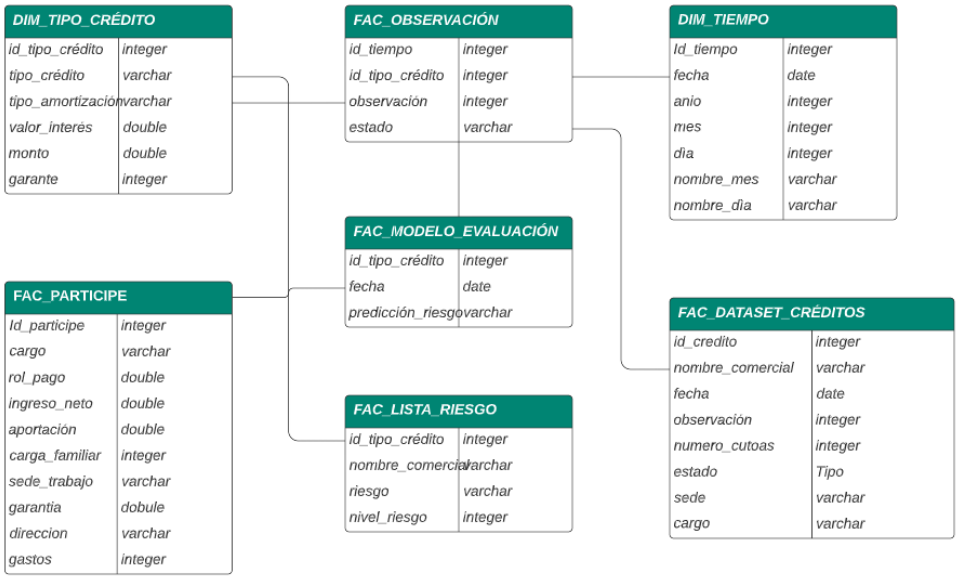
Hechos de dataset créditos: contiene la información de las variables candidatas y las observaciones que serán consideradas como entradas para el desarrollo del trabajo de minería de datos.

Hechos de modelos de evaluación: contiene la información de los resultados que se obtienen de las predicciones de modelos ejecutados.

Hechos de lista riesgos: contiene la información con los niveles de riesgo que se pueden clasificar a los tipos de créditos que se concesionan en el Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE.

Figura 7

Diagrama entidad Relación Multidimensional



Nota: Representa el modelo adecuado para el análisis de los datos de la investigación. Fuente (Elaboración propia, 2022)

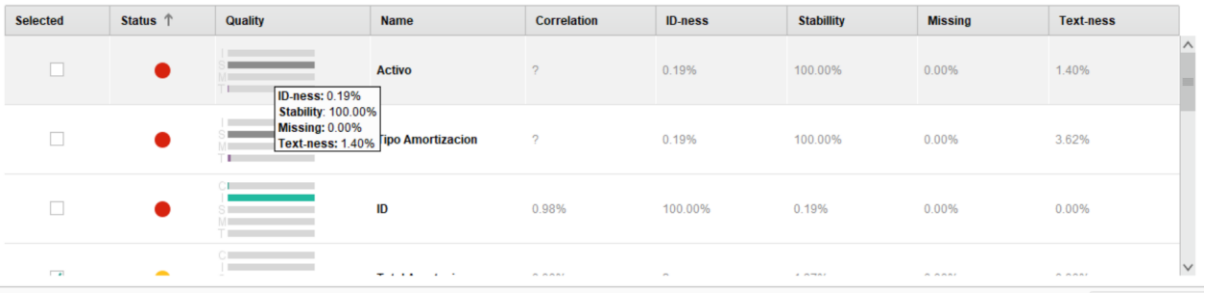
Calidad de los datos

La calidad de los datos es un proceso importante, considerado como un recurso esencial e imprescindible para cualquier desarrollo de análisis de datos, no solo para el proceso de minería de datos. En esta etapa, se verifica la integridad de los datos, se busca eliminar la ambigüedad entre las variables, se establecen reglas de estandarización de datos, se especifican los nuevos tipos de datos para las variables y se procede a considerar como basura a las variables que pueden causar ruido y sesgo durante el proceso de minería, lo que impediría alcanzar la meta planteada.

Para cumplir y garantizar la calidad de los datos se realiza una exploración de los datos mediante técnicas estadísticas que permitirá realizar el seguimiento permitente a los datos permitiendo detectar, identificar y finalmente eliminar los datos que presentan anomalías o que no aportan en el objetivo de la minería, ya que estos afectan en las siguientes fases hacia el descubrimiento de la información.

Figura 8

Detección de calidad de datos fuente



Nota: Se identifica los datos estadísticos que permitan mejorar la calidad de los datos Fuente (Elaboración Propia, 2022).

Fase 3. Modificación (Modify)

En esta fase se realiza la selección y la transformación de los datos en función de las variables que han sido seleccionadas y que cumplieron con los procesos de validación de calidad, que permitan ser un aporte para la transformación a información, permitiendo adaptar el enfoque de la selección y diseño del modelo. En esta fase se extrajo la data de los repositorios de base de datos del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE.

Teniendo como objetivo preparar los datos con la finalidad de generar un nuevo DataSet con todas las variables que tengan un mayor aporte al entendimiento de los datos, dichos datos una vez preparados y depurados mediante las técnicas de estadística, se utilizan como entrada para el modelado de datos y de esta manera empezar con el proceso de construcción de un modelo predictivo que pueda producir los resultados propuestos por la hipótesis.

Exploración del DataSet (conjunto de datos)

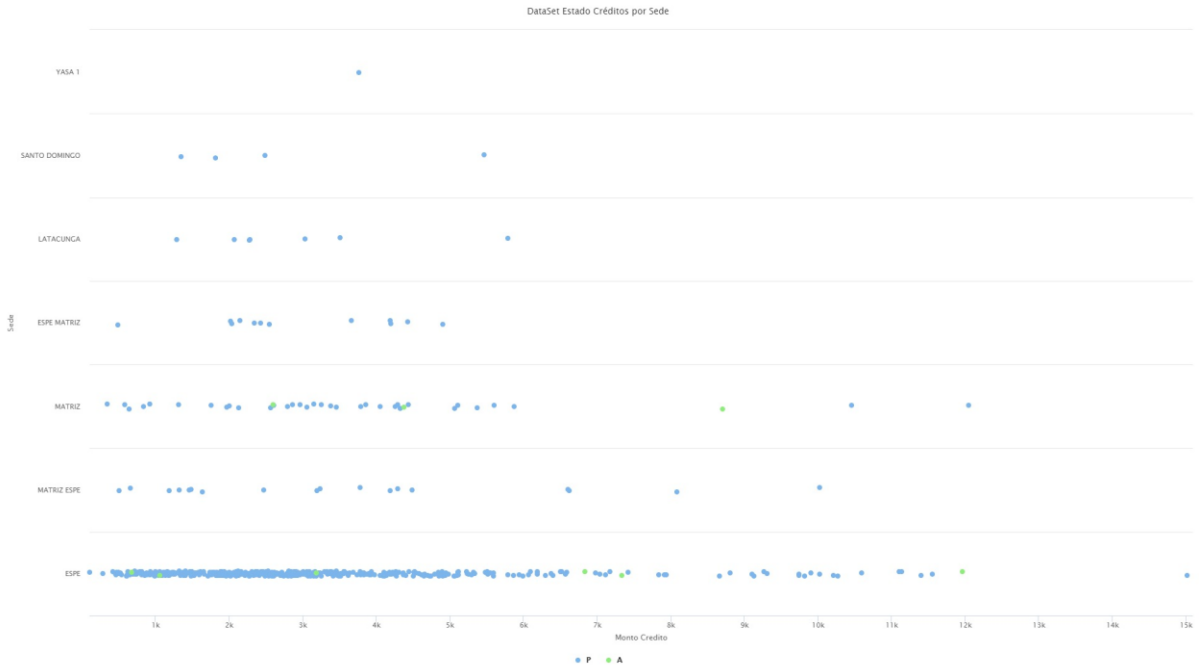
Dado que el conjunto principal de datos son los créditos registrados por parte del personal administrativo del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, en el que se puede visualizar las relaciones entre los créditos desde la fase en

que se realiza la solicitud hasta su finalización, es decir también se identifica los factores que evalúan los agentes de créditos para otorgar el crédito, por ejemplo los roles de pago, capacidad de pago, carga familiar, garantías y otras variables de igual importancia, permitiendo de esta manera identificar la relación entre las variables, los datos corresponden a las sedes del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, claramente delimitados por el tipo de crédito.

Dicho conjunto de datos tiene variables nominales y variables numéricas, el registro de créditos puede ser representado mediante gráficos estadísticos que permita la comprensión de los datos, los datos observados tienen un sesgo que está limitado a la fecha en las fechas en las que se realiza el crédito y el tipo de crédito que es analizando.

Figura 9

Distribución créditos por sede fcpcespe

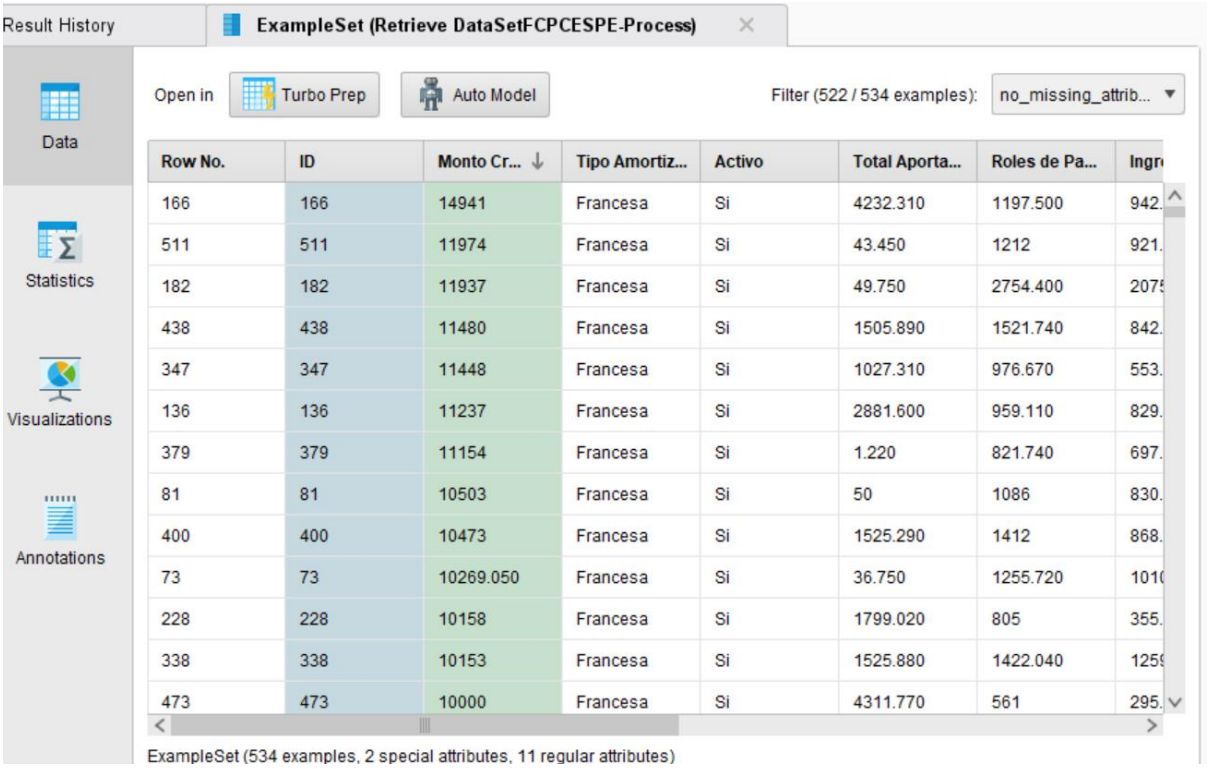


Nota: Muestra La Dispersión de estados de crédito registrados en las diferentes sedes de la fcpcespe

La cantidad de créditos registrados anuales, dan una clara idea sobre cuantas veces se realizan créditos y el estado del mismo, además se puede visualizar en que sedes se realizan más concepciones de créditos y cuáles son los montos aprobados. En el siguiente gráfico, donde se realiza un conteo por “tipo” de crédito.

Figura 10

Dataset fcpcespe



Nota: Dataset de una muestra de los datos de fcpcespe, fuente (elaboración propia, 2022).

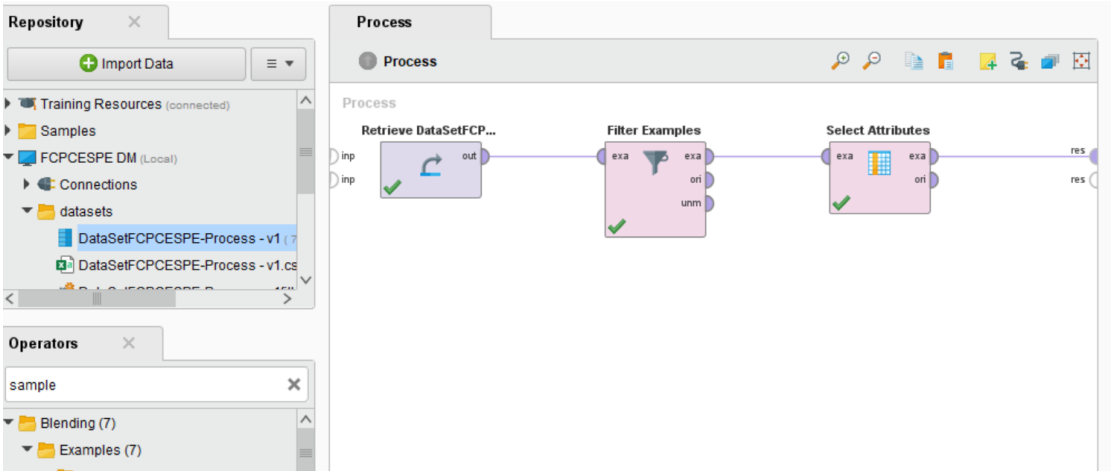
Calidad de dataset (conjunto de datos)

La recopilación de datos proporcionada contenía valores nulos. En esta operación se ignoran todos los valores nulos para todas las variables. Dado que las variables son conjuntos de datos temporales, se procesaron previamente individualmente. Debido a la estructura de los conjuntos de datos-todas las variables fueron recogidas con una temporalidad de un mes, y los valores son los promedios de los valores mensuales, algunos con parcial día y noche-el tratamiento de imputación de

datos que se aplicó a estos conjuntos de datos fue sometido a un proceso de interpolación de datos orientados a series temporales.

Figura 11

Uso de operadores de filtrado



Nota: filtrado de datos a través de los operadores en rapidminer. fuente (elaboración propia, 2022).

Aunque una comprobación superficial podría llevar a uno a creer que los datos son precisos, se debe utilizar una técnica de estadística descriptiva para evaluar los datos. La exploración revela que la variable de "observación" contiene valores atípicos o valores extremos que se desvían de la norma.

La capacidad de determinar si todas las variables están directa o indirectamente relacionadas entre sí es posible gracias a la correlación entre las variables. Cuanto más fuerte es la correlación, o cerca de 1 el coeficiente de correlación es, más estrechamente se conectan las variables, y esto afecta a si se debe incluir una o más de las variables en el modelo. Esto indica que la utilización de una o ambas variables obtendrían el mismo resultado. En el gráfico, se muestran las variables que tienen un alto coeficiente de correlación.

Figura 12

Matriz de correlación de variables

Attribut...	Monto C...	Tipo Cr...	INTERES	Monto +...	Deuda ...	Total Ap...	Roles d...	Ingreso ...	historial...	Número...	Carga F...
Monto Cr...	1	-0.012	-0.012	0.977	0.139	0.008	0.013	0.006	0.006	0.056	-0.004
Tipo Cré...	-0.012	1	1.000	-0.000	0.002	-0.014	-0.013	0.011	0.005	-0.009	0.002
INTERES	-0.012	1.000	1	-0.000	0.002	-0.014	-0.013	0.011	0.005	-0.009	0.002
Monto + i...	0.977	-0.000	-0.000	1	0.164	0.004	0.011	0.010	0.007	0.241	-0.003
Deuda P...	0.139	0.002	0.002	0.164	1	-0.008	0.011	-0.001	-0.021	0.073	-0.000
Total Ap...	0.008	-0.014	-0.014	0.004	-0.008	1	0.012	0.014	0.005	-0.008	0.009
Roles de...	0.013	-0.013	-0.013	0.011	0.011	0.012	1	0.049	-0.011	0.001	-0.004
Ingreso ...	0.006	0.011	0.011	0.010	-0.001	0.014	0.049	1	-0.008	0.021	-0.003
historial ...	0.006	0.005	0.005	0.007	-0.021	0.005	-0.011	-0.008	1	0.005	-0.005
Número ...	0.056	-0.009	-0.009	0.241	0.073	-0.008	0.001	0.021	0.005	1	0.007
Carga F...	-0.004	0.002	0.002	-0.003	-0.000	0.009	-0.004	-0.003	-0.005	0.007	1

*Nota:*1 Correlación entre las variables candidatas del modelo. Fuente (Elaboración propia, 2022)

Se han descubierto que determinadas variables están correlacionadas. Las variables de la correlación son las siguientes:

- Monto Crédito
- Tasa interés
- Ingreso Neto
- Numero Aportaciones
- Roles de Pago
- Monto + Interés
- Deuda Pendiente

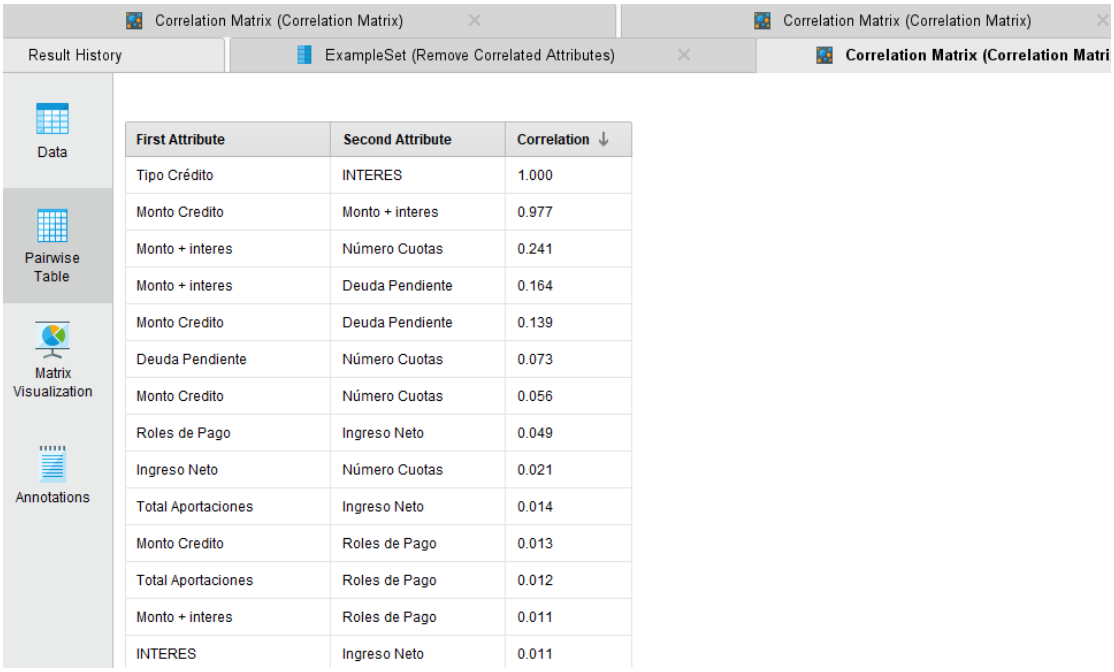
Donde se realiza la correlación entre las variables:

- Tipo de Crédito e interés
- Monto Crédito y Monto + interés
- Monto+ Interés y Número de Cuotas

- Monto+ Interés y Deuda Pendiente
- Monto Crédito y Deuda Pendiente.

Figura 13

Variables con mayor correlación



First Attribute	Second Attribute	Correlation
Tipo Crédito	INTERES	1.000
Monto Credito	Monto + interes	0.977
Monto + interes	Número Cuotas	0.241
Monto + interes	Deuda Pendiente	0.164
Monto Credito	Deuda Pendiente	0.139
Deuda Pendiente	Número Cuotas	0.073
Monto Credito	Número Cuotas	0.056
Roles de Pago	Ingreso Neto	0.049
Ingreso Neto	Número Cuotas	0.021
Total Aportaciones	Ingreso Neto	0.014
Monto Credito	Roles de Pago	0.013
Total Aportaciones	Roles de Pago	0.012
Monto + interes	Roles de Pago	0.011
INTERES	Ingreso Neto	0.011

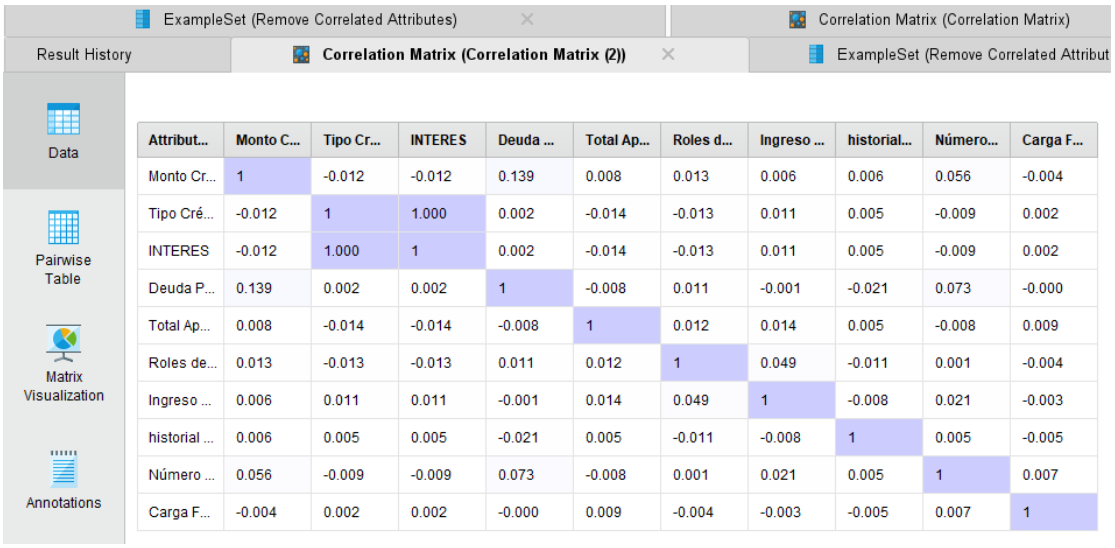
Nota: Identificación de las variables con mayor correlación, fuente (elaboración propia, 2022)

Los siguientes factores se eliminan del procedimiento de modelado en este contexto:

- Activo
- Cargo Participe
- Sede
- Tipo de Amortización

Figura 14

Matriz de correlación reducida



Attribut...	Monto C...	Tipo Cr...	INTERES	Deuda ...	Total Ap...	Roles d...	Ingreso ...	historial...	Número...	Carga F...
Monto Cr...	1	-0.012	-0.012	0.139	0.008	0.013	0.006	0.006	0.056	-0.004
Tipo Cré...	-0.012	1	1.000	0.002	-0.014	-0.013	0.011	0.005	-0.009	0.002
INTERES	-0.012	1.000	1	0.002	-0.014	-0.013	0.011	0.005	-0.009	0.002
Deuda P...	0.139	0.002	0.002	1	-0.008	0.011	-0.001	-0.021	0.073	-0.000
Total Ap...	0.008	-0.014	-0.014	-0.008	1	0.012	0.014	0.005	-0.008	0.009
Roles de...	0.013	-0.013	-0.013	0.011	0.012	1	0.049	-0.011	0.001	-0.004
Ingreso ...	0.006	0.011	0.011	-0.001	0.014	0.049	1	-0.008	0.021	-0.003
historial ...	0.006	0.005	0.005	-0.021	0.005	-0.011	-0.008	1	0.005	-0.005
Número ...	0.056	-0.009	-0.009	0.073	-0.008	0.001	0.021	0.005	1	0.007
Carga F...	-0.004	0.002	0.002	-0.000	0.009	-0.004	-0.003	-0.005	0.007	1

Nota: Las variables con correlación fuerte se excluyen ya que causan ruido en el conjunto de datos. Fuente (elaboración propia,2022)

Al eliminar las variables correlacionadas, la matriz ya no exhibe un alto coeficiente entre las variables, lo que implica que todas las variables son independientes entre sí, mejorando el modelo para el proceso de minería de datos.

Construcción de datasets (conjuntos de datos)

las variables de esta actividad se deducen de otras. No fue esencial crear nuevas variables para este procedimiento, ya que los factores que son cruciales en la concesión de créditos en las diferentes sedes del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, ya han sido tomados en consideración.

Sin embargo, como se ve en la tabla siguiente, el conjunto de datos se ha generado de acuerdo con elementos importantes que incluyen las variables características que representan al participe y a su relación con la concesión de créditos.

Tabla 12

Variables candidatas para la investigación

N	Variable	Descripción
1	Tipo Crédito	Describe la dirección de aplicación del cliente/distrito
2	Interés	Describe la el interés según el tipo y plazo seleccionado para el crédito.
3	Roles de pago Participe	Describe los ingresos mensuales del cliente por parte de la ESPE
4	Tipo de préstamo	Describe al cliente del préstamo/para qué propósito/
5	Numero de Cuotas	Describe el número de cuotas a pagar el crédito.
6	Capital + Interés	Describe la deuda total del préstamo
7	Total, Aportaciones	Describe el valor total de aportaciones del participe.
8	Ingreso Neto	Describe el ingreso total del participe, contemplando ingresos extras y egresos.
9	Historial de Créditos aprobados	Describe el historial de créditos Realizados
10	Carga Familiar	Describe el número de carga familiar del participe
11	Deuda Pendiente	Describe la deuda pendiente de pago.

Nota: Se describen las variables candidatas que se incluirán en el modelo de predicción.

Integración de datasets (conjunto de datos)

De esta forma, la calidad de cada conjunto de datos se trabajó de forma independiente antes de que cada una se comparara de acuerdo con la variable de fecha de giro, que es común a todos los conjuntos de datos elegidos. La creación del conjunto de datos como entrada al modelado, tenía varias entradas o bases de datos, que giraban en torno a 4 aspectos. Aportaciones de los partícipes, los roles de pago del participe, el valor del crédito y el número de cuotas, que son particulares a las observaciones del registro de créditos concedidos en las diferentes sedes, y son el objetivo del estudio,

se utilizaron para proporcionar los cruces de información general. Este conjunto de datos o estos conjuntos de datos son los elegidos para el proceso de modelado de datos. Aunque se pueden ver las variables enlazadas, no se incluyen en esta etapa del proceso de modelado ya que son variables que no proporcionan valor al modelo.

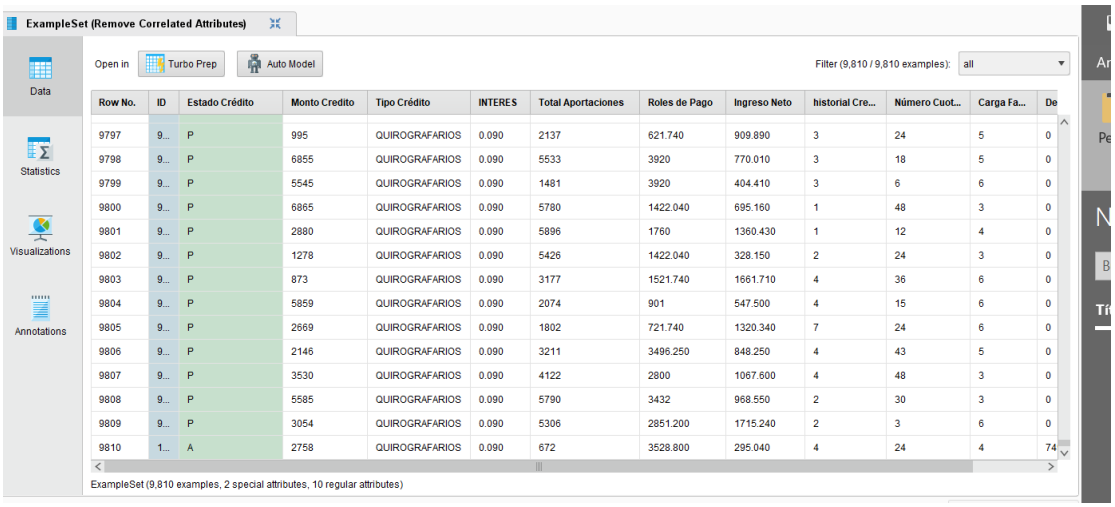
Figura 15

Variables candidatas para el entrenamiento de los modelos

Nota: Una descripción de las variables potenciales que se incluirán en el modelo de predicción. Fuente (elaboración propia,2022)

Figura 16

Dataset FCPCESPE para entrenamiento



Row No.	ID	Estado Crédito	Monto Credito	Tipo Crédito	INTERES	Total Aportaciones	Roles de Pago	Ingreso Neto	historial Cre...	Número Cuot...	Carga Fa...	De
9797	9...	P	995	QUIROGRAFARIOS	0.090	2137	621.740	909.890	3	24	5	0
9798	9...	P	6855	QUIROGRAFARIOS	0.090	5533	3920	770.010	3	18	5	0
9799	9...	P	5545	QUIROGRAFARIOS	0.090	1481	3920	404.410	3	6	6	0
9800	9...	P	6865	QUIROGRAFARIOS	0.090	5780	1422.040	695.160	1	48	3	0
9801	9...	P	2880	QUIROGRAFARIOS	0.090	5896	1760	1360.430	1	12	4	0
9802	9...	P	1278	QUIROGRAFARIOS	0.090	5426	1422.040	328.150	2	24	3	0
9803	9...	P	873	QUIROGRAFARIOS	0.090	3177	1521.740	1661.710	4	36	6	0
9804	9...	P	5859	QUIROGRAFARIOS	0.090	2074	901	547.500	4	15	6	0
9805	9...	P	2669	QUIROGRAFARIOS	0.090	1802	721.740	1320.340	7	24	6	0
9806	9...	P	2146	QUIROGRAFARIOS	0.090	3211	3496.250	848.250	4	43	5	0
9807	9...	P	3530	QUIROGRAFARIOS	0.090	4122	2800	1067.600	4	48	3	0
9808	9...	P	5585	QUIROGRAFARIOS	0.090	5790	3432	968.550	2	30	3	0
9809	9...	P	3054	QUIROGRAFARIOS	0.090	5306	2851.200	1715.240	2	3	6	0
9810	1...	A	2758	QUIROGRAFARIOS	0.090	672	3528.800	295.040	4	24	4	74

Nota: Dataset (conjunto de datos) para entrenamiento de los modelos de Machine Learning. Fuente (Elaboración propia, 2022)

Fase 4. Model (Modelo)

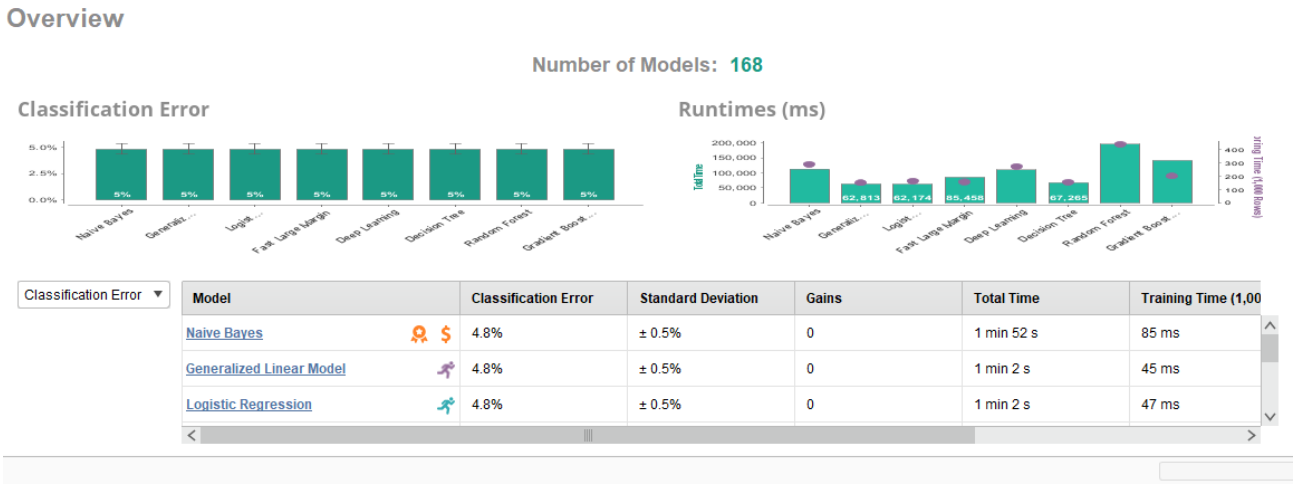
Esta parte es complicada porque considera una variedad de factores como la cantidad de variables, los tipos de datos, los valores de los datos, y los muchos modelos de selección que sirven de base para las predicciones. Para las preguntas de hipótesis de esta sección y el objetivo sugerido de ser satisfechos, las variables y el modelo elegido son cruciales.

Rapidminer Educational Edition en su forma académica es uno de los programas de modelado elegidos debido a la diversidad de modelos y facilidad de uso. Una de las características clave es "auto modelo", que analiza el conjunto de datos elegido y, en función de los factores y los datos, ofrece qué modelos están mejor valorados para pronosticar. Esta característica se utilizó en esta parte para identificar rápidamente los modelos que se desempeñarían mejor con el conjunto de datos sugerido y para ajustar los modelos.

Después de ejecutar el "modelo automático", se descubrió que Deep Learning, Gradient Potencied Trees, Random Forest, Generalized Linear y Decision Tree son los modelos que mejor se adaptan o adaptan al trabajo en predicción con el conjunto de datos suministrado. La lista final será determinada por los envíos de prueba, sin embargo, los primeros contendientes se muestran aquí:

Figura 17

lista de modelos de regresión candidatos



Nota: Resumen con la salida de los valores obtenidos al realizar pruebas en los diferentes modelos candidatos para el aprendizaje. FUENTE (elaboración propia, 2022)

Selección de variables

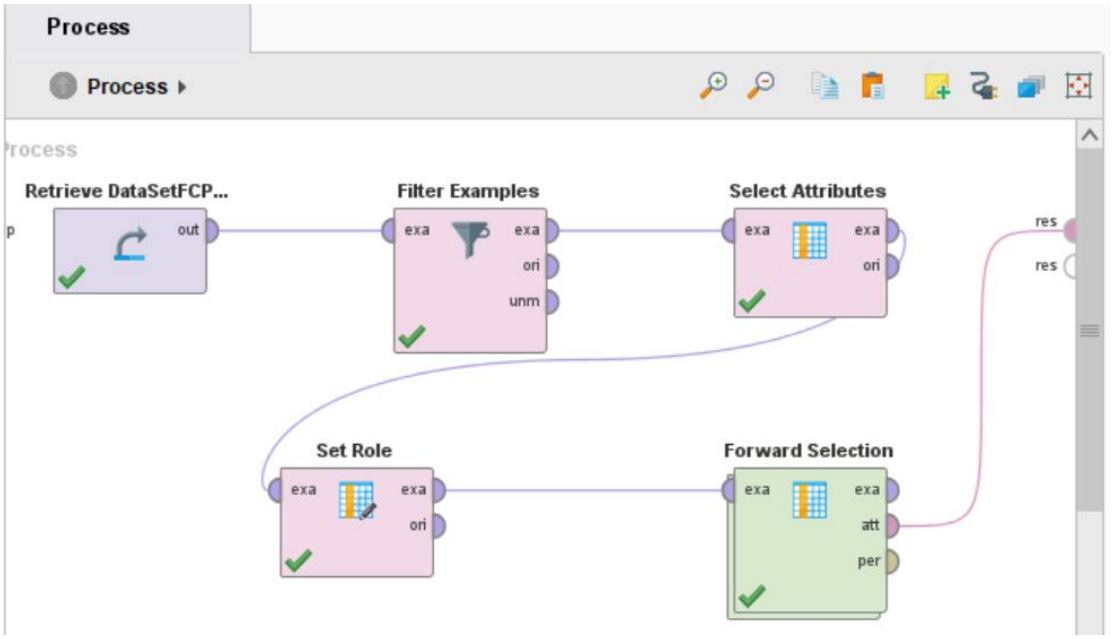
El dataset (conjunto de datos) comprende una serie de factores antes del diseño del modelo, que normalmente no siempre es aceptable para la construcción del modelo. Esto se debe a que ciertas variables pueden inducir el ruido en el modelo, lo cual no es ideal y puede reducir el rendimiento o la precisión de la misma. Los enfoques o procedimientos para seleccionar variables para optimizar el modelo y mejorar la eficacia del modelo de predicción se discuten en esta sección. Debido a la simplicidad de las operaciones, la variable de fecha se dividió en año y mes. Cada técnica da a cada

variable una puntuación o un peso basado en lo esencial que se piensa que es como una variable independiente en el modelo.

Método Forward: es una técnica de selección de variables " comienza con una selección vacía de atributos y añade cada atributo no necesario en el conjunto de datos de cada iteración. Los operadores internos se utilizan para estimar el rendimiento de cada atributo agregado. " (RapidMiner, 2022). El resultado de la ejecución del modelo indica qué variables son adecuadas y cuáles son insuficientes. Los valores 1 y 0 se devuelven en ese orden.

Figura 18

Método forward para la selección de variables



Nota: Técnica para comparar la eficacia de las variables que se incluirían en los modelos de predicción. fuente (elaboración propia, 2022)

Las variables de relevancia en este modelo son: las variables de numero de cuotas, monto crédito, monto + interés, ingreso neto, deuda pendiente. De acuerdo con la técnica, los otros factores simplemente producirán ruido o tendrán una influencia perjudicial en el desempeño del modelo de predicción.

Figura 19

Resultado método forward

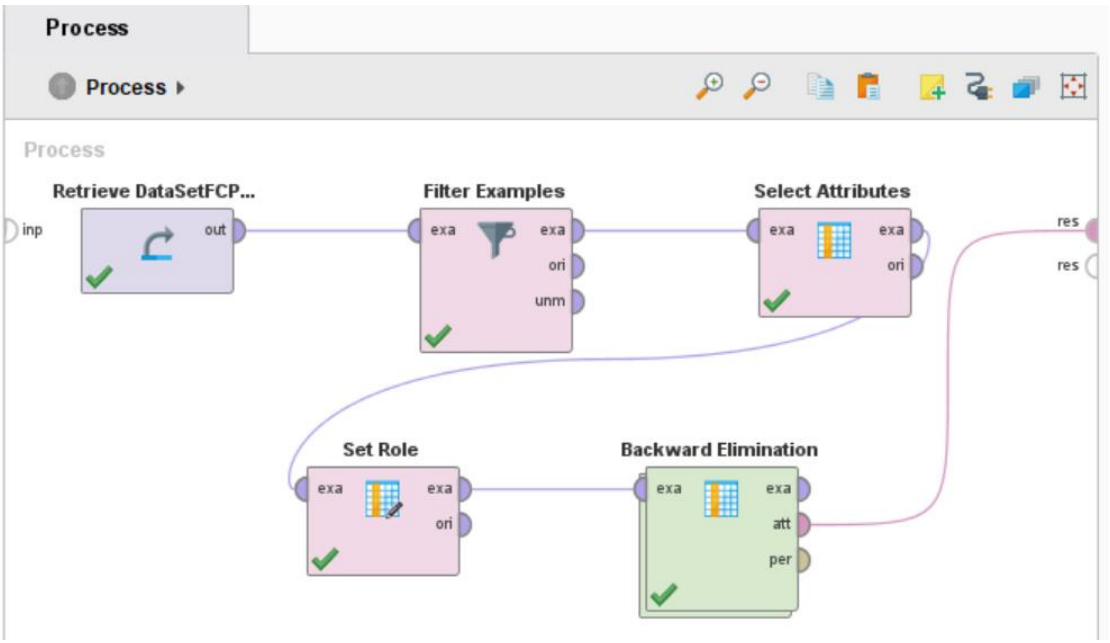
attribute ↓	weight
historial Creditos Aprobados	0
Total Aportaciones	0
Tipo Crédito	0
Roles de Pago	0
Número Cuotas	1
Monto Credito	1
Monto + interes	1
Ingreso Neto	1
INTERES	0
Deuda Pendiente	1
Carga Familiar	0

Nota: El método forward descarta más de la mitad de las variables. Fuente (elaboración propia, 2022)

Método Backward: Esta técnica sigue los mismos pasos que el anterior, pero a la inversa " comienza con todo el conjunto de características y elimina cada atributo restante del conjunto de datos en cada iteración. El rendimiento se calcula utilizando operadores internos como la validación cruzada para cada atributo eliminado. Por último, sólo se elimina de la selección la propiedad con la menor caída de rendimiento. Luego, con la selección actualizada, comienza una nueva ronda ". (RapidMiner, 2022). Las variables con alto rendimiento tomaran el peso de 1, el resto con peso 0.

Figura 20

Método Backward para selección de variables



Nota: Sólo una variable fue descartada por el método backward. Fuente (elaboración propia, 2022).

El resultado de comparar todas las variables al principio de las iteraciones indica que todos ellos, salvo uno, son candidatos para su inclusión en los modelos de predicción.

Figura 21

Resultado método backward

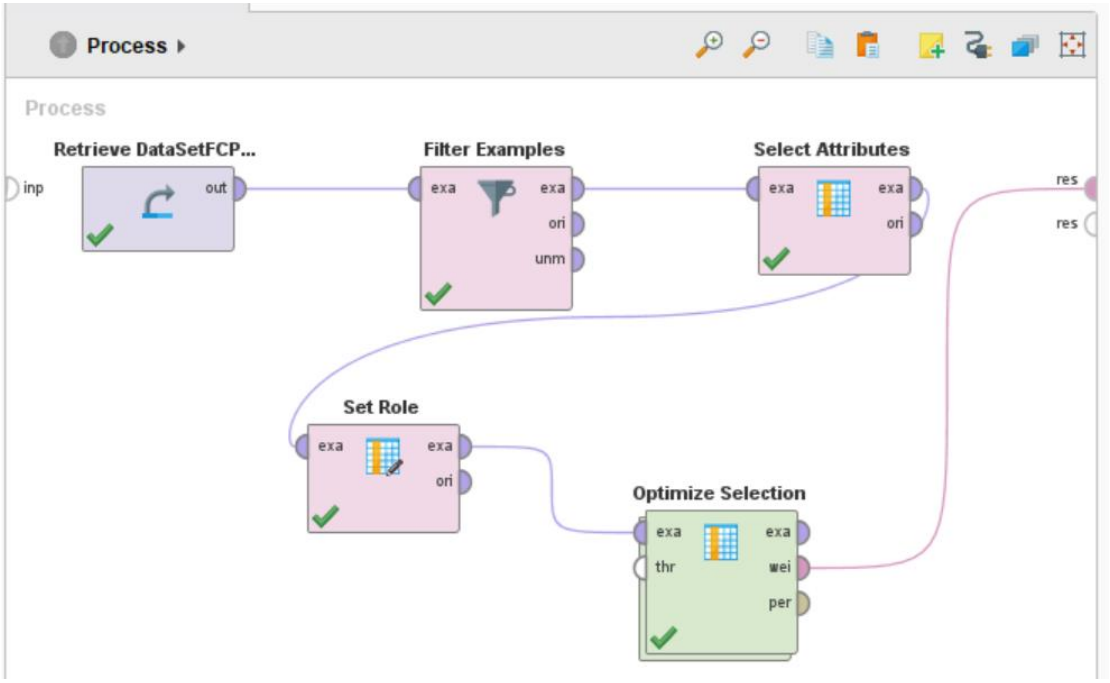
attribute	weight
Tipo Crédito	0
INTERES	1
Monto + interes	1
Deuda Pendiente	1
Total Aportaciones	1
Roles de Pago	1
Ingreso Neto	1
historial Creditos Aprobados	1
Número Cuotas	1
Carga Familiar	1
Monto Credito	1

Nota: Sólo se elimina una variable del conjunto de datos utilizando el método backward. Fuente (elaboración propia, 2022)

Método Optimize Selection: Este enfoque optimiza la selección de variables combinando las ventajas de los métodos anteriores, ya que "pierde las cualidades más relevantes de la colección de datos". Para la selección de características, se utilizan dos métodos deterministas de características interesados, forward y backward. (RapidMiner, 2022). A las variables más eficientes se les asignará un peso de 1, mientras que a las demás se les asignará un peso de 0.

Figura 22

Método Optimize Selection



Nota: Este método es más estricta en términos de selección de variables candidatas. fuente (elaboración propia, 2022).

Al combinar dos métodos de selección, este método eliminó más de la mitad de las variables, lo que implica que añade poco valor al modelo mientras conserva la precisión.

Figura 23

Resultado método Optimize Selection

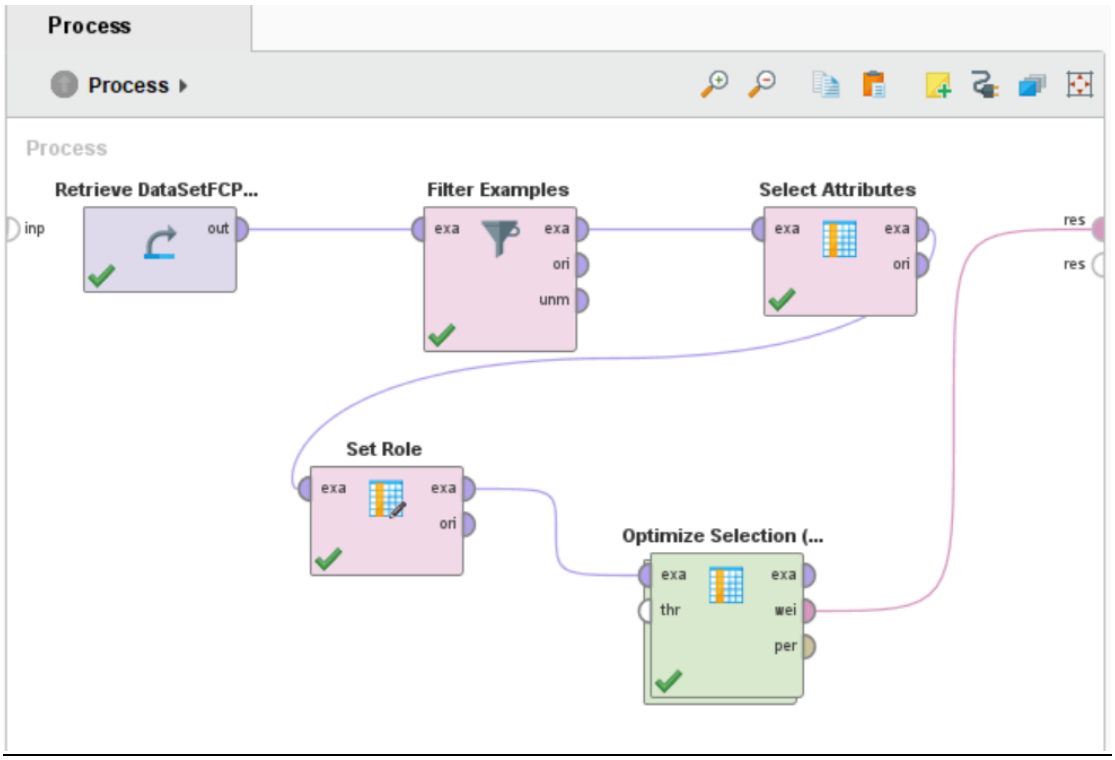
attribute ↓	weight
historial Creditos Aprobados	0
Total Aportaciones	0
Tipo Crédito	0
Roles de Pago	0
Número Cuotas	1
Monto Credito	0
Monto + interes	1
Ingreso Neto	0
INTERES	0
Deuda Pendiente	1
Carga Familiar	0

Nota: En este método aumentaron el número de variables descartados en comparación al método forward. Fuente (elaboración propia, 2022).

Método Optimize Selection (Brute Force): Este método requiere mucho procesamiento y memoria, pero permite localizar a los principales candidatos " Determinando la colección óptima de atributos mediante la evaluación de todas las combinaciones de atributos potenciales Devuelve el subconjunto de cualidades que lograron el mayor rendimiento. Este operador tiene un tiempo de ejecución exponencial puesto que opera en el conjunto de la alimentación en el conjunto de atributos ". (RapidMiner, 2022). Este enfoque se debe utilizar con precaución ya que cuantos más factores compares, más tiempo puede tomar.

Figura 24

Método Optimize selection (brute Force).



Nota: Selección de variables con método Optimize selection (brute Force). Fuente (elaboración propia, 2022).

En comparación con el modelo de backward, esta estrategia ofrece la misma respuesta que las variables tienen rendimientos fuertes. Sólo la variable ineficaz varía.

Figura 25

Salida Método Optimize selection (brute Force).

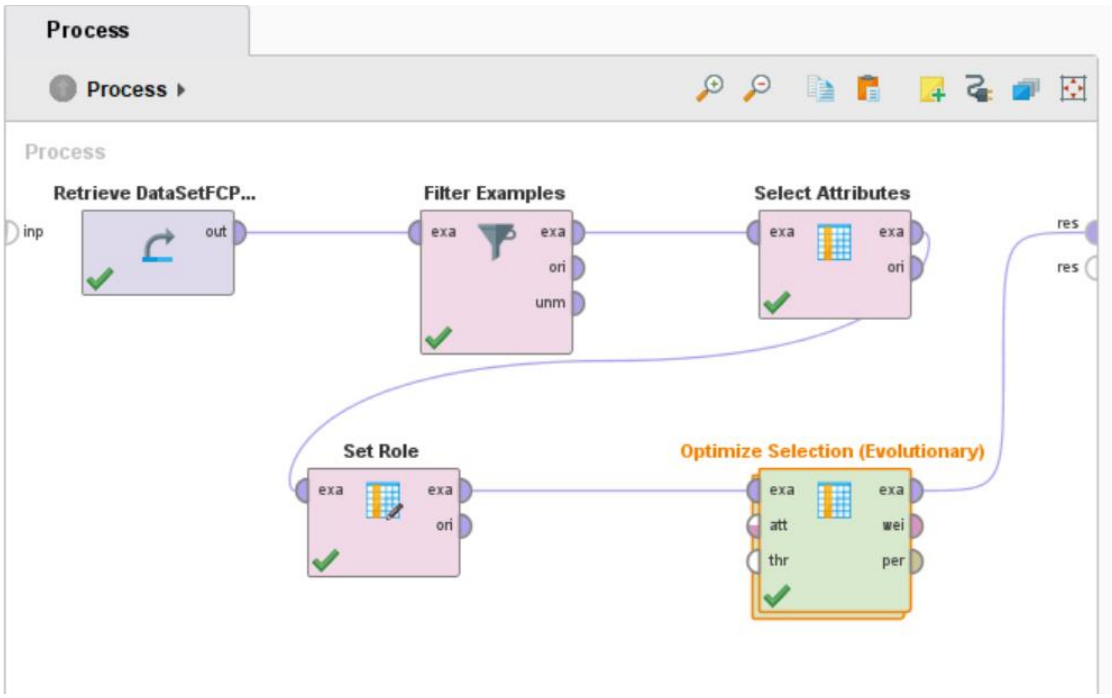
attribute	weight ↓
Deuda Pendiente	1
Tipo Crédito	0
INTERES	0
Monto + interes	0
Total Aportaciones	0
Roles de Pago	0
Ingreso Neto	0
historial Creditos Aprobados	0
Número Cuotas	0
Carga Familiar	0
Monto Credito	0

Nota: Con el método Optimize selection (brute Force) que solo deuda pendiente, tiene alto rendimiento. Fuente (elaboración propia, 2022).

Método Optimize Selection (Evolutionary): Este enfoque optimiza el algoritmo de selección para las variables seleccionadas, o "utiliza heurísticas de búsqueda que replican el proceso de evolución natural". Esta heurística se utiliza a menudo para crear soluciones significativas a los problemas de optimización y búsqueda. Los algoritmos genéticos son la clase más amplia de algoritmos evolutivos (EA), que crean soluciones a problemas de optimización utilizando enfoques inspirados en la evolución natural como la herencia, la mutación, la selección y el crossover” (RapidMiner, 2022).

Figura 26

Selección de variable mediante método Optimize Selection (evolutionary)



Nota: Las variables condicionales se eligen utilizando algoritmos genéticos. Fuente (elaboración propia, 2022).

El método Optimize Selection (evolutionary) (evolutivo), utiliza más de la mitad de las variables como efectos para el modelo de predicción. Varios de los factores rechazados por este procedimiento fueron tenidos en cuenta por los modelos antes mencionados.

Figura 27

Resultados obtenidos mediante aplicación método Optimize selection (evolutionary)

attribute ↓	weight
historial Creditos Aprobados	1
Total Aportaciones	1
Tipo Crédito	0
Roles de Pago	1
Número Cuotas	1
Monto Credito	1
Monto + interes	1
Ingreso Neto	1
INTERES	0
Deuda Pendiente	1
Carga Familiar	0

Nota: La salida de este método reflejo como resultado un aumento de variables útiles frente a las variables de descarte. fuente (elaboración propia, 2022).

Las posibles variables que pueden participar en la creación de los modelos de datos se suministran siguiendo una batería de algoritmos para establecer las características eficientes para ser participantes activos de los modelos de regresión. Una vez creados y/o refinados los modelos, estas variables se incorporarán a los modelos finales.

Tabla 13

Puntaje obtenido por los métodos de selección

Variable	Optimize Selection	Backward	Forward	Brute Force	Evolutionary	Peso
Historial Créditos Aprobados	0	1	0	0	1	2
Total Aportaciones	0	1	0	0	1	2
Tipo Crédito	0	0	0	0	0	0

Roles de Pago	0	1	0	0	1	2
Números de Cuotas	1	1	1	0	1	4
Monto Crédito	0	1	1	0	1	3
Monto + Interés	1	1	1	0	1	4
Ingreso neto	0	1	1	0	1	3
Interés	0	1	0	0	0	1
Deuda Pendiente	1	1	1	1	1	5
Carga Familiar	0	1	0	0	0	1

Nota: Variables seleccionadas por los modelos Fuente (Elaboración propia, 2022).

El total de las puntuaciones suministradas por cada técnica de selección se utiliza para seleccionar las variables; en este ejemplo, las variables cuyos valores agregan 4 -5 puntos, como se indica en la Tabla 12, son tomadas, es decir, aceptadas por cada método. Las variables con una puntuación de 3 están condicionadas y se incluirían en los modelos si no cambiaban el comportamiento de los modelos. La tabla siguiente muestra el resultado de la ejecución de las técnicas de selección:

Tabla 14

Variables aceptadas para los modelos

Variable	Campo	Estado
Deuda Pendiente	Deuda Pendiente	Aceptada
Números de Cuotas	Números de Cuotas	Aceptada
Monto + Interés	Monto + Interés	Aceptada
Números de Cuotas	Números de Cuotas	Condicionada
Ingreso neto	Ingreso neto	Condicionada

Nota: Elección de las variables óptimas para los modelos de regresión. fuente (elaboración propia, 2022).

El resto de las variables fueron eliminadas de consideración debido a su bajo peso en el conjunto de datos. Como resultado, no se utilizarán en el diseño del modelo de regresión.

Selección de los modelos

Para la construcción de modelos, se requiere que el enfoque correcto cumpliese los objetivos especificados y los datos recogidos que compren el amplio contexto del problema; en este estudio, el objetivo es tener una predicción del riesgo de crédito en el Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, con un dataset de 9810 registros de créditos.

Los modelos de predicción requieren un número mínimo de registros en los datos con los que se debe trabajar para que los modelos puedan formarse y evaluarse. Los registros se recogerán desde los repositorios de base de datos de las sedes del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE para esta tesis.

Los modelos se emplean en enfoques de minería de datos. En este estudio, fueron examinados previamente con la herramienta Rapidminer, que permite que los datos se presenten a varios algoritmos para determinar con mayor eficacia cuáles son los mejores algoritmos de predicción que se van a ajustar al conjunto de datos. Los modelos seleccionados se seleccionarán para cada tipo de relación, siempre que se adapten en entornos similares, y se elegirá uno de cada categoría.

Tabla 15

Modelos de predicción

Modelo	Tipo	Selección
Deep Learning	Neuronal	Si
Naive Bayes	Aprendizaje Automático	Si
Gradient Boosted Tree	Árboles	Si
Random Forest	Árboles	No
Decision Tree	Árboles	No

Nota: Selección de modelos de predicción. (elaboración propia, 2022).

Rapidminer puede abordar problemas de clasificación y regresión en la categoría de predicción. Los modelos automatizados pueden ayudarle a evaluar los datos, proporcionar modelos aplicables para resolver problemas y comparar los resultados de estos modelos cuando haya finalizado el cálculo.

Figura 28

Comparación de resultados automodelo

Row No.	Model	Classificatio...	Standard De...	Gains	Total Time	Training Time ...	Scoring Time...
1	Naive Bayes	0.048	0.005	0	112704	85.117	288.226
2	Generalized Linear...	0.048	0.005	0	62813	44.648	151.886
3	Logistic Regression	0.048	0.005	0	62174	47.197	163.609
4	Fast Large Margin	0.048	0.005	0	85458	249.541	156.983
5	Deep Learning	0.048	0.005	0	109556	824.669	272.681
6	Decision Tree	0.048	0.005	0	67265	10.907	154.434
7	Random Forest	0.048	0.005	0	195016	37.920	438.838
8	Gradient Boosted ...	0.048	0.005	0	141301	74.414	203.364

Nota: Comparación de datos estadísticos obtenidos de los modelos generados por auto modelado. Fuente (elaboración propia, 2022).

Diseño del modelo

Las variables que componen los diseños de los modelos y sus relaciones entre sí siguen visualmente el mismo patrón general.

Los modelos se crearán utilizando el programa Rapidminer, cuyos procedimientos para componentes visuales facilitan el desarrollo. Para crear un modelo aceptable, la construcción general de los modelos se compone de procesos concatenados. En esta situación, la división será 70/30, o el 70% de los datos para la formación y el 30% para las pruebas, ya que todo el conjunto de datos actúa como entrenamiento y prueba para el aprendizaje de los modelos. Antes de llegar a la despensa de datos, este conjunto de datos se establece mediante un procedimiento de calidad de datos que se detalla en la arquitectura de la solución.

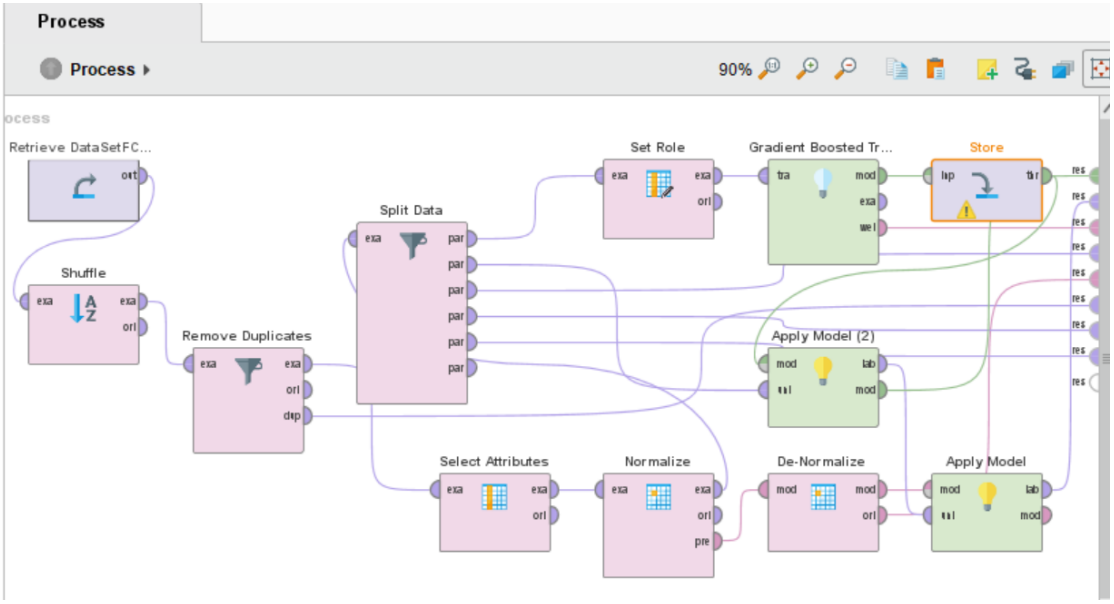
Los datos se mezclan en el siguiente paso para barajar a fondo los datos. Después de eso, los datos se normalizan y las variables que conformarán el modelo se eligen utilizando un conjunto de datos de variables con varias unidades de medida. La función del campo se determina entonces para los datos de entrenamiento cuando se hacen las predicciones; en esta instancia, es la variable "observación" para todos los modelos. La utilización del modelo elegido, cuya salida irá directamente al modelo dimensional, es el paso siguiente.

- El procedimiento puede resumirse del siguiente modo:
- El conjunto de datos es de entrada
- El conjunto de datos de tipos de crédito es barajar
- Las variables finales a utilizar con el modelo se eligen
- Las variables seleccionadas se normalizan y el conjunto de datos se divide en mitades de entrenamiento y de prueba.
- Elija el rol o variable dependiente, o predictor.
- Escriba el modelo utilizando los coeficientes proporcionados por el modelo.
- Desnormalice el conjunto de datos y envíe la salida al origen de la etapa (no es necesario, pero útil para la revisión)

Los modelos creados para la investigación fueron diseñados de la siguiente manera:

Figura 29

Diseño del modelo de regresión Gradiente Boosted tree

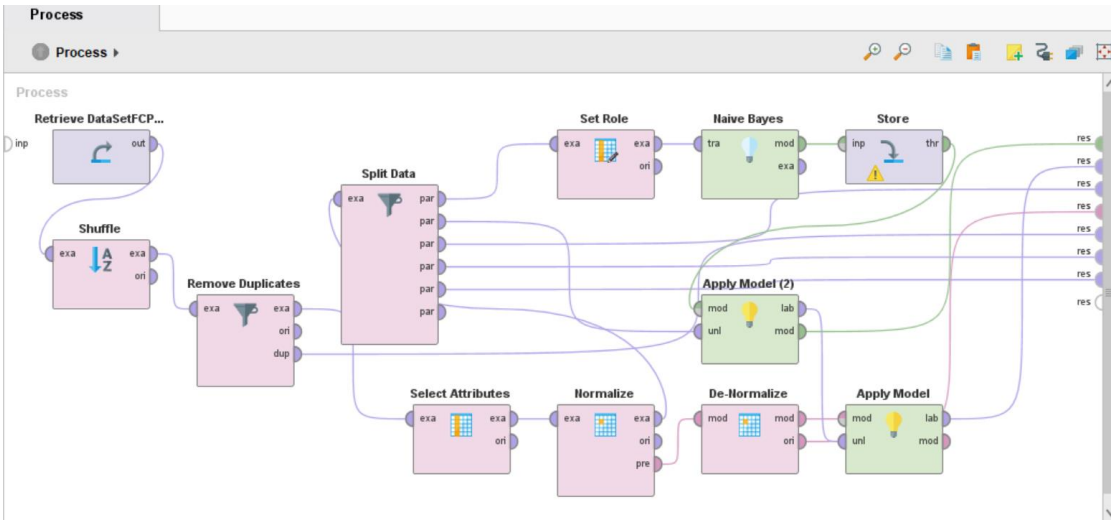


Nota: Proceso de diseño con operadores para el modelo Gradient Boosted Tree. Fuente (elaboración propia, 2022).

El modelo Gradient Boosted Tree fue sintonizado utilizando un total de 100 árboles basados en el tipo y el promedio, debido a que los valores bajos en la media requieren más sintonía. Una profundidad de 4, una tasa de aprendizaje de 0.1, y una función de distribución de Poisson mejoran el modelo. El diseño sigue el método estándar establecido.

Figura 30

Diseño de modelo Naive bayes

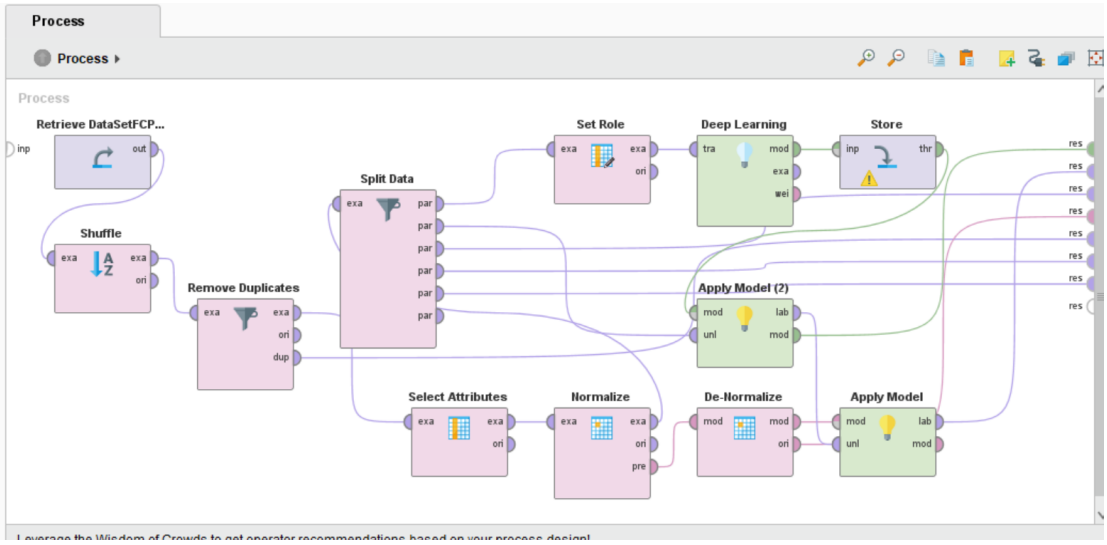


Nota: Proceso de diseño con operadores para el modelo Naive Bayes. Fuente (elaboración propia, 2022).

Para mejorar el modelo, el modelo Naive Bayes utilizó una distribución optimista de Poisson. El diseño del modelo se adhiere al enfoque descrito.

Figura 31

Diseño de modelo Deep Learning.



Nota: Proceso de diseño con operadores para el modelo Deep Learning. Fuente (elaboración propia, 2022).

Con una función de activación de Tanh, tres capas ocultas de 30 neuronas cada una, una rho de 0.999, y una función de distribución de Poisson, el modelo Deep Learning es optimizado. La técnica es seguida por el modelo.

Fase 5. Assessment (Evaluación)

Esta fase prueba cada modelo y las variables relacionadas para determinar qué tan resiliente es cada uno de los modelos elegidos. Para aumentar la precisión de cada modelo que se está utilizando, los modelos también deben mejorarse cambiando sus propios parámetros. Se determinará si los objetivos sugeridos y las hipótesis de investigación se llevarán a cabo mediante la validación de las predicciones realizadas. Los modelos utilizados para el desafío de regresión son los enumerados en la Tabla 13.

Deep Learning (aprendizaje profundo): “se basa en una red neuronal artificial de retroalimentación de múltiples capas que se enseña a través de descenso de gradiente estocástico a través de la propagación retro. La red puede tener un alto número de capas ocultas formadas por neuronas con funciones de activación” (RapidMiner, 2022).

Naive Bayes es un clasificador de baja variación y alta inclinación que puede desarrollar un modelo sólido incluso con un conjunto de datos limitado. Es fácil de usar y computacionalmente barato. La clasificación de texto, incluida la detección de spam, el análisis de sentimientos y los algoritmos de recomposición, es un caso de aplicación común.

La premisa clave de Naive Bayes es que el valor de cualquier atributo es independiente del valor de cualquier otro atributo, dado el valor de la etiqueta (la clase). Esta suposición rara vez es precisa (es "ingenua"), pero la experiencia demuestra que el clasificador Naive Bayes con frecuencia actúa de forma eficaz.

El supuesto de independencia simplifica considerablemente los cálculos necesarios para desarrollar el modelo de probabilidad Naive Bayes.

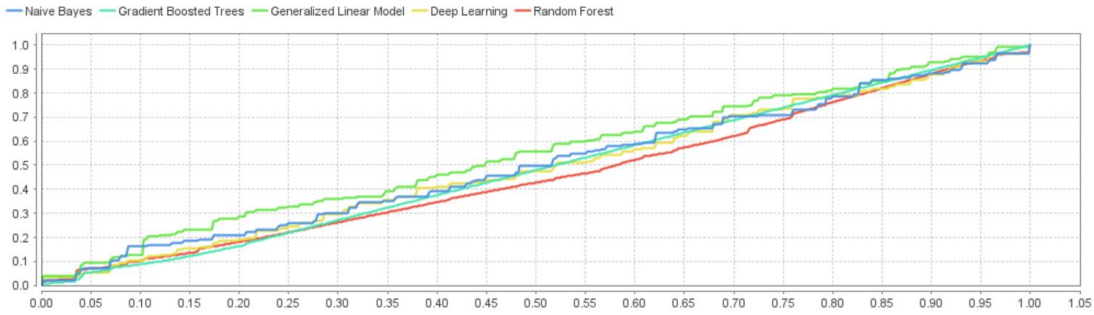
Para completar el modelo de probabilidad, se deben realizar varias suposiciones con respecto a las distribuciones de probabilidad condicional para las características individuales, dada la clase. Para modelar los datos de atributo, este operador emplea la densidad de probabilidad de Gauss (RapidMiner, 2022).

Gradient Boosted Trees: es una colección de modelos de árbol de regresión o clasificación. Ambos son enfoques de conjunto de aprendizaje hacia adelante que proporcionan resultados de predicción mejorando constantemente las estimaciones. El impulso es un enfoque de regresión no lineal versátil que ayuda a mejorar la precisión de los árboles. Una serie de árboles de decisión se forman mediante la aplicación secuencial de algoritmos de clasificación débiles a datos cada vez más alterados, lo que resulta en un conjunto de modelos de predicción débiles. Al tiempo que mejora la precisión de los árboles, reduce su velocidad y su interpretación humana. Para hacer frente a estos retos, el enfoque de aumento de gradiente generaliza la reserva de árboles (RapidMiner, 2022).

Estos tres modelos fueron sometidos a un proceso de pruebas, lo que permitió que los factores seleccionados decidieran un resultado excelente para satisfacer el objetivo previsto. A continuación, se da uno de los modelos probados con una especie seleccionada al azar.

Figura 32

Comparación entre los modelos en razón de ajuste-efectividad



Nota: La efectividad en los modelos seleccionados, se logran en función de un ajuste mayor al valor real. fuente (elaboración propia, 2022).

La eficacia de los modelos se determina por su modificación; cuanto mejor se ajustan los modelos a los datos de prueba, mejor se consiguen los resultados en un entorno real. Una vista superior le permite examinar la modificación en mayor profundidad.

En el aprendizaje supervisado, el error cuadrático medio MSE, la raíz del error cuadrático medio RMSE, y el R-cuadrado se emplean como medidas de evaluación.

Se puede utilizar la siguiente fórmula para calcular el MSE y la RMSE:

Figura 33

Cálculo de error cuadrático medio y raíz del error cuadrático medio

$$MSE = \frac{1}{|D|}\sum_{d\in D}(f(d) - h(d))^2 \qquad RMSE = \sqrt{MSE}$$

Nota 2 Fórmula de cálculo del error CUADRÁTICO (Gironés et al., 2018).

Tabla 16

Comparación error cuadrático

Criterio	GBT	NB	DL

Error cuadrático medio	84.70	865.39	628.02
Raíz del error cuadrático medio	9.25	29.45	25.07
R-cuadrado	0.97	0.73	0.81
Error absoluto medio	7.30	20.91	15.91

Nota 3 error cuadrático y R-cuadraos en modelos seleccionados. (Elaboración propia, 2022).

Los modelos candidatos son excelentes; sin embargo, los resultados de Gradiente Potenciado Árbol y Deep Learning son extremadamente cercanos, colocando el modelo lineal detrás de ambos modelos. Se elige para tener el menor error de la raíz media de la raíz (RMSE) y el R-cuadrado más alto. siendo el modelo GBT en esta situación, con una precisión de (100-7.20, el valor absoluto de error), es decir, 92.72 por ciento, como el candidato a trabajar con los objetivos establecidos en este estudio.

Utilización de conocimiento descubierto

Las salidas de los modelos de minería de datos se analizan para ver si el conocimiento encontrado es único e interesante, y los hallazgos de los modelos se interpretan en relación con la experiencia de los expertos de dominio.

Esta etapa comprende la interpretación de los clientes, el control cruzado de los clientes, la observación del interés y la relación de la información obtenida, así como el análisis del proceso y las diferentes maneras de obtener el conocimiento (Kumar et al., 2019). Con base en la revisión, se puede desarrollar otra etapa de minería de datos híbridos para analizar los hallazgos revelados. Entender los resultados, determinar si la información descubierta es original e interesante, interpretar los resultados por expertos de dominio, y evaluar el efecto de los conocimientos descubiertos son parte del proceso de evaluación.

Desarrollo de un tipo de prototipo

La interfaz de usuario sirve como un canal de comunicación entre el sistema y el usuario final. Como resultado, para desarrollar en el Fondo Complementario Previsional Cerrado de Cesantía de La

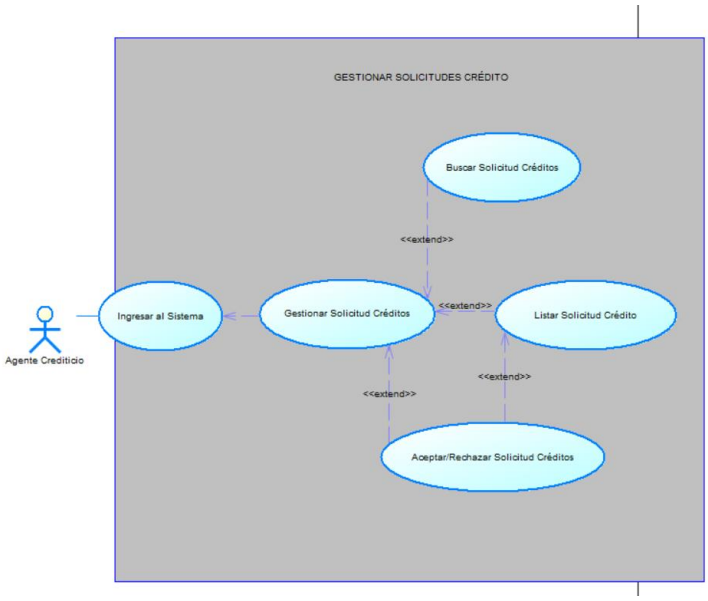
Universidad de las Fuerzas Armadas ESPE, como una herramienta interactiva, la elección fue tomada simplemente refiriéndose a la colección de reglas o comandos. Los ejemplos de información que se deben visualizar incluyen las sanciones de romper el conjunto de reglas.

La recopilación de reglas se obtiene fácilmente al atravesar la salida del árbol de decisiones. El investigador extrajo reglas que fueron pensadas para ser ambiguas, relevantes y originales para los experimentos de dominio, y luego compartió y discutió los hallazgos con oficiales de cartera de préstamos y especialistas en dominio del departamento de crédito.

Los sucesos que se han producido en la parte posterior de la pantalla, así como una explicación de las actividades del sistema. La importancia del componente de interfaz de usuario se deriva del hecho de que los usuarios finales a menudo evalúan en el Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, principalmente en la calidad de la interfaz de usuario en lugar de en el propio sistema. El usuario entra en los datos de evaluación del riesgo y lo triangula con un conjunto de criterios, que incluyen características obligatorias y opcionales.

Figura 34

Caso de Uso de gestionar solicitud crédito agente



Nota: Caso de uso para el diseño de interfaz de formulario para predicción de riesgo, Fuente (elaboración propia, 2022)

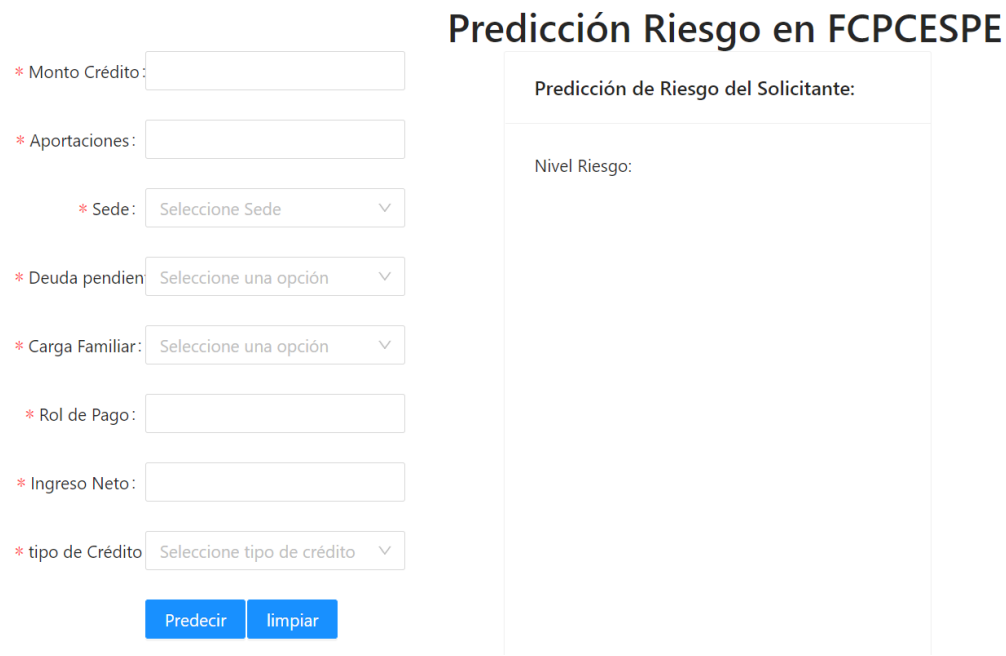
Si el usuario no especifica los datos necesarios en el sistema, la interfaz muestra un mensaje de error y el sistema falla.

Existen dos posibilidades de satisfacer el atributo necesario: Alto Riesgo o Riesgo Bajo. Residencia de la interfaz La pantalla muestra todas las calidades con un recuadro de colocación para la selección alternativa, el importe de la solicitud y la cantidad de depósito.

El formulario indica que son necesarios todos los campos y se disponen de dos opciones de elección: predicción y restablecimiento.

Figura 35

Formulario de predicción vacío.



Nota: Interfaz gráfica del prototipo de formulario de predicción del Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE, Fuente (elaboración propia, 2022)

La salida de interfaz muestra la segunda dimensión de la interfaz. Cuando se espera que el crédito sea altamente riesgoso, la pantalla que muestra la salida muestra los hallazgos de la predicción como "Alto Riesgo". El resultado final se muestra en la siguiente imagen.

Figura 36

Formulario en caso de riesgo alto

Nota: Interfaz gráfica del prototipo de formulario de predicción del fondo complementario previsional cerrado de cesantía de la universidad de las fuerzas armadas Espe, Fuente (elaboración propia, 2022)

La segunda dimensión de la salida de interfaz que muestra la interfaz. Cuando se prevé que el crédito es de bajo riesgo, la pantalla que muestra la salida contiene los resultados de la predicción. El resultado final se muestra en la siguiente imagen.

Figura 37

Formulario en caso de riesgo bajo

Nota: Interfaz gráfica del prototipo de formulario de predicción del fondo complementario previsional cerrado de cesantía de la universidad de las fuerzas armadas Espe, Fuente (elaboración propia, 2022)

Pruebas de aceptabilidad del usuario

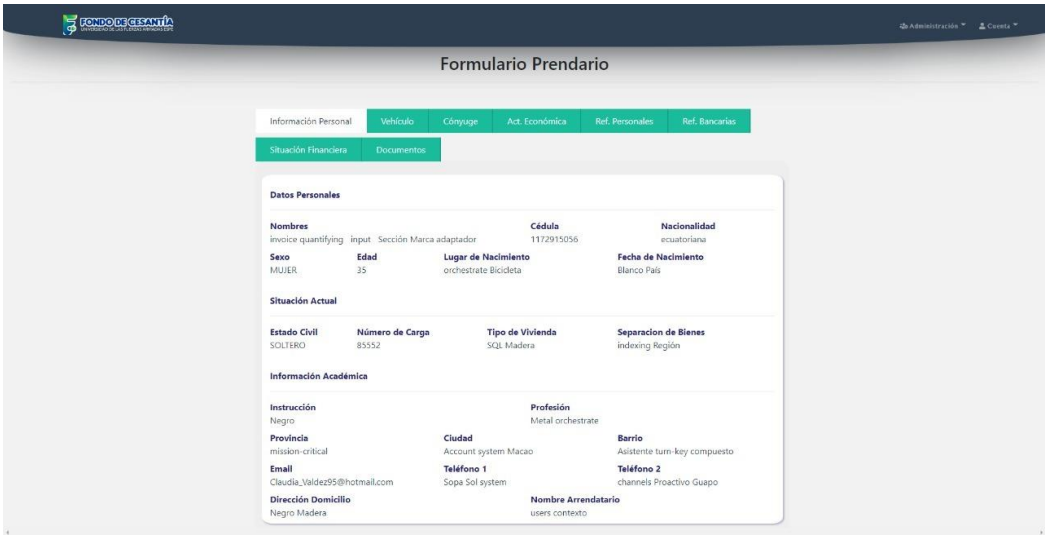
Una vez que se ha completado el proceso de desarrollo del sistema, la etapa siguiente es probar y evaluar el sistema para ver si satisface las necesidades de los usuarios y evaluar el rendimiento del sistema. La complejidad del sistema y otros elementos clave determinan el alcance de las pruebas y evaluaciones completadas, así como la pertinencia de la misma.

Porque el objetivo de la prueba y evaluación del sistema es asegurar que el sistema haga lo que se supone que debe lograr. Los agentes de créditos fueron elegidos en el Fondo Complementario Previsional Cerrado de Cesantía de La Universidad de las Fuerzas Armadas ESPE. Sin embargo, los participantes fueron informados sobre el flujo del sistema y las características del sistema.

El formulario actúa como ayuda en el proceso que mantienen en sus sistemas de gestión de créditos de donde obtienen la data para poder ingresarlos en el formulario y obtener la predicción, por ejemplo, a continuación, se presenta una imagen con el formulario Prendario.

Figura 38

Formulario prendario FCPCESPE



Nota: Formulario del Fondo Complementario Previsional Cerrado de Cesantía de la universidad de las Fuerzas Armadas ESPE. Fuente (elaboración propia, 2022)

La prueba de usabilidad del sistema se ha diseñado para medir la usabilidad del formulario del modelo para predicción de riesgo La usabilidad se utiliza con frecuencia para referirse a la forma en que las personas pueden utilizar las características del sistema. Esto tampoco es una característica de interfaz de usuario unidimensional. Tiene cinco características: capacidad de aprendizaje, eficiencia, numerabilidad, errores (tasa de error comprometida) y satisfacción. Diez participantes fueron testeados por 5 preguntas sobre cada participante, y el resumen fue proporcionado matemáticamente de la siguiente manera: número de preguntas tiempo 10 participantes divididos por 50.

El resultado de las pruebas de usabilidad se muestra como sigue. Los valores para todas las herramientas de medición escala Likert en la tabla son fijos como Totalmente de acuerdo = 5, De acuerdo = 4, Neutral = 3, Desacuerdo = 2 y fuertemente Desacuerdo = 1 basado en la evaluación demostrada cinco preguntas para las diez personas seleccionadas 76% responden firmemente de acuerdo, 16% de acuerdo sólo 8% de los usuarios son vacilantes de usar el software.

Tabla 17

Prueba de aceptación de usuario.

Nº	Pregunta de Evaluación	Fuertemente Desacuerdo	Desacuerdo	Neutral	De acuerdo	Totalmente de acuerdo	Promedio
1	Usaría este formulario con frecuencia				25%	75%	100%
2	Ha encontrado el sistema no complejo					100%	100%
3	el sistema es fácil de usar			5%	15%	75%	95%
4	No es necesario apoyo de una persona técnica para poder usar este sistema			10%	10%	70%	90%
5	El sistema no tiene inconsistencia			25%	25%	30%	80%
6	Total			8%	15%	70%	93%

Nota: Resultados de la prueba de aceptación del usuario. Fuente (elaboración propia, 2022)

Capítulo V

Conclusiones y recomendaciones

Conclusiones

- Muchas de las organizaciones como lo es el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas que tienen presencia en las diferentes sedes de la universidad buscando incentivar a los partícipes a solicitar más créditos, permitiendo de esta manera que sus bases de datos crezcan y tengan un mayor crecimiento para el negocio, lo cual permite aumentar el volumen de datos, volumen que al ser tratado y explorado con técnicas de minerías de datos se convierte en fuentes importantes donde se puede descubrir conocimientos ocultos.
- La predicción del riesgo de crédito a través de reglas que se formularon mediante los modelos creados a partir de la minería de datos, permitieron ser un soporte en el proceso de concesión de créditos, siendo inclusive a futuro un mecanismo que brinde seguridad y confianza para los partícipes, de esta manera incentivando a que soliciten más créditos reduciendo también el riesgo de créditos con problemas.
- Las correlaciones políticas, económicas, sociales y tecnológicas en el sector financiero obligan a los acreedores a utilizar una cantidad sustancial de elementos subjetivos en la identificación de clientes libres de riesgo, ya que resulta difícil de expresar a través de reglas deterministas.
- Esta investigación ha evaluado la aplicación de Minería de Datos mediante Técnicas y Algoritmos en la información crediticia del Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas, para predecir el patrón de créditos de alto riesgo y bajo riesgo mediante el desarrollo de un modelo de clasificación utilizando la herramienta Rapidminer.

Recomendaciones

- A pesar de que los resultados de este estudio fueron alentadores, se recomienda conforme aumente el volumen de datos, realizar más entrenamiento en los modelos y probar otras técnicas de clasificación como la red neuronal y las redes bayesianas (o combinaciones de cualquiera de las técnicas), deberían también emprenderse incluyendo datos antes de la implementación del sistema para tener la imagen completa de la historia del Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas.
- A partir del experimento realizado en esta investigación y trabajos previos, las técnicas de minería de datos podrían contribuir mucho en la identificación de clientes potenciales que podrían ser de alto riesgo, por lo que podría ser más importante utilizar la técnica de minería de datos como una herramienta para el proceso de toma de decisiones en otras palabras el Fondo Complementario Previsional Cerrado de Cesantía de la Universidad de las Fuerzas Armadas podría optimizar su esfuerzo de evaluación de crédito mediante el empleo de tecnología de minería de datos
- Es necesario desarrollar un prototipo de evaluación de riesgo de crédito o sistema de base de conocimiento para la implementación práctica del presente trabajo de investigación académica.

Referencias

Arcand, J.-L., & McDonald, S. (2018). Credit markets with imperfect information: Risk-aversion versus pessimism. Economics Letters, 165, 35-38. https://doi.org/10.1016/j.econlet.2018.01.029

Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017). Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI), 1-9. https://doi.org/10.1109/ICCNI.2017.8123782

Bartosova, V., 2008. Financial analysis and planning, Žilina: EDIS Publishers, University ofZilina, pp. 82.

Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S.Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127–31

Chen, H., Maslar, D. A., & Serfling, M. (2020). Asset redeployability and the choice between bank debt and public debt. Journal of Corporate Finance, 64, 101678. https://doi.org/10.1016/j.jcorpfin.2020.101678

Cole, R. A. (2011). Bank Credit, Trade Credit or No Credit: Evidence from the Surveys of Small Business Finances. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.1540221

Dhankhad, S., Mohammed, E., & Far, B. (2018). Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study. 2018 IEEE International Conference on Information Reuse and Integration (IRI), 122-125. https://doi.org/10.1109/IRI.2018.00025´

Ergungor, O. E. (2001). Theories of bank loan commitments. Economic Review, 37(3), 2-19.

Fondo ESPE. (s. f.). Recuperado 15 de septiembre de 2021, de http://www.fcpcespe.com.ec/

Forough, J., & Momtazi, S. (2021). Ensemble of deep sequential models for credit card fraud detection. Applied Soft Computing, 99, 106883. https://doi.org/10.1016/j.asoc.2020.106883

Fosu, S., Danso, A., Agyei-Boapeah, H., Ntim, C. G., & Adegbite, E. (2020). Credit information sharing and loan default in developing countries: The moderating effect of banking market concentration and national governance quality. Review of Quantitative Finance and Accounting, 55(1), 55-103. https://doi.org/10.1007/s11156-019-00836-1

Get SQL Anywhere—Manage, Synchronize and Exchange Data. (s. f.). Recuperado 15 de septiembre de 2021, de https://www.sqlanywhere.info/EN/

Ghorbani, A., & Farzai, S. (2018). Fraud Detection in Automobile Insurance using a Data Mining Based Approach. 8, 8.

Gololo, I. A. (2017). An Evaluation of the Role of Commercial Banks in Financing Small and Medium Scale Enterprises (SMEs): Evidence from Nigeria. Indian Journal of Finance and Banking, 1(1), 16-32. https://doi.org/10.46281/ijfb.v1i1.82

Goodman S. A dirty dozen: twelve P-value misconceptions. Semin Hematol.2008;45(3):135–40.

Gupta, M. K., & Chandra, P. (2020). A comprehensive survey of data mining. International Journal of Information Technology, 12(4), 1243-1257. https://doi.org/10.1007/s41870-020-00427-7

Hassan, M. K., El Desouky, A. I., Elghamrawy, S. M., & Sarhan, A. M. (2018). Intelligent hybrid remote patient-monitoring model with cloud-based framework for knowledge discovery. Computers & Electrical Engineering, 70, 1034-1048. https://doi.org/10.1016/j.compeleceng.2018.02.032

Hooi et al. - 2016—BIRDNEST Bayesian Inference for Ratings-Fraud Det.pdf. (s. f.). Recuperado 15 de septiembre de 2021, de https://epubs.siam.org/doi/pdf/10.1137/1.9781611974348.56

Instituto Federal Goiano Campus Ceres, GO, Brazil, Sousa, M. de M., Figueiredo, R. S., & Federal University of Goiás, GO, Brazil. (2014). CREDIT ANALYSIS USING DATA MINING: APPLICATION IN THE CASE OF A CREDIT UNION. Journal of Information Systems and Technology Management, 11(2), 379-396. https://doi.org/10.4301/S1807-17752014000200009

Jafar Hamid, A., & Ahmed, T. M. (2016). Developing Prediction Model of Loan Risk in Banks Using Data Mining. Machine Learning and Applications: An International Journal, 3(1), 1-9. https://doi.org/10.5121/mlaij.2016.3101

Kulkarni, P., & Ade, R. (2016). Logistic Regression Learning Model for Handling Concept Drift with Unbalanced Data in Credit Card Fraud Detection System. En S. C. Satapathy, K. S. Raju, J. K. Mandal, & V. Bhateja (Eds.), Proceedings of the Second International Conference on Computer and Communication Technologies (Vol. 380, pp. 681-689). Springer India. https://doi.org/10.1007/978-81-322-2523-2_66

Kumar, N., Jain, S., & Chauhan, K. (2019). Knowledge Discovery from Data Mining Techniques. International Journal of Engineering Research & Technology (IJERT), 7(12), 1-3.

Minnis, M., & Sutherland, A. (2017). Financial Statements as Monitoring Mechanisms: Evidence from Small Commercial Loans: FINANCIAL STATEMENTS AS MONITORING MECHANISMS. Journal of Accounting Research, 55(1), 197-233. https://doi.org/10.1111/1475-679X.12127

Pazmiño-Maji, R. A., García-Peñalvo, F. J., & Conde-González, M. A. (s. f.). Statistical Implicative Analysis Approximation to KDD and Data Mining: 8.

P. C. et. al. Crisp-dm 1.0 - step-by-step data mining guide. https://www.the-modeling-agency.com/crisp-dm.pdf.

Riquelme Santos, J.C., Ruíz, R. y Gilbert, K. (2006). Minería de Datos: Conceptos y Tendencias. Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18.

Sadgali, I., Sael, N., & Benabbou, F. (2019). Performance of machine learning techniques in the detection of financial frauds. Procedia Computer Science, 148, 45-54. https://doi.org/10.1016/j.procs.2019.01.007

Sadikin, M., & Alfiandi, F. (2018). Comparative Study of Classification Method on Customer Candidate Data to Predict its Potential Risk. International Journal of Electrical and Computer Engineering (IJECE), 8(6), 4763. https://doi.org/10.11591/ijece.v8i6.pp4763-4771

S. institute. Enterprise miner - semma. https://bit.ly/2JLIb3z.

S. H. B. S. R. Bulkley, J. Gayle. Adding the where to the who. In 24th SUGI - SAS Users Group International conference conference

Viaene, S., Derrig, R. A., & Dedene, G. (2004). A case study of applying boosting naive bayes to claim fraud diagnosis. IEEE Transactions on Knowledge and Data Engineering, 16(5), 612-620. https://doi.org/10.1109/TKDE.2004.1277822

Weka 3—Data Mining with Open Source Machine Learning Software in Java. (s. f.). Recuperado 15 de septiembre de 2021, de https://www.cs.waikato.ac.nz/ml/weka/index.html

Zaki,M.J., & Wong, L.(2018), Data Mining Techniques, John Wiley & Sons Ltd, Chichester , England

(S/f). Diva-portal.org. Recuperado el 1 de septiembre de 2021, de https://www.diva-portal.org/smash/get/diva2:1250897/FULLTEXT01.pdf

Arnold, K., Gosling, J., & Holmes, D. (2005). The Java programming language. Addison Wesley Professional.

Wohlgethan, E. (2018). SupportingWeb Development Decisions by Comparing Three Major JavaScript Frameworks: Angular, React and Vue. js (Doctoral dissertation, Hochschule für Angewandte Wissenschaften Hamburg).

Hickson, I., & Hyatt, D. (2011). Html5. W3C Working Draft WD-html5-20110525, 53.

Bierman, G., Abadi, M., & Torgersen, M. (2014, July). Understanding typescript. In European Conference on Object-Oriented Programming (pp. 257-281). Springer, Berlin, Heidelberg.

Hamid, A., & Ahmed, T. (2016). Developing prediction model of loan risk in banks using data mining. Machine Learning and Applications: An International Journal.

Patel, B., Patil, H., Hembram, J., & Jaswal, S. (2020). Loan default forecasting using data mining.

Samsir, S., Suparno, S., & Giatman, M. (2020). Predicting the loan risk towards new customer applying data mining using nearest neighbor algorithm.

Windarto, A., & Wanto, A. (2018). ata mining tools| rapidminer: K-means method on clustering of rice crops by province as efforts to stabilize food crops in Indonesia. Indonesia.