UNIVERSIDAD INTERNACIONAL SEK

FACULTAD DE ARQUITECTURA E INGENIERÍAS

Trabajo de fin de carrera titulado:

"Modelo de Clasificación de Riesgo Crediticio Utilizando Random Forest en financiera del Ecuador"

Realizado por:

Juan Freire López

Director del proyecto:

Ing. César Byron Guevara Maldonado, PhD.

Como requisito para la obtención del título de MASTER EN SISTEMAS DE INFORMACIÓN CON MENCION EN DATA SCIENCE

Quito, Agosto 2021

DECLARACION JURAMENTADA

Yo, **Juan Freire López,** con cédula de identidad 1720479045, declaro bajo juramento que el trabajo aquí desarrollado es de mi autoría, que no ha sido previamente presentado para ningún grado a calificación profesional; y, que ha consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración, cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la UNIVERSIDAD INTERNACIONAL SEK, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normativa institucional vigente.

Juan Freire López

C.C: 1720479045

DECLARATORIA

El presente trabajo de investigación titulado:

"MODELO DE CLASIFICACIÓN DE RIESGO CREDITICIO UTILIZANDO RANDOM FOREST EN FINANCIADORA EN EL ECUADOR"

Realizado por:

Juan Freire López

Como requisito para la Obtención del Título de:

MASTER EN TECNOLOGÍAS DE LA INFORMACIÓN CON MENCION EN DATA SCIENCE

Ha sido dirigido por el profesor

Ing. César Byron Guevara Maldonado, PhD.

Quien considera que constituye un trabajo original de su autor

CÉSAR BYRON GUEVARA MALDONADO, PhD.

DIRECTOR

PROFESORES INFORMANTES

Después de revisar el trabajo presentado, lo ha calificado como apto para su defensa oral ante el tribunal examinador.

Ing. DIEGO RIOFRIO LUZCANO

Ing. JOE CARRION JUMBO

Quito, Agosto de 2021

DEDICATORIA

Dedico todo este trabajo a mi familia que me apoyo durante esta etapa de mi vida, pero principalmente a mi madre y mi hermano que estuvieron conmigo brindándome su ayuda y ánimo durante este periodo.

También menciono a mi padre, aunque partió de este mundo hace mucho tiempo su ejemplo ha dejado huella en mí y ha sido una motivación enorme para seguir superándome.

AGRADECIMIENTO

A mi madre por darme su apoyo desde que decidí tomar este gran desafío.

A mi hermano por darme sus consejos y ayuda cuando lo necesité

A mi novia Ana Lucía Andrade por brindarme su ayuda, especialmente al momento de revisar la redacción de este documento.

A mi tutor de tesis, Cesar Guevara por su gran ayuda durante todo este tiempo, desde la creación del artículo, hasta la finalización de este documento.

A Diego Riofrio que me ayudo cuando requerí de su apoyo.

ÍNDICE GENERAL DE CONTENIDOS

ÍNDICE GENERAL DE TABLAS

ÍNDICE GENERAL DE FIGURAS

RESUMEN

Esta investigación, se enfoca en Insotec, organización que se encuentra potenciando los procedimientos de sus áreas, incluyendo la evaluación crediticia que actualmente se realiza manualmente, este procedimiento analiza caso por caso y se aprueban las solicitudes dependiendo de la calificación, esto genera muchos inconvenientes como alta carga operativa y probabilidad de cometer errores en las evaluaciones.

Para mejorar el proceso de evaluación crediticia de los clientes, el estudio propone crear un modelo basado en un algoritmo de inteligencia artificial que clasifique a los clientes de la organización como "buenos" o "malos" pagadores, en función de las diferentes variables consideradas importantes para el análisis.

En la creación de modelo de clasificación se utiliza la base de datos de clientes de crédito de la organización con un histórico desde el año 2017 al 2020, este conjunto de datos cuenta con 18 variables y 63.896 registros. Adicionalmente, para la implementación del modelo de clasificación se utiliza la metodología CRISP-DM.

Posteriormente, se prepara la información con el preprocesamiento de datos, en este paso se utiliza la técnica de eliminación de datos atípicos, con lo cual el conjunto de datos se reduce de 63.896 a 58.247. Finalmente, se seleccionan las variables con mayor importancia con el método de chi cuadrado, en este caso son 8 variables seleccionadas.

El modelo es implementado con *Random Forest*, el cual arroja una precisión mayor al 97%, el porcentaje de error es del 2,8%, con el 2,1% falsos positivos y 11,1% falsos negativos para predecir.

Finalmente, la creación del modelo ayuda a contar con una herramienta adicional que sirve para clasificar automáticamente a los clientes como "buenos" y "malos" pagadores, lo que puede ser utilizado para entregar créditos con más rapidez y con un menor grado de riesgo. Sin embargo, este modelo necesita ser desplegado como lo indica la metodología CRISP-DM, para que el conocimiento obtenido sea aprovechado por el cliente.

Palabras Claves: Random Forest, arboles de decisión, chi cuadrado, clasificación de clientes de crédito.

ABSTRACT

This research focuses on Insotec, a medium-sized entity that dedicates to granting microcredits. The organization is looking forward to enhancing different areas procedures, including the credit evaluation, which is carried out manually. Currently, analysts classify risk on a case-by-case basis, and applications are approved depending on the credit rating. This process generates many drawbacks, such as high operational load and the probability of making wrong evaluations.

This study proposes to enhance the risk evaluation process by creating a model based on an artificial intelligence algorithm that classifies the organization's customers as good or bad payers based on the different variables considered essential for the analysis. Furthermore, the study uses the organization's credit customer database with a historic from 2017 to 2020.

This dissertation proposes creating an algorithmic model using the CRISP-DM methodology to classify the entity's clients. First, the business and the data set are understood.

Second, the information is prepared within the data preprocessing phase. This step applies the outlier elimination technique. In this phase, the dataset reduces from 63,896 to 58,247 records. Afterward, the chi-square technique selects the variables with the most significant importance method, which sets eight variables for the analysis.

When the dataset is ready, the CRISP-DM methodology suggests selecting the techniques most suitable for the model. In this case, the Random Forest algorithm obtains the best precision results compared against other algorithms such as neural networks and decision trees. The Random Forest model throws a precision of 97%, an error of 2.8%, and a rate of 2.1% false positives and 11.1% false negatives to predict. Additionally, the analysis implements ten models with the cross-validation method for the model evaluation phase.

Finally, creating the model leads to having an additional tool, which automates the clients' classification process as good and bad payers. This automation can be used to deliver loans faster and with less risk. However, the deployment phase needs to be executed as mentions de CRISP-DM methodology states, so that the client can make the most of the knowledge obtained by the model.

CAPÍTULO I

INTRODUCCIÓN

En las últimas décadas la economía de los diferentes sectores productivos se ha desarrollado gracias a la expansión de los servicios financieros. Los préstamos brindados por estas entidades dan la oportunidad a los empresarios de obtener ingresos para impulsar los diferentes negocios, aplazando los pagos para los siguientes meses (Lanzarini et al., 2017).

Según el (Banco Mundial, 2021) los préstamos ayudan a impulsar la economía de los diferentes sectores productivos. En el indicador de créditos otorgados al sector privado, se indica que en el 2019 han representado el 132% del producto interno bruto a nivel mundial, en Latinoamérica y el Caribe el 55,6%, mientras que en Ecuador el 42,7%.

Los prestamos indudablemente sirven para potenciar la economía de los diferentes sectores. Sin embargo, en caso de no ser manejados correctamente pueden crear inconvenientes como sobreendeudamiento de los clientes y problemas en cumplir con los pagos, lo que generaría pérdidas en las entidades financieras (Sobarsyah et al., 2020) (Pandey et al., 2017).

Para evitar estos problemas, los préstamos deben ser manejados estratégicamente desde su concepción, por lo que se requiere contar con métodos que ayuden a calificar el riesgo de los clientes correctamente, por este motivo las entidades financieras se apoyan en clasificadores crediticios con modelos estadísticos para medir el riesgo basados en la información registrada. Adicionalmente, se puede reducir el riesgo de la cartera de préstamos mejorando estos procedimientos (Shi et al., 2019).

Los clasificadores de crédito que utilizan procesos estadísticos fueron introducidos en los años cincuenta, actualmente, son utilizados mundialmente y se han convertido en una parte fundamental del proceso de evaluación de crédito. El principal objetivo es analizar el comportamiento de pago de los prestamistas con el fin de reducir el porcentaje de error (Zhang, & Zhang, 2019). Adicionalmente, se puede automatizar los procesos de evaluación de riesgo con algoritmos de inteligencia artificial, estas implementaciones pueden generar beneficios como mejora de la eficiencia en el proceso de solicitud de créditos, reducción de costos e incremento de competitividad. Con el objetivo de obtener

los beneficios mencionados, Bequé and Lessmann (2017) propone la implementación de modelos con algoritmos de inteligencia artificial en el proceso de clasificación de clientes.

Por lo expuesto anteriormente, este estudio utiliza la información de Insotec, el cual pretende brindar una herramienta que facilite el proceso de clasificación crediticia. Existen investigaciones en el área que implementan modelos de clasificación crediticia con algoritmos que utilizan redes neuronales, bayes naive, máquinas de vectores de soporte y árboles de decisión. Sin embargo, en este estudio se desarrolla un modelo con Random Forest basado en las recomendaciones literarias realizadas por Siswanto et al. (2019), Pradhan, Akter, and Al Marouf (2020), Arora and Kaur (2019) y Subasi and Cankurt (2019).

Finalmente, el objetivo de esta investigación es crear un modelo de clasificación de clientes con un alto nivel de precisión, para lo cual se utilizan técnicas de preprocesamiento de datos. Para el modelo se implementa un modelo con *Random Forest* y se compara su resultado versus otros algoritmos como redes neuronales. Adicionalmente, esta investigación va a mejorar los procesos de evaluación del riesgo, la calificación crediticia de clientes y el porcentaje de error en la clasificación de los clientes de la organización.

El documento está estructurado de la siguiente manera: en la sección 1 se presenta el planteamiento del problema, objetivos generales y específicos, justificación y el alcance. En la sección 2, se presenta el marco teórico. En la sección 3, se detallan los trabajos relacionados. En la sección 4, los métodos y materiales utilizados en el estudio. Posteriormente, en la sección 5, se analizan los resultados. Finalmente, la sección 6, contiene las conclusiones y líneas futuras de investigación.

1.1. Planteamiento del Problema

Este estudio utiliza la información de Insotec, que es una organización mediana dedicada a la colocación de créditos en el Ecuador, esta entidad trabaja en el sector micro financiero, y tiene como finalidad potenciar la economía en los diferentes lugares del país. Adicionalmente, la mayoría de la cartera de la organización se encuentra dentro del segmento agropecuario (Insotec, 2021).

Misión "Contribuir al desarrollo de la microempresa mediante la concesión de créditos, brindando acceso a servicios complementarios que mejoran la competitividad y la calidad de vida, con énfasis en el sector agropecuario, bajo criterios de sostenibilidad, responsabilidad social y excelencia en la calidad del servicio." (Insotec, 2021).

Visión "Somos una institución especializada en microfinanzas líder a nivel nacional, por su calidad reconocida en impacto social y su contribución al desarrollo económico en sus territorios, con capacidad de expansión internacional." (Insotec, 2021).

Actualmente, muchas financieras similares a Insotec, han implementado modelos que utilizan técnicas de inteligencia artificial dentro de su proceso de evaluación del riesgo crediticio, esto ha mejorado los resultados de los análisis (Lee & Shin, 2020; Liebergen, 2017). Sin embargo, este procedimiento todavía se realiza de forma manual en Insotec, esto genera una carga operativa alta y los tiempos de respuesta muchas veces son ineficientes a las solicitudes de los clientes. Además de invertir demasiado tiempo en esta tarea, es posible colocar créditos erróneamente, los cuales generan problemas colaterales como incumplimiento en los pagos, incremento de la cartera vencida e incluso afectar a la economía del país si se evidencian estos inconvenientes en las diferentes financieras (Pandey et al., 2017).

Por la gran cantidad de solicitudes y el tiempo reducido para entregar una respuesta a los solicitantes, la probabilidad de cometer errores en la aprobación de créditos es alta. Además, la organización tiene como uno de sus principales objetivos el incremento constante de su cartera. Sin embargo, tanto tiempo invertido en la evaluación crediticia es uno de los principales obstáculos y esto podría generar pérdida de clientes.

La finalidad de este estudio es desarrollar un primer modelo que ayude a clasificar a los clientes actuales y nuevos de la organización, dependiendo de su comportamiento de

pago. Tomando en cuenta estas consideraciones se plantea la siguiente pregunta de investigación. ¿Se puede crear un modelo confiable que utilice inteligencia artificial para la clasificación crediticia en Insotec, que pueda ser utilizado para crear predicciones precisas del comportamiento de “buenos” y “malos” pagadores?

1.2. Objetivos de la Investigación

En esta sección se describen el objetivo general y los objetivos específicos que se han diseñado para resolver el problema planteado en la sección 1.2.

1.2.1. Objetivo General

Desarrollar un modelo de clasificación de los clientes de Insotec aplicando técnicas de aprendizaje de máquina, para disminuir el riesgo crediticio, mejorar los tiempos de respuesta a las solicitudes y la competitividad.

1.2.2. Objetivo Específicos

- Elaborar una revisión de literatura de creación de modelos de clasificación crediticia, utilizando el buscador Google Scholar de las editoriales IEEE, Springer, ACM, Elsevier y Nature de los cuartiles Q1 a Q4, para encontrar estudios similares realizados principalmente desde el 2017 y aplicar las técnicas recomendadas en las investigaciones.
- Seleccionar los métodos y técnicas para el estudio, con los artículos seleccionados, para obtener mejores resultados de precisión.
- Analizar la información del negocio y el conjunto de datos de clientes de la organización, utilizando herramientas y técnicas de análisis de datos, parar eliminar los registros y campos que pueden generar ruido en el análisis, de esta manera se prepara el conjunto de datos que va a ser ingresado al modelo.
- Proponer un modelo de clasificación de clientes, con las técnicas seleccionadas, que obtenga un porcentaje de precisión mayor al 90%.
- Verificar los resultados, utilizando la matriz de confusión del modelo, para conocer la confiabilidad del modelo para predecir a los “buenos” y “malos”.

1.3. Justificación

Las entidades financieras a nivel mundial se desenvuelven en un ambiente de alta competitividad, en el cual están obligadas a adoptar permanentemente medidas para mejorar su eficiencia y mitigar los riesgos (Thanh et al., 2020). En este sentido, se han desarrollado modelos de clasificación crediticia con métodos de aprendizaje automático, que buscan determinar los riesgos que constituyen el otorgamiento de créditos para así gestionarlos de forma proactiva (Liebergen, 2017).

Los modelos de clasificación crediticia que se implementan con algoritmos de inteligencia artificial pueden generar beneficios como aumentar la tasa de retención de clientes, mejorar la productividad, incrementar el margen de ganancias, disminuir el tiempo de respuestas a las solicitudes de los clientes y la posibilidad de ampliar la inclusión de modelos analíticos en otras áreas (Lee & Shin, 2020).

Adicionalmente, con la implementación de estos modelos se puede obtener resultados más precisos en grandes y complejos conjuntos de datos, lo que facilita la calificación constante del riesgo crediticio de cada uno de los clientes dentro de una entidad financiera. Las entidades bancarias que automatizan los diferentes procesos utilizando inteligencia artificial pueden reducir del 20 al 25% de sus costos (Lee & Shin, 2020).

Tomando en cuenta los beneficios mencionados tras la implementación de modelos de clasificación, muchas financieras han invertido tiempo y recursos para incluir estos modelos en sus entidades, además se han realizado varias investigaciones relacionadas a la implantación y evaluación de estos modelos crediticios dentro de instituciones financieras (Siswanto et al., 2019; Ahmad et al., 2017).

Dentro de las investigaciones en la creación de modelos de clasificación crediticia, se experimenta con diferentes tipos de algoritmos como redes neuronales, arboles de decisión, *Rando Forest*, etc (Ziemba et al., 2020; Siswanto et al., 2019; Ahmad et al., 2017).

Sin embargo, en las investigaciones realizadas por Pradhan et al. (2020), Arora and Kaur (2019), Ziemba et al. (2020) y Subasi and Cankurt (2019), se comparan modelos con varios algoritmos, en estos experimentos los modelos que fueron implementados con

Random Forest obtuvieron los mejores resultados de precisión. Los resultados de los estudios mencionados se presentan en la Tabla 1.

Tabla 1: Resultados Investigaciones similares

Investigación	**Resultados con *Random Forest***
Ziemba et al. (2020)	98.00%
Arora & Kaur 2019	97.90%
Pradhan et al. (2020)	85,00%
Subasi & Cankurt (2019)	89,01%

1.4. Alcance

Inicialmente, esta investigación abarca la revisión de la literatura de aprendizaje automático en el área de evaluación de riesgo crediticio, selección de métodos y técnicas de inteligencia artificial.

Posteriormente, se analiza la información de negocio y de los clientes de Insotec para realizar un correcto preprocesamiento de datos de acuerdo con la metodología CRISP-DM. Al conjunto resultado de la fase de preprocesamiento de datos se incluye dentro del modelo de clasificación con *Random Forest* y se lo compara con otras técnicas como redes neuronales, árboles de decisión y máquina de vectores de soporte.

Adicionalmente, se analizan los resultados del modelo de clasificación con *Random Forest* con la matriz de confusión y se evalúa la precisión del modelo la cual debe obtener un resultado mayor al 90% y con una tasa baja de falsos positivos. Adicionalmente, se realizarán experimentos aleatorios para evaluar la robustez del modelo.

Finalmente, el estudio únicamente analizará hasta la fase de evaluación del modelo, esto debido a que la metodología CRIS-DM incluye dentro de sus procesos el despliegue y compartición de la información generada con el cliente. Sin embargo, este paso puede tomar mucho tiempo hasta que el grado de confianza en el modelo de clasificación crediticia sea alto. Por ello, se incluye la fase de aceptación e inclusión del modelo en el proceso de clasificación crediticia dentro de las investigaciones futuras.

CAPÍTULO II

MARCO TEÓRICO

En esta sección se describe las teorías realizadas en previas investigaciones que aportan en la creación de modelo de clasificación crediticia para Insotec.

2.1. Machine Learning

El aprendizaje automático es un tipo de inteligencia artificial que se implementa con métodos computacionales, basados en la experiencia, capaces de detectar patrones y generar predicciones precisas (X.-D. Zhang, 2020).

Esta experiencia se refiere a la información pasada que es utilizada para aprender, dependiendo del tamaño y calidad de la información es posible obtener predicciones exitosas (Mohri et al. 2012).

Existen diferentes tipos de algoritmos de aprendizaje automático y son detallados en esta sección.

2.1.1. Aprendizaje supervisado

El aprendizaje supervisado es una técnica que utiliza algoritmos que trabajan con conjuntos de entrenamiento conformado por clases o etiquetas, en esta técnica los algoritmos aprenden en función del histórico y hacen predicciones de salida dependiendo de la clase (X.-D. Zhang, 2020). En la Figura 1 se muestran varios de los métodos algorítmicos que trabajan bajo el aprendizaje supervisado según (Jiang et al. 2020).

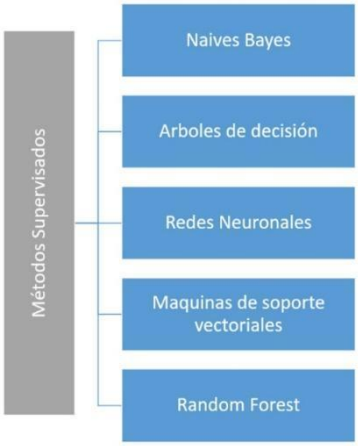

Figura 1: Métodos de aprendizaje Supervisados (Jiang et al., 2020)

2.1.2. Aprendizaje no supervisado

El aprendizaje no supervisado es una técnica donde los conjuntos de entrenamiento no cuentan con clases o etiquetas, por lo tanto, no existe ningún tipo de conocimiento anterior ni una salida predeterminada. En este caso, el conjunto de datos inicial debe ser segmentado por clusters (X.-D. Zhang, 2020). En la Figura 2 se muestran varios de los métodos algorítmicos que trabajan bajo el aprendizaje no supervisado mencionados por Nisioti et al. (2018) y Amruthnath and Gupta (2018).

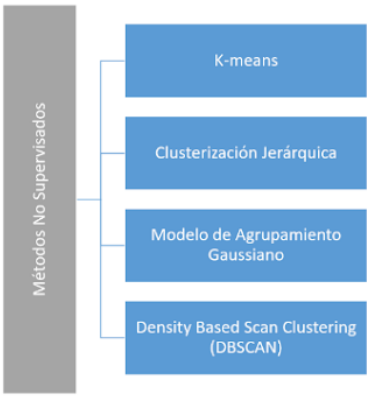

Figura 2: Métodos de aprendizaje no supervisados (Nisioti et al., 2018; Amruthnath & Gupta, 2018)

2.1.3. Aprendizaje por refuerzo

El aprendizaje por refuerzo es una técnica donde un agente aprende en un entorno interactivo mediante prueba y error, con el uso de retroalimentaciones de las acciones propias y experiencias ganadas en el proceso (X.-D. Zhang, 2020). En la Figura 3 se muestran varios de los métodos algorítmicos que trabajan bajo el aprendizaje por refuerzo mencionados por Otoum et al. (2019), Ruiz-Montiel et al. (2017) y Alfakih et al. (2020).

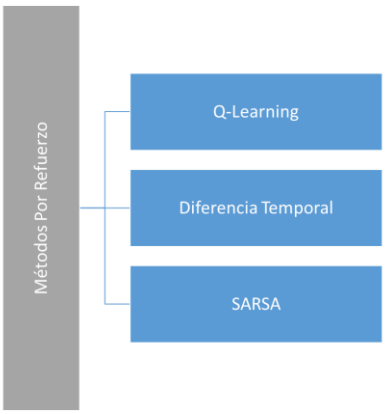

Figura 3: Métodos de aprendizaje por refuerzo (Ruiz-Montiel et al., 2017; Ruiz-Montiel et al., 2017; Ruiz-Montiel et al., 2017)

2.2. Algoritmo de Aprendizaje

Existen varios algoritmos de aprendizaje como redes neuronales, *bayes nayves*, arboles de decisión, máquina de vectores de soporte. Sin embargo, en esta investigación se detallará el método de aprendizaje supervisado de *Random Forest.*

Se detalla únicamente *Random Forest* porque en el punto 4.3.2 se muestra que su precisión es la mejor, además este algoritmo es recomendado en varias investigaciones como se referencia en el punto 1.3.

2.2.1. Arboles Aleatorios

Arboles Aleatorios (*Random Forest)* es un algoritmo de aprendizaje supervisado que surge de la combinación de árboles de decisión sin correlación, que promedia sus resultados. También, es conocido como un método ensamblado, que combina los resultados de los diferentes árboles para obtener un valor para todo el conjunto de árboles (Safari, 2020).

Este algoritmo utiliza la técnica *Bagging* que combina los resultados de diferentes clasificadores para poder crear un resultado único, el cual tendrá una varianza menor para la predicción, comparada con los resultados de los clasificadores individuales. En la Figura 4 se puede observar cómo se realizan varias clasificaciones que se promedian para obtener un resultado único (Wang et al., 2019).

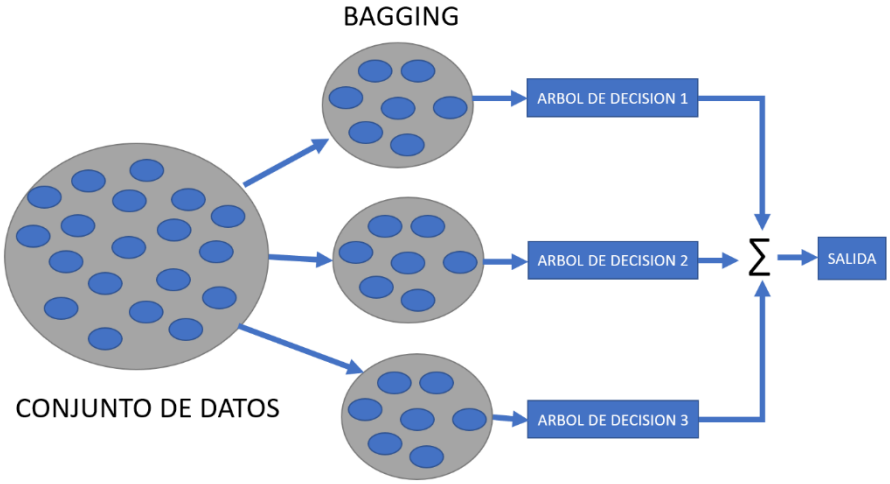

Figura 4: Proceso de Clasificación de Random Forest (Wang et al., 2019)

Esta técnica introducida por Breiman (2001) puede considerar un conjunto de datos x, donde x_i representa a cada una de las iteraciones que son formadas de una muestra aleatoria obtenida del conjunto original. T_b representa cada uno de los árboles y B el

número de árboles que se encuentran en el bosque de la regresión, esta regresión se representa en la Ecuación 1 (Uthayakumar et al. 2020).

$$\hat{y}(x_i) = \frac{1}{B}\sum_{b=1}^{B} \quad T_b(xi) \tag{1}$$

2.3. Machine learning aplicada a entidades bancarias

Después de la crisis financiera mundial, el manejo de riesgos en el sector bancario ha ganado importancia y estas entidades se han visto obligadas a mejorar la gestión de riesgos en los diferentes procesos como identificación, valoración y planificación de respuesta (Liebergen, 2017).

En todos los procesos de riesgo mencionados anteriormente, se requiere utilizar proactivamente la información del negocio para mejorar las estrategias dentro de las entidades bancarias (Liebergen, 2017) (Lee & Shin, 2020). Con el objetivo de mejorar estos procesos se ha incorporado algoritmos de inteligencia artificial, esto ha beneficiado el manejo de las diferentes áreas, principalmente en las relacionadas con estrategia de negocio, marketing, detección del fraude y riesgos crediticios (Liebergen, 2017).

De acuerdo con Lee and Shin (2020) las entidades bancarias que se apoyan de inteligencia artificial pueden reducir del 20 al 25% de sus costos, esto debido a que la automatización de procesos ayuda a disminuir la carga operativa, lo que contribuye a que los colaboradores utilicen su tiempo en tareas estratégicas que agregan mayor valor a las empresas, como, por ejemplo, crear mejores productos para los clientes.

También, los algoritmos de inteligencia artificial ayudan a detectar movimientos inusuales en las cuentas de los clientes, con este procedimiento se puede detectar con mayor facilidad si existe un posible fraude como el lavado de activos (Z. Chen et al., 2018).

Referente al área de riesgos, el uso de aprendizaje automático ha ganado trascendencia, especialmente en la construcción de modelos de riesgo crediticios porque sus análisis son más precisos y capaces de identificar complejos patrones no lineales en conjuntos de datos de gran volumen. Adicionalmente, con estudios más precisos se puede reducir la incertidumbre y el riesgo, además de eliminar la parcialidad de las decisiones humanas (Leo et al. 2019).

2.4. Proceso de aplicación de técnicas en modelo de Machine Learning

En la década de los noventas, con el objetivo de normalizar los procesos de aplicación de las técnicas de inteligencia artificial, se desarrolló la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) que es catalogada como una de las más completas y compone los procesos necesarios para la implementación de proyectos dentro del área de ciencia de datos (Martínez et al. 2019).

De acuerdo con Huber et al. (2019), esta metodología está compuesta por seis fases graficadas en la Figura 5 y son detalladas a continuación:

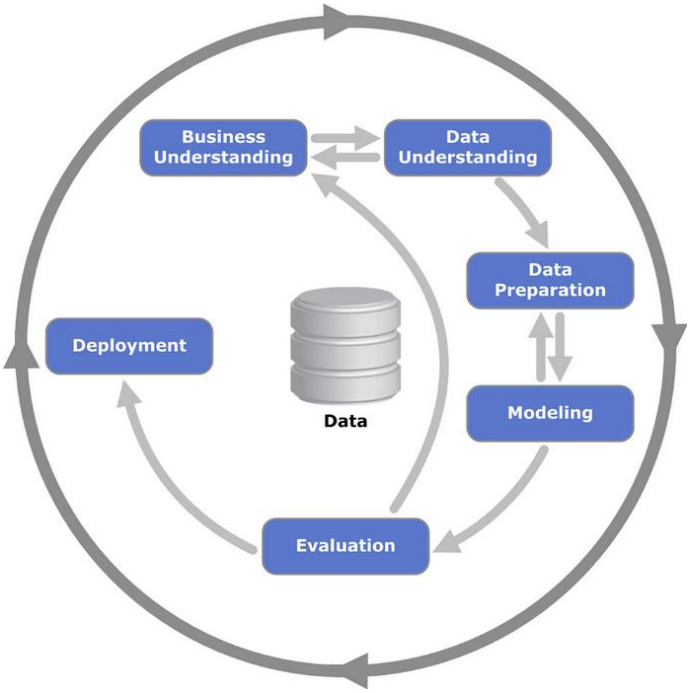

Figura 5: Proceso de la Metodología CRISP-DM (Huber et al., 2019).

Primero, se necesita comprender el funcionamiento del negocio, por lo tanto, el objetivo de esta etapa es conocer la situación actual para diseñar un plan de implementación del proyecto de ciencia de datos.

Segundo, se integrar los datos para comprenderlos, en esta etapa se empieza con la creación del conjunto inicial de datos y se continua con procesos que facilitan el entendimiento de la información, la identificación de los problemas en la calidad de los datos y el reconocimiento de los tipos de datos almacenados.

Tercero, Preparación de los datos, en esta etapa se realiza la selección, limpieza y transformación a los datos brutos, con la finalidad de eliminar el ruido que pueda distorsionar los resultados del estudio.

El objetivo de este procedimiento es crear un conjunto de datos que se puedan interpretar fácilmente para obtener mejores resultados en el análisis.

Cuarto, fase de modelado, en este paso se seleccionan las técnicas que se van aplicar en el modelo, se ajustan los parámetros para que los resultados sean los mejores. Algunas de las técnicas tienen restricciones en el formato de los datos, por lo que se podría regresar a la fase anterior.

Quinto, evaluación, en este paso se evalúan los resultados con base a los objetivos planteados al inicio del proyecto.

Sexto, despliegue, una vez creado el modelo se requiere estandarizar el conocimiento obtenido y compartirlo con el cliente, para que pueda utilizar la información a favor de su empresa.

2.5. Evaluación del modelo

En esta sección se describe la validación cruzada que es el método utilizado para la evaluación del modelo de clasificación crediticia.

2.5.1. Validación Cruzada

La validación cruzada es un método de evaluación de modelos de aprendizaje automático que intenta particionar las bases de datos en conjuntos de entrenamiento y pruebas con un alta imparcialidad e independencia de los datos (Wong & Yang, 2017).

El método más simple de validación cruzada es conocido como holdout, este método divide el conjunto de datos en conjuntos de entrenamiento y pruebas, en este tipo de validación se utiliza el método de entrenamiento para construir el modelo y el conjunto de pruebas para evaluar el mismo. Sin embargo, este método depende mucho de la división inicial por lo que su varianza puede ser alta (Wong & Yang, 2017).

Otro, es el método de validación cruzada k-iteraciones que se encuentra representada en la Figura 6, está técnica divide al conjunto de datos inicial en k subconjuntos de datos, posteriormente, se repite el método holdout k número de veces. Para probar el modelo, se selecciona uno de los k subconjuntos, en la figura se encuentran graficados de color gris, mientras que para entrenar al modelo se utiliza los k-1 restantes que se encuentran de color azul.

Al ejecutar el procedimiento en las k iteraciones, se calcula el resultado final promediando el porcentaje de precisión y error que se obtuvieron en cada una de las iteraciones. Adicionalmente, se recomienda utilizar la validación cruzada con 10-folds, debido a su alta precisión y baja varianza (Wong & Yang, 2017; Niu et al., 2018).

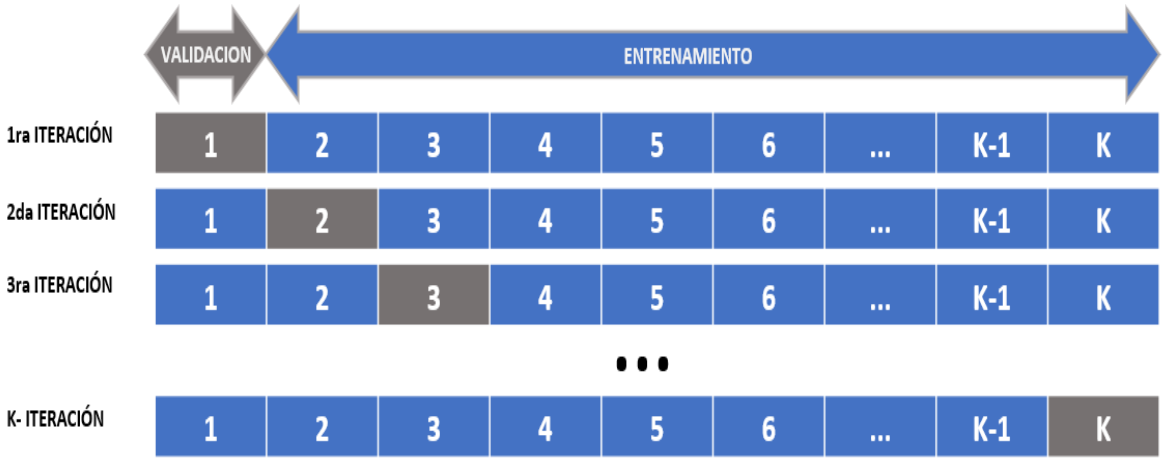

Figura 6: Diagrama validación cruzada k-iteraciones (Niu et al., 2018)

2.6. Pre-procesamiento de datos

La creciente necesidad de contar con información precisa ha sido fundamental para poder realizar estudios analíticos de manera correcta. Sin embargo, para alcanzar este objetivo se debe realizar un correcto pre-procesamiento de los datos, este paso es un gran desafío dentro del área de la ciencia de datos y se ha convertido en uno de los más importantes dentro de la aplicación de técnicas de inteligencia artificial (Gemp et al. 2017).

El pre-procesamiento de datos se encarga de asegurar la calidad de los datos que se ingresarán dentro del modelo, en esta etapa se manejan los errores e inconsistencias que pueden cambiar los resultados del análisis, por esta razón, el 60% del tiempo de un científico de datos es invertido en este procedimiento (Gemp et al., 2017).

Para la correcta ejecución de este proceso se deben realizar las siguientes verificaciones:

Verificar los registros con valores nulos, se debe revisar si existen valores faltantes que pueden generar errores al momento de entrenar al algoritmo, en caso de existir este tipo de casos se puede tomar las siguientes acciones:

- Incluir la moda o la media del conjunto en los espacios faltantes.
- Rellenar los espacios faltantes utilizando una regresión.
- Considerar los datos vacíos como una nueva categoría que puede llamarse indeterminado.
- Eliminar los registros con este tipo de casos. Sin embargo, se debería analizar si la omisión del registro pude tener impacto en el análisis.

Por ejemplo, Lup et al. (2001) colocó "UNKOWN" en lugar del valor nulo, como se muestra en la Tabla 2, este valor se encuentra de color gris.

Tabla 2: Ejemplo de Valor nulo (Lup Low et al., 2001)

ID	STATION	DATE	TEMP	VISIB	GUST	DEWP
0	1001499999	01/01/2019	27.2	1.2	17.1	16.5
1	1001499999	UNKNOWN	88.1	999.9	999.9	63
2	1001499999	01/01/2020	90	100	80	70

Verificar que los datos tengan el formato correcto, cuando se integra información de varias fuentes, existe la posibilidad que los datos tengan inconsistencias o diferentes formatos. Como, por ejemplo, Tang (2014) muestra el siguiente caso representado en las Tablas 3 y 4, se puede observar en la Tabla 3 que los campos CAPITAL y CITY tienen diferente formato en algunos registros, estas inconsistencias se encuentran de color gris.

Tabla 3: Conjunto de datos con diferencia de formato en campos (Tang, 2014)

NAME	COUNTRY	CAPITAL	CITY	CONF
George	China	Beijing	Beijing	Sigmod
Ian	China	Shanghai (beijing)	Hongkong (shanghai)	Icde
Peter	China	Tokyo	Tokyo	Icde
Alan	Usa	Olympia	Olympia	Icde
Jon	Mexico	Mexico	Mexico	Sigmod
Mike	Canada	Toronto (ottawa)	Toronto	Vldb

Para solventar las diferencias de la Tabla 3, se requiere unificar criterios y formatos, en la Tabla 4 se depuraron las inconsistencias, uno de los cambios fue la actualización de Shanghai (Beijing) por Beijing en el campo CAPITAL.

Tabla 4: Conjunto de datos con unificación de formato y criterio (Tang, 2014)

NAME	COUNTRY	CAPITAL	CITY	CONF
George	China	Beijing	Beijing	Sigmod
Ian	China	Beijing	Hongkong	Icde
Mike	Canada	Ottawa	Toronto	Vldb

Eliminar valores duplicados, es posible que algunos registros se repitan y estos valores afecten el entrenamiento del algoritmo, por lo que es vital eliminarlos del conjunto de datos para evitar que generen ruido en el análisis (Lup et al. 2001).

Lup et al. (2001) muestra un ejemplo de registros duplicados en la Tabla 5, en este caso se necesita primero unificar criterios de algunos campos, posteriormente se identifica los registros duplicados y se los elimina.

Tabla 5: Ejemplo de registros duplicados

Name	Address	Sex	Tel. no.
Tan Ah Kow	Blk 555, Bukit Merah, #11-01, S(112555)	M	222 1256
A.K. Tan	Apt Blk 555, B. Merah #11-01, Singapore 112555	M	2221256

Eliminación de outliers, cuando existen datos atípicos que difieren mucho de los demás, se los puede eliminar para que no generen ruido en el análisis. Sin embargo, es necesario evaluar el impacto de su omisión antes de eliminarlos del estudio. En el ejemplo, presentado por Safaei et al. (2020) representado en la Figura 7 se pueden observar a los datos atípicos de color azul.

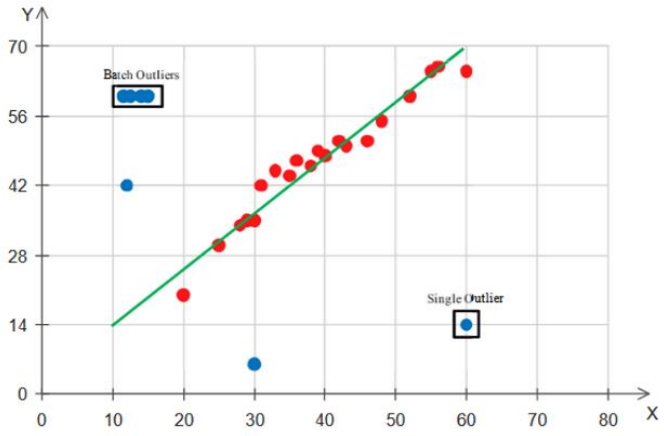

Figura 7: Ejemplo con datos outliers (Safaei et al., 2020)

2.6.1. Selección de Variables

La selección de variables ha demostrado ser efectiva para procesar conjunto datos de gran volumen y mejorar la eficiencia del aprendizaje, esta técnica es parte del pre-procesamiento de datos y se refiere al proceso de seleccionar las características que tengan mayor influencia para el estudio (Cai et al., 2018).

2.6.2.1. Chi Cuadrado

La técnica de Chi cuadrado es una prueba de independencia que ayuda a determinar el grado de correlación existente entre las variables. Se utiliza para reducir el número de columnas de un conjunto de datos, para lo cual se debe seleccionar las variables que tengan mayor grado de dependencia, cuando el valor es mayor existe mayor dependencia con la clase (Ramya & Kumaresan, 2015).

La fórmula propuesta por Hidalgo et al. (2020) y Guevara and Peñas (2020) se muestra en la Ecuación 2, en este caso c son los grados de libertad, x representa los valores del conjunto de datos de clientes y m los valores esperados. Adicionalmente, las m observaciones son clasificadas en k clases, en este caso son 2 y corresponden a los "buenos" y "malos" pagadores.

$$x_c^2 = \sum_{i-1}^{k} \frac{(x_i - m_i)^2}{m_i} \qquad (2)$$

CAPÍTULO III

TRABAJOS RELACIONADOS

4.1. Método de investigación

Para esta investigación se utilizó estudios realizados desde el 2017 en el campo de aprendizaje automático para la clasificación del riesgo crediticio, para el análisis se seleccionaron fuentes confiables de información de editoriales como lo son IEEE, Springer, ACM, Elsevier y Nature.

Para filtrar la información se utilizó el buscador Google Scholar utilizando palabras clave de búsqueda como “credit” + “rating” + “clasification” + “random forest” + client”.

Una vez encontrados los documentos con los filtros de búsqueda mencionados, se realizó una investigación experimental utilizando la base de clientes de Insotec, similar a lo realizado por Xu et al. (2020), en el cual se experimenta con las recomendaciones de los documentos y se compara los resultados con el conjunto de datos utilizado en el estudio.

4.2. Revisión de la literatura

En las últimas décadas, la economía de los diferentes sectores productivos se ha desarrollado gracias a la expansión de los servicios financieros, los préstamos otorgados por estas entidades brindan la oportunidad de obtener ingresos para impulsar los diferentes negocios, aplazando los pagos para los siguientes meses (Lanzarini et al., 2017).

Estas entidades han incluido modelos estadísticos dentro de sus procesos de evaluación, los clasificadores de crédito que utilizan estadísticas fueron introducidos en los años cincuenta, actualmente son utilizados mundialmente y se han convertido en una parte fundamental del proceso de concesión de créditos. El principal objetivo es evaluar el comportamiento de pago de los prestamistas, con el fin de reducir porcentaje de error (Zhang et al. 2019).

Existen investigaciones que implementan algoritmos de clasificación de clientes de crédito, que utilizan redes neuronales, bayes naive, máquinas de vectores de soporte y

árboles de decisión. Sin embargo, en este estudio se seleccionó *Random Forest* porque ha demostrado tener una alta precisión en varias investigaciones (Ziemba et al. 2020).

De varios de los trabajos relacionados Pradhan et al. (2020) compara algunos algoritmos como redes neuronales, arboles de decisión, Máquina de vectores de soporte y *Random Forest* en un conjunto de datos de clientes de crédito, el cual cuenta con 4.600 registros y 47 variables. Primero, se utilizó la técnica de "*feature selection*" la cual redujo el conjunto a 31 variables, en el experimento el algoritmo con *Random Forest* obtuvo mejores resultados con una precisión del 85% y una tasa de error del 10%.

Un experimento similar fue realizado por Arora and Kaur (2019) donde también se compararon varios clasificadores para un conjunto de datos de clientes de crédito de la base "*Lending Club*". Este *dataset* se encuentra en Kaggle, el cual posee 42.530 registros y 143 variables con un histórico desde el año 2007 al 2011. Posteriormente, se aplicaron las técnicas de chi-cuadrado, gain ratio, relieff y Bolasso con lo que el conjunto se redujo a 36 variables. En la aplicación de modelos de clasificación, *Random Forest* obtuvo una exactitud del 97.9%, una AUC o "área bajo la curva ROC (curva de características operativa del recepto)" del 93.4% y una tasa de error de 2.1% aproximadamente.

En otro estudio llevado a cabo por Zhang (2020), se propone un modelo con árboles de decisión C4.5 en una base de clientes de crédito, que cuenta con 1044 registros y 8 atributos. El resultado del modelo fue una precisión del 93.5%, con un porcentaje de error del 8.8%, de los cuales 92 fueron falsos positivos y 0 falsos negativos.

Otro aporte importante fue propuesto por Subasi and Cankurt (2019), para evaluar el comportamiento de pago de los clientes en uno de sus productos de préstamo, en este trabajo se compararon algunos algoritmos como en Arora and Kaur (2019) y Pradhan et al. (2020). En el experimento el conjunto de datos está compuesto por 25,000 registros y 23 variables. Después de aplicar los diferentes algoritmos de clasificación, el modelo con *Random Forest* obtuvo el mayor porcentaje de precisión con el 89.01%, una AUC del 95% y una tasa de error del 11,68%.

CAPÍTULO IV

MÉTODO Y MATERIALES

En esta sección se describe los métodos y materiales utilizados en este estudio, siguiendo el flujo de la metodología CRISP-DM detallada en el punto 2.4, inicialmente se requiere entender al negocio y la información que se utiliza en el estudio, por esta razón primero se detalla y analiza al conjunto de datos utilizado en el análisis. Posteriormente, se describe el preprocesamiento de datos, el modelado y la evaluación.

4.1. Conjunto de datos de Clientes de Crédito

El conjunto de datos contiene información de los clientes de Insotec, del período que comprende desde el 1 de enero del 2017 hasta el 31 de diciembre del 2020. La información está clasificada por "buenos" y "malos" pagadores. Adicionalmente, el conjunto está compuesto por 18 variables y 63.896 registros.

Las variables del estudio se detallan en la Tabla 6, donde se puede observar los tipos de datos, una pequeña descripción de la información y un ejemplo del contenido.

Tabla 6: Detalle de las variables del conjunto de datos.

Variable	Tipo de dato	Tipo de información	Ejemplo
Tipo_prestamo	Texto	Tipo de préstamo entregado	Directo
Año	Entero	Año de ultima transacción	2017
Mes	Texto	Mes de ultima transacción	2
Genero	Entero	Genero	Masculino
Estado_civil	Texto	Estado civil	Soltero
Instrucción	Texto	Nivel de instrucción	Primaria
Atraso	Texto	Atraso máximo en días	30
Edad	Entero	Edad actual	20
Patrimonio	Texto	Patrimonio actual en dólares	10000
Provincia	Texto	Provincia de residencia	Pichincha
Canton	Texto	Cantón de residencia	Quito
Tiempofuncionamiento	Entero	Días que ha existido el negocio	100
Deudainicial	Double	Deuda inicial del ultimo préstamo	2000
Actividad_economica	Texto	Actividad económica	Manufactura
Calificacionactual	Texto	Calificación crediticia	A-1
Utilidad	Entero	Utilidad del negocio en dólares	10000
Ingresos	Entero	Ingresos del negocio en dólares	20000
Gastos	Entero	Gastos del negocio en dólares	10000

Variable	Tipo de dato	Tipo de información	Ejemplo
Es_buen_pagador	Texto	Si califica como buen pagador o no, es la clase que divide al conjunto de datos	BP y MP

4.2. Análisis de gráficas

En esta sección se realiza un análisis descriptivo con los graficas de frecuencia, dispersión y cajas de las variables más importantes del conjunto de datos que se detallan en la sección 4.3.2.

4.2.1. Gráficas de Frecuencia

El conjunto de datos de clientes está conformado por dos clases “buenos” y “malos pagadores” representados en la Figura 8 de color azul y rojo respectivamente, en esta figura se puede observar los diagramas de frecuencia de las variables, en el caso de la clase “buenos pagadores” se cuenta con 58.472 instancias, mientras que los “malos” 5.424 registros, correspondientes al 92% y 8% respectivamente.

Adicionalmente, se observa que la mayoría de los “buenos pagadores” tienen bajos “atrasos”, “calificaciones” buenas en contraste de los “malos pagadores” los cuales mayormente reflejan un “tiempo de funcionamiento” bajo.

Finalmente, en los cuadros de la variable de “ingresos”, la mayoría de los “buenos pagadores” tienen valores medianos al igual que en la variable “gastos”.

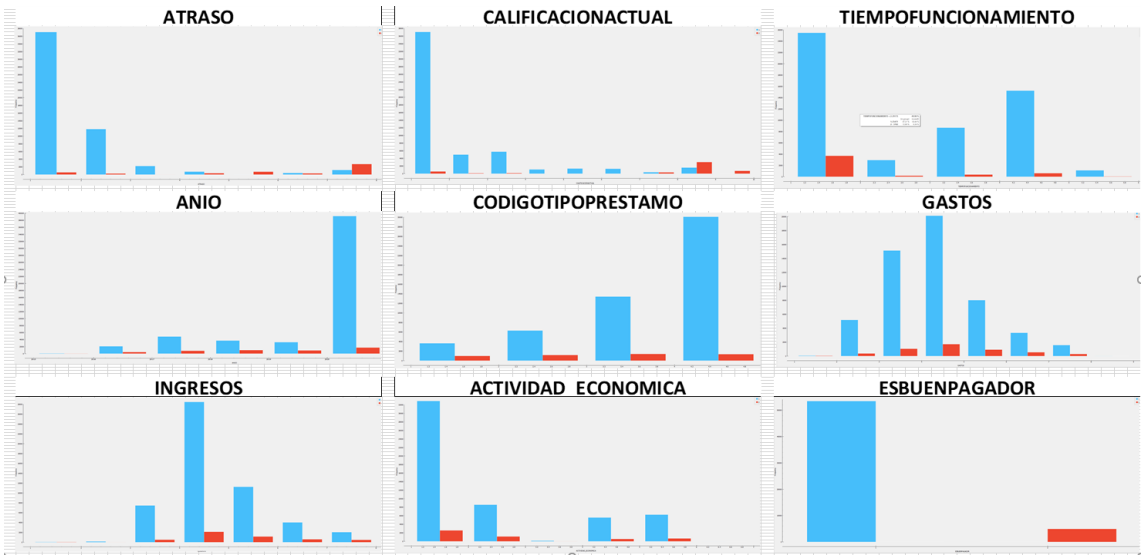

Figura 8: Distribución de datos de la clase

4.2.2. Gráficas de Dispersión

Las gráficas de dispersión de todas las variables se pueden observar en la Figura 9, la cual fue realizada en la herramienta WEKA, esta gráfica permite visualizar que la mayoría de variables del conjunto de datos no tienen correlación, o poseen una baja correlación negativa, como por ejemplo en la gráfica de las variables de "gastos" versus el "tiempo de funcionamiento".

Por otra parte, considerando que los gráficos de "ingreso" versus "gastos" y "atraso" versus "calificación actual" presentan una alta correlación positiva, se concluye que las variables de "atraso", "ingreso", "gastos" y "calificación actual tienen mayor importancia dentro de este estudio. Adicionalmente, se observa que "tiempo de funcionamiento" podría tener un grado moderado de significancia en el estudio. En la Figura 9, se observan subrayadas de color rojo las variables de más relevancia para el análisis.

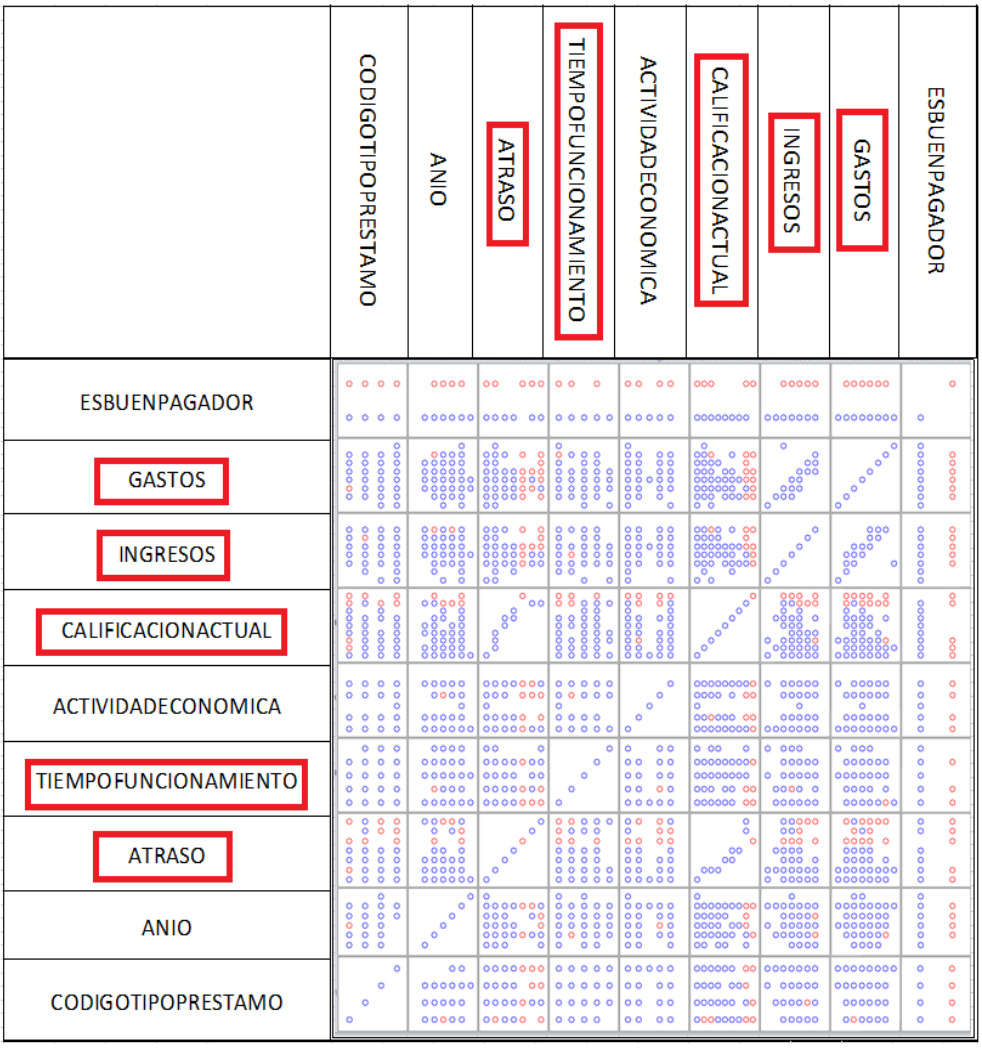

Figura 9: Gráficas de dispersión de las variables

En la tabla 7 se visualiza la matriz de correlación de las variables por Pearson, en esta matriz se observa que las variables con mayor correlación resaltadas en rojo, son “atraso” versus “calificación actual” y “gastos” versus “ingresos”, esto confirma lo mencionado anteriormente. Sin embargo, el tiempo de funcionamiento que fue considerada con significancia moderada anteriormente no tiene valores relevantes en esta matriz.

Tabla 7: Matriz de correlación por Pearson

	ACTIVIDAD ECONOMICA	ANIO	ATRASO	CALIFICACION ACTUAL	CODIGO TIPO PRESTAMO	GASTOS	INGRESOS
ANIO	0.02						
ATRASO	0.066	-0.722					
CALIFICACIONACTUAL	0.078	-0.686	1.912				
CODIGOTIPOPRESTAMO	-0.092	0.35	-0.482	-0.434			
GASTOS	0.504	0.15	0.024	0.023	-0.064		
INGRESOS	0.428	0.112	0.056	0.052	-0.094	1.792	
TIEMPOFUNCIONAMIENTO	-0.216	-0.316	0.005	0.008	-0.114	-0.098	-0.09

4.2.3. Gráficas de Cajas

En la Figura 10, se visualiza las gráficas de cajas de las variables más relevantes, estas visualizaciones fueron realizadas en RStudio que es una herramienta de desarrollo integrado que utiliza el lenguaje R para implementación de procesos computacionales capaces de resolver problemas estadísticos complejos (Ren, 2016).

Las gráficas de las variables se encuentran en función del comportamiento de pago, en referencia al comportamiento de pago, la categoría uno corresponde a los buenos pagados mientras que la categoría dos a los malos pagadores, en estas gráficas se visualiza que en las variables de atraso, calificación actual, tiempo de funcionamiento, año, gastos, ingresos y actividad económica se presentan datos atípicos, estos se encuentran representados por círculos.

Se observa en el literal “a” de la Figura 10, que la media del atraso para los “buenos pagadores” se encuentra en la categoría uno, por otro lado, el promedio para los “malos pagadores” se encuentra en el rango 7.

Es interesante visualizar en el literal “b”, que la variable “calificación actual” tiene un comportamiento similar al “atraso" representada en el literal “a”. Sin embargo, en este

caso se observa que la mayoría de los “malos pagadores” están dentro del rango número 8.

Adicionalmente, se observa que los promedios de los “buenos” y “malos” pagadores son similares para las variables de “ingresos”, “gastos” y “actividad económica” pertenecientes a los literales “g”, “f” y “h” respectivamente.

Finalmente, analizando la gráfica de tiempo de funcionamiento del literal “c”, se puede concluir que los “buenos pagadores” se encuentran distribuidos en las diferentes categorías, en contraste, los “malos pagadores” que se encuentran dentro de la primera categoría que pertenece a los negocios con poco tiempo de funcionamiento.

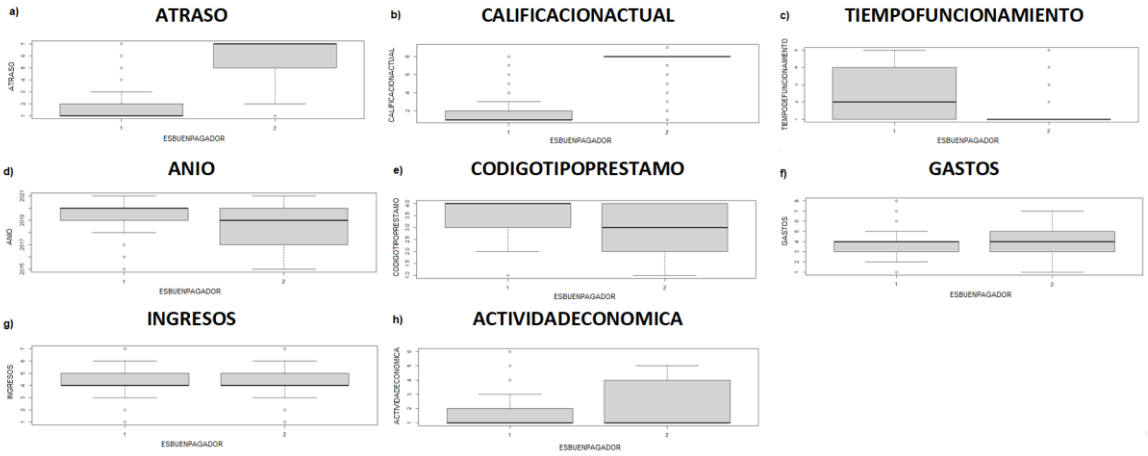

Figura 10: Gráficas de cajas de las variables seleccionadas

4.3. Método

En esta sección se detalla el método utilizado en el estudio, donde se aplica la metodología CRISP-DM, inicialmente se realiza el entendimiento del negocio y del conjunto de datos que se menciona en el punto 4.1. Posteriormente, se realiza la preparación de la información, modelado y evaluación.

4.3.1. Preprocesamiento de datos

La etapa de preprocesamiento de datos tiene una importancia muy significativa en el análisis de datos, ya que en la misma se elimina el ruido, instancias y variables que no aportan valor al estudio (Kotsiantis et al. 2006).

En este paso se elimina los datos atípicos para evitar que exista ruido en el análisis, para realizar este procedimiento se utiliza Orange, esta herramienta es un framework visual

que utiliza lenguaje Python, sus usuarios pueden realizar los diferentes procesos del aprendizaje automático con herramientas visuales sin la necesidad de entender el código (Demšar et al., 2004).

En la figura 11 se observa el procedimiento de creación del modelo, desde la carga del conjunto inicial de datos hasta el ingreso al algoritmo de modelado *Random Forest.* En este proceso, primero se carga la información para lo que se utiliza el componente "*File*" representado con la letra A, segundo se elimina los outliers con el componente "*Outliers*" representado por la letra B.

Una vez eliminados los datos atípicos se selecciona las variables con mayor correlación utilizando el componente "*Rank*" representado con la letra D, en este paso se utiliza Chi cuadrado para seleccionar las variables. Finalmente, se selecciona las variables con mayor significancia en el conjunto de datos con el componente "*Select Columns*" representado con la letra C. Posteriormente, se ingresa la información seleccionada al algoritmo con el componente "*Random Forest"* representado con la letra E.

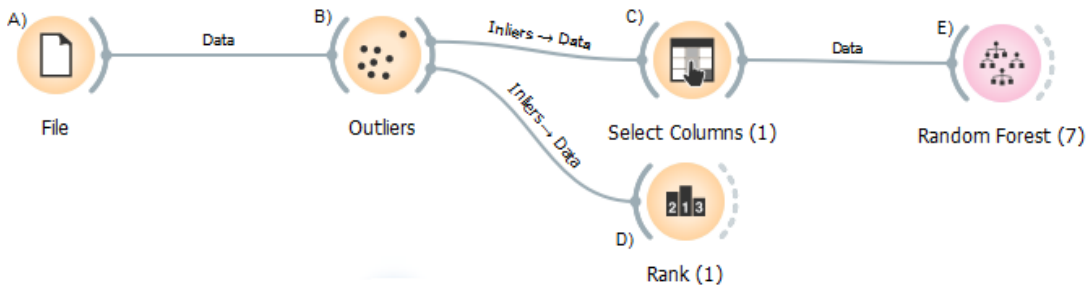

Figura 11: Procesos de limpieza y carga de datos

Como se observa en la Figura 11 primero se elimina los datos atípicos, un ejemplo de este tipo de datos se puede observar en la Figura 12, en este caso se evalúa el "tiempo de funcionamiento" (se categoriza con 1 al negocio con menos tiempo y a 5 con mayor tiempo de funcionamiento) versus los "ingresos" (1 representa el menor ingreso, mientras que 8 al mayor nivel de ingreso) en función de la clase (buenos y malos pagadores de color azul y rojo respectivamente).

Se puede evidenciar que existe un registro que difiere mucho de los demás, porque cuenta con un "tiempo de funcionamiento" muy bajo y sus "ingresos" son muy altos. Este dato es omitido para el análisis como se observa en la figura 13.

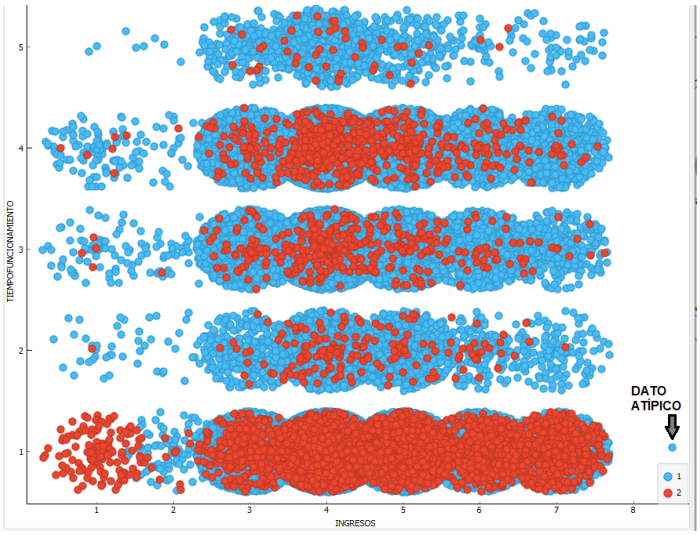

Figura 12: Dato atípico

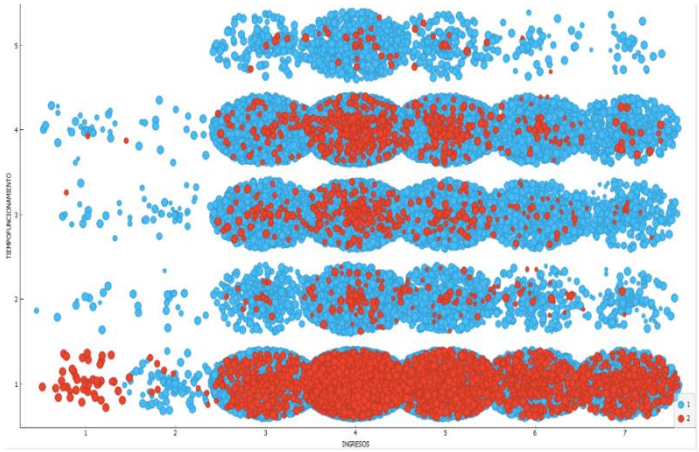

Figura 13: Eliminación de dato atípico

Después de realizar este procedimiento se encontraron 5.649 datos atípicos, el conjunto se redujo de 63896 a 58.247 registros. Posteriormente, se aplica la técnica del chi cuadrado para la selección de variables para este proceso también se utiliza Orange, donde solo 8 de las 18 variables tienen relevancia.

En el análisis, las variables con mayor correlación con la clase son atraso y calificación actual con valores de 29.010 y 24.751 respectivamente, estas variables tienen mayor nivel de importancia para identificar a los “buenos” y “malos” pagadores.

Por otro lado, las variables “tiempo funcionamiento”, “año”, “tipo préstamo”, “gasto”, “ingresos” y “actividad económica” obtuvieron valores de 1.429, 1.288, 938, 405, 253 y 176, estos resultados son mayores a 100 por lo que también fueron incluidos dentro de los campos seleccionados. Por otro lado, las variables “edad”, “instrucción”,

"deudainicial", "utilidad", "mes", "patrimonio", "estadocivil", "género" y "provincia" obtuvieron valores menores a 100 y fueron excluidos del análisis. Los resultados de todas las variables al aplicar la técnica de chi-cuadrado se pueden observar en la Tabla 7.

Tabla 8: Resultado de Variables aplicando chi-cuadrado

VARIABLE	VALOR
ATRASO	29010.2
CALIFICACIONACTUAL	24751.2
TIEMPOFUNCIONAMIENTO	1429.4
ANIO	1288.8
CODIGOTIPOPRESTAMO	938.4
GASTOS	405.3
INGRESOS	253.4
ACTIVIDAD_ECONOMICA	176.8
EDAD	38.3
INSTRUCCION	37.9
DEUDAINICIAL	36.6
CANTON	32.8
UTILIDAD	18.8
MES	17.4
PATRIMONIO	10.8
ESTADO_CIVIL	3.3
GENERO	2.3
PROVINCIA	0.6

Después de la selección de variables con chi cuadrado se define el vector resultado de la siguiente manera $d_n = \{a, ca, tf, an, tp, g, i, ae\}$. La variable *a* representa a los "atrasos", *ca* a la "calificación actual", *tf* el "tiempo de funcionamiento" del negocio, *an* el "año" de la última transacción, *tp* el "tipo de préstamo" de la última transacción, *g* al "gasto" del negocio, *i* al "ingreso" y *ae* a la "actividad económica" del cliente.

En la tabla 8 se despliegan los valores de las variables que conforman el vector resultado provenientes de las variables seleccionadas después de aplicar la técnica de chi cuadrado.

Tabla 9: Resultado de Variables aplicando chi-cuadrado

VARIABLE DEL VECTOR	VALOR
a	29010.2
ca	24751.2
tf	1429.4
an	1288.8

VARIABLE DEL VECTOR	VALOR
tp	938.4
g	405.3
i	253.4
ae	176.8

El vector resultado, que está compuesto de las variables mostradas en la Tabla 8, es ingresado al modelo utilizando Orange como se muestra en la figura 11, donde posteriormente se lo configura para que utilice un algoritmo basado en *Random Forest.*

4.3.2. Modelo con *Random Forest* vs Otros Algoritmos

De acuerdo con la metodología CRISP-DM, una vez preparada la información, se continúa con la fase de modelado, en esta etapa se seleccionan las técnicas que se utilizarán dentro del modelo (Martínez et al. 2019). En esta investigación, primero se realiza una revisión literaria para conocer las recomendaciones en investigaciones similares. Posteriormente, se comparan los resultados de los diferentes algoritmos.

En investigaciones similares se recomienda *Random Forest* debido a su alta precisión, Como los presentados por Arora and Kaur (2019), Pradhan et al. (2020) y Ziemba et al. (2020). En los estudios mencionados, los modelos con *Random Forest* obtuvieron los mejores resultados en comparación a otros que utilizaron diferentes técnicas como redes neuronales y arboles de decisión.

Al realizar una comparación similar en el estudio como se observa en la Tabla 9, el modelo que utilizó *Random Forest* obtuvo la precisión más alta con 97,20%, por otro lado, los modelos de redes neuronales, árboles de decisión y Máquina de vectores de soporte obtuvieron precisiones de 97,1%, 96,8% y 91,6% respectivamente. Por los resultados y las recomendaciones de la literatura se selecciona *Random Forest* para crear el modelo.

Tabla 10: Comparación resultados de algoritmos

Algoritmo	Precisión
Random Forest	97,20%
Redes Neuronales	97,10%
Arboles de decisión	96,80%
Máquina de vectores de soporte	91,60%

Como se mencionó en la sección 4.2.1, se seleccionaron 8 variables y 58.247 instancias para ser ingresadas en el modelo, para la configuración del modelo se recomiendan 50 árboles (Mercadier & Lardy, 2019). Sin embargo, cuando se prueba el modelo con 10, 15, 50 y 100 árboles, los resultados de precisión son similares por lo que esta configucación no tiene mayor impacto en el modelo.

4.3.3. Validación Cruzada

Una vez creado el modelo, se evalúa los resultados de acuerdo a la metodología CRISP-DM, para este proceso se ocupa RStudio.

Para la evaluación del modelo, se utiliza la validación cruzada, se prueba desde 5 hasta 15 modelos, los resultados han mostrado precisiones similares desde los k=10. Se utiliza 90% de los registros para entrenamiento, mientras que el 10% es utilizado para las pruebas.

El promedio de los resultados de los 10 modelos es igual a 97,29%, estos resultados se observan en la Tabla 10. Posteriormente, se aplica la prueba de normalidad a los resultados de los 10 modelos, el conjunto de datos obtiene un P-valor = 0,60 por lo que se aprueba la hipótesis nula y se concluye que el conjunto de datos posee una distribución normal.

Tabla 11: Resultados validación cruzada k=10

#	CLASIFICADOS CORRECTAMENTE	CLASIFICADOS INCORRECTAMENTE	PRECISION
1	56,601	1,646	97.17%
2	56,512	1,735	97.02%
3	56,831	1,416	97.57%
4	56,811	1,436	97.53%
5	56,901	1,346	97.69%
6	56,681	1,566	97.31%
7	56,472	1,775	96.95%
8	56,611	1,636	97.19%
9	56,621	1,626	97.21%
10	56,651	1,596	97.26%

CAPÍTULO V

RESULTADOS

En esta sección se describen el resultado general del modelo. Adicionalmente, se ingresan datos de prueba que contienen dos perfiles de clientes de la institución, y se utilizan para evaluar la precisión del modelo.

5.1. Resultados del modelo

Para el experimento se creó 10 modelos con *Random Forest,* para la evaluación se utiliza la validación cruzada con una división 90-10 como se menciona en la sección 4.3.3, los resultados de la matriz de confusión del promedio de los 10 modelos se visualizan en la Tabla 11.

Tabla 12: Matriz de transición de las instancias

		PREDICCION		
		BUENOS PAGADORES	**MALOS PAGADORES**	∑
ACTUAL	**BUENOS PAGADORES**	52773	480	53253
	MALOS PAGADORES	1146	3848	4994
	∑	53919	4328	58247

La precisión del modelo general es del 97,29% como se explica en la sección 4.2.4. Adicionalmente, en la matriz de confusión detallada en la tabla 12 se observa lo siguiente:

- 52773 verdaderos positivos, clientes que son “buenos pagadores” y el modelo los clasifica como “buenos pagadores”.
- 3848 verdaderos negativos, clientes que son “malos pagadores” y el modelo los clasifica como “malos pagadores”.
- 480 falsos positivos, clientes que son “buenos pagador”, pero son clasificados como “malos pagadores”,
- 1146 Falsos negativos, clientes que son “buenos pagadores”. Sin embargo, son clasificados como “malos pagadores”.
- Al dividir los registros correctamente clasificados versus el total de instancias de cada clase se obtiene una precisión para los “buenos pagadores” del 97,9%, mientras que para los “malos pagadores” la precisión es del 89% lo que indica que el modelo tiene más dificultad para clasificar a los “malos pagadores”.

5.1. Validación de datos de ejemplo

Para la validación de los datos de ejemplo se utilizó Orange, donde se ingresa la información de clientes potenciales en el modelo de clasificación de clientes. Posteriormente, la herramienta los clasifica automáticamente como “buenos” o “malos” pagadores dependiendo de las variables, en esta sección se ingresaron dos perfiles de clientes que son descritos a continuación:

Los resultados del perfil 1 se pueden observar la Tabla 12, este perfil pertenece a un cliente que tiene un nivel de atraso alto, una calificación C1(considerado como mala), no tiene tiempo de funcionamiento de su negocio y por el momento no genera utilidad.

Se puede concluir del perfil 1, que la persona no tiene un negocio establecido y en otros emprendimientos ha fracasado por lo que sus utilidades son bajas y su calificación no es buena.

El modelo clasifica al cliente del perfil 1 con un 76% de ser un “mal” pagador y un 24% de ser un “buen” pagador, esta predicción es acertada. Sin embargo, el porcentaje de precisión no es lo suficientemente alta para aceptar la clasificación de modelo sin realizar otras validaciones.

Tabla 13: Perfil 1 de solicitante de crédito

CLASE	BUEN PAGADOR	MAL PAGADOR
PREDICCION	76%	24%

Los resultados de predicción del perfil 2 se observan en la tabla 13, este perfil pertenece a un cliente que tiene registrado un atraso de 100 días, una calificación de B-2(catalogada como regular), un tiempo de funcionamiento de 12 meses, y gastos bajos al igual que sus ingresos,

El modelo clasifica a la persona del perfil 2 con un 94% de ser un “buen” pagador y un 6% de ser “mal” pagador, esta clasificación es correcta, ya que el atraso lo registró en época de pandemia.

Sin embargo, el cliente es considerado dentro de los “buenos pagadores” por su experiencia en el negocio. Además, el cliente genera una utilidad constante.

Adicionalmente, el resultado del modelo es confiable, debido a que el nivel de precisión es del 94%.

Tabla 14: Perfil 2 de solicitante de crédito

CLASE	BUEN PAGADOR	MAL PAGADOR
PREDICCION	94%	6%

CAPÍTULO VI

CONCLUSIONES Y TRABAJOS FUTUROS

Primero se realizó una investigación de literatura utilizando principalmente las plataformas de GOOGLE SCHOLAR, en estos buscadores se encontraron artículos relacionados con la implementación de modelos algorítmicos utilizando inteligencia artificial dentro del campo de clasificación de riesgo crediticio en entidades financieras. Esta información permitió encontrar metodologías y técnicas para la implementación del modelo de clasificación con la información de Insotec.

Posteriormente, se implementó un modelo utilizando la metodología CRISP-DM, después de comprender el negocio y la información utilizada en el análisis, se preparó el conjunto de datos, en donde se redujo el conjunto de datos de 63.896 a 58.247 registros, y se seleccionaron 8 variables para ingresarlas al modelo.

Una vez preparada la información, según la metodología *CRISP-DM* se seleccionan los métodos y técnicas. En este caso se escogió el método ensamblado de *Random Forest* por recomendaciones bibliográficas y se lo comparó versus otros algoritmos como árboles de decisión, redes neuronales y máquina de vectores de soporte.

Al realizar la comparación de los resultados de precisión de los algoritmos, se evidenció que los resultados de *Random forest* fueron los mejores para el conjunto de datos, ya que se obtuvo una precisión del 97,2% y una tasa de error del 2.8%. Adicionalmente, el porcentaje de error para clasificar a los “malos” y “buenos” pagadores fue del 11% y 2% respectivamente. Estos resultados permiten clasificar a los “buenos” clientes. Sin embargo, se dificulta identificar a los “malos pagadores” y es necesario mejorar este porcentaje para tener mayor confiabilidad en el modelo.

Finalmente, el promedio de los resultados de la validación cruzada de los 10 modelos es igual a 97,29%, al aplicar la prueba de normalidad a los resultados de estos modelos, se obtuvo un p-valor de 0,6 por lo que se concluye que el conjunto es normal.

Para futuras investigaciones, se debería evaluar la posibilidad de incluir una nueva categoría en la clase, el modelo actual únicamente cuenta con dos clases como lo son los “buenos” y “malos” pagadores, por lo que sería interesante segmentar el conjunto de datos

con una tercera clase que sea intermedia, de esta manera se podría analizar si la precisión mejora.

El estudio utilizó la metodología *CRISP-DM* para la implementación del modelo. Sería importante evaluar el sexto paso referente al despliegue del modelo de clasificación dentro de la organización, en este paso se evalúa el grado de impacto del modelo en la organización y de aceptación.

Insotec es una organización con un nivel de riesgo muy alto por lo que la inclusión del modelo dentro de los procedimientos de clasificación de riesgo debe ser paulatino para incrementar la confianza y disminuir el riesgo. En este proceso se puede modificar al modelo con la inclusión de otras variables que puedan ser importantes para la calificación de los clientes, por ejemplo, variables externas que no se disponían en este estudio, como información de big data, por ejemplo, créditos activos en otras instituciones financieras.

Finalmente, el estudio se enfocó a estudiar las técnicas de chi cuadrado y *Random Forest*, pero sería interesante combinarlas en el modelo para analizar los resultados. Por ejemplo, en la fase de modelado se podría combinar varios algoritmos como redes neuronales o máquina de vectores de soporte, con el fin de comparar sus resultados y seleccionar los modelos combinados que obtengan la mayor precisión como lo realizaron Arora & Kaur (2019) y Safari (2020).

BIBLIOGRAFIA

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, *147*, 77–89. https://doi.org/10.1016/j.enbuild.2017.04.038

Alfakih, T., Hassan, M. M., Gumaei, A., Savaglio, C., & Fortino, G. (2020). Task Offloading and Resource Allocation for Mobile Edge Computing by Deep Reinforcement Learning Based on SARSA. *IEEE Access*, *8*, 54074–54084. https://doi.org/10.1109/ACCESS.2020.2981434

Amruthnath, N., & Gupta, T. (2018). A research study on unsupervised machine learning algorithms for early fault detection in predictive maintenance. *2018 5th International Conference on Industrial Engineering and Applications (ICIEA)*, 355–361. https://doi.org/10.1109/IEA.2018.8387124

Arora, N., & Kaur, P. (2019). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, *86*, 105936. https://doi.org/10.1016/j.asoc.2019.105936

Banco Mundial. (2021). Banco Mundial. In Banco Mundial. https://www.bancomundial.org/es/home

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, *86*, 42–53. https://doi.org/https://doi.org/10.1016/j.eswa.2017.05.050

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Chen, Z., Teoh, E. N., Nazir, A., Karuppiah, E. K., Lam, K. S., & others. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, *57*(2), 245–285.

Demšar, J., Zupan, B., Leban, G., & Curk, T. (2004). Orange: From Experimental Machine Learning to Interactive Data Mining. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Knowledge Discovery in Databases: PKDD 2004* (pp. 537–539). Springer Berlin Heidelberg.

Gemp, I., Theocharous, G., & Ghavamzadeh, M. (2017). Automated Data Cleansing through Meta-Learning. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 4760–4761.

Guevara, C., & Peñas, M. S. (2020). Surveillance Routing of COVID-19 Infection Spread Using an Intelligent Infectious Diseases Algorithm. *IEEE Access*, *8*, 201925–201936. https://doi.org/10.1109/ACCESS.2020.3036347

Hidalgo, J., Guevara, C., & Yandún, M. (2020). Generation of User Profiles in UNIX Scripts Applying Evolutionary Neural Networks. In I. Corradini, E. Nardelli, & T. Ahram (Eds.), *Advances in Human Factors in Cybersecurity* (pp. 56–63). Springer International Publishing.

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2019). DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, 403–408. https://doi.org/10.1016/j.procir.2019.02.106

Insotec. (2021). CREDITO PARA EL DESARROLLO. In *Insotec*. https://www.insotec-ec.com/

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, *1*, 111–117.

Lanzarini, L., Villa Monte, A., Bariviera, A. F., & Jimbo Santana, P. (2017). Simplifying credit scoring rules using LVQ + PSO. *Kybernetes*, *46*, 8–16. https://doi.org/10.1108/K-06-2016-0158

Lee, I., & Shin, Y. J. (2020). Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons*, *63*(2), 157–170. https://doi.org/10.1016/j.bushor.2019.10.005

Leo, M., Sharma, S., & Maddulety, K. (2019). Machine Learning in Banking Risk Management: A Literature Review. *Risks*, *7*(1), 29. https://doi.org/10.3390/risks7010029

Liebergen, B. (2017). Machine learning: a revolution in risk management and compliance? *Journal of Financial Transformation*, *45*, 60–67.

Lup Low, W., Li Lee, M., & Wang Ling, T. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems*, *26*(8), 585–606. https://doi.org/10.1016/S0306-4379(01)00041-2

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández Orallo, J., Kull, M., Lachiche, N., Ramírez Quintana, M. J., & Flach, P. A. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 1. https://doi.org/10.1109/TKDE.2019.2962680

Mercadier, M., & Lardy, J. P. (2019). Credit spread approximation and improvement using random forest regression. *European Journal of Operational Research*, *277*(1), 351–365. https://doi.org/10.1016/j.ejor.2019.02.005

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning*. The MIT Press.

Nisioti, A., Mylonas, A., Yoo, P. D., & Katos, V. (2018). From Intrusion Detection to Attacker Attribution: A Comprehensive Survey of Unsupervised Methods. *IEEE Communications Surveys Tutorials*, *20*(4), 3369–3388. https://doi.org/10.1109/COMST.2018.2854724

Niu, M., Li, Y., Wang, C., & Han, K. (2018). RFAmyloid: A Web Server for Predicting Amyloid Proteins. *International Journal of Molecular Sciences*, *19*(7). https://doi.org/10.3390/ijms19072071

Otoum, S., Kantarci, B., & Mouftah, H. (2019). Empowering Reinforcement Learning on Big Sensed Data for Intrusion Detection. *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 1–7. https://doi.org/10.1109/ICC.2019.8761575

Pandey, T. N., Jagadev, A. K., Mohapatra, S. K., & Dehuri, S. (2017). Credit risk analysis using machine learning classifiers. *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, 1850–1854. https://doi.org/10.1109/ICECDS.2017.8389769

Pradhan, M. R., Akter, S., & Al Marouf, A. (2020). Performance Evaluation of Traditional Classifiers on Prediction of Credit Recovery. In T. Sengodan, M. Murugappan, & S. Misra (Eds.), *Advances in Electrical and Computer Technologies* (pp. 541–551). Springer Singapore.

Ramya, R., & Kumaresan, S. (2015). *Analysis of feature selection techniques in credit risk assessment*. 1–6. https://doi.org/10.1109/ICACCS.2015.7324139

Ruiz-Montiel, M., Mandow, L., & Pérez-de-la-Cruz, J. L. (2017). A temporal difference method for multi-objective reinforcement learning. *Neurocomputing*, *263*, 15–25. https://doi.org/10.1016/j.neucom.2016.10.100

Safaei, M., Asadi, S., Driss, M., Boulila, W., Alsaeedi, A., Chizari, H., Abdullah, R., & Safaei, M. (2020). A Systematic Literature Review on Outlier Detection in Wireless Sensor Networks. *Symmetry*, *12*(3). https://doi.org/10.3390/sym12030328

Safari, M. J. S. (2020). Hybridization of multivariate adaptive regression splines and random forest models with an empirical equation for sediment deposition prediction in open channel flow. *Journal of Hydrology*, *590*, 125392. https://doi.org/10.1016/j.jhydrol.2020.125392

Shi, B., Zhao, X., Wu, B., & Dong, Y. (2019). Credit rating and microfinance lending decisions based on loss given default (LGD). *Finance Research Letters*, *30*, 124–129. https://doi.org/10.1016/j.frl.2019.03.033

Siswanto, Abdussomad, Gata, W., Wardhani, N. K., Gata, G., & Prasetvo, B. H. (2019). The Feasibility of Credit Using C4.5 Algorithm Based on Particle Swarm Optimization Prediction. *2019 6th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 416–421. https://doi.org/10.23919/EECSI48112.2019.8977074

Sobarsyah, M., Soedarmono, W., Yudhi, W. S. A., Trinugroho, I., Warokka, A., & Pramono, S. E. (2020). Loan growth, capitalization, and credit risk in Islamic banking. *International Economics*, *163*, 155–162. https://doi.org/10.1016/J.INTECO.2020.02.001

Subasi, A., & Cankurt, S. (2019). Prediction of default payment of credit card clients using Data Mining Techniques. *2019 International Engineering Conference (IEC)*, 115–120. https://doi.org/10.1109/IEC47844.2019.8950597

Tang, N. (2014). Big Data Cleaning. In L. Chen, Y. Jia, T. Sellis, & G. Liu (Eds.), *Web Technologies and Applications* (pp. 13–24). Springer International Publishing.

Thanh, N. D., Duong, L. T. T., & That, N. H. (2020). Investigating determinants of competitiveness of the retail banking service in Vietnam: a customer approach. *Journal of International Economics and Management*, *20*(1), 80–100.

Uthayakumar, J., Vengattaraman, T., & Dhavachelvan, P. (2020). Swarm intelligence based classification rule induction (CRI) framework for qualitative and quantitative approach: An application of bankruptcy prediction and credit risk analysis. *Journal of King Saud University - Computer and Information Sciences*, *32*(6), 647–657. https://doi.org/10.1016/j.jksuci.2017.10.007

Wang, S., Ren, W., Zhang, Y., & Liang, F. (2019). Random Forest Classifier for Distributed Multi-plant Order Allocation. In G. Q. Huang, C.-F. Chien, & R. Dou (Eds.), *Proceeding of the 24th International Conference on Industrial Engineering and Engineering Management 2018* (pp. 123–132). Springer Singapore.

Wong, T.-T., & Yang, N.-Y. (2017). Dependency Analysis of Accuracy Estimates in k-Fold Cross Validation. *IEEE Transactions on Knowledge and Data Engineering*, *29*(11), 2417–2427. https://doi.org/10.1109/TKDE.2017.2740926

Zhang, W., He, H., & Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring. *Expert Systems with Applications*, *121*, 221–232. https://doi.org/10.1016/j.eswa.2018.12.020

Zhang, X.-D. (2020). Machine learning. In *A Matrix Algebra Approach to Artificial Intelligence* (pp. 223–440). Springer.

Ziemba, P., Radomska-Zalas, A., & Becker, J. (2020). Client evaluation decision models in the credit scoring tasks. *Procedia Computer Science*, *176*, 3301–3309. https://doi.org/10.1016/j.procs.2020.09.068

ANEXOS

a. Carta de Aceptación

El artículo fue enviado a "IHSED 2021-4th International Conference on Human Systems Engineering and Design: Future Trends and Applications" que tendrá lugar el 23-25 de septiembre de 2021 en Dubrovnik - Croacia. Adicionalmente, el artículo científico entra en la categoría Q4 de Scopus.

ACCEPTANCE LETTER

Juan FREIRE LOPEZ
Universidad Internacional SEK, Ecuador
jgfreire.mdat@uisek.edu.ec

May 9, 2021

Dear Juan FREIRE LOPEZ,

We are pleased to inform you that your submission has been accepted for Oral presentation at the 12th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences to be held at Virtual Conference, Manhattan, New York, United States of America, 25-29, July, 2021. (http://ahfe.org)

Paper ID#: 115

Paper Title: Application of Decision Tree to Banking Classification Model

The acceptance decision is based on peer-reviews conducted by conference chairs and assigned reviewers from the scientific committee.

[For inclusion in the AHFE 2021 Conference Proceedings and program, at least one unique registration per paper or poster is required].

Whether this submission is a paper presentation or poster demonstration, your full paper (optional) will be included in the Conference Proceedings if submitted along with the signed Springer consent to publish agreement form by the posted deadline.

We look forward to seeing you in Virtual Conference, Manhattan, New York!

Sincerely,

AHFE 2021 Administration

Questions? Please send to support@ahfe.org
Conference website: http://ahfe.org

b. Autorización de uso de información

INS-PE-087-2021

Quito, 25 de Junio de 2021

Señores
UNIVERSIDAD INTERNACIONAL SEK
Presente.

Carta de Autorización

Yo, **Carlos Andrés Holguín Sánchez** con No. de cédula 171096092-1 en calidad de Representante Legal del Instituto de Investigaciones Socio Económicas y Tecnológicas INSOTEC, con No. de RUC 1790456064001 autorizo a **Juan Freire** con No. De cédula 1720479045, para que utilice la información de Insotec en su proyecto de tesis. Sin embargo, en el documento no se presentará información confidencial de los clientes ni del negocio, únicamente los resultados de efectividad al aplicar las diferentes técnicas de análisis de datos en el conjunto de datos de la organización.

Atentamente,

Econ. Carlos Andrés Holguín
C.I: 171096092-1

Pinta E6-29 y Rábida, Edificio la Pinta, 2do. Piso · Telfs.: (593-2) 254 3260 / 255 0527
Casilla: 17-07-9817 · Quito · Ecuador
www.insotec-ec.com