ISBN: 978-628-95207-4-3. ISSN: 2414-6390. Digital Object Identifier: 10.18687/LACCEI2023.1.1.1148

# Intelligent predictive model applying Data Mining strategies for a credit evaluation of a commercial company.

Grecia-Castañeda Rojas[1], Schleiffer-Canales Carreño[2], Christian Ovalle[3] Erick Humberto Rabanal Chávez[4]
[1,2,3,4]Universidad Privada del Norte, Perú, N00061951@upn.pe, N00059084@upn.pe, denis.ovalle@upn.pe, erick.rabanal@upn.edu.pe

*Abstract-This scientific article presents a predictive model developed by the Orange software, to evaluate the credit capacity of customers, through their transactions of a commercial company, with the aim of preventing delinquency and lack of cash flow. The model is guided by the SEMMA methodology and uses neural network, logistic regression and decision tree algorithms, and its accuracy was measured by performance indicators. The results showed that the decision tree algorithm achieved an accuracy of 99%, which demonstrates the efficiency of the model and the prediction if the client will comply with the payment.*

*In addition, a significant decrease in the time required to assess the creditworthiness of clients was identified after the implementation of the intelligent predictive model. Before the model, 9 human operations were required to assess credit, while after the model it was reduced to only 6 human operations. This translated into a reduction in operating time of 33.33%. In addition, the implementation of the predictive model also made it possible to significantly reduce the time required to complete the first workflow. Before the model, the collection process could take from 60 to 240 days, but after the implementation of the model, the collection time was reduced to only 60 days. In addition, the implementation of the model was also able to completely eliminate delinquent customers, indicating a significant improvement in the company's credit risk management and productivity improvement.*

*Keywords- Predictive model, Process mining, Decision tree, Collections, Machine Learning*

**Digital Object Identifier:** (only for full papers, inserted by LACCEI).
**ISSN, ISBN:** (to be inserted by LACCEI).
**DO NOT REMOVE**

# Modelo predictivo inteligente aplicando estrategias de Minería de Datos para una evaluación crediticia de una empresa comercial.

Grecia-Castañeda Rojas[1], Schleiffer-Canales Carreño[2], Christian Ovalle[3] Erick Humberto Rabanal Chávez[4]
[1,2,3,4]Universidad Privada del Norte, Perú, N00061951@upn.pe, N00059084@upn.pe, denis.ovalle@upn.pe, erick.rabanal@upn.edu.pe

***Resumen-En este artículo científico se presenta un modelo predictivo desarrollado el software Orange, para evaluar la capacidad crediticia de los clientes, a través de sus transacciones de una empresa comercial, con el objetivo de prevenir la morosidad y la falta de flujo de efectivo. El modelo se guía de la metodología SEMMA y utiliza algoritmos de redes neuronales, regresión logística y árbol de decisión, y se midió su precisión mediante indicadores de rendimiento. Los resultados mostraron que el algoritmo de árbol de decisión logró una exactitud del 99%, lo que demuestra la eficacia del modelo y de la predicción si el cliente cumplirá con el pago.***

***Además, se identificó una disminución significativa en el tiempo requerido para evaluar la capacidad crediticia de los clientes después de la implementación del modelo predictivo inteligente. Antes del modelo, se requerían 9 operaciones humanas para evaluar el crédito, mientras que después del modelo se redujo a solo 6 operaciones humanas. Esto se tradujo en una reducción del tiempo de operación del 33.33%. Además, la implementación del modelo predictivo también permitió reducir significativamente el tiempo necesario para completar el primer flujo de trabajo. Antes del modelo, el proceso de cobro podía tomar de 60 a 240 días, pero después de la implementación del modelo, el tiempo de cobro se redujo a solo 60 días. Además, la implementación del modelo también logró eliminar completamente los clientes morosos, lo que indica una mejora significativa en la gestión del riesgo crediticio de la empresa y mejora de la productividad.***

***Palabras Claves- Modelo predictivo, Minería de procesos, Árbol de decisión, Cobranzas, Machine Learning.***

## I. Introducción

En el presente, para las pequeñas, medianas empresas y, a veces, para las empresas más grandes, llegar al usuario final puede ser un problema en algunos casos. Por tanto, las comercializadoras son precisamente las encargadas de hacer de intermediarios entre el fabricante o productor y el consumidor final.

Según [1], en el país las empresas del rubro comercial representan el 46,69%. No obstante, una de las operaciones para ampliar la cartera de clientes es el otorgamiento de créditos a las compañías, lo cual es una actividad con alto riesgo, ya que existe una alta morosidad evidenciada de las empresas con sus proveedores, solo a mayo de 2022, unos 8,5 millones de peruanos tienen deuda morosa por un total de S/ 30.900 millones, según el último Informe de Morosidad de Equifax. Para Brachfiel, la morosidad surge del incumplimiento de obligaciones de pago que impactan negativamente en el desempeño financiero de la empresa emisora de crédito [2].

De la misma forma, el riesgo de crédito son pérdidas por incumplimiento de obligaciones contractuales (créditos o bonos), mora por diversos factores, que pueden incluir fluctuaciones repentinas en los mercados de activos financieros, situaciones de falta de liquidez, falta de ejecución de garantías o cobros, según [3].

Por lo tanto, el objetivo de este trabajo es realizar una simulación que permita la predicción inteligente al momento de realizar el proceso de evaluación de ampliación de créditos, para que así se evite la pérdida de liquidez en los negocios.

Como es conocido, un proceso de negocio consiste en un conjunto de actividades que se realizan para cumplir con los objetivos de este mismo. Estas actividades están integradas en un marco organizacional y técnico que aportan una ejecución eficiente. La gestión de estos procesos de negocio es fundamental porque asegura y permite la identificación de los recursos utilizados. Actualmente los sistemas de información están cada vez más interconectados con las operaciones de los negocios, donde se registran datos valiosos. Sin embargo, las empresas desconocen, carecen de herramientas o tienen problemas para extraer el valor útil de estos datos, lo cual no les permitirá tener una mejor toma de decisiones para sus negocios.

Según estudio de IDC Latinoamérica, el uso de los datos se convirtió en uno de los puntos clave en la agenda de los CEO´s y de los principales ejecutivos de las industrias, lo cual motivó a las empresas peruanas a incrementar en más de 30% su presupuesto para iniciativas con tecnologías como Big Data & Analytics, entre otras.[4]

En el mercado peruano, las empresas están adoptando cada vez más la Inteligencia Artificial como una medida para reducir costos y automatizar procesos. Esta tendencia hacia la transformación digital ha impulsado la implementación de herramientas y técnicas basadas en algoritmos que automatizan las actividades empresariales. Una de las técnicas fundamentales es el Data Mining, que permite analizar grandes volúmenes de datos para descubrir patrones y tendencias

valiosa. La minería de procesos, una técnica dentro del Data Mining, se enfoca en analizar los procesos y brindar a las organizaciones la capacidad de explorar y comprender a fondo sus procesos empresariales. analizar los registros de actividades y el flujo de trabajo, para identificar ineficiencias, cuellos de botella y áreas de mejora en los procesos. Esto permite tomar decisiones informadas para optimizar los procesos y mejorar el rendimiento.

Por consiguiente, en este trabajo se propone el uso de las técnicas del data mining y se formuló la siguiente pregunta *¿Cómo el modelo inteligente predecirá la evaluación de créditos a futuro?*

## II. Estado del arte

En un estudio [5], donde solicitaron los datos a la cooperativa multipropósito de Filipinas, se extrajo 1000 instancias de conjunto de datos de los cuales 900 se destinaron para entrenamiento y 100 para la predicción. Su objetivo principal fue emplear distintas técnicas útiles de minería de datos para la predicción de incumplimiento de préstamos. Se utilizó en la fase de entrenamiento J48 con un factor de confianza 0.5, donde se obtuvo una alta precisión del 76.85%. Los k vecinos más cercanos (k-NN) obtuvieron los resultados más altos (78.38%) en las variantes de IBK, Naive Bayes cuya precisión fue de 76.65% y logística que tiene una precisión de 77.31% k-NN 3. La implantación de estos algoritmos en el conjunto de prueba dio 48 no morosos y 52 morosos. Por otra parte, la logística arrojó 44 no morosos y 56 morosos. Dicho estudio pudo implementar diferentes algoritmos de minería de datos supervisados y no supervisados para identificar el mejor clasificador para un conjunto de datos, el cual fue el J48.

Por otro lado, [6] nos plantean en su revisión sistemática para identificar los métodos y tecnologías de minería de datos utilizados en el contexto de la planificación y programación de la producción, los tres enfoques principales para optimizar un taller de producción utilizando la minería de datos, los cuales son: clasificación basada en la metodología de trabajo, clasificación basada en CPS y clasificación basada en técnicas de datos, para optimizar el piso de producción en el contexto de la industria 4.0.

### *A. Minería de procesos*


Fig.1 La minería de procesos reside en la intersección de la ciencia de datos y la ciencia de procesos.

Van der Aalst [7], nos explica que la minería de procesos es un campo de investigación donde se encuentra la inteligencia computacional y la minería de datos, por un lado, y el modelo y análisis de procesos, por el otro. El propósito es encontrar, monitorear y optimizar procesos reales a través de la extracción de conocimiento de registro de eventos.

### *B. Minería de datos*

Peréz [8], nos muestra que la minería de datos implica el análisis y la interpretación automática de comportamientos, patrones, tendencias, predicciones y otras funciones inteligentes integradas en los datos.

### *C. Metodología SEMMA*

El método SEMMA (Sample, Explore, Modify, Model and Assess), fue propuesta por el Instituto SAS en 2012, lo define como un proceso de muestrear, explorar, modificar, modelar y evaluar, aplicado a una gran cantidad de datos almacenados, que posibilita el descubrimiento de patrones como herramienta de apoyo al negocio. Según SAS, SEMMA es más que un método de minería de datos, es un conjunto de herramientas funcionales que se enfocan en los aspectos de autodesarrollo de los modelos de minería.

A continuación, se describe las cinco fases de la metodología:

**Muestrear:** En esta etapa se obtienen muestras de los datos que sean representativas para el análisis, pero de tamaño adecuado para poder manipularlas en un tiempo y con unos recursos razonables.

**Explorar:** Análisis preliminar de los datos, obteniendo unas primeras conclusiones sobre su morfología, tendencias, etc., para ayudarnos a decidir qué camino seguir.

**Modificar:** En esta fase se modifican los datos, se aplican transformaciones y realizan selecciones para crear las variables ya orientadas al proceso de selección del modelo.

**Modelar:** Aplicando Modelos de Minería de Datos, se obtienen funciones o combinaciones de las variables de elegidas como predictoras, que nos ayudan a predecir la variable objetivo.

**Evaluar:** En esta última fase se evalúa la utilidad y fiabilidad de los insights obtenidos con el Modelo, y se estima su rendimiento.

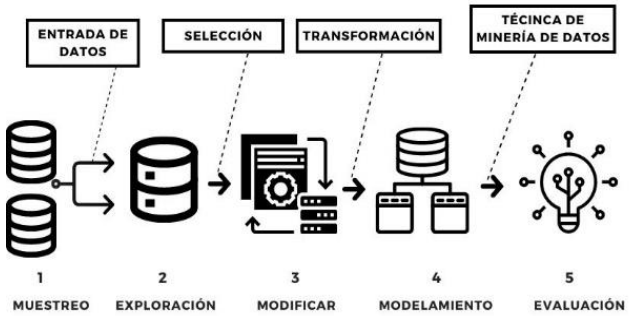


Fig.2 Visión general de los pasos de la metodología SEMMA

*D. Inteligencia artificial*

Rouhiainen [9], nos dice que la inteligencia artificial es la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones, ya que pueden analizar grandes volúmenes de información a la vez. Asimismo, la proporción de errores es significativamente menor en las máquinas que realizan las mismas tareas que sus contrapartes humanas.

*E. Machine learning*

El machine learning o aprendizaje automático es la rama de la inteligencia artificial que dota a las máquinas la habilidad de "aprender" a partir del análisis con la capacidad de identificar patrones en datos masivos, elaborar predicciones (análisis predictivo**)** y apoyar en la toma decisiones con la mínima intervención humana; personas y máquinas trabajan de la mano.

*F. Algoritmos de aprendizaje automático supervisados y no supervisado*

En el aprendizaje automático se emplea dos tipos de técnicas: el aprendizaje supervisado, donde un algoritmo usa datos "etiquetados", para encontrar una función que, dadas las variables de entrada les proporcione las etiquetas de salida correspondiente. El algoritmo se entrena con un "histórico" de datos y así "aprende" a asignar la etiqueta de salida adecuada a un nuevo valor, es decir, predice el valor de salida.

Por el contrario, el aprendizaje no supervisado está entrenando un modelo con datos sin procesar y sin etiquetar. Como sugiere el nombre, el aprendizaje automático no supervisado no requiere tanta intervención humana en comparación con el aprendizaje supervisado. Se requiere una persona para establecer los parámetros del modelo, como la cantidad de puntos de clúster, pero el modelo puede manejar grandes conjuntos de datos de manera eficiente sin supervisión humana.

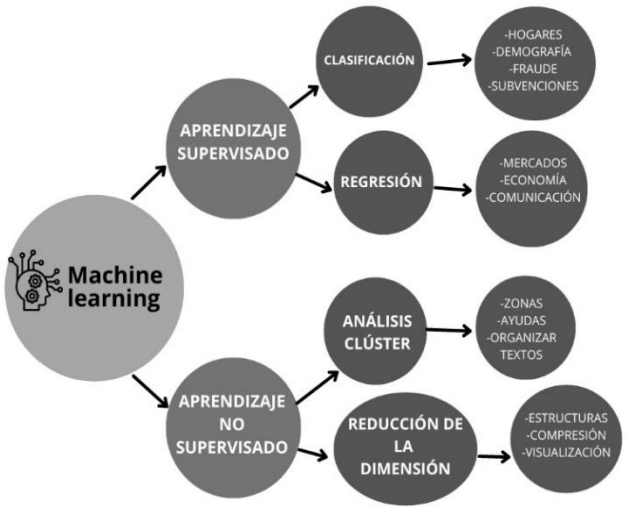


Fig. 3 Entre las técnicas de aprendizaje automático se incluyen el aprendizaje supervisado y el aprendizaje no supervisado

*G. Software de minería de datos Orange*

Orange es un programa informático para realizar minería de datos y análisis predictivo desarrollado en la facultad de informática de la Universidad de Ljubljana. Consta de una serie de componentes desarrollados en C++ que implementan algoritmos de minería de datos, así como operaciones de preprocesamiento y representación gráfica de datos.

*H. Árbol de decision*

Es un algoritmo que sirve para clasificar información y, más adelante, evaluar los diferentes escenarios. El concepto de árbol de decisión describe que es un modelo predictivo basado en los posibles resultados de elegir alternativas. En otras palabras, un árbol de decisiones es un mapa de los posibles resultados de una serie de decisiones interrelacionadas.

*I. Matriz de Confusión*

Una matriz de confusión, también conocida como matriz de error, es una tabla resumida que se utiliza para evaluar el rendimiento de un modelo de clasificación. El número de predicciones correctas e incorrectas se resumen con los valores de conteo y se desglosan por cada clase.

Herramienta para analizar el rendimiento de algoritmos de aprendizaje supervisado.


Fig. 4 Matriz de confusión binaria

A estas 4 selecciones se conocen como la matriz de confusión.

- Verdadero positivo: El valor real es positivo y la prueba predijo que era positivo.
- Verdadero negativo: El valor real es negativo y la prueba predijo asimismo que el resultado era negativo
- Falso negativo: El valor real es positivo, y la prueba predijo que el resultado es negativo.
- Falso positivo: El valor real es negativo y la prueba predijo que el resultado es positivo

*J. Vista Minable*

Una Vista Minable es la conso- lidación en una única tabla de todas las observaciones y los atributos sobre los que se aplicarán los algoritmos de minería de datos.

*K. Análisis Predictivo*

El análisis predictivo es el estudio de datos actuales e históricos para predecir el futuro. Se utiliza una mezcla de técnicas matemáticas, estadísticas y de machine learning avanzadas para analizar los datos y así determinar y extrapolar las tendencias ocultas.

## III. METODOLOGÍA

La presente investigación se sitúa en el tipo aplicada con un enfoque mixto, cuenta con un diseño cuasiexperimental con un alcance predictivo para el cual se utilizará la técnica del modelo matemático y estadístico del algoritmo.

Sampieri [10], nos dice que un diseño cuasiexperimental tiene como objetivo probar hipótesis causales manipulando (al menos) una variable independiente, donde las unidades de estudio no pueden asignarse aleatoriamente a grupos por razones logísticas o éticas.

Se plantea implementar un modelo inteligente basado utilizando la herramienta minería de procesos, para predecir la evaluación de créditos.

Para diseñar el modelo inteligente de predicción, se procede con los siguientes pasos

### *A. Análisis comparativo de metodologías*

En el análisis detallado de las metodologías de la data mining empleadas en estudios anteriores, se encontraron metodologías de minería de datos, como KDD, SEMMA y CRISP-DM, por lo cual se procedió a realizar la investigación de cada una y un análisis comparativo, para determinar cuál se utilizará para el modelo.

TABLA 1

COMPARACIÓN DE GENERALIDADES DE CADA METODOLOGÍA.

| | KDD | SEMMA | CRISP-DM |
|---|---|---|---|
| Enfoque | Orientado a la identificación de patrones más favorables para cierta tarea. | Orientado al desarrollo del proceso de minería de datos. | Orientado a los objetivos empresariales. |
| Uso | Productos enfocados en patrones de datos | Ligado a productos SAS | Metodología abierta y gratuita |
| Metodología | De patrones arquitectónicos orientados a datos. | Metodología aún no definida | Metodología de gestión de proyectos |
| Complejidad | Más complejo de implementar que los otros dos, tiene una cantidad de considerable de fases a desarrollar. | Simple y bastante ágil, sus fases están más implementadas a desarrollo ágil. | Es el menos complejo de entender y aplicar, cuenta con una curva de adaptabilidad muy amplia para cualquier desarrollador. |
| # de fases de desarrollo | 9 | 6 | 6 |
| Siglas | Knowledge Discovery in Dtabases | Sample, Explore, Modify, Model and Access | Cross.industry Standard Process |
| Relevancia | Baja | Media | Alta |

Realizado el análisis entre estas tres metodologías de minería de datos, se determinó que para el desarrollo de este trabajo se utilizará la metodología SEMMA, porque se centra más en las características técnicas del desarrollo del proceso, es decir es más puntual sobre el proceso, ya que se tiene analizado el problema de la empresa.

En el estudio [11], introdujo la clásica metodología SEMMA, la cual indica su importancia, ya que ayudó a analizar las características de comunicación de la red sobre la base de registros de tráfico y registros de base de datos.

### *B. Modelo predictivo para la evaluación de créditos*

Para diseñar el modelo inteligente de predicción, se elaborará los siguientes pasos:

Paso 1: Recolectar la información del ERP de la empresa sobre las transacciones de ventas y clientes.

Paso 2: Preparación y transformación de datos y variables con el fin de obtener la vista minable para el modelo,

Paso 3: En el software Orange se enlazarán y aplicarán a la vista minable los algoritmos de predicción (árbol de decisión, regresión logística y redes neuronales), con el fin de tener una proyección del rendimiento de cada uno.

Paso 4: Se aplicarán métricas de rendimiento a los algoritmos trabajados, con la finalidad de determinar y elegir el que tenga una mejor precisión en la predicción de evaluación de créditos.

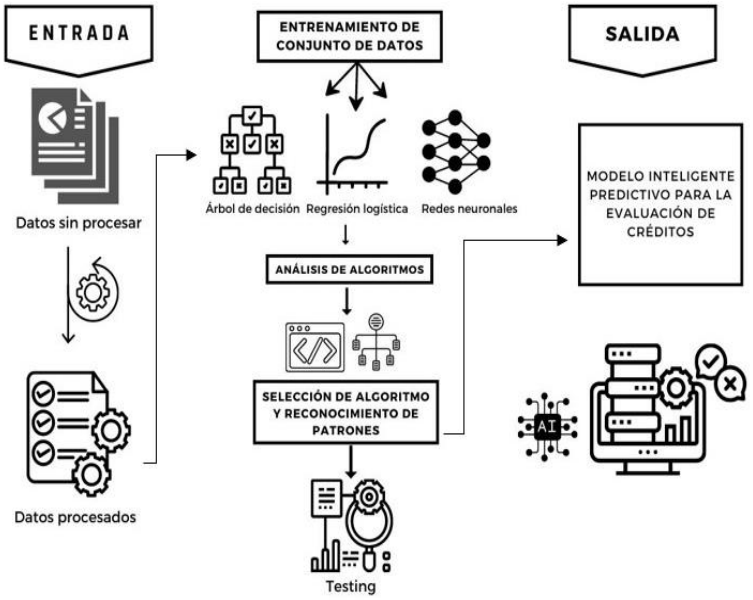


Fig.5 Modelo predictivo inteligente para la evaluación de créditos

### *C. Desarrollo de la metodología SEMMA*

En tal sentido, los pasos para el desarrollo de este trabajo que busca implementar un modelo predictivo para la evaluación de ampliación de créditos, utilizando la metodología SEMMA y las estrategias de minería de procesos, que además ayudará cuantificar el impacto en las lagunas de este proceso, son los siguientes:

1) Recopilación y muestreo de datos:

Del software ERP STARSOFT GE, se reúne la información relevante sobre los clientes y sus pagos, como historial crediticio, ingresos, gastos, días de mora, etc.

- Se extrajo el reporte de ventas de todo el año 2022
- Se extrajo el reporte de cuentas por cobrar hasta el 31 de diciembre del 2022
- Se hizo un cruce de datos con la información de ambos reportes
- Se realizo un Excel con el cruce de ambos reportes donde se identificaron los siguientes elementos:

TABLA 2

RESUMEN DE LA RECOLECCIÓN DE INFORMACIÓN OBTENIDA EN LA BASE DE DATOS DE LA EMPRESA.

| Atributo | Tipo | Valores |
|---|---|---|
| Días de mora | Discreta | -1 al -350 |
| Fecha de vencimiento | Fecha | 01/01/2021 al 31/12/2021 |
| Estado | Categórica | Pagado - No pagado |
| Tipo de pago | Categórica | Crédito - Al contado |
| Tipo de documento | Categórica | Factura - Boleta |
| Serie | Entero | Numeral |
| Fecha de emisión | Fecha | 01/01/2021 al 31/12/2021 |
| Nombre del cliente | Cualitativa | |
| Rubro | Categórica | Alimenticio, Agroexportación, Aeronáutica, Arquitectura, Comercial, Construcción, Educación, Eléctrico, Electromecánico, Ensayos técnicos, Industrial, Inmobiliario, Logística, Manufactura, Mecánica, Saneamiento, Seguridad, Tecnología, Turismo, Siderúrgica, Metalmecánica, Textil y Transporte. |
| IGV. dólares | Cuantitativa. N | |
| Vta. dólares | Cuantitativa. N | |
| Total dólares | Cuantitativa. N | |
| Tipo de cambio del día | Cuantitativa. N | |
| IGV. soles | Cuantitativa. N | |
| Vta. soles | Cuantitativa. N | |
| Total, soles | Cuantitativa. N | |

2) Exploración y limpieza de datos:

Después, se observa que hay 7443 registros almacenados en la base de datos de la cartera de la empresa y se divide en clientes que pagan a crédito y al contado, compuesta de 16 variables.
Luego, se evaluó la relación de las variables, para que permita construir el modelo y se realizó uno de los puntos más importantes que es la eliminación y limpieza de la información obtenida, donde se procedió a eliminar los datos (transacciones) atípicos que no aportaban valor para el estudio, como por ejemplo las ventas a clientes que no se identificaron y los datos con valores nulos, que pudo ser debido a la falla en el momento del llenado de información del cliente.


Fig. 6 Muestra de las variables eliminadas para la realización del estudio.

3) Selección y transformación de variables:

A continuación, es te paso se observa que con la eliminación de las columnas que generaban ruido y no aportaban valor para la construcción de la vista minable y la limpieza de transacciones de "personas naturales", que no ayudaban en la generación de patrones, nuestra muestra quedó con 5000 registros de datos para el estudio. También, se determinó que el estudio no va dirigido al cliente si no a las transacciones de cada uno, lo cual se estableció que ‘'Tipo de cliente‘, ‘'Total dólares", "Total soles", ‘’ Nombre del cliente’’, "Días de mora" y " Estado de cuenta" serían los candidatos para la formación de nuestra vista minable. Por otro lado, se creó una categoría de estado que indica si el cliente "paga" o "no paga", para la construcción del modelo predictivo, a la vez, para encontrar patrones para el análisis se procedió a transformar, agrupar y a clasificar las variables, dejando los siguientes datos:

TABLA 3

DESCRIPCIÓN DE LAS VARIABLES MODIFICADAS PARA EL ESTUDIO.

| | Variable | Función |
|---|---|---|
| Modelado de variables para el análisis | Tipo de cliente | Clasifica a los clientes y sus cuentas si son de crédito o al contado. |
| | Venta en dólares | Esta variable nos indica si el cliente pagó en dólares o no. |
| | Nivel total en soles de venta | Esta variable se le transformo a una variable ordinal por categoría y se le determinó:<br>-Si va 0 a 10000 es bajo.<br>-Si va 10001 a 20k es mediano.<br>-Si va 20001 en adelante es de categoría alto. |
| | Días de mora | Determina los días que lleva un cliente por no cancelar su cuota en la fecha establecida. |
| | Estado | Indica si el cliente ‘’paga’’ o ‘’no paga’’ la cuenta. |
| | Nombre del cliente | Indica el nombre de cada empresa con la que se trabaja. |

4) Modelamiento:

Seguidamente, para desarrollar la construcción de nuestro modelo utilizamos el software Orange, ya que, sus herramientas y funcionalidades que son: Procesamiento de datos, modelos predictivos, métodos de descripción de datos, gráficos y validación del modelo, permite realizar un estudio confiable. Por ejemplo, unas de las principales funciones del software son que:

- Los usuarios pueden crear sus propios flujos de trabajo interactivos con el objetivo de analizar y visualizar los datos con mayor amplitud.
- Permite rediseñar y adaptar la herramienta a las necesidades del usuario y/o de la empresa.
- La visualización de la información puede realizarse en distintos formatos, diagramas de dispersión, gráficos de barras, árboles o redes y mapas de color, lo cual permite mostrar con mayor claridad los resultados para interpretar de mejor forma la información.

Entonces, para la construcción del análisis con las 6 variables de la vista minable descritas anteriormente, se realiza los siguientes pasos en el software:

Primero: Se selecciona "Abrir archivo" en la barra de herramientas, se busca el archivo deseado en la computadora y se selecciona. Luego, se da clic en "Abrir” y se carga el archivo Excel con los datos de las variables ya modificadas para la creación de modelo predictivo en el software Orange.

Segundo: Se determina el ´´Target´´ que es la variable objetivo que se desea predecir. Se determinó que la variable estado de cuenta es el target y nos tiene que dar un resultado, ya sea si el cliente "Paga" o "No paga" la transacción. También, "Nombre del cliente" se pone como "meta", ya que no aporta mucho en generar patrones, pero si ayuda a identificar al cliente.

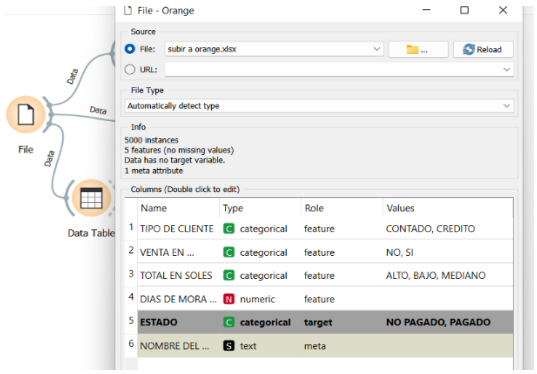


Fig. 7 Muestra de la selección de variables objetivo de estudio en el software Orange.

Tercero: Para la aplicación de algoritmos en nuestra vista minable en el software, se determinó que se evaluará con los métodos: Regresión logística, árbol de decisión, redes neuronales y K- Means, ya que, el software acepta nuestros datos con estos métodos.

TABLA 4

RESUMEN DE DESCRIPCIÓN DE ALGORITMOS QUE SE UTILIZARON PARA LA CONSTRUCCIÓN DEL MODELO

| Tipo de aprendizaje | Algoritmo | Fórmula del algoritmo |
|---|---|---|
| Supervisado | **Regresión logística:** La regresión logística es un algoritmo de clasificación binaria supervisada en el aprendizaje automático. | P(Y)=11+e−(b0+b1X1)<br>Donde:<br>y: Variable dependiente o a predecir<br>x: Variable independiente<br>a: Es la pendiente<br>b: La constante que debe ser determinada. |
| | **Árbol de decisión:** Se trata de dividir el conjunto en subconjuntos homogéneos, basados en ganancia de información. | Los pasos son: Seleccionar, dividir, repetir y asignar.<br>Fórmula de ganancia de información:<br>H(S) - ∑(S_i/S) * H(S_i). |
| | **Redes neuronales**: Son modelos que utilizan capas de neuronas interconectadas para procesar datos y realizar predicciones complejas. | Los pasos son: Entrada, propagación, activación, salida, error y ajuste.<br>La fórmula general de una neurona es:<br>y = f(∑(w_i*x_i) + b). |
| **No supervisado** | **K- Means:** Es un algoritmo de clustering no supervisado que agrupa un conjunto de datos en K grupos. | Los pasos son: Asignar puntos de datos al centroide más cercano, mover los centroides al centro de los puntos asignados y repetir hasta que los centroides converjan. |

De la misma forma, en la figura 8 se puede observar los widgets empleados y la secuencia de enlaces del modelo. Lo cual, para la aplicación de estos métodos se siguió los siguientes pasos: Selección de los widgets en "modelo" (regresión lineal, árbol de decisión y redes neuronales), luego seleccionamos en las herramientas " sin supervisar" el widget ‘'K-Means" y todos los widgets seleccionados los enlazamos con el widget "archivo" (el que se importó primero al software). También, en las herramientas del software "visualizar" seleccionamos los widgets " visor de árboles", donde se enlaza con árbol de decisión y " gráfico de dispersión" se enlaza con K- Means, para que represente gráficamente los datos que se están analizando y facilite la comprensión de ello.

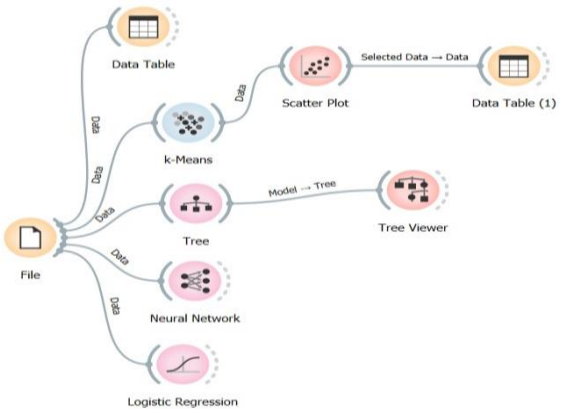


Fig. 8 Muestra de la selección de algoritmos y visualizaciones del modelo predictivo.

Cuarto: Por último, para evaluar el rendimiento del modelo de aprendizaje automático y analizar la precisión de la predicción del modelo seleccionamos en las herramientas de "evaluar" los widgets "prueba y puntuación" y "matriz de confusión", a la vez enlazamos los widgets "regresión lineal, árbol de decisión, redes neuronales y archivo" con el widget "prueba y puntuación", el cual se termina enlazando con el widget "matriz de confusión".

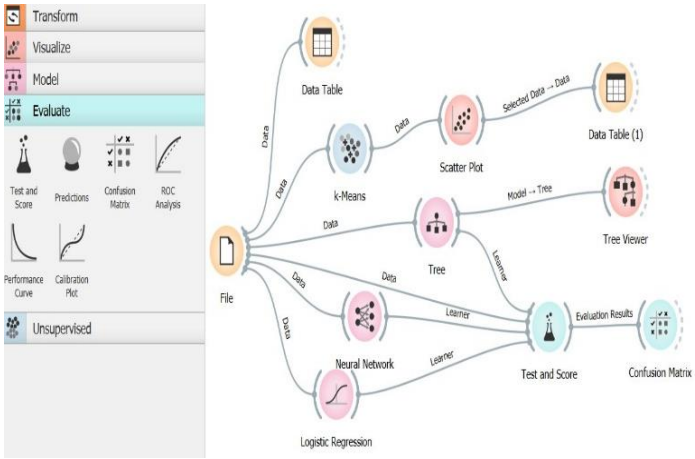


Fig. 9 Muestra de la construcción del modelo predictivo.

## IV. RESULTADOS

### *A. Resultados obtenidos de los algoritmos en el software*

En la figura 10 se evidencia que en la evaluación del modelo con "prueba y puntuación" del software, nos da como resultado el nivel de precisión e indica que se debe utilizar el algoritmo de árbol de decisión que es de 99%, por lo que evidencia un alto nivel para realizar la predicción al querer otorgar créditos.

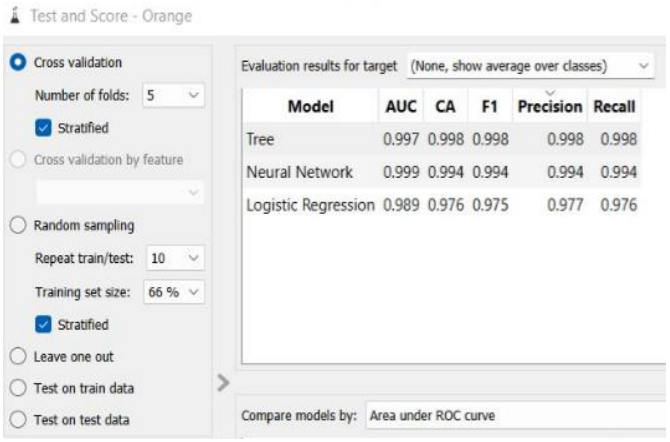


Fig. 10 Muestra de los resultados de la evaluación de algoritmos para el modelo predictivo en el software Orange.

TABLA 5

INDICADORES DE RESULTADO DE LOS TRES ALGORITMOS

| | | | |
|---|---|---|---|
| PRESICIÓN | **99.8%** | **99.4%** | **97.7%** |
| AUC | **99.7%** | **99.9%** | **98.9%** |
| EXHAUSTIVIDAD | **99.8%** | **99.4%** | **97.6%** |

En base a los resultados, al comparar los tres modelos, se puede notar que el árbol de decisión tiene un mejor rendimiento que los otros dos modelos. Esto se puede inferir porque en la Figura 11 el gráfico de la curva ROC está más cerca del borde superior izquierdo y su AUC es el más alto de los tres modelos. Esto sugiere que el árbol de decisión es el mejor modelo para clasificar los casos positivos y negativos en general. Sin embargo, la red neuronal y la regresión logística también tienen un buen rendimiento, su AUC es ligeramente inferior en comparación con el árbol de decisión, sin embargo, es importante tener en cuenta que la elección del mejor modelo dependerá de la aplicación específica y los requisitos de rendimiento deseados.

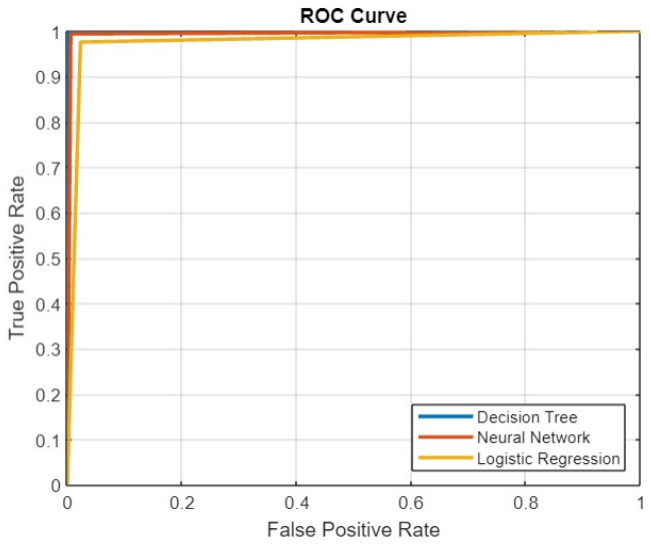


Fig. 11 Desempeño de los algoritmos considerados en la Curva Roc

Por otro lado, como se muestra en la tabla 6, se calculó matemáticamente la exactitud y la exhaustividad de los tres algoritmos utilizados en el modelo predictivo. La exactitud mide la proporción de predicciones correctas, mientras que la exhaustividad mide la proporción de casos positivos correctamente identificados.
Donde:
Exactitud = VP +VN / VP + FN + FP + TN
Exhaustividad = VP /VP +FN

TABLA 6

INDICADORES DE RESULTADO OBTENIDO MATEMATICAMENTE

| Resultados de exactitud y exhaustividad de los algoritmos | | | |
|---|---|---|---|
| Redes neuronales | | | |
| VP | 95,2 | Nivel de exactitud | 0,976 |
| FN | 4,8 | | |
| FP | 0 | Nivel de exhaustividad | 0,952 |
| VN | 100 | | |
| Regresión logística | | | |
| VP | 82 | Nivel de exactitud | 0,910 |
| FN | 18 | | |
| FP | 0 | Nivel de exhaustividad | 0,820 |
| VN | 100 | | |
| Árbol de decisión | | | |
| VP | 98,1 | Nivel de exactitud | 0,994 |
| FN | 1,2 | | |
| FP | 0 | Nivel de exhaustividad | 0,988 |
| VN | 100 | | |

De la Figura 12, se puede determinar que los valores de 98.8% y 100% corresponden a los valores predictivos de forma correcta por el modelo tanto los verdaderos positivos como los verdaderos negativos. Mientras que los valores 1.2% y 0.0% corresponden a los casos en los que la prueba predijo de manera erróneas.

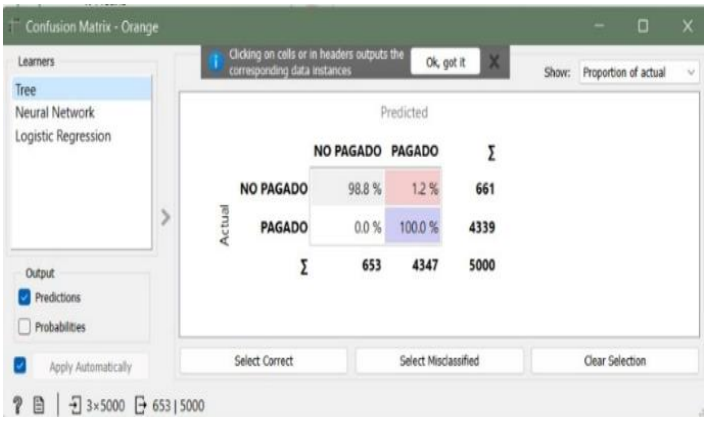


Fig. 12 Muestra de los resultados de la matriz de Confusión del árbol de decisión.

Se obtuvo un nivel de exhaustividad del 99% lo cual demuestra que la cantidad de verdaderos positivos que el modelo a clasificado en función del número total de valores positivos ha sido determinado de manera correcta

Por consiguiente, para la demostración del entrenamiento del modelo predictivo se selecciona en evaluar el widget "predicciones" y se enlaza con los widgets "archivo" y "árbol de decisión", ya que en la evaluación de algoritmos nos indica que se debe trabajar con árbol de decisión por una mayor confiabilidad

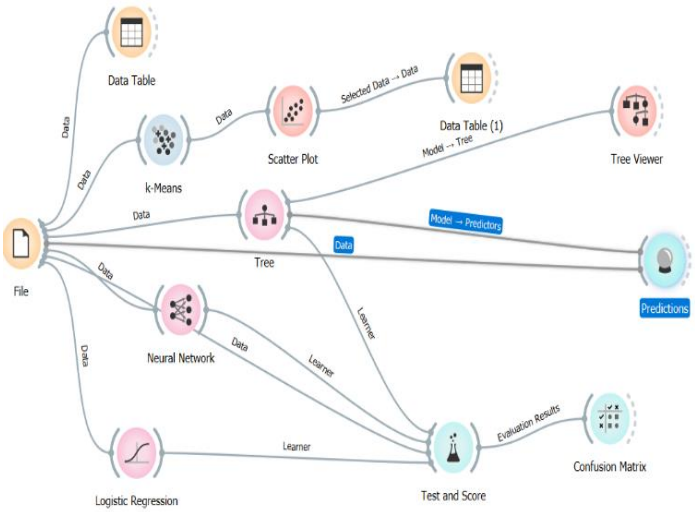


Fig. 13 Muestra de la selección de widgets y enlaces del entrenamiento del modelo predictivo.

Como muestra, en la figura 14 al darle clic en predicciones nos abre una ventana y podemos observar el comportamiento del modelo predictivo, donde la columna "tree o árbol" nos indica si el cliente por sus transacciones pagará o no pagará el crédito

Predictions - Orange

Show probabilities for: Classes in data — Show classification errors

| | Tree | error | ESTADO | OMBRE DEL CLIEN | TIPO DE CLIENTE | ENTA EN DÓLARE | TOTAL EN SOLES | DE MORA (AL 31 / |
|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | ALS PERU S.A. | CREDITO | NO | BAJO | 214 |
| 2 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | MASTER DRILLI... | CREDITO | SI | BAJO | 37 |
| 3 | 0.02 : 0.98 → PAGADO | 0.023 | PAGADO | FIBRAFIL S.A. | CREDITO | NO | BAJO | 0 |
| 4 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | LININGS S.A. | CREDITO | NO | BAJO | -25 |
| 5 | 0.02 : 0.98 → PAGADO | 0.023 | PAGADO | ALS PERU S.A. | CREDITO | NO | BAJO | 0 |
| 6 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | LININGS S.A. | CREDITO | NO | BAJO | 58 |
| 7 | 0.02 : 0.98 → PAGADO | 0.021 | PAGADO | MASTER DRILLI... | CREDITO | SI | BAJO | 0 |
| 8 | 0.02 : 0.98 → PAGADO | 0.021 | PAGADO | MASTER DRILLI... | CREDITO | SI | BAJO | 0 |
| 9 | 0.02 : 0.98 → PAGADO | 0.021 | PAGADO | MASTER DRILLI... | CREDITO | SI | BAJO | 0 |
| 10 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | ALS PERU S.A. | CREDITO | NO | BAJO | -15 |
| 11 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | LININGS S.A. | CREDITO | NO | BAJO | -26 |
| 12 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | MOLINO EL TRI... | CREDITO | NO | BAJO | 181 |
| 13 | 0.02 : 0.98 → PAGADO | 0.021 | PAGADO | MASTER DRILLI... | CREDITO | SI | BAJO | 0 |
| 14 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | MOLINO EL TRI... | CREDITO | NO | BAJO | 181 |
| 15 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | LININGS S.A. | CREDITO | NO | BAJO | 21 |
| 16 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | LININGS S.A. | CREDITO | NO | BAJO | 78 |
| 17 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | LININGS S.A. | CREDITO | NO | BAJO | 62 |
| 18 | 0.02 : 0.98 → PAGADO | 0.023 | PAGADO | ALS PERU S.A. | CREDITO | NO | BAJO | 0 |
| 19 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | MOLINO EL TRI... | CREDITO | NO | BAJO | 133 |
| 20 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | MOLINO EL TRI... | CREDITO | NO | BAJO | 181 |
| 21 | 1.00 : 0.00 → NO PAGA... | 0.000 | NO PAGADO | MOLINO EL TRI... | CREDITO | NO | BAJO | 114 |

Fig. 14 Pequeña muestra de los datos arrojados por el modelo predictivo en el software.

solicitado y a su costado en la columna "error" indica si la predicción fue mayor o menor que el valor real en cada dato.

Para mejorar la visualización de los datos importados en el software, se da clic en el widget "gráfico de dispersión" o "Scatter plot", donde se ubica el eje vertical el tipo de cliente y en el eje horizontal los grupos de "Pagado", "no pagado" y en el tercero la mezcla de ambos. Además, nos muestra con el símbolo " X" que los primeros dos grupos están las transacciones nivel bajo y el grupo 3 o "C3" están las transacciones nivel alto, mediano y bajo.

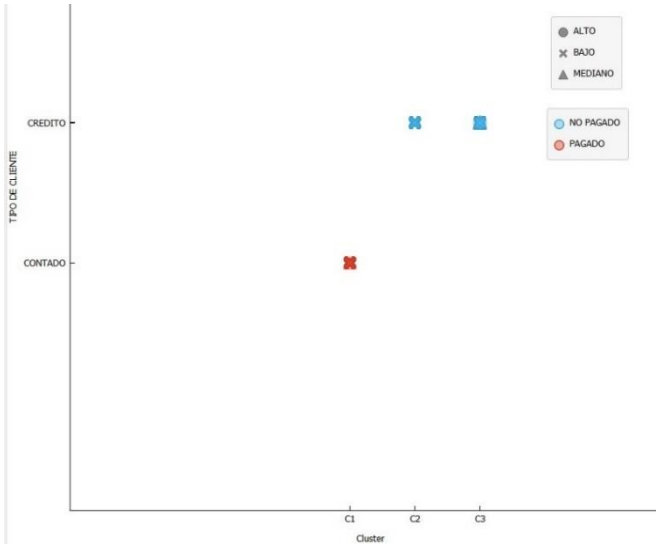


Fig. 15 Gráfico de dispersión de los grupos del modelo predictivo.

Luego, al visualizar el gráfico de distribución con los datos ya dados de la predicción realizada, nos indica que con la aplicación del modelo se tendrá pocos clientes morosos.

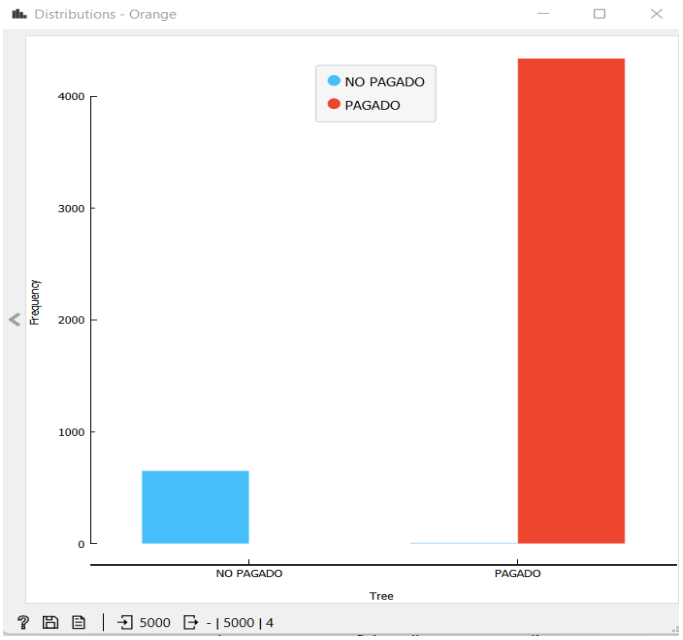


Fig. 16 Gráfico de distribución que indica los resultados de la aplicación del modelo predictivo.

## *B. Resultados obtenidos del flujo del proceso*

El proceso actual del área de créditos y finanzas cuenta de 8 a 9 operaciones dependiendo de la morosidad del cliente. El tiempo que toma este flujo oscila entre 60 a 240 días

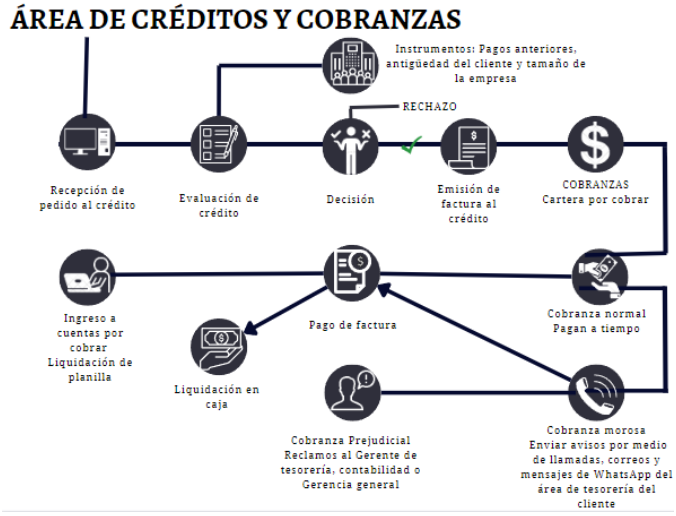


Fig. 17 Flujo del proceso actual de créditos y cobranza

Después, con la implementación del modelo inteligente predictivo en la evaluación de créditos a futuro se seguirá el

siguiente flujo de operación como se muestra en la siguiente figura.

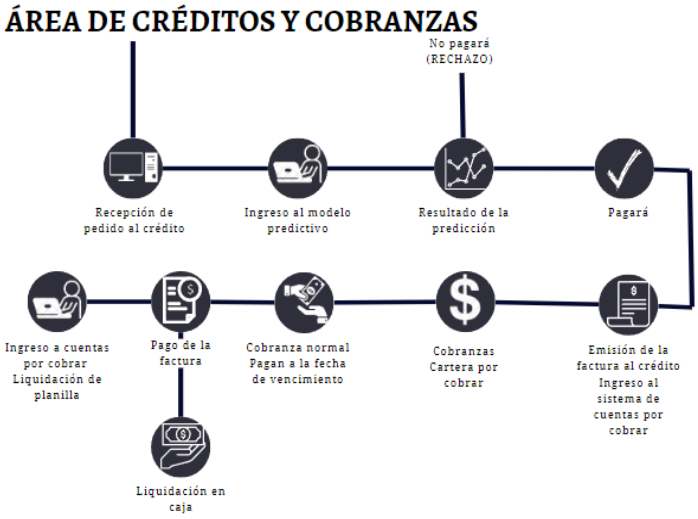


Fig. 18 Flujo del proceso con la aplicación del modelo predictivo.

Como podemos ver en la figura 18, el número de operaciones se redujo a 6 actividades humanas. Logrando reducir en 33.3% las operaciones y tomando un tiempo de 60 días para la liquidación de caja y considerando cero días de mora.

## V. CONCLUSIONES

En esta investigación, se construyó un modelo predictivo utilizando el programa Orange y la implementación del árbol de decisión. Este modelo se basó en 5000 registros de transacciones y obtuvo una precisión del 99%, lo que sugiere que es altamente confiable para predecir con éxito los valores de las variables objetivo en un conjunto de datos desconocido a futuro.

Sumado a esto, con la aplicación del modelo que ha sido altamente efectiva se logró reducir las actividades a solo 6 operaciones humanas. Esto tendrá un gran impacto en la productividad y eficiencia del proceso, al reducir los tiempos de espera, los errores, los costos, y mejorar la calidad del servicio final.

Cabe destacar, que la precisión del modelo predictivo puede verse afectada por la calidad y cantidad de datos utilizados en su entrenamiento. Por ello, se recomienda llevar a cabo una validación continua del modelo para asegurarse de que esté actualizado y mantenga su nivel de precisión ante posibles cambios en los datos.

Adicionalmente, es importante tener presente que los resultados obtenidos en este modelo predictivo no necesariamente pueden aplicarse a otras carteras de clientes, ya que estas pueden presentar características y comportamientos de pago distintos. Por tanto, se aconseja realizar pruebas y validaciones adicionales del modelo en nuevas carteras, antes de tomar decisiones importantes basadas en su uso.

## REFERENCIAS


[1] Instituto Nacional de Estadística e Informática (INEI), "Estadísticas del comercio exterior del Perú", presentación en el Panel III del Evento Internacional sobre Estadísticas de Comercio Exterior, agosto 2015. [En línea]. Disponible en: https://unstats.un.org/unsd/trade/events/2015/aguascalientes/9.-Panel%20III%20-%20Presentation%201%20-%20INEI%20Peru.pdf

[2] P. J. Brachfield, Instrumentos Para Gestionar y Cobrar Impagados: Las Herramientas Indispensables Para la Gestión Práctica de Impagados. Barcelona: Profit Editorial, 2012.

[3] A. J. McNeil, R. Frey y P. Embrechts, Quantitative risk management: Concepts, techniques and tools: Revised edition, Inglaterra: Princeton University Press, 2015.

[4] I. Baufest,"Digitalización: Empresas Peruanas Elevaron Inversión En 30% en el último Año", Agencia peruana de noticias,2015. [En línea]. Disponible en: Digitalización: empresas peruanas elevaron inversión en 30% en último año | Noticias | Agencia Peruana de Noticias Andina
[Accedido: 07-feb-2023]

[5] J. C. Alejandrino, J. J. P. Bolacoy y J. V. B. Murcia, "Supervised and unsupervised data mining approaches in loan default prediction", International Journal of Electrical and Computer Engineering (IJECE), vol. 13, no.2, pp 1837-1847,2022, doi: 10.11591/ijece.v13i2.pp1837-1847

[6] P. Martins, Z. Yahouni y G. Alpan. "Literature review on using data mining in production planning and scheduling within the context of cyber physical systems", Journal of Industrial Information Integration, vol.28, no.1, pp. 28-32,2022, doi: 10.1016/j.jii.2022.100371

[7] Van der Aalst, "Data Science", in Process Mining, Ed Berlín: Springer, 2016, pp. 3-23

[8] M. Pérez, Minería De Datos A Través De Ejemplos, Madrid: RC Libros, 2014.

[9] L. Rouhiainen, “Introducción a la Inteligencia Artificial”, en Inteligencia Artificial, L. P. Rouhiainen, Ed. Barcelona: Planeta, 2018, pp. 16-20

[10] R. H. Sampieri, Metodologia de La Investigacion, 6a ed. Mc Graw Hill, Mexico D.F, 2014, pp. 128-129

[11] Z. Ying, W. Song, W. Hao, C. Zepeng y L. Xuejun. "Exploración visual de eventos de seguridad de red basada en SEMMA", Journal of Zhejiang University, vol. 49, no.2, pp. 131-140,2022, doi: 10.3785/j.issn.1008-9497.2022.02.001

[12] W. Premchaiswadi y P. Porouhan, "Process modeling and decision mining in a collaborative distance learning environment", Decision Analytics, vol. 2, no.1, pp. 42-76,2015, doi: 10.1186/s40165-015-0015-5

[13] J. C. Alejandrino, J. J. P. Bolacoy y J. V. B. Murcia, "Supervised and unsupervised data mining approaches in loan default prediction", International Journal of Electrical and Computer Engineering (IJECE), vol. 13, no.2, pp 1837-1847,2022, doi: 10.11591/ijece.v13i2.pp1837-1847

[14] P. M. Seeger, Z. Yahouni y G. Alpan, "Literature review on using data mining in production planning and scheduling within the context of cyber physical systems", Journal of Industrial Information Integration, vol.28, no.1, pp. 20-25,2022, doi: 10.1016/j.jii.2022.100371

[15] D. Kotios, G. Makridis, G. Fatouros y D. Kyriazis, "Deep learning enhancing banking services: A hybrid transaction classification and cash flow prediction approach", Journal of Big Data, vol. 9, no.1, pp. 70-99,2022, doi: 10.1186/s40537-022-00651-x

[16] A. Alonso Robisco y J. M. Carbó Martínez, "Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction", Financial Innovation, vol. 8, no.1, pp. 35-70,2022, doi: 10.1186/s40854-022-00366-1