Universidad Internacional de La Rioja (UNIR)

Escuela de Ingeniería

Máster en Análisis y Visualización de Datos Masivos

Comparativa de técnicas Machine Learning sobre comportamiento de pago de clientes con cuentas por cobrar.

Trabajo Fin de Máster

Presentado por: Pesantez Chuqui, Martha Cecilia

Director/a: García Sánchez, Pablo

Ciudad: Cuenca, Ecuador
Fecha: 19 de Septiembre del 2019

Resumen

El presente proyecto tiene como objetivo identificar el algoritmo con mayor exactitud usando una comparativa de técnicas de *Machine Learning* para detectar el comportamiento de pago de clientes con cuentas por cobrar con la finalidad de ayudar a la gestión de cobranzas, en base a una correcta predicción a focalizar sus recursos y esfuerzos en clientes con mayor riesgo de mora y reducir las acciones de intervención en clientes que si pagarán la próxima cuota mensual del crédito. La metodología utilizada esta basada en *KDD* que parte de obtener el conocimiento de dominio en el área de cobranzas, un análisis exploratorio con técnicas de visualización para identificar el conjunto de datos objetivo totalmente anonimizado, siguiendo con la fase de preparación y pre procesamiento de datos que incluye; limpieza de datos, creación de campos calculados, transformación de variables categóricas en numéricas, selección de atributos entre otros, para aplicar minería de datos que permita encontrar patrones de información y el mejor modelo de predicción. Usando varios algoritmos de clasificación y métricas de evaluación se determina que los algoritmos J48 y Random Forest presentan mejores y similares resultados en rendimiento, precisión y exactitud con la técnica de selección de atributos *CorrelationAttributeEval*, siendo J48 el escogido por el tiempo de ejecución y su tasa de error. En conclusión, el uso de una metodología ha facilitado el entendimiento y desarrollo del presente proyecto, la fase de preparación y pre procesamiento de datos conjuntamente con la técnica de selección de atributos ha sido de gran importancia en la generación del modelo, lo cual ha repercutido en la calidad de desempeño del modelo final, la importancia de usar técnicas de visualización en el análisis exploratorio de los datos y la evaluación de métricas de desempeño ha sido fundamental en la toma de decisiones.

Palabras Clave: cuentas por cobrar, aprendizaje automático, cobranzas, comportamiento de pago, selección de atributos.

Abstract

The following project aims to identify algorithm accurately using a comparison of *Machine Learning* techniques to detect payment behavior of customers with accounts receivable with the purpose of helping the collection management based on a correct prediction focused on its resources and efforts in clients with greater risk of tardiness of payment, and to reduce the actions of intervention in clients that will pay the next pending monthly credit payments on time. The methodology used is based on KDD that begins with obtaining domain knowledge in the area of collections, an exploratory analysis with visualization techniques for identifying the fully anonymized set target data, continuing with the phase of preparation and pre-processing of data that includes; data cleaning, creation of calculated fields, transformation of categorical variables into numerical and feature selection among others, to apply data mining that allows us to find patterns of information and the best model prediction. Using various classification algorithms and evaluation metrics it is determined that the J48 and Random Forest algorithms present better and similar results in performance, precision and accuracy with the technique of *CorrelationAttributeEval* feature selection, with J48 being chosen for its time of execution and its low error rate. In conclusion, the use of a methodology has facilitated the understanding and development of this project, the preparation phase and pre data processing together with the feature selection techniques has been of great importance in the generation of the model, which has had an impact on the quality of performance of the final model, the importance of using visualization techniques in the analysis, data exploration and evaluation of performance metrics has been critical in the decision making.

Keywords: accounts receivable, Machine Learning, collections, payment behavior, feature selection.

Índice de Tablas

Índice de Ilustraciones

Índice de Ecuaciones

Índice de Contenido

1 Introducción

La recuperación de cartera de clientes es una dificultad financiera para las empresas cuando no se la gestiona de forma correcta, por lo tanto es necesario gestionarla de la manera más efectiva, identificando a sus clientes deudores y su comportamiento de pago para mejorar el proceso de cobro. Zeing et al. (2008) señala que "la experiencia a través de múltiples industrias muestran que la gestión efectiva de *AR* (cuentas por cobrar) y el desempeño financiero general de las empresas están positivamente correlacionados". (Zeng, Melville, Lang, Boier-Martin, & Murphy, 2008, pág. 1).

> "Una empresa que no convierte en efectivo sus cuentas por cobrar se queda sin los recursos suficientes para el correcto funcionamiento de sus ciclos operativos de producción y venta, lo cual puede, por una parte, conducirla hacia la escasez de recursos y detener sus ciclos operativos, y por la otra, a multiplicar los clientes deudores". (Morales Castro & Morales, 2014, pág. 144)

La competitividad obliga a las empresas a desarrollar políticas flexibles de ventas a crédito para ganar participación en el mercado. Sin embargo hay varios factores que pueden afectar a que el cliente no pague regularmente las cuotas de su crédito, desde la falta de empleo hasta la cultura del no pago.

Por lo tanto, dichos deudores tienen dificultades para liquidar sus deudas y el departamento de créditos y cobranzas de la empresa debe adoptar sistemas de recaudo de pagos, como el contacto telefónico y domiciliario de los clientes, segmentando a los clientes en base a su tiempo de morosidad y numero de pagos vencidos.

Hoy en día las nuevas segmentaciones se consideran en base al comportamiento histórico del cliente y con ello se busca predecir su conducta de pago. Una vez identificado dichos segmentos se busca definir estrategias más adecuadas para la recuperación de cartera en base a su probabilidad de pago (Deloitte, 2012).

Por otra parte el crecimiento exponencial de los datos gracias a las nuevas tecnologías y a la frecuencia de su uso se vuelve un reto cada vez mayor en las empresas que buscan una mejora competitiva y permanencia en el mercado, ya que requieren de métodos capaces para extraer conocimiento útil y valioso de grandes volúmenes de información, como el análisis predictivo para construir sistemas que permitan aprender de los datos para identificar patrones y predecir resultados futuros con una mínima intervención humana (Umaquinga, Saltos, & Peluffo, 2017).

Durante el desarrollo del presente proyecto se ha seguido una metodología de trabajo basado en *KDD* – descubrimiento de conocimiento de bases de datos (Usama Fayyad, 1996), desde la recolección de datos de todas la facturas a crédito, con un análisis exploratorio y conocimiento previo de la empresa de estudio, usando la técnica de visualización de datos.

El pre procesamiento inicial de los datos, ha comprendido el proceso ETL (extracción de datos de las diferentes tablas transaccionales, la transformación y limpieza de datos) que incluye limpieza y formateo de datos, transformación de variables categóricas a numéricas, creación de campos calculados, ya que para el estudio se tenía un 15% de clientes con más de un factura a crédito por lo que era necesario agrupar las variables del historial de pago.

Todo el proceso de pre-procesamiento ha tomado un 80% del tiempo del desarrollo del presente proyecto, lo cual ha repercutido en la calidad del modelo de predicción generado, donde se obtiene un nivel de exactitud alto casi en todos los algoritmos de clasificación a excepción de *KNN* (K vecinos más cercanos), con un nivel de exactitud arriba del 92% y una tasa de error entre el 6 al 8% sin selección de atributos.

Pero al aplicar selección de atributos (*FS*) (García Gutiérrez, 2016), usando las técnicas de *CfsSubsetEval*, *CorrelationAttributeEval* y *InfoGainAttributeEval* al dataset obtenido en el pre procesamiento, se redujo la dimensionalidad de atributos; cada técnica de selección de atributos descartó atributos menos relevantes e inútiles y se aplicó los algoritmos de clasificación usados para la comparativa, obteniendo una mejora en cada métrica de evaluación y sobre todo se redujo considerablemente el tiempo de ejecución, obteniendo modelos más simples.

Se aplicaron 7 algoritmos de clasificación J48, Regresión Logística, Naive Bayes, K vecino más cercano (*KNN*), Maquina de Vectores de Soporte (*SVM*), Multilayer Perceptron y Random Forest al dataset completo y a los 3 dataset resultantes de la selección de atributos, se evaluaron indicadores de desempeño como la exactitud, precisión, *ROC*, media armónica, tasa de error, falsos positivos y falsos negativos.

En conclusión, Random Forest y J48 con la técnica de selección de atributos CorrelationAtributeEval han obtenido los mejores resultados en exactitud y precisión, con un bajo nivel de error en la predicción de falsos positivos. Sin embargo, J48 ha tenido un mejor desempeño en el tiempo de ejecución con un 94% de exactitud y una tasa de error de 6%. Es importante tener un número pequeño de falsos positivos para evitar que se pierda de gestionar a los clientes deudores, con mayor riesgo de pago y que pueden ser clasificados erróneamente como clientes que pagan.

La empresa donde se desarrolla el presente proyecto es una compañía limitada de capital privado, la misma que ofrece líneas de crédito en la compra de todas sus líneas de productos tales como línea blanca, línea café, computadores, o celulares entre otros.

Durante el desarrollo del proyecto no se publicará datos de identificación de la empresa ni datos que puedan afectar la intimidad de los clientes, se trabajará solo con datos estadísticos, valores de saldos, cobro, pocos datos personales y de gestiones de cobro que permita identificar los factores o características de pago de los clientes. De acuerdo al *RGPD* (Reglamento General de Protección de Datos) en ningún caso los datos personales han sido comprometidos, ya que la empresa facilitadora de los datos se ha encargado de securizarlos y proporcionar el dataset anonimizado.

1.1 Justificación

La empresa actualmente no puede predecir el comportamiento de pago de sus clientes, para clasificarlos entre buenos pagadores y morosos, solo se basa en los reportes de los balances mensuales que indican el segmento de mora actual del cliente, datos del crédito y comportamiento de pago histórico; pero es necesario determinar el comportamiento de pago de un cliente en los próximos pagos, para anticipar y mejorar la gestión de cobro, para lo cual se requieren analizar varios métodos de clasificación analizando sus patrones y características.

Para el presente proyecto se realizará un análisis de los clientes que tienen créditos de consumo y que realizan el pago en cuotas mensuales, en base a la gestión de recuperación de cartera que el departamento de créditos y cobranzas realiza mediante gestiones telefónicas y domiciliarias.

Se plantea construir un modelo de clasificación para identificar el comportamiento de pago de los clientes en base a variables propias del crédito, que se toman en cuenta para valorar el riesgo del cliente al momento de otorgar dicho crédito, variables tales como: edad, estado civil, tipo de vivienda, cantidad de cuotas, valor de la cuota, saldo de crédito, entre otros.

También variables propias de la gestión de cobranzas como numero de gestiones efectivas o no efectivas, domiciliarias y telefónicas, y, variables de comportamiento de mora que resultan de la conducta de pago y que dan la idea del comportamiento de pago del cliente, como cuotas pagadas, cuotas en mora, mora máxima trimestral, mora máxima anual, numero de cuotas pagadas con mora, entre otros.

Una gestión de cobranzas en base a una correcta predicción, permite focalizar los recursos y esfuerzos del departamento de cobranzas a los clientes donde existe mayor riesgo y

basa su estrategia en pronósticos y conocimiento de su cartera para una correcta toma de decisiones.

Todo esto ayudaría a obtener una mayor recuperación en la cartera de crédito en mora y mejorar sus niveles de índice de morosidad. Para lograr este objetivo la estrategia debe adecuarse al tipo de segmento al que va dirigida y optimizar los recursos materiales y humanos para realizar una gestión de cobranzas efectiva.

1.2 Planteamiento del problema

Se plantea realizar una comparativa de técnicas *Machine Learning* para la clasificación del comportamiento de pago de clientes con cuentas por cobrar, de una empresa que otorga créditos en la venta de productos de línea blanca, línea café, computadores, celulares entre otros, permitiendo pagar el crédito en cuotas mensuales, para que permita identificar el comportamiento de pago mes a mes de los clientes recurrentes y nuevos que ingresan en gestión de recuperación y cobranzas.

Estos modelos permitirán clasificar al cliente considerando las variables del comportamiento histórico del cliente y con estas se busca predecir su comportamiento de pago. Una vez identificado el comportamiento de pago de un cliente, se podrá definir estrategias más adecuadas para la recuperación de cartera, por ejemplo reducir las acciones de intervención a los clientes que probablemente paguen a tiempo e intensificar las acciones a los clientes que puedan caer en mora.

El análisis de información de clientes con diferentes conductas de pago permite a los modelos de predicción aprender las diferentes características que destacan en los clientes que pagan o no pagan una cuota mensual. La variable de respuesta se codifica como "si paga" (1) si el cliente paga el monto total de la cuota, y como "no paga" (0) si el cliente no paga o paga solo una parte de la cantidad que le corresponde pagar en la cuota mensual.

Con la finalidad de ayudar a la gestión de cobranzas de la empresa, se desea detectar patrones crediticios y de comportamiento de pago que puedan afectar en la mora, y así ayudar a disminuir los índices de morosidad de la empresa, mediante el uso de la minería de datos y aprendizaje automático se extrae conocimiento útil y valioso de grandes volúmenes de datos, utilizando el descubrimiento de conocimiento de bases de datos (KDD) (Umaquinga, Saltos, & Peluffo, 2017) .

Lezcano (2002) afirma que “El proceso de KDD usa algoritmos de Minería de Datos para extraer (identificar) lo que se considera como conocimiento de acuerdo a la especificación

de ciertos parámetros usando una base de datos junto con pre-procesamientos y post procesamientos." (Jiménez & Giraldo, 2013, pág. 43)

El proceso de selección de algoritmos se basa en diferentes factores. Inicialmente se ejecutan todos los algoritmos de clasificación incluidos en el dataset original con las variables pre-seleccionadas, luego se aplica los mismos algoritmos a cada dataset con selección de atributos, obteniendo la matriz de confusión con falsos positivos y falsos negativos de los que se obtiene la precisión de cada algoritmo aplicado, en base a los datos obtenidos se realiza una comparativa de todas las técnicas de aprendizaje supervisado, para aplicar el algoritmo de mejor precisión en la clasificación del comportamiento de pago de clientes por cobrar.

1.2.1 Objetivo Principal

La principal contribución de esta investigación o TFM es encontrar la mejor forma de hacer predicciones precisas sobre el comportamiento de pago de clientes con facturas a crédito, basada en datos históricos mediante una comparativa de soluciones, para encontrar el mejor modelo posible de predicciones.

El estudio se realizará en una empresa de venta de productos de línea blanca, línea café, computadores, celulares entre otros que otorga la facilidad de pago a crédito, para predecir si un cliente paga o no su cuota mensual, en base a sus patrones y características similares, lo cual ayudará a mejorar el tratamiento de la gestión de cobranzas enfocándose en los clientes con mayor riesgo de mora.

Los datos históricos que se revisaron corresponden a los créditos de consumo otorgados durante el periodo de enero de 2016 hasta diciembre 2018, para obtener una muestra real el corte de pago se extrajo hasta 30 de abril de 2019, por posibles pagos de clientes con créditos o saldos pendientes a diciembre de 2018.

1.2.2 Objetivos Específicos

Para la ejecución de dicha investigación se realizara un paso a paso de la siguiente metodología de trabajo propuesta:

a) Recolección de datos de clientes con crédito y su comportamiento de pago proveniente de fuentes primarias.
b) Preparación y pre procesamiento de datos para la selección de variables y clase.
c) Discriminar las variables pre-seleccionadas usando técnicas de Selección de Atributos.
d) Aplicación de varios modelos de clasificación.
e) Validación y selección del mejor modelo de clasificación.

1.3 Glosario de términos usados dentro del proyecto

Etapas de cobranza: dependiendo del tiempo transcurrido del no cumplimiento del pago de las cuentas por cobrar tienen a caer los clientes en dos diferentes etapas prejudiciales y judiciales.

Riesgo crediticio: está relacionado con la incertidumbre sobre los créditos aprobados a terceros y la recuperación del mismo en los términos acordados por la empresa.

Clientes morosos: son aquellos clientes que no pagan en la fecha de vencimiento sus deudas, aunque es posible que paguen después.

Mora: La mora es el retraso culpable o deliberado en el cumplimiento de una obligación o deber.

Cartera: en la gestión de créditos y cobranzas se refiere a las cuentas por cobrar.

1.4 Estructura del Trabajo

El presente trabajo de investigación está organizado de la siguiente manera:

La **sección 2** revisa el estado del arte con un análisis de trabajos anteriores y similares y nuestra contribución, también se realiza un análisis de la situación actual de la empresa, y el marco teórico que corresponde a las técnicas de minería de datos que se pretende aplicar.

La **sección 0** presenta la metodología de trabajo a seguir desde la recolección de datos, pre procesamiento de datos, selección de atributos, aplicación de modelos de clasificación y métricas de validación y selección.

En la **sección 4** se analiza los resultados obtenidos al aplicar los modelos de clasificación propuestos en el dataset completo y los dataset con selección de atributos, para encontrar el mejor modelo y técnica de predicción.

La **sección 5** presenta las conclusiones y futuras líneas de trabajo.

2 Contexto y Estado del arte

Este capítulo inicia comprendiendo los principales conceptos de la administración de cobranzas, para luego profundizar y conocer la forma de trabajo de la empresa de estudio y sus necesidades, de manera que permita entender mejor el trabajo a desarrollar.

Así mismo, se realiza un análisis de trabajos anteriores o similares y nuestra contribución, se presenta una introducción acerca de los varios métodos de clasificación y las técnicas de selección de atributos, que son frecuentemente utilizados en la minería de datos y que serán utilizados en nuestro trabajado investigativo.

2.1 Administración de Cobranzas

Una venta no está completa hasta que se realiza el cobro y ahí es donde empieza a funcionar la cobranza de los créditos de la empresa, que en caso de no gestionar correctamente se corre el riesgo de la falta de liquidez y atrasos en los saldos de las cuentas por cobrar, por lo que el éxito o fracaso de una empresa puede verse determinada por la eficiencia con que se recupere sus cuentas por cobrar (Morales Castro & Morales, 2014).

La administración de una cartera de cuentas por cobrar tiene como actividad fundamental la prevención, a través de conocer mejor a los clientes y tener un especial cuidado al momento de otorgar un crédito, otra medida es administrar de una manera óptima la cartera de los clientes y determinar estrategias para cuando los clientes no cumplen sus pagos, siendo muy consistentes con la situación del mercado, la económica y el tipo de cliente; todo esto conduce a tener una cobranza eficiente y oportuna.

2.1.1 Fases de la Cobranza

Morales Castro & Morales (2014) indica como fases de cobranza las siguientes:

1. **Prevención**: Acciones encaminadas a evitar el incumplimiento de pago de un cliente, es decir, disminuir el riesgo de mora en las cuentas por cobrar.
2. **Cobranza**: Acciones encaminadas a recobrar adeudos en instancias tempranas de mora.
3. **Recuperación**: Acciones encaminadas a recobrar adeudos de créditos en mora con bastante tiempo.
4. **Extinción**: Acciones encaminadas a registrar contablemente las cuentas por cobrar como saldadas, cuando los clientes pagan los adeudos respectivos.

La gestión de cobranzas se basa en segmentación y en utilización de canales de contacto como llamadas telefónicas, visitas domiciliaras, envío de SMS masivos con el fin de obtener una respuesta positiva por parte del cliente para el cumplimiento de sus obligaciones.

Cuando a pesar de todo el esfuerzo que ha realizado el gestor de cobranzas, no se logra recuperar las cuentas por cobrar durante un tiempo establecido, se origina nuevos gastos adicionales como la cobranza prejudicial y judicial a través de juicios que determinen otras acciones de cobro.

2.2 Compresión del trabajo

La empresa fuente del estudio ofrece créditos de consumo para que las facturas de compra sean pagadas de 1 a 30 cuotas mensuales, previa autorización y análisis del analista de créditos. Previo a la obtención del crédito, el departamento de Créditos y Cobranzas realiza una categorización del cliente que permita analizar su capacidad de pago; una vez aprobado el crédito se factura al cliente el producto y se difiere el pago en cuotas mensuales que el cliente debe cancelar mensualmente a partir de la primera fecha de vencimiento.

Dicha empresa de venta de productos línea blanca, línea café, computadores, celulares entre otros, con sede en Ecuador y sucursales a nivel nacional, opera con un esquema tradicional de cobro porque da el mismo tratamiento a todos sus clientes y su estrategia está basada en la recuperación de saldos de cartera.

Actualmente su gestión de cobro no se basa en un predicción de posibles clientes morosos, que permita focalizar los esfuerzos donde existe mayor riesgo, basando su estrategia en pronósticos y conocimiento previo de la cartera, obteniendo así una mayor recuperación y un menor índice de morosidad, lo cual puede generar costos elevados en recursos humanos y técnicos.

En su caso, realiza acciones de recaudación genérica sin tener en cuenta detalles del cliente, por ejemplo todos los clientes son contactados en intervalos fijos aunque paguen a tiempo, también es cierto que mientras más tarde se contacta con el cliente hay menos probabilidad de pago a tiempo de la cuota.

Pero el contacto repetitivo a los clientes puede conducir a disminuir la satisfacción del cliente, de esta manera se quiere ayudar a predecir el comportamiento de pago de un cliente, para tomar acciones correctivas en base al pago del cliente durante la fase preventiva y administrativa de cobranzas de facturas a crédito.

El objetivo es adecuarse al tipo de segmento al que va dirigido y optimizar los recursos humanos y materiales que intervienen en el proceso para lo que es necesario segmentar su cartera de clientes, en clientes que pagan o no pagan una cuota mensual.

La cobranza eficiente orientada al cliente, se refiere al uso de conocimiento sobre el desempeño y características del mismo, donde la probabilidad de deterioro del

comportamiento de pago defina las estrategias de cobro para encontrar un algoritmo y una metodología que permita gestionar eficientemente la gestión de cobranza, usando la analítica de datos en base a predicción de datos (García, y otros, 2018).

2.3 Trabajos similares y contribución en esta investigación

En la Tabla **1** se presentan algunos proyectos de investigación similares para análisis de comportamiento de pago de los clientes, pago de facturas y la diferencia que presenta con el presente proyecto de investigación.

Tabla 1. *Resumen de trabajos Similares. Elaboración propia*

	Proyecto de investigación 1 Cabezas Arias, Lady Pamela y Parra Romo, Nelson Antonio. (2017).	Proyecto de Investigación 2 Jácome Jara, Marco Santiago. (2014).	Proyecto de Investigación 3 Cheong, Michelle L. F. and SHI, Wen. (2018)	Proyecto de Investigación 4 Zeng, Sai & Melville, Prem & A. Lang, Christian & M. Boier-Martin, Ioana & Murphy, Conrad. (2008).
Tema de investigación	Aplicación de un modelo Perceptrón Multicapa de redes neuronales artificiales para la clasificación del comportamiento de pago en clientes en mora en una entidad de cobranza	Construcción de un modelo estadístico para calcular el riesgo de deterioro de una cartera de microcréditos y propuesta de un sistema de gestión para recuperación de la cartera en una empresa de cobranzas	Customer level predictive modeling for accounts receivable to reduce intervention actions	Using predictive analysis to improve invoice-to-cash collection
Técnica de minería de datos utilizada	Redes Neuronales	Regresión Logística, regresión logística mejorada con árboles de decisión, regresión basada en distancias, K-NN	Árbol de probabilidad, árbol de clasificación errónea, Regresión, Regresión polinomial, Red neuronal, Regresión con red neuronal y modelo de conjunto.	C4.5 Arboles de decisión, Clasificador Naïve Bayes, Regresión logística, PART, Boosting decision stumps.
Metodología para proyecto de minería utilizada	No define metodología.	Knowlege Discovery in Databases - KDD	No define metodología	KDD
Total registros usados	50000	2931	10.562	Dataset de facturas de 4 empresas
Herramienta para minería de datos	R Studio	SPSS y WEKA	SAS Enterprise	No indica la herramienta usada.
Año de publicación de la investigación	2017	2014	2018	2008
Objetivo planteado que se resalta	Identificación de diferentes tipos de clientes presentes en la cartera de recuperación por su comportamiento de pago, usando el modelo perceptron multicapa para una institución financiera, que determina si un cliente es bueno o malo. Analiza créditos solo a cuotas de 24 meses.	Construir modelos estadísticos de score para calcular el riesgo de deterioro en los clientes de cartera de microcréditos. Analiza créditos a cuotas de 12 meses.	Predecir si un cliente paga a tiempo o no facturas mediante una medida de pureza para determinar si el cliente es bueno (pureza=1) o malo (pureza=0)	Modelos predictivos para predecir si una factura se pagara a tiempo o se retrasará, de 1 a 30 días, 31 a 60 días, 61 a 90 días o más de 90 días de retraso.

Conclusiones que se resaltan	-El modelo con mayor exactitud fue el de 3 neuronas. -Se establece que mientras mayor sean los días de mora menor probabilidad de pago sin importar el ingreso, el número de llamadas o cargas familiares del deudor.	-La mejor metodología fue el clasificador K-NN porque tiene-mayor precisión en la predicción. -El método de regresión logística mejorada con árboles decisión fue el mejor modelo para entender cómo se comporta la cartera por su capacidad de explicación.	-El mejor modelo de predicción fue la red neuronal. - La tasa de clasificación errónea es más adecuada para datos categóricos. - La red neuronal tuvo la menor suma de errores al cuadrado y error cuadrático promedio.	-Además de las características de la factura, el comportamiento de pago histórico del cliente proporciona una mejora significativa en la exactitud de la predicción. -Un modelo específico por empresa tuvo mayor precisión en la predicción que un modelo unificado para las 4 empresas.
Diferencias con relación al presente proyecto de investigación.	-Objetivo: Comparativas de varias técnicas machine learning para clasificar los clientes por su comportamiento de pago. -Análisis de clientes con más de una factura a crédito y con diferentes cuotas de pago. -Pre selección y Selección de variables usando feature selection. -Herramienta utilizada. -Propuesta de una metodología de desarrollo. -Variables o factores analizados.	-Mayor cantidad de datos. -Objetivo: Comparativas de varias técnicas machine learning para clasificar los clientes por su comportamiento de pago. -Análisis de clientes con más de una factura a crédito y con diferentes cuotas de pago. -pre selección y Selección de variables usando feature selection. -Herramienta utilizada. -Propuesta de una metodología de desarrollo. -Variables o factores analizados.	-Mayor cantidad de datos. -Objetivo: Comparativas de varias técnicas machine learning para clasificar los clientes por su comportamiento de pago. -Análisis de clientes con más de una factura a crédito y con diferentes cuotas de pago. -pre selección y Selección de variables usando feature selection. -Herramienta utilizada. -Propuesta de una metodología de desarrollo. -Variables o factores analizados.	-Objetivo: Comparativas de varias técnicas machine learning para clasificar los clientes por su comportamiento de pago. -Análisis de clientes con más de una factura a crédito y con diferentes cuotas de pago. -Pre selección y Selección de variables usando feature selection. -Herramienta utilizada. -Propuesta de una metodología de desarrollo. -Variables o factores analizados.

Una vez analizados los trabajos descritos en la Tabla **1**, vemos que la aportación de este trabajo respecto al resto es realizar una comparativa de técnicas *Machine Learning* con y sin selección de atributos, siguiendo una metodología de desarrollo basada en KDD para encontrar la mejor técnica y modelo de clasificación.

Además de indicar que los atributos y los datos con los que se pretende trabajar son completamente distintos y propios de la empresa de estudio.

En resumen las contribuciones que se pretenden realizar con el desarrollo del presente proyecto investigativo son las siguientes:

- Identificación de algoritmos de mayor precisión para predecir el comportamiento de pago mensual de clientes con créditos vigentes, usando variables de cliente, variables de crédito y variables de gestión de cobranza.
- Hacer una comparativa de algoritmos con atributos o características iniciales y por selección de atributos.
- Brindar información sobre qué técnica, metodología y algoritmos se pueden utilizar en la predicción de comportamiento de pago de clientes con cuentas por cobrar, para el ámbito de CRÉDITOS y COBRANZAS. Los resultados de esta investigación podrán servir como referencia para otras investigaciones similares, pudiéndose aumentar el número de variables y el tamaño de los datos.
- Detectar factores que faciliten determinar de forma temprana si un cliente tiene riesgo de caer en mora o no, para en base a ello tomar las medidas correctivas de gestión de cobro.
- Generar una metodología que permita extraer conocimiento útil y valioso de grandes volúmenes de datos

2.4 Metodología KDD (Knowledge Discovery in Databases)

El proceso *KDD* (descubrimiento de conocimiento de bases de datos) es interactivo e iterativo, Brachman y Anand (1996) dan una visión práctica del proceso enfatizando la naturaleza interactiva del proceso. Indica que sus pasos básicos son los siguientes (Usama Fayyad, 1996):

1. Desarrollar una comprensión del domino de la aplicación e identificar el objetivo del proceso KDD desde el punto de vista del cliente.
2. Seleccionar un conjunto de datos objetivo, enfocarse en un subconjunto de variables y muestras de los datos donde se va a realizar el descubrimiento.
3. Limpieza y pre procesamiento de datos. Las operaciones básicas incluyen eliminar el ruido, manejar campos faltantes entre otros.

4. Reducción y proyección de datos, con métodos de reducción y transformación de datos, el número efectivo de variables puede ser reducido o transformado.
5. Igualar los objetivos del proceso KDD a un método de minería de datos particular. Por ejemplo: resumen, clasificación, regresión, agrupamiento, etc.
6. Análisis exploratorio y selección del modelo e hipótesis, se decide qué modelos y parámetros pueden ser los apropiados
7. Minería de datos: búsqueda de patrones de interés en una representación particular o en un conjunto de representaciones incluyendo reglas de clasificación o árboles, regresión y agrupamiento (clustering).
8. Interpretar los patrones minados. Posiblemente volviendo a cualquiera de los pasos 1 a 7 para iteración adicional. Se puede visualizar los patrones extraídos o visualizar los datos dados de los modelos extraídos.
9. Actuando sobre el conocimiento descubierto, utilizando el conocimiento directamente, incorporando el conocimiento en otros sistemas para acción adicional o simplemente documentando e informarlo a las partes interesadas.

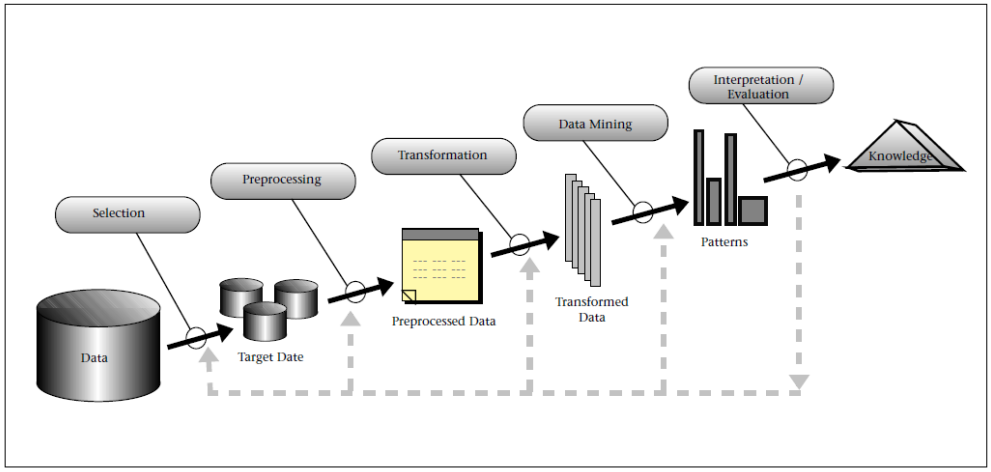


Ilustración 1. Pasos a seguir de un proceso KDD. Usama Fayyad, G. (1996). Recuperado de https://pdfs.semanticscholar.org/fca8/b47e29e75d0676af8ec1b36a1dd29c9a1e6e.pdf

2.5 Minería de Datos

Giudici (2003) afirma “La minería de datos es el proceso de selección, exploración, y modelado de grandes volúmenes de datos para descubrir regularidades o relaciones que al principio se desconocen, con el objetivo de obtener resultados claros y útiles para el propietario de la base de datos”. La base de la minería de datos está en la inteligencia artificial, estadística, aprendizaje automático y los sistemas de bases de datos.

La minería de datos es una herramienta útil para tomar decisiones estratégicas y desempeña un papel importante en la segmentación de mercados, servicios al cliente, detección de fraudes, otorgamiento de créditos y su evaluación (García, y otros, 2018).

Los métodos más utilizados para realizar tareas de predicción de clientes por su comportamiento de pago son los métodos de clasificación. Es un método que clasifica a los miembros de un conjunto dado de instancias en algunos grupos en términos de sus características. En la literatura hay muchos algoritmos de predicción, cada uno con un conjunto de parámetros diferentes que puede afectar los resultados obtenidos.

Se han elegido métodos de predicción bien conocidos implementados en la herramienta Weka, ya que estos métodos son ampliamente conocidos y pueden ser útiles para resolver problemas similares de predicción de comportamiento de pago de clientes a crédito y que se detallan a continuación.

Arboles de Decisión (*Decision Tree*): Los arboles de decisión son algoritmos donde los datos se dividen continuamente de acuerdo con un determinado parámetro. El árbol se explica por dos entidades llamadas nodos y hojas de decisión. El nodo más alto es considerado el nodo raíz, las hojas son las decisiones o los resultados finales y los nodos decisión son donde se dividen los datos y dirigen una decisión a tomar. Los arboles de decisión se utilizan cuando la variable respuesta es cuantitativa discreta o cualitativa. Los arboles de clasificación se basan en maximizar la medida de pureza de las variables respuesta de las de observaciones. La ventaja de este método es simple de entender y explicar, pero la limitación de este modelo es que no se puede generalizar una estructura de diseño para un contexto, para otros contextos (Keramati & Yousefi, 2011) (Suca, Córdova, Condori, Cayra, & Sulla, 2016).

Árbol de decisión C4.5 es conocido como J48 en la herramienta Weka. Es un método iterativo que va colocando los posibles valores de los atributos por su ganancia de información, cuando todas caigan en una clasificación y no exista ambigüedad entonces se asigna un nodo raíz.

Las ventajas de un árbol de decisión C4.5 con respecto a ID3 son (Suca, Córdova, Condori, Cayra, & Sulla, 2016):

- Manejo de atributos continuos y discretos. C4.5 para manejar atributos continuos crea un umbral y luego se divide la lista en aquellos cuyo valor de atributo sea superior al umbral y los menores o iguales a él.

- Maneja datos faltantes, los datos faltantes son marcados con signos de interrogación y no son usados en el cálculo de la ganancia y la entropía.
- Manejo de atributos con costos diferentes.
- C4.5 realiza la poda del árbol después de la creación, eliminando las ramas que no ayudan, reemplazándolos con los nodos hoja.

Redes neuronales (*Artificial neural networks*): Keramati & Yousefi afirman "Las redes neuronales artificiales son modelos estadísticos no lineales basados en la función del cerebro humano. Son herramientas poderosas para el modelado de relaciones de datos desconocidos. *ANN* es capaz de reconocer el patrón complejo entre las variables de entrada y salida y predecir el resultado de nuevos datos de entrada independientes". (Keramati & Yousefi, 2011, pág. 418)

Clasificador bayesiano: El clasificador de Bayes es un clasificador probabilístico simple basado en la aplicación del teorema de Bayes (estadísticas bayesianas) con fuerza (naive), describe la probabilidad de un evento, basado en el conocimiento previo de las condiciones que podrían estar relacionadas con el evento. (Keramati & Yousefi, 2011)

Las ventajas del clasificador *Naive Bayes* incluye: menor coste computacional, fácil de entender al ser un clasificador basado en probabilidades, y si los datos son discretos solo requiere un paso de procesamiento (Adnan, Husain, & Rashid, 20012). La ventaja de este método es que requiere una pequeña cantidad de datos de entrenamiento para estimar los parámetros (medias y varianzas de las variables) necesarias para la clasificación. La principal desventaja de este modelo es que la precisión predictiva está altamente correlacionada con este supuesto.

Regresión logística: la regresión logística es una forma de regresión lineal, que puede predecir una salida discreta de un conjunto de variables que pueden ser continuas, discretas, dicotómicas o una combinación de cualquiera de ellas. Generalmente la variable dependiente o respuesta es dicotómica. La relación entre el predictor y la variable respuesta no es una función lineal en la regresión logística. Keramati & Yousefi (2011) afirma que "Las ventajas de este método son que la regresión logística no asume la linealidad de la relación entre las variables independientes y la variable dependiente, no requiere variables distribuidas normalmente y la debilidad del modelo es que las variables independientes se relacionan linealmente con el logit de la variable dependiente". (Keramati & Yousefi, 2011, pág. 418)

K vecino más cercano (*KNN*): es un clasificador no paramétrico basado en el aprendizaje por similitud. Se recopila un conjunto de datos de entrenamiento, para este conjunto de

datos de entrenamiento se introduce una función de distancia entre la variable explicativa de las observaciones. Para cada nueva observación, este método explora el espacio de patrones para los k vecinos más cercanos que están más cerca de la observación en términos de distancia entre las variables explicativas. La nueva observación se asigna a la clase que pertenece las *KNN* a esa clase (Keramati & Yousefi, 2011).

Máquina de vectores de soporte (*SVM*): la máquina de vectores de soporte es una técnica de clasificación que involucra tres elementos. Una fórmula de puntuación que es una combinación lineal de atributos seleccionados para el problema de clasificación, una función objetiva que considera tanto las muestras de entrenamiento como las de prueba para optimizar la clasificación de nuevos datos, un algoritmo de optimización para determinar los parámetros óptimos de entrenamiento de la función objetivo de la muestra. Las ventajas del método son que, en el caso no paramétrico, SVM requiere una estructura de datos sin suposiciones como la distribución normal y la continuidad. SVM puede realizar una asignación no lineal de un espacio de entrada original a un espacio de atributos de alta dimensionalidad y este método es capaz de predicciones continuas y categóricas. Las debilidades de este método son que son difíciles de interpretar a menos que las características interpretables y las formulaciones estándares no contengan especificación de las restricciones de negocios. (Keramati & Yousefi, 2011)

2.6 Selección de Atributos.

La técnica de selección de atributos es considerada como un algoritmo de pre procesamiento, que nos ayuda a reducir el número de atributos que conforma el dataset original (datos en bruto) con el fin de mejorar el rendimiento y precisión de los algoritmos de aprendizaje. Ejemplos: algoritmos de inducción de modelos, inferencia de reglas, o de agrupamiento (clustering). Se realizan tareas de adecuación de los datos, eliminación de atributos redundantes o no relevantes, creación y combinación de atributos existentes, normalización de atributos o transformación de los mismos para mejorar el procesado (García Gutiérrez, 2016).

“El objetivo del proceso de selección de atributos es elaborar un ranking de características en función a su relevancia predictiva, lo que permite la reducción de la dimensionalidad de los datos pudiéndose entonces crear modelos que describen mejor el vector de clases observado”. (García Gutiérrez, 2016, p. 1).

Desafortunadamente no existe una única técnica que siempre obtenga los mejores resultados porque esta elección se debe hacer de forma empírica por parte de personas expertas en el área.

Según (Guyon & Elisseff, 2003) la utilización de métodos de selección de características aporta las siguientes ventajas:

- Mejora en el rendimiento de los algoritmos de aprendizaje en tiempo y en memoria, porque requiere menos CPU y recursos de memoria.
- Disminución de los costes asociados a la recopilación de datos.
- Mejora la predicción de los resultados obtenidos partiendo solo de atributos relevantes.

Las técnicas de selección de atributos más utilizadas son los métodos de filtrado y los métodos de envoltura. Los métodos de envoltura usan el algoritmo de clasificación para medir la importancia de los atributos que componen el conjunto de datos por lo tanto podemos decir que la selección de atributos por envoltura depende en gran medida del algoritmo que hemos elegido de antemano.

Los métodos de envoltura tienen generalmente mejor desempeño que los métodos de filtrado ya que al mejorar el algoritmo de clasificación mejoramos el proceso de selección de atributos, pero tienen un coste computacional demasiado alto. Los métodos de filtrado son independientes del algoritmo de aprendizaje y suelen ser computacionalmente más simples y paralelizables, porque evalúan únicamente la relevancia de los atributos y su relación con la clase sin ocupar el algoritmo de clasificación.

2.6.1 Técnicas de selección de atributos

Kittler (1986) señala que “La extracción de características es un enfoque basado en la transformación, por lo tanto, transforma el significado original de las características. Este método implica la creación de un subconjunto de nuevas características mediante la combinación de las existentes; por lo tanto, se emplea cuando la semántica del conjunto de datos original no es necesaria para ningún proceso futuro” (Castillo et al, 2016, pág. 5).

Liu y Motoda (1998) señala que “FS intenta conservar el significado del conjunto de características original. Es parte de lo que se denomina técnicas de preservación semántica, conocidas como enfoques basados en la selección.” (Castillo et al, 2016, pág. 6).

En el problema que aborda el presente trabajo de investigación, es muy importante mantener la semántica del conjunto de variables iniciales y por lo tanto, se han aplicado los métodos de FS (selección de características), puesto que interesa saber que atributos de los clientes definen su comportamiento de pago.

“Los métodos de selección de atributos trabajan sobre subconjuntos del conjunto total de atributos y tratan de medir la calidad y la relevancia de dichos subconjuntos. Se trata no solo

de intentar evaluar la relevancia de cada atributo por separado sino del subconjunto como un todo ya que dos atributos tomados de forma conjunta pueden aportar aquella información más relevante de la que aportan considerándolos por separado.” (García Gutiérrez, 2016, p. 9)

WEKA (*Waikato Environment for Knowledge Analysis*) es uno de los diversos sistemas computacionales que permiten aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, permitiendo aplicar las técnicas de selección de atributos. Esta herramienta separa el proceso en 2 partes (Brownlee, 2016):

- ***Evaluador de atributos (Attribute Evaluator):*** El evaluador de atributos es una técnica que cada atributo del dataset conocido como feature es evaluado en el contexto de una variable de salida (clase). Este método es una técnica por el cual se prueba diferentes combinaciones de atributos en el conjunto de datos para llegar a una lista corta de características escogidas.
- ***Método de búsqueda (Search Method):*** describe la forma estructurada en que se estudia el conjunto de posibles subconjuntos de características, en función de los resultados del evaluador.

Hay muchos algoritmos implementados por Weka de entre ellos se ha considerado los más relevantes para el tema de investigación.

- **Correlation Feature Selection (*CfsSubsetEval o CSE*):** Elabora un ranking de subconjuntos de atributos de acuerdo a la correlación estadística, buscando atributos que están poco correlacionados entre ellos, pero que tienen una buena correlación con la clase. Por lo tanto, los atributos irrelevantes serán ignorados porque mantendrán una muy baja o nula correlación con la clase. Por otra parte, la información redundante será penalizada ya que el atributo redundante tendrá una alta correlación con una o varias de las características o atributos restantes (García Gutiérrez, 2016). Este evaluador es aplicado con el método de búsqueda BestFirst.
- **Correlation Atribute Eval (*CA*):** Evalúa los atributos con respecto a la clase objetivo. El método de correlación de Pearson se utiliza para medir la correlación entre los atributos de cada uno y el atributo clase. Considera los atributos nominales en base al valor y cada valor actúa como un indicador. Se combina este atributo evaluado con el algoritmo de búsqueda Ranker. “Se puede calcular la correlación entre cada atributo y la variable de salida y seleccionar solo aquellos atributos que tienen una correlación positiva o negativa de moderada a alta (cerca -1 a 1) y

eliminar aquellos atributos con una correlación baja (valores cercanos a cero)" (Brownlee, 2016).

- **Information Gain Attribute (*IG*):** Se puede calcular la ganancia de información (entropía) para cada atributo con respecto a la clase objetivo. "Los valores de entrada varían desde 0 (sin información) a 1 (máxima información). Aquellos atributos que aporten más información tendrán un mayor valor de ganancia de información y podrán seleccionarse, mientras aquellos que no agreguen mucha información tendrán una puntuación más baja y podrán eliminarse". (Brownlee, 2016).

  Se combina con el método de búsqueda Ranker y se puede calcular con la siguiente formula (Gnanambal, Thangaraj, Meenatchi, & Gayathri, 2018).

  InfoGain (Clase, Atributo) = H (Clase) - H (Clase | Atributo)

  Donde H representa la Entropía.

- **GainRatio Attribute (*GA*):** Este método mide la importancia de los atributos con respecto a la clase objetivo sobre la base de la relación de ganancia. Y puede ser calculado por la siguiente formula, (Gnamambal , Thangaraj, Meenatchi, & Gayathri, 2018).

  GainR (Clase, Atributo) = (H (Clase) - H (Clase | Atributo)) / H (Atributo) donde H representa la Entropía

Todos los métodos de selección de atributos utilizan algún tipo de algoritmo de búsqueda para encontrar los mejores atributos dentro del conjunto de datos completo, Weka usa los siguientes algoritmos de búsqueda:

- **GreedyStepwise:** Realiza una búsqueda voraz sin vuelta atrás (*backtracking*), es decir, selecciona el mejor de todos, luego la mejor pareja que lo incluye, luego el mejor trio que incluya a los atributos anteriores y así hasta que la solución ya no mejora (García Gutiérrez, 2016).
- **BestFirst**: Igual que Greedy es una búsqueda en profundidad, pero aplicando vuelta atrás (*backtracking*) hasta un límite de retrocesos. El subconjunto de atributos es evaluado usando una métrica y el valor obtenido se guarda como una cota, se quitan otros atributos del conjunto original, siguiendo un esquema ordenado de eliminación de atributos; cada subconjunto obtenido es evaluado. Si algún subconjunto obtiene una evaluación igual o peor que la cota, se detiene la exploración es decir se realiza una poda. Por otro lado si todos los subconjuntos evaluados resultan mejor que la

cota, se actualiza la cota con el nuevo valor y se repite el proceso hasta que no haya ramas que explorar tal como una búsqueda en árbol. Este procedimiento ahorra tiempo y es considerado una solución óptima (García Gutiérrez, 2016).

- **Ranker:** Se utiliza cuando evaluamos atributos por separado de uno en uno, se evalúa de manera independiente cada atributo que tenga correlación con el valor objetivo (clase) ordenando y eliminando los atributos menos valorados.

2.7 Métricas de Evaluación

La evaluación de los algoritmos de aprendizaje automático es un parte esencial en la ejecución de cualquier estudio investigativo. Cada modelo puede proporcionar resultados satisfactorios cuando se evalúa con una métrica, por ejemplo la precisión, pero puede dar resultados deficientes cuando se evalúa en comparación con otras métricas (Mishra, 2018). Por lo tanto, sería mejor evaluar la calidad de un modelo de aprendizaje automático mediante varias métricas que evalúan el rendimiento del modelo, tales métricas nos permitirá comparar distintos modelos y elegir el mejor.

Los algoritmos de clasificación tienen un conjunto de métricas de evaluación que se detallan a continuación.

Matriz de Confusión: Dadas n clases, la matriz de confusión tiene un tamaño de nxn y sus elementos p_{ij} indican el número de instancias de la clase i que han sido clasificadas por el modelo como de la clase j. Donde, las filas tienen la clase real de la instancia y las columnas tienen la clase estimada por el clasificador (Mishra, 2018).

```
=== Confusion Matrix ===

 a b   <-- classified as
 8 1 | a = yes
 1 4 | b = no
```

Ilustración 2. Matriz de Confusión en Weka.

Una matriz de confusión nos permite visualizar mediante una tabla la distribución los errores cometidos por un algoritmo de clasificación.

Tabla 2. *Matriz de Confusión. Elaboración Propia*

	Clase Predicha	
Clase real	**Si**	**No**
Si	**Verdadero Positivo (VP)**	**Falso Negativo (FN)**
No	**Falso Positivo (FP)**	**Verdadero Negativo(VN)**

- **VP (Verdaderos positivos):** instancias correctamente reconocidas por el modelo.
- **FN (Falsos negativos):** instancias positivas que el modelo las clasifica como negativas.
- **FP (Falsos positivos):** instancias que son negativas pero el modelo dice que son positivas.
- **VN (verdaderos negativos):** instancias que son negativas y correctamente reconocidas como tales.

Partimos de que **N** es el número del conjunto de datos de entrenamiento

$$N = VP + FP + VN + FN$$

Ecuación 1. Suma total del conjunto de datos de entrenamiento

El número de instancias clasificadas correctamente es la suma de la diagonal de la matriz y el resto están clasificadas incorrectamente.

Exactitud: La **exactitud (*accuracy*)** es el porcentaje de instancias del conjunto de datos de prueba que son clasificadas correctamente, se puede calcular tomando el promedio de los valores que se encuentran en la diagonal principal de la matriz de confusión, es decir,

$$Exactitud = \frac{VP + VN}{N}$$

Ecuación 2. Formula de la Exactitud

La tasa de error (*error rate*) es el porcentaje de instancias del conjunto de datos que son clasificadas incorrectamente.

$$Tasa\ de\ Error = 1 - Exactitud = 1 - \frac{VP + VN}{N}$$

Ecuación 3. Formula de la tasa de error.

Que equivale a la siguiente formula.

$$Tasa\ de\ Error = \frac{FP + FN}{N}$$

Ecuación 4. Formula equivalente de la tasa de error.

La **precisión** es el porcentaje de instancias positivas respecto al total predichas como positivas.

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

Ecuación 5. Formula de la Precisión

Área bajo la curva (*AUC – Area Under Curve*)

El área bajo la curva es una de las métricas más utilizadas para la evaluación de los algoritmos de clasificación binaria. “El AUC de un clasificador es igual a la probabilidad de que el clasificador clasifique un ejemplo positivo elegido al azar más alto que un ejemplo negativo elegido al azar” (Mishra, 2018). Antes de definir el AUC se debe entender los dos siguientes términos.

- **La tasa de verdaderos positivos (*TP Rate*)** es la proporción de instancias positivas correctamente clasificadas, con respecto a todas las instancias positivas, también llamado ***Recall*, alcance y sensibilidad**.

$$Tasa\ Verdaderos\ Positivos\ (TP\ Rate) = \frac{VP}{FN + VP}$$

Ecuación 6. Tasa de verdaderos Positivos

- **La tasa de falsos positivos (*FP Rate*)** que es la proporción de instancias negativas que se consideran erróneamente como positivas, con respecto a todas las instancias negativas, también llamada como **especificidad** (*specificity*).

$$Tasa\ Falsos\ Positivos\ (FP\ Rate) = \frac{FP}{FP + VN}$$

Ecuación 7. Tasa de falsos positivos

La tasa de falsos positivos y la tasa de verdaderos positivos tienen valores en el rango [0, 1], con valores de umbral como (0.00, 0.02, 0.04,…., 1.00) dibujados en una gráfica. “AUC es el área bajo la curva de la gráfica de la Tasa de falsos positivos versus tasa de verdaderos positivos en diferentes puntos en [0, 1]” (Mishra, 2018).

Por lo tanto cuanto mayor sea este valor mejor es el rendimiento de nuestro modelo.

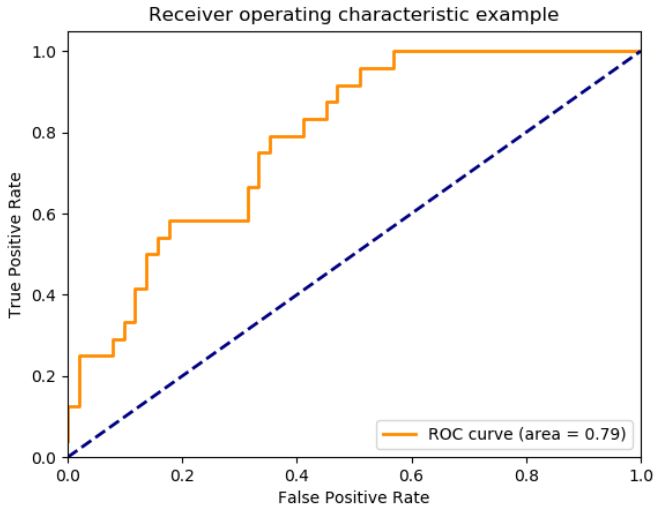


Ilustración 3. ROC área. (Mishra, Metrics to Evaluate your Machine Learning Algorithm, 2018). Recuperado de https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

F Measure

Se considera la Media armónica entre la precisión y recall. Su rango de puntación es [0,1] y dice que tan preciso es un clasificador (cuantas instancias clasifica correctamente) así como que tan robusto es (que no pierda un número significativo de instancias). Cuanto mayor es este puntaje mayor es el rendimiento de nuestro modelo (Mishra, 2018).

$$F\ measure = 2 * \frac{1}{\frac{1}{Precisión} + \frac{1}{Recall}}$$

Ecuación 8. F Measure o F1 Score

F Measure trata de encontrar el balance entre precisión y recall.

2.8 Conclusiones

Al finalizar este capítulo se ha brindado una introducción al problema a resolver ya que es importante conocer el modelo de negocio del presente trabajo a desarrollar, se ha considerado importante realizar una comparativa entre el trabajo actual y algunos trabajos ya existentes, destacando la contribución de la presente investigación, además se ha realizado una descripción teórica de los métodos de selección de características, algoritmos de clasificación a usar y sus respectivas métricas de evaluación. En el siguiente capítulo se continúa con la metodología a seguir para el cumplimiento de la presente investigación.

3 Metodología de trabajo

No existe una metodología estándar a seguir para resolver el problema general de la predicción de clientes de cartera de crédito por probabilidad de pago, sin embargo para el desarrollo del presente trabajo se aplicará el proceso denominado Descubrimiento de conocimiento en base de datos (*Knowlege Discovery in Databases - KDD*), el cual es un proceso metodológico para encontrar un modelo valido, útil y entendible que describa patrones de acuerdo a la información y como modelo se interpreta que es la representación que intenta explicar ese patrón de los datos (Usama Fayyad, 1996). En la *Ilustración **4*** se resume los pasos a seguir.

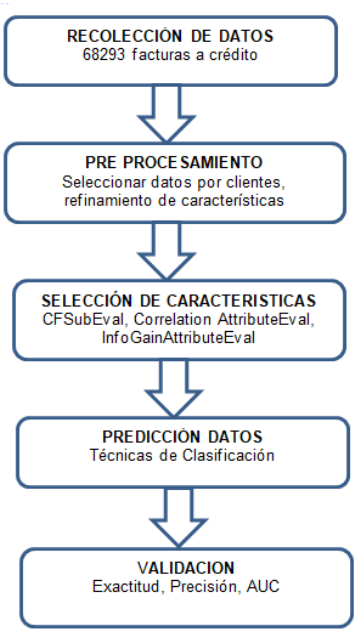


Ilustración 4. Pasos a seguir en la metodología propuesta.

Por lo tanto, después de la recolección de datos y un pre procesamiento inicial mediante un proceso de limpieza de datos y selección de variables relevantes, otras técnicas de pre procesamiento serán aplicadas como la Selección de atributos (*Feature Selection - FS*), que es una técnica de reducción de dimensionalidad efectiva, capaz de descartar las características menos relevantes o no útiles (Castillo, et al., 2016).

“Los métodos FS verifican la relevancia de todas las variables de los patrones, con el objetivo de reducir el conjunto de características en un conjunto de datos, para obtener modelos inteligibles, en un menor tiempo debido a su menor complejidad, y generalmente sin pérdida de precisión. Esta tarea también puede ayudar a identificar los indicadores principales a ser considerados por un experto humano para hacer predicciones.” (Castillo, et al., 2016, p. 4).

Después que se haya llevado a cabo el proceso de selección de características, se aplica diferentes modelos de clasificación tanto al conjunto de datos completo como a cada conjunto de datos reducido en base a la selección de atributos. Los modelos de predicción serán evaluados por medio de métricas y experimentos.

3.1 Recolección de datos.

El primer paso a seguir es la recolección de los datos de clientes con crédito y su comportamiento de pago provenientes de fuentes primarias, para lo cual se requiere conocer la situación actual del negocio, realizar un análisis exploratorio y mediante el uso de herramientas de visualización como Tableau comprender los datos con que se cuenta.

3.1.1 Situación actual del negocio

La empresa en estudio donde se desarrolla el proyecto de investigación es de tipo privada, pertenece al ámbito comercial y está enfocada a la venta de productos línea blanca, línea café, computadores, celulares entre otros, otorga facilidades de pago a sus clientes permitiéndoles realizar compras a crédito y realizar el pago en cuotas mensuales.

Analizando las ventas a crédito que la empresa ha realizado a nivel nacional entre el periodo 2016 hasta Abril 2019 se obtiene que un 70% de sus clientes tienen una factura a crédito y el 30% restante ha tenido más de una factura a crédito durante este periodo como se visualiza en la *Ilustración **5***.

La muestra de que se dispone es suficiente para identificar el comportamiento de pago de los clientes mes a mes con cuentas por cobrar. Por lo tanto, partimos de un total de 68.293 registros, correspondientes a facturas a crédito de 57.059 clientes en total, con datos estadísticos y de comportamiento de pago totalmente anonimizado.

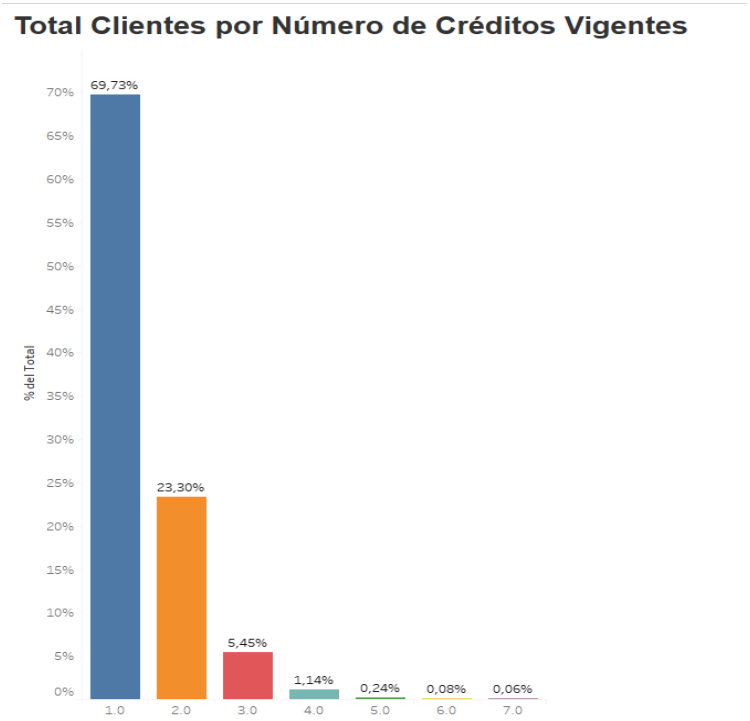


Ilustración 5. Total clientes por número de créditos. Elaboración Propia

Además como se observa en la *Ilustración* ***6*** las cuotas de pago son diferidas entre 1 a 30 meses siendo a 24 y 13 meses los planes más utilizados, por lo tanto un cliente puede tener más de una factura a crédito y con diferentes planes de financiamiento pero el pago debe corresponder al total de cuotas financiadas independientemente de la fecha de vencimiento de cada crédito.

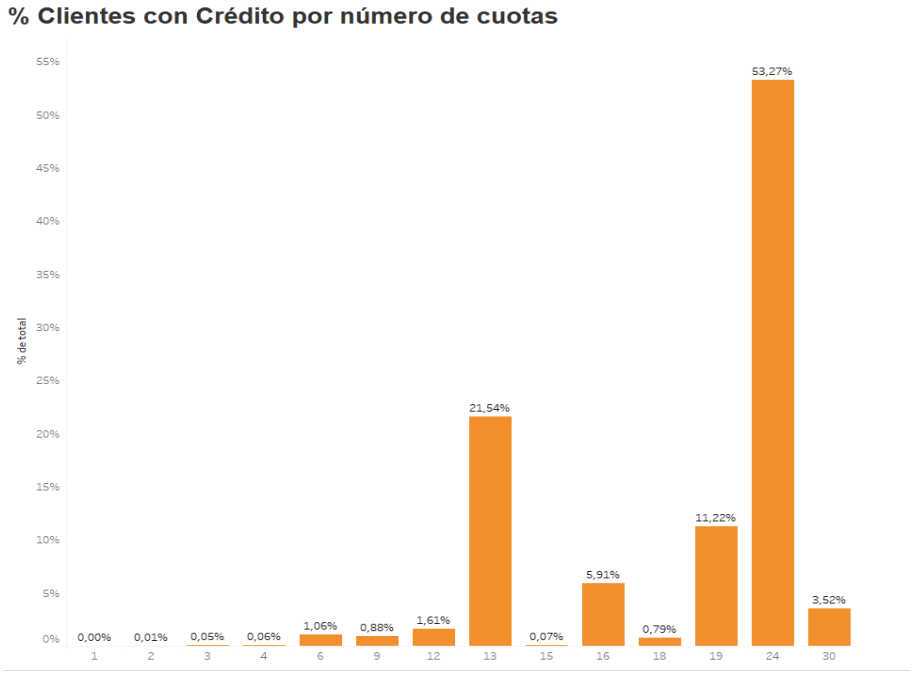


Ilustración 6. Total clientes por número de cuotas. Elaboración Propia

Los clientes presentan distintos comportamientos en su forma de pago, la gestión de recuperación de cartera realiza la mayor cantidad de acciones de cobranza a todos los clientes de la cartera por igual, sin distinguir a los clientes entre contactables, incontactables, buenos y malos e incluso pueden quedar clientes sin gestionar.

La cartera de cobranzas se gestiona cada mes con una asignación de primer nivel al call center (llamada telefónica al cliente), los clientes que pasaron el primer nivel de cobranza y que no han pagado sus deudas son asignados a los cobradores de ruta en base a los criterios que determina el Administrador de Cobranzas.

El call center se encarga de realizar contacto telefónico con el cliente y el cobrador de ruta de realizar la gestión domiciliaria, clasificando la gestión en efectiva o no efectiva, siendo efectiva una gestión cuando existe un compromiso de pago, contacto directo con el cliente, cliente ya canceló entre otros, y una gestión no efectiva es cuando el cliente es difícil, el cliente no contesta o no hay contacto con el cliente.

3.1.2 Comprensión de los datos

Para la clasificación de clientes de cartera por su probabilidad de pago es necesario identificar las variables que aportan mayor información para el modelo de predicción, agrupar las que aportan poca información y descartar las que no proporcionan información relevante para el análisis, también es necesario crear nuevas variables a partir de la información obtenida.

Es importante aclarar que la relación cliente-crédito puede ser de uno a varios, es decir un cliente puede tener varios créditos, razón por la cual es necesario crear variables calculadas para agrupar valores, cuotas, gestiones y el peor comportamiento de pago histórico del cliente en base a sus facturas de crédito, porque puede darse el caso de clientes con más de una factura a crédito con diferente comportamiento de pago.

3.1.2.1 Recolección de datos iniciales

Se parte de una muestra de datos totalmente anonimizados que corresponden a créditos otorgados entre el periodo 2016 hasta el 2018 pero con corte de pago hasta Abril de 2019 y que tengan saldos pendientes, para en base al comportamiento de pago histórico clasificar a los clientes como “si paga” o “no paga” la siguiente cuota mensual.

Empezamos identificando la fuente de datos y los atributos que intervienen en el proceso de crédito y cobranzas de los clientes acorde a la Tabla **3**.

Tabla 3. *Resumen de Recolección de datos iniciales. Elaboración Propia*

Fuente de datos	Atributos	Tipo	Descripción
Comprobante	Agencia	Nominal	Agencia donde se obtuvo el crédito
	Fecha de crédito	Fecha	Fecha en la que se otorga el crédito
	Monto del crédito	Numérico	Valor total del crédito
	Cuotas	Numérico	Número de cuotas que se difiere el crédito
	Garante	Nominal	Indica si el cliente tiene un garante para el crédito.
	Entrada	Numérico	Valor de la entrada que paga el cliente por el crédito. (total crédito=total factura – valor entrada)
	Fecha vencimiento	Fecha	Fecha que vence la primera cuota
	Línea de producto	Nominal	Línea del producto que aplica el crédito: blanca, café, celulares, motos, computadoras otros.
Cliente	Código del cliente	Numérico	Código de identificación del cliente
	Fecha de nacimiento	Fecha	Fecha de nacimiento del cliente
	Es empleado	Nominal	Indica si el cliente es empleado
	Dependientes	Numérico	Numero de cargas familiares
	Tipo de vivienda	Nominal	Tipo de vivienda
	Sexo	Nominal	
	Instrucción	Nominal	Instrucción académica.
	Actividad económica	Nominal	Actividad económica del cliente
Pago comprobante	Fecha de pago	Nominal	Fecha de pago de cada uno de las cuotas del crédito
	Días de atraso	Numérico	Número de días que se atrasa o se adelante en el pago de las cuotas del crédito.
	Valor pagado	Numérico	Valor que ha pagado de la cuota mensual.
	Es pagado	Boolean	Indica si la cuota está pagada o pendiente.
Movimientos gestiones	Fecha gestión	fecha	Fecha de la gestión domiciliaria o telefónica
	Tipo de gestión	Nominal	Tipo de gestión que realiza el personal de cobranzas (efectiva o no efectiva)
Balance mensual	Segmento de cartera en mora	Nominal	Segmento de cartera en mora que la empresa clasifica al cliente cada mes.
	Cuotas pagadas	numérico	Numero de cuotas pagadas hasta el cierre de mes
	Cuotas en mora	Numérico	Numero de cuotas en mora hasta el cierre de mes.
	Valor por vencer	Numérico	Valor de la cuota que esta por vencer.
	Valor vencido	Numérico	Valor vencido que no ha pagado el cliente en su fecha de vencimiento.
	Saldo crédito	Numérico	Saldo pendiente de pago del crédito.

3.2 Preparación y pre procesamiento de datos.

Una vez obtenido los datos iniciales y con un conocimiento previo de su contenido, es necesario realizar una preparación de los datos antes de aplicar las distintas técnicas de minería de datos, lo que repercutirá en los resultados obtenidos que dependen en gran medida de la calidad y preparación de los datos. En esta fase se incluye la limpieza, formateo de datos, conversión de atributos categóricos a numéricos, transformación a datos calculados, pre selección de variables independientes y definición de clase respuesta; para continuar en la siguiente fase con el proceso de selección de atributos.

Todo el proceso ETL (extracción de datos de diferentes tablas transaccionales, la transformación y limpieza de datos) es programado en PSLQL – Oracle 11g, en donde se creó un paquete con distintos procesos y funciones para obtener los valores finales con lo que se va a trabajar, incluye limpieza de datos, formateo de datos, campos calculados, creación e inserción de datos en tablas resumen.

3.2.1 Limpieza de datos

En este paso se va a resolver redundancias, detección y eliminación de outliers, discriminar o reemplazar valores nulos.

A los siguientes campos se han aplicado limpieza de datos:

ESTADOCIVIL: se encontró 2 registros de clientes cuyos datos están en nulos. En este caso, al ser pocos los clientes con valores nulos, se realiza una corrección manual de los datos utilizando datos del historial.

TIPOVIVIENDA: se encontró 2 registros de clientes cuyos datos están en nulos. De igual forma, se realiza una corrección manual, al no encontrarse datos en el historial del cliente se elimina esos registros.

INSTRUCCIÓN: se encontró 2872 registros de clientes cuyos datos están en nulos. Al considerarse un gran número de registros con nulos y siendo una variable no relevante se decide no usar este campo.

AGENCIA: se encontró que la empresa cuenta con 90 agencias y existen casos de clientes con más de un crédito en agencias distintas y al ser una variable no relevante se decide no usar este campo.

OCUPACIONLABORAL: se encontró 4 registros de clientes cuyos datos están en nulos, al ser pocos registros con valores nulos, se realiza una corrección manual revisando este campo en el historial del cliente.

EDAD: se encontró un total de 43 registros con ouliers desde -5971 hasta 1844 por lo que son datos extremos que es imposible que sean valores correctos, se procede a revisar en la tabla histórica de CLIENTE la fecha nacimiento y calcular la edad a la fecha del crédito.

Usando consultas SQL Y funciones como count(), min() y max() de ORACLE se pudo obtener los valores mínimos y máximos, se encontró que existían valores que eran atípicos al resto de la muestra como el caso de algunos clientes que tenían más de 2 créditos vigentes y representan solo un 2.6% de la muestra total, por lo tanto se decidió excluirlos del análisis para evitar que esos valores de suma de cuotas y diferidos de cuotas afecte al análisis final, dejando solo en la muestra los datos de clientes con uno y dos créditos vigentes.

Tabla 4. *Porcentaje Total de Clientes por número de créditos.*

# créditos	Clientes por Crédito	Porcentaje
1	47619	83,46%
2	7957	13,95%
3	1241	2,17%
4	194	0,34%
5	33	0,06%
6	9	0,02%
7	6	0,01%
	57059	100,00%

Se trató de clasificar las gestiones de cobranzas en domiciliarias y telefónicas pero al existir una gran cantidad de tipos de gestiones que maneja la empresa y no se distinguen si son telefónicas o domiciliarias, se decidió agrupar el número de gestiones en la variable TOTALGESTIONES, ya que se cuenta con un total de 159 tipos de gestiones y solo 2 con la descripción “domicilio“ pero hay la posibilidad de que al visitar al cliente en el domicilio cayó dentro de otro tipo de gestión como por ejemplo “deja mensaje” y eso no se puede diferenciar.

Se define a la clase como una variable dicotómica que indica el comportamiento de pago del cliente basándose en una predicción de pagos mes a mes. La variable de respuesta se codifica como "pagada" si el cliente paga el monto total de la cuota mensual y como "no pagada" si el cliente no paga o paga solo una parte de la cantidad.

3.2.2 Transformación de datos

Este paso consiste en generar nuevos campos o atributos, en base a los campos iniciales identificados previamente, con la finalidad de facilitar el análisis y procesamiento de los datos, además de disminuir la existencia de errores que se pueden presentar en los

resultados que se generan. Tomando en cuenta que se tiene que agrupar los datos por cliente, debemos reducir los datos de las facturas, cuotas y valores pagados, analizando el peor comportamiento de pago del cliente para segmentarlo en la condiciones de mora y pago a tiempo. Los campos calculados se listan en la Tabla **5**.

Tabla 5 *Variables calculadas o construidas. Elaboración Propia.*

VARIABLE	DESCRIPCIÓN	CÁLCULO
EDAD	Edad del cliente	Cálculo de la edad entre la fecha de nacimiento y fecha de crédito.
CUOTAS	Número de cuotas que difiere el crédito	Si el cliente tiene más de un crédito vigente, sumatoria de cuotas por cada crédito.
ENTRADA	Valor de entrada que paga el cliente por crédito.	Si el cliente tiene más de un crédito vigente, sumatoria de los valores de entrada que el cliente pagó.
VALORCUOTA	Valor de cuota mensual a pagar.	Si el cliente tiene más de un crédito vigente, sumatoria de los valores de cuota que paga mensualmente.
VALORCREDITO	Valor del crédito otorgado.	Si el cliente tiene más de un crédito vigente, sumatoria de los valores de créditos otorgados.
CUOTASPAGADAS	Cuotas pagadas hasta el periodo de cierre.	Si el cliente tiene más de un crédito vigente, sumatoria de las cuotas pagadas por cada crédito.
CUOTASMORA	Cuotas en mora (que no pago en la fecha de vencimiento)	Si el cliente tiene más de un crédito vigente, sumatoria de las cuotas en mora por cada crédito.
VALORVENCIDO	Valor vencido que corresponde a las cuotas en mora.	Si el cliente tiene más de un crédito vigente, sumatoria de los valores vencidos de cuotas en mora por cada crédito.
VALORXVENCER	Valor por pagar en las próximas cuotas.	Si el cliente tiene más de un crédito vigente, sumatoria de los valores por vencer por cada crédito.
SALDOCREDITO	Saldo pendiente de pago entre lo pagado y lo que falta pagar.	Si el cliente tiene más de un crédito vigente, sumatoria de los saldos por cada crédito.
NUMCREDITOS	Total créditos vigentes	Sumatoria de créditos vigentes
MORAULTMES	Segmento de mora del mes anterior a la fecha de corte.	Se calcula el segmento de mora del mes anterior, si el cliente tiene más de un crédito, el peor segmento de mora.
MORAULTTRIM	Segmento de mora del último trimestre a la fecha de corte.	Se calcula el segmento de mora del último trimestre, si el cliente tiene más de un crédito, el peor segmento de mora

MORAULTSEM	Segmento de mora del último semestre a la fecha de corte.	Se calcula el segmento de mora del último semestre, si el cliente tiene más de un crédito, el peor segmento de mora
MORAULTANIO	Segmento de mora del último año a la fecha de corte.	Se calcula el segmento de mora del último año, si el cliente tiene más de un crédito, el peor segmento de mora
CUOTASPAGATIEMPO	Numero de cuotas que pagó a tiempo por cada crédito.	Sumatoria de cuotas que pagó a tiempo antes o durante la fecha de vencimiento.
CUOTASPAG1ATRAS	Número de cuotas que pagó con atraso de 1 a 30 días.	Sumatoria de cuotas que pagó con atraso de 1 a 30 días después de la fecha de vencimiento.
CUOTASPAG31ATRAS	Número de cuotas que pagó con atraso de 31 a 60 días.	Sumatoria de cuotas que pagó con atraso de 31 a 60 días después de la fecha de vencimiento.
CUOTASPAG61ATRAS	Número de cuotas que pagó con atraso de 61 a 90 días.	Sumatoria de cuotas que pagó con atraso de 61 a 90 días después de la fecha de vencimiento.
CUOTASPAG91ATRAS	Número de cuotas que pagó con atraso de 91 días en adelante	Sumatoria de cuotas que pagó con atraso de 91 días en adelante después de la fecha de vencimiento.
GESTIONESEFEC	Número de gestiones efectivas	Sumatoria total de gestiones efectivas
GESTIONESNOEFEC	Número de gestiones no efectiva	Sumatoria total de gestiones no efectivas
DIAVENCIMIENTO	Día de vencimiento	Si el cliente tiene más de un crédito vigente, se calcula una media para obtener la fecha de vencimiento.

3.2.3 Formatear Datos

Se sabe que las variables categóricas ocultan y enmascaran mucha información interesante en un conjunto de datos. Por lo tanto, es importante utilizar métodos probados para tratar con variables categóricas que pueden mejorar el desempeño del modelo. Si no se lo hace, muchas veces se perderá de encontrar las variables más importantes del modelo y extraer la mayor cantidad de información posible (Ray, 2015).

En base a ello vamos a usar métodos probados para tratar los atributos categóricos y transformarlos en numéricos de manera que no perdamos información valiosa y el desempeño del modelo sea el mejor (Ray, 2015).

Además se debe considerar que algunos algoritmos pueden trabajar con variables categóricas directamente pero muchos algoritmos de aprendizaje automático no funcionan directamente con variables categóricas y requieren que todos sus valores de entrada y

salida sean numéricos ya que producen mejores resultados (Brownlee, 2017). Ejemplo en Python la biblioteca "*sklearn*" requiere atributos numéricos (Ray, 2015).

LINEAPRODUCTO: La empresa tiene 21 líneas de producto entre ellas: blanca, café, motos, celulares, computadores y otras; Por lo tanto, podríamos tener 21 variables categóricas nominales que podrían afectar el rendimiento del modelo (Ray, 2015). Por experiencia usando la lógica de negocio se sabe que esas líneas de producto se pueden clasificar en alto riesgo, medio riesgo y bajo riesgo. Por lo tanto, utilizamos la técnica de combinar niveles en grupos similares y un orden natural para indicar el riesgo en el pago del crédito, realizamos una codificación numérica de 1= BAJO RIESGO ,2 = MEDIO RIESGO y 3=ALTO RIESGO

Agrupando en la línea de alto riesgo la línea de productos celulares porque se considera un bien que se deprecia rápido, e incluso el cliente puede no querer pagar por este motivo o por robo, pérdida o falla técnica y así cada línea según el riesgo en el pago de crédito.

ESTADOCIVIL: Para este caso se ha tomado la decisión de crear variables binarias usando el método Dummy Coding por cada nivel de la variable categórica, donde la presencia de un nivel toma el valor de 1 y la ausencia de un nivel toma el valor 0 (Ray, 2015). Las variables resultantes son: ESSOLTERO_N, ESCASADO_N, ESDIVORCIADO_N, ESVIUDO_N, ESUNIONLIBRE_N.

Se evitó convertir la variable categoría directamente porque el modelo puede tomar la importancia teniendo en cuenta el orden.

TIPOVIVIENDA**:** Esta variable categórica tiene 6 tipos de vivienda, lo cual constituye 6 niveles de la variable categórica. Usando el método de reducción de niveles podemos combinar algunos niveles basándonos en la lógica empresarial, que agrupa niveles similares y por la distribución de frecuencias; por lo tanto, combinamos los niveles que tienen una frecuencia inferior a 5% (Ray, 2015).

Combinamos “Propio con deuda” y “Propio sin deuda” en una sola categoría denominada “Propia”; combinamos “Arrienda”, “Anticresis” y “Alquila” en un sola categoría denominada “Arrendada” y dejamos la categoría “Familiar” tal como estaba porque tiene un alta distribución de frecuencias.

Por lo tanto, primero se redujo los niveles de las variables categóricas mediante el uso de la técnica de combinación de niveles y luego se utilizó la codificación binaria (*Dummy Coding*) por cada nivel resultante, creando las siguiente variables binarias VIVIENDAPROPIA_N,

VIVIENDARRENDADA_N, VIVIENDAFAMILIAR_N. Conocido este método también como *one-hot encoding* (Brownlee, 2017).

***OCUPACIONLABORAL*:** Esta variable categórica tiene 8 tipos o niveles. Igual que la variable anterior se va a combinar los niveles por la lógica empresarial y la distribución de frecuencias. Combinamos “Independiente”, “Comerciante Formal”, “Comerciante Informal” y “Negocio Propio” en el nivel “Independiente”, combinamos “otros”, “Giros” y “Montepio” en el nivel “Otros” y dejamos la categoría “Dependiente” como tal.

Igual que la variable anterior luego de reducir los niveles por agrupación se realizó la codificación binaria (*Dummy Coding*) creando las siguientes variables binarias DEPENDIENTELABORAL_N, INDEPENDIENTELABORAL_N y OTROSLABORAL_N.

***SEXO*:** Para esta variable categórica, se usa el método de codificación entera 0= Femenino y 1=Masculino.

***MORAULTMES, MORAULTTRIM, MORAULTSEM Y MORAULTANIO*:** Estas variables categóricas indica los segmentos por días de mora (retraso en el pago) del cliente, en base a los datos históricos de pago de cliente se puede sustituir estas variables categóricas por las respectivas variables numéricas que tienen los días de mora del último mes, ultimo trimestre, último semestre y último año y que a su vez son variables calculadas.

ESEMPLEADO y RESULTADO (atributo clase): Estas variables pueden ser tratadas con el método de Codificación Binaria (*Dummy Coding*), la presencia de un nivel se representa con 1 y la ausencia se representa con 0 (Ray, 2015). Quedando 1=SI, 0=NO.

***ULTIMAGESTION**:* Para esta variable categórica, se usa el método de codificación entera con orden 0=sin gestión, 1=gestión no efectiva y 2=gestión efectiva.

Tabla 6. Resumen de la *Conversión de variables categóricas a numéricas.*

VARIABLE CATEGORICA	VARIABLE NUMERICA	METODO DE TRANSFORMACIÓN
LINEAPRODUCTO	LINEAPRODUCTO_N	Combine Levels using Business Logic
ESTADOCIVIL	ESSOLTERO_N	Dummy Coding
	ESCASADO_N	
	ESDIVORCIADO_N	
	ESVIUDO_N	
	ESUNIONLIBRE_N	
TIPOVIVIENDA	VIVIENDAPROPIA_N	Combine Levels using Business Logic and frequency or response rate, Dummy coding
	VIVIENDARRENDADA_N	
	VIVIENDAFAMILIAR_N	
OCUPACIONLABORAL	DEPENDIENTELABORAL_N	Combine Levels using Business Logic and frequency or response rate, Dummy coding
	INDEPENDIENTELABORAL_N	
	OTROSLABORAL_N	
SEXO	SEXO_N	Convert to number with Label Encoder
MORAULTMES	MAXDIASATRASOULTMES	Reemplazo por variable numérica que representa el total de días de atraso del último mes.
MORAULTTRIM	MAXDIASATRASOULTTRIM	Reemplazo por variable numérica, que representa el máximo de días de atraso en el último trimestre.
MORAULTSEM	MAXDIASATRASOULTSEM	Reemplazo por variable numérica, que representa el máximo de días de atraso en el último semestre.
MORAULTANIO	MAXDIASATRASOULTANIO	Reemplazo por variable numérica, que representa el máximo de días de atraso en el último año.
ESEMPLEADO	ESEMPLEADO_N	Dummy Coding
RESULTADO	RESULTADO_N	Dummy Coding
ULTIMAGESTION	ULTIMAGESTION_N	Convert to number with Label Encoder

3.2.4 Integrar y pre-seleccionar atributos

Luego de realizar los pasos de recolección, descripción, exploración, limpieza, integración y formateo de los atributos iniciales, se han pre seleccionado y construido una vista compacta con los registros a utilizar, así como los atributos finales pre seleccionados antes del proceso de selección de atributos y de minería de datos. Quedando con un total de 50.995 registros de clientes con facturas a crédito totalmente anonimizados.

Obsérvese la Tabla **7** en donde se muestra un resumen de cada campo, con su tipo, valor mínimo, valor máximo y el catálogo de valores que contendrá cada campo.

Tabla 7. *Pre selección de atributos. Elaboración propia*.

DESCRIPCION	VARIABLE	TIPO	MIN	MAX	VALOR
Calificación de pago del cliente (Clase predictora)	RESULTADO_N	Numérico	0	1	0=No Paga, 1=Si Paga
Estado civil convertido en variables binarias por cada nivel categórico.	ESSOLTERO_N	Numérico	0	1	0=No, 1=Si
	ESCASADO_N	Numérico	0	1	0=No, 1=Si
	ESUNIONLIBRE_N	Numérico	0	1	0=No, 1=Si
	ESVIUDO_N	Numérico	0	1	0=No, 1=Si
	ESDIVORCIADO_N	Numérico	0	1	0=No, 1=Si
Tipo de vivienda convertido en variables binarias por cada combinación de niveles categóricos resultantes.	VIVIENDAPROPIA_N	Numérico	0	1	0=No, 1=Si
	VIVIENDARRENDADA_N	Numérico	0	1	0=No, 1=Si
	VIVIENDAFAMILIAR_N	Numérico	0	1	0=No, 1=Si
Edad del cliente al obtener el crédito, en caso de tener más de un crédito, se obtendrá la edad del último crédito.	EDAD	Numérico	18	83	Rango de valores desde 18 a 83
Ocupación laboral del cliente convertido en variables binarias por cada combinación de niveles categóricos resultantes.	DEPENDIENTELABORAL_N	Numérico	0	1	0=No, 1=Si
	INDEPENDIENTELABORAL_N	Numérico	0	1	0=No, 1=Si
	OTROSLABORAL_N	Numérico	0	1	0=No, 1=Si
Indica si el cliente es empleado o no de la empresa	ESEMPLEADO_N	Numérico	0	1	1=si, 0=no
Sexo del cliente	SEXO_N	Numérico	0	1	1=Masculino, 0=Femenino
Línea de producto que aplica crédito el cliente, si tiene más de un crédito se obtendrá el que mayor riesgo tenga.	LINEAPRODUCTO_N	Numérico	1	3	1=Línea de producto de bajo riesgo, 2=línea de producto de medio riesgo, 3=Línea de producto de alto riesgo.
Total de cuotas que tiene el cliente de acuerdo al número de créditos vigentes.	CUOTAS	Numérico	2	60	Rango de valores desde 2 -60
Suma del valor de entrada que paga el cliente por	ENTRADA	Numérico	0	1030	Rango de valores desde 0 - 1030

cada crédito vigente					
Suma de cuotas mensuales que el cliente tiene con saldo>0	VALORCUOTA	Numérico	10	635	Rango de valores desde 10 - 635
Suma de todos los créditos otorgados al cliente que estén con saldo>0	VALORCREDITO	Numérico	57	7740	
Total de cuotas pagadas de los créditos con saldo>0	CUOTASPAGADAS	Numérico	0	46	
Total de cuotas en mora de total de créditos con saldo>0	CUOTASMORA	Numérico	0	46	
Total de valores de cuotas vencidas o en mora en créditos con saldo >0	VALORVENCIDO	Numérico	0	4550	
Total de valores de cuotas por vencer en créditos con saldo >0	VALORXVENCER	Numérico	0	5775	
Total de saldo pendiente de los créditos otorgados al cliente	SALDOCREDITO	Numérico	0	5775	
Total de créditos vigentes que tiene el cliente con saldo>0	NUMCREDITOS	Numérico	1	2	
Máximo días de atraso en el último mes	MAXDIASATRASOULTMES	Numérico	0	1166	Rango de valores de 0 a 1166
Máximo días de atraso en el último trimestre	MAXDIASATRASOULTTRIM	Numérico	0	1258	Rango de valores de 0 a 1258
Máximo días de atraso en el último semestre	MAXDIASATRASOULTSEM	Numérico	0	1258	Rango de valores de 0 a 1258
Máximo días de atraso en el último año	MAXDIASATRASOULTANIO	Numérico	0	1258	Rango de valores de 0 a 1258
Numero de cuotas que el cliente ha pagado a tiempo sin retraso dentro de la fecha de vencimiento del crédito.	CUOTASPAGATIEMPO	Numérico	0	38	Rango de valores de 0 a 38
Numero de cuotas que el cliente ha pagado con atraso de 1 a 30 días después de la fecha de vencimiento del crédito.	CUOTASPAG1ATRAS	Numérico	0	38	

Numero de cuotas que el cliente ha pagado con atraso de 31 a 60 días después de la fecha de vencimiento del crédito.	CUOTASPAG31ATRAS	numérico	0	20	
Numero de cuotas que el cliente ha pagado con atraso de 61 a 90 días después de la fecha de vencimiento del crédito.	CUOTASPAG61ATRAS	Numérico	0	14	
Numero de cuotas que el cliente ha pagado con atraso de 91 a 240 días después de la fecha de vencimiento del crédito.	CUOTASPAG91ATRAS	Numérico	0	21	
Número total de gestiones de cobranzas efectivas que tiene el cliente durante todo el periodo analizado. (telefónicas y domiciliarias)	GESTIONESEFEC	Numérico	0	190	
Número total de gestiones de cobranzas no efectivas que tiene el cliente durante todo el periodo analizado. (telefónicas y domiciliarias)	GESTIONESNOEFEC	Numérico	0	269	
indica si tiene ultima gestión de cobranzas y si es efectiva o no efectiva	ULTIMAGESTION_N	Numérico	0	2	0=sin gestión, 1=gestión no efectiva, 2= gestión efectiva
Día De vencimiento, corresponde al día de cada mes que el cliente debe pagar.	DIA_VENCIMIENTO	Numérico	1	31	

3.3 Selección de atributos

Se va a proceder a discriminar las variables pre-seleccionadas usando técnicas de selección de atributos, ya que las ventajas de realizar una selección de atributos son varias, desde el aumento en la velocidad de ejecución y modelos más simples que ayudan a un mejor conocimiento y entendimiento de los datos (García Gutiérrez, 2016).

Se considera un atributo irrelevante cuando el conocimiento de su valor no aporta nada para predecir la clase objetivo. Los atributos redundantes son aquellos que no aportan suficiente información y por lo tanto pueden ser eliminados. La redundancia es definida en términos de dependencia que existe entre los atributos, por lo tanto los atributos que son altamente correlacionados se dicen que son redundantes y se puede prescindir de aquellos (García Gutiérrez, 2016).

En este caso, entre las variables que describen el comportamiento de pago de un cliente, algunas pueden afectar las predicciones a medida que se proporciona información, mientras que otras pueden no ser necesarias e incluso redundantes. Por lo tanto el objetivo de selección de atributos es identificar los atributos que son irrelevantes y/o redundantes y eliminarlos para obtener las ventajas antes mencionadas.

3.3.1 Técnicas de selección de atributos.

Aplicamos las técnicas de selección de atributos explicadas en la **sección 2.6** usando la herramienta Weka. Se usarán 3 métodos de selección de atributos, uno de ellos con una evaluación de subconjunto de atributos y el resto con una evaluación unitaria de atributos.

- ***CfsSubsetEval*** con el método de búsqueda *Bestfirst*, valida valores numéricos, nominales discretos y continuos y es considerado un método de evaluación de subconjuntos de atributos basado en correlación.
- ***CorrelationAttributeEval*** con el método de búsqueda *Ranker*, evalúa el valor de cada atributo midiendo su correlación (*Pearson*) con respecto a la clase, valida valores numéricos, nominales discretos y continuos. Empíricamente solo seleccionaremos los atributos o variables que tienen una correlación moderada alta con valores cercanos a 1 o -1 y descartamos los cercanos a 0 usando como umbral limite 0.3 para seleccionar los atributos relevantes y descartar el resto de atributos no considerados relevantes.
- ***InfoGainAttributeEval*** al igual que el anterior evaluador se usa el método de búsqueda *Ranker*, evalúa el valor de cada atributo midiendo la ganancia de información (entropía) con respecto a la clase. Los valores que aportan más información tendrá un valor mayor de ganancia de información cercano a 1 y los

valores cercanos a cero serán los valores a excluir, empíricamente usamos como umbral limite 0.2 para seleccionar los atributos relevantes y descartar el resto.

Partimos de 38 atributos de entrada y 1 atributo de salida que es la clase a predecir, definidos en la fase de pre procesamiento de datos. Para realizar el proceso de selección de atributos usaremos los atributos de entrada conocidos como predictores. El resultado final se puede observar en la Tabla **8**.

Tabla 8. *Selección de Atributos. Elaboración Propia.*

Variable / Atributo	CfsSubsetEval (CSE)	CorrelationAttributeEval (CA)	InfoGainAttributeEval (IG)
ESSOLTERO_N			
ESCASADO_N			
ESUNIONLIBRE_N			
ESVIUDO_N			
ESDIVORCIADO_N			
VIVIENDAPROPIA_N			
VIVIENDARRENDADA_N			
VIVIENDAFAMILIAR_N			
EDAD	X		
DEPENDIENTELABORAL_N			
INDEPENDIENTELABORAL_N			
OTROSLABORAL_N			
ESEMPLEADO_N			
SEXO_N			
LINEAPRODUCTO_N			
CUOTAS			
ENTRADA			
VALORCUOTA			
VALORCREDITO			
CUOTASPAGADAS			
CUOTASMORA	X	X	X
VALORVENCIDO	X	X	X
VALORXVENCER			X
SALDOCREDITO			
NUMCREDITOS			
MAXDIASATRASOULTMES	X	X	X
MAXDIASATRASOULTTRIM	X	X	X
MAXDIASATRASOULTSEM		X	X
MAXDIASATRASOULTANIO	X	X	X
CUOTASPAGATIEMPO	X	X	
CUOTASPAG1ATRAS			
CUOTASPAG31ATRAS		X	
CUOTASPAG61ATRAS		X	
CUOTASPAG91ATRAS	X		
GESTIONESEFEC		X	X
GESTIONESNOEFEC		X	X
ULTIMAGESTION_N	X	X	X
DIA_VENCIMIENTO			

Se puede observar en la selección de atributos (*CSE, CA, IG*) que existen atributos comunes que son seleccionados, que indican que tienen una mayor correlación y ganancia de información con respecto a la clase objetivo o respuesta, sin embargo la comparación de algoritmos no se realizará con aquellos atributos comunes, sino con cada uno de los conjuntos de datos seleccionados en los métodos de selección de atributos aplicado.

Para saber que técnica de selección de atributos es la más eficiente, se deberá ejecutar cada modelo con selección de atributos de cada técnica y comparar los resultados en base a métricas de evaluación.

3.4 Métodos de clasificación

Para llevar a cabo el objetivo de la presente investigación, se realizó minería de datos usando el paquete de software WEKA. *"(Waikato Environment for Knowledge Analysis)* es un entorno para experimentación de análisis de datos que permite aplicar, analizar y evaluar las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario" (García Gutiérrez, 2016, p. 2).

Se generó el fichero de datos dataset_completo.arff en el formato de datos de Weka (formado arff) y 3 dataset con los atributos seleccionados por cada método de selección de atributos, cada dataset contendrá un total de 52.995 registros que representan a los clientes con crédito que forma parte de nuestra muestra de estudio pre procesado.

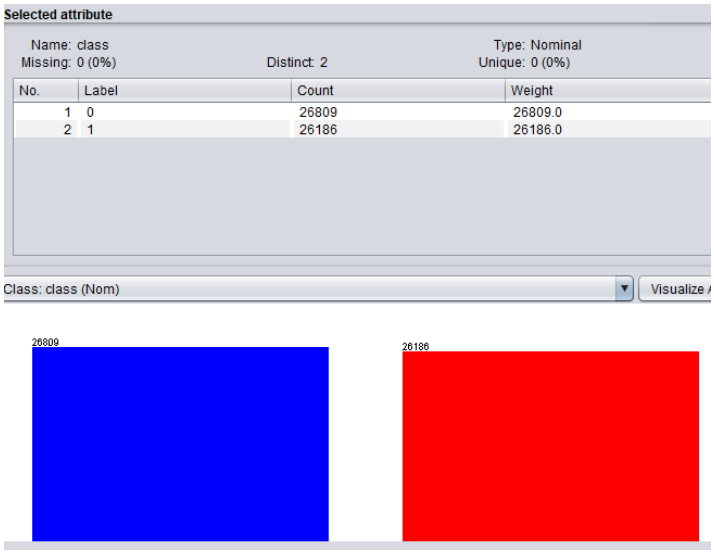


Ilustración 7. Distribución de la clase en Weka

Una vez terminado el proceso de pre-procesamiento y selección de atributos podemos constatar que el dataset completo se encuentra perfectamente balanceado al presentar la siguiente distribución.

Tabla 9. *Distribución de la clase del dataset completo.*

Clase	**Dataset Completo**	
0	50,61%	26.809
1	49,39%	26.186
TOTAL		52.995

Para el desarrollo del presente proyecto de investigación, se va a aplicar una comparativa de los algoritmos propuestos en la **sección 2.5** que corresponde a los diferentes algoritmos que tiene la aplicación WEKA, los datos ya han sido seleccionados y todos manejan atributos numéricos, ya que se realizó la transformación de datos categóricos a numéricos usando métodos probados para su conversión como codificador de etiquetas (Label Encoder), combinación de niveles (combine levels), codificación ficticia (dummy coding) entre otros (Ray, 2015).

Por lo que hemos escogido algoritmos que sean compatibles con los datos de entrada, ya que al tratar de discretizar variables continuas puede provocar un sesgo o perdida de información, porque se debe establecer rangos de valores cuyo criterio de asignación puede variar y esto a su vez puede alterar la exactitud de los algoritmos de clasificación asociados (Ray, 2015).

La evaluación y comparativa de los algoritmos se va a desarrollar con 4 conjuntos de datos, de los cuales, 3 conjuntos de datos tienen solos los atributos seleccionados en cada método de selección de atributos y el cuarto conjunto será el dataset completo con todos los atributos preseleccionados. Entonces se va a evaluar cada algoritmo de clasificación antes y después de la selección de atributos.

Todos los experimentos se han realizado con sus parámetros por defecto y utilizando la validación cruzada estándar de 10 veces (Kohavi, 1995). Con esta técnica se estima con que precisión el modelo predictivo se desempeñara en la práctica, limitando al mismo tiempo el problema de sobreajuste. El conjunto de datos se divide en 10 partes donde el proceso de entrenamiento se lleva a cabo en 9 partes y después de esto, el modelo resultante prueba a la izquierda la décima parte. Este proceso se repite 10 veces, entrenando los modelos en 9 partes diferentes. Al final del proceso, se promedian los 10 resultados de evaluación de las partes de prueba (Castillo, et al., 2016).

Todos los experimentos se llevaron a cabo en una CPU Intel ® Core (TM) i7-4500U a 2.40 Ghz con una memoria de 12GB con sistema operativo Windows 10 Profesional de 64 bits, WEKA versión 3.8.3, Java JRE_1.8.0_181 y Tableau 2018.3 con licencia de estudiante.

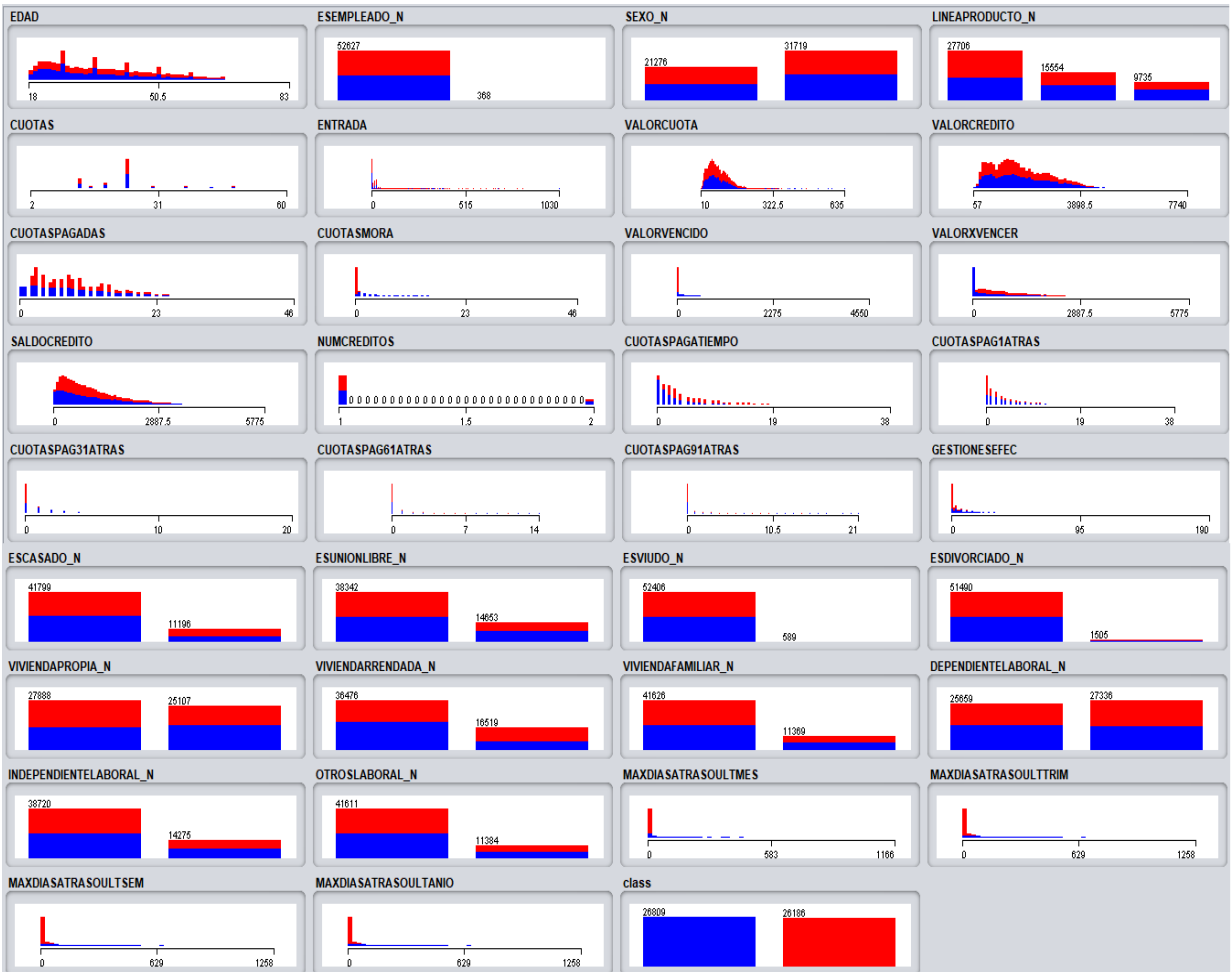


Ilustración 8. Visualización de la distribución de datos de los atributos del dataset completo en Weka.

3.4.1 Métodos de clasificación

A continuación se detallan los algoritmos de aprendizaje automático utilizados en el tema de investigación incluido su nombre estándar y el nombre utilizado en weka.

Algoritmos de aprendizaje automático lineal: Suponen que el atributo predicho es una combinación lineal de los atributos de entrada.

- Regresion Logistica: function.Logistic

Algoritmos de aprendizaje automático no lineales: Los algoritmos no lineales no hacen suposiciones solidas sobre la relación entre los atributos de entrada y el atributo de salida que se predice.

- Naive Bayes: bayes.NaiveBayes
- Árbol de decisión (C4.5): árboles J48
- K-vecinos más cercanos (KNN): lazy.IBk
- Máquina de vectores de soporte (SVM): functions.SMO
- Red neuronal: functions.MultilayerPerceptron

Algoritmos de aprendizaje automático conjunto: Los algoritmos de aprendizaje automático conjunto combinan las predicciones de múltiples modelos para hacer predicciones más robustas.

- Bosques Aleatorios (Random Forest): trees.RandomForest

Inicialmente todas las pruebas se realizaron mediante la ejecución de algoritmos de forma manual en Weka, para lo cual se ha usado el componente *KnowledgeFlow* de Weka que permite desarrollar un flujo de datos en batch para el tratamiento de datos usando los algoritmos de clasificación indicados.

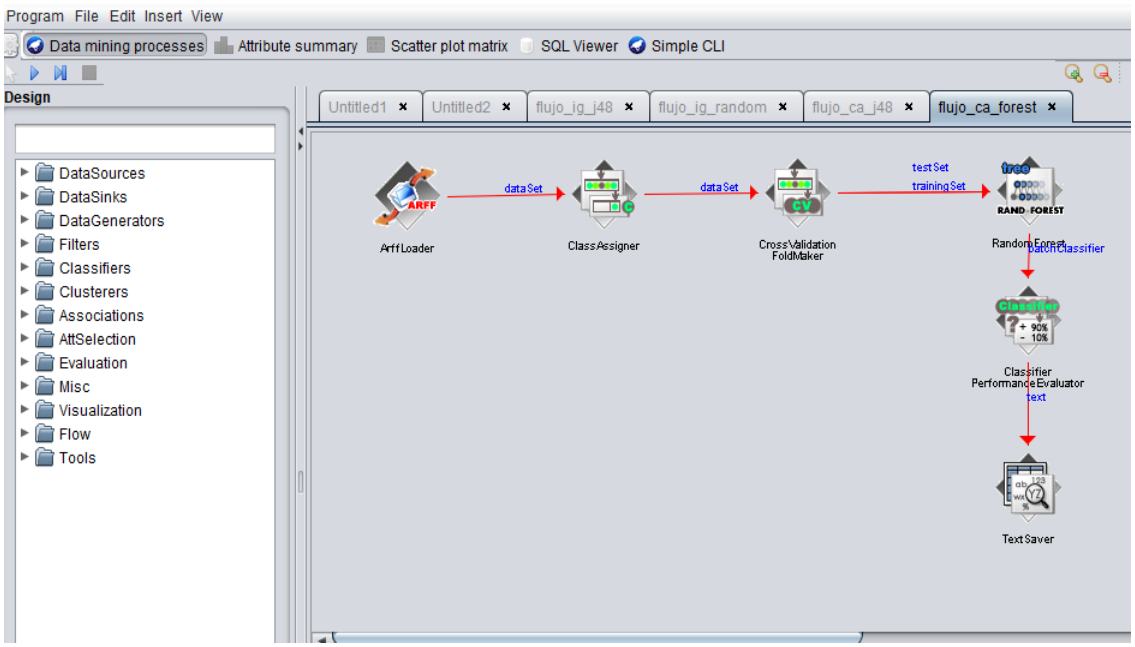
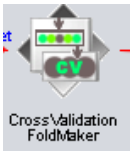


Ilustración 9. Flujo desarrollado para ejecución del algoritmo Random Forest en Weka.

Se puede seleccionar varios componentes WEKA de una barra de herramientas, colocarlos en un layout y conectarlos para formar un flujo de conocimiento para el procesamiento y análisis de datos. A continuación se explica los componentes principales usados para la generación del flujo de datos.

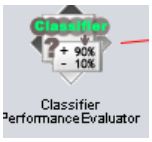
Este componente permite cargar ficheros de datos en formato .arff que utiliza Weka.

Este componente determina que atributo del conjunto de datos cargado pertenece a la clase.



Este componente divide el conjunto de datos en dos dataset uno entrenamiento y otro de pruebas usando validación cruzada de 10 folds.

Este componente determina el algoritmo de clasificación a usar.



Este componente a partir de los resultados obtenidos en el algoritmo de clasificación genera los datos de evaluación del mismo como matriz de confusión, recall, TP Rate, FP Rate, etc.

Este componente guarda los resultados del componente anterior en un documento de texto.

3.5 Conclusiones

Para el desarrollo del presente trabajo ha sido necesario seguir una metodología de trabajo para desarrollar y llevar a cabo cada una de las tareas que permitan determinar el algoritmo de clasificación de mayor precisión para el comportamiento de pago de clientes con cuentas por cobrar.

Se partió de la recolección de datos de facturas a crédito, utilizando análisis exploratorio y técnicas de visualización se identificó que un 30% de la muestra total de clientes tienen más de una factura a crédito, se identificó información histórica del pago de los clientes con cuentas por cobrar y las variables a usar que incluye datos demográficos, datos del crédito, datos de gestión de cobranzas y comportamiento histórico de pago de cada cliente.

De toda la información disponible, era necesario identificar las variables que aportan mayor información, agrupar las que aportan poca información y descartar las que no proporcionan información relevante para el modelo de predicción, todo esto en la fase de pre

procesamiento de datos que incluye limpieza, transformación, creación de campos calculados para el caso de clientes con más de una factura a crédito.

Para el tratamiento de variables categóricas ha sido necesario aplicar métodos de conversión a valores numéricos, para evitar que se pierda información valiosa al momento de realizar la predicción del modelo, evitando que se enmascare información interesante en el conjunto de datos, de manera que se extraiga la mayor cantidad de información posible.

Se ha realizado la aplicación de métodos de selección de atributos utilizando 3 técnicas de selección que evalúan tanto al conjunto completo de atributos como a cada atributo individualmente, todo esto con la finalidad de encontrar los atributos más relevantes y eliminar dichos atributos que no aportan en la predicción del modelo.

Para identificar la mejor técnica *Machine Learning* se ha considerado importante evaluar al conjunto completo de datos resultante de la fase de pre procesamiento y los 3 conjuntos de datos resultantes de aplicar cada técnica de selección de atributos. En el siguiente capítulo se lleva a cabo dicha comparativa y evaluación de técnicas con la finalidad de encontrar el modelo de predicción de mayor y mejor exactitud.

4 Experimentos y Resultados

Esta sección presenta los experimentos realizados en cada conjunto de datos, sus respectivas evaluaciones y conclusiones usando la herramienta de Business Intelligence Tableau 2018.3.6 de 64 bits, para la toma de decisiones en base a los datos obtenidos en la ejecución de cada algoritmo antes y después de la selección de atributos.

Todos los algoritmos de clasificación se ejecutaron usando validación cruzada de 10 folders que permite dividir el conjunto de datos, en datos de entrenamiento y validación en 10 iteraciones de manera que se evalúa todo el conjunto de datos y podemos conocer cómo se comporta nuestro modelo en la vida real. Todos los algoritmos de clasificación se ejecutaron con la parametrización por defecto que viene en Weka, dejando planteado realizar en el futuro una comparativa más completa con análisis de parámetros.

Partimos de que se quiere predecir el comportamiento de pago del cliente en cada cuota mensual, permitiendo predecir si el cliente paga o no la cuota del próximo mes. La variable de respuesta se codifica como “si paga” (1) si el cliente paga el monto total de la cuota y como “no paga” (0) si el cliente no paga o solo paga una parte de la cantidad que le corresponde a la cuota mensual.

La evaluación y comparativa de algoritmos se desarrolló con 4 conjunto de datos, de los cuales, 3 conjunto de datos tienen los atributos seleccionados en cada método de selección de atributos (CSE, CA y IG) respectivamente y el cuarto conjunto de datos es el dataset completo con todos los atributos pre seleccionados. Por lo tanto, se evalúa cada algoritmo de clasificación ejecutado con la herramienta Weka antes y después de la selección de atributos usando las métricas de evaluación de exactitud, precisión, área bajo la curva, media armónica y tasa de error.

4.1 Comparativa y Evaluación del Conjunto de datos completos

Nuestra clase positiva sería (1) y la clase negativa (0). Utilizando las métricas de evaluación indicadas en la **sección 2.7** se realiza la comparativa de evaluación de los algoritmos de clasificación indicados en la **sección 2.5** usando el dataset completo que contiene a todos los atributos pre-seleccionados en la **sección 3.2.4,** previo al uso de los métodos de selección de atributos. Los resultados obtenidos se indican en la Tabla **10**.

Tabla 10. *Métricas de Evaluación de algoritmos de clasificación con dataset completo*

Clasificador	Exactitud	Precisión	TP Rate	FP Rate	F-Measure	ROC Area	Tiempo
J48	0.930	0.928	0.931	**0.070**	0.929	0.950	00:00:33
Logistic	0.939	0.924	0.955	0.077	0.939	0.988	00:00:28
NaiveBayes	0.924	0.889	0.966	0.118	0.926	0.979	**00:00:03**
KNN	0.836	0.819	0.858	0.185	0.838	0.836	00:02:13
SVM	0.937	0.913	**0.964**	0.089	0.938	0.937	00:49:02
Multilayer	0.932	0.920	0.945	0.080	0.932	0.983	00:55:33
Random	**0.940**	**0.929**	0.953	0.071	**0.940**	**0.990**	00:03:28

Se puede observar que al aplicar los algoritmos de predicción sobre el dataset completo se obtiene un nivel de exactitud alto casi en todos los algoritmos aplicados a excepción de KNN (K vecinos más cercanos), con un nivel de exactitud arriba del 0.924 lo que corresponde al 92% de instancias clasificadas correctamente que es considerado un nivel alto de predicción.

La visualización de los datos de evaluación que se presenta en la *Ilustración **10*** nos muestra claramente una comparativa visual de la ejecución de cada algoritmo con el dataset completo, donde se puede observar que los algoritmos de *J48, Regresión Logistica, Naive Bayes y Random Forest* son los algoritmos que mejor desempeño tienen y en menor tiempo de ejecución.

El valor ROC es la representación de la compensación entre las tasas de falsos positivos y falsos negativos. F-Measure, que es otra medida para evaluar la efectividad, es la media armónica de la precisión y el recall lo cual *Random Forest* muestra un mejor desempeño aunque el tiempo de ejecución no parece ser el mejor.

Ilustración 10. Comparativa Visual de métricas de evaluación por Clasificador - Dataset Completo.

Pero es necesario evaluar la tasa error para determinar que algoritmo resulta ser conveniente para la predicción que se quiere realizar. Por lo tanto, en la Ilustración **11** se representa el valor absoluto de la suma de los errores obtenidos por cada algoritmo sumando los falsos positivos y falsos negativos y usando la *Ecuación **3*** que corresponde a la tasa de error, podemos revisar el error en la exactitud que se obtiene con cada algoritmo y evaluar sus valores.

.

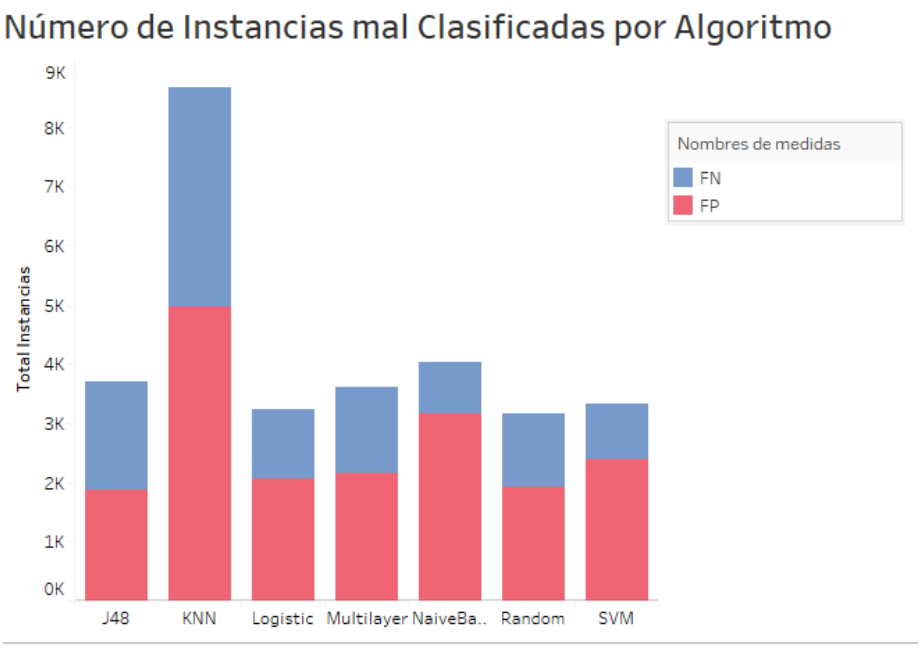


Ilustración 11. *Número de Instancias mal clasificadas por Algoritmo*

En general, como se puede observar los algoritmos tienen a sacar más falsos positivos, lo cual puede ser un riesgo en el modelo de predicción, porque se estaría identificando como clientes que pagan a clientes que pueden caer en mora y esto evitaría que el gestor de cobranzas realice una correcta gestión telefónica o domiciliaria en aquel cliente mal clasificado. Por lo que es mejor detectar "no paga" cuando realmente si paga (falso negativo) que detectar "si paga" cuando realmente no lo es (falso positivo).

También se debe considerar que un cliente que paga la próxima cuota y es clasificado como que "no paga" puede generar gastos en la gestión de cobranzas porque estaría el gestor de cobranzas gastando recursos en un cliente no peligroso de caer en mora y dejando de gestionar a los clientes con verdadero peligro de mora, inclusive puede afectar en la satisfacción de cliente que puede verse presionado sin justificación alguna.

Tabla 11. *Tasa de Error en la ejecución de algoritmos para el Dataset Completo.*

Clasificador	FP	FN	Total	FP%	FN%	Tasa de Error
J48	**1881**	1819	3700	**3.55%**	3.43%	6.98%
Logistic	2052	1181	3233	3.87%	2.23%	6.10%
Naive Bayes	3159	**880**	4039	5.96%	**1.66%**	7.62%
KNN	4971	3715	4.516	9.38%	7.01%	8.52%
SVM	2391	944	3364	4.51%	1.78%	6.35%
Multilayer	2157	1445	3602	4.07%	2.73%	6.80%
Random	1915	1243	**3158**	3.61%	2.35%	**5.96%**

En la Tabla **11** se puede observar que el mayor nivel de exactitud obtenido corresponde a al algoritmo Random Forest ya que su tasa de error es de 5.96% siendo el nivel más bajo de error, luego le sigue el algoritmo de Regresión Logística conocido en Weka como Logistic, luego el algoritmo de Maquina de Vectores de Soporte (SVM) y el algoritmo J48, todos ellos con una tasa de error menor al 7%.

Por otra parte se puede observar que el mayor índice de exactitud teniendo en cuenta solo el menor número de falsos positivos corresponde al algoritmo J48 con un 3.55% del total de la tasa de error y le sigue el algoritmo Random Forest con un 3.61% del total de la tasa de error y por último el algoritmo Logistic con un 3.87% que tienen una tasa de error menor al 4% con respecto a los falsos positivos.

En cambio el mayor índice de exactitud teniendo en cuenta solo el menor número de falsos negativos corresponde a los Algoritmos Naive Bayes, SVM y Logistic pero en cambio estos algoritmos tienen un tasa de error alta en los falsos positivos por lo que estarían descartados. Quedando solo los algoritmos Random Forest y J48 como de mejor precisión y menor de tasa de error en falsos positivos y falsos negativos.

Pero nos falta comparar cuanta mejora tienen dichos algoritmos con cada técnica de selección de atributos para encontrar la mejor técnica *Machine Learning* para la predicción del comportamiento de pago de clientes con cuentas por cobrar.

4.2 Comparativa y Evaluación conjunto de datos – selección de atributos CfsSubsetEval

Utilizando las métricas de evaluación indicadas en la **sección 2.7** se realiza la comparativa de evaluación de los algoritmos de clasificación indicados en la **sección 2.5** usando el dataset solo con los atributos seleccionados que corresponde a la técnica CfsSubsetEval (CSE) que se indica en la Tabla **8** por lo tanto el resto de atributos serán descartados para esta evaluación. Los resultados obtenidos se indican en la Tabla **12.**

Tabla 12. *Métricas de Evaluación de algoritmos de clasificación con selección de atributos (CSE)*

Clasificador	Exactitud (CSE)	Precision (CSE)	TP Rate (CSE)	FP Rate (CSE)	F-Measure (CSE)	ROC Area (CSE)	Tiempo (CSE)
J48	**0.94**	0.927	0.953	0.073	**0.94**	0.985	0:00:09
Logistic	0.936	0.913	**0.963**	0.09	0.937	0.987	0:00:10
NaiveBayes	0.924	0.883	0.977	0.127	0.927	0.98	**0:00:00**
KNN	0.915	0.912	0.916	0.086	0.914	0.945	0:01:26
SVM	0.937	0.912	0.965	0.091	0.938	0.937	0:09:25
Multilayer	0.936	0.928	0.944	0.072	0.936	0.985	0:09:11
Random	0.934	**0.932**	0.935	**0.066**	0.934	**0.988**	0:02:09

Se puede observar que al aplicar los algoritmos de predicción sobre el dataset con selección de atributos (CSE) el tiempo de ejecución se redujo considerablemente llegando a tener como máximo 9 minutos de ejecución con respecto a los 51 minutos en el algoritmo Multilayer Perceptron que se realizó en la comparativa anterior con el dataset completo.

También se puede observar que las métricas de evaluación de los algoritmos KNN y Multilayer Perceptron mejoraron considerablemente en todas las métricas de evaluación.

La visualización de los datos de evaluación que se presenta en la *Ilustración 12* nos muestra claramente una comparativa visual de la ejecución de cada algoritmo con el dataset con selección de atributos CSE, donde se puede observar que el algoritmo J48 tiene un mejor nivel en exactitud y F-Measure lo que representa una mejor exactitud en la clasificación mientras que Random Forest tiene la mejor precisión y mantiene la mejor área bajo la curva lo cual indica que la probabilidad del modelo es mucho más alta para la clase positiva que para la negativa.

Naive Bayes a pesar de que el tiempo de ejecución es el más bajo llegando a cero su rendimiento y precisión no es el mejor siendo J48 el algoritmo más adecuado por tiempo y rendimiento seguido por Random Forest que bajo considerablemente el tiempo de ejecución usando la técnica de selección de atributos (CSE), aunque todavía sigue siendo un poco alto en comparación con J48.

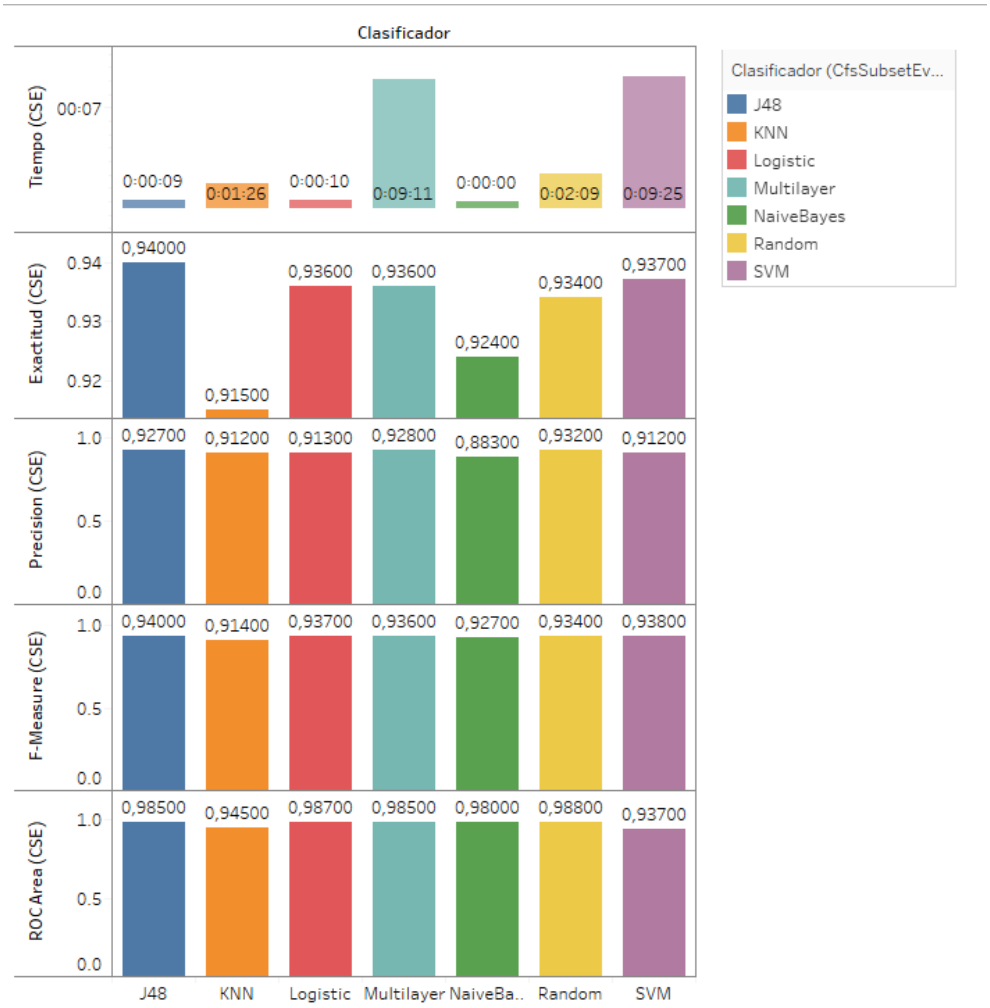


Ilustración 12. Comparativa Visual de métricas de evaluación por Clasificador con selección de Atributos (CSE)

De igual manera se evalúa la tasa de error para determinar que algoritmo resulta ser más conveniente para la predicción que se quiere realizar. Por lo tanto, en la *Ilustración* ***13*** se representa el valor absoluto de la suma de errores obtenidos por cada algoritmo sumando los falsos positivos y falsos negativos y usando la *Ecuación* ***3*** que corresponde a la tasa de error, podemos revisar el error en la exactitud tanto en la predicción de verdaderos como falsos positivos.

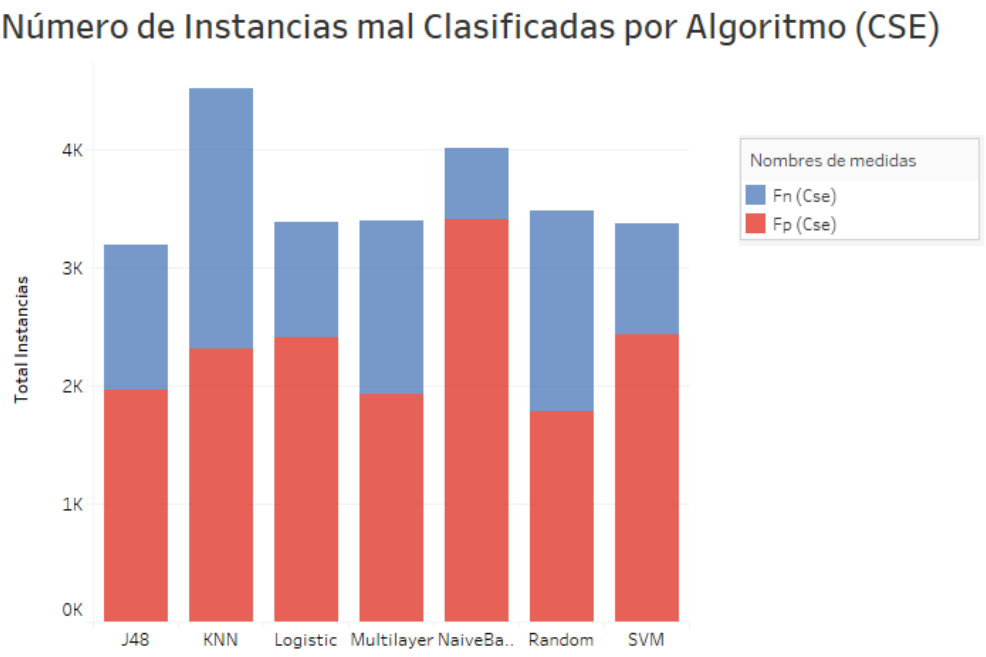


Ilustración 13. Número de Instancias mal clasificadas por Algoritmo (CSE)

Los falsos positivos se redujeron en los algoritmos KNN, Multilayer Perceptron y Random Forest mientras que en el resto de algoritmos aumentaron los falsos positivos pero no de forma considerable con respecto a la evaluación de algoritmos que se realizó con el dataset completo en la sección anterior.

Tabla 13. *Tasa de Error en la ejecución de algoritmos con selección de Atributos CSE*

Clasificador	FP (CSE)	FN (CSE)	Total (CSE)	FP%(CSE)	FN% (CSE)	Tasa Error % (CSE)
J48	1964	1223	**3187**	3.71%	2.31%	**6.01%**
Logistic	2406	975	3381	4.54%	1.84%	6.38%
NaiveBayes	3402	**602**	4004	6.42%	**1.14%**	7.56%
KNN	2316	2200	4516	4.37%	4.15%	8.52%
SVM	2436	928	3364	4.60%	1.75%	6.35%
Multilayer	1929	1465	3394	3.64%	2.76%	6.40%
Random	**1776**	1700	3476	**3.35%**	3.21%	6.56%

En la Tabla **13** se puede observar que el mayor nivel de exactitud obtenido corresponde al algoritmo J48 con una tasa de error de 6.01%, luego le sigue SVM con 6.35%, Logistic con 6.38%, Multilayer Perceptron con 6.40% y Random Forest con 6.56%. Todos con una tasa de error menor a 7%, la técnica *CseSubsetEva*l con selección de atributos considera a J48

como el algoritmo de mejor rendimiento a diferencia de la evaluación anterior con el dataset completo que consideró a Random Forest.

Por otra parte se puede observar que el mayor índice de exactitud teniendo en cuenta solo la tasa de error en los falsos positivos corresponde al algoritmo *Random Forest* con un 3.35%, luego le sigue *Multilayer Perceptron* con 3.64% y J48 con 3.71%, todos ellos con una tasa de error menor a 4% con respectos a los falsos positivos.

En cambio el mayor índice de exactitud teniendo en cuenta solo el menor número de falsos negativos corresponde a los algoritmos Naive Bayes con 1.14%, luego SVM 1.75% y Logistic 1.84% pero en cambio tienen un mayor número de falsos positivos por lo que estarían descartados. Quedando los algoritmos J48 con 2.31%, Multilayer con 2.76% y Random Forest con 3.21% como seleccionados.

4.3 Comparativa y Evaluación conjunto de datos – selección de atributos CorrelationAttributeEval

Utilizando las métricas de evaluación indicadas en la **sección 2.7** se realiza la comparativa de evaluación de los algoritmos de clasificación indicados en la **sección 2.5** usando el dataset solo con los atributos seleccionados que corresponde a la técnica *CorrelationAttributeEval (CA)* que se indica en la Tabla **8** por lo tanto el resto de atributos serán descartados para esta evaluación. Los resultados obtenidos se indican en la Tabla **14.**

Tabla 14. *Métricas de Evaluación de algoritmos de clasificación con selección de atributos (CA)*

Clasificador	Exactitud (CA)	Precisión (CA)	TP Rate (CA)	FP Rate (CA)	F-Measure (CA)	ROC Area (CA)	Tiempo (CA)
J48	**0.94**	0.927	0.954	0.074	**0.94**	0.983	0:00:13
Logistic	0.937	0.915	0.962	0.088	0.938	0.987	0:00:11
NaiveBayes	0.925	0.893	**0.965**	0.113	0.927	0.979	**0:00:00**
KNN	0.914	0.913	0.912	0.085	0.913	0.945	0:01:27
SVM	0.936	0.914	0.962	0.089	0.937	0.937	0:10:01
Multilayer	0.936	0.93	0.941	0.069	0.935	0.986	0:08:35
Random	0.937	**0.933**	0.94	**0.066**	0.936	**0.988**	0:01:55

Se puede observar que al aplicar los algoritmos de predicción sobre el dataset con selección de atributos aplicando la técnica *CorrelationAttributeEval* los valores de tiempo, precisión y exactitud tienen una ligera variación de mejora respecto a la técnica antes analizada de *CsfSubsetEval*. Los niveles de Exactitud mejoraron lo cual indica que hay un mayor número

de instancias clasificadas correctamente con respecto a la técnica de selección de atributos antes comparada, al igual que la precisión tuvo mejoras en algunos algoritmos lo que se traduce en una mejora tanto en la media armónica (F-measure) y el ROC área.

La visualización de los datos de evaluación que se presenta en la *Ilustración* ***14*** nos indica claramente la mejora indicada anteriormente con una comparativa visual de la ejecución de cada algoritmo con el dataset aplicado selección de atributos *CorrelationAttributeEval* (CA).

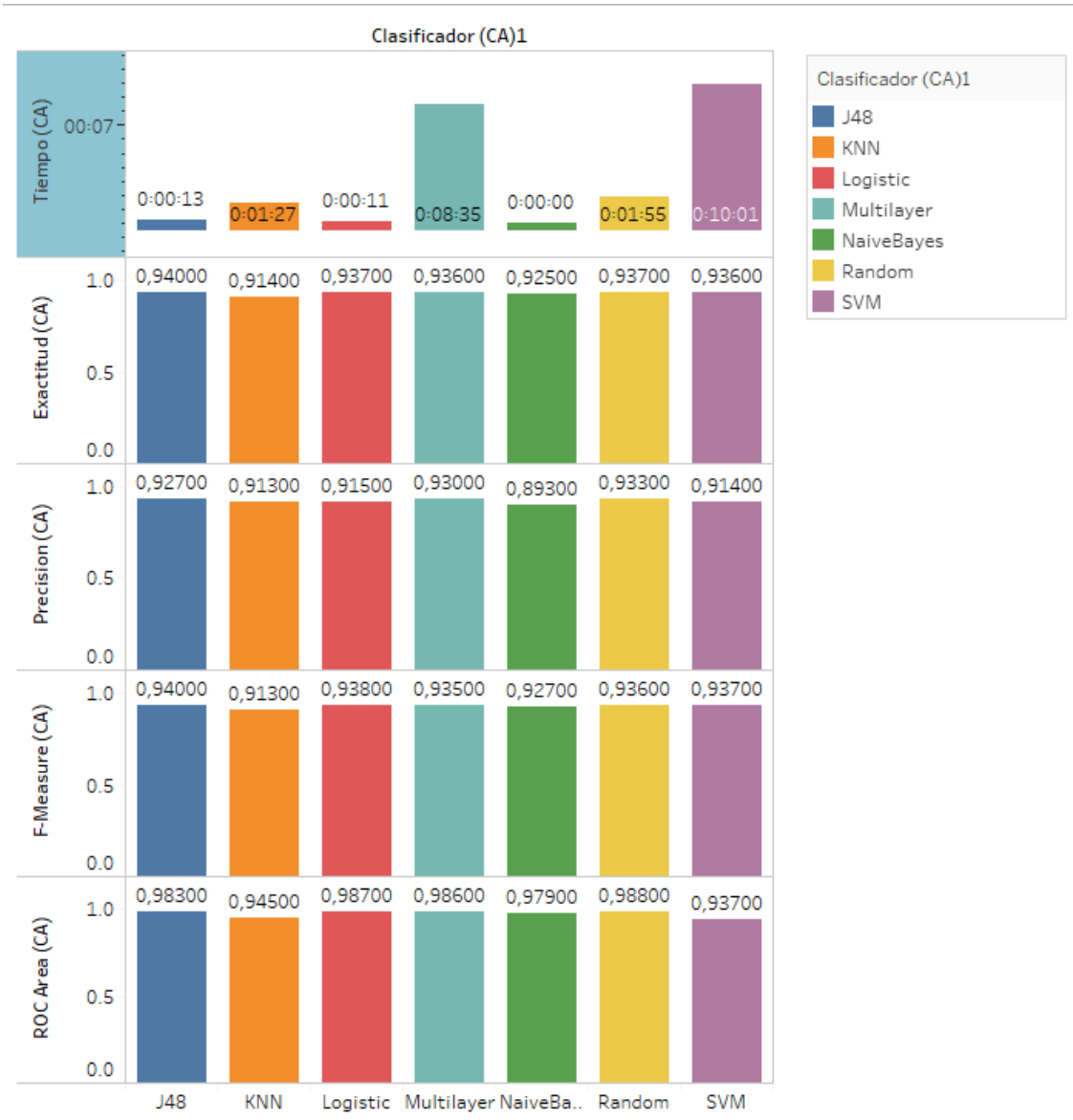


Ilustración 14. Comparativa Visual de métricas de evaluación por Clasificador con selección de Atributos (CA)

De igual manera se evalúa la tasa de error para determinar que técnica y algoritmo de *Machine Learning* es el mejor. Por lo tanto, en la *Ilustración* ***15*** se representa el valor absoluto de la suma de errores obtenidos por cada algoritmo sumando los falsos positivos y falsos negativos y usando la *Ecuación* ***3*** que corresponde a la tasa de error en la exactitud.

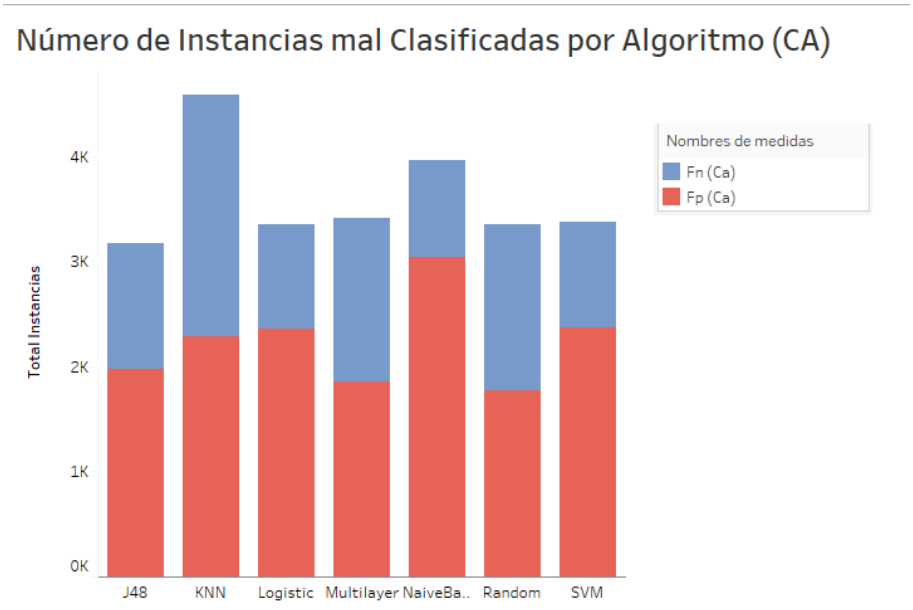


Ilustración 15. Número de Instancias mal clasificadas por Algoritmo (CA)

Los falsos positivos se redujeron en todos los algoritmos excepto J48 que se mantiene y los faltos negativos se redujeron en los algoritmos Random Forest, J48, KNN y Logistic con respecto a la técnica de selección de atributos de CsfSubsetEval (CSE).

Tabla 15. *Detalle del nivel de Exactitud de los algoritmos con selección de Atributos (CA)*

Clasificador	FP (CA)	FN (CA)	Total (CA)	FP% (CA)	FN% (CA)	Tasa de Error %(CA)
J48	1978	1192	**3170**	3.73%	2.25%	**5.98%**
Logistic	2355	**991**	3346	4.44%	**1.87%**	6.31%
NaiveBayes	3036	924	3960	5.73%	1.74%	7.47%
KNN	2286	2296	4582	4.31%	4.33%	8.65%
SVM	2375	1005	3380	4.48%	1.90%	6.38%
Multilayer	1858	1547	3405	3.51%	2.92%	6.43%
Random	**1769**	1582	3351	**3.34%**	2.99%	6.32%

En la Tabla **15** se puede observar que el nivel de exactitud obtenido corresponde al algoritmo J48 con una tasa de error de 5.98%, luego le sigue Logistic con 6.31% y Random Forest con 6.32%. Todos con una tasa de error menor a 7%. Con esta técnica de selección de atributos *CorrelationAttributeEval* (CA) se obtuvo una menor tasa de error con respecto a la técnica de selección de atributos con *CsfSubsetEval* (CSE).

Por otra parte se puede observar que el mayor índice de exactitud teniendo en cuenta solo la tasa de error en los falsos positivos corresponde al algoritmo Random Forest con un 3.34%, luego le sigue Multilayer Perceptron con 3.51% y J48 con 3.73%.

En cambio el mayor índice de exactitud teniendo en cuenta solo el menor número de falsos negativos corresponde a los algoritmos Naive Bayes con 1.74%, luego SVM 1.90% y Logistic 1.87% pero en cambio tienen un mayor número de falsos positivos por lo que estarían descartados. Quedando los algoritmos J48% con 2.25%, Multilayer Perceptron con 2.92% y Random Forest con 2.99%.

También si analizamos Multilayer Perceptron a pesar de haber tenido una mejora considerable en las métricas de exactitud y tasa de error el tiempo de ejecución es demasiado alto con respecto a J48 y Random Forest, por lo que debe ser descartado.

Igual que en las comparaciones anteriores a pesar de que Naive Bayes tenga un tiempo de ejecución bajo de 0 segundos su tasa de error es alta y su nivel de desempeño y precisión no es el mejor.

4.4 Comparativa y Evaluación conjunto de datos – selección de atributos InfoGainAttributeEval

Utilizando las métricas de evaluación indicadas en la **sección 2.7** se realiza la comparativa de evaluación de los algoritmos de clasificación indicados en la **sección 2.5** usando el dataset solo con los atributos seleccionados que corresponde a la técnica InfoGainAttributeEval (IG) que se indica en la Tabla **8** por lo tanto el resto de atributos serán descartados para esta evaluación. Los resultados obtenidos se indican en la Tabla **16.**

Tabla 16**.** *Métricas de Evaluación de algoritmos de clasificación con selección de atributos (IG)*

Clasificador	Exactitud (IG)	Precision (IG)	TP Rate (IG)	FP Rate (IG)	F-Measure (IG)	ROC Area (IG)	Time (IG).
J48	**0.939**	0.922	0.959	0.079	0.94	0.984	0:00:09
Logistic	0.936	0.913	0.963	0.089	**0.937**	0.986	0:00:11
NaiveBayes	0.925	0.889	**0.969**	0.118	0.927	0.98	**0:00:00**
KNN	0.918	0.918	0.916	0.08	0.917	0.935	0:01:14
SVM	0.936	0.913	0.963	0.09	**0.937**	0.937	0:06:57
Multilayer	0.935	0.93	0.94	0.069	0.935	0.985	0:07:33
Random	0.936	**0.932**	0.939	**0.067**	0.935	**0.988**	0:02:09

Se puede observar que al aplicar los algoritmos de predicción sobre el dataset con selección de atributos aplicando la técnica *InfoGainAttributeEval* (IG) los valores de tiempo de ejecución, precisión y exactitud disminuyen con una ligera variación a diferencia de Random Forest donde el tiempo de predicción del algoritmo aumenta ligeramente respecto a la técnica antes analizada de *CorrelationAttributeEval*.

La visualización de las métricas de evaluación con selección de atributos usando la técnica *InfoGainAttributeEval* la podemos observar en la *Ilustración* ***16***, donde la exactitud es similar en casi todos los algoritmos de clasificación al igual que la precisión que solo tiene una ligera variación con respecto a la técnica de selección de atributos *CorrelationAttributeEval* (CA) antes analizada.

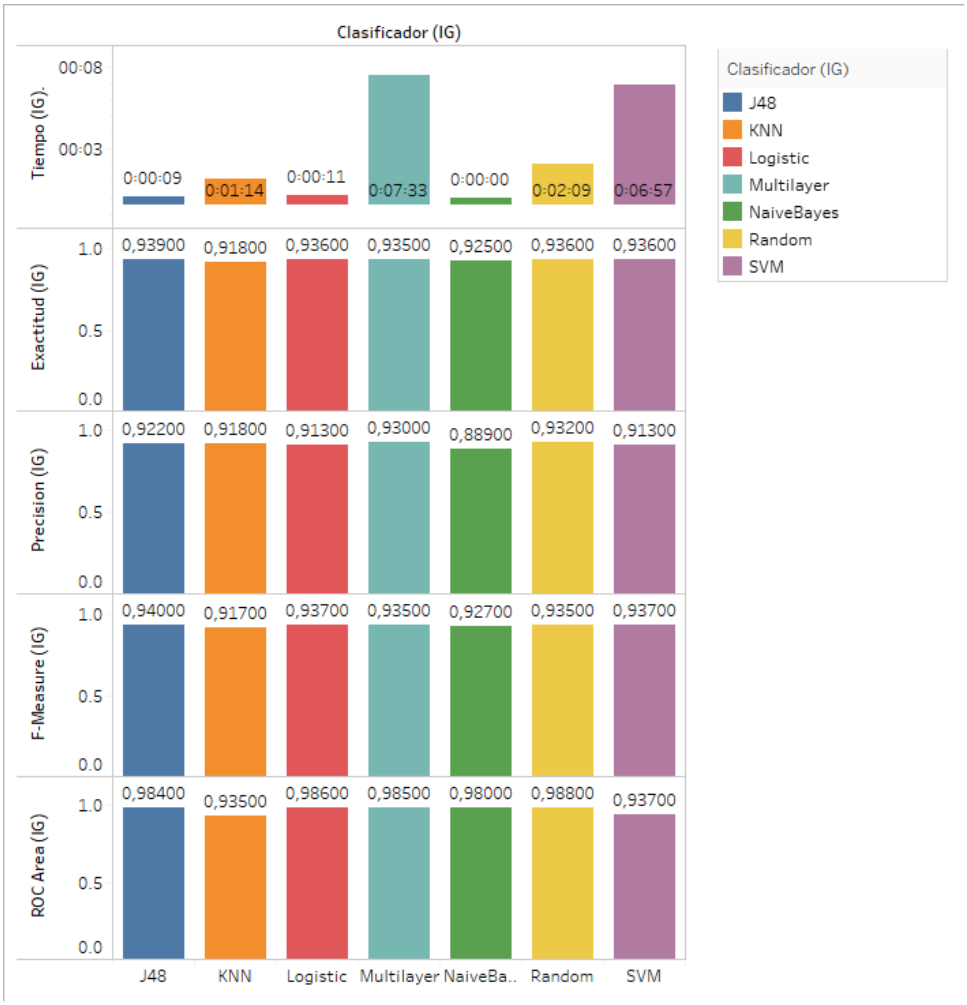


Ilustración 16. Comparativa Visual de métricas de evaluación por Clasificador con selección de Atributos (IG)

De igual manera se evalúa la tasa de error con la técnica de *InfoGainAttributeEval*, en la *Ilustración* ***17*** se representa el valor absoluto de la suma de errores obtenidos por cada algoritmo sumando los falsos positivos y falsos negativos y usando la *Ecuación* ***3*** podemos obtener la tasa de error en la predicción de cada algoritmo.

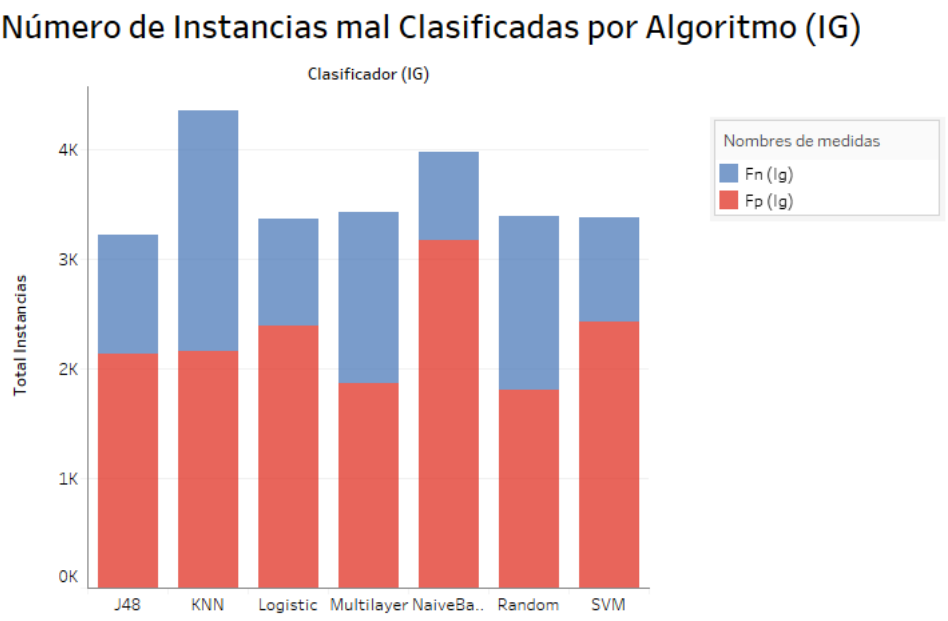


Ilustración 17. Número de Instancias mal clasificadas por Algoritmo (IG)

Los falsos positivos aumentaron en la mayoría de algoritmos excepto KNN que disminuye y Multilayer Perceptron que se mantiene con respecto a las dos técnicas de selección de atributos que son *CsfSubsetEval* y *CorrelationAttributeEval* antes analizadas.

Tabla 17. *Detalle del nivel de Exactitud de los algoritmos con selección de Atributos (IG)*

Clasificador	FP (IG)	FN (IG)	Total (IG)	FP% (IG)	FN% (IG)	Tasa de Error% (IG)
J48	2130	1086	**3216**	4.02%	2.05%	**6.07%**
Logistic	2392	974	3366	4.51%	1.84%	6.35%
NaiveBayes	3161	**815**	3976	5.96%	**1.54%**	7.50%
KNN	2152	2198	4350	4.06%	4.15%	8.21%
SVM	2418	956	3374	4.56%	1.80%	6.37%
Multilayer	1858	1562	3420	3.51%	2.95%	6.45%
Random	**1797**	1595	3392	**3.39%**	3.01%	6.40%

En la Tabla **17** se puede observar que la tasa de error aumenta con respecto a las dos otras técnicas de selección de atributos antes analizadas como son *CfsSubsetEval* y *CorrelationAttributeEval* y se mantiene J48 como el mejor algoritmo.

4.5 Conclusiones

Al terminar la fase de Experimentos y Resultados se puede concluir que tanto el algoritmo J48 y Random Forest han presentado los mejores resultados en las métricas de precisión, exactitud, área ROC y F-measure con los valores más altos y con una baja tasa de error tanto en la predicción de falsos positivos y falsos negativos. Si bien al evaluar el dataset completo se tiene buenos resultados, al momento de aplicar las técnicas de selección de atributos las métricas de evaluación en todos los algoritmos de comparación mejoran y el tiempo de ejecución junto con la tasa de error disminuyen considerablemente.

Al evaluar y comparar las métricas de evaluación en la ejecución de todos los algoritmos en cada técnica de selección de atributos se determina que la técnica de *CorrelationAttributeEval* ha presentado los mejores resultados tanto en el algoritmo J48 y Random Forest, el tiempo de ejecución y tasa de error ha sido un determinante clave para escoger a J48 como el mejor algoritmo de predicción sobre el comportamiento de pago de clientes con cuentas por cobrar.

J48 con *CorrelationAttributeEval* tiene una exactitud del 94,018%, precisión 92.27% y una tasa de error del 5.98% siendo 3.73% la tasa de error en los falsos positivos y 2.25% la tasa de error en los falsos negativos, todo esto se ejecuta en un tiempo de 00:00:13 a diferencia de Random Forest que tiene una exactitud de 93.7%, precisión de 93.3%y una tasa de error de 6.32% donde 3.34% corresponde a los falsos positivos y 2.99% corresponde a los falsos negativos y todo esto se ejecuta en un tiempo de 00:01:59.

En definitiva, los dos algoritmos tanto J48 y Random Forest con los parámetros por defecto de Weka presentan buenos resultados y bastantes similares pero en tiempo de ejecución J48 es mucho más rápido y por eso se plantea como la primera opción.

5 Conclusiones y Trabajos Futuros

Durante el desarrollo de este proyecto investigativo se ha trabajado en un sistema de predicción sobre el comportamiento de pago de clientes con cuentas por cobrar basado en datos históricos de una empresa de venta a crédito en todas sus líneas de producto tales como línea blanca, línea café, computadores, o celulares entre otros.

Para el desarrollo del mismo se ha seguido una metodología de trabajo basado en el proceso KDD - descubrimiento de conocimiento de bases de datos (Usama Fayyad, 1996). Se partió de la recolección y comprensión de los datos de clientes con crédito y sus pagos históricos, se ha trabajado en gran parte en la preparación y pre procesamiento de los datos que incluye la limpieza, formateo de los datos, creación de campos calculados, pre selección de variables y definición de la clase respuesta.

Además se ha visto la necesidad de tratar atributos categóricos de manera que no se pierda información valiosa y que el desempeño del modelo de predicción sea el mejor (Ray, Analytics Vidhya, 2015). Todo esto ha repercutido en los resultados obtenidos que dependen en gran medida de la calidad y preparación de los datos, y la importancia de contar con una metodología que ha facilitado el entendimiento y desarrollo de este proyecto.

La técnica de selección de atributos (*Feature Selection* - FS) es considerada una técnica de reducción de dimensionalidad efectiva, capaz de descartar los atributos menos relevantes o no útiles (García Gutiérrez, 2016). Al usar esta técnica se determinó las variables más relevantes para predecir el comportamiento de pago descartando aquellas variables no relacionadas o superfluas lo cual ha resultado en una mejora al aplicar los algoritmos de clasificación obteniendo mejores resultados en sus métricas de evaluación y reduciendo considerablemente el tiempo de ejecución por lo que se obtuvo modelos más simples.

Con la técnica de selección de atributos FS inclusive algunos algoritmos mejoraron notablemente en tiempo y exactitud a pesar de utilizar los parámetros por defecto de cada algoritmo de clasificación, por lo que se obtuvo modelos muchos más simples.

Realizar una comparativa de técnicas *Machine Learning* usando los algoritmos de clasificación J48, Regresión Logística, Naive Bayes, K vecino más cercanos (KNN), Maquina de Vectores de Soporte (SVM), Multilayer Perceptron y Random Forest sobre el dataset pre procesado y con 3 dataset resultantes de aplicar técnicas de selección de atributos ha permitido utilizar métricas de evaluación que validan el desempeño y rendimiento de cada modelo hasta encontrar el más adecuado para el trabajo propuesto.

También, ha sido de gran importancia conocer a profundidad el dominio del trabajo propuesto respecto a la gestión de cobranzas, lo que ha permitido tomar decisiones con criterio y basados en el modelo de negocio tanto en la fase de pre procesamiento como en la evaluación de técnicas de *Machine Learning*.

En conclusión, los algoritmos *Random Forest* y *J48* con la técnica de selección de atributos *CorrelationAtributeEval* han obtenido los mejores resultados en exactitud, precisión, F-measure, ROC área y tasa de error, con un bajo nivel de error en la predicción de falsos positivos ya que es importante tener un número pequeño de falsos positivos para evitar que se pierda de gestionar a los clientes deudores y con mayor riesgo de pago y que pueden ser clasificados erróneamente como clientes que pagan.

Siendo el algoritmo J48 la mejor opción para aplicar en la predicción del comportamiento de pago de clientes con cuentas por cobrar por el tiempo de ejecución y tasa de error que es menor a comparación de Random Forest. Sin embargo, no se debe descartar la opción de usar el algoritmo Random Forest y se podría verificar si al ajustar algunos parámetros en Weka el tiempo de ejecución de dicho algoritmo mejora considerablemente sin afectar la exactitud y precisión actual.

Analizando el árbol de decisión resultante de aplicar el algoritmo J48 sobre la muestra de datos con selección de atributos *CorrelationAttributeEval* como se indica en la *Ilustración **18*** y en los Anexos de la **Sección 7.1** queda demostrado que los arboles de decisión son una excelente herramienta de ayuda para la toma de decisiones para el usuario final, ya que son precisos y bastante fáciles de entender y aplicar.

También pueden proporcionar información sobre las variables de decisión que tienen mayor influencia en el comportamiento de pago, el modelo resultante ha determinado las siguientes variables como de gran importancia MAXDIASATRASOULTTRIM, VALORVENCIDO, CUOTASMORA, GESTIONESEFEC, CUOTASPAGATIEMPO, ULTIMAGESTION_N.

Dichas variables indican que es más seguro que un cliente pague la siguiente cuota cuando ha tenido un máximo de 5 días de atraso en el último trimestre, en caso de que el cliente tenga más de 5 días de atraso y un valor vencido de menos de 7$ sin cuotas de mora también tienen la probabilidad de pagar la siguiente cuota, lo cual indica la importancia de gestionar al cliente de manera oportuna para evitar retrasos que a la larga puede conducir a una mora, porque con el tiempo el cliente puede dejar de pagar las cuotas.

El modelo también indica que cuando un cliente tiene cuotas en mora es mejor que haya tenido un histórico de gestiones efectivas mayor a 7, lo que indica que el cliente puede tener un mayor compromiso de pago a pesar de tener atrasos en su pago.

Cuando el cliente tiene un máximo de días de atraso del último trimestre mayor a 81 días y cuotas en mora es más probable que ya no pague las siguientes cuotas. También se indica que mientras más días de atraso tenga y las gestiones efectivas sean menores a 7 el cliente tiende a no pagar su deuda.

Por lo tanto, la metodología propuesta que consiste en seleccionar atributos y aplicar diferentes modelos de clasificación se podría implementar en cualquier problema de clasificación, regresión o agrupación.

Como futuros trabajos, podría ser interesante comparar los resultados obtenidos en las predicciones con otros métodos de computación flexible como algoritmos genéticos, sistemas recomendadores o lógica difusa. Además de realizar una comparativa más completa que permita ajustar los parámetros de cada clasificador para obtener un mejor desempeño.

Pero algo muy importante sería plantearse un cuadro de mando integral usando *Business Intelligence* que permita clasificar y segmentar las cobranzas por antigüedad, cantidad total vencida, agrupar por tiempo de mora, geografía entre otras mediante indicadores estratégicos y ligados a planes de acción que permitan controlar y monitorear el proceso de cobranzas de una manera más eficiente.

Un cuadro de mando integral nos ayudaría a determinar cuan efectivo ha sido nuestro modelo de predicción en base al comportamiento de pago del cliente en el tiempo, como fase preventiva en la gestión de cobranzas.

Otro trabajo futuro es gestionar automáticamente las rutas de cobro para las gestiones domiciliarias usando algoritmos genéticos o heurísticos combinando con el presente trabajo de predicción de comportamiento de pago mensual.

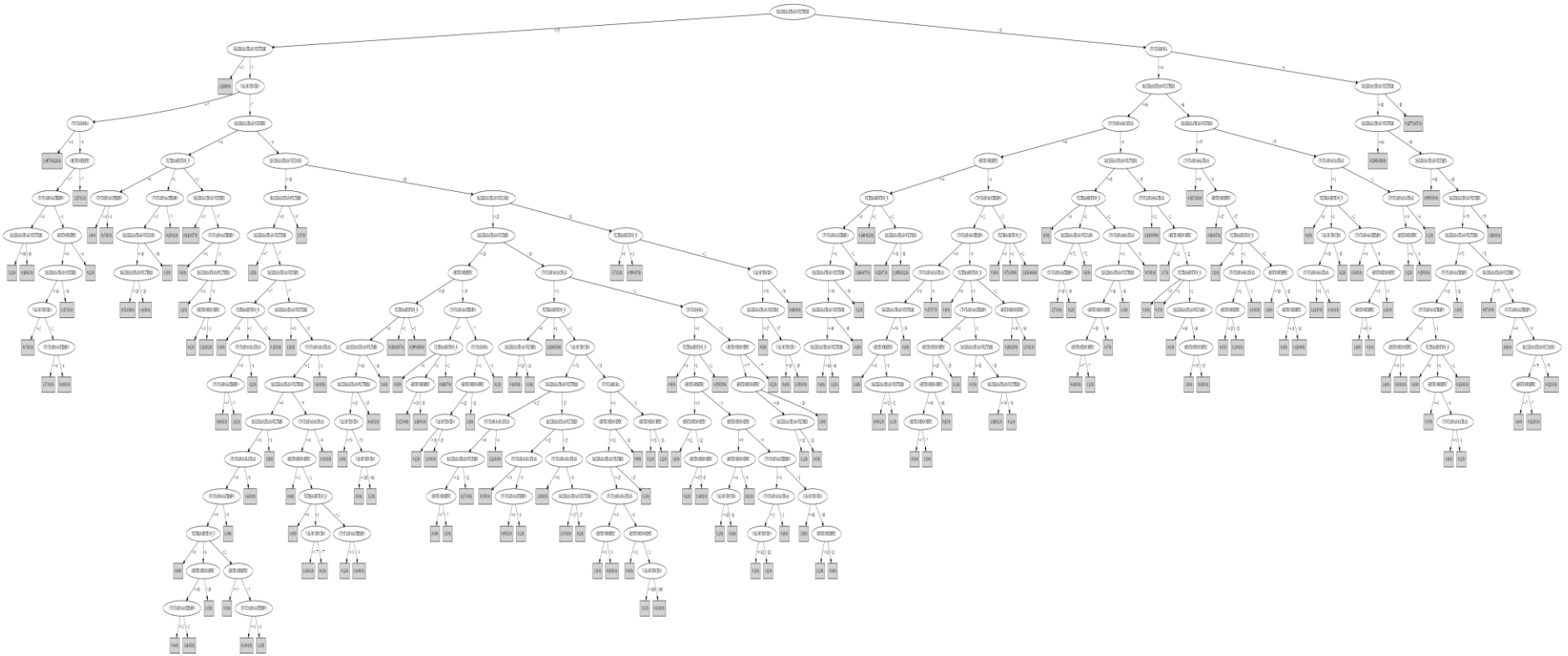


Ilustración 18. Árbol de Decisión generado en WEKA con J48 y FS - CorrelationAttributeEval. Elaboración Propia**.**

6 Referencias

Adnan, M., Husain, W., & Rashid, A. (20012). Hybric Approaches Using Decision Tree, Naive Bayes, Means and Euclidean Distances for Childhood Obesity Prediction. *International Journal of Software Engineering and Its Applications*, 99-106.

Brownlee, J. (13 de Julio de 2016). *How to Perform Feature Selection With Machine Learning Data in Weka*. Recuperado el Julio de 2019, de Machine Learning Mastery: https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/

Brownlee, J. (13 de Julio de 2016). *Machine Learning Mastery.* Recuperado el 07 de 2019, de How to Perform Feature Selection With Machine Learning Data in Weka: https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/

Brownlee, J. (28 de Junio de 2017). *Why One-Hot Encode Data in Machine Learning?* Obtenido de Machine Learning Mastery: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

Cabezas Arias, L. P., & Parra Romo, N. A. (2017). *Aplicación de un modelo Perceptrón Multicapa de redes neuronales artificiales para la clasificación del comportamiento de pago en clientes en mora en una entidad de cobranza.* Obtenido de Repositorio Institucional de la Universidad de las Fuerzas Armadas ESPE: http://repositorio.espe.edu.ec/handle/21000/13734

Castillo, P., Mora, A., Faris, H., Merelo, J., Garcia-Sanchez, P., Fernández-Ares, A., . . . García-Arenas, M. (2016). Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment. *Knowle dge-Base d Systems*.

Cheong, M. L., & SHI, W. (2018). Customer level predictive modeling for accounts receivable to reduce intervention actions. *Proceedings of the 14th International Conference on Data Science (ICDATA 2018)* (págs. 23-29). Las Vegas: Research Collection School Of Information Systems.

Corso, C. L. (s.f.). Aplicación de algoritmos de clasificación supervisada usando Weka. *Universidad Tecnológica Nacional, Facultad Regional Córdoba*.

Deloitte. (2012). *Tendencias de cobranza y recuperación de cartera en el sector financiero a partir de la crisis.* Mexico.

García Gutiérrez, J. A. (2016). *Comenzando con Weka: Filtrado y seleccion de atributos basada en su relevancia descriptiva para la clase.*

García Gutiérrez, J. A. (2016). Por qué debería mi empresa invertir en Big Data?: proyecto formativo y motivación.

García, J., Molina, J., Berlanga, A., Patricio, M., Bustamante, Á., & Padilla, W. (2018). *Ciencia de Datos Tecnicas Analiticas y Aprendizaje Automatico en un enfoque práctico.* Bogotá: Alfaomega Colombiana S.A.

Giudici, P. (2005). *Applied Data Mining: Statistical Methods for Business and Industry.* John Wiley & Sons.

Gnanambal, S., Thangaraj, M., Meenatchi, V., & Gayathri, V. (2018). Classification Algorithms with Attribute Selection: an evaluation study using WEKA. *Int. J. Advanced Networking and Applications*, 3640-3644.

Guyon, I., & Elisseff, A. (2003). *An introduction to varible and feature selection.*

Jácome Jara, M. A. (23 de Diciembre de 2014). *Construcción de un modelo estadístico para calcular el riesgo de deterioro de una cartera de microcréditos y propuesta de un sistema de gestión para la recuperación de la cartera en una empresa de cobranzas.* Obtenido de BIBDIGITAL: http://bibdigital.epn.edu.ec/handle/15000/9194

Jiménez, J., & Giraldo, J. (2013). Caracterización del Proceso de Obtención de Conocimiento y algunas metodologías para crear proyectos de minería de datos . *ResearchGate*, 42-44.

Keramati, A., & Yousefi, N. (2011). A proposed Classification of Data Mining Techniques in Credit Scoring. 416-424.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI'95 Proceedings of the 14th international joint conference on Artificial* (págs. 1137-1143). Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc.

Mishra, A. (24 de Febrero de 2018). *Metrics to Evaluate your Machine Learning Algorithm*. Recuperado el 24 de Julio de 2019, de Towards Data Science:

https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Morales Castro, J. A., & Morales, C. A. (2014). *Economía Digital Crédito y Cobranza.* Mexico: Grupo Editorial Patria.

Peiguang, H. (20 de Mayo de 2015). Predicting and Improving Invoice-to-Cash Collection Through. Massachusetts, Estados Unidos.

Ray, S. (26 de Noviembre de 2015). *Simple Methods to deal with Categorical Variables in Predictive Modeling*. Obtenido de Analytics Vidhya: https://www.analyticsvidhya.com/blog/2015/11/easy-methods-deal-categorical-variables-predictive-modeling/

Suca, C., Córdova, A., Condori, A., Cayra, J., & Sulla, J. (2016). Comparación de Algoritmos de Clasificación para la predicción de casos de obesidad infantil. 1-9.

Umaquinga, A., Saltos, T., & Peluffo, D. (2017). Herramientas de análisis de grandes volúmenes de datos para la toma de decisiones empresariales y fidelización del cliente. *Revista de Estrategias del Desarrollo Empresarial*, 31-37.

Usama Fayyad, G. P.-S. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Maganize*, 37-54.

Zeng, S., Melville, P., Lang, C., Boier-Martin, I., & Murphy, C. (2008). Using Predictive Analysis to Improve Invoice-to-Cash Collection. 1043-1050.

7 Anexos

7.1 Salida en Weka del algoritmo J48 y FS – CorrelationAttributeEval

MAXDIASATRASOULTTRIM <= 35

| MAXDIASATRASOULTTRIM <= 5: 1 (20296.0)000000

| MAXDIASATRASOULTTRIM > 5

| | VALORVENCIDO <= 7

| | | CUOTASMORA <= 0: 1 (4973.0/1226.0)

| | | CUOTASMORA > 0

| | | | GESTIONESEFEC <= 7

| | | | | CUOTASPAGATIEMPO <= 1

| | | | | | MAXDIASATRASOULTTRIM <= 10: 1 (2.0)

| | | | | | MAXDIASATRASOULTTRIM > 10: 0 (10.0/1.0)

| | | | | CUOTASPAGATIEMPO > 1

| | | | | | GESTIONESEFEC <= 6

| | | | | | | MAXDIASATRASOULTMES <= 14

| | | | | | | | VALORVENCIDO <= 2: 0 (5.0/1.0)

| | | | | | | | VALORVENCIDO > 2

| | | | | | | | | CUOTASPAGATIEMPO <= 6: 1 (7.0/1.0)

| | | | | | | | | CUOTASPAGATIEMPO > 6: 0 (3.0/1.0)

| | | | | | | MAXDIASATRASOULTMES > 14: 1 (17.0/1.0)

| | | | | | GESTIONESEFEC > 6: 0 (2.0)

| | | | GESTIONESEFEC > 7: 1 (17.0/2.0)

| | VALORVENCIDO > 7

| | | MAXDIASATRASOULTMES <= 6

| | | | ULTIMAGESTION_N = 0

| | | | | CUOTASPAGATIEMPO <= 4: 1 (4.0)

| | | | | CUOTASPAGATIEMPO > 4: 0 (5.0/1.0)

| | | | ULTIMAGESTION_N = 1

| | | | | CUOTASPAGATIEMPO <= 5

| | | | | | MAXDIASATRASOULTANIO <= 36

| | | | | | | MAXDIASATRASOULTTRIM <= 20: 0 (52.0/18.0)

| | | | | | | MAXDIASATRASOULTTRIM > 20: 1 (4.0/1.0)

| | | | | | MAXDIASATRASOULTANIO > 36: 1 (3.0)

| | | | | CUOTASPAGATIEMPO > 5: 0 (29.0/2.0)

| | | | ULTIMAGESTION_N = 2

| | | | | MAXDIASATRASOULTMES <= 5: 0 (111.0/37.0)

| | | | | MAXDIASATRASOULTMES > 5

| | | | | | CUOTASPAGATIEMPO <= 1: 0 (4.0)

| | | | | | CUOTASPAGATIEMPO > 1

| | | | | | | MAXDIASATRASOULTTRIM <= 6: 1 (2.0)

| | | | | | | MAXDIASATRASOULTTRIM > 6

| | | | | | | | GESTIONESNOEFEC <= 1: 0 (2.0)

| | | | | | | | GESTIONESNOEFEC > 1: 1 (11.0/2.0)

| | | MAXDIASATRASOULTMES > 6

| | | | MAXDIASATRASOULTANIO <= 10

| | | | | MAXDIASATRASOULTMES <= 9

| | | | | | MAXDIASATRASOULTTRIM <= 7: 1 (8.0)

| | | | | | MAXDIASATRASOULTTRIM > 7

| | | | | | | MAXDIASATRASOULTMES <= 7

| | | | | | | | ULTIMAGESTION_N = 0: 0 (0.0)

| | | | | | | | ULTIMAGESTION_N = 1

| | | | | | | | | CUOTASPAG31ATRAS <= 0

| | | | | | | | | | CUOTASPAGATIEMPO <= 5: 0 (8.0/1.0)

| | | | | | | | | CUOTASPAGATIEMPO > 5: 1 (2.0)

| | | | | | | | | CUOTASPAG31ATRAS > 0: 1 (2.0)

| | | | | | | | ULTIMAGESTION_N = 2: 0 (8.0/1.0)

| | | | | | | MAXDIASATRASOULTMES > 7

| | | | | | | | MAXDIASATRASOULTTRIM <= 8: 1 (13.0)

| | | | | | | | MAXDIASATRASOULTTRIM > 8

| | | | | | | | | CUOTASPAG31ATRAS <= 1

| | | | | | | | | MAXDIASATRASOULTTRIM <= 9

| | | | | | | | | | MAXDIASATRASOULTMES <= 8

| | | | | | | | | | | CUOTASPAG31ATRAS <= 0

| | | | | | | | | | | | CUOTASPAGATIEMPO <= 9

| | | | | | | | | | | | | ULTIMAGESTION_N = 0: 0 (0.0)

| | | | | | | | | | | | | ULTIMAGESTION_N = 1

| | | | | | | | | | | | | | GESTIONESNOEFEC <= 10

| | | | | | | | | | | | | | | CUOTASPAGATIEMPO <= 2: 0 (4.0)

| | | | | | | | | | | | | | | CUOTASPAGATIEMPO > 2: 1 (4.0/1.0)

| | | | | | | | | | | | | | GESTIONESNOEFEC > 10: 1 (3.0)

| | | | | | | | | | | | | ULTIMAGESTION_N = 2

| | | | | | | | | | | | | | GESTIONESEFEC <= 5: 0 (3.0)

| | | | | | | | | | | | | | GESTIONESEFEC > 5

| | | | | | | | | | | | | | | CUOTASPAGATIEMPO <= 4: 0 (3.0/1.0)

| | | | | | | | | | | | | | | CUOTASPAGATIEMPO > 4: 1 (2.0)

| | | | | | | | | | | | | CUOTASPAGATIEMPO > 9: 1 (4.0)

| | | | | | | | | | | | CUOTASPAG31ATRAS > 0: 1 (6.0/1.0)

| | | | | | | | | | | MAXDIASATRASOULTMES > 8: 1 (8.0)

| | | | | | | | | MAXDIASATRASOULTTRIM > 9

| | | | | | | | | CUOTASPAG31ATRAS <= 0

| | | | | | | | | | GESTIONESNOEFEC <= 2: 0 (4.0)

| | | | | | | | | | | GESTIONESNOEFEC > 2

| | | | | | | | | | | | ULTIMAGESTION_N = 0: 1 (0.0)

| | | | | | | | | | | | ULTIMAGESTION_N = 1

| | | | | | | | | | | | | VALORVENCIDO <= 77: 1 (4.0/1.0)

| | | | | | | | | | | | | VALORVENCIDO > 77: 0 (2.0)

| | | | | | | | | | | | ULTIMAGESTION_N = 2

| | | | | | | | | | | | | CUOTASPAGATIEMPO <= 3: 0 (2.0)

| | | | | | | | | | | | | CUOTASPAGATIEMPO > 3: 1 (6.0/1.0)

| | | | | | | | | | CUOTASPAG31ATRAS > 0: 0 (3.0/1.0)

| | | | | | | | | CUOTASPAG31ATRAS > 1: 0 (6.0/1.0)

| | | | | MAXDIASATRASOULTMES > 9: 1 (55.0)

| | | | MAXDIASATRASOULTANIO > 10

| | | | | MAXDIASATRASOULTANIO <= 28

| | | | | | MAXDIASATRASOULTMES <= 20

| | | | | | | GESTIONESEFEC <= 19

| | | | | | | | ULTIMAGESTION_N = 0

| | | | | | | | | MAXDIASATRASOULTMES <= 16

| | | | | | | | | | MAXDIASATRASOULTTRIM <= 15

| | | | | | | | | | | VALORVENCIDO <= 53: 1 (9.0)

| | | | | | | | | | | VALORVENCIDO > 53

| | | | | | | | | | | | VALORVENCIDO <= 102: 0 (3.0)

| | | | | | | | | | | | VALORVENCIDO > 102: 1 (2.0)

| | | | | | | | | | MAXDIASATRASOULTTRIM > 15: 0 (6.0/1.0)

| | | | | | | | | MAXDIASATRASOULTMES > 16: 1 (6.0)

| | | | | | | | ULTIMAGESTION_N = 1: 0 (361.0/167.0)

| | | | | | | | ULTIMAGESTION_N = 2: 0 (399.0/196.0)

| | | | | | | GESTIONESEFEC > 19

| | | | | | | | CUOTASPAGATIEMPO <= 7

| | | | | | | | | ULTIMAGESTION_N = 0: 0 (0.0)

| | | | | | | | | ULTIMAGESTION_N = 1

| | | | | | | | | | GESTIONESEFEC <= 33: 0 (25.0/4.0)

| | | | | | | | | | GESTIONESEFEC > 33: 1 (10.0/3.0)

| | | | | | | | | ULTIMAGESTION_N = 2: 0 (44.0/7.0)

| | | | | | | | CUOTASPAGATIEMPO > 7

| | | | | | | | | CUOTASMORA <= 1

| | | | | | | | | | GESTIONESNOEFEC <= 21

| | | | | | | | | | | VALORVENCIDO <= 29: 0 (2.0)

| | | | | | | | | | | VALORVENCIDO > 29: 1 (3.0/1.0)

| | | | | | | | | | GESTIONESNOEFEC > 21: 1 (8.0)

| | | | | | | | | CUOTASMORA > 1: 0 (2.0)

| | | | | | MAXDIASATRASOULTMES > 20

| | | | | | | CUOTASPAG31ATRAS <= 1

| | | | | | | | ULTIMAGESTION_N = 0

| | | | | | | | | MAXDIASATRASOULTMES <= 24: 0 (4.0/1.0)

| | | | | | | | | MAXDIASATRASOULTMES > 24: 1 (3.0)

| | | | | | | | ULTIMAGESTION_N = 1: 1 (110.0/34.0)

| | | | | | | | ULTIMAGESTION_N = 2

| | | | | | | | | VALORVENCIDO <= 74

| | | | | | | | | | MAXDIASATRASOULTTRIM <= 25

| | | | | | | | | | | CUOTASPAG31ATRAS <= 0

| | | | | | | | | | | | MAXDIASATRASOULTMES <= 21

| | | | | | | | | | | | | GESTIONESEFEC <= 7: 0 (4.0)

| | | | | | | | | | | | | GESTIONESEFEC > 7: 1 (3.0)

| | | | | | | | | | | | MAXDIASATRASOULTMES > 21: 1 (17.0/3.0)

| | | | | | | | | | | CUOTASPAG31ATRAS > 0: 1 (21.0/4.0)

| | | | | | | | | | MAXDIASATRASOULTTRIM > 25

| | | | | | | | | | MAXDIASATRASOULTMES <= 25

| | | | | | | | | | | CUOTASPAG31ATRAS <= 0: 0 (5.0/1.0)

| | | | | | | | | | | CUOTASPAG31ATRAS > 0

| | | | | | | | | | | | CUOTASPAGATIEMPO <= 6: 0 (9.0/1.0)

| | | | | | | | | | | | CUOTASPAGATIEMPO > 6: 1 (2.0)

| | | | | | | | | | MAXDIASATRASOULTMES > 25

| | | | | | | | | | | CUOTASPAG31ATRAS <= 0: 1 (8.0/1.0)

| | | | | | | | | | | CUOTASPAG31ATRAS > 0

| | | | | | | | | | | | MAXDIASATRASOULTTRIM <= 27: 1 (5.0/1.0)

| | | | | | | | | | | | MAXDIASATRASOULTTRIM > 27: 0 (2.0)

| | | | | | | | VALORVENCIDO > 74

| | | | | | | | | CUOTASMORA <= 1

| | | | | | | | | | GESTIONESNOEFEC <= 11

| | | | | | | | | | | MAXDIASATRASOULTMES <= 25

| | | | | | | | | | | | CUOTASPAG31ATRAS <= 0

| | | | | | | | | | | | | GESTIONESEFEC <= 1: 1 (3.0)

| | | | | | | | | | | | | GESTIONESEFEC > 1: 0 (8.0/1.0)

| | | | | | | | | | | | CUOTASPAG31ATRAS > 0

| | | | | | | | | | | | | GESTIONESNOEFEC <= 2: 0 (4.0)

| | | | | | | | | | | | | GESTIONESNOEFEC > 2

| | | | | | | | | | | | | | VALORVENCIDO <= 100: 1 (2.0)

| | | | | | | | | | | | | | VALORVENCIDO > 100: 0 (3.0/1.0)

| | | | | | | | | | | MAXDIASATRASOULTMES > 25: 1 (3.0)

| | | | | | | | | | GESTIONESNOEFEC > 11: 0 (9.0)

| | | | | | | | | CUOTASMORA > 1

| | | | | | | | | | GESTIONESNOEFEC <= 13: 0 (2.0)

| | | | | | | | | | GESTIONESNOEFEC > 13: 1 (2.0)

| | | | | | CUOTASPAG31ATRAS > 1

| | | | | | | | CUOTASMORA <= 1

| | | | | | | | | ULTIMAGESTION_N = 0: 0 (0.0)

| | | | | | | | | ULTIMAGESTION_N = 1

| | | | | | | | | | GESTIONESEFEC <= 3

| | | | | | | | | | | GESTIONESNOEFEC <= 12: 1 (6.0)

| | | | | | | | | | | GESTIONESNOEFEC > 12

| | | | | | | | | | | | GESTIONESNOEFEC <= 15: 0 (2.0)

| | | | | | | | | | | | GESTIONESNOEFEC > 15: 1 (4.0/1.0)

| | | | | | | | | | GESTIONESEFEC > 3

| | | | | | | | | | | GESTIONESNOEFEC <= 9

| | | | | | | | | | | | GESTIONESNOEFEC <= 6

| | | | | | | | | | | | | VALORVENCIDO <= 62: 1 (2.0)

| | | | | | | | | | | | | VALORVENCIDO > 62: 0 (4.0)

| | | | | | | | | | | | GESTIONESNOEFEC > 6: 0 (11.0)

| | | | | | | | | | | GESTIONESNOEFEC > 9

| | | | | | | | | | | | CUOTASPAGATIEMPO <= 1

| | | | | | | | | | | | | CUOTASPAG31ATRAS <= 2

| | | | | | | | | | | | | | VALORVENCIDO <= 112: 0 (2.0)

| | | | | | | | | | | | | | VALORVENCIDO > 112: 1 (2.0)

| | | | | | | | | | | | | CUOTASPAG31ATRAS > 2: 0 (6.0)

| | | | | | | | | | | | CUOTASPAGATIEMPO > 1

| | | | | | | | | | | | | VALORVENCIDO <= 68: 1 (8.0)

| | | | | | | | | | | | | VALORVENCIDO > 68

| | | | | | | | | | | | | | GESTIONESEFEC <= 12: 1 (2.0)

| | | | | | | | | | | | | | GESTIONESEFEC > 12: 0 (4.0)

| | | | | | | | | ULTIMAGESTION_N = 2: 0 (59.0/29.0)

| | | | | | | | CUOTASMORA > 1

| | | | | | | | | GESTIONESNOEFEC <= 77

| | | | | | | | | GESTIONESNOEFEC <= 20

| | | | | | | | | | MAXDIASATRASOULTMES <= 21: 1 (2.0)

| | | | | | | | | | MAXDIASATRASOULTMES > 21: 0 (3.0)

| | | | | | | | | GESTIONESNOEFEC > 20: 1 (8.0)

| | | | | | | | | GESTIONESNOEFEC > 77: 0 (2.0)

| | | | | MAXDIASATRASOULTANIO > 28

| | | | | | ULTIMAGESTION_N = 0: 1 (7.0/2.0)

| | | | | | ULTIMAGESTION_N = 1: 0 (99.0/35.0)

| | | | | | ULTIMAGESTION_N = 2

| | | | | | | VALORVENCIDO <= 56

| | | | | | | | MAXDIASATRASOULTSEM <= 17: 0 (8.0)

| | | | | | | | MAXDIASATRASOULTSEM > 17

| | | | | | | | | VALORVENCIDO <= 19: 0 (4.0)

| | | | | | | | | VALORVENCIDO > 19: 1 (19.0/3.0)

| | | | | | | VALORVENCIDO > 56: 0 (48.0/6.0)

MAXDIASATRASOULTTRIM > 35

| CUOTASMORA <= 0

| | MAXDIASATRASOULTTRIM <= 66

| | | CUOTASPAG61ATRAS <= 0

| | | | GESTIONESEFEC <= 6

| | | | | ULTIMAGESTION_N = 0

| | | | | | CUOTASPAGATIEMPO <= 1

| | | | | | | MAXDIASATRASOULTTRIM <= 56

| | | | | | | | MAXDIASATRASOULTTRIM <= 49

| | | | | | | | | MAXDIASATRASOULTTRIM <= 44: 0 (4.0)

| | | | | | | | | MAXDIASATRASOULTTRIM > 44: 1 (2.0)

| | | | | | | | MAXDIASATRASOULTTRIM > 49: 0 (8.0)

| | | | | | | MAXDIASATRASOULTTRIM > 56: 1 (2.0)

| | | | | | CUOTASPAGATIEMPO > 1: 1 (164.0/55.0)

| | | | | ULTIMAGESTION_N = 1: 0 (246.0/123.0)

| | | | | ULTIMAGESTION_N = 2

| | | | | | MAXDIASATRASOULTSEM <= 38: 0 (28.0/7.0)

| | | | | | MAXDIASATRASOULTSEM > 38: 1 (308.0/132.0)

| | | | GESTIONESEFEC > 6

| | | | | CUOTASPAGATIEMPO <= 2

| | | | | | CUOTASPAGATIEMPO <= 0

| | | | | | | CUOTASPAG31ATRAS <= 0

| | | | | | | | MAXDIASATRASOULTTRIM <= 54

| | | | | | | | | GESTIONESEFEC <= 8: 1 (4.0)

| | | | | | | | | GESTIONESEFEC > 8

| | | | | | | | | | MAXDIASATRASOULTTRIM <= 52: 0 (9.0/2.0)

| | | | | | | | | | MAXDIASATRASOULTTRIM > 52: 1 (2.0)

| | | | | | | | MAXDIASATRASOULTTRIM > 54: 0 (4.0)

| | | | | | | CUOTASPAG31ATRAS > 0: 0 (37.0/7.0)

| | | | | | CUOTASPAGATIEMPO > 0

| | | | | | | ULTIMAGESTION_N = 0: 0 (0.0)

| | | | | | | ULTIMAGESTION_N = 1

| | | | | | | | CUOTASPAGATIEMPO <= 1

| | | | | | | | | GESTIONESNOEFEC <= 29

| | | | | | | | | | GESTIONESNOEFEC <= 10

| | | | | | | | | | | GESTIONESNOEFEC <= 7: 0 (3.0)

| | | | | | | | | | | GESTIONESNOEFEC > 7: 1 (3.0)

| | | | | | | | | | GESTIONESNOEFEC > 10: 0 (13.0)

| | | | | | | | | GESTIONESNOEFEC > 29: 1 (2.0)

| | | | | | | | CUOTASPAGATIEMPO > 1

| | | | | | | | | MAXDIASATRASOULTTRIM <= 38: 0 (5.0)

| | | | | | | | MAXDIASATRASOULTTRIM > 38

| | | | | | | | | MAXDIASATRASOULTTRIM <= 56: 1 (10.0/2.0)

| | | | | | | | | MAXDIASATRASOULTTRIM > 56: 0 (2.0)

| | | | | | | ULTIMAGESTION_N = 2

| | | | | | | | GESTIONESNOEFEC <= 34: 0 (81.0/29.0)

| | | | | | | | GESTIONESNOEFEC > 34: 1 (5.0/1.0)

| | | | | CUOTASPAGATIEMPO > 2

| | | | | | ULTIMAGESTION_N = 0: 0 (0.0)

| | | | | | ULTIMAGESTION_N = 1: 0 (71.0/30.0)

| | | | | | ULTIMAGESTION_N = 2: 1 (134.0/64.0)

| | | CUOTASPAG61ATRAS > 0

| | | | MAXDIASATRASOULTTRIM <= 45

| | | | | ULTIMAGESTION_N = 0: 0 (3.0)

| | | | | ULTIMAGESTION_N = 1

| | | | | | MAXDIASATRASOULTANIO <= 72

| | | | | | | CUOTASPAGATIEMPO <= 10: 1 (7.0/1.0)

| | | | | | | CUOTASPAGATIEMPO > 10: 0 (2.0)

| | | | | | MAXDIASATRASOULTANIO > 72: 0 (4.0)

| | | | | ULTIMAGESTION_N = 2

| | | | | | CUOTASPAG61ATRAS <= 1

| | | | | | | MAXDIASATRASOULTTRIM <= 41

| | | | | | | | GESTIONESNOEFEC <= 10

| | | | | | | | | GESTIONESNOEFEC <= 7: 0 (4.0/1.0)

| | | | | | | | | GESTIONESNOEFEC > 7: 1 (2.0)

| | | | | | | | GESTIONESNOEFEC > 10: 0 (7.0)

| | | | | | | MAXDIASATRASOULTTRIM > 41: 1 (4.0)

| | | | | | CUOTASPAG61ATRAS > 1: 0 (5.0/1.0)

| | | | MAXDIASATRASOULTTRIM > 45

| | | | | CUOTASPAG61ATRAS <= 2: 1 (119.0/30.0)

| | | | | CUOTASPAG61ATRAS > 2

| | | | | | GESTIONESNOEFEC <= 22: 1 (7.0)

| | | | | | GESTIONESNOEFEC > 22

| | | | | | | ULTIMAGESTION_N = 0: 0 (0.0)

| | | | | | | ULTIMAGESTION_N = 1: 0 (5.0)

| | | | | | | ULTIMAGESTION_N = 2

| | | | | | | | MAXDIASATRASOULTANIO <= 60: 0 (3.0)

| | | | | | | | MAXDIASATRASOULTANIO > 60

| | | | | | | | | GESTIONESNOEFEC <= 35: 1 (4.0)

| | | | | | | | | GESTIONESNOEFEC > 35: 0 (4.0/1.0)

| | MAXDIASATRASOULTTRIM > 66

| | | MAXDIASATRASOULTSEM <= 97

| | | | CUOTASPAG61ATRAS <= 0: 0 (367.0/26.0)

| | | | CUOTASPAG61ATRAS > 0

| | | | | GESTIONESEFEC <= 17: 0 (86.0/17.0)

| | | | | GESTIONESEFEC > 17

| | | | | | ULTIMAGESTION_N = 0: 1 (1.0)

| | | | | | ULTIMAGESTION_N = 1

| | | | | | | CUOTASPAG61ATRAS <= 1

| | | | | | | | GESTIONESEFEC <= 28: 0 (3.0)

| | | | | | | | GESTIONESEFEC > 28: 1 (3.0/1.0)

| | | | | | | CUOTASPAG61ATRAS > 1: 1 (3.0/1.0)

| | | | | | ULTIMAGESTION_N = 2

| | | | | | | GESTIONESEFEC <= 20: 1 (6.0)

| | | | | | | GESTIONESEFEC > 20

| | | | | | | | GESTIONESEFEC <= 24: 0 (4.0)

| | | | | | | | GESTIONESEFEC > 24: 1 (16.0/6.0)

| | | MAXDIASATRASOULTSEM > 97

| | | | CUOTASPAG31ATRAS <= 2

| | | | | ULTIMAGESTION_N = 0: 0 (1.0)

| | | | | ULTIMAGESTION_N = 1

| | | | | | VALORVENCIDO <= 38

| | | | | | | CUOTASPAG61ATRAS <= 1: 1 (21.0/5.0)

| | | | | | | CUOTASPAG61ATRAS > 1: 0 (3.0/1.0)

| | | | | | VALORVENCIDO > 38: 0 (2.0)

| | | | | ULTIMAGESTION_N = 2

| | | | | | CUOTASPAGATIEMPO <= 0: 0 (4.0/1.0)

| | | | | | CUOTASPAGATIEMPO > 0

| | | | | | | GESTIONESNOEFEC <= 3

| | | | | | | | GESTIONESEFEC <= 4: 1 (6.0)

| | | | | | | | GESTIONESEFEC > 4: 0 (3.0)

| | | | | | | GESTIONESNOEFEC > 3: 1 (13.0)

| | | | CUOTASPAG31ATRAS > 2

| | | | | CUOTASPAG31ATRAS <= 6

| | | | | | GESTIONESEFEC <= 4: 1 (2.0)

| | | | | | GESTIONESEFEC > 4: 0 (19.0/4.0)

| | | | | CUOTASPAG31ATRAS > 6: 1 (2.0)

| CUOTASMORA > 0

| | MAXDIASATRASOULTTRIM <= 81

| | | MAXDIASATRASOULTTRIM <= 64: 0 (2008.0/330.0)

| | | MAXDIASATRASOULTTRIM > 64

| | | | MAXDIASATRASOULTMES <= 68: 0 (995.0/19.0)

| | | | MAXDIASATRASOULTMES > 68

| | | | | MAXDIASATRASOULTMES <= 79

| | | | | | MAXDIASATRASOULTTRIM <= 71

| | | | | | | CUOTASPAGATIEMPO <= 11

| | | | | | | | CUOTASPAGATIEMPO <= 1

| | | | | | | | | GESTIONESNOEFEC <= 9: 1 (6.0)

| | | | | | | | | GESTIONESNOEFEC > 9: 0 (8.0/2.0)

| | | | | | | | CUOTASPAGATIEMPO > 1

| | | | | | | | | ULTIMAGESTION_N = 0: 0 (0.0)

| | | | | | | | | ULTIMAGESTION_N = 1

| | | | | | | | | | GESTIONESEFEC <= 4: 0 (5.0)

| | | | | | | | | | GESTIONESEFEC > 4

| | | | | | | | | | | CUOTASPAG31ATRAS <= 1: 1 (4.0)

| | | | | | | | | | | CUOTASPAG31ATRAS > 1: 0 (2.0)

| | | | | | | | | ULTIMAGESTION_N = 2: 0 (11.0/1.0)

| | | | | | | CUOTASPAGATIEMPO > 11: 1 (3.0)

| | | | | | MAXDIASATRASOULTTRIM > 71

| | | | | | | MAXDIASATRASOULTMES <= 75: 0 (97.0/5.0)

| | | | | | | MAXDIASATRASOULTMES > 75

| | | | | | | | CUOTASPAGATIEMPO <= 0: 0 (10.0)

| | | | | | | | CUOTASPAGATIEMPO > 0

| | | | | | | | | MAXDIASATRASOULTANIO <= 79

| | | | | | | | | | GESTIONESEFEC <= 7: 1 (6.0)

| | | | | | | | | | GESTIONESEFEC > 7: 0 (12.0/3.0)

| | | | | | | | | MAXDIASATRASOULTANIO > 79: 0 (12.0/1.0)

| | | | | MAXDIASATRASOULTMES > 79: 1 (10.0/1.0)

| | MAXDIASATRASOULTTRIM > 81: 0 (20774.0/37.0)