Predicción de clientes efectivos en la gestión de carteras de cobranza castigada en la empresa InteliBPO S.A.S a través de modelos de aprendizaje automático

Autores

Yazmin Loraine Ortiz Numpaque

Elkin Felipe Ramírez González

Director

Elio H. Cables Pérez Pérez, Ph.D

Universidad Antonio Nariño

Facultad de Ingeniería de Sistemas

Especialización en Gobierno de Datos

Bogotá D.C.

2022

Índice de contenidos

Índice de Figuras

Índice de tablas

Resumen

Generar valor con los datos es un punto crucial en cualquier organización para destacar sobre la competencia y así mismo continuar innovando, por lo tanto, el presente proyecto de grado tiene como objeto hacer uso de algoritmos de aprendizaje automático supervisado enfocados en la clasificación como los son K-NN, máquinas de soporte vectorial, bosques aleatorios y árboles de decisión, con el propósito de predecir los clientes que serán efectivos en la gestión de cobranza a realizar por la organización InteliBPO S.A.S para las carteras en etapa de castigo, sobre la base de los datos registrados en el tercer trimestre del año 2022.

Se realiza una serie de procesos que incluyen el análisis exploratorio de los datos, selección de los atributos más representativos de los clientes que aporten valor a los modelos, una etapa de entrenamiento y finalmente un análisis de los resultados obtenidos con el propósito de seleccionar el modelo que sea más consistente con la predicción de registros efectivos y que pueda contribuir de forma positiva a la generación de estrategias de gestión más asertivas presentándose como una herramienta de apoyo a la gestión realizada por el área de operaciones de la organización.

Palabras clave: *predicción, algoritmos de aprendizaje automático.*

Abstract

Generating value with data is a crucial point in any organization to stand out from the competition and continue to innovate, therefore, this degree project aims to make use of supervised machine learning algorithms focused on classification such as K-NN, vector support machines, random forests and decision trees, with the purpose of predicting the clients that will be effective in the collection management to be carried out by the InteliBPO S.A.S organization for portfolios in the penalty stage, based on the data recorded in the third quarter of the year 2022.

A series of processes are carried out that include the exploratory analysis of the data, selection of the most representative attributes of the clients that add value to the models, a training stage and finally an analysis of the results obtained with the purpose of selecting the model. that it is more consistent with the prediction of effective records and that it can contribute positively to the generation of more assertive management strategies, presenting itself as a management support tool carried out by the organization's operations area.

Keywords: *prediction, machine learning algorithms.*

1. Introducción

Gracias al auge y crecimiento exponencial de la información, se evidencia con mayor frecuencia que las empresas otorgan gran valor a sus datos, pues entienden que el adecuado tratamiento de los mismos genera valor en la organización, mayor cantidad de ingresos, favorece la retención de clientes y un mejor entendimiento de su público objetivo, lo que conlleva a realizar con el presente proyecto un aporte que logre generar un impacto positivo a nivel estratégico.

Para el caso de estudio desarrollado, la contribución se enfoca a la gestión de carteras de cobranzas, orientado en específico a la realizada por la empresa InteliBPO S.A.S que se especializa en la gestión de grandes volúmenes de usuarios a través de canales digitales. Se busca optimizar y dar valor a la recuperación de cartera a través de la aplicación y uso de modelos de aprendizaje automático que contribuyan a predecir la efectividad en carteras en etapa de castigo en la organización, haciendo uso de algoritmos de clasificación como k-vecinos más próximos, máquinas de soporte vectorial, bosques aleatorios y árboles de decisión.

La implementación de modelos de aprendizaje automático permitió abrir la puerta a nuevas estrategias no solo de gestión en el marco operativo, sino también, a nivel de negocio.

2. Descripción y formulación del problema

La Compañía InteliBPO S.A.S es un Virtual Contact Center, prestadores de servicios de estrategias de comunicación B2B y B2C, dedicada en una de sus líneas de servicio, a la recuperación de cartera a través de gestiones estratégicas de contacto por medio de múltiples canales digitales.

Dentro de sus procesos, la unidad encargada de gestión llega a recibir mensualmente dependiendo del tipo de contrato carteras de recuperación en rangos desde los 10.000 usuarios hasta de 400.000 para ser gestionados diariamente.

En la actualidad los procesos de gestión son ejecutados por los gerentes de operaciones de manera empírica, es decir, generan las estrategias de gestión en base a sus conocimientos previos y/o experiencias con carteras de características similares, se infiere, de acuerdo a lo analizado, que durante el tratamiento dado a las bases de clientes puedan haber patrones y comportamientos de los clientes que son pasados por alto y que hace falta identificarlos con el fin de mejorar la efectividad mejorando los indicadores de recuperación de cartera. Adicionalmente, aunque se establecen inicialmente metas de efectividad junto al cliente que suministra la cartera, actualmente la organización no cuenta con un método que permita identificar si los registros serán efectivos y en base a esto enfocar los esfuerzos de gestión.

Teniendo en cuenta lo anteriormente expresado se identifica la siguiente problemática:

¿Cómo identificar los posibles clientes que serán efectivos en las carteras de cobranza a gestionar por InteliBPO S.A.S que están en etapa de castigo a partir de los datos registrados en el tercer trimestre del año 2022?

3. OBJETIVOS

3.1 Objetivo General

Predecir los clientes que serán efectivos en la gestión de cobranza a realizar por la organización de InteliBPO S.A.S para las carteras en etapa de castigo a través de modelos de aprendizaje automático sobre la base de los datos registrados en el tercer trimestre del año 2022.

3.2 Objetivos Específicos

- Identificar las variables más significativas del set de datos objeto de estudio a través de un análisis exploratorio.

- Estudiar diferentes algoritmos de aprendizaje automático supervisado que permitan realizar tareas de clasificación.

- Aplicar los algoritmos de aprendizaje automático referentes a clasificación en base a los datos de cartera de cobranza castigada del tercer trimestre del año 2022.

- Evaluar los resultados que generan el entrenamiento de los modelos de aprendizaje automático para la predicción de registros con gestión efectiva.

4. Marco referencia

4.1 Marco teórico

4.1.1 Gestión de cobranza

La cobranza hace referencia a todas aquellas acciones que llevan a cabo las entidades vigiladas con el propósito de recuperar la cartera a través del pago por parte de los clientes de las obligaciones que tienen asociadas. Las gestiones de cobranza y las actividades que esto involucra, se pueden ejecutar directamente o por medio de una empresa especializada que preste el servicio, cualquiera de ellas conforme a las normas vigentes garantizando el debido proceso y la protección de los derechos de los usuarios (Colcob, 2018).

La gestión de cobranza se trata de adelantar procesos eficientes que tienen por objeto la transformación de las cuentas por cobrar en liquidez, en el tiempo más corto, oxigenando de manera positiva el flujo de caja de la organización. El tratamiento eficiente en el cobro de la cartera de cobranza, se entiende como un proceso organizado de hitos, etapas o procesos que se ajustan a cada tipo de organización, región, sector o país, puesto que no todas las empresas poseen la misma naturaleza, como tampoco los deudores morosos lo son (Debitia, 2021).

La cobranza requiere de una política adecuada y enfocada a los distintos canales (llamadas telefónicas, mail, sms, whatsApp u otros medios) en este caso la empresa InteliBPO S.A.S con un enfoque digital y buscando seguir el estado de sus facturas o créditos, la obligación de pago y ofrecer opciones de normalización y pago trazara e implementara dichas políticas y lineamientos.

La finalidad en el proceso de cobranza se enfoca en el equilibrio entre cobrar rápido y mantener la relación con el cliente para evitar pérdidas en las entidades. Por lo tanto, con los problemas económicos que trajo consigo la pandemia, y a su vez la inflación con la subida de tasas de interés podrían conllevar a una posible recesión económica mundial, según un nuevo estudio integral del Banco Mundial, lo que obliga a implementar sistemas de cobro más eficaces, y un sistema que aporte valor requiere plantear modelos que permitan la toma de decisiones acertadas, eficientes, y que contemple un alto grado de automatización, además de contar con el desarrollo de variadas estrategias para asegurar el recaudo (Sun, Wiering & Petkov, 2014).

4.1.2 Canales digitales

Es evidente que la digitalización avanza a gran velocidad; y las empresas que avanzan online han conseguido aumentar sus clientes mejorando y posicionando su imagen de marca.

En la actualidad los canales digitales entregan servicios, comunican y venden por medio de dispositivos como computadores, tablets o teléfonos celulares brindando múltiples beneficios en la actualidad, y son enfocados para ser accesibles a los clientes desde cualquier punto de la geografía y a cualquier hora (Grau H, 2017).

En la actualidad en la gestión de contacto se identifican dos canales: los canales propios compuestos por página web empresarial o corporativa, atención en tienda en linea propia B2B o B2C o la app de manejo comercial, y del otro lado está el canal externo compuesto por los ahora llamados tiendas de terceros o por su nombre en inglés marketplace. Estos canales permiten: actividad y accesibilidad 24/7 donde los clientes podrán realizar cualquier solicitud.

Un beneficio clave de la digitalización en las organizaciones modernas, y a su vez la correspondiente automatización de procesos, es que se incrementan los niveles de calidad en el servicio y sin necesidad de depender de una persona al frente de ello, con lo anterior se reduce la inversión de trabajo y recurso al interior de las empresas (Grau H, 2017).

Para las organizaciones que adoptan este tipo de modelos, adicional a la utilidad que encuentren en ellos se debe tener claridad de no confundir los canales indicados con los usados tradicionalmente, ejemplo el teléfono, una comunicación formal, un documento o catálogo impreso o en formato pdf, un folleto. Tampoco con otros que en la teoría tienen componente digital, pero que tienen similitud a los tradicionales porque requieren de un recurso humano que los gestione diariamente, se habla de email, Whatsapp, Telegram, y otras redes sociales.

A continuación, y con la finalidad de acercar al lector de la presente investigación se resaltan los siguientes términos.

- ASR: Reconocimiento de voz automático.
- AVI: Agente virtual inteligente
- IVR: Respuesta de voz interactiva.
- CHATBOT: Aplicación asistente que usa una comunicación con los usuarios a través de mensajes de texto.

4.1.3 Aprendizaje automático

El aprendizaje automático es una ramificación de la Inteligencia Artificial, que permite a un sistema el aprendizaje a partir de los datos, contrario del aprendizaje por medio de la programación explícita. Estos sistemas se enfocan en aprender o mejorar su rendimiento de acuerdo a los datos que lo alimentan, con

el propósito de mejorar la experiencia de los usuarios o incluso de los sistemas para que sea más eficiente, seguro y fluido (Oracle, 2022).

Los sistemas de información que hacen uso de aprendizaje automático, se ajustan y mejoran a sí mismos de manera continua a través de los datos que reciben, lo que convierten en experiencia generando cada vez resultados más precisos (Google Cloud, 202). Aprendizaje automático visto como un modelo, corresponde a la salida de información generada al momento de entrenar los algoritmos con datos. Posterior al entrenamiento, al disponer de un modelo con una entrada, se generará una salida. Como ejemplo, un algoritmo predictivo tendrá como salida un modelo predictivo. Posteriormente, al momento de contar con el modelo predictivo con datos, su retorno será un pronóstico basado en los datos del modelo entrenado.

A continuación, se describen algunos algoritmos de aprendizaje automático.

4.1.3.1 K vecinos más próximos

Consiste en un algoritmo que predice un punto o ubicación de los datos sobre la base de un conjunto de datos disponibles. El caracter o letra K indica el total de vecinos más próximos. Cuando identifica una cantidad numerosa de vecinos K y que los mismos hacen parte de una nueva clase de un grupo específico o particular, entonces los puntos pronosticados recientemente de igual manera también son ingresados en la clase indicada. La distancia entre cada punto de datos en la clase y el nuevo punto de datos predicho puede ser calculado utilizando la fórmula de Manhattan, distancia de Hamming y distancia euclidiana (Cunningham & Delany, 2014).

4.1.3.2 Máquinas de soporte vectorial

Una Máquina de Soporte Vectorial conocida por sus siglas SVM, es un algoritmo de clasificación que aprende de la superficie decisión de dos clases distintas de puntos de entrada y se basa en el concepto de hiperplano. Cumple la función de clasificador para una sola clase, los vectores de soporte entregan una descripción la cual está en capacidad de generar una frontera de decisión al contorno del dominio de los datos sometidos al aprendizaje, muy práctica cuando los datos no son linealmente separables, porque trata que ninguno de los datos de conocimiento se encuentre por fuera de dicha frontera. La separación de datos en las diferentes clases y la formación de grupos es realizada por la función de frontera la cual lo realiza al momento que es traída al espacio de entrada (Garcia Z, 2017).

4.1.3.3 Árboles de decisión

Un algoritmo basado en árboles con una estructura jerárquica, que proporciona la función de clasificación, para predecir una categoría en base a las observaciones o matriz de características; y de regresión que buscan predecir un valor cuantitativo. Mediante la aplicación de este procedimiento se consigue reducir las variaciones de un método de aprendizaje automático. De acuerdo a lo anterior, un método natural para reducción de la varianza y mejorar la precisión de un método de aprendizaje, es tomar en repetidas veces las muestras del conjunto de datos de entrenamiento, se elabora un modelo predictivo de forma separada en cada conjunto y las predicciones resultantes son promediadas; obteniendo como resultado un único modelo de aprendizaje de baja varianza (James, et al., 2013).

4.1.3.4 Bosques aleatorios

Este algoritmo del aprendizaje automático supervisado se ha convertido en uno de los más populares y usados con mayor frecuencia gracias a su nivel de precisión, flexibilidad y simplicidad. Es altamente adaptable por su naturaleza no lineal y es usado para ejecuciones de clasificación y regresión, aunque su mayor limitación es que llega a presentar problemas de sobreajuste. Recibe su nombre por su naturaleza de crecimiento de árboles de decisión y la fusión de estos logran entregar predicciones con un mayor nivel de precisión, mientras que un solo árbol de manera individual entregará resultados reducidos, y el bosque asegura mayor precisión y cantidad de grupos y decisiones, adicional y como beneficio agrega aleatoriedad al modelo entregando mejores características entre subconjuntos, todos beneficios que favorecen a los modelos elaborados por los científicos de datos (Tibco, 2022).

4.1.4 Herramientas y lenguaje de programación

4.1.4.1 Python

Es un lenguaje de programación con características de eficiencia, fácil integración, aprendizaje y manejo de alto nivel, orientado a objetos, con una semántica dinámica integrada, enfocado primordialmente en desarrollos web, aprendizaje automático, ciencia de datos y en diversas aplicaciones informáticas (Bahit,2012).

Las herramientas tecnológicas a utilizar en el desarrollo del presente trabajo serán ejecutadas con Python, que por medio de su lenguaje de alto nivel permite realizar los desarrollos necesarios conjuntamente alimentado por el Dataset en formato MS Excel entregará como salida los resultados de los algoritmos.

4.1.4.2 Jupyter

La aplicación Jupyter Notebook trabaja bajo una interfaz web, cliente servidor, de código abierto dispuesta para trabajar audio, video, texto, imágenes y la ejecución de código en variados lenguajes. Ejecutada por comunicación con un núcleo de cálculo. Su uso enfocado a la creación y entrenamiento de modelos de aprendizaje automático, modelización estadística, visualización de datos y depuración de datos otorga una gran versatilidad a los especialistas, quienes tendrán una ventaja al utilizar esta herramienta de código abierto y, gracias a los variados intérpretes de otros lenguajes integrados, será viable la creación de código de forma sencilla en otros tipos de lenguaje (Granado, et al., 2018).

4.2 Estado del Arte

Esta sección del documento hace mención a las investigaciones previas encontradas y que se relacionan o guardan similitud en cuanto a la implementación de inteligencia artificial para la gestión de carteras de cobranza. Debido a que es un tema que cada vez toma mayor relevancia por las múltiples ventajas, así como asertividad en la toma de decisiones, en la forma de interactuar con los usuarios, en el recaudo enfocando esfuerzos y recursos de forma apropiada, se realizan hallazgos con diferentes enfoques.

De acuerdo al proyecto “Software gestor de cobranza usando inteligencia artificial”, se identifica una metodología basada en la creación de un software teniendo en cuenta los módulos de parametrización para la segmentación, priorización, gestión y asignación de la cartera, así como otros frentes de cobranza

como lo son la gestión de cartas, módulo de reportes, módulo de gestión externa y la tercerización de la cartera. Todo lo anterior planeando realizar el análisis de pasos y el estudio para desarrollar un software que brinde una solución inteligente de cobranzas para entidades de crédito y consumo aplicando inteligencia artificial (Gonzalez, et al., 2018).

Con la implementación de un modelo para gestión de cobranza enfocado a compañías con una amplia capacidad operativa en el manejo de tecnología, que administran grandes volúmenes de cartera a través de múltiples canales. Se busca optimizar la recuperación de cartera mediante la aplicación de algoritmos que permiten hacer segmentaciones, identificación o predicción. Además, define automáticamente los incentivos y remuneración para los diferentes canales con base en criterios y cumplimiento de gestión y volúmenes de recuperación de cartera. El sistema inteligente trabaja bajo un modelo de licenciamiento tradicional que brinda a las compañías autonomía en el manejo de datos, infraestructura y la libertad necesaria y capacidad de adaptación a las necesidades particulares (Gonzalez, et al., 2018).

En el trabajo “Analítica de datos aplicada a la cobranza de cartera” los autores llevaron sus esfuerzos en el almacenamiento de la información de forma centralizada a través de un Data Warehouse y un dashboard con información actualizada, en cuanto a los modelos de aprendizaje realizaron su enfoque en la probabilidad que tiene un deudor de pagar su obligación para enfocar la gestión en esos mismos (Montoya J, 2019).

Referente a la etapa de crédito, el punto en la implantación de Inteligencia Artificial, de programas informáticos que utilizan este modelo y que produce comunicaciones con los clientes, actualmente se concentra en la digitalización en procesos internos referidos a la solicitud de crédito, la recolección y análisis de datos, y las respuestas sobre requerimientos, condiciones acordadas, saldos a la

fecha, canales y formas de pago. La fase de recuperación de cartera vencida aún se encuentra separada de los procesos de transformación, con mínimas interacciones en la sección de cobro administrativo y etapas de mora inicial, con la aplicación de herramientas como, mensajes de voz y mensajes de texto con recordatorios de pago. En la actualidad la aplicación de los modelos de inteligencia artificial como de aprendizaje automático son usados de manera básica para los procesos de cobranza sin embargo todo apunta a que en el mediano y corto plazo habrá crecimiento en su utilización. Uno de los capítulos especializados de la inteligencia artificial contempla el contacto con los deudores en las organizaciones orientadas a la cobranza de obligaciones. Los sistemas utilizados tienen accesibilidad a los grandes datasets y compilan información de cada una de las interacciones; también, hacen contacto por diferentes canales como los chats, las páginas web, llamadas o correo electrónico (Montoya Yepes, 2019).

Con similitud se encuentra también el trabajo “Pronóstico del cumplimiento de pago de los clientes usando aprendizaje automático” en el que se referencian variados modelos de clasificación, así como otras metodologías comúnmente utilizadas. Las metodologías utilizadas, son superiores en sus métricas de precisión al análisis de regresión logística, modelo usado actualmente por la empresa. Como resultado del análisis se concluye que se cuenta con otros modelos de regresión logística que entregan una mayor confiabilidad ajuste y precisión sobre los datos resultantes y resultados finales (Campos Z, 2020).

Al área responsable de analítica de la empresa se les presenta y hace entrega del documento guía de uso de jupyter por medio del cual implementaron una buena práctica para hacer seguimiento y control de forma trimestral. En dicho documento por medio del backtesting analizaron qué comportamiento tiene el modelo, haciendo la medición de los umbrales y del rendimiento establecidos en su inicio, evidenciando alertas o si hay estabilidad en ellos, o detectando variación en las

métricas. Para el presente trabajo fueron establecidos los parámetros de precisión, f1 score, sensibilidad y exactitud.

El modelo de clasificación fue desarrollado con el uso de bosques aleatorios; este algoritmo fue seleccionado por sus características adecuadas en cuanto a la precisión y exactitud de sus métricas, sumado a lo anterior son tenidos en cuenta la eficiencia en los tiempos de ejecución. El modelo contenido en un Jupyter está programado con una periodicidad semanal con lo que se conseguirá segmentar de la base de cartera para el desarrollo de las estrategias de la alta gerencia. Este modelo otorgará a la organización un mejoramiento en la eficiencia con el uso de sus recursos al momento de la cobranza de cartera. Aunque muchas empresas dedicadas a la gestión de cobro esperan a que sus clientes se encuentren en periodo de mora para tomar acciones, con el uso de los prototipos propuestos, se ha evidenciado que es de mayor beneficio conocer anticipadamente aquellos clientes que son propensos a entrar en mora con lo que se enfocarán los recursos en este segmento de cliente (Campos Z, 2020).

Las tres investigaciones mencionadas previamente presentan un panorama sobre las técnicas, herramientas y algoritmos objeto de estudio aplicadas, que están estrechamente relacionados a la gestión de cobranza, teniendo como factor común la utilización de inteligencia artificial para tomar provecho del alto volumen de datos para realizar la máxima extracción de información proveniente de ellos.

Si bien el enfoque guarda similitud, por la naturaleza de la actividad, la gestión de cobranza y sus procesos son variables y cada empresa los aborda de distintas formas dependiendo su capacidad, número de clientes, tipo de carteras y canales de comunicación. Haber analizado temas como segmentación, priorización, gestión y asignación de la cartera, la aplicación de algoritmos, y los resultados en la segmentación, identificación o predicción, pasando por modelos en jupyter, nos entregaron pautas y un feedback para el desarrollo de nuestras labores para

predecir la efectividad en carteras en etapa de castigo en la organización, haciendo uso de los algoritmos de clasificación como k-vecinos más próximos, máquinas de soporte vectorial, bosques aleatorios y árboles de decisión.

4.3 Impacto

Tras la aplicación de los algoritmos de aprendizaje automático para la predicción de clientes más efectivos en carteras de cobranza castigada se espera un impacto positivo en la efectividad de gestión de la empresa InteliBPO S.A.S igual o superior al 37% en el segmento de cartera castigada, así como un aumento del 20% en los indicadores de recaudo trazados por la alta Gerencia.

Se espera que con la aplicación de los modelos de aprendizaje automático se agilicen y mejoren los procesos de toma de decisiones por su capacidad de clasificación e identificación, labores que antes de su implementación se llevaban a cabo manualmente demandando tiempo y recursos, aprovechables en otros procesos de la compañía.

4.4 Componente de Innovación

En un entorno de economía globalizada, la aplicación de modelos de predicción constituye una herramienta primordial para entregar soluciones a los frecuentes problemas que las empresas gestoras de cobranza enfrentan y en la solución de factores que impiden un recaudo efectivo y eficiente, la industria del software ofrece una mayor velocidad de captura y procesamiento de información por ende mayor valor para los procesos y empresas en general. Los recientes cambios mundiales y regionales inclinan a las compañías modernas a reconsiderar sus procesos en razón a la notoria y dinámica complejidad de sus entornos y a

causa de ello la disminución en su competitividad y disminución en la efectividad de sus modelos de negocios.

Una solución presentada en la teoría y en los datos de la realidad, señala a la innovación como un factor importante de perdurabilidad y continuidad empresarial, sin embargo, en algunos casos la innovación es un sinónimo de dificultad para las empresas. La capacidad de innovar se debe constituir en un recurso de las empresas que al igual que sus operaciones financieras, comerciales y de producción, debe ser implementado y gestionado de la misma manera e importancia y contemplada dentro de la misma estrategia de negocio. Así, dar el paso hacia una economía de alto valor y con una temática orientada a la innovación, requiere que el mundo de las tecnologías de información y comunicaciones sea competitivo e integrado con los sectores económicos en pro del crecimiento económico y fortalecimiento de las compañías.

Se concluye entonces que un modelo de análisis predictivo, que contiene información sobre los hábitos y comportamiento de clientes, utiliza la información y considera otras variables externas, aporta herramientas a la organización para la acertada toma de decisiones y entrega una visión adecuada para cumplir con estrategias de cobro de acuerdo a la segmentación realizada sobre los clientes. Un modelo de predicción, ayuda a identificación de clientes, e incrementa la rentabilidad de la cartera de la empresa.

5. Metodología

Con el propósito de alcanzar los objetivos propuestos en el desarrollo de este proyecto de aplicación se utilizaron los fundamentos establecidos en la metodología CRISP-DM, cuyas siglas significan Cross-Industry Standard Process for Data Mining (IBM, 2021). Sin embargo, se estructuró en las siguientes fases, tal como se muestra en la figura 1.

Figura 1. *Fases de la metodología del proyecto*

Fuente: Elaboración propia tomado de Draw.io (2022)

A continuación, se describen de forma detallada las diferentes fases que conforman la metodología utilizada.

5.1 Entendimiento del negocio

Esta fase se enfocará en tener claridad sobre el objetivo y los requisitos del negocio, para entender el contexto y así mismo enfocarlo en el problema de minería de datos que se quiere solucionar.

5.2 Extracción y entendimiento de los datos

Se comienza con la recolección de datos inicial para realizar un análisis exploratorio que permita familiarizarse con los datos, entenderlos con base al negocio o evidenciar problemas con la calidad de los datos que brinde una visión más general del potencial y viabilidad que tienen con respecto al problema a solucionar.

5.3 Preparación de datos

Esta fase de preparación de datos se convertirá en la definición del dataset que se implementará para construir el modelo de aprendizaje automático, lo cual incluirá las tareas de selección de atributos, registros, limpieza y transformación de los datos a como las acepta las herramientas de modelado.

5.4 Construcción de modelo

Se realiza una selección y aplicación de diferentes algoritmos o técnicas de aprendizaje automático enfocados al problema que se requiere dar solución. Dependiendo de la técnica seleccionada se deben configurar los parámetros más óptimos y si es necesario volver a la fase anterior para ajustar de nuevo los datos para lograr mejores resultados.

5.5 Evaluación e interpretación

Con la construcción de uno o varios modelos, se debe evaluar si los resultados obtenidos satisfacen el objetivo de negocio o si, por el contrario, se requiere revisar de nuevo alguna variable o consideración del negocio que puede ser importante.

6. Desarrollo de la propuesta

A continuación, se realiza una descripción de cada uno de los procesos y actividades, que involucra cada etapa de la metodología planteada en el apartado anterior.

6.1 Entendimiento del negocio

La empresa InteliBPO S.A.S a través de sus múltiples contratos con diferentes empresas recibe asignaciones de carteras de cobranza castigada para su respectiva gestión. Sin embargo, uno de los mayores desafíos a los que se enfrenta la compañía está en determinar qué estrategias de gestión debe realizar para que sea más efectiva y en qué segmento de clientes enfocar los esfuerzos tanto operacionales como de recursos económicos, teniendo en cuenta, que la mayor característica de estas carteras es que los usuarios son morosos y por ende el ritmo de contacto, generación de promesas y recaudo es más demorado e implica mayores esfuerzos.

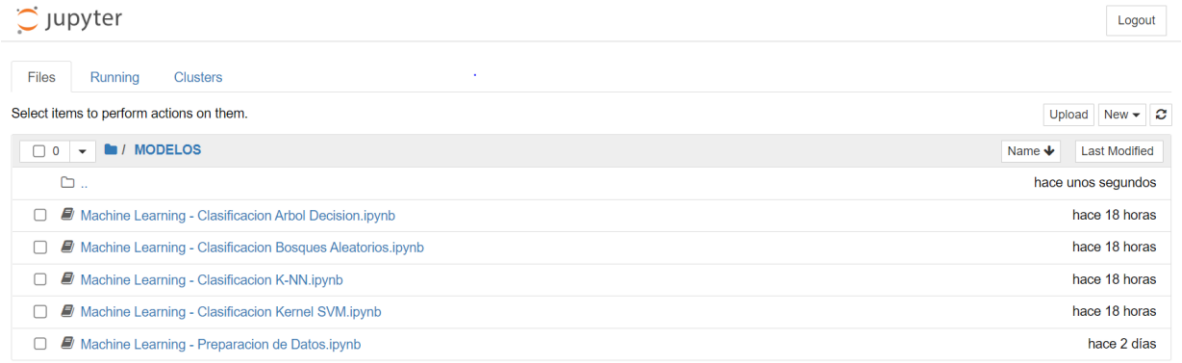
Por lo tanto, un primer paso para avanzar en la automatización de estrategias de gestión eficaces, está en lograr predecir que futuros clientes serán efectivos o no, para así mismo comenzar a tomar decisiones más asertivas para el negocio.

6.2 Extracción y entendimiento de los datos

Para realizar el análisis inicial de los datos sobre los cuales se trabajó, se hizo uso de Jupyter Notebook que se encuentra configurado en un servidor de desarrollo en la nube, en el cual, se creó un proyecto que contiene los notebooks de entendimiento, preprocesamiento de datos y los modelos generados, los cuales se

desarrollaron en el lenguaje de programación python, como se evidencia en la figura 2.

***Figura 2**. Estructura del proyecto en jupyter*



Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

Para iniciar con el proceso de extracción de los datos se creó el archivo Machine Learning - Preparacion Datos.ipynb, donde se realiza la importación de las librerías de pandas, numpy y matplotlib, esta última siendo la que permite que los gráficos sean renderizados directamente en el Notebook.

En esta etapa de la metodología se realizó la extracción de los datos en un dataframe de pandas con la lectura de un archivo .csv cuya fuente se encuentra en una ruta de la sftp, a la cual tiene acceso el servidor, y en la cual se guardan reportes históricos de gestión. La información del reporte seleccionado para analizar contiene 67 columnas y 81.594 registros de gestiones realizadas en un periodo de tiempo de tres meses a carteras de cobranzas en etapa de castigo.

Para el análisis exploratorio de los datos se utilizó pandas profiling, que es un módulo open source de Python con el que se generó un informe en formato web (html) a partir del resultado generado. Con la información obtenida se evidencio

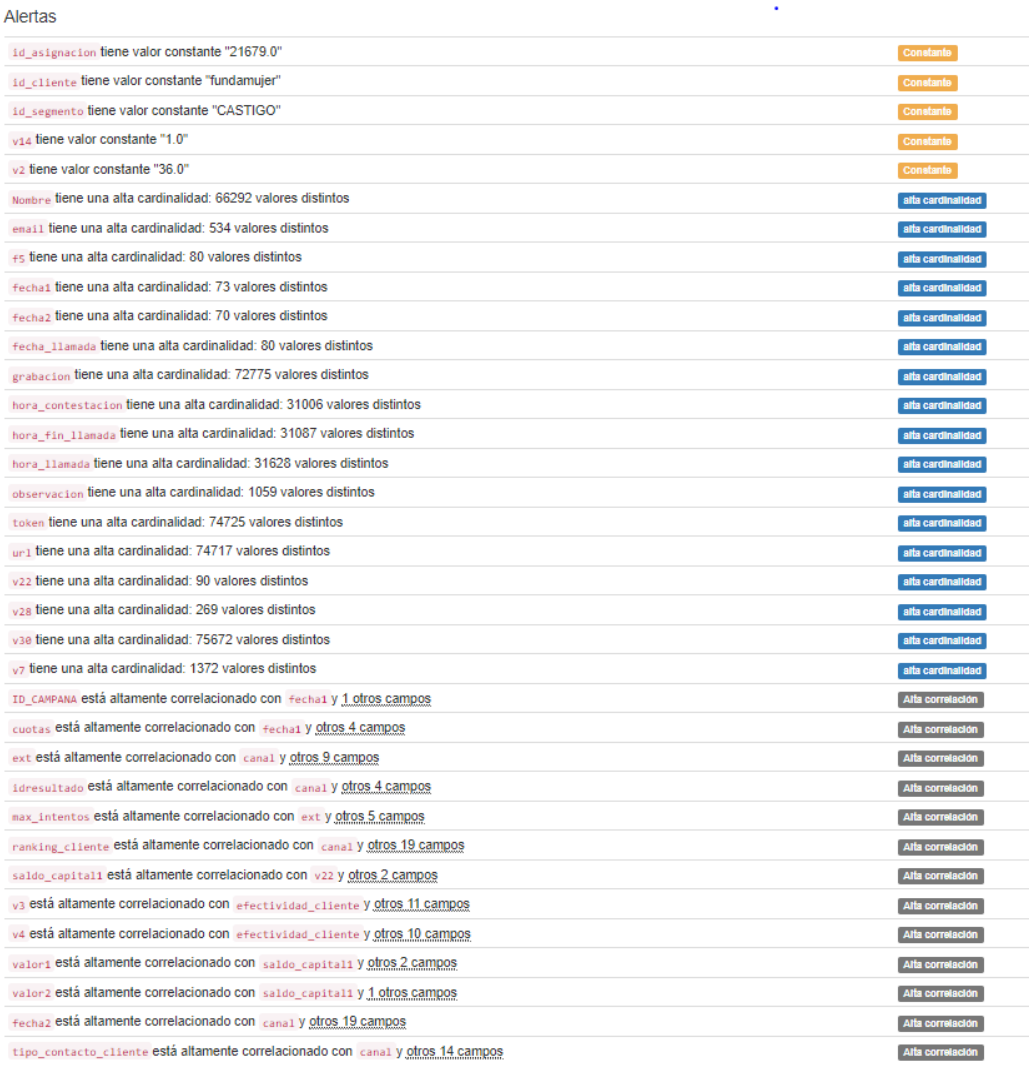
la estructura del dataset que cuenta con 47 variables categóricas, 12 numéricas y 12 más que no fue posible definir; la información no contiene registros duplicados y teniendo en cuenta todos los campos del dataset hay un 23,4% de campos vacíos (Ver figura 3).

***Figura 3**. Estadísticas generales del conjunto de datos*

Fuente: Elaborado con Pandas Profiling Report
Report generated by YData

Como característica importante del reporte generado están las alertas, las cuales se analizaron para determinar los datos que aportan valor al objetivo del proyecto, ya que se evidencia las columnas del dataset que presentan valores constantes mostrando el valor del campo como por ejemplo id_segmento cuyo valor constante es 'CASTIGO'; los campos que presentan alta cardinalidad con su respectiva información de la cantidad de valores únicos; los campos con alta correlación en los que se indican los campos con los que refleja dependencia; los campos que presentan valores faltantes indicando el total de registros y porcentaje; así como los campos que presentan un tipo de dato no admitido y se debe verificar si necesitan limpieza o mayor análisis (Ver figura 4).

Figura 4. *Características y alertas generales del conjunto de datos*



Alertas

id_asignacion tiene valor constante "21679.0"	Constante
id_cliente tiene valor constante "fundamujer"	Constante
id_segmento tiene valor constante "CASTIGO"	Constante
v14 tiene valor constante "1.0"	Constante
v2 tiene valor constante "36.0"	Constante
Nombre tiene una alta cardinalidad: 66292 valores distintos	alta cardinalidad
email tiene una alta cardinalidad: 534 valores distintos	alta cardinalidad
f5 tiene una alta cardinalidad: 80 valores distintos	alta cardinalidad
fecha1 tiene una alta cardinalidad: 73 valores distintos	alta cardinalidad
fecha2 tiene una alta cardinalidad: 70 valores distintos	alta cardinalidad
fecha_llamada tiene una alta cardinalidad: 80 valores distintos	alta cardinalidad
grabacion tiene una alta cardinalidad: 72775 valores distintos	alta cardinalidad
hora_contestacion tiene una alta cardinalidad: 31006 valores distintos	alta cardinalidad
hora_fin_llamada tiene una alta cardinalidad: 31087 valores distintos	alta cardinalidad
hora_llamada tiene una alta cardinalidad: 31628 valores distintos	alta cardinalidad
observacion tiene una alta cardinalidad: 1059 valores distintos	alta cardinalidad
token tiene una alta cardinalidad: 74725 valores distintos	alta cardinalidad
url tiene una alta cardinalidad: 74717 valores distintos	alta cardinalidad
v22 tiene una alta cardinalidad: 90 valores distintos	alta cardinalidad
v28 tiene una alta cardinalidad: 269 valores distintos	alta cardinalidad
v30 tiene una alta cardinalidad: 75672 valores distintos	alta cardinalidad
v7 tiene una alta cardinalidad: 1372 valores distintos	alta cardinalidad
ID_CAMPANA está altamente correlacionado con fecha1 y 1 otros campos	Alta correlación
cuotas está altamente correlacionado con fecha1 y otros 4 campos	Alta correlación
ext está altamente correlacionado con canal y otros 9 campos	Alta correlación
idresultado está altamente correlacionado con canal y otros 4 campos	Alta correlación
max_intentos está altamente correlacionado con ext y otros 5 campos	Alta correlación
ranking_cliente está altamente correlacionado con canal y otros 19 campos	Alta correlación
saldo_capital1 está altamente correlacionado con v22 y otros 2 campos	Alta correlación
v3 está altamente correlacionado con efectividad_cliente y otros 11 campos	Alta correlación
v4 está altamente correlacionado con efectividad_cliente y otros 10 campos	Alta correlación
valor1 está altamente correlacionado con saldo_capital1 y otros 2 campos	Alta correlación
valor2 está altamente correlacionado con saldo_capital1 y 1 otros campos	Alta correlación
fecha2 está altamente correlacionado con canal y otros 19 campos	Alta correlación
tipo_contacto_cliente está altamente correlacionado con canal y otros 14 campos	Alta correlación

Fuente: Elaborado con Pandas Profiling Report
Report generated by YData

Los datos presentan información de la mejor gestión realizada a los clientes de carteras de cobranza teniendo en cuenta todos los canales de gestión que maneja la empresa que son AVI, sms, email, chatbot, TuAcuerdo, IVR, chat y asesor. Por lo tanto, para entender el negocio y la información del informe se organizó el diccionario de datos descrito en la tabla 1.

Tabla 1. Diccionario de datos del informe de gestión

Campo	Ejemplo	Descripción	Tipo de datos
ID_CAMPANA	1663167028	Identificador de la campaña de gestión	Numérico
Identificación	1065663241	Número de identificación del usuario	Numérico
Nombre	YAZMIN ORTIZ	Nombre y apellido del usuario	Texto
canal	ASESOR	Canal de gestión	Texto
cuotas	1	Numero de cuotas a ofrecer	Numérico
dias_mora	1383	Días de mora que tiene la obligación	Numérico
efectividad_cliente	EFECTIVO	Indica si la gestión fue efectiva o no para el cliente	Texto
efectividad_intelibpo	EFECTIVO	Indica si la gestión fue efectiva o no para la empresa	Texto
email	yaz.154@hotmail.com	Correo electrónico del usuario	Alfanumérico
empresa	Banco XXXX	Empresa que asigno el registro	Texto
ext	9020	Numero de extensión en caso de transferencia	Numérico
f5	2022-09-22 13:20:00	Fecha y hora en que se insertó el registro	Fecha y hora
fecha1	2022-09-14	Primera opción de fecha para pago	Fecha

fecha2	2022-09-14	Segunda opción de fecha para pago	Fecha
fecha_llamada	2022-09-22	Fecha en que se generó la gestión	Fecha
gestionados_cliente	GESTIONADO	Indica si el usuario se considera gestionado para el cliente	Texto
gestionados_intelibpo	GESTIONADO	Indica si el usuario se considera gestionado para la empresa	Texto
grabacion		Para las gestiones que involucran llamada el nombre de la grabación	Alfanumérico
homologacion_cliente	YA PAGO	Descripcion del idresultado para el cliente	Texto
homologacion_intelibpo	AL DIA	Descripcion del idresultado para la empresa	Texto
hora_contestacion	11:53:11	Hora en que se tiene contacto con el usuario	Hora
hora_fin_llamada	11:53:11	Hora en que finaliza el contacto con el usuario	Hora
hora_llamada	11:53:11	Hora en que se inició la gestión	Hora
id_asignacion	21679	Identificador de la asignación a la que pertenece el usuario	Numerico
id_cliente	Sucursal XXX	Cliente al que pertenece el usuario	Texto
id_segmento	CASTIGO	Segmento al que pertenece el usuario	Texto
idresultado	80101	Identificador del resultado de gestión	Numérico

		según el canal	
intento	1	Numero de intento en que se realizado la gestión de la campaña	Numérico
max_intentos	13	Cantidad de gestiones que lleva en el tiempo de la asignación que ha generado costos	Numérico
moment	RECORDACION DE PAGO	Clasificador del momento de gestión en que se encuentra el registro	Alfanumérico
observacion	TITULAR CONFIRMA QUE YA CUMPLIO CON EL ACUERDO DE PAGO, ENVIA SOPORTE DE PAGO POR MEDIO DE WHATSAPP	Comentarios o información adicional	Alfanumérico
previous_moment	XAFINAR NOMBRE	Clasificador del momento previo de gestión en que se encuentra el registro	Alfanumérico
producto1	512181013857	Numero de la obligación del usuario	Alfanumérico
ranking_cliente	1	Ranking del idresultado de las gestiones (entre menor el número más efectiva la gestión) para el cliente	Numérico
resultado		Descripción del idresultado	Texto
saldo_capital1	208886	Saldo a capital del usuario	Numérico
telefono	573044008743	Número de teléfono de contacto del usuario	Numérico

tipo_contacto_cliente	CONTACTO DIRECTO	Tipo de contacto para el cliente	Texto
tipo_contacto_intelibpo	CONTACTO DIRECTO	Tipo de contacto para la empresa	Texto
token	1312128690	Identificador único del usuario	Numérico
url	xxxxxxxx.com/xx/1312128690	Url de la pagina de autogestión de la empresa con el token	Alfanumérico
v1	0	Variable opcional	Alfanumérico
v2	36	Variable opcional	Alfanumérico
v3	5	Variable opcional	Alfanumérico
v4	12	Variable opcional	Alfanumérico
v5		Variable opcional	Alfanumérico
v6		Variable opcional	Alfanumérico
v7	JAIR ALEJANDRO RODRIGUEZ	Variable opcional	Alfanumérico
v8	senora	Variable opcional	Alfanumérico
v9		Variable opcional	Alfanumérico
v10		Variable opcional	Alfanumérico
v12		Variable opcional	Alfanumérico
v14		Variable opcional	Alfanumérico
v15	PACIFICO	Variable opcional	Alfanumérico
v16	BI_AZUCENA	Variable opcional de contexto	Alfanumérico
v17		Variable opcional	Alfanumérico
v22	Credito Empresarial	Variable opcional	Alfanumérico
v23		Variable opcional	Alfanumérico

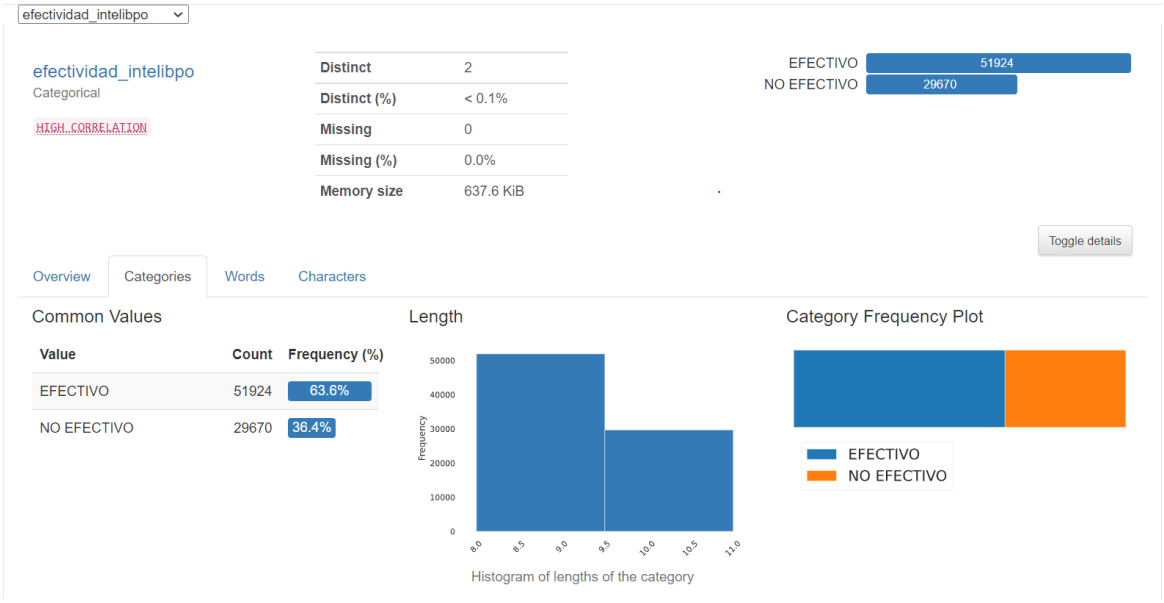
v24		Variable opcional	Alfanumérico
v25		Variable opcional	Alfanumérico
v26	330260	Variable opcional	Alfanumérico
v27	FEMENINO	Variable de genero	Alfanumérico
v28	Cali	Variable de región	Alfanumérico
v29		Variable opcional	Alfanumérico
v30	ANA TERESA HERRERA SUAREZ	Variable con nombre completo del usuario	Alfanumérico
valor1	208887	Valor ofrecido para el pago	Numérico
valor2	568753	Valor total de la deuda	Numérico

Fuente: Elaboración propia

6.3 Preparación de datos

Teniendo en cuenta el negocio y el análisis anterior de las variables del dataset, en esta etapa de la metodología, lo primero que se determinó fue que el campo efectividad_intelibpo es la variable objetivo o variable dependiente a contemplarse en los modelos de aprendizaje automático, la cual es categórica y tiene dos posibles valores: EFECTIVO con un total de 51.924 registros que corresponden al 63.6% del dataset y el NO EFECTIVO con 29.670 registros correspondientes al 36.4% (Ver figura 5).

***Figura 5**. Características y gráficas exploratorias de la variable efectividad del conjunto de datos*

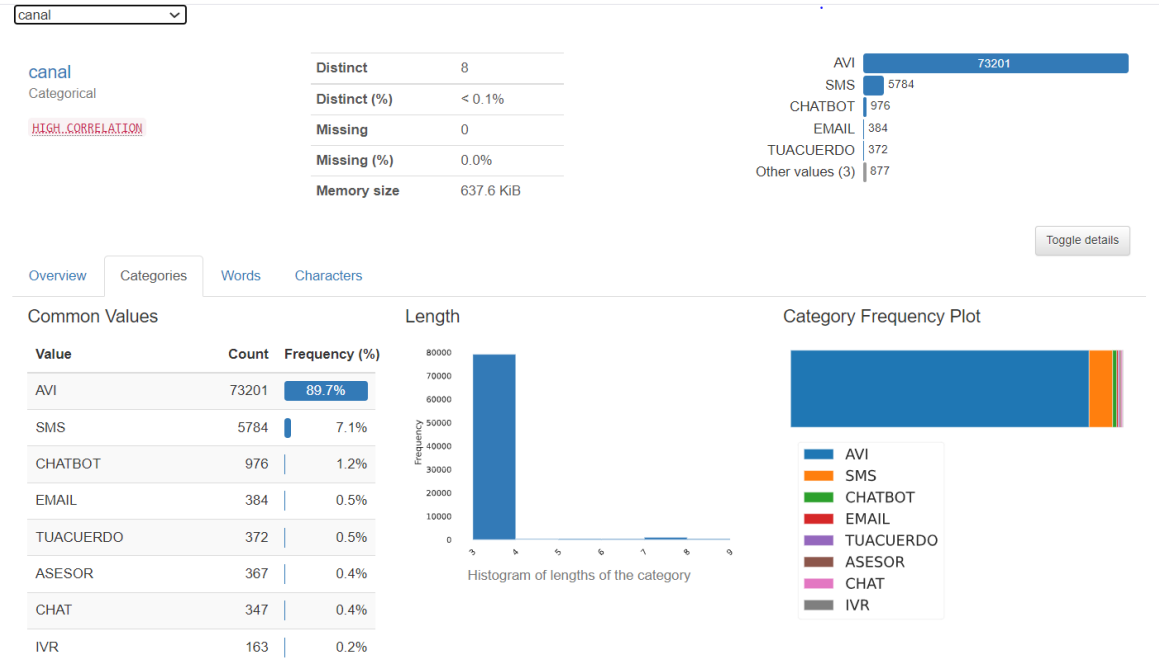


Fuente: Elaborado con Pandas Profiling Report
Report generated by YData

Con respecto a los atributos a utilizar en los modelos de aprendizaje automático se definió que podían aportar valor las variables:

1. **canal**: Indica el canal por el cual fue gestionado el usuario. Según el análisis exploratorio es una variable categórica, no presenta valores nulos, ni datos que presenten el mismo significado pero que varíe en la sintaxis. Se observa una inclinación considerable por la gestión con el canal AVI, lo que a nivel de negocio es consistente, debido a que, a través de este, se realizan gestiones masivas mientras que algunos de los otros canales son de autogestión (Ver figura 6).

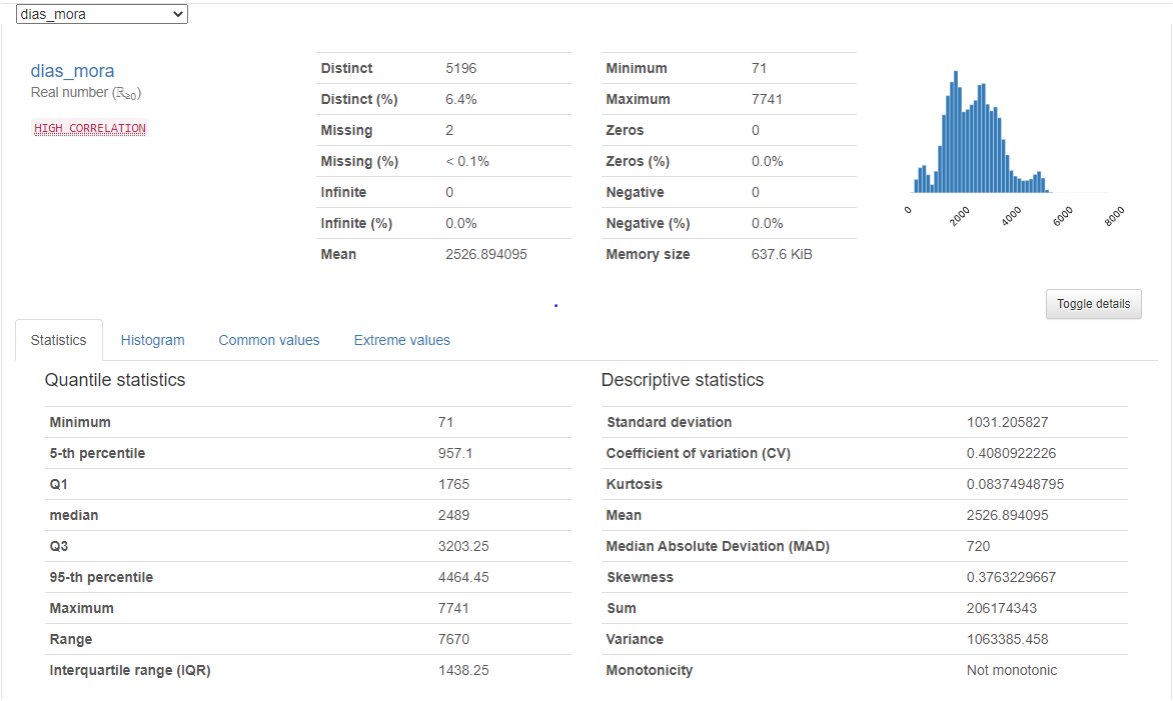
***Figura 6**. Características y gráficas exploratorias de la variable canal del conjunto de datos*



Fuente: Elaborado con Pandas Profiling Report
Report generated by YData

2. **dias_mora**: Representa la cantidad de días en mora que tiene el producto asociado al usuario y por el cual se encuentra en etapa de castigo. Su valor máximo es de 7.670, presenta una mediana de 2.489 y un percentil 95 de 4.464,45 que se observa junto a su histograma en la figura 7.

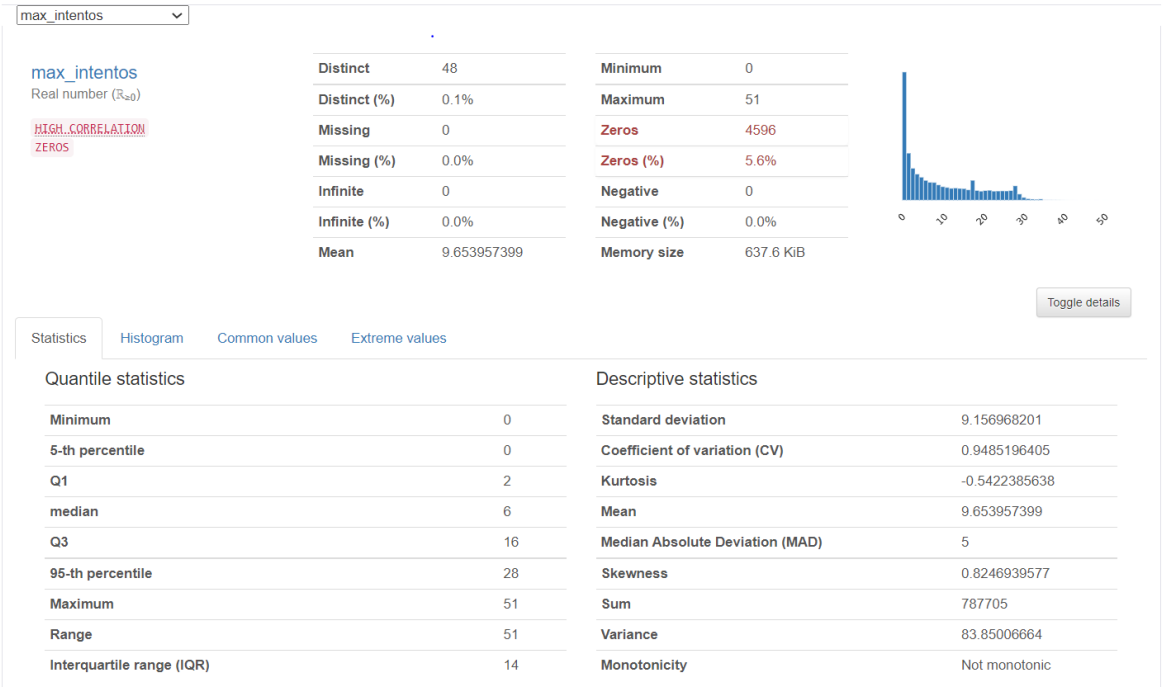
Figura 7. *Características y gráficas exploratorias de la variable días mora del conjunto de datos*



Fuente: Elaborado con Pandas Profiling Report
Report generated by YData

3. **max_intentos**: Corresponde a la cantidad de intentos en cuya gestión se ha generado algún costo, por lo tanto, al observar que hay una cantidad considerable de registros con el valor en cero (0), se valida que es un dato acorde al negocio y que tiene correlación con la efectividad, ya que, por ejemplo, en los casos en que no hay contacto con el cliente y que a su vez no genero un costo implica generalmente que la gestión no fue efectiva por lo que hay una correlación (Ver figura 8).

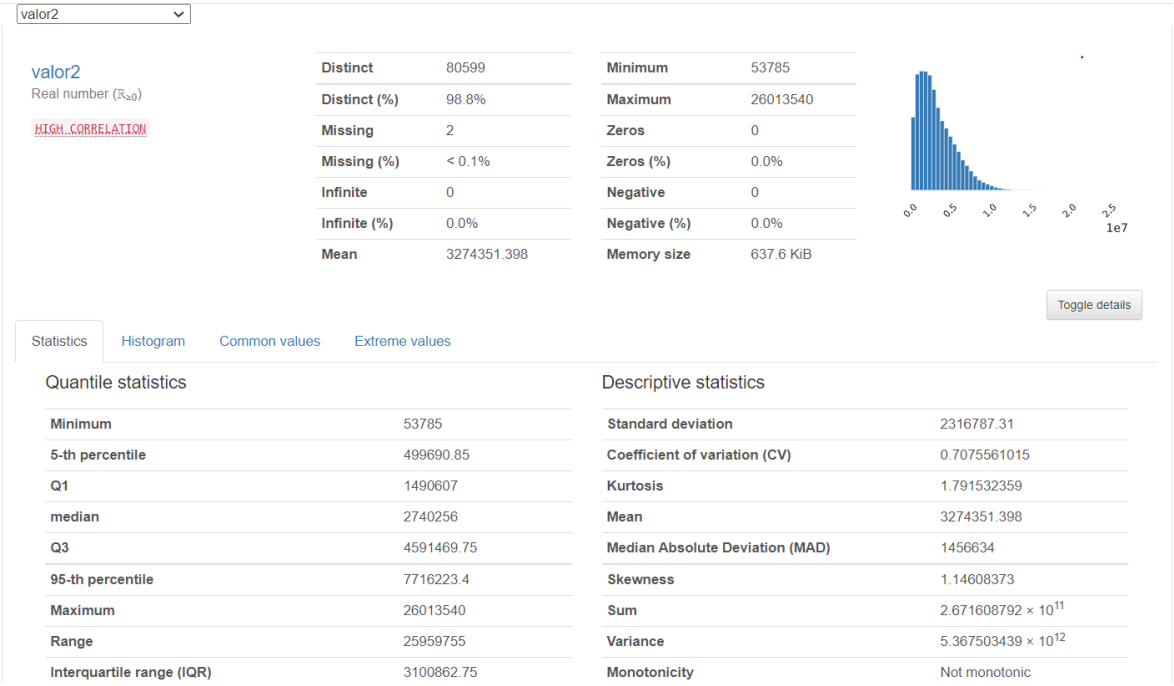
***Figura 8**. Características y gráficas exploratorias de la variable max_intentos del conjunto de datos*



Fuente: Elaborado con Pandas Profiling Report
Report generated by YData

4. **valor2**: Determina el valor total de la deuda que presenta el usuario y según el análisis exploratorio hay un valor mínimo de $53.785, un máximo de $26.013.540 y una mediana de $2.740.256 (Ver figura 9).

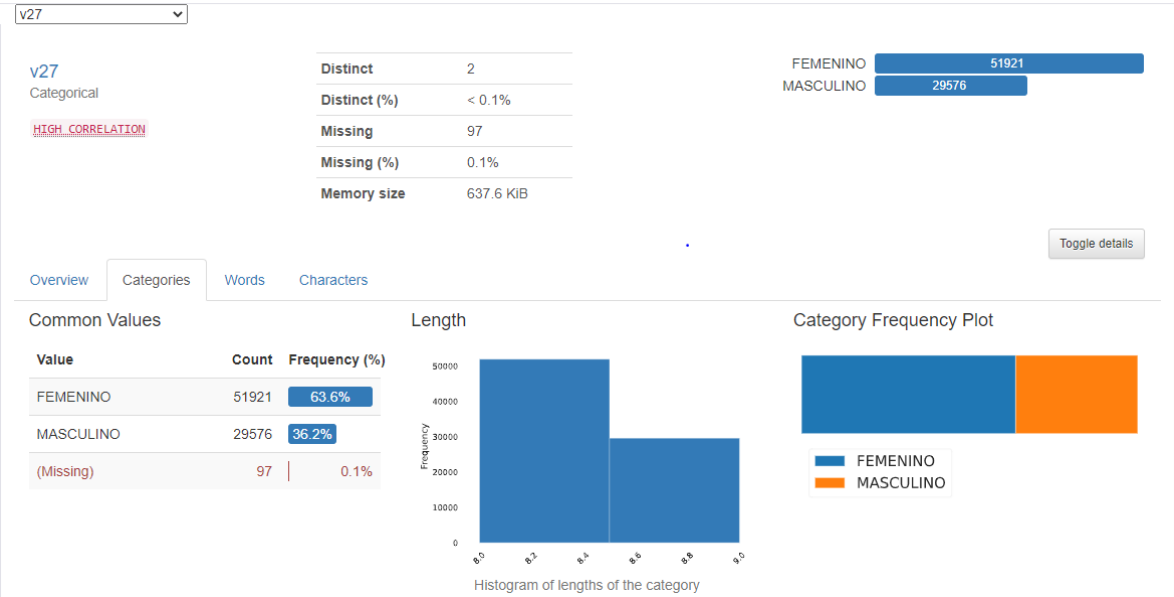
***Figura 9**. Características y gráficas exploratorias de la variable valor2 del conjunto de datos*



Fuente: Elaborado con Pandas Profiling Report
Report generated by YData

5. **v27**: Representa el género del usuario en donde el 63.6% corresponde al FEMENINO, el 36.2% al MASCULINO y el restante a registros que presentan valores nulos (Ver figura 10).

***Figura 10**. Características y gráficas exploratorias de la variable v27 (género) del conjunto de datos*



Fuente: Elaborado con Pandas Profiling Report
Report generated by YData

La preparación de los datos para el modelo se generó en el jupyter un dataframe con las variables definidas anteriormente y se realizó una limpieza de valores nulos de la variable v27 (género) ya que hay 97 registros que se evidencian sin datos correspondiente al 0,12%, lo que no representa una pérdida de datos significativa que influya negativamente en el resultado de los algoritmos.

Para continuar con el proceso de carga y preprocesado de datos se definió el valor de **X** con el conjunto de datos del dataframe anterior de las variables independientes escogidas para entrenar el modelo; y el valor de **y** con la variable dependiente (efectividad_intelibpo). Para poder entrenar los modelos, se preparó las variables categóricas que se van a utilizar, convirtiéndolas en datos numéricos comprensibles para los algoritmos con la función **LabelEncoder**, y adicional, para

los datos de *canal* y *v27* (género) que son de tipo categóricas nominal se utilizó la función **OneHotEncoder** para convertirlas en variables dummy.

Con los datos transformados y listos para utilizarse en los algoritmos, se procedió a dividir el dataset en el conjunto de entrenamiento con un 75% y el conjunto de testing con un 25% del total de los datos. Además, de agregar el argumento *stratify* con el valor de **y** para solventar el problema de datos desequilibrados, evitando que la división del mismo se incline en dejar con cierto valor alguno de los conjuntos de datos (Ver figura 11).

Figura 11. *Distribución del conjunto de datos de entrenamiento y testing*

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0, stratify=y)
```

Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

Por último, se realiza un escalado de variables importando la función **StandardScaler** sobre los datos tanto de entrenamiento como de testing (Ver figura 12).

Figura 12. *Escalado de variables de entrenamiento y test*

```
from sklearn.preprocessing import StandardScaler
sc_X = StandardScaler()
X_train = sc_X.fit_transform(X_train)
X_test = sc_X.transform(X_test)
```

Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

6.4 Construcción de modelo

El tipo de implementación de aprendizaje automático que se enfoca en el objetivo del proyecto que es el de predecir futuros clientes efectivos, es de tipo supervisado con un problema de clasificación donde la variable objetivo es categórica, como se describió en el punto anterior de la metodología. Se hizo uso de varios algoritmos de aprendizaje supervisado para poder evaluar a partir de los resultados obtenidos cuál cumplía mejor el objetivo propuesto.

6.4.1 Modelo K-NN

Se utilizó el algoritmo de K-NN (K vecinos más cercanos) importando la clase *KNeighborsClassifier* de Scikit-Learn y se invoca configurando los parámetros de *n_neighbors* con un valor de 9 para indicar que será la cantidad de vecinos más cercanos para clasificar; la *métrica* minkowki; y *p* con el valor 2 para la distancia euclidean (Ver figura 13).

Figura 13. *Entrenamiento con modelo de K-NN*

```
from sklearn.neighbors import KNeighborsClassifier
```

```
classifier = KNeighborsClassifier(n_neighbors = 9, metric = "minkowski", p = 2)
classifier.fit(X_train, y_train)
```

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=None, n_neighbors=9, p=2,
           weights='uniform')
```

Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

6.4.2 Modelo Kernel SVM

Para el algoritmo de máquina de soporte vectorial con kernel se importó la clase *SVC* de Scikit-Learn configurando como parámetros el kernel radial para generar una separación no lineal (Ver figura 14).

***Figura 14**. Entrenamiento con modelo de máquina de soporte vectorial*

```
from sklearn.svm import SVC
```

```
classifier = SVC(kernel = "rbf", random_state = 0)
classifier.fit(X_train, y_train)
```

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
  kernel='rbf', max_iter=-1, probability=False, random_state=0,
  shrinking=True, tol=0.001, verbose=False)
```

Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

6.4.3 Modelo Árbol de decisión

Para el algoritmo de árbol de decisión se importó la clase DecisionTreeClassifier de la librería tree de sklearn en cuyos parámetros se definió el criterio de entropía que mide la calidad de las divisiones para que los nodos finales del árbol sean lo más homogéneos posible (Ver figura 15).

***Figura 15**. Entrenamiento con modelo de árbol de decisiones*

```
from sklearn.tree import DecisionTreeClassifier
```

```
classifier = DecisionTreeClassifier(criterion = "entropy", random_state = 0)
classifier.fit(X_train, y_train)
```

```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=0,
            splitter='best')
```

Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

6.4.4 Bosques Aleatorios

Por último, se utilizó el algoritmo de bosques aleatorios importando RandomForestClassifier de la librería ensemble de sklearn, en el que se configuró el número de estimadores en 10, que significa que esta será la cantidad de árboles que formarán parte del bosque, y el criterio de división de entropía, siendo el mismo que en los árboles de decisiones (Ver figura 16).

***Figura 16**. Entrenamiento con modelo de bosques aleatorios*

```
from sklearn.ensemble import RandomForestClassifier

classifier = RandomForestClassifier(n_estimators = 15, criterion = "entropy", random_state = 0)
classifier.fit(X_train, y_train)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='entropy',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=15, n_jobs=None,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```

Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

Realizado el entrenamiento de cada uno de los algoritmos se calculó la predicción en el jupyter con respecto al modelo, con el conjunto de datos de testing (Ver figura 17).

***Figura 17**. Predicción del conjunto de test en los modelos de aprendizaje automático*

```
y_pred  = classifier.predict(X_test)

print(y_test)
print(y_pred)
```

Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

6.5 Interpretación de resultados

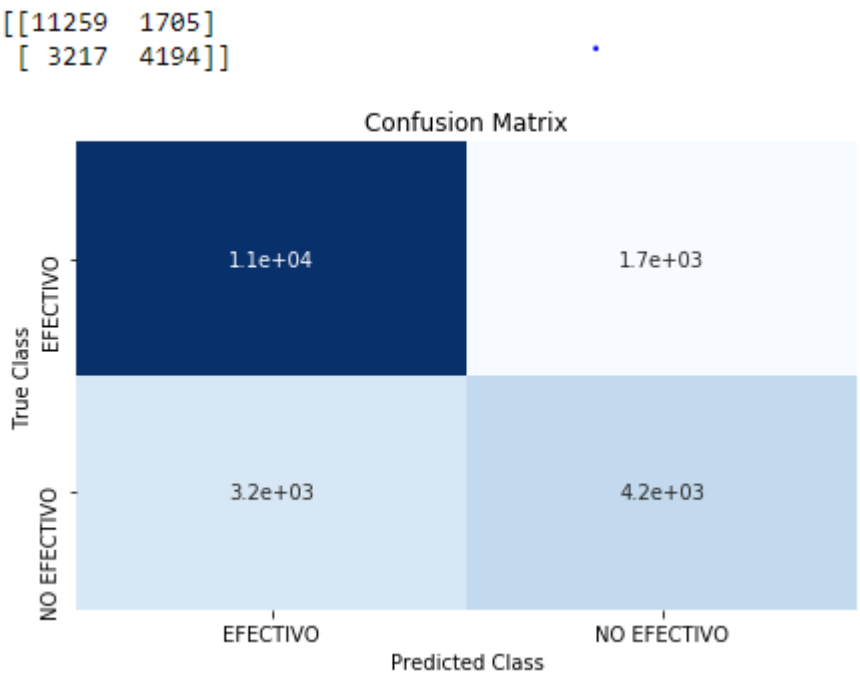
Como último paso de la metodología planteada se realizó una evaluación e interpretación de los modelos de aprendizaje realizados a partir de la matriz de confusión y las principales métricas de clasificación.

6.5.1 Evaluación del modelo K-NN

Según los resultados obtenidos en el modelo K-NN con la matriz de confusión (Ver figura 18) se llegó a las siguientes interpretaciones:

- Se logró predecir correctamente como verdaderos EFECTIVOS un total de 11.259 registros, mientras que 3.217 registros que también lo eran los clasificó como NO EFECTIVOS lo que se conoce como un error de tipo II en estadística.
- El modelo predijo correctamente un total de 4.194 registros como NO EFECTIVOS, mientras que 1.705 registros que también lo eran los clasificó como EFECTIVOS lo que se conoce como un error de tipo I en estadística.
- Hay una precisión del 77.78% y una sensibilidad del 86.85% para la predicción de registros EFECTIVOS, lo que puede interpretarse como que el algoritmo predice adecuadamente la categoría, pero también incluye registros de la otra opción.
- Para los registros NO EFECTIVOS hay una precisión del 71.10% y una sensibilidad del 56.59%, que indica que el algoritmo no detecta la categoría muy bien, pero cuando lo hace una cantidad considerable es confiable.
- Se evidencia que la exactitud del modelo es de 75.84% y el score f1 es de 75.14%.

Figura 18. Matriz de confusión del modelo K-NN



Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

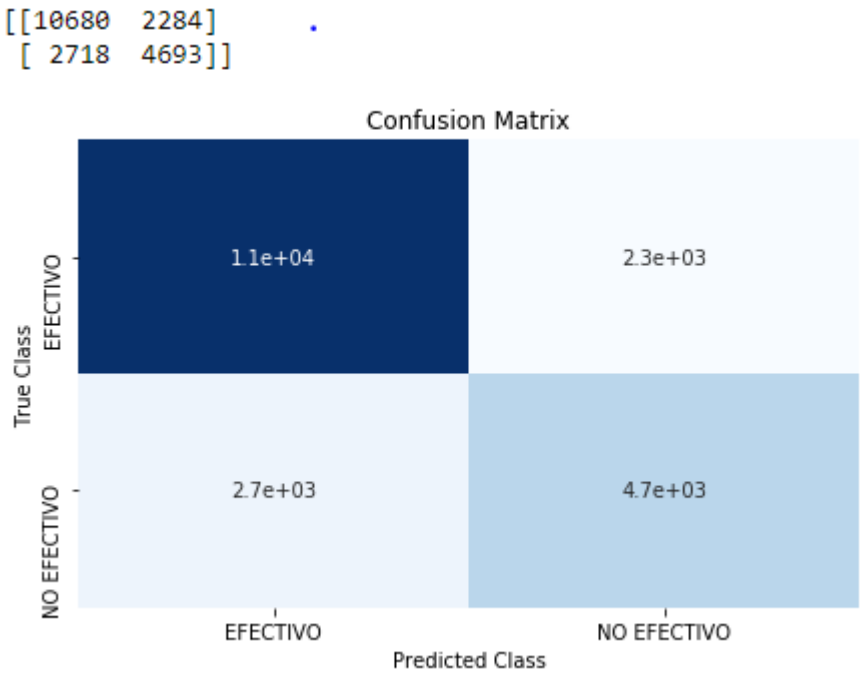
6.5.2 Evaluación del modelo Kernel SVM

Con los resultados obtenidos en el modelo de máquina de soporte vectorial y la generación de la matriz de confusión (Ver figura 19) se llegó a las siguientes interpretaciones:

- Se logró predecir correctamente como verdaderos EFECTIVOS un total de 10.680 registros, mientras que 2.718 registros que también lo eran los clasificó como NO EFECTIVOS.
- El modelo predijo correctamente un total de 4.693 registros como NO EFECTIVOS, mientras que 2.284 registros que también lo eran los clasificó como EFECTIVOS.

- Hay una precisión del 79.71% y una sensibilidad del 82.38% para la predicción de registros EFECTIVOS, lo que puede interpretarse como que el algoritmo predice adecuadamente la categoría, pero también incluye registros de la otra opción.
- Para los registros NO EFECTIVOS hay una precisión del 67.26% y una sensibilidad del 63.32%, que indica que el algoritmo no logra clasificar adecuadamente la categoría correcta.
- Se evidencia que la exactitud del modelo es de 75.45% y el score f1 es de 75.28%.

Figura 19. *Matriz de confusión del modelo de máquina de soporte vectorial*



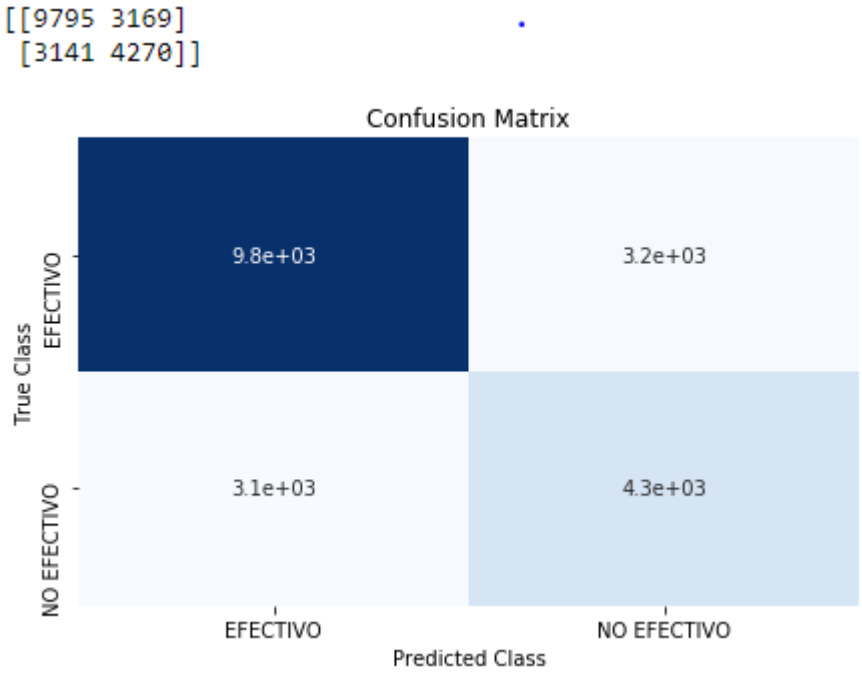
Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

6.5.3 Evaluación del modelo de árbol de decisión

Con los resultados obtenidos en el entrenamiento del modelo de árbol de decisiones y con la generación de la matriz de confusión (Ver figura 20) se llegó a las siguientes interpretaciones:

- Se logró predecir correctamente como verdaderos EFECTIVOS un total de 9.795 registros, mientras que 3.141 registros que también lo eran los clasificó como NO EFECTIVOS.
- El modelo predijo correctamente un total de 4.270 registros como NO EFECTIVOS, mientras que 3.169 registros que también lo eran los clasificó como EFECTIVOS.
- Hay una precisión del 75.72% y una sensibilidad del 75.56% para la predicción de registros EFECTIVOS, lo que puede interpretarse como que el algoritmo muchas veces no logra predecir adecuadamente la categoría.
- Para los registros NO EFECTIVOS hay una precisión del 57.40% y una sensibilidad del 57.62%, que indica que el algoritmo no logra clasificar adecuadamente la categoría correcta.
- Se evidencia que la exactitud del modelo es de 69.03 % y el score f1 es de 69.04%.

***Figura 20**. Matriz de confusión del modelo de árbol de decisión*



Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

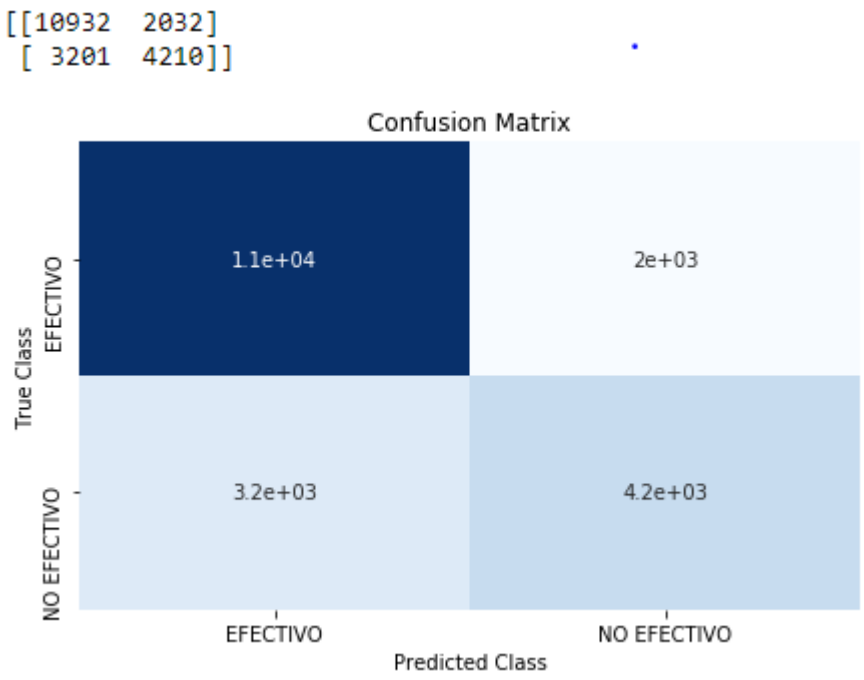
6.5.4 Evaluación del modelo de bosques Aleatorios

Por último, con los resultados obtenidos en el modelo de bosques aleatorios y la generación de la matriz de confusión (Ver figura 21) se llegó a las siguientes interpretaciones:

- Se logró predecir correctamente como verdaderos EFECTIVOS un total de 10.932 registros, mientras que 3.201 registros que también lo eran los clasificó como NO EFECTIVOS.
- El modelo predijo correctamente un total de 4.210 registros como NO EFECTIVOS, mientras que 2.032 registros que también lo eran los clasificó como EFECTIVOS.

- Hay una precisión del 77.35% y una sensibilidad del 84.33% para la predicción de registros EFECTIVOS, lo que puede interpretarse como que el algoritmo predice adecuadamente la categoría, pero también incluye registros de la otra opción.
- Para los registros NO EFECTIVOS hay una precisión del 67.45% y una sensibilidad del 56.81%, que indica que el algoritmo no logra clasificar adecuadamente la categoría correcta.
- Se evidencia que la exactitud del modelo es de 74.32% y el score f1 es de 73.77%.

Figura 21. *Matriz de confusión del modelo de bosques aleatorios*



Fuente: Elaboración propia tomado de Jupyter Notebook Python, 2022

7. Conclusiones

A partir de las gestiones registradas en las bases de datos en el último trimestre, fueron identificados a través de un análisis exploratorio el dataset con los principales atributos de los clientes de carteras de cobranza castigada que generan valor al objeto de estudio, con los cuales se pudo aplicar diferentes algoritmos enfocados al aprendizaje supervisado logrando realizar tareas de clasificación para lograr predecir los registros efectivos según la gestión.

Los modelos son útiles pero no se puede esperar un 100% de confiabilidad, por lo que es importante determinar según el objetivo propuesto qué porcentaje de exactitud y precisión son válidos para comenzar a implementar y poder sacar provecho de los resultados obtenidos, por lo tanto, con la evaluación e interpretación que se realizó de los cuatro (4) diferentes modelos realizados se concluyó que el que mejor se comportaba fue el algoritmo de k-NN con una exactitud de 75.84% y una precisión de 75.35%, valores que se toman como un muy buen resultado inicial, sobre el cual comenzar a enfocar las estrategias de gestión con el área de operaciones, aún más, teniendo en cuenta que no se contaba con ninguna herramienta que generará resultados de este tipo.

Se acepta el porcentaje de error que conlleva el modelo, porque al tratarse de un tema de cobranza, esto no conlleva a que los registros no sean gestionados, sino por el contrario, a que con la nueva información con la que va a contar la compañía se pueda implementar medidas diferentes de gestión, como son el tema de cantidad de intentos, horarios o selección de canales para la contactabilidad, que permitan mejorar los indicadores de efectividad y recaudo.

8. Referencias

Colcob (2018). Guía de mejores prácticas en la gestión de cobranza. Obtenido de https://colcob.com/images/pdf2018/20180227guiabuenaspracticascobranza.pdf

Debitia (2021). ¿Qué es la cartera de cobranza?. Obtenido de https://debitia.com.ar/que-es-la-cartera-de-cobranzas/#:~:text=El%20proceso%20de%20cobranzas%2C%20o,de%20un%20Pol%C3%ADtica%20de%20Cobranza

Brief (15/09/2022). Is a Global Recession Imminent?. The world bank. Obtenido de https://www.worldbank.org/en/research/brief/global-recession

Z. Sun, M. A. Wiering and N. Petkov (2014). Classification system for mortgage arrear management. IEEE Conference on Computational Intelligence for Financial Engineering & Economics (CIFEr), 2014, pp. 489-496, doi: 10.1109/CIFEr.2014.6924113

Grau H, (22/11/2017). ¿Qué es un canal digital? La guía definitiva. Telematel. Obtenido de https://www.telematel.com/blog/canal-digital-que-es-telematel/

Oracle. ¿Qué es el aprendizaje automático? Recuperado de https://www.oracle.com/co/artificial-intelligence/machine-learning/what-is-machine-learning/

Google Cloud. ¿Qué es el aprendizaje automático? Recuperado de https://cloud.google.com/learn/what-is-machine-learning?hl=es-419

Pádraig Cunningham and Sarah Jane Delany. 2021. k-Nearest Neighbour Classifiers - A Tutorial.ACM Comput. Surv. 54, 6, Article 128 (July 2021). Obtenido de https://dl.acm.org/doi/pdf/10.1145/3459665

Garcia Z, 2017. Implementación en MATLAB del algoritmo MTS para problemas de predicción con salidas compuestas. Pag, 29, obtenido de https://repositorio.uci.cu/bitstream/123456789/8091/1/TD_08853_17.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Tibco. ¿Qué es un bosque aleatorio?. Recuperado de https://www.tibco.com/es/reference-center/what-is-a-random-forest

Bahit E, (2012). Python para principiantes. Recuperado de http://46.101.4.154/Libros/El%20Lenguaje%20Python.pdf

Aristizabal A, Botache D, Concha C, Gonzalez D, (2018). Software gestor de cobranza usando inteligencia artificial de la corporación universitaria unitec durante el periodo 2.017 – 2.018. Recuperado de https://repositorio.unitec.edu.co/bitstream/handle/20.500.12962/1257/Software%20gestor%20de%20cobranza%20usando%20inteligencia%20artificial.pdf?sequence=1&isAllowed=y

Montoya J, (2019). Analítica De Datos Aplicada A La Cobranza De Cartera. Recuperado de https://repository.eafit.edu.co/bitstream/handle/10784/13894/JuanDavid_MontoyaYepes_2019.pdf?sequence=1&isAllowed=y

Campos Zuleyca (2020). Pronóstico del cumplimiento de pago de los clientes usando aprendizaje automático. Recuperado de https://repositorio.unal.edu.co/bitstream/handle/unal/78297/642171.2020.pdf?sequence=5&isAllowed=y

Cobos Carlos, Zuñiga Jhon, Guarin Juan, León Elizabeth, Mendoza Martha, & Universidad Nacional de Colombia. Facultad de Ingeniería. (2018). CMIN - herramienta case basada en CRISP-DM para el soporte de proyectos de minería de datos.

Granado Eduardo, Diaz Elena (2018). *Manual de uso de jupyter notebook para aplicaciones docentes*. Universidad Complutense de Madrid. Recuperado de https://eprints.ucm.es/id/eprint/48304/1/ManualJupyter.pdf

IBM (17/08/2021). Guía de CRISP-DM de IBM SPSS Modeler. Edicion 18, release 4. Obtenido de https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf