UNIVERSIDAD TÉCNICA DE MACHALA
FACULTAD DE INGENIERÍA CIVIL

MAESTRÍA EN SOFTWARE

APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO EN EL ANÁLISIS DE CARTERA DE UNA EMPRESA PÚBLICA DE AGUA, ALCANTARILLADO Y ASEO

MIRANDA GALLEGOS JORGE LUIS

VIRTUAL

TUTOR(A) ING. BERTHA MAZÓN OLIVO
COTUTOR ING. EDUARDO TUSA

MACHALA
2023

PENSAMIENTO

“El conocimiento es poder”

-Francis Bakon

DEDICATORIA

En primer lugar, doy gracias a Dios por la oportunidad de tenerme en este camino del saber, a mis padres Laura Gallegos, Luis Miranda por el apoyo y confianza que me brindan desde muy pequeño, Patricia Miranda por siempre estar presente ante cualquier situación y a mi esposa Karen Cueva e hijos Jorge Alejandro y Thiago Emiliano que han sido mi motivo principal para continuar en este camino, camino que nos ha tocado duro pero juntos de la mano de Dios hemos aprendido que las enseñanzas de Dios vienen en forma de pruebas muy grandes.

AGRADECIMIENTOS

A mi tutora, Ing. Bertha Mazón por su confianza y paciencia. A mi madre Laura Gallegos, mi esposa Karen Cueva por su apoyo incondicional en este camino y a todos los que me han podido dar la mano de alguna u otra forma

RESPONSABILIDAD DE AUTORÍA

Yo, Jorge Luis Miranda Gallegos con C.C./C.I./Pasaporte 0704720705, declaro que el trabajo de "APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO EN EL ANÁLISIS DE CARTERA DE UNA EMPRESA PÚBLICA DE AGUA, ALCANTARILLADO Y ASEO", en opción al título de Magister en Maestría en Software, es original y auténtico; cuyo contenido: conceptos, definiciones, datos empíricos, criterios, comentarios y resultados son de mi exclusiva responsabilidad.

MIRANDA GALLEGOS JORGE LUIS

C.C./C.I./Pasaporte 0704720705

Machala, 2023/05/07

APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO EN EL ANÁLISIS DE CARTERA DE UNA EMPRESA PÚBLICA DE AGUA, ALCANTARILLADO Y ASEO

INFORME DE ORIGINALIDAD

7%	7%	3%	2%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	www.derechoecuador.com Fuente de Internet	1%
2	www.ame.gob.ec Fuente de Internet	1%
3	repository.eafit.edu.co Fuente de Internet	1%
4	risti.xyz Fuente de Internet	1%
5	datospdf.com Fuente de Internet	1%
6	www.ucv.edu.pe Fuente de Internet	1%
7	sriagral.uabc.mx Fuente de Internet	<1%
8	www.researchgate.net Fuente de Internet	<1%

CERTIFICACIÓN DEL TUTOR

Yo, Bertha Eugenia Mazón Olivo con C.C./C.I./Pasaporte 0603100512; tutor del trabajo de "APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO EN EL ANÁLISIS DE CARTERA DE UNA EMPRESA PÚBLICA DE AGUA, ALCANTARILLADO Y ASEO", en opción al título de Magister en Maestría en Software, ha sido revisado, enmarcado en los procedimientos científicos, técnicos, metodológicos y administrativos establecidos por el Centro de Posgrado de la Universidad Técnica de Machala (UTMACH), razón por la cual doy fe de los méritos suficientes para que sea presentado a evaluación.

Bertha Eugenia Mazón Olivo

C.C. /C.I ./Pasaporte 0603100512

Machala, 2023/05/07

CESIÓN DE DERECHOS DE AUTOR

Yo, Jorge Luis Miranda Gallegos con C.I. 0704720705, autor del trabajo de titulación "APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO EN EL ANÁLISIS DE CARTERA DE UNA EMPRESA PÚBLICA DE AGUA, ALCANTARILLADO Y ASEO.", en opción al título de MAGISTER EN SOFTWARE, declaro bajo juramento que:

- El trabajo aquí adjunto es de mi autoría, que no ha sido presentado previamente para ningún grado o calificación profesional. En consecuencia, asumo una responsabilidad frente a cualquier reclamo o demanda por parte de terceros de manera exclusiva.
- Sexo la Universidad técnica de Machala de forma exclusiva con referencia a la obra en formato digital los derechos de:
    a. Hola incorporar la mencionada obra en el repositorio institucional para su demostración a nivel mundial, respetando lo establecido por la licencia *Creative Commons Atribution no commercial Igual 4.0 International (CC BY NCSA 4.0);* la Ley de Propiedad intelectual del Estado ecuatoriano y el régimen institucional.
    b. Adecuarla a cualquier formato o tecnología de uso internet, así como correspondiéndome como autor la responsabilidad de velar por dichas adaptaciones con la finalidad de que no desnaturalice el contenido o sentido de la misma.

Firmado electrónicamente por:
JORGE LUIS MIRANDA GALLEGOS

JORGE LUIS MIRANDA GALLEGOS

C.I. 0704720705

RESUMEN

En la actualidad las nuevas tendencias e innovaciones tecnológicas como la aplicación de aprendizaje automático son de gran importancia para las empresas e industrias para impulsar una eficiencia en la toma de decisiones. En ese sentido la Empresa Pública de Agua, Alcantarillado y Aseo (EPAAA) del cantón Pasaje, busca apoyar sus decisiones basándose en el análisis de cartera para la gestión de recaudaciones de la empresa, ante esta problemática se desarrolló un Sistema de Soporte de Decisiones (DSS) aplicando inteligencia de negocios, minería de datos y aprendizaje automático, utilizando la metodología *CRISP-DM* la cual consta de 6 fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Con la implementación del DSS es posible determinar qué está sucediendo en cuanto a la cartera y la gestión de recaudaciones; además, se puede realizar la caracterización de los clientes, clasificar los clientes por sector y ciclo, clasificación de clientes pagadores, clasificación de clientes deudores y apoyó para la gestión de cartera. El DSS está compuesto de una interfaz *dashboard* que consta de gráficos estadísticos, clasificaciones y predicciones de los abonados. Los resultados obtenidos con la implementación del DSS en la EPAAA permitió un 94.44% de eficiencia en la toma de decisiones, resultado que se obtuvo mediante una encuesta a los colaboradores que toman decisiones en la EPAAA. En conclusión, este trabajo nos permite mejorar los tiempos de respuesta de los colaboradores generando una eficiente gestión de recaudaciones, el DSS logró repotenciar los conocimientos sobre la cartera permitiendo caracterizar a los clientes, un logro importante es que se pudo clasificar a los clientes en función a la deuda donde producto de esto, se pudo tomar mejores decisiones.

PALABRAS CLAVES: minería de datos, aprendizaje automático, sistema de toma de decisiones, servicio de agua potable

ABSTRACT

Currently, new trends and technological innovations such as the application of machine learning are of great importance for companies and industries to promote efficiency in decision making. In this sense, the Public Water, Sewerage and Cleaning Company (EPAAA) of the Pasaje canton, seeks to support its decisions based on the portfolio analysis for the company's collection management, in view of this problem, a Decision Support System was developed (DSS) applying business intelligence, data mining and machine learning, using the CRISP-DM methodology which consists of the phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. With the implementation of the DSS it is possible to determine what is happening in terms of the portfolio and collection management; In addition, it is possible to characterize the clients, classify the clients by sector and cycle, classification of paying clients, classification of debtor clients and support for portfolio management. The DSS is composed of a dashboard interface consisting of statistical graphs, classifications and subscriber predictions. The results obtained with the implementation of the DSS in the EPAAA allowed a 94.44% efficiency in decision-making, a result that was obtained through a survey of the collaborators who make decisions in the EPAAA. In conclusion, this work allows us to improve the response times of the collaborators, generating an efficient management of collections, the DSS managed to strengthen the knowledge about the portfolio, allowing to characterize the clients, an important achievement is that the clients could be classified according to to the debt where as a result of this, better decisions could be made.

KEYWORDS: data mining, machine learning, decision making system, drinking water service

Índice General

Lista de Ilustraciones

Lista de Tablas

LISTA DE ABREVIATURAS

EPAAA – Empresa Pública de Agua, Alcantarillado y Aseo

ANN – Redes Neuronales Artificiales

GAD – Gobierno Autónomo Descentralizado

ETL – Extracción, transformación y carga

SVM – Soporte de Maquina de Vectores

KNN – Algoritmo de K vecinos más cercanos

MV – Varianza Media

BI – Inteligencia de Negocios

DBN – Red de creencias Profundas

DSS – Sistema de soporte de Decisiones

INTRODUCCIÓN

Con el fin de mejorar la gestión de recaudaciones y análisis de cartera la **importancia del tema** es desarrollar un Sistema de Soporte de Decisiones (DSS) en la Empresa Pública de Agua, Alcantarillado y Aseo (EPAAA) del cantón Pasaje, El Oro – Ecuador aplicando aprendizaje automático, se decidió usar los datos históricos ya almacenados para apoyar la toma de decisiones de los mandos tácticos y estratégicos de la institución. Para ello fue necesario la implementación de modelos de analítica de datos creando un nuevo almacén de datos [1] que permitiera acceder de forma eficiente a la información mediante una minería de datos [2], esta nació con el propósito de aprovechar la cantidad de datos que se almacena, la potencia de nuevos ordenadores para las operaciones de análisis de los mismos [3] [4] [5] y evaluar los puntos débiles que se deben mejorar al momento de hacer la gestión del usuario.

La predicción de datos es una herramienta eficiente para evaluar riesgos en el mercado realizando predicción financiera [6], se propone desarrollar un DSS en la EPAAA de Pasaje, mediante el análisis de cartera empleando algoritmos de aprendizaje automático para una gestión eficiente de recaudaciones [7] encargado de hacer la segmentación de los deudores por las características que este posee, además predecir las probabilidades de pago y una estrategia adecuada para la recuperación de cartera.

Las deudas en una empresa son consideradas como un grupo principal en los servicios de las instituciones y una gran fuente de ganancia, donde se destaca la efectividad de aplicar inteligencia de negocios, minería de datos [8] y aprendizaje automático como arboles de decisiones y redes neuronales, otros algoritmos ANN que se utiliza para detectar transacciones fraudulentas en algunas instituciones financieras, modelos forestales aleatorios y de línea de base simples para evaluar el riesgo de dar préstamos [9].

Las empresas en la actualidad buscan implementar estrategias y una de las estrategias según la encuesta del Instituto de Alto Rendimiento de Accenture en Estados Unidos muestra que las industrias han implementado equipamiento con aprendizaje automático [10].

En la creación de la Empresa el 30 de octubre del 2014, esta heredó la data de un sistema transaccional que fue migrado desde el GAD Pasaje hacia la EPAAA de Pasaje, mucha

de esta data es duplicada e incompleta, producto de ésta migración la empresa desde su inicio ya contaba con una gran cantidad de cartera vencida, el tiempo en la generación de reportes consolidados para enfrentar ésta gestión de cartera conllevan demasiado tiempo, siendo esta actualmente la **problemática que enfrenta** la empresa. Este proceso de migración provocó que las estrategias que se implementen para un adecuado manejo de cartera y gestión de recaudación sean ineficientes. Además, se identifica la falta de pago de ciertos sectores del cantón, información homogénea siendo el resultante el incremento de cartera y una tardía respuesta en la toma de decisiones de la EPAAA, falta de indicadores en cuanto a recaudación neta de una emisión de facturación para así poder identificar si el porcentaje de recaudación de esa emisión es normal. Actualmente solo se puede identificar la recaudación que se realiza en el mes, independientemente de las emisiones de facturación que se haya recaudado, comparativas entre recaudaciones. Además, se necesita poder conocer estados de recaudación por categorías, es prescindible conocer que cada cuenta de agua potable tiene definida una categoría y que según su categoría esta tendrá un valor de facturación. En ese sentido se puede definir que estas son las **causas que originan el problema científico.**

Producto de esto resultó la **formulación del problema** de este trabajo y es en **¿Cómo apoyar las decisiones basadas en el análisis de cartera para la gestión de recaudaciones en la EPAAA?.**

El **objeto de estudio** de este trabajo es la gestión de cartera en la Empresa Pública de Agua, Alcantarillado y Aseo. Además, El **campo de acción** de este trabajo se delimita a la aplicación de las tecnologías: inteligencia de negocios, minería de datos y aprendizaje automático, en el desarrollo del DSS para la EPAAA.

Este trabajo tiene como **objetivo general** desarrollar un sistema de soporte de decisiones en la EPAAA de Pasaje, mediante el análisis de cartera empleando algoritmos de aprendizaje automático para una gestión eficiente de recaudaciones. Además, se desarrolló los siguientes objetivos específicos:

- Caracterizar algoritmos y técnicas de minería de datos y aprendizaje automático, que permitan el análisis de cartera y gestión de recaudación mediante una revisión sistemática de la lectura.

- Evaluar los modelos de aprendizaje automático implementados mediante métricas de rendimiento en la toma de decisiones en la EPAAA de Pasaje.
- Implementar un *DSS* con interfaz *dashboard* en la EPAAA de Pasaje que facilite el análisis de cartera y gestión de recaudación.

Se establece como **hipótesis**, **el desarrollo de un DSS en la EPAAA de Pasaje, con algoritmos de aprendizaje automático mejora la forma en la que se gestiona la cartera y recaudaciones. Además, garantiza el nivel de satisfacción de usuarios**. Esta gestión de análisis de cartera y recaudaciones se medirá mediante la satisfacción del usuario al ejecutar una encuesta a 12 colaboradores de la EPAAA, quienes intervienen directamente en la toma de decisiones

Se **justifica** el trabajo reflejando con claridad las fuentes teóricas que llevan al problema sobre como apoyar las decisiones basadas en el análisis de cartera para la gestión de recaudaciones en la EPAAA del cantón Pasaje. Con el fin de cumplir el objetivo de este trabajo en desarrollar un DSS en la EPAAA de Pasaje, mediante el análisis de cartera empleando algoritmos de aprendizaje automático para una gestión eficiente de recaudaciones. Este problema nos conduce a elaborar un diseño de investigación que se desarrolla en 3 etapas que son el enfoque, alcance y análisis. Donde según Hernández-Sampieri [11] el enfoque de esta investigación es cuantitativo debido a que se pudo establecer una población de usuarios para poder medir resultados objetivamente y así poder medir la satisfacción y efectividad del DSS, el alcance de esta investigación es correlacional gracias a que se pudo comparar más de un indicador como por ejemplo el total de la deuda con el total meses adeudados y se pudo identificar relaciones entre ellas y el análisis de esta investigación es cuasiexperimental ya que ponemos a prueba la hipótesis planteada en este trabajo.

Para el desarrollo del DSS no se utilizó una muestra en específico, sino se utilizó los datos reales de la EPAAA que corresponde al número de clientes activos a los cuales se les emiten cartas de pago al mes, por esta razón se determina que contamos con una población finita de 18670 clientes activos mensuales, estos clientes corresponden al periodo comprendido entre el enero 2002 y febrero 2023. Es decir que **se obtuvo una población de 4704840 de datos** para el desarrollo del DSS. Además, contamos con una segunda unidad de investigación de 12 usuarios que también toman decisiones en la EPAAA que

comprenden las áreas de tesorería, cartera, gerencia, comercial y financiero; a los cuales se les aplicó una entrevista.

Además, se desarrolló el DSS aplicando metodología CRISP-DM ya que su metodología es idónea para llevar a cabo las tecnologías a implementar en este proyecto como son la minería de datos, inteligencia de negocios y aprendizaje automático, esta metodología consta de 6 fases como son entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y despliegue.

El trabajo desarrollado es de un **impacto** positivo desde el lado político – empresarial, por tal motivo es muy importante el desarrollo, implementación y puesta en marcha del DSS en la EPAAA del cantón Pasaje, mejorando los servicios a la comunidad porque se ejecutarán tomas de decisiones más rápidas y eficientes, la empresa mejora en la gestión de recaudación por haber mejorado en el análisis de su información generando un aporte importante a la empresa con la implementación del DSS aplicando aprendizaje automático.

El **propósito** de la aplicación de aprendizaje automático para la gestión de cartera en la EPAAA, se debe a la necesidad de determinar la cantidad de deuda que tienen los abonados con respecto al tiempo de cancelación de la deuda, clasificar abonados que tengas problemas de deudas incobrables y no mezclarla con la cartera vencida real. Se llama deuda incobrable porque esta deuda no tiene un abonado deudor y también clasificar deudas homogéneas. Es decir, hay repetidos valores de deudas cargados a una sola persona.

Con la implementación de estos algoritmos se ayudará a identificar el comportamiento de un abonado, debido a que no existen cortes de agua potable por acumular deuda, dando paso a que el abonado pueda atrasarse en los pagos de agua potable ya sea de 3, 6 o 9 meses o por años, con la implementación de aprendizaje automático podremos saber que usuarios son pagadores o deudores. Podremos clasificar la deuda, sobre todo clasificaremos que parroquia y que ruta es pagador o deudor y de tal manera identificar si existe anomalías en cuanto al servicio de agua potable. El propósito más importante es que estas herramientas permitan tomar decisiones.

Este **documento se estructura** en cuatro capítulos, la introducción que da una idea al lector sobre a dónde va la propuesta y la importancia de la misma. En el Capítulo I, se

compone de los antecedentes históricos, conceptuales y contextuales, todos enfocados en el marco de la investigación realizada enfocados en estudios similares. El Capítulo II indica la metodología a seguir, los materiales y métodos. En el Capítulo III se detallan los resultados, fundamentados en los aportes prácticos. En el Capítulo IV se realiza la evaluación y discusión de los resultados obtenidos sobre los hallazgos y finalmente las conclusiones.

CAPÍTULO I. MARCO TEÓRICO

En este capítulo se realizó el estado del arte aplicando la metodología de revisión sistemática de la literatura. La organización de este capítulo comienza con los antecedentes históricos.

1.1 Antecedentes Históricos

La EPAAA de Pasaje fue creada el 30 de octubre de 2014, cuyo objetivo de la Empresa es la prestación de los servicios públicos de agua potable, alcantarillado, depuración de aguas residuales, barrido, limpieza, recolección, transporte, tratamiento y disposición final de residuos sólidos no peligrosos y peligrosos, sus servicios complementarios, conexos y afines que pudieren ser considerados de interés colectivo, otros servicios que resuelva el directorio, así como la gestión de sectores estratégicos, el aprovechamiento sustentable de recursos naturales o de bienes públicos y en general al desarrollo de actividades económicas conexas a su actividad que correspondan al Estado.

Antes de su creación los servicios básicos lo administraban el GAD Pasaje mediante una dirección, al momento de la creación de la EPAAA todos los datos e información acerca del giro de negocio fue migrada a una nueva base de datos transaccional. Este panorama puede dar problemas si se suma información desconocida o errónea: Claves catastrales, el panorama competitivo o los cambios en el comportamiento de los consumidores, datos del cliente incompletos, direcciones.

El manejo de herramientas como inteligencia de negocios, minería de datos y aprendizaje automático generan un gran apoyo en cuanto a la toma de decisiones de los altos mandos y mandos medios. En el 2005 D. Conti y A. Rodríguez realizaron un estudio sobre una combinación de dos criterios para evaluar carteras en este caso, cartera de inversión, el primero es para medir el rendimiento en función de la ganancia esperada y el segundo mide el riesgo de la varianza [12]. Basado en estos dos criterios sobre todo en la ganancia esperada de un activo cualquiera está definida como la diversificación se requiere identificar que el coeficiente de correlación entre activos sea negativo o aproximado a cero, ya que así se reduce el riesgo de cartera y de esa manera se compensan las pérdidas con las ganancias de otros rubros registrados en la cartera.

Este estudio los autores generaron carteras de alto rendimiento donde la aplicación de Redes Neuronales creo una cierta eficiencia para clasificar acciones que conforman la cartera de inversiones y se puede sustentar que el uso de Redes Neuronales puede ser aplicada como una técnica alternativa para carteras de inversión.

Las técnicas de minería de datos como por ejemplo la clasificación que es uno de los datos muy importantes de la técnica minería de datos y la agrupación son utilizadas para estudiar las correlaciones entre fenómenos de las deudas. Por lo tanto el método de la clusterización es un método en que el objetivo primordial es la clasificación de los datos, identificando grupos similares [13], donde implementaron en el año 2013 en un modelo para la crisis de la deuda soberana. Los autores utilizaron técnicas como C 4.5, CART, Logistic Model Trees, Random Forest, Alternating Decision Tree, Naïve Bayesian Classifier, Bayesian Logistic Regresión donde al momento de procesar datos obtuvieron como conclusión un árbol binario con múltiples aspectos sobre las interdependencias entre tasas de inflación y PIB, obtuvieron una probabilidad debido a la clasificación óptima entre el conjunto de datos, reglas de clasificación y reglas de parada.

El uso de modelos de prevención de riesgos financieros basados en análisis estadístico tradicional, nace de la minería de datos ya que hace que cada vez se genere más información de big data [14], donde Khemakhem y Boujelbene construyeron modelos de predicción de riesgo empresarial utilizando redes neuronales artificiales y el método de árbol de decisiones [15], estas son soluciones analíticas dinámicas, construidas mediante software para obtener ventajas competitivas, puede discernir las situaciones rápidamente y ajustarlas según sea necesario para tomar decisiones más concretas. Los autores de esta investigación usaron la técnica SMOTE ya que ayuda a mejorar la estabilidad de los árboles de decisiones y de las redes neuronales artificiales. Además, mejoraron el rendimiento de los datos desequilibrados, frente a la problemática a resolver pudieron concluir sobre la rentabilidad y capacidad de pago y la solvencia con ratios, indicadores importantes en cuanto a la insolvencia de la empresa, los resultados ayudaron a sugerir la importancia sobre estudios en cuanto al riesgo empresarial, además se valida la importancia del uso de estas técnicas en cuanto a la predicción en riesgos financieros.

Ya en el 2017 los métodos de aprendizaje automático normalmente se dividieron en dos fases, la fase de entrenamiento y la fase de predicción, en la primera fase se construye un modelo predictivo para aprender el patrón de las características de entrada y sus salidas.

En la predicción se aplica para predecir las salidas de lo desconocido [16]. En una gestión de carteras de rentas vitalicias variables grandes Wei Xu, Yuehuan Chen, Conrad Coleman, Thomas F. Coleman utilizaron *Matching Machine Learning* como enfoque y en entrenamiento de la máquina de aprendizaje se realizó mediante regresión. Como resultado pudieron obtener que los cálculos fueron precisos y que se optimiza el tiempo de cálculo de las operaciones que el que requiere las simulaciones anidadas.

En un enfoque de fusión de máquinas de aprendizaje y carteras en el 2018 F. D. Paiva, R. T. N. Cardoso, G. P. Hanaoka, y W. M. Duarte propusieron un estudio de toma de decisiones donde aplicaron un modelo de fusión de un clasificador basado en el aprendizaje con el método de la máquina de vectores de soporte SVM y la varianza media (MV) que es método clásico de selección de cartera, donde sus resultados dieron un alto rendimiento para el modelo SVM + MV [17], el rendimiento de clasificación que obtuvieron los autores fueron superiores a la probabilidad de ocurrencia de los eventos calculados, también el clasificador discrimina potencialmente cuando necesita trabajar con objetivos superiores. El modelo presentó resultados importantes en el desempeño de rentabilidad, fue satisfactorio mejor que otros modelos.

En una revisión en el 2018 sobre las deudas de cartera en el aprendizaje automático, los autores recomiendan aprender varios modelos, es decir combinar las variaciones en lugar de encontrar la mejor, en este caso aplicaron 3 variaciones del conjunto bagging, boosting y blending [18]. Para nuestra propuesta se eligió el aprendizaje automático [19] donde se sabe que la agrupación de clusters es un problema en los análisis de datos más conocidos y estudiados. Donde se constituye un área de investigación clave en el campo del aprendizaje, donde no existe supervisión alguna de cómo debe manejarse la información que se estudia. Donde la tarea de agrupar un conjunto de datos *K* lo definen como agrupamiento particional [20].

En la actualidad 2021 han propuesto la utilización del modelo XGBoost de aprendizaje automático para deudas a largo plazo y corto plazo [21], ya que este algoritmo se basa en la aproximación de funciones mediante la optimización y se llegó a comparar modelos de aprendizaje automático supervisado y no supervisado, donde el clasificador no supervisado de la red de creencias profundas (DBN) y el modelo hibrido DBN-SVN puedo generar pronósticos más precisos que XGBoost [22], este segundo era capaz de generar en cambio predicciones más precisas. Los autores concluyeron que estas técnicas

son capaces de predecir dificultades financieras.

Preguntas de investigación

Se plantean las siguientes preguntas para este proyecto a desarrollar, en relación a la investigación de la minería de datos [9], inteligencia de negocios y aprendizaje automático:

¿Cuáles son los beneficios del uso de modelos de aprendizaje automático en el análisis de cartera y gestión de recaudación en la EPAAA?

¿Qué impacto genera en la EPAAA la implementación de algoritmos basados en inteligencia de negocios y minería de datos para la toma de decisiones?

¿Cuál es la situación actual de la empresa sin la implementación de un DSS para el análisis de cartera y gestión de recaudación?

¿Qué técnicas o herramientas serian idóneas a usar para una buena implementación de algoritmos basados en inteligencia de negocios y minería de datos para la toma de decisiones en la EPAAA?

Criterios de inclusión o Exclusión

Estos criterios son un control que sirven para verificar que los resultados que se vayan a obtener se encuentren apegadas a las necesidades de la presente investigación:

Inclusión

- Estudios relacionados al aprendizaje automático.
- Estudios relacionados a la cartera vencida.
- Estudios relacionados a la minería de datos
- Estudios relacionados a la ciencia de datos
- Publicaciones a partir del año 2005.
- Fuentes de artículos científicos y libros

Exclusión

- Estudios no referentes al aprendizaje automático
- Estudios duplicados

- Fuentes de terceros

Línea de tiempo

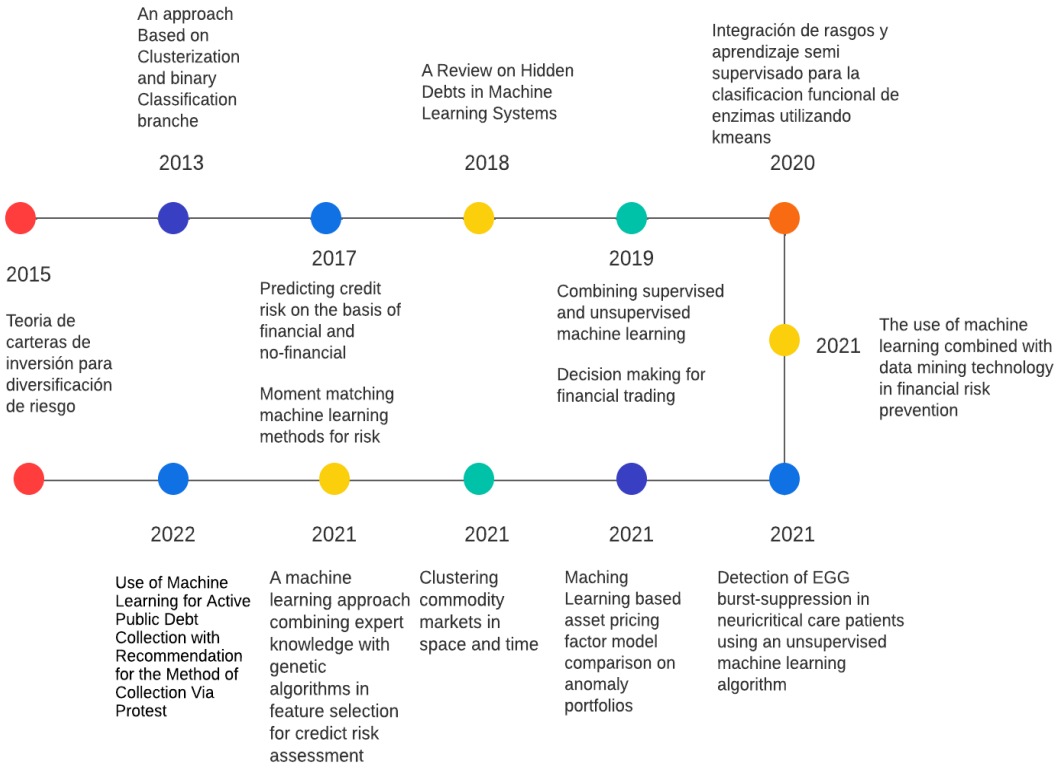


Ilustración 1 Línea de tiempo

1.2 Antecedentes conceptuales y referenciales

Para este apartado de antecedentes conceptuales haremos referencia sobre las variables de investigación (sección 2.1) que estudiaremos para realizar este proyecto.

Las técnicas seleccionadas para el desarrollo del trabajo se han determinado por la revisión sistemática de la lectura, debido a que se ajustan a los datos de la EPAAA y al giro de negocio; por ser datos financieros se sabe que las deudas crecen linealmente y la necesidad de clasificación de cartera, clientes y ciclos. Los algoritmos a utilizar basados en la investigación referentes a la inteligencia de negocios, minería de datos y aprendizaje automático se describen a continuación:

- Correlación
- Regresión Lineal
- Clustering
  - K-Means
- Clasificación
  - SVM
  - Perceptrón Multicapa
  - Regresión Logística
  - Arboles de decisión
  - Clasificador de Naive Bayes
  - Clasificador K-NN
  - Clasificador de bosques aleatorios

Cabe recalcar que estos modelos permiten detectar anomalías, por ejemplo, transacciones inusuales, evitar fraudes, defectos en el manejo de la cartera, entre otros. Estos algoritmos se implementaron en el DSS, utilizando lenguaje *Python* haciendo uso de una librería muy potente de aprendizaje automático de código abierto *scikit-learn* [23]. Los modelos de aprendizaje automático se subdividen en dos como lo muestra la **Ilustración 2**.

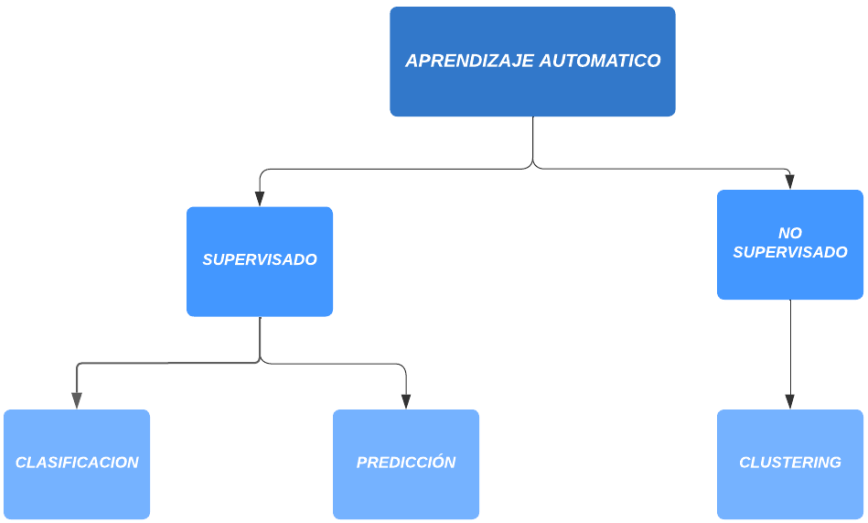


Ilustración 2 Aprendizaje automático

1.2.1 Minería de datos

La Minería de datos es una disciplina que complementa la inteligencia de negocios y se fundamenta en la estadística, matemáticas y ciencias computacionales. Diagnostica patrones, clasifica y predice eventos que pueden ocurrir en una empresa, además que como una tarea principal la minería de datos ejecuta procesos exploratorios y de análisis de la información [2].

La minería de datos [24] [25] es parte de la ciencia de datos el cual se caracteriza por encontrar patrones, son una herramienta eficiente para poder enfrentar la gran cantidad de datos que posee una empresa [26], además de que nos permite poder tener una ventaja competitiva sobre las demás empresas donde no han apostado por una transformación tecnológica [27] en sus empresas.

1.2.1.1 Correlación

La correlación es un método estadístico que nos ayuda a comparar variables, es decir si una variable incrementa en compañía del aumento de otra, se dice que la correlación es directa o positiva [28]. En cambio, si el aumento de una variable depende de la disminución de otra se dice que es inversa o negativa, para este caso se supone que las variables que tiene la EPAA son simétricas.

1.2.1.2 Regresión Lineal

Es un modelo muy usado para analizar comportamientos de variables tanto de entrada y de salida donde se podría establecer predicciones y estimaciones. Esta variable es idónea para implementar en la EPAAA [29]. Es ideal aplicar este modelo por la cartera de deuda que maneja la EPAAA el cual tiene un crecimiento lineal.

1.2.1.3 Clustering

Clustering es una técnica de aprendizaje automático no supervisada que resuelve problemas de agrupamiento, nos permite descubrir patrones de objetos según su posición [30]. La utilización de esta técnica se debe a la complejidad, dimensionalidad del dataset de la EPAAA, donde es idóneo su utilización por el gran volumen de datos que posee la empresa.

1.2.2 Inteligencia de Negocios

La inteligencia de negocios ya es aplicada en muchos giros de negocios, el cual representa una ventaja competitiva, gracias a la ayuda de la transformación tecnológica donde esto permite generar procesos de calidad que generan modelos de gestión con gran madurez. La inteligencia de negocios en conjunto con la analítica de datos son un factor importante para el rendimiento empresarial y para la toma de decisiones [31].

En épocas que no contaban con inteligencia de negocios [32], el análisis de la información la realizaban de forma intuitiva, ahora las empresas con la cantidad de información que tienen y muchas sin saber qué hacer con esta información, la inteligencia de negocios llego para convertir toda esa información en conocimiento [33]. Es una disciplina joven donde nace a partir del año 2000 y se consolida la inteligencia de negocios 2.0 mediante plataformas BI gestionando y analizando información [34].

ETL

El ETL [35] por sus siglas hace referencia a la extracción, transformación y carga de los datos, para este trabajo se utilizó *pentaho data integration* donde en primer lugar se procedió a seleccionar la información que se necesitaba para el DSS, posterior se realizó una transformación de los datos seleccionados y por último en este proceso se extrajo todos esos datos, dando paso a la creación de una nueva base de datos limpia.

DATA WAREHOUSE

Un *data warehouse* [36] es un repositorio de los nuevos datos, es decir un almacén de datos que se generaron mediante el ETL anterior, donde se encuentra unificada toda la información integral diseñada para que las consultas sean más fáciles de escribir, brindando al usuario un rendimiento notable.

Dashboard

El dashboard es una herramienta para aumentar la gestión de la información donde esa información se monitorea, analiza y se muestra mediante claves de desempeño que se reflejan en indicadores o métricas para conocer el estado de la empresa y poder dar un seguimiento eficaz. El dashboard [37] se caracteriza porque esa información se personaliza, además que la información que se obtiene se presenta con gráficos en tiempo

real.

1.2.3 Aprendizaje Automático

El aprendizaje automático se divide en dos áreas importantes [38], en este caso como se había mencionado en el área de aprendizaje automático [42] no supervisado donde se han realizado investigaciones importantes, una de ellas realizó una evaluación cuantitativa en línea de la profundidad del coma en pacientes de UCI, como resultado dieron que el rendimiento del algoritmo fue estable y con alta precisión en un grupo heterogéneo de pacientes, para este caso utilizaron el algoritmo de agrupación espectral usando el algoritmo k-medias en dos grupos y también se usó el paquete de *python sklearn.cluster*.*SpectralClustering* [39], asi como también K-nearest para predicción de nuevos datos.

En otro caso se realizó el aprendizaje automático para identificar el aprendizaje de los alumnos utilizando técnicas de minería de uso web y algoritmos de aprendizaje, entre ellos en este caso se utilizó algoritmos de agrupación de modos K y un algoritmo basado en el FSLSM, donde como resultado dieron que basado en el enfoque que implementaron funciono bien [40]. En una comparación de modelos para el aprendizaje automático en carteras de anomalías con distintas técnicas de regresión lineal con ligeras modificaciones con una versión regularizada, elastic-net y su extensión impulsada, tuvieron un aumento en el rendimiento predictivo [43] en las carteras de anomalías, modelos más complejos como random forest, gradient boosted tree, and neural network based predictors no tuvieron estos resultados [41].

Los riesgos crediticios también se han podido analizar mediante algoritmos de aprendizaje automático, ya que se verifica el incumplimiento al momento de realizar pagos donde en este enfoque han hecho énfasis en el historial crediticio, capacidad de pago, capital, condición de préstamos y el garante. En este caso para la transformación de datos los autores utilizaron el conjunto de datos, lo estandarizaron para tener una media de 0 y desviación estándar de 1. El enfoque también se conoce como método de puntuación z [21], en la división de datos, utilizaron una validación cruzada estratificada con k=10 que se utiliza para dividir el conjunto de entrenamiento y para la agrupación utilizaron clustering, donde el algoritmo de k-medias se utiliza para agrupar características y par aun tomador de decisiones k-means.

Máquina de Soporte Vectorial (SVM)

Esta técnica son entrenadas por algoritmos que realizan optimización convexa y son idóneas para la clasificación de dos clases: fase de entrenamiento donde selecciona datos de entrenamiento, extrae atributos y entrena al clasificador y la fase de reconocimiento el modelo ya entrenado asigna los nuevos datos de entrada [44].

Perceptrón Multicapa

Este modelo también conocido como modelo de red neuronal artificial, es un modelo que emula a un sistema neuronal en el procesamiento de información [45]. Es una red neuronal unidireccional comprendida por al menos 3 capas con su algoritmo de entrenamiento llamado retropropagación [46].

Árboles de decisión

Este modelo es muy utilizado para enfrentar problemas de clasificación supervisada, el cual consta de nodos internos que tiene uno o más atributos de prueba, aristas que llevan a otros nodos y la hoja se refiere a una etiqueta de clase [47]. A pesar que con grandes conjuntos de datos necesitan mucho tiempo para procesar y memoria para almacenar todo el conjunto de entrenamiento.

Regresión Logística

Es una técnica que proviene netamente de la estadística, un método básico para enfrentar problemas de clasificación utilizados para estimar probabilidad de verificar si una instancia pertenece a una clase particular o no [48]. Considerada como una red neuronal en miniatura, dado que funciona bien cuando se enfrenta a muchos datos y no son muy complejas las interrelaciones.

Clasificador de Naive Bayes

Es un método de tareas de clasificación simples basada en el teorema de Bayes donde funciona según la probabilidad condicional [49]. La probabilidad condicional establece que suceda un evento dada el suceso de otro evento, algo interesante que se puede hacer con este algoritmo es calcular la probabilidad condicional.

Clasificador K-NN

Conocido como algoritmo de k vecinos cercanos, es un clasificador no paramétrico en reconocimiento de patrones el cual utiliza proximidad para realizar predicciones sobre la agrupación de datos. Este algoritmo hace referencia en que las propiedades de una variable X de entrada son parecidos a los datos de su vecindad [50]. Este algoritmo además permite trabajar con datos mezclados y se conoce que la mayoría de estos algoritmos se pueden utilizar para descripciones numéricas [51].

Clasificador de bosques aleatorios

Este algoritmo puede emplearse tanto para clasificación como para la regresión, combina predicciones de diferentes árboles. Este método utiliza una variación *boosting* llamada *boosting aggregation* o *bagging.* Donde primero se crean los árboles aquellos que son construidos con muestras aleatorias de datos y recoge las modas de clases predichas como pronósticos [52]. Los resultados de los árboles suelen ser profundos, de múltiples niveles y sobre ajustan los datos.

1.3 Antecedentes Contextuales

La EPAAA se desarrolla en torno a la prestación de los servicios de agua potable, alcantarillado, depuración de aguas residuales, barrido, limpieza, recolección, transporte, tratamiento y disposición final de residuos sólidos no peligrosos y peligrosos, sus servicios complementarios, conexos y afines que pudieren ser considerados de interés colectivo, otros servicios que resuelva el directorio, así como la gestión de sectores estratégicos, el aprovechamiento sustentable de recursos naturales o de bienes públicos y en general al desarrollo de actividades económicas conexas a su actividad que correspondan al Estado, los mismos que se prestarán en base a los principios de obligatoriedad, generalidad, uniformidad, eficiencia, universalidad, accesibilidad, regularidad, calidad, responsabilidad, continuidad, seguridad y precios equitativos.

La empresa pública de agua, alcantarillado y aseo orientará su acción con criterios de eficiencia, racionalidad y rentabilidad social, preservando el ambiente, promoviendo el desarrollo sustentable, integral y descentralizado de las actividades económicas de acuerdo con la Constitución.

En la actualidad la empresa pública de agua, alcantarillado y aseo maneja su cartera mediante llamadas telefónicas y mediante notificaciones a los abonados con deudas, donde el único identificador es que el abonado debe cumplir con un mes de atraso en sus pagos, para que esta sea clasificada como cartera vencida para la empresa y de esa manera generalizada se realiza los dos ejercicios mencionados anteriormente por parte de cartera para poder recuperar fondos.

La empresa actualmente maneja 4 rubros en su facturación como son: agua, alcantarillado, cargos fijos agua, cargos fijos alcantarillado y actualmente cuenta con 18670 abonados. Años anteriores existía un rubro adicional llamado servicios administrativos que se eliminó en una nueva ordenanza en el 2019. Además, mediante una entrevista su pudo conocer sobre la necesidad de mejorar la gestión de cartera y la gestión de recaudación.

Tabla 1Entrevista del estado actual de la EPAAA

Entrevista		**1**	**2**	**3**	**4**	**5**
p1	¿La información del sistema le permite conocer lo recaudado en el mes de una sola emisión?	12	0	0	0	0
p2	¿Puede ejecutar reportes inmediatos de un ciclo en cuanto a su cartera y gestión de recaudación de una emisión?	8	4	0	0	0
p3	¿EL sistema actual le permite obtener valores recaudados clasificándolos por categorías?	12	0	0	0	0
p4	¿Los tiempos para poder obtener la información antes mencionada son eficientes?	6	4	2	0	0
p5	¿La información actual le permite generar acciones inmediatas?	3	4	4	1	0
p6	¿EL sistema actual le permite conocer el comportamiento de un abonado en particular en cuanto a su gestión de pago?	4	4	4	0	0
p8	¿Es necesario crear más herramientas que faciliten y optimicen su trabajo para una mejor gestión de recaudación?	0	0	0	4	8
p9	¿El sistema actual es eficiente en cuanto a la gestión de cartera vencida?	9	3	0	0	0
p10	¿Es necesario mejorar la gestión de cartera y la gestión de recaudación de la EPAAA?	12	0	0	0	0
TOTAL		66	19	10	5	8
%		55	15,83	8,33	4,17	6,67

Basada en esta entrevista podemos justificar la problemática de la EPAAA sobre la necesidad a resolver problemas que puedan ayudar a mejorar la eficiencia de la EPAAA como se muestra en la **ilustración 3**.

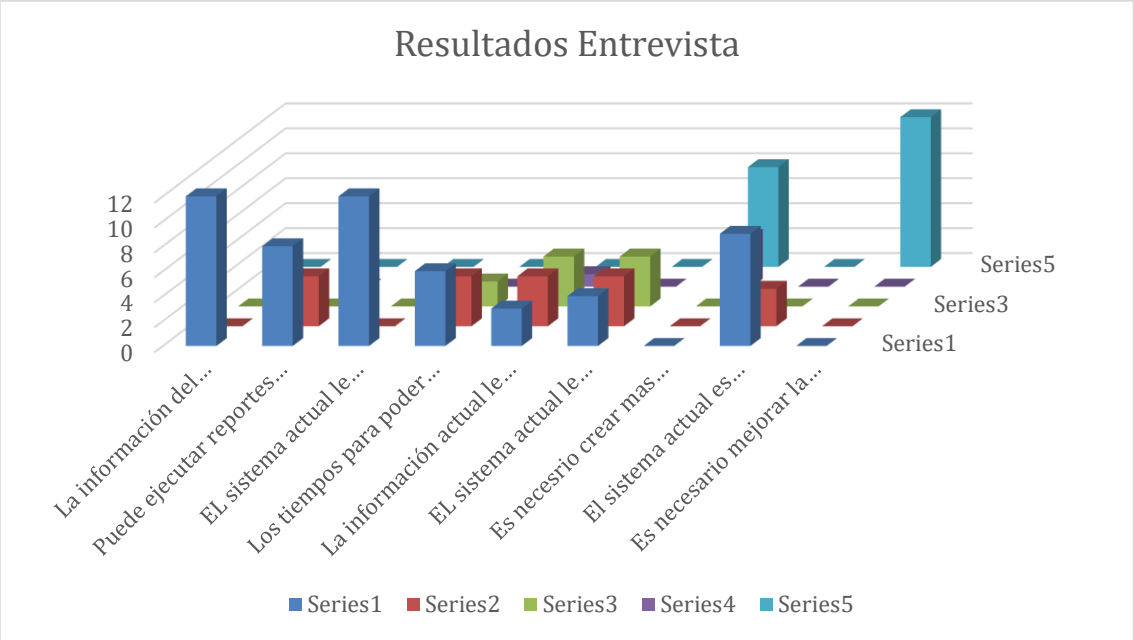


Ilustración 3 Resultados de la entrevista

Los entrevistados son los altos mandos y mandos medios de la empresa que intervienen en la gestión de recaudación y la gestión de cartera, donde el 45% de los entrevistados están nada satisfechos de como manejan actualmente sus labores, el 15.83% se encuentran poco satisfechos y el 8.33% están medianamente satisfecho. El 16.67% manifiestan que requieren mejoras y creación de herramientas que ayuden a la gestión de recaudación y gestión de cartera.

1.3.1 Propuesta de solución y contribuciones

Se desarrollará un DSS en la EPAAA de Pasaje, mediante el análisis de cartera empleando algoritmos de aprendizaje automático para una gestión eficiente de recaudaciones, de esta manera se podrá brindar un apoyo para los altos mandos en la toma de decisiones. Esta estrategia contribuirá para los intereses de la empresa con respecto a un mejor manejo de la cartera y también al interés colectivo ya que se podrá evidenciar falencias y a su vez responder de manera más eficiente ante sucesos que se podrán identificar.

Permitirá automatizar procesos, mejorar en la toma de decisiones y optimizar las operaciones, donde se podrá simplificar aquellas tareas repetitivas lo que permitirá

ahorrar recursos y tiempo. Se podrá analizar grandes cantidades de información y de la misma manera extraer patrones que no son simples a la vista del colaborador de la empresa, además podrá mejorar los costos de producción y en otros casos evaluar costos de producción.

CAPÍTULO II. MATERIALES Y MÉTODOS

2.1 Paradigma, diseño y tipo de investigación

Analizando la información de la gestión de cartera y de recaudación de la EPAAA, se busca medir la correlación de los indicadores que conforman la gestión de recaudación y de cartera como el total meses adeudados, sus ciclos o parroquias, la categoría a la que pertenece el abonado, el abonado, el sector donde está ubicada la cuenta, la cantidad de meses pagados; con la deuda total, total emitido, total recaudado. Es necesario medir el tipo de correlación que existe entre estos indicadores antes mencionados. [53] [27].

El **paradigma** [54] que se determinó para el presente trabajo de investigación, basado en el libro de metodología de investigación según R. Hernández Sampieri [28], se lo define como Cuantitativo – Cuasi Experimental como se muestra en la **Ilustración 4**. Además, esta investigación está basada en la revisión sistemática de la literatura y en los objetivos que se requieren alcanzar, se toma en consideración el cumplimiento de cada objetivo específico para llegar al objetivo general de la tesis, con el fin de garantizar el desarrollo de la propuesta.

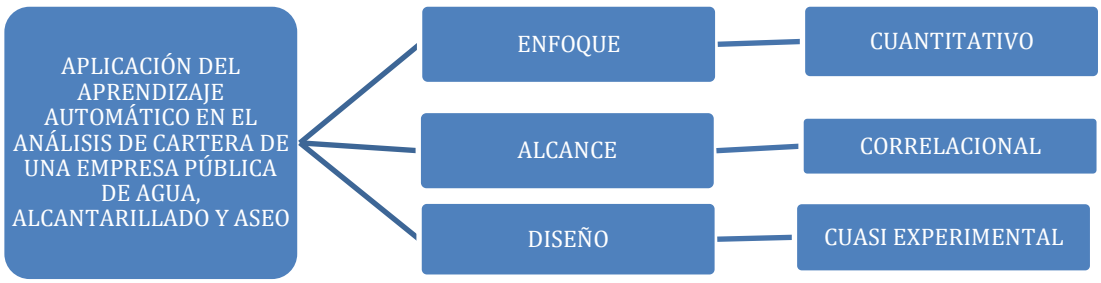


Ilustración 4 Tipo de Estudio

Con el **enfoque** cuantitativo como se muestra en la **ilustración 5**, vamos a utilizar la recopilación de todos datos para evidenciar una hipótesis, basada en la medición numérica y también en el análisis estadístico, con el fin de identificar patrones de comportamiento y certificar teorías.

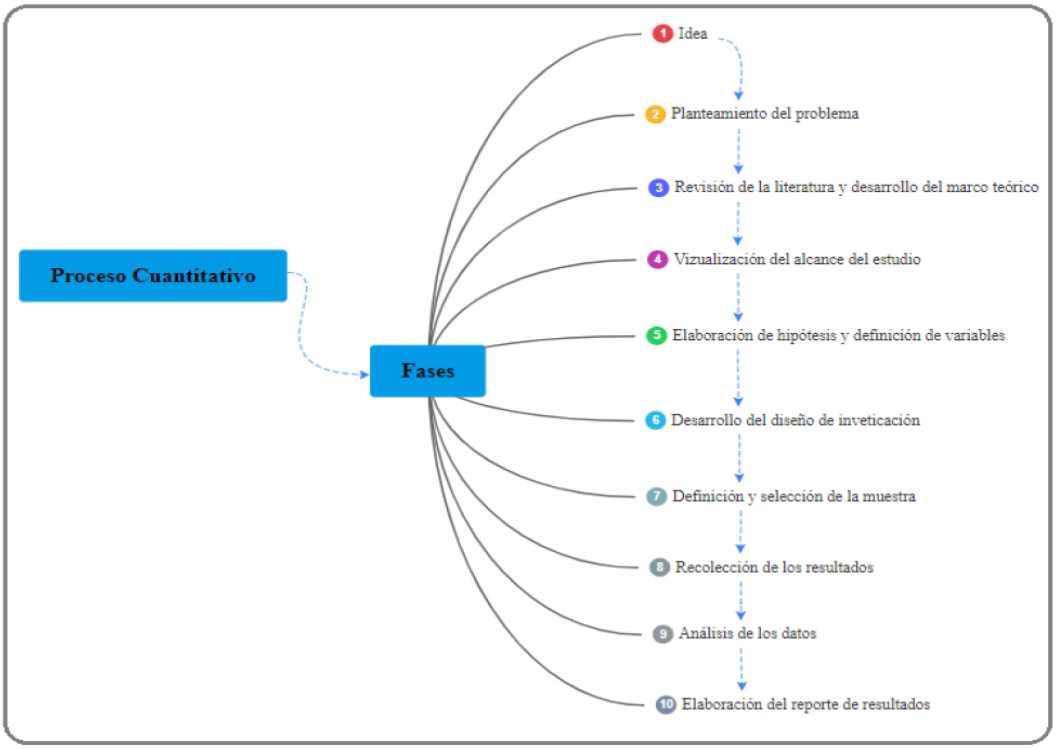


Ilustración 5 Proceso cuantitativo

En el **diseño** cuasi experimental no puede realizar asignación aleatoria de sujetos (variable independiente) a grupos o condiciones, porque los grupos ya existen [28]. Se manipula a propósito, por lo menos una variable independiente para de esta forma observar su efecto sobre una o más variables dependientes como se indica en la **ilustración 6**, es decir que se identifica la relación de la variable independiente sobre la variable dependiente [55].

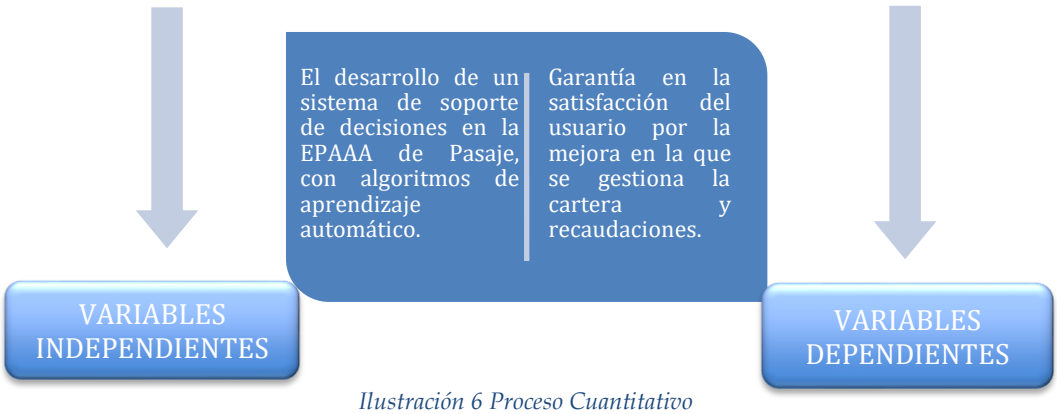


Ilustración 6 Proceso Cuantitativo

2.2 Calculo de la población y muestra

Para esta investigación la población se subdivide en dos unidades de investigación, por un lado, corresponden al número de clientes activos a los cuales se les emiten cartas de pago al mes, por esta razón se determina que contamos con una población finita de 18670 clientes activos. En este sentido este proyecto se ha desarrollado con la información relacionada a la generación de cartera por el cobro de los servicios básicos que ofrece la EPAAA de Pasaje correspondiente al periodo comprendido entre el enero 2002 y febrero 2023 con un total de 4704840 registros. Además, la segunda unidad de investigación es la población de 12 colaboradores que toman decisiones en la EPAAA para mejorar la gestión de cartera y recaudación, estos se encuentran en las áreas de tesorería, cartera, comercial, financiero y gerencia; a los cuales se les aplicó una entrevista.

2.3 Operacionalización de Variables

Variable Independiente: El desarrollo de un sistema de soporte de decisiones en la EPAAA de Pasaje, con algoritmos de aprendizaje automático.

Tabla 2 Operacionalización de variable independiente

CONCEPTUALIZACIÓN	**MÉTRICAS**
Es el software por implementar para la ayuda en la toma de decisiones de la EPAAA, con la ayuda de técnicas de análisis de datos y técnicas predictivas.	Análisis de técnicas de datos. Análisis de técnicas predictivas.

Variable Dependiente: Garantía en la satisfacción del usuario por la mejora en la que se gestiona la cartera y recaudaciones

Tabla 3 Operacionalización de variable dependiente

CONCEPTUALIZACIÓN	MÉTRICAS
Es el desempeño del personal de la EPAAA y las mejoras operacionales en cuanto a la gestión de recaudación que se obtuvo con el DSS.	Satisfacción del usuario. Eficacia.

2.4 Métodos Teóricos

Los métodos teóricos nos brindan la facilidad de conocer más sobre el tema que se está investigando y posterior comprenderlo mejor. La investigación teórica son el pilar para la existencia de nuevas ideas [56].

2.4.1 Método analítico - sintético

Este método se caracteriza por el análisis de la investigación donde segmenta aquello que se estudia sobre lo que se puede evidenciar [57]. La inclusión de aprendizaje automático en la EPAAA analizará el comportamiento de los clientes con respecto a la toma de decisiones para la cartera y la gestión de recaudaciones. En este caso vamos a interactuar mucho con las categorías debido a que los clientes vamos a segmentarlos por categorías.

De la misma manera tenemos los ciclos de como esta segmentado la facturación, los ciclos son parroquias o sectores definidos para una emisión de facturación mensual en cuanto al servicio de agua, alcantarillado que provee la empresa.

2.4.2 Metodología CRISP-DM

La metodología *CRISP-DM* es una metodología idónea para este caso de estudio ya que tiene gran aceptación cuando se trata de minería de datos, esta metodología *CRISP-DM (Cross Industry Standard Process for Data Mining)* [58] es promovida principalmente por IBM. Posee una combinación de las buenas prácticas que consta de 6 pasos como se muestra en la **ilustración 7**, generando una implementación exitosa en iniciativas de ciencia de datos [29].

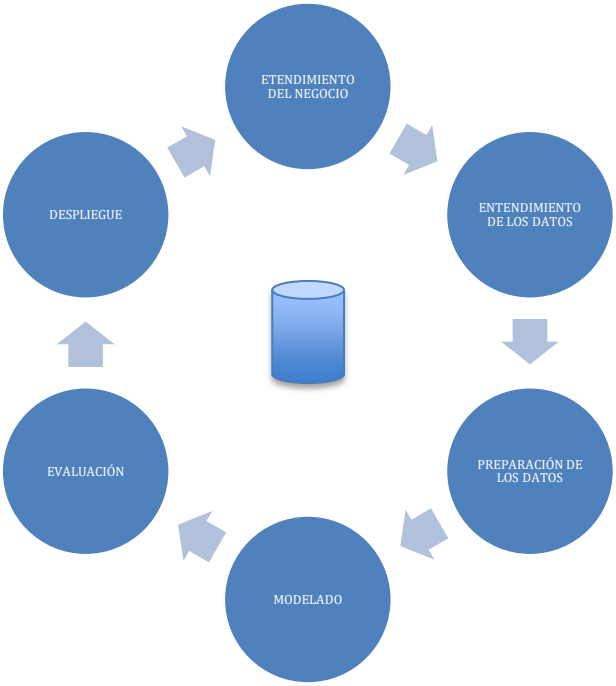


Ilustración 7 Metodología CRISP-DM

2.4.2.1 Comprensión del negocio

Esta fase en la EPAAA es muy importante porque si no se tiene una correcta comprensión del problema o el giro de negocio para la construcción del sistema de soporte de decisiones de nada servirán las etapas que la presiden.

2.4.2.2 Comprensión de datos

La segunda fase de esta metodología debemos obtener los datos con los que cuenta la EPAAA, es decir para este trabajo se utilizará la totalidad de la información de la empresa. Una vez teniendo los datos debemos identificar, describir y explorar los datos.

2.4.2.3 Preparación de los datos

La construcción del *dataset* se realiza mediante la selección de los datos que se van a transformar en este proyecto, realizamos la limpieza de los datos, creación de indicadores y la transformación de los datos.

2.4.2.4 Modelado

En esta fase de modelado, se define las características que tendrá el *dashboard* aplicando inteligencias de negocios, minería de datos, además se seleccionarán las técnicas de algoritmos de aprendizaje automático el cual se implementarán, posterior también se seleccionan los datos de prueba para así poder obtener el modelo.

2.4.2.5 Evaluación

Esta fase se desarrollará la calidad que van a tener los modelos a ejecutarse basadas en el análisis de las métricas que arrojarán cada una de ellas.

2.4.2.6 Despliegue

Se refiere a la implementación del modelo, también en la documentación que se realiza para la ejecución del mismo.

2.5 Métodos Empíricos

Los métodos empíricos hacen referencia a la experiencia como en la observación y experimentación o manipulación física de aquello que se quiere conocer [59]. En este sentido se pudo analizar mediante la experiencia [60] en la EPAAA la tardía respuesta en cuanto a los análisis de cartera por la dificultad en el manejo del gran volumen de datos y la poca información de datos procesados para la toma de decisiones.

En el desarrollo se apoyó también con entrevistas a los colaboradores de la EPAAA y encuestas de satisfacción del DSS, con el propósito de realizar la recopilación y análisis de requerimientos, necesidades urgentes para mejorar sus labores, elaboración de preguntas referentes al giro de negocio e implantar indicadores para la evaluación.

Una vez entendiendo las necesidades de áreas que intervienen en el análisis de la cartera y la gestión de recaudaciones, se experimentó [56] con los datos, donde se desarrolló una base de datos inédita con información limpia mediante un ETL, de la misma manera la experimentación para la selección de las variables a utilizar como fuente de información del DSS. Además, se experimentó con parámetros de algoritmos de aprendizaje; esto gracias a la población de datos utilizada que son todos los datos de la EPAAA que corresponden al número de clientes activos.

2.6 Técnicas Estadísticas

Para el presente análisis de datos estadísticos se aplicó técnicas de estadística descriptiva como gráficos estadísticos de barra, inteligencia de negocios con gráficos estadísticos mediante un *dashboard*, minería de datos con análisis de correlación, regresión lineal con representación gráfica y con deducciones estadísticas a partir de modelos de aprendizaje automático que arrojaron resultados de las métricas de los algoritmos implementados.

CAPÍTULO III. RESULTADOS DEL PROCESO DE DESARROLLO DEL DSS PARA LA EPAAA DE PASAJE

En este capítulo se procedió a realizar la ejecución de la metodología *CRISP-DM* que consta de 6 fases, mediante la aplicación de *Python* haciendo uso de una librería muy potente de aprendizaje automático de código abierto scikit-learn, para así generar métricas de medición [23], donde se ponen en manifiesto el desarrollo de sus 6 fases. Para poder lograr este proyecto se necesitó generar una base de datos limpia en postgresql mediante un proceso ETL, con equipos idóneos cuya característica es un procesador RYZEN 7 series 6000 con RAM DDR5 de 16GB y tarjeta gráfica de 4GB.

3.1 Comprensión del negocio

La EPAAA necesita apoyar las decisiones de los altos mandos y mandos técnicos ejecutando un eficiente análisis de cartera para la gestión de las recaudaciones (Capítulo I). Existe mucha información, pero poco conocimiento en cómo manejarla. Por ejemplo, se puede conocer cuánto se recaudó en el mes, pero no se puede segmentar si todo lo recaudado de ese mes pertenece al ciclo de facturación en curso. En un análisis de cartera vencida anual solo obtenemos valores de los años cursados y no podemos conocer del estado de la cartera vencida actual.

Tabla 4 Encuesta al Personal Administrativo Financiero

AREA	No	PREGUNTA
Personal Administrativo Financiero de la Empresa Pública de Agua, Alcantarillado y Aseo - EPAAA	**P1**	¿Puede segmentar las cartas de pago para conocer rubros puntuales como agua y alcantarillado?
	P2	¿Puede conocer lo recaudado de una emisión de facturación puntual?
	P3	¿Puede obtener un histórico de lo recaudado y deudas de una emisión de facturación puntual?
	P4	¿Puede conocer lo recaudado por un ciclo o parroquia en particular de una emisión de facturación puntual?
	P5	¿Puede conocer lo recaudado de una categoría especifica de una emisión de facturación en particular?
	P6	¿Puede conocer el comportamiento de una cuenta sobre lo recaudado, deudas y otros datos propios de la cuenta de agua?
	P7	¿Puede comparar anualmente la recaudación y la deuda que tiene la EPAAA
	P8	¿Está conforme con la información que obtiene del sistema transaccional de la EPAAA?

Se identifica necesidades del personal Administrativo Financiero y de altos manos del estado actual en cuanto al manejo y el aprovechamiento de la información mediante una encuesta como se muestra en la **Tabla 1**, además del conocimiento que esta información le permite obtener.

Para esto se propone desarrollar un sistema de toma de decisiones aplicando mediante el análisis de cartera empleando algoritmos de aprendizaje automático para una gestión eficiente de recaudaciones.

3.2 Comprensión de datos

La EPAAA tiene su base de datos transaccional con datos reales como lo muestra la **tabla 1** e **ilustración 8**, donde entre sus tablas importantes para la obtención de nuestros resultados están:

Tabla 5 Tablas de base de datos de EPAAA

Tablas	Detalles
abonado	Tabla donde almacena el número de cuenta, el id del cliente, id de la categoría, etc.
ruta	Tabla donde almacena las rutas de la cuenta
sector agua	Tabla que almacena el sector de la cuenta
ciclo	Ciclo o Parroquia a la que pertenece la cuenta de agua
cliente	Detalle de la información del propietario de la cuenta
Tipo de categoría	Categoría a la que pertenece la cuenta
Agua liquidación	Información de las lecturas de agua potable
Agua emisión	Tabla donde almacena los códigos y fechas de las emisiones de facturación
factura	Tabla donde almacena todas las facturas de la cuenta de agua pagadas o impagas.
factura detalle	Contiene los rubros a cobrar y valores de la carta de pago
rubro	Contiene el detalle al que se le adjuntará un valor, ejemplo agua potable, alcantarillado.
agua emisión ruta	Contiene la ruta, sector de la emisión que se factura

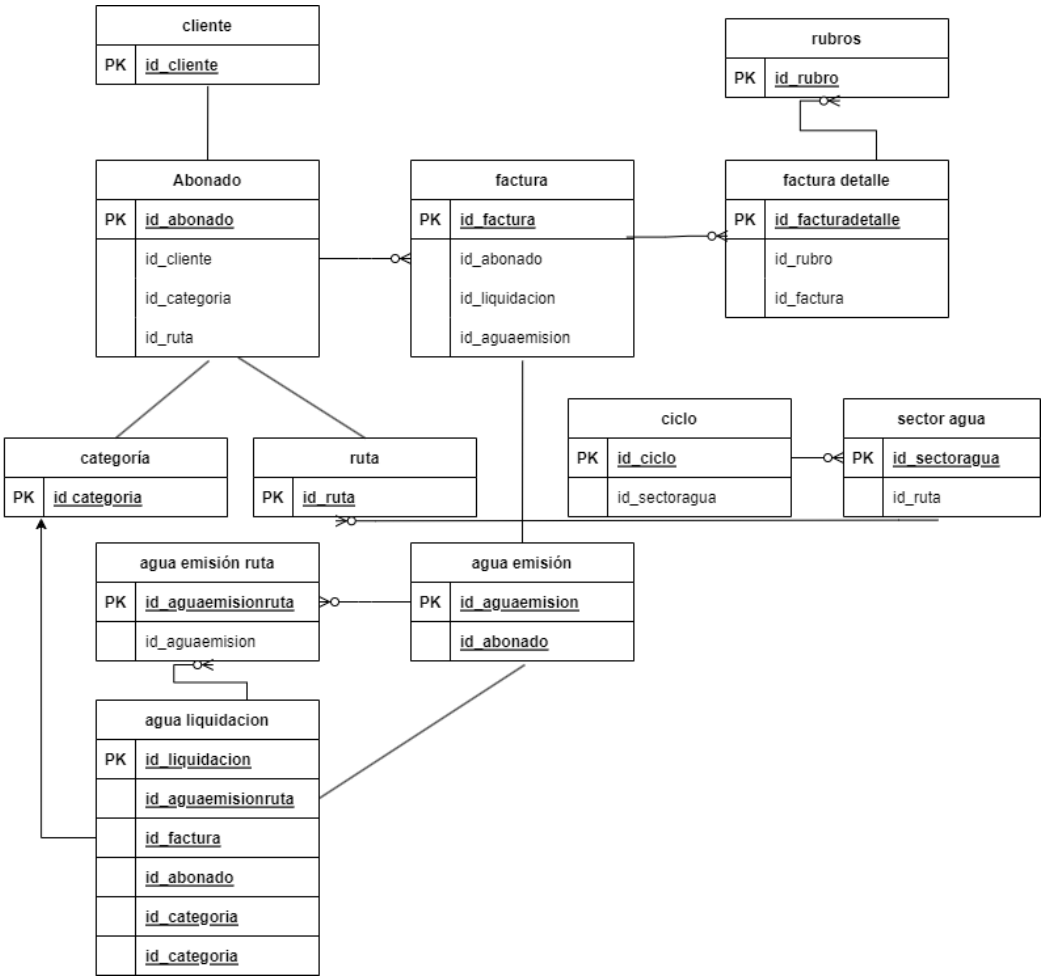


Ilustración 8 Entidades de base de datos transaccional de EPAAA

La agrupación y selección de información de estas tablas tiene como resultado un *datamart* que contiene la estructura idónea para el desarrollo del DSS. Debido a la gran cantidad de información que poseen era necesario crear un ETL, donde en primer lugar basado en la experiencia se realizó una selección de datos mediante un query como se identifica en la **tabla 5**, cabe resaltar que no se puede publicar el query aplicado en el ETL para la extracción, por proteger ese tipo de información por confidencialidad de la EPAAA. así se pudo extraer la información que requerimos para así generar un solo registro y poder implementar un DSS.

Con la obtención de este datamart se puede enfrentar de mejor manera la información duplicada en cuanto a los clientes, ya que por ejemplo se obtiene un abonado activo con un id de cliente que contiene valores de deuda, además que parte de los errores solventados con el datamart es identificar de mejor manera a un usuario con valores por n cantidad de veces, donde solo ese usuario acumula gran cantidad de deuda. A este

usuario se lo clasifico y de igual manera se lo representa en el DSS para calcular la deuda solo de este usuario que es alta.

Debido a que la base de datos de la EPAAA es una base transaccional y por el gran volumen de información que existe en cuanto a la información de cartera de la empresa. Se implementó un proceso ETL como se indica en la **ilustración 9**. Esta información se procesó en *Pentaho Data Integration* [61], el cual mediante un conjunto de pasos se pudo brindar soluciones a las problemáticas planteadas. En primera instancia como todo proceso ETL fue la extracción de los datos de la base de datos transaccional que se mencionó anteriormente, donde mediante un query se extrae información de las tablas antes mencionadas en la **tabla 5**, donde para extraer esa información se estableció conexiones a la base de datos ingresando credenciales parra las conexiones ODBC/JDBC de la base de datos de la EPAAA, posterior se procedió a su transformación de los datos que necesitaba extraer. Es decir, se transforma los datos en relación a lo que se necesita obtener para la carga en una base de datos limpia.

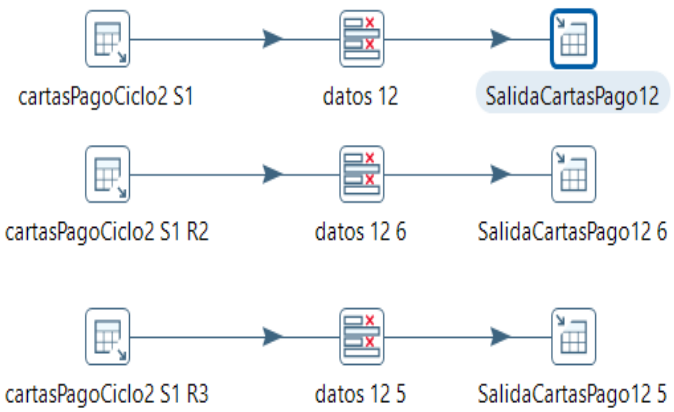


Ilustración 9 Pentaho ETL

La extracción de los datos provocó afectaciones minusculas en los sistemas transaccionales, los datos se extrajeron de sus diferentes fuentes donde posterior fueron cargadas, posterior son preparados los datos para su transformación de los nuevos datos fuentes. Los datos con registros problemas se ha garantizado la facil localización para la carga de los datos limpios como se presenta en la **ilustración 10** y esta información es reflejada en un *Dashboard*. Además permitió que la carga de los datos se realicen en una

sola tabla que contiene un solo registro detallado de la cartera de cada abonado.

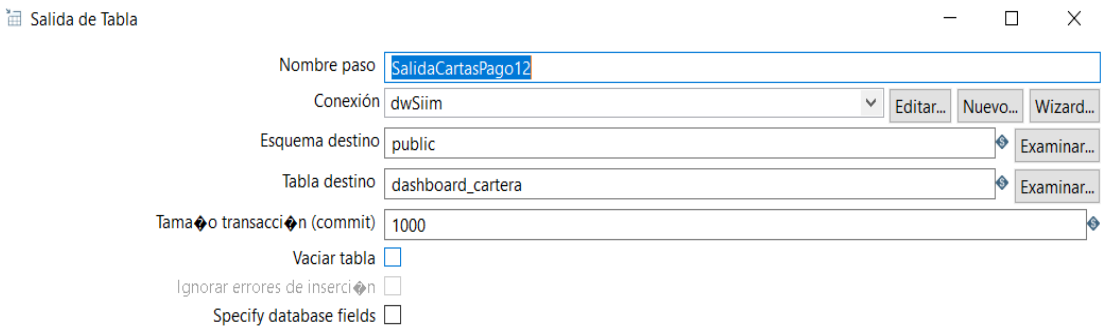


Ilustración 10 Traspaso de datos limpios

Para la creación del *dataset* de los datos de prueba y entrenamiento se utiliza la base limpia donde se maneja los siguientes campos que muestra la **tabla 6** y la vista creada a partir de la base de datos limpia **tabla 7**, para el *dashboard* y para entrenamiento del algoritmo.

Tabla 6 Tabla de datos 1

CarteraAgua		
Columna	**Definicion**	**Detalle**
id	Integer	Identificador de la tabla
modulo	bigint	Identificador de las cartas de pago de agua y alcantarillado
idciclo	bigint	Identificador del ciclo o parroquia
idsector	bigint	Identificador del sector perteneciente al ciclo
idruta	bigint	Identificador de la ruta perteneciente al sector
secuencia	bigint	numeración de ubicación de la cuenta del abonado
ordenruta	bigint	identificación de ubicación de la cuenta del abonado
emision	character varying(20)	Número de la emisión facturada
pagado	Integer	Identificador de deudor o pagado
clavepredial	text	Es la clave catastral
propietario	text	Nombre del propietario
categoria	character varying(20)	Tipo de categoría del predio
id_categoria	integer	identificador del tipo de categoría del predio
cuenta	bigint	Cuenta del abonado
abonadocliente	bigint	Identificador del propietario
direccionubicacion	text	Dirección del propietario
ruta	text	Ruta de la cuenta
sector	text	Sector al que pertenece la cuenta
ciclo	text	Ciclo de donde pertenece la cuenta
rutasecuencia	text	Identificador de donde se ubica la cuenta
lecturaanterior	numeric	Consumo anterior de agua

Tabla 6 tabla de datos 1 (continuación)

CarteraAgua		
Columna	**Definicion**	**Detalle**
lecturaactual	numeric	Consumo actual de agua
subruta	text	Identificador de donde se ubica la cuenta
idfactura	bigint	Código de facturación de la cuenta
fechacreacion	timestamp	Fecha de la facturación de la emisión
anio	numeric	Año de la facturación de la emisión
mes	numeric	Mes de la facturación de la emisión
fechacobro	timestamp	Fecha de cobro de la facturación de la carta de pago
aniocobro	numeric	Año de cobro de la facturación de la carta de pago
mescobro	numeric	Mes de cobro de la facturación de la carta de pago
consumoagua	numeric	Valor de consumo de agua
alcantarillado	numeric	Valor de consumo de alcantarillado
multas	numeric	Valor de multas
desc3edad	numeric	Descuento por tercera edad
servadm	numeric	Valor de servicios administrativos
descdiscapacidad	numeric	Descuento por discapacidad
interes	numeric	Valor de interés
convenios	numeric	Convenios realizados por deudas
otrosrubros1	numeric	cargos fijos de agua y alcantarillado
ultimaemision	character varying(20)	ultima emisión generada

Esta tabla contiene un resumen de la cartera de la empresa y se puede visualizar todas las emisiones de cartas de pago que ha realizado la empresa. De ahí podemos clasificar teniendo en cuenta que se obtiene sectores de los usuarios y tipo de categoría que maneja cada cliente. A partir de esta tabla ya resumida se pudo generar una vista donde se obtienen de manera resumida las deudas, rubros totales de la historia de cartera.

Tabla 7 Vista de datos

Rubro Total	
Columna	**Detalle**
Cuenta	Contiene el identificador único de la cuenta que contiene la deuda
Mesesdeuda	Es la cantidad de meses que tiene deuda esa cuenta
Totalagua	Cantidad de deuda total de deuda de agua
Totalalcantarillado	cantidad total de deuda de alcantarillado
interes	cantidad total de interés
total_tarifafija	Cantidad total de deuda de tarifa fija
total_servadm	Cantidad total de deuda de servicios administrativos
total_terceraedad	Cantidad total de deuda de tercera edad
total_discpacidad	Cantidad total de deuda de discapacidad
totaldeuda	Es la cantidad total de deuda de la cuenta
ciclo	Nombre de la parroquia a la que pertenece la cuenta
idciclo	identificador de la parroquia a la que pertenece la cuenta
pagado	0 significa impago, 1 significa pagado
propietario	nombre del propietario de la cuenta
categoria	tipo de categoría con que se clasifica el tipo de vivienda o comercio
id_categoria	identificador del tipo de categoria
idsector	Identificador del sector donde se ubica el usuario
Idruta	Identificador del sector donde se ubica el usuario

3.3 Preparación de los datos

La construcción del *dataset* se realizó con toda la información limpia, se trasladó a archivos *csv* como se expone en la **ilustración 11**, para poder ser leídos por el sistema de manera más eficiente y poder generar los datos de manera más rápida en el *dashboard* [62].

Ilustración 11 dataset

Previamente cargamos en *csv* toda la información para su posterior carga y desarrollar los *dashboard* y el aprendizaje de datos. Posterior preparamos los datos que vamos a extraer para su visualización general en el *dashboard* [63], como mostramos en la **ilustración 12** e **ilustración 13**.

```
# # Carga de datos
#
# Para este caso, se cargarán los libros de 'carteraAgua.csv'


carteraAgua = pd.read_csv("data/bi_1_CA.csv")
rubrosTotales = pd.read_csv("data/bi_1_RT.csv")
```

Ilustración 12 Carga de datos

```
# ## Preparando los datos para el dashboard


# Extraemos los ciclos
ciclos_grouped = carteraAgua.groupby(["ciclo"])
ciclos_list = [key for key, _ in ciclos_grouped]

ciclos_pair = ciclos_grouped[["ciclo","idciclo"]].max()

# extraemos las categorías
categoria_grouped = carteraAgua.groupby(["categoria"])
categoria_list = [key for key, _ in categoria_grouped]

# extraemos los años de emisión
anios_emision_grouped = carteraAgua.groupby(["anio_emision"])
anios_emision_list = [key for key, _ in anios_emision_grouped]

# extraemos los meses de emisión
mes_emision_grouped = carteraAgua.groupby(["mes_emision"])
mes_emision_list = [key for key, _ in mes_emision_grouped]



ciclos_ids =  ciclos_grouped["idciclo"].max().reset_index()


# Vuelve a los datasets interactivos
carteraAgua_interactive = carteraAgua.interactive()
```

Ilustración 13 Preparación de datos

Para poder realizar la preparación de datos, en *clustering* se tomó en cuenta ciertos criterios para su procesamiento, en este caso "totaldeuda, mesesdeuda, id_ciclo, id_categoria". A diferencia del algoritmo de regresión lineal ahí se tomó en cuenta los meses de deuda *versus* el total de la deuda de una categoría y ciclo respectivamente.

Los datos de entrenamiento nos permiten graficar tendencias y métricas en cuanto a los algoritmos que se procesan, además tener una visualización más idónea de cómo se comporta la cartera de la EPAAA.

3.4 Modelado

En esta fase de modelado, se definió las características que tendrá el *dashboard* aplicando inteligencia de negocios, minería de datos, además los algoritmos de aprendizaje automático el cual se implementaron en esta sección utilizando el *dataset* y los datos para el entrenamiento del algoritmo.

Los algoritmos Utilizados son:

- Correlación
- Regresión Lineal
- Clustering
    - K-Means
- Clasificación
    - SVM
    - Perceptrón Multicapa
    - Regresión Logística
    - Arboles de decisión
    - Clasificador de Naive Bayes
    - Clasificador K-NN
    - Clasificador de bosques aleatorios

Como lo habíamos indicado en el entrenamiento de datos, se cargan los *datasets* previamente, para entrenamiento y predicción de los mismos.

El **propósito** de la aplicación de estas técnicas se debe a la necesidad de determinar la cantidad de deuda que tienen los abonados con respecto al tiempo de cancelación de la deuda, clasificar abonados que tengas problemas de deudas incobrables para poder

identificar esta cartera vencida y no mezclarla con la cartera vencida real. Se llama deuda incobrable porque esta deuda no tiene asignación a algún deudor en particular y también clasificar deudas homogéneas. Es decir, hay repetidos valores de deudas cargados a una sola persona.

La ayuda de estos algoritmos ayudará a identificar el comportamiento de un abonado, debido a que no existen cortes de agua potable dando paso a que el abonado pueda atrasarse en los pagos de agua potable ya sea de 3, 6 o 9 meses o por años, con la implementación de aprendizaje automático podremos saber que usuarios son pagadores o deudores. Podremos clasificar que parroquia y que ruta es pagador o deudor y de tal manera identificar si existe anomalías en cuanto al servicio de agua potable. El propósito más importante es que estas herramientas permitan tomar decisiones y así poder optimizar tiempos de respuesta.

Correlación

La correlación lineal se da entre las variables de un conjunto de datos, siendo 1 una correlación proporcional (cuanto más incrementa A, más incrementará B), -1 correlación inversa (cuanto más incrementa A, más decrementa B), y 0 ninguna correlación como se indica en la **ilustración 14 y 15**.

```
# ## Correlación
# Se extraen todas las variables numéricas y se analiza su coorelación
# Extraer los campos numéricos
datos_correlacion = pd.read_csv("data/correlacion.csv")
datos_correlacion= datos_correlacion.drop(columns=["Unnamed: 0"], axis=1)

# Extraemos la matriz de coorrelación
grafico_correlacion_simple = datos_correlacion.corr().round(3)
grafico_correlacion_simple

np.triu(np.ones(grafico_correlacion_simple.shape)).astype(bool)

grafico_correlacion_simple = grafico_correlacion_simple.where
(np.triu(np.ones(grafico_correlacion_simple.shape)).astype(bool))
grafico_correlacion_simple = grafico_correlacion_simple.fillna(0).round(2)
```

Ilustración 14 Correlación

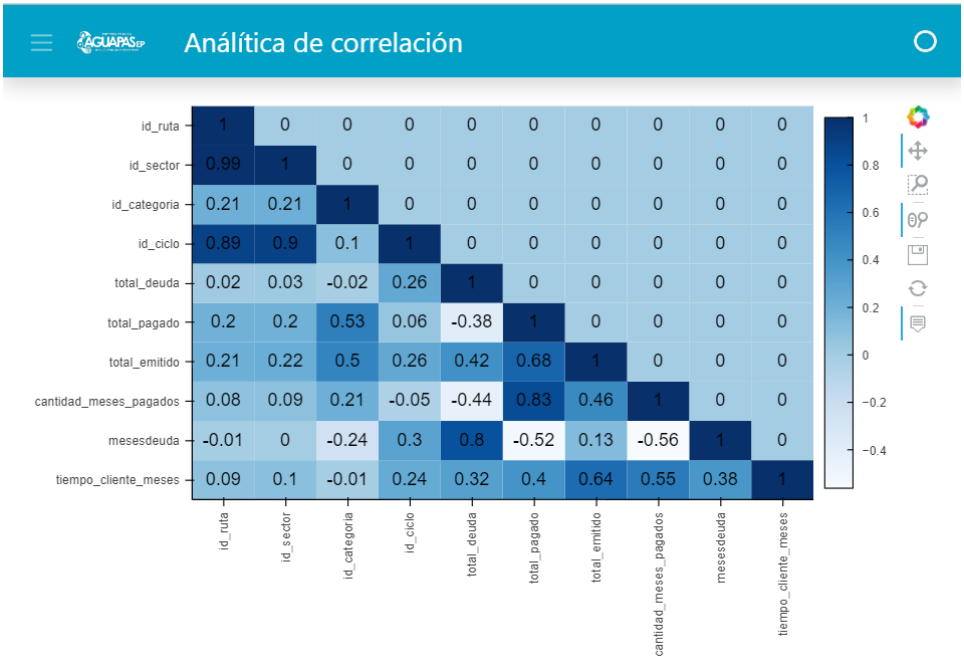


Ilustración 15 Matriz de Correlación

Para el presente análisis, se tomaron las dimensiones numéricas del *dataset*, y mediante este análisis, se pretende encontrar relaciones ocultas entre las distintas variables. El gráfico corresponde a un mapa de calor de la matriz de correlación de Pearson[64] que compara las variables y muestra sus coeficientes de correlación. La diagonal, muestra 1 siempre, pues corresponde a la comparación entre las mismas variables.

Se alimentó la matriz de correlación con datos agrupados por cuenta, es decir, cada fila del *dataset*, corresponde a información de una cuenta.

Entre las relaciones destacadas se resalta:

- Total_pagado x cantidad_meses_pagados: se observa una correlación positiva, pues ambos están estrechamente relacionados
- Total_deuda x meses deuda: se observa una correlación positiva, pues ambos están estrechamente relacionados
- Id_sector, id_ciclo e id_ruta: la forma de explicar esta correlación, es que dado que conforme el transcurso del tiempo, se agregaron paulatinamente nuevos sectores y nuevos ciclos juntos, es decir, a la vez que se crea un ciclo, se deben crear sectores, por tanto, los identificadores aumentarán de forma similar

Con esta información, se puede rescatar, que basado en la cantidad de meses pagos, se puede modelar el crecimiento de la deuda, pues la relación es lineal.

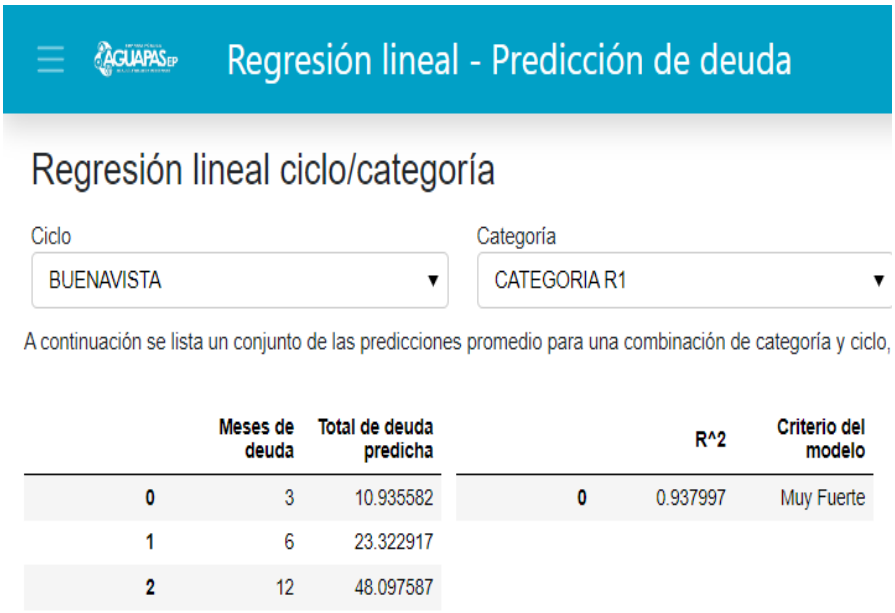
Regresión Lineal

Este modelo se basa en la relación de dependencia entre una variable dependiente Y, y una variable independiente X. Una vez analizada esta correlación nos dio pie para poder realizar la relación sobre meses de deuda en el eje de las X y total deuda en el eje de las Y. La predicción se realiza para 3, 6 y 12 posteriores como se muestra en la **ilustración 16, 17 y 18**.

```
# ## Definir la función del modelo


def bind_build_model_ciclo_cat(categoria_widget,ciclo_widget):

    rubros_filtrados = rubrosTotal[
            (rubrosTotal["categoria"] == categoria_widget)
            & (rubrosTotal["Ciclo"] == ciclo_widget)
        ]

    rubros_filtrados = rubros_filtrados.rename(columns={"mesesdeuda":"Meses de deuda"})

    x_data = rubros_filtrados[["Meses de deuda"]]
    y_data = rubros_filtrados["totaldeuda"]

    predicciones = pd.DataFrame({"Meses de deuda": [3, 6, 12]})

    modelo = LinearRegression().fit(x_data, y_data)
    # hacer predicciones para 3, 6 y 12 meses posteriores
    predicciones["Total de deuda predicha"] = modelo.predict(predicciones)
    # r2 = modelo.score(x_data, y_data)

    return predicciones
```

Ilustración 16 Regresión Lineal

Para esto, se procesaron los datos de las cuentas hasta la emisión anterior de la facturación, es decir, cada fila del *dataset*, es una cuenta que se analizará.



	Meses de deuda	Total de deuda predicha
0	3	10.935582
1	6	23.322917
2	12	48.097587

	R^2	Criterio del modelo
0	0.937997	Muy Fuerte

Ilustración 17 Resultado Criterio R2

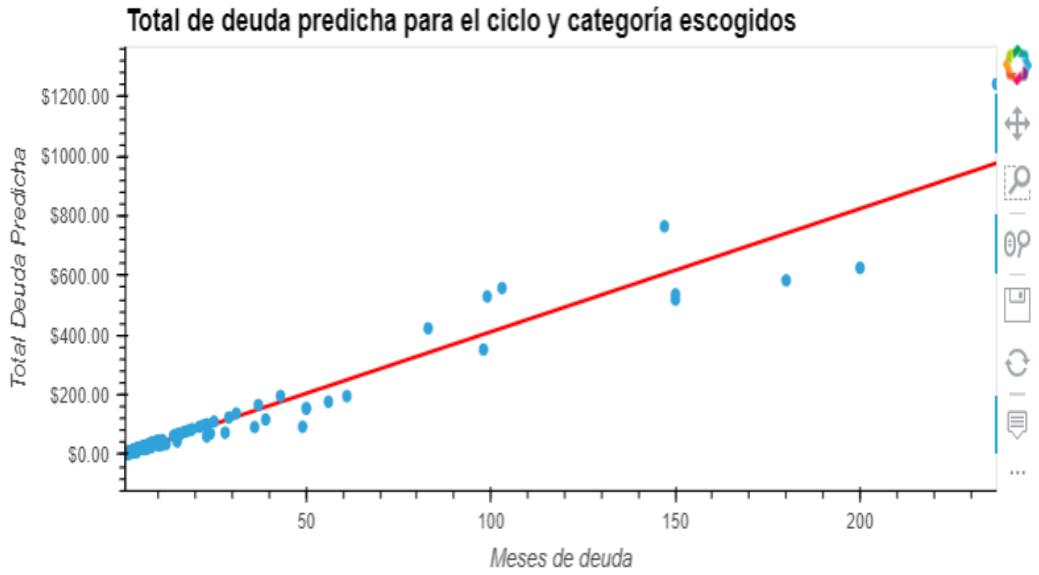


Ilustración 18 Regresión Lineal

Con esto se obtuvo un modelo regresión lineal simple, que se entrena basado en los filtros de la parte superior, es decir (Ciclo y categoría), es decir, se alimenta al modelo, con los datos filtrados por ciclo y categoría, por lo que el modelo será distinto dependiendo de la selección.

Clustering

El aprendizaje automático estudia y aplica modelos de aprendizaje en diferentes manifestaciones, en este caso aplicaremos K-means [65]. Este método es uno de los más conocidos en cuando al aprendizaje automático, en este algoritmo se ha utilizado el método del codo donde nos indicó que la cantidad ideal para el algoritmo es el uso de 3 clusteres, como se puede identificar en la **ilustración 19**.

```
from sklearn.cluster import KMeans
# Aplicar k-means para determinar valor de k
inertias = []
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k)
    kmeans.fit(df_norm)
    inertias.append(kmeans.inertia_)

codo_data = pd.DataFrame({"Inercia":inertias})

codo_metodo = codo_data.hvplot(
    title="Gráfica del codo"
)
```

Ilustración 19 Clustering

```
# Binding de la tabla de resumen
def bind_cluster_sumary(nro):
    clusters = KMeans(n_clusters=nro)
    clusters.fit(df_norm)
    rubrosTotal["clusters"] = clusters.labels_
    # determinamos la cantidad de clientes en cada cluster
    conteo_elementos = rubrosTotal.groupby("clusters").count()["cuenta"]
    # Extraemos los centroides de cada clusters
    conteo_elementos =conteo_elementos.reset_index().rename
    (columns={"cuenta":"Cant. cuentas"})[["clusters","Cant. cuentas"]]
    # Agrupamos los clusters
    clusters_agrupados = rubrosTotal.groupby("clusters")
    # Calculamos la deuda promedio de cada cluster
    deuda_promedio_cluster = clusters_agrupados
    ["total_deuda"].mean().reset_index().sort_values("clusters")
    deuda_promedio_cluster = deuda_promedio_cluster.rename(columns={"total_deuda":"Deuda Promedio"}, )
```

Ilustración 20 Cálculos de deuda promedio

Para el desarrollo se utilizaron las columnas de la vista (total_deuda, meses_deuda, id_ciclo, id_categoría) como se aprecia en la **ilustración 20**. Cabe resaltar que se incrustó en la aplicación, un input numérico, para tener la capacidad de escoger la cantidad de *clusters* a generar, desde 1 hasta 4 que sería la cantidad de dimensiones de los datos.

Dado que son 4 dimensiones de entrada, no es posible hacer una gráfica que represente todas ellas a la vez, por lo que se optó por múltiples gráficos de dos dimensiones para comprender de alguna manera la morfología de los resultados.

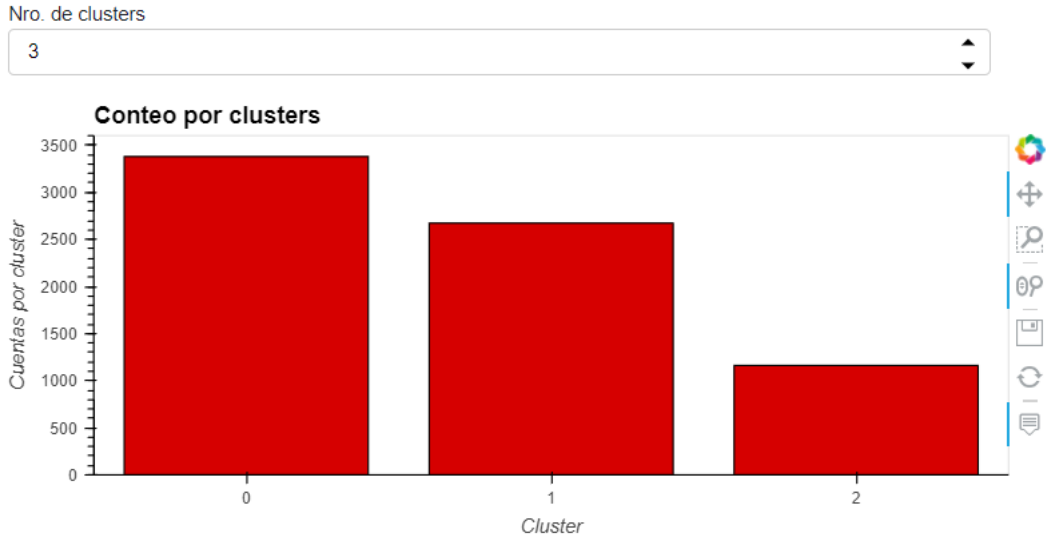


Ilustración 21 Clusters

El gráfico superior, corresponde a un diagrama de barras que representa la cantidad de cuentas en cada *cluster*, siendo el de menos cuentas el 2, y el que tiene más cuentas el 0.

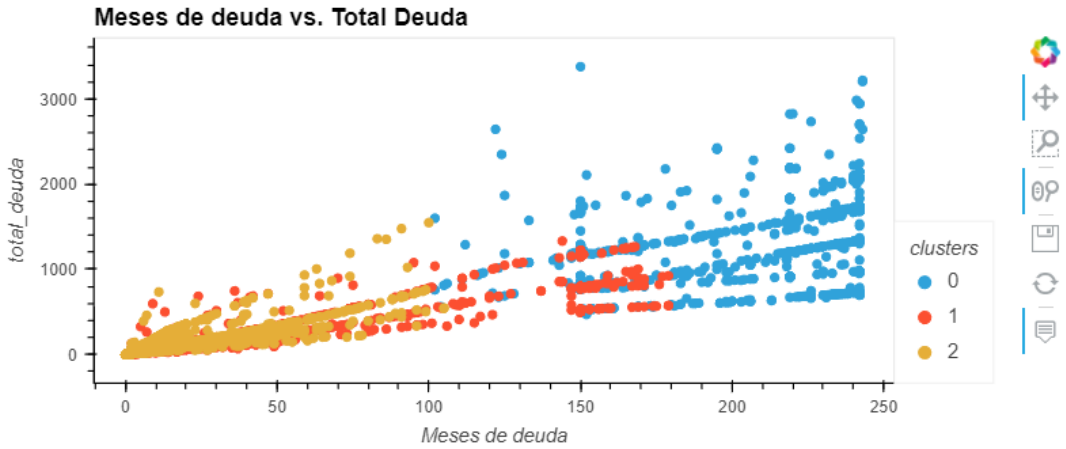


Ilustración 22 Clustering

En la presente **ilustración 22**, se pueden observar varias características, la primera de ellas, es que existe una tendencia a la linealidad en el crecimiento de la deuda, de modo que a más mese de deuda, mayor su incremento. Así mismo, se resalta que existen velocidades de incremento distintas, posiblemente condicionadas por la categoría de los clientes, siendo este un determinante para las tasas de interés.

Respecto al algoritmo, se observa que al haberse configurado 3 clusters, generó 3 grupos de datos, el primero que tomó como dato de relevancia los meses de deuda, agrupando las cuentas con más deuda, seguido del 1 que acumuló mayormente los meses promedio y el 0 con las cuentas que tienen menos meses de duda. Es importante mencionar que, al existir otras dimensiones dentro de los datos de entrada, los grupos no están perfectamente delimitados, pues existen otros criterios a considerarse, sin embargo, se determina que los meses de deuda es uno de los más relevantes.

Basado en la gráfica anterior, también se puede remarcar, que el cluster 0, que cuenta con la mayor deuda, es el que menos cuentas tiene, lo que nos indica, que la mayor cantidad de deudores, se encuentra en un grupo reducido de cuentas, y que el cluster 2, que corresponde mayormente a cuentas con deudas de entre 70 a 170 meses, son el grupo mayoritario, lo que indica que se debe prestar especial atención a ese grupo de clientes, dado que su deuda puede significar gran parte de la cartera pendiente de cobro, y su cantidad de meses de deuda, es alta.

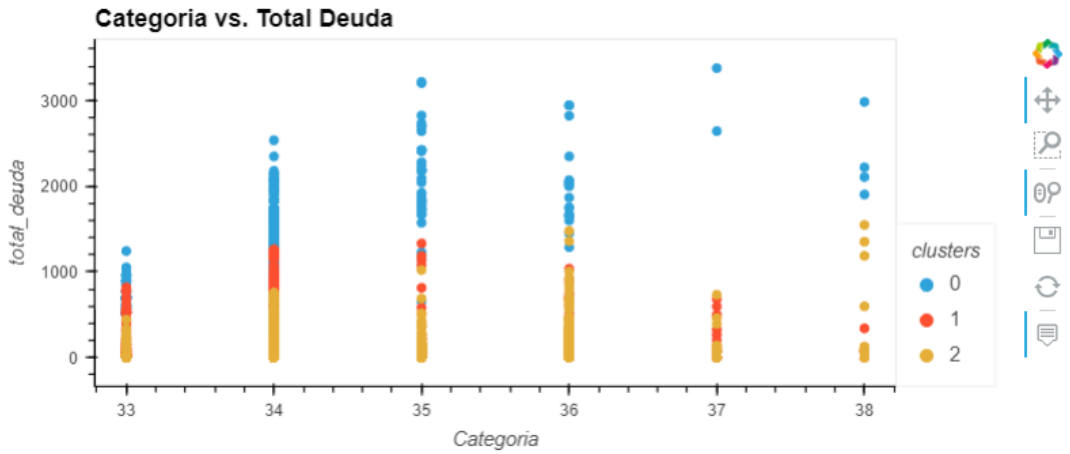


Ilustración 23 Clustering K-means

En la ilustración 23, se seccionaron las categorías de cliente *versus* la cantidad de deuda. Se puede resaltar, que en todas las categorías excepto la 33, existe parte del clister 0, correspondiente a las cuentas con deuda más antiguas, con especial atención en las categorías 34 y 35, y con valores atípicos en las categorías 37 y 38.

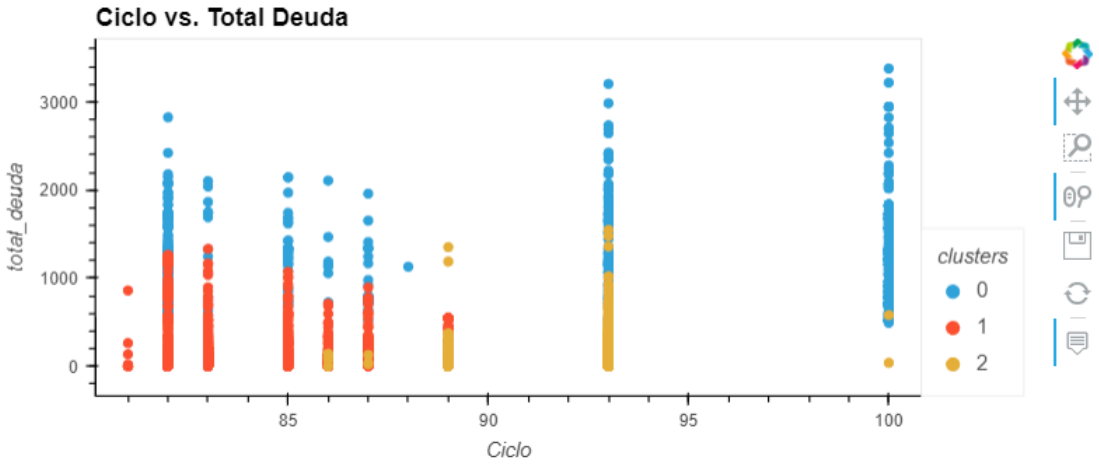


Ilustración 24 Clustering K-means

En la **ilustración 24**, se compararon el ciclo vs el total de deuda, en lo que se puede remarcar, que buena parte de las cuentas con mayor cantidad de meses de deuda, se concentran en el ciclo 100, debido a un error de migración, en este ciclo se acumulan cuentas con solamente deuda, sin embargo, siguen perteneciendo a la cartera vencida, por lo que se debe considerar.

Clasificación

En aprendizaje automático la clasificación proporcionó modelos fundamentales para resolver problemas de clasificación, en este sentido vamos a emplear los siguientes modelos de clasificación:

- SVM
- Perceptrón Multicapa
- Regresión Logística
- Arboles de decisión
- Clasificador de Naive Bayes
- Clasificador K-NN
- Clasificador de bosques aleatorios

```
models = {
    'MLPClassifier': mlp,
    'LogisticRegression': lr,
    'SVM': svm,
    'DecisionTreeClassifier': dtc,
    'GaussianNB': nb,
    'KNeighborsClassifier': knn,
    'RandomForestClassifier': clf
}
```

Ilustración 25 Modelos

Al implementar estos modelos de la **ilustración 25**, lo que hacemos es ingresar los *dataset* como se muestra en la **ilustración 26** y posterior dividimos en dos conjuntos, los datos entrenamiento y los datos de prueba.

```
# Cargamos los datos del Dataset
rubrosTotal = pd.read_csv("data/redes_neuronales.csv")
df = rubrosTotal[['cuenta', 'tiempo_cliente_meses', 'mesesdeuda',
       'cantidad_meses_pagados', 'total_emitido', 'total_pagado',
       'total_deuda', 'id_ciclo', 'id_categoria', 'pagado']]
df.groupby("pagado").count()

# ## Redes Neuronales

# Importamos las librerías necesarias
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, recall_score, precision_score
from sklearn.model_selection import train_test_split


# Eliminamos cualquier dato faltante (NaN)
df = df.dropna()

# Dividimos el dataframe en dos conjuntos: un conjunto de entrenamiento y un conjunto de prueba
X = df.drop(['pagado', 'cuenta'], axis=1)
y = df['pagado']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)
```

Ilustración 26 dataset Clasificación

Una red neuronal se construye gracias a un conjunto de nodos interconectados y se organizan en capas, es un clasificador binario, a todos los clasificadores se les envió el mismo *dataframe* en los ejes de entrenamiento para su comparación.

Estos modelos se entrenan previamente con datos agrupados por cuenta, es decir, se contabilizan meses de deuda, meses pagados, total de deuda, etc. Estos modelos, se entrenan con el mismo *dataset* de entrenamiento, que corresponde al 60% de los datos, y testeados con el 40% de los datos. Se calculan métricas para conocer la efectividad y eficiencia de los métodos utilizando los datos de prueba, y se muestran en la parte inferior. Así mismo, se muestra la predicción individual al seleccionar un cliente, que determinará si ese cliente se considera como deudor frecuente o un pagador frecuente.

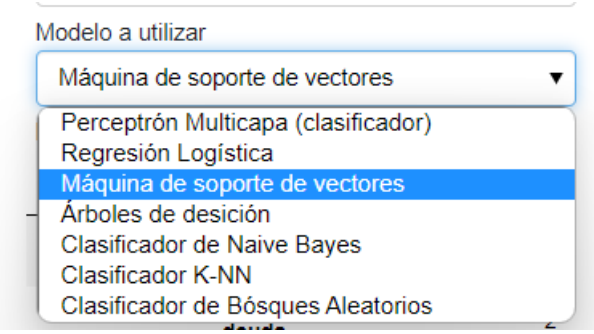


Ilustración 27 Métodos de Clasificación

Se resalta en la **ilustración 27**, el que los datos fueron filtrados hasta la emisión anterior de febrero 2023, debido a que cuando se genera una nueva emisión, inicia como impaga, y existirá inevitablemente un número grande de cuentas impagas.

Cabe resaltar que el modelo con métricas menos sesgadas es la máquina de soporte de vectores. El resto de modelos, se presume que pueden tener sesgo de *overfiting* (sobreajuste), pero se justifica su resultado por la naturaleza de los datos y el comportamiento de los clientes ya que existe gran cantidad de deuda mayormente en los últimos meses y también es posible por la gran cantidad de información que estos procesan aproximadamente 4704840 registros.

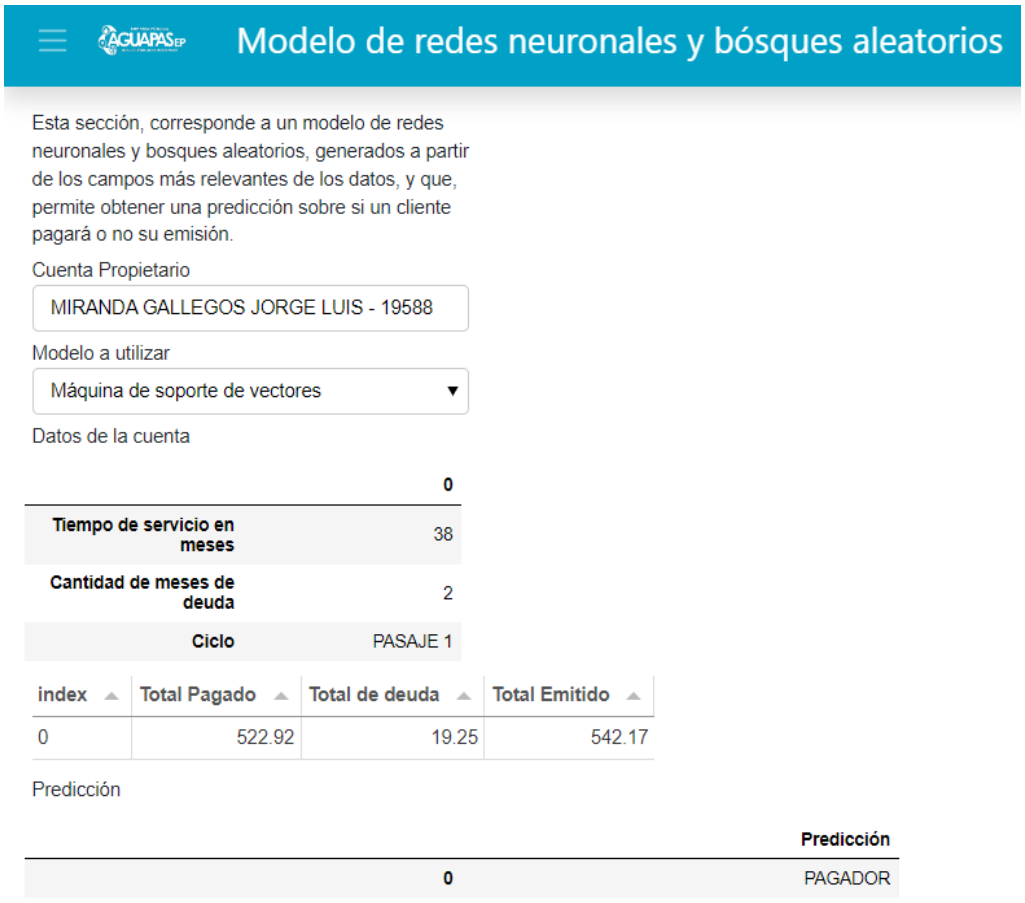


Ilustración 28 Clasificación SVM

En la **ilustración 28**, se escoge una cuenta asociada a un cliente y el tipo de modelo que deseamos utilizar para la predicción y en la **ilustración 29**, la predicción a mostrar.

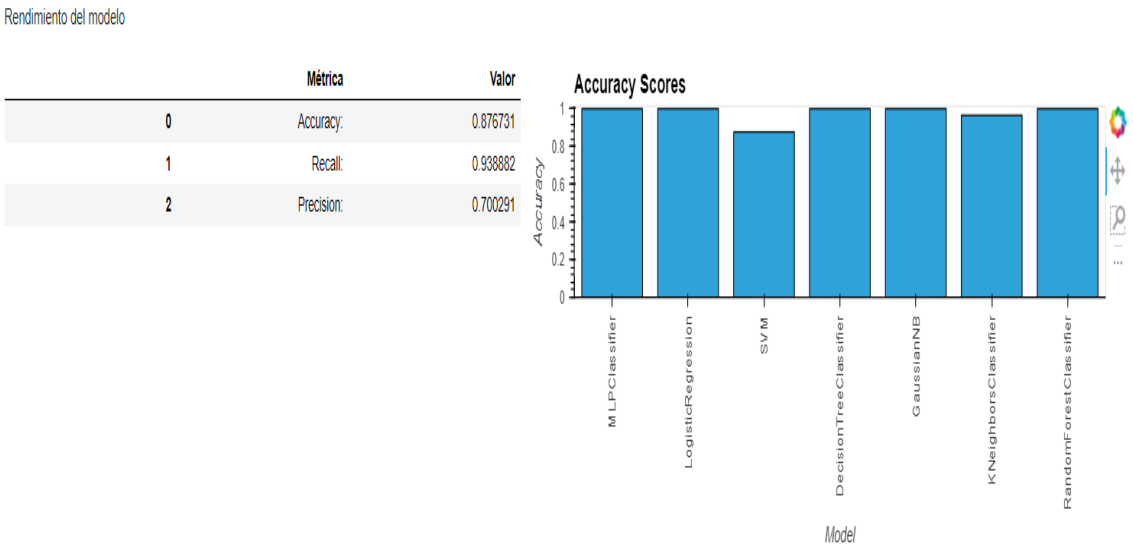


Ilustración 29 Métricas de SVM

Las métricas del modelo seleccionado *Acurracy*, *Recall* y *Precision*, en este caso podemos manifestar que todos los modelos son aplicables para el presente proyecto como se detallan en la **ilustración 30, 31, 32, 33, 34 y 35**.

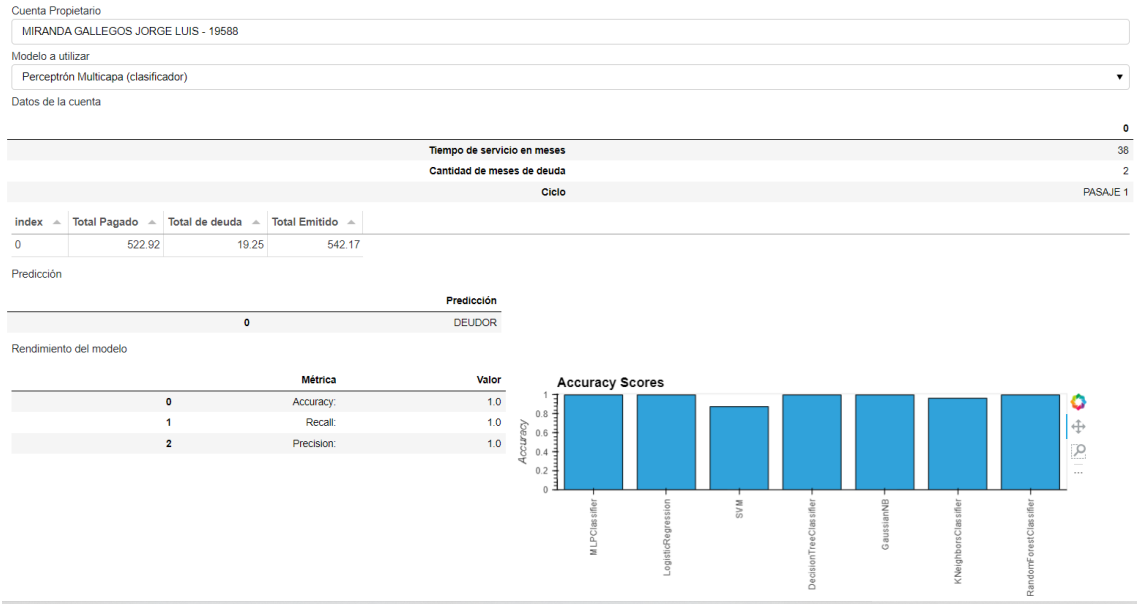


Ilustración 30 Perceptrón Multicapa

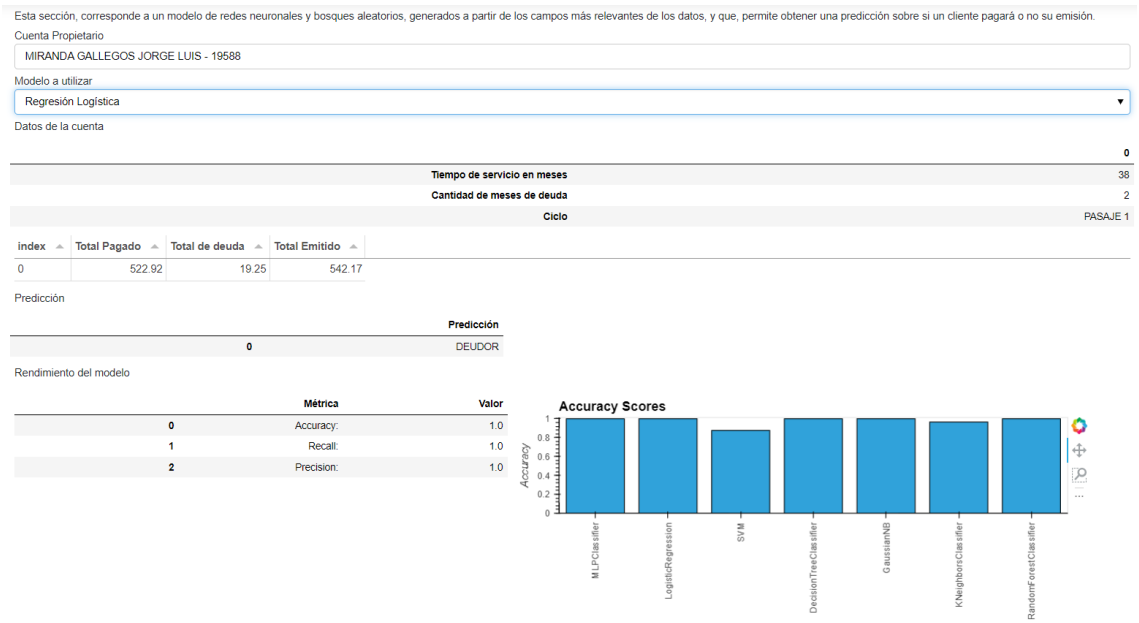


Ilustración 31 Regresión Logística

Esta sección, corresponde a un modelo de redes neuronales y bosques aleatorios, generados a partir de los campos más relevantes de los datos, y que, permite obtener una predicción sobre si un cliente pagará o no su emisión.

Cuenta Propietario

MIRANDA GALLEGOS JORGE LUIS - 19588

Modelo a utilizar

Árboles de desición

Datos de la cuenta

	0
Tiempo de servicio en meses	38
Cantidad de meses de deuda	2
Ciclo	PASAJE 1

index	Total Pagado	Total de deuda	Total Emitido
0	522.92	19.25	542.17

Predicción

	Predicción
0	DEUDOR

Rendimiento del modelo

	Métrica	Valor
0	Accuracy:	1.0
1	Recall:	1.0
2	Precision:	1.0

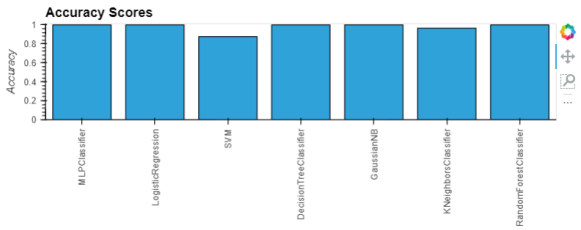


Ilustración 32 Árboles de decisión

Esta sección, corresponde a un modelo de redes neuronales y bosques aleatorios, generados a partir de los campos más relevantes de los datos, y que, permite obtener una predicción sobre si un cliente pagará o no su emisión.

Cuenta Propietario

MIRANDA GALLEGOS JORGE LUIS - 19588

Modelo a utilizar

Clasificador de Naive Bayes

Datos de la cuenta

	0
Tiempo de servicio en meses	38
Cantidad de meses de deuda	2
Ciclo	PASAJE 1

index	Total Pagado	Total de deuda	Total Emitido
0	522.92	19.25	542.17

Predicción

	Predicción
0	DEUDOR

Rendimiento del modelo

	Métrica	Valor
0	Accuracy:	1.0
1	Recall:	1.0
2	Precision:	1.0

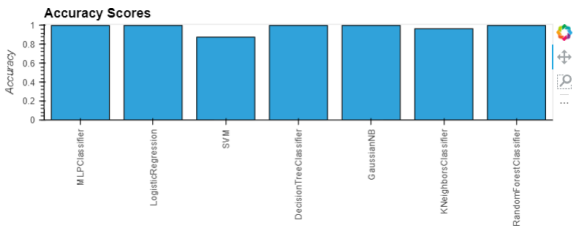


Ilustración 33 Clasificador de Naive Bayes

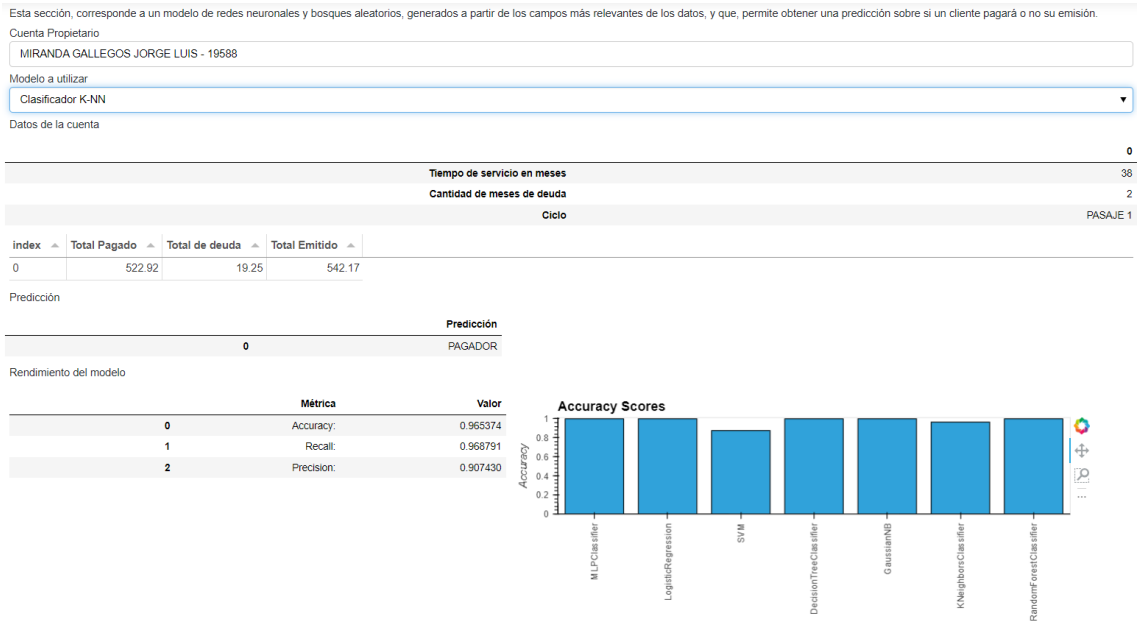


Ilustración 34 Clasificador K-NN

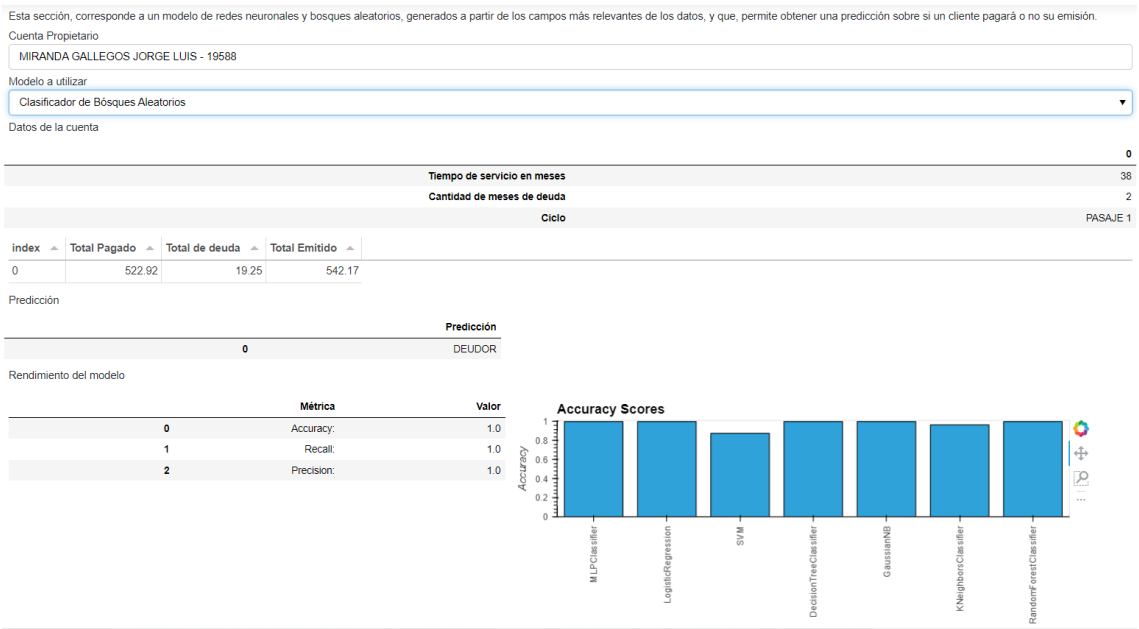


Ilustración 35 Clasificador Bosques Aleatorios

La posible explicación de este fenómeno, quizás se deba a la forma de los datos. Dado que, en el caso de la empresa de agua potable, al no contar con micro-medición y dado que los cortes de servicio no se efectúan de forma instantánea, es bastante habitual que existan cuentas con emisiones impagas acumuladas, de modo que, en nuestros datos refleja un buen porcentaje de cuentas deudoras, además que la información que se pone

a consideración es desde el año 2022, lo que de alguna manera puede sesgar los modelos y generar fenómenos como el *overfitting*.

3.5 Evaluación

Esta fase se presentan los resultados obtenidos de los algoritmos planteados previamente, según sus respectivas métricas; estos son coeficiente de correlación, regresión lineal, clustering y clasificación, con los cuales se apoyarán las decisiones para el análisis de la cartera y la gestión de recaudación en la EPAAA. Para ello la evaluación de los modelos se los explica con más detenimiento en el Capítulo IV.

3.6 Despliegue

El despliegue del proyecto se lo realiza en una *intranet*, ya que por temas recursos de la EPAAA no se lo puede desplegar en servidor de producción en la web y poder acceder desde cualquier punto, para su despliegue en *intranet* y uso procederemos a realizar los siguientes pasos:

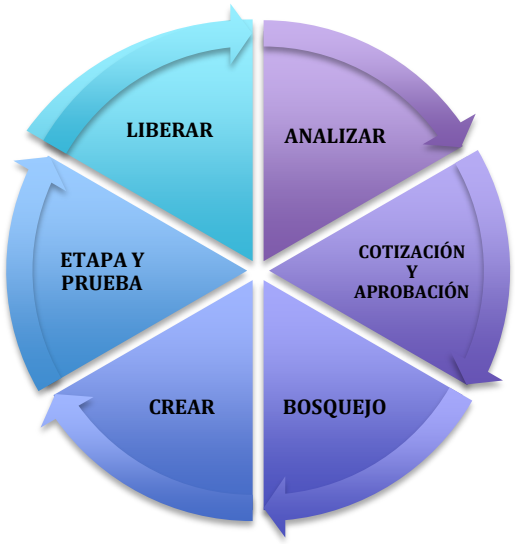


Ilustración 36 Despliegue del Proyecto

CAPÍTULO IV. DISCUSION DE RESULTADOS

En el presente capitulo se describe el proceso de evaluación realizado al DSS implementado en la EPAAA junto con la evaluación de modelos de aprendizaje automático. Se aplicó una encuesta para la evaluación del DSS a los colaboradores de la EPAAA. Para evaluar los modelos de aprendizaje automático se utilizaron las métricas de exactitud.

4.1 Análisis de Resultados

Con el fin de obtener resultados sobre el manejo del análisis de cartera y la gestión de recaudación, se ha evaluado la usabilidad del DSS implementado en la EPAAA. Para lo cual se realizó una encuesta a 12 colaboradores que intervienen en temas financieros y estratégicos, que implican áreas como tesorería, cartera, financiero y área comercial. La medición de las preguntas se las realizará mediante la escala Likert como lo indica en la siguiente **tabla 8**.

Tabla 8 Escala de encuesta

ESCALA	
1	No Satisfecho
2	Poco Satisfecho
3	Moderadamente satisfecho
4	Muy Satisfecho
5	Extramadamente Satisfecho

Esta encuesta al tabular los datos que se muestra en la **tabla 9**, evalúa la efectividad del DSS en cuanto a la utilidad, usabilidad y satisfacción de los colaboradores en cuanto al análisis de la cartera y la gestión de recaudación. Además, evalúa los tiempos de respuesta de las áreas en la EPAAA.

Tabla 9 Encuesta de Resultados

Encuesta	1	2	3	4	5
El sistema permite obtener la información de manera rápida y eficaz.	0	0	0	0	12
El Sistema permite realizar un análisis de cartera.	0	0	0	0	12
Mejoró la gestión de recaudación que ejecuta su área.	0	0	0	2	10
El desempeño del sistema mejoró el desempeño de sus labores.	0	0	0	1	11
La implementación del sistema incremento la eficiencia en la toma de decisiones.	0	0	0	1	11
Fue positivo los resultados obtenidos con la implementación del sistema en la empresa	0	0	0	0	12
TOTAL	0	0	0	4	68
%	0	0	0	5.56	94.44

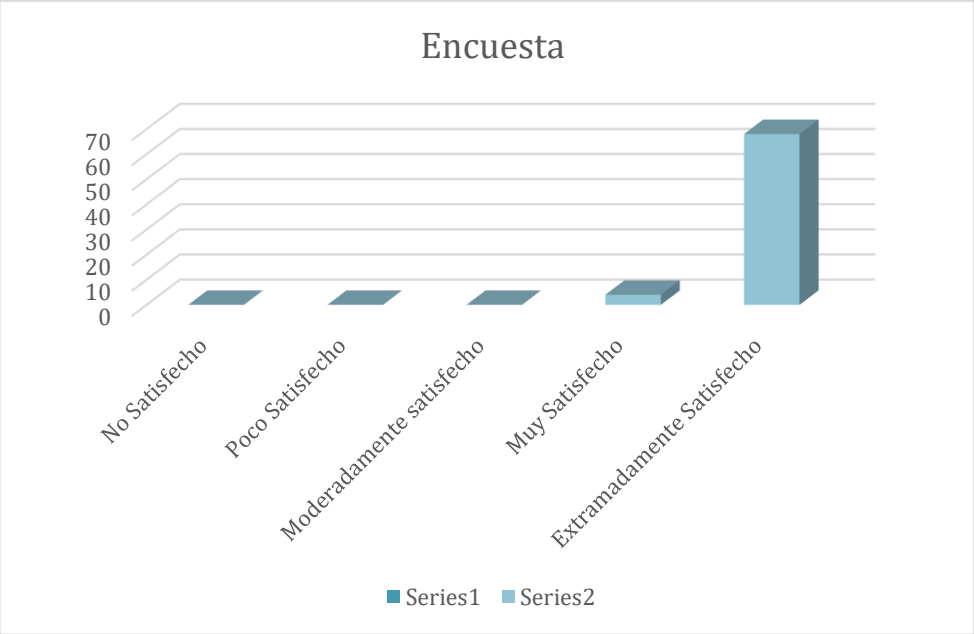


Ilustración 37 Gráfico de Resultados

Análisis

De acuerdo con los resultados de la **ilustración 37** se puede establecer que el 94.44% del personal que interviene en trabajos en cuanto a la toma de decisiones para el análisis de cartera y gestión de recaudación esta extremadamente satisfecho con desarrollo del DSS y los resultados obtenidos, el 5.55% están muy satisfechos ya que quieren experimentar más con el sistema por el nivel de actualización del sistema con respecto a los datos.

4.2 Interpretación de Resultados

Análisis de la Inteligencia de Negocios para el análisis de cartera y gestión de recaudación en la EPAAA

Parte de la solución aplicando inteligencia de negocios mediante un *dashboard*, apoya en gran medida el análisis de datos de cartera debido a que muestra resultados puntuales y de manera sencilla. La información se visualiza de manera general como indica la **ilustración 38**, donde podemos visualizar el año de emisión de facturación y el mes, debido a que las facturaciones de servicios básicos son de carácter mensual, de esta manera podemos visualizar netamente la cantidad de lo recaudado por la EPAAA en una emisión, de la misma manera se segmenta los valores de la carta de pago tanto como para agua potable y como alcantarillado, el top de los deudores de la empresa según la **ilustración 39**.

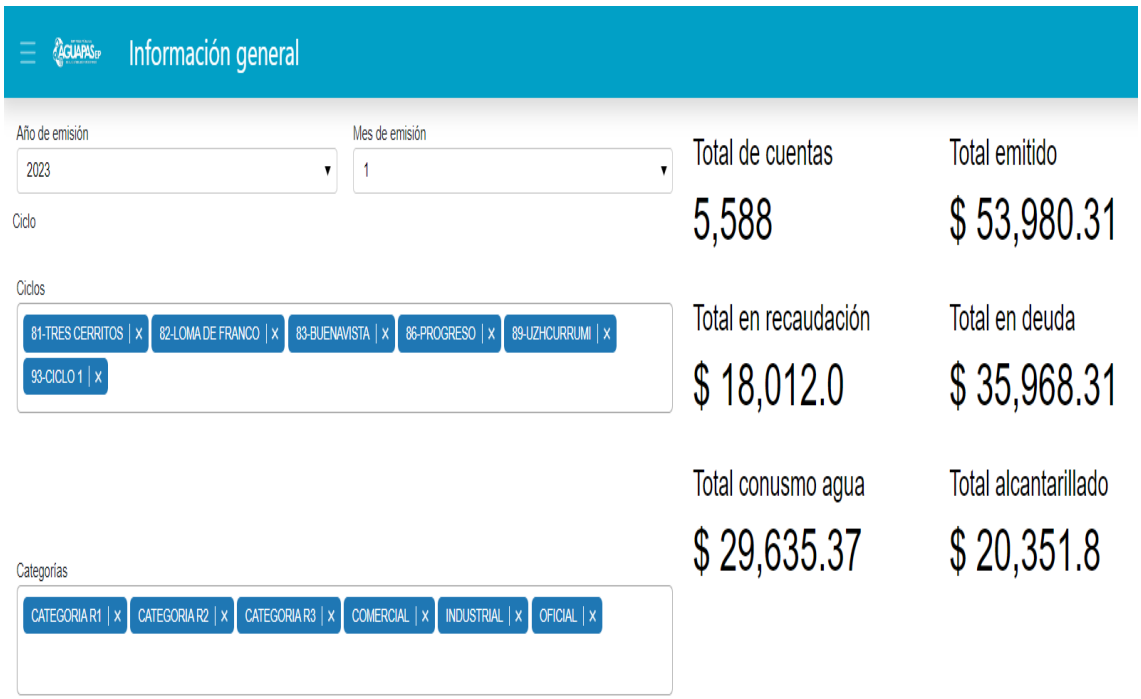


Ilustración 38 Información General

Top 10 deudores

Propietario - cuenta	Meses de deuda	Total de deuda	Ciclo	categoria
AEROFUMIGADORA FUMISANDRA S.A. - 9505	238	17,370.71	LOMA DE FRANCO	INDUSTRIAL
TOBAR BOLIVAR - 3239	243	3,203.58	PASAJE 1	CATEGORIA R3
MUNICIPIO DE PASAJE BAÃ□OS PUBLICOS - 6754	241	2,983.59	PASAJE 1	OFICIAL
DESCONOCIDO DESCONOCIDO - 10789	220	2,825.01	LOMA DE FRANCO	CATEGORIA R3
ROMERO CABRERA JOSE OLMEDO - 13109	226	2,733.94	PASAJE 1	CATEGORIA R3
SAQUICARAY JUAN DE DIOS - 1579	242	2,692.55	PASAJE 1	CATEGORIA R3
ABAD VALLEJO WILLIAM ABRAHAM HRDS - 13175	122	2,642.83	PASAJE 1	INDUSTRIAL

Ilustración 39 Top 10 deudores

También vamos a poder conocer información por cuenta y del año de emisión del cual se necesite información, debido que existe la necesidad de conocer detalles de una cuenta en particular como se muestra en la **ilustración 40 y 41**, esto ayudará para diferentes estrategias de negocio como la recuperación de cartera de un cliente en particular o empresas.

Ilustración 40 Información por Cuenta

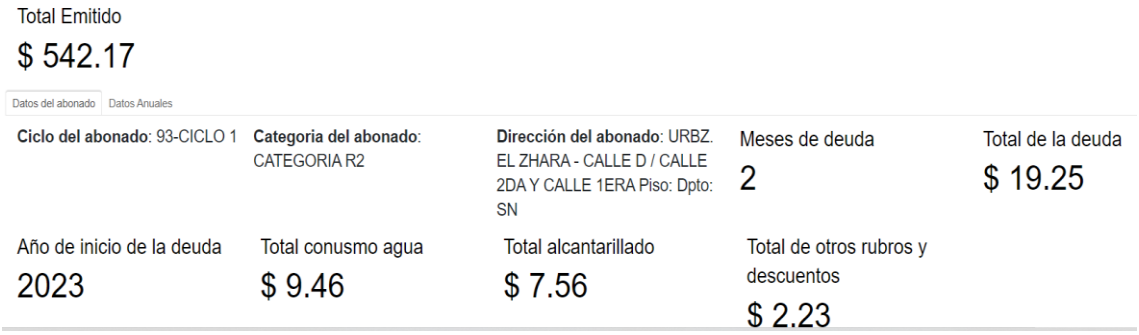


Ilustración 41 Detalle de la búsqueda

De la misma manera podremos conocer información anual como se detalla en la **ilustración 42, 43 y 44**, donde podremos realizar una comparativa de recaudaciones de años donde se puede segmentar por la categoría de la cuenta del usuario y en la comparativa nos muestra en detalle mensual, es decir por emisión. Información de recaudaciones versus deuda de la EPAAA, en detalle esta información nos muestra un histórico de las recaudaciones y las deudas. Además, un análisis puntual respecto a las recaudaciones versus deuda de un año donde se puede segmentar por la categoría de la cuenta del usuario.

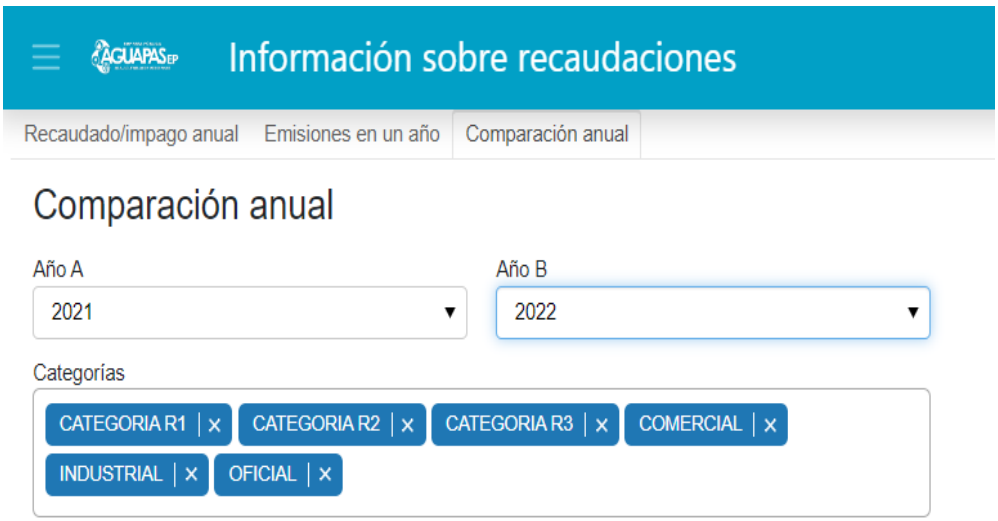


Ilustración 42 Elección de filtros

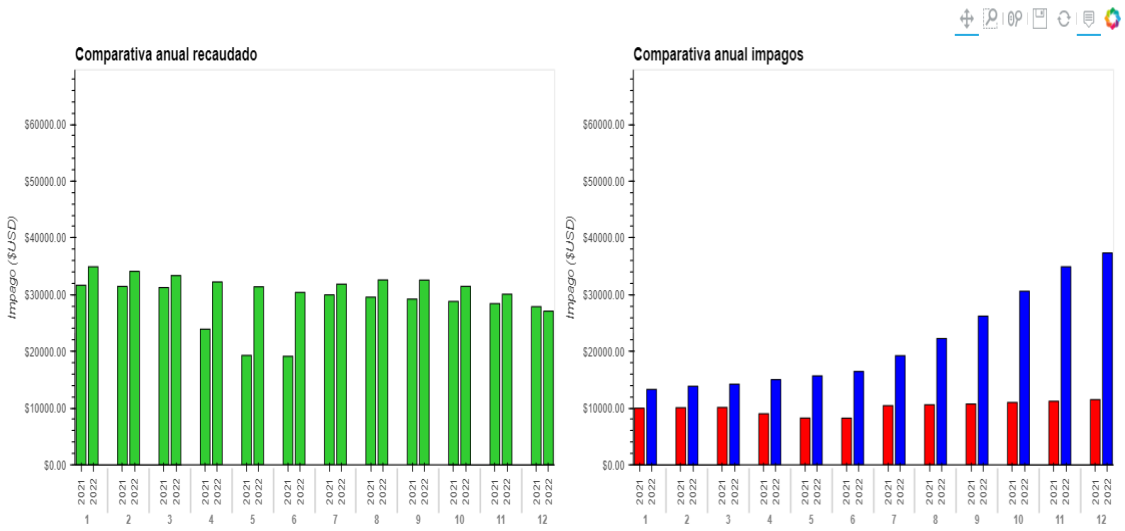


Ilustración 43 Comparación Anual

Ilustración 44 Comparación Histórica

De igual manera podemos revisar los Impagos versus las personas que se encuentran al día, donde estas métricas podremos filtrarlas por categorías, año de emisión y mes de

emisión, este criterio podremos visualizarlo por ciclos (parroquias del cantón pasaje) y por categoría como se detalla en la **ilustración 45 y 46**.

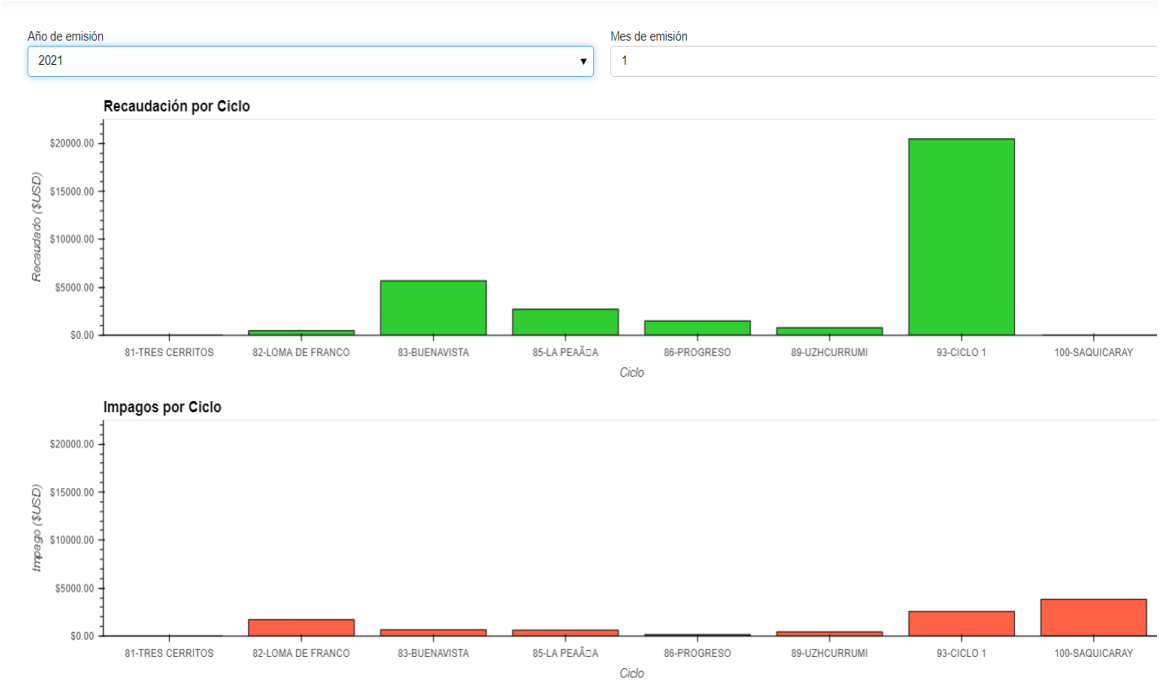


Ilustración 45 Recaudación por Ciclo

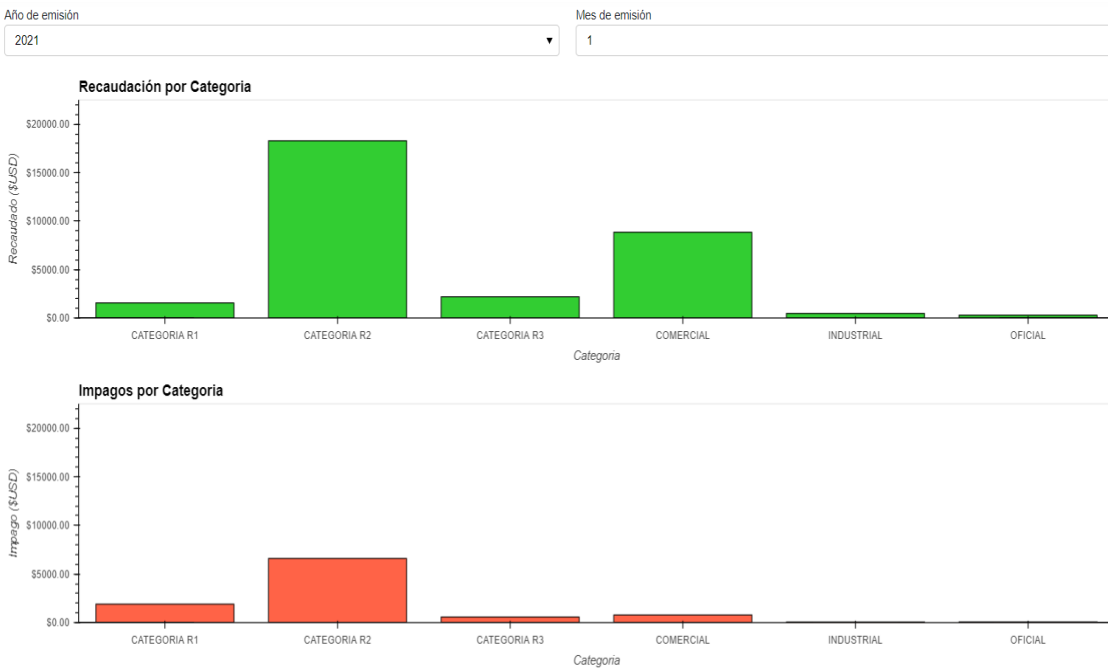


Ilustración 46 Recaudación por categoría

Análisis de Correlación para el análisis de cartera y gestión de recaudación en la EPAAA

En el caso de correlación, al ser una técnica de análisis de datos, no nos permite evaluarle como se evaluaría un modelo; sin embargo, es parte fundamental para descubrir y confirmar posibles relaciones entre los datos, particularmente relaciones lineales que, si bien no es el caso de todos los datos, habilita una base para poder decidir si es factible utilizar regresiones lineales.

Análisis de aplicación de Regresión Lineal para el análisis de cartera y Gestión de Recaudación en la EPAAA

Considerando como precedente los resultados obtenidos con la matriz de correlación, se decidió implementar un modelo de regresión lineal con la finalidad de estimar la deuda tomando como valor de entrada la cantidad de mese de deuda. Esto es posible debido a que, en muchos casos, la deuda crece de manera lineal, pues al no existir micro medición muchas cuentas incrementan su deuda basada en las tarifas base e intereses. El modelo demostró con un R^2 de aproximado 0.90 (pues el valor varía dependiendo el ciclo y categoría con los que se genere el modelo), que es posible y viable realizar predicciones de deuda a partir de un modelo de regresión lineal simple como se muestra en la **ilustración 47**.

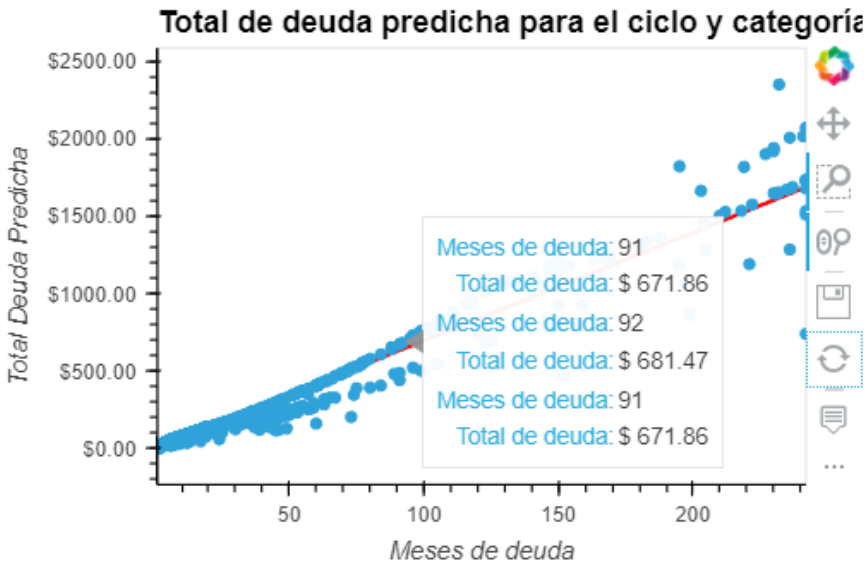


Ilustración 47 Cartera Predicha

Análisis de aplicación de Clustering para el análisis de cartera y Gestión de Recaudación en la EPAAA

La aplicación de técnicas de clustering [66] como K-means, nos habilita la posibilidad de conocer la morfología de los datos en cuanto a características refiere, es decir, nos permite generar grupos de manera rápida para posteriormente poder examinar los grupos obtenidos y extraer información relevante de esta segmentación. En este caso particular, se logró descubrir que la mayor parte de la deuda, y de cuentas con deuda, se encuentra en aquellas cuentas con una deuda media, con rangos de deuda de entre 70 a 170 meses, y aquellas cuentas con deudas mayores, son el grupo minoritario. Esto es un dato relevante pues se estimaba por conocimiento empírico, que la mayor cantidad de cuentas con deuda se encontraba entre los clientes con deudas pequeñas y recientes, demostrando así la importancia de las técnicas de análisis de datos dentro de la toma de decisiones en una organización. Un ejemplo de la aplicación es la **ilustración 48**.

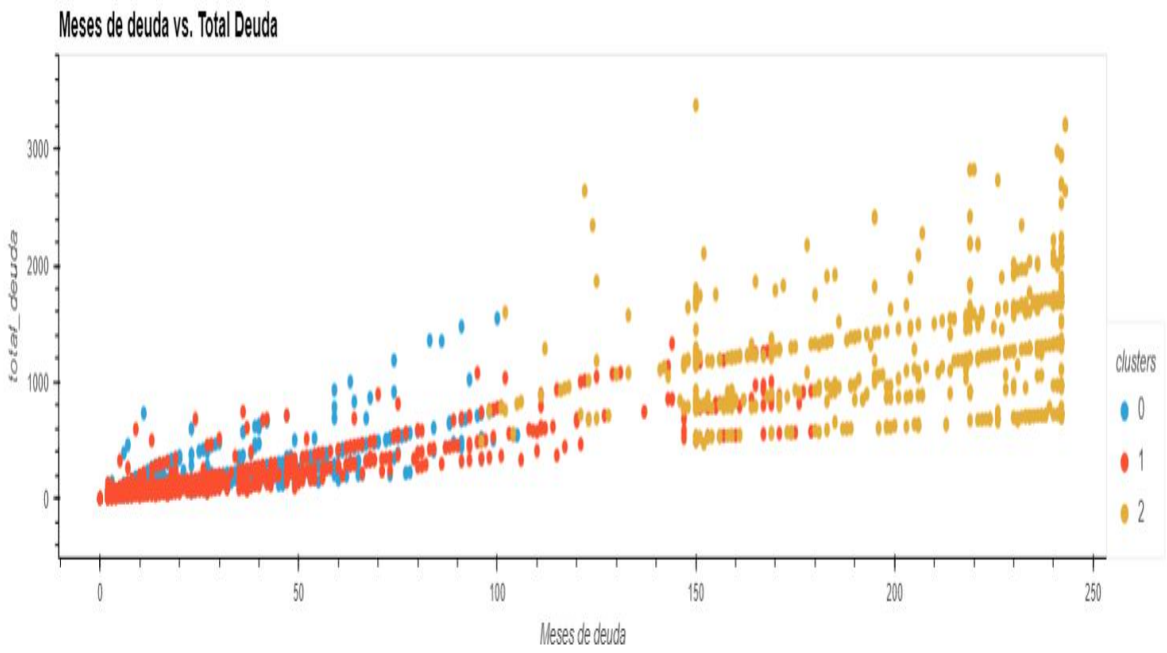


Ilustración 48 Meses de deuda vs total deuda

Análisis de aplicación de Clasificación para el análisis de cartera y Gestión de Recaudación en la EPAAA

Finalmente, el objetivo del uso de técnicas de aprendizaje automático, es conocer y modelar las características generales de una cuenta, para determinar si esta pagará o no durante las emisiones siguientes, lo que habilita decisiones particulares sobre dicho cliente. Se logró modelar de manera eficaz con todas las técnicas propuestas, por ejemplo el modelo conseguido mediante *support vector machine* [67], siendo uno de los 7 modelos aplicados en este trabajo, responde de buena manera a la realidad del caso, con métricas de rendimiento aceptables y generalizando correctamente los datos de entrada como se muestra en la **Tabla 10**.

Tabla 10 Modelos de Clasificación

Modelos de Clasificación			
Modelos	**Accuracy**	**Recall**	**Precisión**
Perceptrón Multicapa	1	1	1
Regresión Logistica	1	1	1
Máquina de Soporte de Vectores	0,877	0,938	0,7
Arboles de decisión	1	1	1
Clasificador de Naive Bayes	1	1	1
Clasificador K-NN	0,965	0,969	0,907
Clasificador de Bosques Aleatorios	1	1	1

Se justifica en primera instancia estos resultados por tener grandes cantidades de datos y pocas clases es por ello que refleja un *overfitting.* En una comparativa general se puede manifestar que todos modelos han tenido una aceptación en cuanto a la cartera de la EPAAA, gracias a esto se ha podido identificar patrones de comportamiento y apoyarse en una herramienta para simplificar la toma de decisiones en el análisis de la cartera y en la gestión de recaudación, el tiempo de respuesta para el conocimiento de la información es eficiente.

Tabla 11 Comparación de algoritmos

APLICACIÓN DE APRENDIZAJE AUTOMÁTICO				
ALGORITMO	**MÉTODOS**	**ESPERADO**	**CRITERIOS**	**DETALLE**
Regresión Lineal	Regresión Lineal	0,8 - 1	0,9	criterio R^2
Clustering	kmeans	0,5 - 1	0,535	coefinciente de silhouette
Clasificación	Perceptrón Multicapa	0,8 - 1	1	accuracy
Clasificación	Regresión Logística	0,8 - 1	1	accuracy
Clasificación	Máquina de Soporte de Vectores	0,8 - 1	0,877	accuracy
Clasificación	Árboles de decisión	0,8 - 1	1	accuracy
Clasificación	Clasificador de Naive Bayes	0,8 - 1	1	accuracy
Clasificación	Clasificador K-NN	0,8 - 1	0,965	accuracy
Clasificación	Clasificador de Bosques Aleatorios	0,8 - 1	1	accuracy

Podemos manifestar que, mediante la aplicación de estos 3 algoritmos y basados en la encuesta de satisfacción realizada, son aceptables para este proyecto como lo muestra la **tabla 11**.

Por otro lado, una vez realizado el análisis de la encuesta en cuanto a la satisfacción del desarrollo del sistema, es pertinente realizar la interpretación de los resultados en cuanto a la hipótesis planteada.

Corroboración de la Hipótesis

Como hipótesis de investigación tenemos:

El desarrollo de un sistema de soporte de decisiones en la EPAAA de Pasaje, con algoritmos de aprendizaje automático permitirá una gestión eficiente de recaudaciones.

De esta hipótesis se derivan la hipótesis nula y la alternativa:

Ho: El desarrollo de un sistema de soporte de decisiones en la EPAAA de Pasaje, con algoritmos de aprendizaje automático permitirá una gestión de recaudaciones menor al 80% de eficiencia.

H1: El desarrollo de un sistema de soporte de decisiones en la EPAAA de Pasaje, con algoritmos de aprendizaje automático permitirá en un 80%, una gestión eficiente de recaudaciones.

Los resultados se han obtenido según el criterio de satisfacción baja o satisfacción alta que se encuentran en escalas de "no satisfecho", "poco satisfecho", "moderadamente satisfecho", "muy satisfecho" y "Extremadamente satisfecho", como se muestra en la **Tabla 12**.

Tabla 12 Taba de satisfacción

	No Satisfecho	Poco Satisfecho	Moderadamente satisfecho	Muy Satisfecho	Extramadamente Satisfecho
	0	0	0	4	68
Total	72				
Total de Satisfacción	0		72		
Total porcentaje de satisfacción	0%		100%		
	Satisfacción Baja		Satisfacción alta		

En cuanto a la obtención de resultados, se puede evidenciar que los valores corresponden a una escala de "satisfacción alta" con el 100% de satisfacción, estos valores son determinantes para la prueba de hipótesis.

Prueba para saber si aplica a una prueba de hipótesis (Bernoulli)

$np0 \geq 5, n\ (1 - p0) \geq 5$

p0= x/n = 72/72 = 1

$72(1) \geq 5, 72(1\text{-}1) \geq 5$

$72 \geq 5, 0 \geq 5$

En la prueba de hipótesis se evidencia que no se cumplen las condiciones, en la condición dos no aplica, por consiguiente, no amerita hacer una prueba de hipótesis debido a que el resultado de satisfacción es del 100%.

Previamente al desarrollo del DSS en una encuesta realizada, se podía evidenciar la ausencia de resultados eficientes para realizar un correcto análisis de cartera y gestión de recaudación, con el desarrollo de un DSS aplicando aprendizaje automático, aumento el

rendimiento del personal considerando un 100% de “satisfacción alta” ante la implementación del mismo.

Con la implementación de un DSS con interfaz dashboard y la aplicación de los algoritmos de aprendizaje automático se obtuvo resultados aceptables, aunque grandes cantidades de datos y pocas clases provoquen un *overfitting*. El trabajo desarrollado permitió realizar una comparativa de estos algoritmos y evaluarlos a través del tiempo de ejecución debido a la gran cantidad de información que posee una base de datos inédita mencionada anteriormente, donde estos resultados permitieron predecir comportamiento de un usuario respecto a la deuda que mantiene con la EPAAA. Permitiendo caracterizar los algoritmos y técnicas de minería de datos y aprendizaje automático, que permitieron el análisis eficiente de cartera y gestión de recaudación.

Para trabajos futuros se recomienda implementar técnicas los cuales se pueda agrupar los datos de la EPAAA y le permitan realizar otras pruebas con respectos a los algoritmos de clasificación.

CONCLUSIONES

- Aplicando una revisión sistemática de literatura se pudo lograr el primer objetivo específico, donde se pudo traer estudios similares para poder caracterizar los algoritmos y técnicas de minería de datos y aprendizaje automático, que permitieron el análisis de cartera y gestión de recaudación y de la misma manera herramientas de ayuda para implementar la inteligencia de negocios mediante un *dashboard* que apoyó al análisis de cartera y la gestión de recaudación de la EPAAA.

- Se pudo realizar una evaluación de los algoritmos utilizados y métodos usados en la implementación del DSS, donde la regresión lineal obtuvo un criterio de 0.90 considerado Fuerte. En *clustering* se logró descubrir que la mayor parte de la deuda y de cuentas con deuda, se encuentra en aquellas cuentas con una deuda media, con rangos de deuda de entre 70 a 170 meses y aquellas cuentas con deudas mayores, son el grupo minoritario. Además, nos refleja un coeficiente de silhouette de 0.535 que representa la calidad de agrupamiento que obtuvo el algoritmo, donde el resultado nos indica que es un valor considerable entre separación de los cluster. Para clasificación todos los modelos tuvieron buena aceptación, un ejemplo es la máquina de soporte de vectores cuya exactitud es de 0.877 el cual se considera aceptable, logrando el segundo objetivo específico planteado.

- Se pudo apoyar las decisiones de altos mandos, gracias a la obtención de información mediante la implementación un DSS con interfaz *dashboard* en la EPAAA donde facilitó el análisis de cartera y gestión de recaudación. Es decir se pudo conocer la cantidad de recaudación total de una emisión puntual y así mismo la cantidad de cartera generada, donde se pudieron realizar comparativas del costo de inversión mensual en una planta de una parroquia del cantón Pasaje y la cantidad de ingreso por concepto de recaudación de agua potable de la misma, pudiendo tomar decisiones en cuanto a recursos tanto humano como económico para dicha parroquia La Peaña.

- El desarrollo del DSS ayudo a segmentar los usuarios homogéneos, se puede conocer de informaciones puntuales que antes era imposible como lo es la

recaudación por categorías, el total de recaudación en el mes de una emisión puntual, mejorando los tiempos de respuestas a la hora de tomar decisiones y como tomar decisiones con lugares en la cual la recaudación es baja.

RECOMENDACIONES

- Se recomienda la aplicación de aprendizaje automático, específicamente la utilización de modelos de clasificación en la EPAAA para tener un análisis puntual prediciendo el comportamiento de un cliente, ya que esto ayudará a clasificarlos como pagadores o deudores y poder generar un apoyo en la toma de decisiones de altos mandos y mandos técnicos, así mismo *clustering* debido a que podremos clasificar por clusters a las parroquias o ciclos y poder identificar su comportamiento referente a la cartera al transcurrir el tiempo. La regresión lineal nos ayudará debido a que en este caso la deuda crece linealmente y es posible segmentar el comportamiento para cierta cantidad de meses.

- Se recomienda la aplicación de inteligencia de negocios mediante un *dashboard* en la EPAAA ya que esta es una herramienta donde se puede incrementar el conocimiento de los altos mandos y mandos medios, debido a que la cantidad de información ahora es aprovechada y apoya en gran medida a la toma de decisiones en la empresa.

- Siendo la minería de datos un recurso propicio para poder identificar patrones y relaciones en cuanto a volúmenes de datos, es recomendable su utilización ya que esto apoya a la toma de decisiones de la EPAAA debido a que la gran cantidad de datos se transforman en conocimiento.

BIBLIOGRAFÍA

[1] P. Mikalef, M. Boura, G. Lekakos, y J. Krogstie, «The role of information governance in big data analytics driven innovation», *Inf. Manage.*, vol. 57, n.º 7, p. 103361, nov. 2020, doi: 10.1016/j.im.2020.103361.

[2] B. MAZÓN-Olivo, M. JARAMILLO-Paredes, O. ROMERO-Hidalgo, A. Borja, M. AGUIRRE-Benalcazar, y M. CONTENTO-Segarra, «Tecnologías de Inteligencia de Negocios y Minería de datos para el análisis de la producción y comercialización de cacao», p. 15.

[3] H. F. Vallejo Ballesteros, E. Guevara Iñiguez, y S. R. Medina Velasco, «Minería de Datos», *RECIMUNDO*, vol. 2, n.º Esp, pp. 339-349, feb. 2018, doi: 10.26820/recimundo/2.esp.2018.339-349.

[4] «Call for Papers: Special Issue on 'AI and machine learning in finance'», *Quant. Finance*, vol. 18, n.º 4, p. 533, abr. 2018, doi: 10.1080/14697688.2018.1444134.

[5] G. P. Clarkson y A. H. Meltzer, «PORTFOLIO SELECTION: A HEURISTIC APPROACH*», *J. Finance*, vol. 15, n.º 4, pp. 465-480, dic. 1960, doi: 10.1111/j.1540-6261.1960.tb02764.x.

[6] P. Vats y K. Samdani, «Study on Machine Learning Techniques In Financial Markets», en *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, Pondicherry, India: IEEE, mar. 2019, pp. 1-5. doi: 10.1109/ICSCAN.2019.8878741.

[7] K. Lee, D. Booth, y P. Alam, «A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms», *Expert Syst. Appl.*, vol. 29, n.º 1, pp. 1-16, jul. 2005, doi: 10.1016/j.eswa.2005.01.004.

[8] Instituto Federal Goiano Campus Ceres, GO, Brazil, M. de M. Sousa, R. S. Figueiredo, y Federal University of Goiás, GO, Brazil, «CREDIT ANALYSIS USING DATA MINING: APPLICATION IN THE CASE OF A CREDIT UNION», *J. Inf. Syst. Technol. Manag.*, vol. 11, n.º 2, pp. 379-396, ago. 2014, doi: 10.4301/S1807-17752014000200009.

[9] M. Becha, O. Dridi, O. Riabi, y Y. Benmessaoud, «Use of Machine Learning Techniques in Financial Forecasting», en *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*, Tunis, Tunisia: IEEE, feb. 2020, pp. 1-6. doi: 10.1109/OCTA49274.2020.9151854.

[10] J. Zhao, «Efficiency of corporate debt financing based on machine learning and convolutional neural network», *Microprocess. Microsyst.*, vol. 83, p. 103998, jun. 2021, doi: 10.1016/j.micpro.2021.103998.

[11] R. Hernández Sampieri y C. P. Mendoza Torres, *Metodología de la investigación: las rutas cuantitativa, cualitativa y mixta*, First edition. México: McGraw-Hill Education, 2018.

[12] D. Conti y A. Rodríguez, «Teoría de carteras de inversión para la diversificación del riesgo: enfoque clásico y uso de redes neuronales artificiales (RNA)», vol. 26, n.º 1, p. 9, 2005.

[13] S. Stancu, A. M. Constantin, O. M. Predescu, y S. V. Stancu, «Sovereign Debt Crisis – An Approach Based on Clusterization and Binary Classification Branche», *Procedia - Soc. Behav. Sci.*, vol. 93, pp. 1926-1930, oct. 2013, doi: 10.1016/j.sbspro.2013.10.142.

[14] B. Gao, «The Use of Machine Learning Combined with Data Mining Technology in Financial Risk Prevention», *Comput. Econ.*, feb. 2021, doi: 10.1007/s10614-021-10101-0.

[15] S. Khemakhem y Y. Boujelbene, «Predicting credit risk on the basis of financial and non-financial variables and data mining», *Rev. Account. Finance*, vol. 17, n.º 3, pp. 316-340, ago. 2018, doi: 10.1108/RAF-07-2017-0143.
[16] W. Xu, Y. Chen, C. Coleman, y T. F. Coleman, «Moment matching machine learning methods for risk management of large variable annuity portfolios», *J. Econ. Dyn. Control*, vol. 87, pp. 1-20, feb. 2018, doi: 10.1016/j.jedc.2017.11.002.
[17] F. D. Paiva, R. T. N. Cardoso, G. P. Hanaoka, y W. M. Duarte, «Decision-making for financial trading: A fusion approach of machine learning and portfolio selection», *Expert Syst. Appl.*, vol. 115, pp. 635-655, ene. 2019, doi: 10.1016/j.eswa.2018.08.003.
[18] D. K. Chaudhary, S. Srivastava, y V. Kumar, «A Review on Hidden Debts in Machine Learning Systems», en *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, Bangalore, India: IEEE, ago. 2018, pp. 619-624. doi: 10.1109/ICGCIoT.2018.8753081.
[19] J. M. Chen, M. U. Rehman, y X. V. Vo, «Clustering commodity markets in space and time: Clarifying returns, volatility, and trading regimes through unsupervised machine learning», *Resour. Policy*, vol. 73, p. 102162, oct. 2021, doi: 10.1016/j.resourpol.2021.102162.
[20] Y. G. Valle, D. Galpert, R. Molina-Ruiz, y G. Aguero-Chapin, «Integración de rasgos y aprendizaje semi-supervisado para la clasificación funcional de enzimas utilizando K-medias de Spark», vol. 14, n.º 4, p. 29, 2020.
[21] P. Z. Lappas y A. N. Yannacopoulos, «A machine learning approach combining expert knowledge with genetic algorithms in feature selection for credit risk assessment», *Appl. Soft Comput.*, vol. 107, p. 107391, ago. 2021, doi: 10.1016/j.asoc.2021.107391.
[22] Y.-P. Huang y M.-F. Yen, «A new perspective of performance comparison among machine learning algorithms for financial distress prediction», *Appl. Soft Comput.*, vol. 83, p. 105663, oct. 2019, doi: 10.1016/j.asoc.2019.105663.
[23] C. Albon, «Machine Learning with Python Cookbook : Practical Solutions From Preprocessing to Deep Learning», p. 366.
[24] A. T. Clavijo, «APLICACION DE LA MINERIA DE DATOS SOBRE BASES DE DATOS TRANSACCIONALES».
[25] D. Lemus-Delgado y R. Pérez Navarro, «Ciencia de datos y estudios globales: aportaciones y desafíos metodológicos», *Colomb. Int.*, n.º 102, pp. 41-62, abr. 2020, doi: 10.7440/colombiaint102.2020.03.
[26] J. Machicao, «La ciencia de datos para tomar mejores decisiones», 2022, doi: 10.13140/RG.2.2.15779.53289.
[27] A. M. Del Do, A. Villagra, y D. Pandolfi, «Desafíos de la Transformación Digital en las PYMES», *Inf. Científicos Téc. - UNPA*, vol. 15, n.º 1, pp. 200-229, mar. 2023, doi: 10.22305/ict-unpa.v15.n1.941.
[28] Ramirew W, «El empleo de la correlación lineal simple en Epidemiología Veterinaria», *redalyc.org*.
[29] M. Juan y G. Astorga, «Aplicación de modelos de regresión lineal para determinar las armónicas de tensión y corriente».
[30] Z. M. Rodriguez, «Machine Learning Process to Determine the Social Demand».
[31] C. W. García Estrella, E. Barón Ramírez, y S. K. Sánchez Gárate, «La inteligencia de negocios y la analítica de datos en los procesos empresariales», *Rev. Científica Sist. E Informática*, vol. 1, n.º 2, pp. 38-53, jul. 2021, doi: 10.51252/rcsi.v1i2.167.
[32] A. J. G. Morales, «Inteligencia de negocios, una ventaja competitiva para las organizaciones», *Cienc. Tecnol.*.

[33] J. A. R. Caracúin, «UNIVERSIDAD DE SAN CARLOS DE GUATEMALA».
[34] B. Mazon-Olivo, I. Ramirez, y A. Pan, «Ciencia de datos en el sector agropecuario», 2018.
[35] N. D. Duque Méndez, E. J. Hernández Leal, Á. M. Pérez Zapata, A. F. Arroyave Tabares, y D. A. Espinosa Gómez, «Modelo para el proceso de extracción, transformación y carga en bodegas de datos. Una aplicación con datos ambientales», *Cienc. E Ing. Neogranadina*, vol. 26, n.º 2, pp. 95-109, ago. 2016, doi: 10.18359/rcin.1799.
[36] M. Jarke, M. A. Jeusfeld, C. Quix, y P. Vassiliadis, «Architecture and quality in data warehouses: An extended repository approach», *Inf. Syst.*, vol. 24, n.º 3, pp. 229-253, may 1999, doi: 10.1016/S0306-4379(99)00017-4.
[37] M. A. Valles Coral, L. M. Hidalgo Macedo, y J. C. Santa-María, «MONITOREO BASADO EN DASHBOARD Y SU EFECTO EN EL CUMPLIMIENTO DE LOS ESTÁNDARES DE ACREDITACIÓN», *Crescendo*, vol. 10, n.º 1, p. 161, jun. 2019, doi: 10.21895/incres.2019.v10n1.10.
[38] A. Géron, «Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow», p. 851.
[39] G. Narula, M. Haeberlin, J. Balsiger, C. Strässle, L. L. Imbach, y E. Keller, «Detection of EEG burst-suppression in neurocritical care patients using an unsupervised machine learning algorithm», *Clin. Neurophysiol.*, vol. 132, n.º 10, pp. 2485-2492, oct. 2021, doi: 10.1016/j.clinph.2021.07.018.
[40] O. E. Aissaoui, Y. E. A. El Madani, L. Oughdir, y Y. E. Allioui, «Combining supervised and unsupervised machine learning algorithms to predict the learners' learning styles», *Procedia Comput. Sci.*, vol. 148, pp. 87-96, 2019, doi: 10.1016/j.procs.2019.01.012.
[41] M. Fang y S. Taylor, «A machine learning based asset pricing factor model comparison on anomaly portfolios», *Econ. Lett.*, vol. 204, p. 109919, jul. 2021, doi: 10.1016/j.econlet.2021.109919.
[42] I. C. P. Verona y L. A. García, «Una revisión sobre aprendizaje no supervisado de métricas de distancia», vol. 10, n.º 4, p. 26, 2016.
[43] Á. Farias Pinheiro, D. S. Da Silveira, y F. B. D. Lima Neto, «Use of Machine Learning for Active Public Debt Collection with Recommendation for the Method of Collection Via Protest», en *Artificial Intelligence and Applications*, Academy and Industry Research Collaboration Center (AIRCC), may 2022, pp. 99-108. doi: 10.5121/csit.2022.120909.
[44] R. González, A. Barrientos, M. Toapanta, y J. Del Cerro, «Aplicación de las Máquinas de Soporte Vectorial (SVM) al diagnóstico clínico de la Enfermedad de Párkinson y el Temblor Esencial», *Rev. Iberoam. Automática E Informática Ind. RIAI*, vol. 14, n.º 4, pp. 394-405, oct. 2017, doi: 10.1016/j.riai.2017.07.005.
[45] L. Quiñones Huatangari, L. Ochoa Toledo, N. Kemper Valverde, O. Gamarra Torres, J. Bazán Correa, y J. Delgado Soto, «Red neuronal artificial para estimar un índice de calidad de agua», *Enfoque UTE*, vol. 11, n.º 2, pp. 109-120, abr. 2020, doi: 10.29019/enfoque.v11n2.633.
[46] H. Vivas, H. J. Martınez, y R. Perez, «Structured Secant Method for the Multilayer Perceptron Training», *Rev. Cienc.*, n.º 2, 2014.
[47] A. Franco-Arcega, J. A. Carrasco-Ochoa, G. Sánchez-Díaz, y J. F. Martínez-Trinidad, «Decision Tree based Classifiers for Large Datasets», vol. 15, n.º 2, 2013.
[48] D. A. López-Sarmiento, H. C. Manta-Caro, y N. E. Vera-Parra, «Clasificador basado en una máquina de vectores de soporte de mínimos cuadrados frente a un

clasificador por regresión logística ante el reconocimiento de dígitos numéricos», *TecnoLógicas*, n.º 31, p. 37, nov. 2011, doi: 10.22430/22565337.99.
[49] E. López-Pezoa, A. Cáceres-Estigarribia, S. A. Grillo, y E. Herrera, «Accuracy evaluation of Naive Bayes and Logistic Regression for classification with binary attributes and classes», *Rep. Científicos FACEN*, vol. 13, n.º 1, pp. 73-84, jun. 2022, doi: 10.18004/rcfacen.2022.13.1.73.
[50] A. Zapata-Tapasco, S. Pérez-Londoño, y J. Mora-Flórez, «Método basado en clasificadores k-NN parametrizados con algoritmos genéticos y la estimación de la reactancia para localización de fallas en sistemas de distribución.».
[51] G. Sánchez-Díaz, U. E. Escobar-Franco, L. R. Morales, I. Piza-Dávila, C. Aguirre-Salado, y A. Franco-Arcega, «Incremental k most similar neighbor classifier for», 2013.
[52] A. M. Padilla-Ospina, J. E. Medina-Vásquez, y J. H. Ospina-Holguín, «Métodos de aprendizaje automático en los estudios prospectivos desde un ejemplo de la financiación de la innovación en Colombia», *Rev. Investig. Desarro. E Innov.*, vol. 11, n.º 1, pp. 9-21, ago. 2020, doi: 10.19053/20278306.v11.n1.2020.11676.
[53] «1-4-HERNÁNDEZ SAMPIERI».
[54] F. González, «¿QUÉ ES UN PARADIGMA? ANÁLISIS TEÓRICO, CONCEPTUAL Y PSICOLINGÜÍSTICO DEL TÉRMINO», vol. 20, 2005.
[55] P. Fernández, G. Vallejo, P. Livacic-Rojas, y E. Tuero, «Validez Estructurada para una investigación cuasi-experimental de calidad. Se cumplen 50 años de la presentación en sociedad de los diseños cuasi-experimentales», *An. Psicol.*, vol. 30, n.º 2, pp. 756-771, may 2014, doi: 10.6018/analesps.30.2.166911.
[56] E. Ortiz, «Los Niveles Teóricos y Metodológicos en la Investigación Educativa», *Cinta Moebio*, n.º 43, pp. 14-23, mar. 2012, doi: 10.4067/S0717-554X2012000100002.
[57] D. A. Luis, «DE LA INVESTIGACIÓN EN LA EDUCACIÓN SUPERIOR», 2021.
[58] J. J. Espinosa Zúñiga, «Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública», *Ing. Investig. Tecnol.*, vol. 21, n.º 1, pp. 1-13, ene. 2020, doi: 10.22201/fi.25940732e.2020.21n1.008.
[59] A. Rodríguez Jiménez y A. O. Pérez Jacinto, «Métodos científicos de indagación y de construcción del conocimiento», *Rev. Esc. Adm. Negocios*, n.º 82, pp. 175-195, jul. 2017, doi: 10.21158/01208160.n82.2017.1647.
[60] L. Rojas Rubio y C. Meneses Villegas, «Una comparación empírica de algoritmos de aprendizaje automático versus aprendizaje profundo para la detección de noticias falsas en redes sociales», *Ingeniare Rev. Chil. Ing.*, vol. 30, n.º 2, pp. 403-415, jun. 2022, doi: 10.4067/S0718-33052022000200403.
[61] L. Díaz-De-la-Paz, J. L. García-Mendoza, B. E. López-Porrero, L. Manuela, y W. Lemahieu, «Técnicas para capturar cambios en los datos y mantener actualizado un almacén de datos», vol. 9, n.º 4, 2015.
[62] A. Pérez Acosta, M. Moreno Espino, y R. Bandón Casamayor, «Goal-oriented dashboard's requirements with i: a case study», *Ingeniare Rev. Chil. Ing.*, vol. 24, n.º 4, pp. 680-689, oct. 2016, doi: 10.4067/S0718-33052016000400012.
[63] A. M. Posada y M. A. Ampuero, «Herramienta de soporte a un sistema de métricas e indicadores para la gestión de proyectos», vol. 7, n.º 2, 2013.
[64] G. M. Curbelo, «PARA EL ANÁLISIS DE CORRELACIÓN Y CONCORDANCIA EN EQUI- POS DE MEDICIONES SIMILARES», 2016.

[65] Md. Zubair, Md. A. Iqbal, A. Shil, M. J. M. Chowdhury, M. A. Moni, y I. H. Sarker, «An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling», *Ann. Data Sci.*, jun. 2022, doi: 10.1007/s40745-022-00428-2.
[66] A. Y. Al-Omary y M. S. Jamil, «A new approach of clustering based machine-learning algorithm», *Knowl.-Based Syst.*, vol. 19, n.º 4, pp. 248-258, ago. 2006, doi: 10.1016/j.knosys.2005.10.011.
[67] F. Provost y T. Fawcett, «Data Science for Business».
[68] «Clarkson y Meltzer - 1960 - PORTFOLIO SELECTION A HEURISTIC APPROACH.pdf».