SISTEMA DE RECOMENDACIÓN DE DOSIS DE COAGULANTE EN LA PLANTA DE TRATAMIENTO EL TABLAZO DEL MUNICIPIO DE POPAYÁN UTILIZANDO APRENDIZAJE AUTOMÁTICO

JAIDY VANESSA FERNÁNDEZ ALVAREZ Y DANIELA GRANADA SALAZAR

Tesis de pregrado en Ingeniería electrónica y telecomunicaciones

Director:

PhD. Cristhian Figueroa Martínez

CoDirector:

PhD. Juan Carlos Corrales

Asesor:

PhD. Mauricio Ramírez

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones

Departamento de Telemática

Línea de investigación e-ambiente

Popayán, Agosto 2023

JAIDY VANESSA FERNANDEZ ALVAREZ
DANIELA GRANADA SALAZAR

SISTEMA DE RECOMENDACIÓN DE DOSIS DE COAGULANTE PARA LA PLANTA DE TRATAMIENTO EL TABLAZO DEL MUNICIPIO DE POPAYÁN UTILIZANDO APRENDIZAJE AUTOMÁTICO

Tesis presentada a la Facultad de Ingeniería Electrónica y Telecomunicaciones de la Universidad del Cauca para la obtención del Título de

Ingeniero en Electrónica y Telecomunicaciones

Director:
PhD. Cristhian Figueroa

CoDirector:
PhD. Juan Carlos Corrales

Asesor:
Ing. Mauricio Ramírez

Popayán
2023

Agradecimientos

En primer lugar, deseamos expresar nuestro más profundo agradecimiento a nuestras familias. Su inquebrantable apoyo y aliento han sido fundamentales en cada etapa de este proceso y han sido el pilar que nos ha sostenido en los momentos más desafiantes.

De igual manera, nuestra gratitud al ingeniero Cristhian Figueroa. Su paciencia, dedicación y rigurosa guía han sido esenciales para la solución de nuestras dudas y la dirección correcta en el desarrollo de este proyecto. Valoramos inmensamente su compromiso y apoyo constante en nuestra formación académica y profesional.

Un reconocimiento especial al jefe de producción de la planta de tratamiento de agua "El Tablazo". Su generosidad al compartir no sólo los datos, sino también sus valiosos conocimientos prácticos, ha enriquecido significativamente nuestro trabajo. Gracias a su experiencia y orientación, pudimos fortalecer y validar la aplicabilidad de nuestro proyecto, obteniendo resultados altamente satisfactorios. Apreciamos sinceramente su disponibilidad, colaboración y amabilidad incesantes que trascienden más allá del ámbito profesional.

Finalmente, a todos aquellos que de alguna manera contribuyeron a la realización de este trabajo de grado, su aporte ha sido invaluable. Gracias por creer en nosotras y en lo que estábamos tratando de lograr.

Resumen Estructurado

Antecedentes: El agua, vital para la existencia humana y estrechamente vinculada a la salud pública, es tratada en plantas especializadas, como "El Tablazo" en Popayán, para asegurar su calidad. La coagulación, uno de los procesos esenciales en estas plantas, depende de una precisa dosificación del coagulante para ser efectiva y evitar riesgos sanitarios. Aunque en "El Tablazo" se utiliza la técnica de pruebas de jarra, esta puede ser susceptible a errores humanos y no adaptarse a cambios abruptos en la calidad del agua. La emergencia de modelos de Aprendizaje Automático (ML) promete mayor precisión en esta tarea, pero las investigaciones recientes a menudo pasan por alto el impacto de las variables meteorológicas en la coagulación y calidad del agua.

Objetivo: Desarrollar un sistema de recomendación de la dosis de coagulante para el proceso de potabilización del agua en la planta "El Tablazo" del municipio de Popayán utilizando técnicas de aprendizaje automático.

Métodos: En este estudio, se propone la adopción de avanzados modelos de ML que consoliden y analicen de manera integral datos hidrológicos y meteorológicos. La principal motivación de esta integración es diseñar un robusto sistema de recomendación que permita determinar de forma precisa la dosis de coagulante necesaria. A través de esta propuesta, se aspira no solo a garantizar y mejorar significativamente la calidad del agua potable, sino también a lograr una reducción notable en los costos asociados al tratamiento y potabilización del agua.

Resultados: A lo largo de la investigación, se logró consolidar un exhaustivo conjunto de datos, el cual comprende información tanto hidrológica como meteorológica, específicamente de la planta de tratamiento "El Tablazo". Adicionalmente, se desarrollaron y refinaron dos modelos predictivos para la determinación de la dosis de coagulante: uno orientado a la clasificación y otro a la regresión. Estos modelos, fruto de rigurosos análisis, fueron posteriormente entregados al jefe de producción de la mencionada planta de tratamiento, con el objetivo de ser implementados y evaluados en un entorno real y operativo, asegurando su aplicabilidad y eficacia en escenarios prácticos.

Conclusión: A lo largo del meticuloso estudio que fusionó variables hidrológicas y meteorológicas, se resaltó la importancia cardinal de las condiciones climáticas en determinar la calidad del agua y la precisión en la dosificación de coagulantes. La implementación de modelos avanzados, en particular el algoritmo de Bosques Aleatorios (RF), demostró ser un instrumento prevalente y robusto, capaz de manejar la complejidad de

dichas variables y ofrecer predicciones consistentes. Esta investigación no solo subraya la imperativa integración de datos ambientales en la gestión del agua, sino que también establece a RF como una herramienta clave, posicionando a la planta de tratamiento "El Tablazo" en la frontera de las innovaciones en tratamiento de agua.

Palabras clave: Coagulación, dosificación de coagulante, aprendizaje automático, datos meteorológicos, predicción, regresión y clasificación.

Structured Abstract

Background: Water, vital for human existence and closely linked to public health, is treated in specialized plants, such as "El Tablazo" in Popayán, to ensure its quality. Coagulation, one of the essential processes in these plants, relies on accurate coagulant dosing to be effective and avoid health risks. Although "El Tablazo" employs the jar test technique, it can be susceptible to human errors and may not adapt to abrupt changes in water quality. The emergence of Machine Learning (ML) models promises greater accuracy in this task, but recent research often overlooks the impact of meteorological variables on coagulation and water quality.

Objective: Develop a recommendation system for coagulant dosing in the water purification process at the "El Tablazo" plant in the municipality of Popayán using machine learning techniques.

Methods: In this study, the adoption of advanced ML models is proposed to comprehensively consolidate and analyze hydrological and meteorological data. The main motivation for this integration is to design a robust recommendation system that accurately determines the required coagulant dose. Through this proposal, the goal is not only to ensure and significantly improve the quality of drinking water but also to achieve a notable reduction in water treatment and purification costs.

Results: Throughout the research, an exhaustive data set was consolidated, encompassing both hydrological and meteorological information, specifically from the "El Tablazo"treatment plant. Additionally, two predictive models were developed and refined for determining the coagulant dose: one focused on classification and the other on regression. These models, the result of rigorous analysis, were subsequently provided to the production manager of the aforementioned treatment plant, aiming to be implemented and evaluated in a real and operational environment, ensuring their applicability and effectiveness in practical scenarios.

Conclusion: Throughout the meticulous study that merged hydrological and meteorological variables, the cardinal importance of climatic conditions in determining water quality and accuracy in coagulant dosing was highlighted. Implementing advanced models, particularly the Random Forests (RF) algorithm, proved to be a prevalent and robust tool, capable of handling the complexity of these variables and providing consistent predictions. This research not only emphasizes the imperative integration of environmental data in water management but also establishes RF as a key tool, positioning the "El

Tablazo" treatment plant at the forefront of innovations in water treatment.

Keywords: Coagulation, coagulant dosing, machine learning, meteorological data, prediction, regression, and classification.

Contenido

Lista de figuras

Lista de Tablas

Lista de Acrónimos

AI	**Artificial Intelligence**	**Inteligencia Artificial**	95
AIASD	**Artificial Intelligence for Aluminum Sulfate Dosing**	**Inteligencia Artificial para la Dosificación de Sulfato de Aluminio**	9
ANFIS	**Adaptive Neuro-Fuzzy Inference System**	**Sistema de Inferencia Neurodifuso Adaptativo**	10
ANN	**Artificial Neural Network**	**Redes Neuronales Artificiales**	9
ANOVA	**Analysis of Variance**	**Analisis de Varianza**	10
BP	**Back-Propagation**	**Retropropagación**	12
CCD	**Central Composite Design**	**Diseño Compuesto Central**	10
CF	**Cascading Feedforward**	**Alimentación Hacia Adelante en Cascada**	13
CNN	**Convolutional Neural Network**	**Red Neuronal Convolucional**	9
CRISP-DM	**Cross Industry Standard Process for Data Mining**	**Proceso Estándar de la Industria Cruzada para Minería de Datos**	19
CRISP-ML(Q)	**Cross Industry Standard Process for Machine Learning**	**Proceso Estándar de la Industria Cruzada para el Aprendizaje Automático**	19
DOC	**Dissolved Organic Carbon**	**Carbono Orgánico Disuelto**	15
DT	**Decision Tree**	**Árboles de Decisión**	96

ELM	**Extreme Learning Machine**	**Máquina de Aprendizaje Extremo**	12
ERT	**Extremely Randomized Tree**	**Árboles Extremadamente Aleatorios**	13
FIS	**Fuzzy Inference System**	**Sistema de Inferencia Difusa**	10
FNN	**Feedforward Neural Network**	**Red Neuronal de Alimentación Hacia Adelante**	12
GA	**Genetic Algorithm**	**Algoritmo Genético**	11
GB	**Gradient Boosting**	**Aumento de Gradiente**	97
KNN	**K-Nearest Neighbors**	**K-Vecinos más Cercanos**	96
LightGBM	**Light Gradient Boosting Machine**	**Máquina Ligera de Reforzamiento del Gradiente**	11
LSTM	**Long-Short Term Memory**	**Memoria a Corto-Largo Plazo**	12
MAE	**Mean Absolute Error**	**Error Absoluto Medio**	22
MAR	**Missing At Random**	**Ausentes al Azar**	41
MCAR	**Missing Completely At Random**	**Ausentes Completamente al Azar**	41
MICE	**Multiple Imputation by Chained Equations**	**Imputación Múltiple por Ecuaciones Encadenadas**	42
MNAR	**Missing Not At Random**	**Ausentes No al Azar**	41
ML	**Machine Learning**	**Aprendizaje automático**	2
MLP	**Multi-Layer Perceptron**	**Perceptrón Multicapa**	9
MLR	**Multiple Linear Regression**	**Regresión Lineal Múltiple**	13
MSE	**Mean Squared Error**	**Error Cuadrático Medio**	22
NARX	**Nonlinear AutoRegressive eXogenous**	**No Lineal Autoregresivo con Variables Exógenas**	12

PAC	**Polyaluminium Chloride**	**Cloruro de Polialuminio**	12
PAHCS	**Polialuminium Chloride Hydroxide Silicate**	**Polialuminio Cloruro Hidrosilicato**	13
PID	**Proportional, Integral, Derivative**	**Proporcional Integral Derivado**	14
PLSR	**Partial Least Squares Regression**	**Regresión de Mínimos Cuadrados Parciales**	13
RB	**Radial Basis**	**Base Radial**	13
RBP	**Reverse Back-Propagation**	**Retropropagación Inversa**	12
RF	**Random Forest**	**Bosques Aleatorios**	10
RMSE	**Root Mean Squared Error**	**Raiz del Error Cuadrático Medio**	22
SCADA	**Supervisory Control And Data Acquisition**	**Control de supervision y Adquisición de Datos**	83
SMOTE	**Synthetic Minority Over-sampling Technique**	**Técnica de sobremuestreo minoritario sintético**	50
SOSM	**Second Order Slip Mode**	**Modo de Deslizamiento de Segundo Orden**	10
SVM	**Support Vector Machine**	**Máquina de Vectores de Soporte**	10
TOC	**Total Organic Carbon**	**Carbono Orgánico Total**	14
WTP	**Water Treatment Plants**	**Planta de tratamiento de Agua**	1

Capítulo 1.

Introducción

1.1 Antecedentes.

El agua es un recurso vital para la vida humana, y su calidad es crucial para la salud pública. De hecho, la Asamblea General de las Naciones Unidas ha ratificado el derecho universal al acceso a agua potable segura y saludable con el fin de minimizar los riesgos para la salud asociados con el consumo de agua contaminada [2].

Las plantas de tratamiento de agua (WTP, *Water Treatment Plants*) son responsables de gestionar las propiedades físicas, químicas y microbiológicas del agua durante el proceso de potabilización. Su objetivo es garantizar que el agua sea segura para el consumo humano.

En consecuencia, para asegurar la calidad del agua, es esencial contar con un control riguroso y sistemático de los procesos de tratamiento en las WTP. Este proceso incluye varias etapas, tales como oxigenación, coagulación, decantación, filtración y desinfección o cloración [3]. El proceso de coagulación es particularmente crucial, ya que neutraliza las cargas eléctricas de las partículas suspendidas en el agua. Esto resulta en la formación de flóculos, grupos de partículas, que pueden ser eliminados mediante sedimentación o filtración [4].

La determinación precisa de la dosis de coagulante es clave para la mejora de la calidad del agua y para la eliminación eficaz de las partículas en suspensión. Diversos factores pueden influir en la dosificación, incluyendo las características del agua y las condiciones específicas del tratamiento. Una dosificación incorrecta puede llevar a una coagulación ineficiente, generar subproductos no deseados, aumentar los costos de tratamiento y presentar posibles problemas de salud pública [5, 6]. Por lo tanto, el cálculo y ajuste correcto de la dosis de coagulante se vuelve indispensable para garantizar la calidad y eficiencia del proceso de tratamiento de agua potable.

Actualmente existen diversas técnicas para determinar la dosis adecuada de coagulante entre las cuales se encuentran las pruebas de jarra, detectores de corriente de transmisión, análisis de calidad de agua y los modelos matemáticos. Las pruebas de jarra corresponden a la técnica más utilizada en las plantas de tratamiento de agua potable. Esta técnica simula en laboratorio las condiciones de una planta real con el fin de cal-

cular la dosis óptima de coagulante, la turbiedad residual y el tamaño del flóculo [7, 8]. Los detectores de corriente de transmisión miden la conductividad del agua y realizan ajustes a la dosis de coagulante según las fluctuaciones detectadas [9]. Los análisis de calidad del agua establecen la dosis necesaria de coagulante para reducir la turbiedad y otros contaminantes. Para esto consideran parámetros de calidad que cambian con las condiciones climáticas y estacionales [5]. Finalmente, los modelos matemáticos predicen la calidad del agua tratada y optimizan el proceso de coagulación, minimizando el uso de coagulante y reduciendo costos operativos [10].

En Colombia, la Resolución 2115-2007 y el Decreto 1575 de 2007 [11] establecen las guías, criterios y umbrales de funcionamiento de las plantas de tratamiento del país. A nivel local, en la ciudad de Popayán (Cauca), la mayor planta de tratamiento de agua potable es *"El Tablazo"*. Esta planta tiene la responsabilidad de garantizar que la calidad del agua suministrada a la mayoría de la ciudad cumpla con las regulaciones locales y nacionales para el consumo humano. Este desafío implica la implementación y el mantenimiento de una serie de procesos de tratamiento de agua, siendo uno de los más críticos el proceso de coagulación. No obstante, determinar la dosis correcta de coagulante presenta desafíos significativos y depende de diversos factores, lo cual puede afectar la eficiencia y la efectividad del proceso de tratamiento.

1.2 Planteamiento del problema.

En la actualidad la planta de "El Tablazo" determina la dosis de coagulante utilizando la técnica de pruebas de jarra. No obstante, esta técnica es altamente dependiente de la habilidad y experiencia de los operadores y no puede adaptarse rápidamente a cambios repentinos en las propiedades del agua. Lo anterior puede resultar en dosificaciones inexactas de coagulante, el cual en exceso produce altos niveles de aluminio residual en el agua tratada [12].

En este sentido, los modelos de Aprendizaje Automático (ML, *Machine Learning*) han surgido como una alternativa más eficiente y precisa para determinar la dosis de coagulante. Estos modelos pueden aprender con la experiencia y generar respuestas o sugerencias de manera más precisa. No obstante, en el presente estudio se encontró que los trabajos de investigación estudiados basados en modelos de ML han utilizado exclusivamente datos hidrológicos, pero no consideran variables meteorológicas (ver sección 2.2). Estas variables meteorológicas pueden influir en las características del agua que llega a la planta de "El Tablazo" y, por consiguiente, en el proceso de coagulación [13].

1.3 Hipótesis

La hipótesis del presente trabajo de investigación sostiene que, al implementar modelos de ML que integren tanto datos hidrológicos como meteorológicos, se puede optimizar el proceso de dosificación de coagulante en la planta de tratamiento de agua "El Tablazo".

En consecuencia, se propone un sistema de recomendación de dosis de coagulante que tiene en cuenta tanto los datos hidrológicos como los meteorológicos para mejorar

la precisión y la eficiencia del proceso de coagulación. Al superar las limitaciones de las técnicas actuales, se espera optimizar la calidad del agua y reducir los costos de tratamiento.

1.4 Objetivos.

1.4.1 Objetivo general.

Desarrollar un sistema de recomendación de la dosis de coagulante para el proceso de potabilización del agua en la planta el Tablazo del municipio de Popayán utilizando técnicas de aprendizaje automático.

1.4.2 Objetivos específicos.

- Caracterizar las variables que más impactan los procesos de coagulación y floculación del agua destinada a potabilización.
- Construir modelos de estimación de dosis de coagulante para el proceso potabilización de agua en la planta del Tablazo departamento del Cauca a partir de algoritmos de aprendizaje automático.
- Evaluar los modelos de aprendizaje automático construidos mediante el análisis de métricas de error.

1.5 Aportes del Proyecto

El presente trabajo de grado realiza los siguientes aportes:

- Conjunto de datos que contiene variables de agua cruda, datos de tratamiento y datos de agua tratada de la planta de agua potable "El Tablazo".
- Mecanismo de fusión de datos. (Disponible en repositorio en línea: `https://github.com/jvanessafdez/coagulant-dose`
- Modelos de predicción de dosis de coagulante.
- Artículo: ***Estimation of Water Turbidity in Drinking Water Treatment Plants Using Machine Learning Based on Water and Meteorological Data*** publicado en *Environmental Sciences Proceedings* disponible en linea: `https://www.mdpi.com/2673-4931/25/1/89`
- Artículo: ***Sistema de recomendación de dosis de coagulante para la planta de agua potable "El Tablazo"*** enviado a *IEEE Latin America Transactions* Pagina web: `https://latamt.ieeer9.org/`

1.6 Contenido de la monografía.

El presente trabajo de grado está compuesto por 5 capítulos los cuales se describen a continuación.

- **Capítulo 2. Estado actual del conocimiento**

  Proporciona una visión exhaustiva de la literatura y los trabajos relacionados, facilitando una comprensión más profunda y contextualizada del problema de investigación que se aborda.

- **Capítulo 3. Materiales y métodos**

  Presenta una descripción exhaustiva de los materiales y métodos empleados durante la realización de esta investigación. Se detalla de manera meticulosa cada paso seguido, proporcionando así una visión clara del proceso llevado a cabo en el estudio.

- **Capítulo 4. Resultados y discusión.**

  Se exhiben los resultados de los diferentes modelos desarrollados. Se evalúa cada uno de los modelos y se selecciona el mejor de ellos. Finalmente, se presentan los hallazgos y logros obtenidos durante la realización de la estancia de investigación.

- **Capítulo 5. Conclusiones y trabajos futuros.**

  Finalmente, se presentan las conclusiones derivadas de este estudio, proporcionando un panorama integral de los hallazgos y descubrimientos claves. Además, se establecen directrices para investigaciones futuras y desarrollos subsiguientes, aprovechando el conocimiento acumulado durante el proceso de elaboración de este trabajo de grado en modalidad de investigación.

Capítulo 2.

Estado Actual del Conocimiento.

El proyecto en cuestión se apoya en una amplia variedad de fuentes documentales que sirven como base de conocimiento para su desarrollo y orientación hacia el objetivo principal de recomendar la dosis adecuada de coagulante en la planta de tratamiento de agua potable "El Tablazo" mediante el uso de técnicas de ML. La sección 2.1 presenta los conceptos generales asociados a esta temática, mientras que la sección 2.2 expone los trabajos previos relacionados con el tema principal.

2.1 Conceptos Generales

2.1.1 Producción de agua potable

La producción de agua potable es un proceso que consta de varias etapas para asegurar la purificación y la desinfección del agua. En la planta de tratamiento "El Tablazo", se lleva a cabo mediante una serie de operaciones específicas como se puede observar en la Figura 2.1.

La primera etapa es la oxigenación del agua, en la cual se utiliza un aireador tipo cascada. Este dispositivo permite el paso de aire a través del agua, mejorando su calidad y sabor [14].

La segunda etapa es la coagulación, en la cual se utiliza una unidad de mezcla rápida y cuatro floculadores hidráulicos para facilitar la formación de flóculos. Estos se forman a partir de la adición de productos químicos y ayudan a eliminar las partículas en suspensión en el agua [14].

La tercera etapa es la decantación, la cual está representada en la Figura 2.1 por los sedimentadores. La planta de tratamiento de "El Tablazo" cuenta con cuatro sedimentadores de placas paralelas, los cuales permiten la separación de los flóculos formados en la etapa anterior. Los flóculos más pesados se asientan en el fondo de los sedimentadores, mientras que el agua clarificada se dirige a la siguiente etapa [14].

Por último, la cuarta etapa es la filtración, la cual utiliza ocho unidades de filtración rápida para remover cualquier partícula en suspensión restante. Este proceso se lleva a cabo

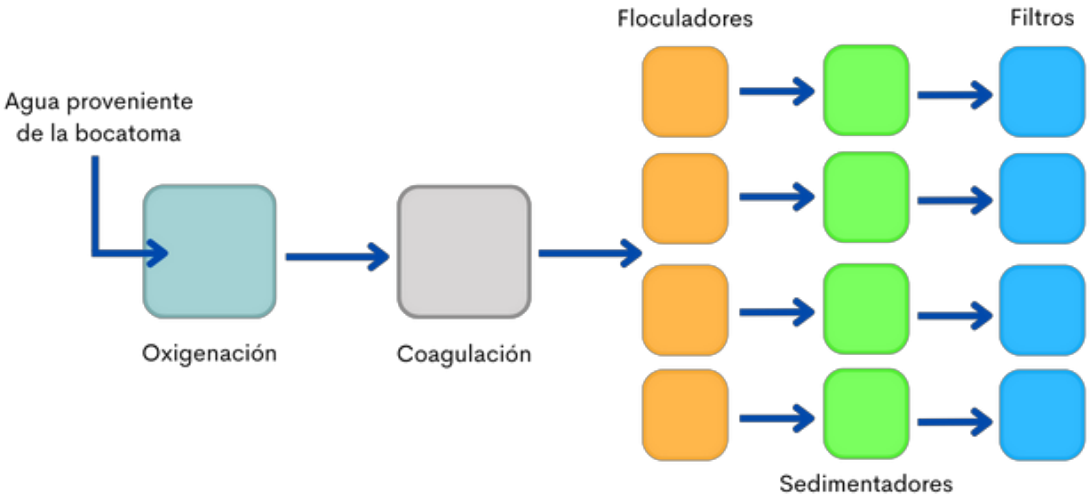

Figura 2.1: Primeras fases de potabilización de la planta de tratamiento de Agua "El Tablazo". *Fuente Propia.*

mediante diversos tipos de filtros que retienen las partículas indeseadas y permiten el paso del agua purificada [14].

2.1.2 Aplicación de dosis de coagulante en plantas de tratamiento de agua potable

El proceso de coagulación juega un papel crucial en el tratamiento del agua potable, ya que neutraliza las fuerzas electrostáticas repulsivas que mantienen las partículas coloidales dispersas en el agua. Para lograr esta neutralización, se utilizan ciertas sales de aluminio o hierro llamadas coagulantes, que contienen cationes trivalentes capaces de unirse a las cargas eléctricas negativas de las partículas. De esta manera, las partículas se agrupan y sedimentan permitiendo la formación de flóculos o grupos de sedimentos que pueden ser eliminados mediante procesos de sedimentación o filtración [15].

Por esta razón, la dosis de aplicación de coagulante es esencial en la eliminación de partículas en suspensión y por ende en la mejora de la calidad del agua potable. Cabe resaltar que la cantidad necesaria para una coagulación efectiva varía ampliamente según las características del agua y las condiciones específicas del tratamiento.

Los estudios realizados por [5] han demostrado que la dosificación del coagulante es influenciada por diversos factores, entre ellos la concentración y características de las partículas en suspensión. Por ejemplo, aguas con partículas más pequeñas pueden requerir dosis más altas de coagulante para lograr una coagulación efectiva. Asímismo, el pH del agua, la turbiedad, color y alcalinidad también pueden impactar en la dosificación del coagulante.

Adicionalmente, las investigaciones realizadas por N. Valentin, T. Denoeux y F. Fotoohi[6], revelaron que una cantidad insuficiente de coagulante puede afectar la eficiencia de la planta de tratamiento y no lograr los objetivos de calidad del agua requeridos. Por otro lado, una cantidad excesiva de coagulante puede sobrecargar el proceso de tratamiento,

disminuir la eficiencia del proceso de coagulación, provocar la formación de subproductos no deseados, aumentar los costos de tratamiento y generar preocupaciones de salud pública. En consecuencia, resulta crucial ajustar la dosis de coagulante de manera precisa para garantizar la calidad y la eficiencia del proceso de tratamiento de agua potable. La Figura 2.2 muestra una fotografía del dosificador de coagulante usado en la planta de tratamiento "El Tablazo".

Figura 2.2: Dosificador de coagulante de la planta de tratamiento de Agua "El Tablazo". *Fuente propia*

2.2 Trabajos relacionados.

El estudio de el estado actual del conocimiento en la determinación de dosis de coagulante en agua potable fue realizada a través de un mapeo sistemático. Para esto utilizamos las guía propuesta por Petersen [16] para mapeos sistemáticos con el fin de obtener una visión general del área de conocimiento e identificar posibles brechas y oportunidades de investigación. Adicionalmente, se utilizó la guía de revisión sistemática propuesta por Kitchenham [17] para identificar estudios relevantes en este campo.

Los dos trabajos guía siguen un procedimiento en cinco pasos: (A) definir la pregun-

ta de investigación, (B) realizar la búsqueda bibliográfica, (C) seleccionar los estudios relevantes, (D) clasificar los estudios seleccionados y (E) extraer y sintetizar la información. A continuación, se describen en detalle los pasos aplicados en el contexto de este estudio.

A. **Pregunta de Investigación:** El objetivo principal del presente trabajo consiste en desarrollar un sistema de recomendación para dosificación de coagulante en la planta de tratamiento "El Tablazo". Dicho sistema requiere datos históricos recolectadas sobre parámetros hidrológicos y meteorológicos analizados a través de modelos de aprendizaje automático.

   Por lo tanto, es fundamental para este trabajo conocer los principales estudios que han usado estas temáticas en el área de interés. Consecuentemente, se plantea la siguiente pregunta de investigación: ¿Cuáles son los métodos y modelos de predicción utilizados para determinar la dosis necesaria de coagulante en plantas de tratamiento de agua potable y cuál es la efectividad de estos métodos y modelos en términos de rendimiento y calidad del agua tratada?

B. **Estrategia de búsqueda:** Para realizar la revisión sistemática de este trabajo de investigación se consultaron las siguientes bases de datos de investigación científica Scopus, Web of Science y Google Scholar.

C. **Selección de Estudios:** Los criterios usados para la selección de estudios fueron los siguientes:

   Aceptación: Aquellos estudios que usen los términos "dosis de coagulante", "dosis de floculante", "dosis de aluminio", "predicción" o "estimación", "aprendizaje automático" y/o alguna técnica específica de este campo.

   Rechazo: Aquellos estudios que no usen los términos "dosis de coagulante", "dosis de floculante", "dosis de aluminio", "predicción" o "estimación", "aprendizaje automático" y/o alguna técnica específica de este campo.

   Con estos criterios de aceptación y rechazo se procede a realizar la búsqueda en las bases de datos seleccionadas usando la siguiente cadena de búsqueda.

   Cadena de búsqueda: ("coagulant dose" OR "coagulant dosage" OR "coagulation dosage" OR "coagulation dose" OR "coagulation treatment") AND ("water treatment" OR "water purification" OR "water disinfection") AND ("prediction model" OR "predictive model" OR "predictive method" OR "machine learning" OR "artificial intelligence" OR "data mining")

   A través de este proceso, se encontraron 1095 artículos. Se aplicó un filtro basado en el título y las palabras clave, lo que permitió reducir la cantidad de artículos a 376.

D. **Clasificación de Estudios:** Con el fin de optimizar la pertinencia y calidad del análisis, se llevó a cabo una revisión minuciosa de los resúmenes y conclusiones de los 376 estudios obtenidos en la fase anterior. Para esto, se seleccionaron únicamente aquellos trabajos que cumplían los criterios de selección descritos anteriormente. Este proceso riguroso de selección nos llevó a un conjunto final de 65 artículos de gran relevancia para el estudio en cuestión.

Posteriormente, se realizó una clasificación de los 65 artículos resultantes según dos criterios principales: el tipo de problema que se abordó y los algoritmos que se utilizaron para resolverlo. Este enfoque permitió agrupar los estudios de manera que se pudieran identificar tendencias y comparar metodologías en trabajos similares.

Los resultados de este análisis detallado, que incluye tanto la clasificación de los estudios como una descripción del grupo se presentan en la Tabla 2.1. Esta tabla proporciona una visión clara y estructurada de la literatura en el campo de interés, permitiendo una comprensión más fácil y rápida de los problemas abordados y las técnicas de ML utilizadas en cada uno de los estudios seleccionados. Una descripción acerca de las principales técnicas de ML se encuentra en el Anexo A.

Problema	Modelo	Estudios	Descripción
Clasificación	Redes Neuronales Artificiales (ANN, *Artificial Neural Network*)	[18]	Esta investigación se centra en una clasificación para predecir la dosis de coagulante basándose en imágenes y haciendo uso de una Red Neuronal Convolucional (CNN, *Convolutional Neural Network*).
	Sistema Experto	[19]	Esta investigación está centrada en la clasificación de datos relativos a la calidad y cantidad de agua, tanto cruda como tratada. Para ello, se lleva a cabo la implementación de un sistema experto llamado Inteligencia Artificial para la Dosificación de Sulfato de Aluminio (AIASD, *Artificial Intelligence for Aluminum Sulfate Dosing*). La validez de este sistema se confirma mediante la aplicación de la prueba de Turing.
Optimización	ANN	[20]	Esta investigación se centra en la optimización de costos en los proceso de potabilización haciendo uso de una Red Neuronal de Perceptrón Multicapa (MLP, *Multi-Layer Perceptron*)
	Metodología de Superficie de Respuesta (RSM, *Response Surface methodology*)	[21]	Esta investigación se centra en la optimización de las pruebas en el proceso de coagulación haciendo uso de RSM

Problema	Modelo	Estudios	Descripción
Optimización y Control	Bosques Aleatorios (RF, *Random Forest*)	[22]	Esta investigación está centrada en un método de control de optimización en tiempo real basado en RF y un controlador de Modo de Deslizamiento de Segundo Orden (SOSM, *Second Order Slip Mode*).
	Máquina de Vectores de Soporte (SVM, *Support Vector Machine*)	[23]	Esta investigación está centrada en un modelo de control de dosificación de coagulante basado en SVM optimizada por el algoritmo mejorado de saltos de rana mezclados (ISFLA, *Improve Shuffled Frog Leaping Algorithm*).
Optimización y Regresión	RSM	[24] y [25]	Estas investigaciones se enfocan en la optimización del proceso de coagulación, para lo cual implementan RSM. Además, se emplean otros métodos como el Diseño Compuesto Central (CCD, *Central Composite Design*) y el Análisis de Varianza (ANOVA, *Analysis of Variance*). Estos enfoques permiten explorar y comprender de manera eficaz la interacción entre las variables del proceso y las respuestas resultantes.
Regresión	Sistema Adaptativo de Inferencia Neuro-Difusa (ANFIS, *Adaptive Network-based Fuzzy Inference System*)	[26]	Esta investigación está enfocada en el proceso de aplicación de dosis de coagulante. Se utiliza un enfoque integrado que combina Sistema de Inferencia Neurodifuso Adaptativo (ANFIS, *Adaptive Neuro-Fuzzy Inference System*) con el método de agrupamiento K-medias. Esta metodología tuvo resultados prometedores, particularmente durante la temporada de lluvias.
	Sistema de Inferencia Difusa (FIS, *Fuzzy Inference System*)	[27]	Esta investigación se centra en la predicción de la dosis de coagulante mediante el uso de FIS impulsado por datos. Al comparar este sistema con otros métodos de aprendizaje automático, se comprobó que el FIS superó en rendimiento a las alternativas evaluadas.

Problema	Modelo	Estudios	Descripción
	Máquina Ligera de Reforzamiento del Gradiente (LightGBM, *Light Gradient Boosting Machine*)	[28]	Esta investigación se centra en desarrollar un sistema inteligente para la dosificación, que integra el reconocimiento de imágenes, la LightGBM y las capacidades de autoaprendizaje.
	RF	[29]	Esta investigación está centrada en la implementación y efectividad del modelo de RF, optimizado a través del Algoritmo Genético (GA, *Genetic Algorithm*), en la predicción en tiempo real de la dosis de coagulante necesaria, teniendo en cuenta las variables relacionadas con la calidad del agua.

Problema	Modelo	Estudios	Descripción
	ANN	[30, 31, 32, 33, 34, 35, 36, 6, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59]	Estas investigaciones se enfocan en problemas de regresión empleando varias arquitecturas de red neuronal. MLP es la arquitectura más prevalente, utilizada en los estudios citados en [30], [31], [32], [33], [34], [35], [44], [45] y [46]. Otras arquitecturas notables incluyen la Retropropagación Inversa (RBP, *Reverse Back-Propagation*) [37], Red Neuronal de Alimentación Hacia Adelante (FNN, *Feedforward Neural Network*) [47], Máquina de Aprendizaje Extremo (ELM, *Extreme Learning Machine*) [36], [48], arquitectura Elman [49] y la Memoria a Corto-Largo Plazo (LSTM, *Long-Short Term Memory*) [50]. La optimización de esos modelos es poco mencionada, pero estudios como [44] hace uso de el modelo No Lineal Autoregresivo con Variables Exógenas (NARX, *Nonlinear AutoRegressive eXogenous*) y el estudio [60] utiliza GA para optimizar la red neuronal. En cuanto a los coagulantes, la mayoría de los estudios emplean sulfato de aluminio y Cloruro de Polialuminio (PAC, *Polyaluminium Chloride*), aunque hubo uno que menciona la Moringa [41] como coagulante de estudio.
	GAMTF	[61]	Esta investigación está enfocada en la implementación de un modelo de aprendizaje profundo para series de tiempo multivariadas, que emplea la atención gráfica. Este modelo, que utiliza datos a largo plazo, tiene como objetivo mejorar los sistemas de apoyo a la toma de decisiones en los procesos de tratamiento de agua. La singularidad del modelo radica en su capacidad para predecir de manera simultánea tanto la dosis necesaria de coagulante como la turbidez del agua decantada, utilizando las propiedades del agua cruda como entrada.

Problema	Modelo	Estudios	Descripción
	VARIOS MODELOS	[62], [63], [64], [65], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77]	Estas investigaciones comparan distintos modelos para predecir la dosis de coagulante, destacando el uso frecuente de ANN. MLP se ha aplicado en gran medida [62], [66], [68], [70], [71], [73], [74], [75], [76], [77]. Otras arquitecturas incluyen la Retropropagación (BP, *Back-propagation*) [64], FNN [68], [75], Alimentación Hacia Adelante en Cascada (CF, *Cascading Feed-forward*), Base Radial (RB, *Radial Basis*) [68], [75], [77] y la LSTM [74]. Además, se han incorporado modelos para comparar, entre los que destacan ANFIS [62], [63], [66], [77], Regresión [64], [71], [72], RSM [69], SVM [69], [74], Árboles Extremadamente Aleatorios (ERT *Extremely Randomized Trees*) [70], Regresión Lineal Múltiple (MLR, *Multiple Linear Regression*) y Regresión de Mínimos Cuadrados Parciales (PLSR, *Partial Least Squares Regression*) [71]. El Alumbre es el coagulante más frecuentemente utilizado, aunque otros como PAC [62] y Polialuminio Cloruro Hidrosilicato (PAHCS, *Polialuminium Chloride Hydroxide Silicate*) [77] también se mencionan en los estudios.

Tabla 2.1: Clasificación de los trabajos relacionados. *Fuente propia*

Adicionalmente en estos 5 problemas destacan las siguientes investigaciones:

- La investigación [18] abordó la optimización del proceso de coagulación en el tratamiento del agua mediante el uso de CNN, específicamente enfocándose en un problema de clasificación relacionado con el análisis de imágenes. Los investigadores aplicaron una CNN para predecir el rendimiento de las pruebas de jarra, un método tradicionalmente empleado en el estudio de la coagulación. Utilizaron agua artificial y grabaron imágenes del floc durante las pruebas, construyendo modelos para predecir la turbidez de los sobrenadantes a partir de estas imágenes y niveles de turbidez.

  Durante el aprendizaje, todos los modelos alcanzaron una precisión del 100 %, indicando una eficacia adecuada de la CNN para la extracción de características de las imágenes de floc. Al examinar el sobreajuste mediante datos de prueba, la precisión de predicción de todos los modelos superó el 96 %. Al

aplicar el modelo a muestras de agua natural, la precisión bajó pero aún se mantuvo alta (90 %).

Además, los investigadores descubrieron que las imágenes capturadas durante la mezcla rápida (primeros 100 segundos) proporcionaron suficiente información para garantizar la confiabilidad del modelo. Por último, sugieren que se deben recoger más datos de muestras de agua de diversas fuentes y condiciones para mejorar la robustez del modelo.

- El estudio [21] exploró la optimización de la coagulación en el tratamiento del agua utilizando RSM y CCD. Los experimentos evaluaron la remoción de turbidez y Carbono Orgánico Total (TOC, *Total Organic Carbon*) en agua, con alumbre como coagulante y el pH ajustado antes de su agregación. Se realizaron 13 pruebas con dos factores (dosis de alumbre y pH de coagulación) y se desarrollaron modelos matemáticos de regresión cuadrática para predecir la eficiencia de remoción. Los modelos, que tenían coeficientes de determinación (R2) superiores al 90 %, predecían la máxima remoción de turbidez y TOC a diferentes pH y dosis de alumbre. La optimización del proceso reveló que una remoción simultánea del 92 % de turbidez y 39,5 % de TOC se podía lograr con una dosis de alumbre de 44 mg/L y un pH de 7,6.

  El estudio también informa sobre un método alternativo para predecir las dosis óptimas de alumbre y la calidad del agua tratada utilizando modelos ANN en el sur de Australia. Los modelos ANN demostraron un alto rendimiento, con valores R2 entre 0,90 y 0,98 para los modelos que predicen turbidez, color y absorbancia ultravioleta en el agua tratada, y un valor R2 de 0,94 para el modelo que predice las dosis óptimas de alumbre. Estos modelos fueron implementados en dos herramientas de simulación que permiten a los operadores controlar las tasas de dosificación de alumbre automáticamente y en tiempo real.

  Estos enfoques pueden facilitar la optimización del tratamiento de agua, garantizando la calidad y reduciendo los costos y riesgos para la salud.

- El trabajo [22] abordó el desafío de optimizar y controlar el proceso de coagulación en el tratamiento de agua potable. El control convencional Proporcional Integral Derivado (PID, *Proportional, Integral, Derivative*) ha demostrado ser insuficiente debido a las frecuentes fluctuaciones en la calidad del agua cruda y al gran retardo que caracteriza este proceso.

  Para superar estos desafíos, los investigadores proponen un enfoque de control compuesto que combina el algoritmo RF y un controlador SOSM. Esta combinación permite un ajuste más preciso y oportuno de la dosis de coagulante en respuesta a los cambios en la calidad del agua.

  El algoritmo de RF se utiliza para proporcionar un control anticipado, ayudando a predecir los cambios en la calidad del agua y ajustar la dosis de coagulante de manera proactiva. Por otro lado, el controlador SOSM funciona como un controlador de retroalimentación, ajustando la dosis de coagulante para adaptarse a la no linealidad y la incertidumbre del proceso de coagulación.

Las pruebas experimentales demuestran que este enfoque de control compuesto supera al control PID convencional y otros métodos en términos de adaptabilidad a los cambios en la calidad del agua y precisión del control para una turbidez estable del efluente.

Además, este sistema de control optimizado puede resultar en un ahorro significativo en la dosis de coagulante, ya que permite un control más preciso de la turbidez del efluente. De este modo, la investigación aborda eficazmente el problema de optimización y control en el proceso de coagulación del tratamiento de agua potable.

- La investigación [25] se centró en la optimización y regresión del proceso de coagulación-floculación en el tratamiento de agua potable. Se utiliza RSM en lugar de los métodos convencionales de pruebas de jarras para determinar las dosis óptimas de coagulante y pH.

  El estudio busca maximizar la eficiencia de eliminación de turbidez y carbono orgánico disuelto (DOC, *Dissolved Organic Carbon*) mediante la comparación de diferentes coagulantes y la aplicación del RSM. Se desarrollan modelos cuadráticos para las respuestas de turbidez y DOC, y se encuentran las condiciones óptimas para cada una de ellas.

  Los resultados muestran que la combinación óptima de dosis de coagulante y pH para la máxima eliminación de turbidez y DOC es de 0.11 mM de PAC a pH 7.4 y 0.15 mM de alumbre a pH 6.6. Estas condiciones logran una eliminación de turbidez del 91.4 % y una eliminación de DOC del 31.2 % con PAC, y del 86.3 % y 34.3 % respectivamente con alumbre.

  El estudio demuestra que el RSM es efectivo para obtener predicciones precisas y optimizar el proceso de coagulación-floculación. Los resultados experimentales coinciden con las predicciones del modelo, lo que demuestra los beneficios de este enfoque en términos de eficiencia y reducción del número de experimentos necesarios.

- El trabajo [63] se enfocó en el problema de regresión en el proceso de coagulación para la producción de agua potable. La dosificación de coagulantes tiene una relación no lineal con las características del agua cruda, lo que dificulta su control mediante métodos convencionales. Para superar estas limitaciones, se propone utilizar ANFIS para modelar la dosificación de coagulantes en una planta de tratamiento de agua potable en Boudouaou, Argelia.

  Se utilizan seis variables de calidad del agua cruda y la dosis de alumbre para construir el modelo de dosificación de coagulante. Se comparan dos sistemas neurodifusos basados en ANFIS: ANFIS-GRID y ANFIS-SUB.

  Los resultados muestran que el modelo ANFIS-SUB tiene un menor error cuadrático medio y un mayor coeficiente de correlación en comparación con el modelo ANFIS-GRID. Esto indica que ANFIS-SUB supera a ANFIS-GRID en términos de precisión y capacidad para resolver el problema en cuestión.

  El estudio destaca la importancia de modelar sistemas en el tratamiento de agua potable y la dificultad de controlar y predecir los procesos de coagula-

ción. El enfoque ANFIS proporciona una solución efectiva al modelar la dosificación de coagulante, lo que permite una mejor comprensión y predicción del comportamiento del sistema. El modelo ANFIS-SUB, en particular, muestra un rendimiento superior en la predicción de la dosificación de coagulante.

En conclusión, este estudio demuestra que el enfoque ANFIS es eficaz para modelar y predecir la dosificación de coagulante en el tratamiento de agua potable. Los resultados indican que el modelo ANFIS-SUB es más confiable y preciso que el ANFIS-GRID. Este enfoque ofrece beneficios significativos al reemplazar el método convencional de prueba de jarras, al ser más rápido, rentable y capaz de aplicarse en tiempo real.

2.2.1 Resultados y Análisis

En esta sección, se presentan los resultados de los pasos A, B, C y D del mapeo y revisión sistemática de la literatura, tal como se detalló previamente. Además, se incluye el paso E de este proceso, que tiene como objetivo resaltar los resultados obtenidos y realizar un análisis detallado de los mismos.

E. **Extracción de Datos y Síntesis:** Se puede observar que el problema de clasificación fue el que aportó la menor cantidad de estudios en el proceso previamente realizado. Sin embargo, sus contribuciones son muy valiosas. En cuanto al problema de optimización, este ocupa el siguiente lugar en términos de escasez de estudios y se encontró entrelazado con otros tipos de problemas, como los de control y regresión. Por otro lado, el problema de regresión fue el que aportó la mayor cantidad de estudios, evidenciando una amplia diversidad de modelos y técnicas explorados en dichos trabajos.

2.2.2 Brechas y Aportes

En esta sección se presentan las brechas localizadas en la revisión de las investigaciones seleccionadas.

Artículo	Brechas
[18]	Se llevó a cabo el estudio dentro de un ambiente controlado, centrando la atención en un problema de clasificación vinculado con el análisis de imágenes, con el objetivo de optimizar el proceso de coagulación. A pesar de que se consideran imágenes y datos derivados de pruebas de jarra, el estudio no incorpora la variabilidad estacional y climática. Esto significa que no se evalúan los cambios en las características del agua a lo largo del año que pueden ser provocados por factores estacionales o climáticos.

Artículo	Brechas
[21]	Este estudio se concentra en la optimización del proceso de coagulación a través de experimentos con agua sintética. Sin embargo, hay que destacar que dicha metodología podría no reflejar de manera precisa las variaciones y complejidades inherentes al agua natural, sobre todo considerando la influencia de diversos contaminantes y condiciones ambientales. Además, al igual que el problema anterior, la falta de consideración de la variabilidad estacional y climática en las características del agua puede limitar la aplicabilidad y precisión de los resultados en condiciones reales.
[22]	El estudio aborda de manera integral la optimización y control del proceso de coagulación en el tratamiento de agua, poniendo un especial énfasis en el uso del algoritmo RF y concentrándose en la regulación de la turbidez del agua a través del ajuste de la dosificación del coagulante. No obstante, sería beneficioso explorar y evaluar otras técnicas de aprendizaje automático para llevar a cabo esta tarea. La comparación de resultados obtenidos con distintos métodos podría conducir a la creación de un sistema más robusto y eficiente. Adicionalmente, es posible que existan otras variables dentro del proceso de tratamiento del agua que, al ser consideradas, podrían mejorar aún más la eficacia y eficiencia del proceso de tratamiento.
[25]	El enfoque de este estudio radica en la optimización del proceso de coagulación-floculación mediante el uso de RSM, con especial atención en las variables de control, que son la dosis de coagulante y el pH. No obstante, es importante resaltar que en la práctica, existen otras variables que pueden tener una influencia considerable en el proceso de coagulación-floculación. Entre estas variables se incluyen aspectos inherentes al agua cruda y factores meteorológicos, los cuales, no han sido considerados en el estudio. Además, se puede ampliar el horizonte de investigación al explorar la aplicación de algoritmos de ML para optimizar aún más el proceso.
[63]	El estudio consiguió evidenciar que la implementación de modelos ANFIS en plantas de tratamiento de agua potable tiene beneficios considerables, en particular para determinar la dosificación óptima de coagulante basándose en seis variables de calidad del agua en línea. Sin embargo, este estudio no consideró variables meteorológicas, las cuales pueden reflejar variaciones estacionales o climáticas que podrían influir en las propiedades del agua. Además, se puede fortalecer aún más la validez de este estudio comparando los resultados obtenidos con ANFIS con los de otros tipos de modelos.

Tabla 2.2: Brechas de las investigaciones destacadas. *Fuente propia.*

El presente trabajo de grado, se enfoca en la brecha relacionada con la incorporación de variables meteorológicas que tienen el potencial de influir en las características del agua. Al considerar estas variables, el estudio integra la variabilidad climática local, esforzándose por desarrollar un modelo que refleje condiciones

más realistas y pertinentes.

Este estudio no sólo se distingue por la inclusión de variables meteorológicas, sino también por su enfoque dual: compara varios modelos desde las perspectivas de regresión y clasificación. Aunque estos enfoques procesan los mismos datos desde perspectivas distintas, comparten el mismo objetivo de recomendar la dosis de coagulante necesaria teniendo en cuenta ciertas características del agua y el lugar de estudio.

La singularidad del conjunto de datos reside en su estructura, en la cual la mayoría de los valores se distribuyen en intervalos de cinco unidades. Esta disposición es idónea para convertir la variable continua en una categórica, permitiendo así el análisis desde una perspectiva clasificatoria sin comprometer su esencia numérica. Además, el empleo de ambos enfoques sobre el mismo conjunto de datos introduce un elemento de validación cruzada. La consistencia o complementariedad entre los modelos de regresión y clasificación refuerza la confianza en las conclusiones obtenidas.

Con esto, no sólo se estaría ofreciendo un enfoque adicional para el estudio, sino también abriendo caminos para futuras investigaciones que puedan beneficiarse de combinar estas técnicas.

2.3 Resumen

Este capítulo abordó la explicación de los fundamentos teóricos relacionados con los procedimientos de la planta de tratamiento, enfocandose en la etapa de coagulación y la dosificación de coagulante. Para entender el estado actual del conocimiento acerca de la determinación de dosis de coagulante, se llevaron a cabo un mapeo sistemático y una revisión exhaustiva, fundamentándose en las directrices establecidas por Petersen y Kitchenham. Estas revisiones se emprendieron para analizar trabajos cruciales vinculados al problema de investigación delineado en el Capítulo 1.

El procedimiento del mapeo sistemático engloba etapas como la definición de la pregunta de investigación, la realización de la búsqueda bibliográfica, seleccionar los estudios relevantes, clasificarlos y sintetizar la información obtenida. Como resultado, se identificaron 65 artículos relevantes para el estudio, clasificados según el tipo de problema abordado que para estos estudios incluyó la regresión, clasificación, y optimización; y se identificaron los principales algoritmos de ML utilizados en cada uno de los estudios. Finalmente, se identifican brechas en la literatura revisada.

El presente trabajo de grado se distingue por abordar la inclusión de variables meteorológicas en el proceso de dosificación del coagulante, reconociendo la influencia que puede tener el clima en las características del agua. Además, se destaca el enfoque dual del estudio, que compara modelos desde perspectivas de regresión y clasificación, buscando ofrecer una recomendación precisa y adaptada a las condiciones locales.

Capítulo 3.

Materiales y Métodos

Este capítulo describe el método empleado para desarrollar el sistema de recomendación de dosis de coagulante para la planta de agua potable "El Tablazo". Inicialmente, se destaca la adopción del Proceso Estándar de la Industria Cruzada para el Aprendizaje Automático (CRISP-ML(Q), *Cross Industry Standard Process for Machine Learning*), como la estructura fundamental que orienta las fases de este proyecto. Una descripción meticulosa de este enfoque se ofrece en la sección 3.1, detallando cada una de las etapas que prescribe este marco de trabajo.

Posteriormente, en la sección 3.2, se expone cómo se aplicaron concretamente los principios y etapas de CRISP-ML(Q) en la construcción del mencionado sistema de recomendación. Esta sección brinda una visión detallada y específica de la ejecución de cada fase dentro del contexto del proyecto.

3.1 Marco de Trabajo CRISP-ML(Q).

El marco de trabajo CRISP-ML(Q), como se detalla en el artículo citado [78], se propone como un modelo de proceso estándar para el desarrollo de aplicaciones de ML, destacando su compatibilidad con el Proceso Estándar de la Industria Cruzada para la Minería de Datos (CRISP-DM, *Cross Industry Standard Process for Data Mining*). Este modelo integra la garantía de calidad en cada fase y tarea, con el objetivo de minimizar los riesgos que puedan afectar la eficacia y el éxito de la aplicación de ML.

Una de las principales innovaciones de CRISP-ML(Q) es la inclusión de una fase de monitoreo y mantenimiento para atender los riesgos de deterioro del modelo en entornos dinámicos, lo que amplía el alcance del modelo de proceso en comparación con CRISP-DM. Además, combina la comprensión del negocio y de los datos en una sola fase, dada la interrelación entre estas dos actividades.

Siguiendo las indicaciones del marco CRISP-ML(Q), en cada fase se aplica el aseguramiento de calidad para verificar el cumplimiento de los estándares actuales y las mejores prácticas de la industria y la academia. En este sentido, se definen los objetivos de una fase genérica, se instancian pasos y tareas específicas y se identifican los riesgos

potenciales. La Figura 3.1 detalla el procedimiento de aseguramiento de calidad.

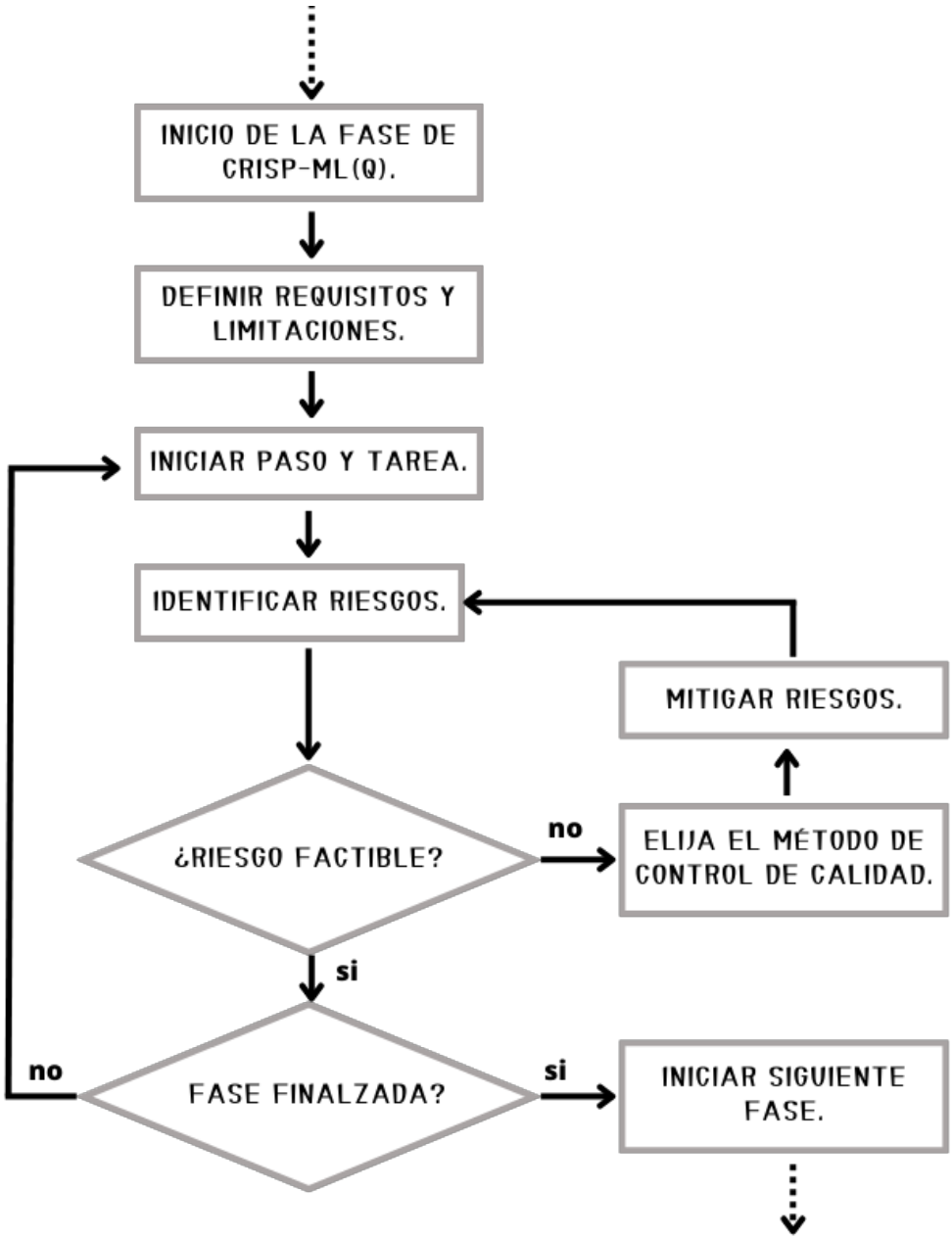

Figura 3.1: Metodología de aseguramiento de calidad (Figura Adaptada de [1].

El marco de trabajo CRISP ML(Q), tal y como se detalla en [78], se centra en mejorar el éxito y la eficiencia de las aplicaciones de ML, proporcionando una guía a lo largo del ciclo de vida de dichas aplicaciones. La Figura 3.2 muestra las seis etapas primordiales de este marco de trabajo y se detallan a continuación.

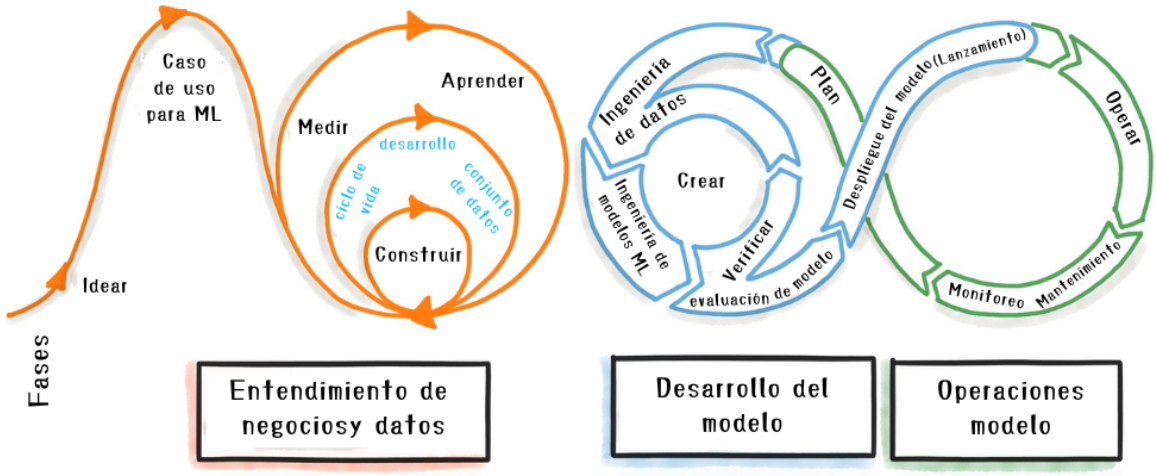

Figura 3.2: Proceso de ciclo de vida de desarrollo de aprendizaje automático. (Figura tomada de [1]

- **Entendimiento de Negocios y Datos:** La primera fase implica entender el negocio y los datos, incluyendo definición del alcance, establecimiento de criterios de éxito, verificación de factibilidad, recolección y revisión de calidad de los datos.
- **Preparación de los Datos:** Este proceso consiste en seleccionar, limpiar, construir y estandarizar los datos para que estén aptos para la fase de modelado. Este proceso puede ser dinámico y necesitar ajustes basados en hallazgos de las etapas de modelado o implementación.
- **Modelado:** Crucial para el éxito, esta etapa comprende la revisión de literatura, selección de algoritmos, entrenamiento del modelo, compresión del modelo y optimización.
- **Evaluación:** Esta fase incluye validación de rendimiento, determinación de robustez, incremento de explicabilidad y comparación de resultados con los criterios de éxito. Si los criterios no se cumplen, se puede retroceder a etapas anteriores o detener el proyecto.
- **Despliegue:** En esta fase, se consideran la elección de hardware, evaluación bajo condiciones de producción, garantía de aceptación del usuario y usabilidad, minimización de riesgos de errores inesperados y estrategia de despliegue.
- **Monitorización y Mantenimiento:** Finalmente, esta etapa requiere mantenimiento del modelo para prevenir degradación del rendimiento, monitoreo del modelo, actualización del modelo, control de actualizaciones y implementación de una estrategia de despliegue.

En la implementación de este marco de trabajo, se abordó hasta la fase de evaluación. Las etapas de despliegue y mantenimiento están fuera del alcance del presente trabajo de grado. No obstante, se planea como trabajo futuro para continuar este trabajo en la planta de tratamiento "El Tablazo", quienes están interesados en realizar las dos fases restantes.

3.2 Implementación de CRISP-ML(Q).

En la presente sección, se detalla la aplicación de CRISP-ML(Q), expuesta en el segmento previo, con el objetivo de desarrollar el sistema de recomendación de dosis de coagulante para la planta de tratamiento "El Tablazo".

3.2.1 Alcance de la Aplicación de ML

El alcance de este proyecto es aplicar técnicas de ML para construir un sistema de recomendación. Este sistema, sustentado en un modelo predictivo, determina de manera precisa y eficaz la cantidad necesaria de coagulante para el tratamiento de agua en la planta de "El Tablazo", considerando variables hidrológicas y meteorológicas relevantes. A través de este proceso, se evaluaron varios modelos y se seleccionó el que ofrecía el mejor rendimiento, culminando en la entrega y evaluación de los resultados del sistema de recomendación. Es importante mencionar que el despliegue del sistema y la creación de un panel de visualización quedaron fuera del alcance de este trabajo.

Criterios de Éxito

- **Criterios de Éxito del Negocio:** El propósito de implementar un sistema de recomendación para la dosificación de coagulante radica en disminuir la cantidad de pruebas de jarra requeridas para hallar la dosis adecuada. Al lograrlo, no sólo se optimizarán los tiempos de operación sino que también se prevendrá la aplicación excesiva de coagulante para reducir los costos de producción.

  Es importante señalar que el alcance del presente trabajo de grado no considera la evaluación de la eficacia de este sistema en producción. No obstante, en el marco de este proyecto, se ha diseñado el sistema de recomendación y un tablero de control del cual se encuentra ofreciendo orientación al acueducto para el despliegue exitoso del sistema en sus instalaciones.

- **Criterios de Éxito de ML:** Para los modelos que tratan el problema desde una perspectiva de clasificación, el criterio de éxito se establece en incrementar la precisión, la exhaustividad y el puntaje f1 de aquellas clases individuales que se encuentren por debajo del promedio. El rendimiento de estos modelos es posteriormente comparado con el del modelo base, es decir, sin procesar previamente.

  Por su parte, aquellos modelos que tratan el problema desde un enfoque de regresión, su criterio de éxito se basa en obtener un valor de Error Cuadrático Medio (MSE, *Mean Squared Error*), Raiz del Error Cuadrático Medio (RMSE, *Root Mean Squared Error*) y Error Absoluto Medio (MAE, *Mean Absolute Error*) que no exceda el 5 % de la media de la variable objetivo en el conjunto de prueba. Esta meta propone un límite superior aceptable para el error de las predicciones, garantizando un rendimiento satisfactorio del modelo en términos de precisión.

Factibilidad.

El logro de los objetivos planteados es factible con los recursos actuales. En cuanto a la obtención de datos, la planta de tratamiento "El Tablazo" ha proporcionado un conjunto de datos hidrológicos amplio y el Instituto de Estudios de Hidrología, Meteorología y Medio Ambiente IDEAM ofrece datos meteorológicos de libre acceso en su plataforma web. Estos datos son fundamentales para entrenar los modelos de ML que usamos en este proyecto.

Respecto a las herramientas de software, el desarrollo se llevó a cabo a través de cuadernos de Jupyter, la cual es una plataforma código abierto de escritura de scripts principalmente en python. Esto aseguró que el proyecto se pudiera realizar sin enfrentar gastos extras relacionados con licencias de software.

3.2.2 Fuentes de Datos.

Los datos utilizados en este proyecto provienen de dos fuentes distintas: la planta de tratamiento "El Tablazo" y el Instituto de Estudios de Hidrología, Meteorología y Medio Ambiente IDEAM.

Respecto a "El Tablazo", se recibieron 120 hojas de cálculo en distintos formatos (.xls y .xlsm). Estas hojas de cálculo se encuentran almacenadas en 10 carpetas, cada una de ellas lleva por nombre el año al que corresponden sus registros internos. Dentro de cada carpeta, existen 12 hojas de cálculo que representan los meses del año respectivo, y cada una de ellas lleva por nombre el mes que abarca. De este modo, se cuenta con informes mensuales desde enero de 2013 hasta diciembre de 2022.

Las hojas de cálculo contienen una gama diversa de datos relativos a variables hidrológicas tanto de agua cruda como de agua tratada. Entre estos datos se incluyen mediciones de caudal, turbiedad, color, pH, por nombrar solo algunos. Adicionalmente, estos archivos también proveen información sobre las cantidades específicas de diversos productos químicos que se han aplicado al agua durante su proceso de potabilización, tales como coagulantes, cal y cloro. Las hojas de cálculo no son compartidos en la presente monografía por cuestiones de confidencialidad ya que pertenecen al acueducto de la ciudad de Popayán y requieren de su autorización para publicarse.

En relación a los datos meteorológicos, se ha recurrido a datos de libre acceso proveídos a través del banco de datos hidrometeorológicos del IDEAM [1].

En el presente trabajo de grado, se seleccionaron los datos de las estaciones "EL TABLAZO - AUT [26015010]" y "AEROPUERTO G L VALENCIA [26035030]", abarcando el periodo desde el año 2013 hasta el 2022. En el conjunto de datos obtenido, se notó que algunas variables no disponen de información continua desde el inicio del periodo, sino que comienzan a finales del año 2016. Esta situación se presentó particularmente en el caso de la variable de precipitación. Además, cabe destacar que la disponibilidad de datos para todas las variables solo se extiende hasta octubre de 2022.

[1] El banco de datos hidrometeorológicos del IDEAM está disponbile a través del enlace: `http://dhime.ideam.gov.co/webgis/home/`

3.2.3 Construcción del Conjunto de Datos.

Como primer paso, se llevó a cabo una exhaustiva revisión de la calidad de los datos recolectados de la planta de tratamiento "El Tablazo". Este proceso es fundamental para asegurar la fiabilidad y validez de los análisis posteriores. En este sentido, se realizó un análisis meticuloso de las características inherentes a estos datos.

Se efectuó una revisión detallada de los nombres asignados a las columnas en cada hoja de cálculo, con el propósito de garantizar su coherencia y uniformidad. Cuando se detectaron errores ortográficos, se corrigieron para garantizar que cada columna se identificara de manera correcta y coherente.

Además, se verificaron los nombres de las hojas en cada archivo. En caso de detectarse errores ortográficos, estos se corrigieron para garantizar que cada hoja, correspondiente a un día del mes, se identificara simplemente por su número.

Una vez finalizada la edición y organización de los archivos, se llevó a cabo una lectura y verificación de la integridad de los datos a través de un cuaderno de Jupyter. Durante esta revisión, se identificaron y corrigieron errores de digitación, como números mal escritos, caracteres extraños, y otros errores que pudieran afectar la calidad y precisión de los datos. A partir de esta revisión, se generó un conjunto de datos inicial. El código fuente desarrollado en este proyecto para este proceso se encuentra disponible en el repositorio de GitHub[2]

Analisis Exploratorio de los Datos (EDA, *Exploratory Data Analysis*)

La fase de análisis exploratorio de los datos sirvió como una primera aproximación para comprender la estructura y calidad de la información con la que se contaba. Esta etapa inicial incluyó acciones críticas para garantizar la integridad de los datos subsiguientes. Para ello, se llevaron a cabo las siguientes etapas: identificación del tipo de variables y análisis de datos faltantes.

- **Identificación de Variables:** La Tabla 3.1 muestra una descripción de cada una de las variables con las que cuenta el conjunto de datos creado anteriormente. Esto determina cómo realizar el análisis de los datos según sea su tipo.

Variable	Descripción	Tipo
COAGULANTE_DOSIS	Dosis de coagulante aplicada al agua. Esta se mide en mg/L	flotante
FECHA	Fecha de cuando el dato fue tomado.	fecha
HORA	Hora de cuando el dato fue tomado.	entero
CAUDAL	Cantidad de agua de entrada en L/s.	flotante
CAL_1RIA_KG	Cantidad de cal (hidróxido de calcio) agregada en el equipo de aplicación.	flotante
CAL_1RIA_DOSIS	Dosis de cal antes de aplicar un tratamiento, en mg/L.	flotante

[2]Código fuente para la creación del conjunto de datos disponible en: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/02_Dataframe_Restructuration.ipyn`

AGUA_CRUDA_P.H	p.H del agua antes de aplicar un tratamiento.	flotante
AGUA_CRUDA_COLOR	Color del agua antes de aplicar un tratamiento. Se mide en Unidades de Platino Cobalto (PCU, *Platinum-Cobalt Units*)	flotante
AGUA_CRUDA_NTU	Turbiedad del agua antes de aplicar un tratamiento. Se mide en Unidad nefelométrica de turbiedad (NTU, *Nephelometric Turbidity Unit*)	flotante.
AGUA_CRUDA_ALCALINIDAD	Alcalinidad del agua antes de aplicar un tratamiento. Se mide en mg/L de carbonato de calcio.	flotante
AGUA_CRUDA_CONDUCTIVIDAD	Conductividad del agua antes de aplicar un tratamiento. Se mide en µS/cm.	flotante
COAGULANTE_GRANULADO	Cantidad de Kilogramos de coagulante granulado agregado en el equipo de aplicación.	flotante
COAGULANTE_LIQUIDO	Cantidad de mg/L de coagulante líquido agregado en el equipo de aplicación.	flotante
AGUA_TRATADA_CLORO	Cloro después de aplicar un tratamiento. Se mide en mg/L.	flotante
AGUA_TRATADA_ALCALINIDAD	Alcalinidad del agua después de aplicar un tratamiento. Se mide en mg/L de carbonato de calcio.	flotante
AGUA_TRATADA_P.H	p.H del agua después de aplicar un tratamiento.	flotante
AGUA_TRATADA_COLOR	Color del agua después de aplicar un tratamiento. Se mide en PCU	flotante
AGUA_TRATADA_NTU	Turbiedad del agua después de aplicar un tratamiento. Se mide en NTU	flotante

Tabla 3.1: Conjunto de datos inicial. *Fuente Propia.*

- **Análisis de Datos Faltantes:** Se verificó la existencia de datos nulos en el conjunto de datos, buscando valores como "null", "NaN" (Not a Number) o incluso espacios en blanco. Con el apoyo de la librería `Pandas` de Python, se determinó la cantidad y el tipo de dato de cada variable, así como la presencia de datos nulos y el volumen total de registros en el conjunto de datos. La Figura 3.3 muestra que el conjunto consta de 87648 registros, que se extienden desde el índice 0 hasta el índice 87647. Este contiene 18 variables, con diferentes cantidades de datos nulos en las variables. En cuanto a los tipos de datos, se identificaron float64 e int64 para los valores numéricos y datetime64 para valores de tiempo.

  Se elaboró una tabla para detallar el porcentaje de datos nulos de cada variable,

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 87648 entries, 0 to 87647
Data columns (total 18 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   FECHA                     87648 non-null  datetime64[ns]
 1   HORA                      87648 non-null  int64
 2   CAUDAL                    87380 non-null  float64
 3   CAL_1RIA_KG               29 non-null     float64
 4   CAL_1RIA_DOSIS            587 non-null    float64
 5   AGUA_CRUDA_P.H            86723 non-null  float64
 6   AGUA_CRUDA_COLOR          81227 non-null  float64
 7   AGUA_CRUDA_NTU            86984 non-null  float64
 8   AGUA_CRUDA_ALCALINIDAD    68372 non-null  float64
 9   AGUA_CRUDA_CONDUCTIVIDAD  60089 non-null  float64
 10  COAGULANTE_GRANULADO      5897 non-null   float64
 11  COAGULANTE_LIQUIDO        135 non-null    float64
 12  COAGULANTE_DOSIS          44393 non-null  float64
 13  AGUA_TRATADA_CLORO        77295 non-null  float64
 14  AGUA_TRATADA_ALCALINIDAD  36156 non-null  float64
 15  AGUA_TRATADA_P.H          84606 non-null  float64
 16  AGUA_TRATADA_COLOR        78531 non-null  float64
 17  AGUA_TRATADA_NTU          84771 non-null  float64
dtypes: datetime64[ns](1), float64(16), int64(1)
memory usage: 12.7 MB
```

Figura 3.3: Información del conjunto de datos. *Fuente propia.*

facilitando su visualización. Esta tabla se muestra en la Figura 3.4. Es significativo destacar que la variable objetivo, `COAGULANTE_DOSIS`, presentaba un 49.35 % de ausencia de datos, lo que llevó a una indagación profunda para identificar el motivo de estos valores nulos. A través de una entrevista con el jefe de la División de Producción del acueducto de Popayán, Ingeniero Mauricio Ramírez, se intentó esclarecer esta situación. Los detalles completos de la entrevista se hallan en el Anexo B.

El Ingeniero Ramírez explicó un detalle crucial: cuando la turbiedad del agua en la planta de tratamiento "El Tablazo", es menor a 3, tradicionalmente no se aplica coagulante. Esta decisión proviene de la experiencia práctica de los operadores, y por lo tanto, en estas situaciones, la dosis de coagulante no se anota en las hojas de cálculo, resultando en la ausencia de datos. En términos simples, no se añade coagulante y, por lo tanto, su valor equivaldría a cero. Además, se incorporó una nueva directriz desde 2021 que modifica esta práctica, estableciendo que se debe añadir coagulante al agua entre las 8 de la noche y las 8 de la mañana, sin importar la turbiedad, para mantener una calidad constante del agua.

Dada la información proporcionada, la técnica de imputación de datos fue implementada en el proyecto de grado en cuestión. El objetivo de la imputación de datos es reemplazar valores ausentes con estimaciones relevantes [79]. En este contexto, se optó por imputar valores de cero para la dosis de coagulante en situaciones en las que no se añadió coagulante debido a la baja turbiedad del agua, basándose directamente en el testimonio del Ingeniero Ramírez.

Para registros previos al año 2021, se imputaron ceros en las dosis de coagulante cuando la turbiedad era inferior a 3. Durante los años 2021 y 2022, se adoptó un enfoque similar, ajustándose al horario de 8 de la mañana a 8 de la noche,

	Variables	% de nulos	N de nulos	N de NO nulos
0	CAL_1RIA_KG	99.966913	87619	29
1	COAGULANTE_LIQUIDO	99.845975	87513	135
2	CAL_1RIA_DOSIS	99.330276	87061	587
3	COAGULANTE_GRANULADO	93.271951	81751	5897
4	AGUA_TRATADA_ALCALINIDAD	58.748631	51492	36156
5	COAGULANTE_DOSIS	49.350812	43255	44393
6	AGUA_CRUDA_CONDUCTIVIDAD	31.442817	27559	60089
7	AGUA_CRUDA_ALCALINIDAD	21.992516	19276	68372
8	AGUA_TRATADA_CLORO	11.812021	10353	77295
9	AGUA_TRATADA_COLOR	10.401835	9117	78531
10	AGUA_CRUDA_COLOR	7.325894	6421	81227
11	AGUA_TRATADA_P.H	3.470701	3042	84606
12	AGUA_TRATADA_NTU	3.282448	2877	84771
13	AGUA_CRUDA_P.H	1.055358	925	86723
14	AGUA_CRUDA_NTU	0.757576	664	86984
15	CAUDAL	0.305769	268	87380
16	HORA	0.000000	0	87648
17	FECHA	0.000000	0	87648

Figura 3.4: Porcentaje de nulos de variables del conjunto de datos. *Fuente propia.*

siguiendo la nueva política de la planta de tratamiento. El código fuente para este proceso de imputación está disponible en el repositorio de GitHub[3].

Gracias a este método, se mantuvo la integridad del conjunto de datos, reflejando fielmente las operaciones de la planta. El resultado fue un conjunto de datos coherente, en el cual los valores no registrados o ausentes se reemplazaron de acuerdo con las directrices de la planta. Es importante resaltar que, tras la reestructuración, se identificó una presencia marcada de ceros en la variable de dosis de coagulante, como se puede observar en la Figura 3.5. Este valor se convirtió en el más frecuente en todo el conjunto.

EDA comparativo

La situación descrita en la sección anterior fue abordada realizando utilizando dos conjuntos de datos, uno incluyendo los datos con las imputaciones de ceros, y otro sin las imputaciones.

A continuación, se llevó a cabo un EDA comparativo, el cual comprendió tres procedimientos: revisión de datos faltantes en ambos conjuntos, análisis univariable con la dosis de coagulante como variable principal, y análisis multivariable.

Como resultado de la imputación de ceros, se observó una notable reducción de datos nulos en la variable `COAGULANTE_DOSIS`, descendiendo de un 49.35 % a un 7.04 %, tal

[3]Código fuente en: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/02_Dataframe_Restructuration.ipynb`

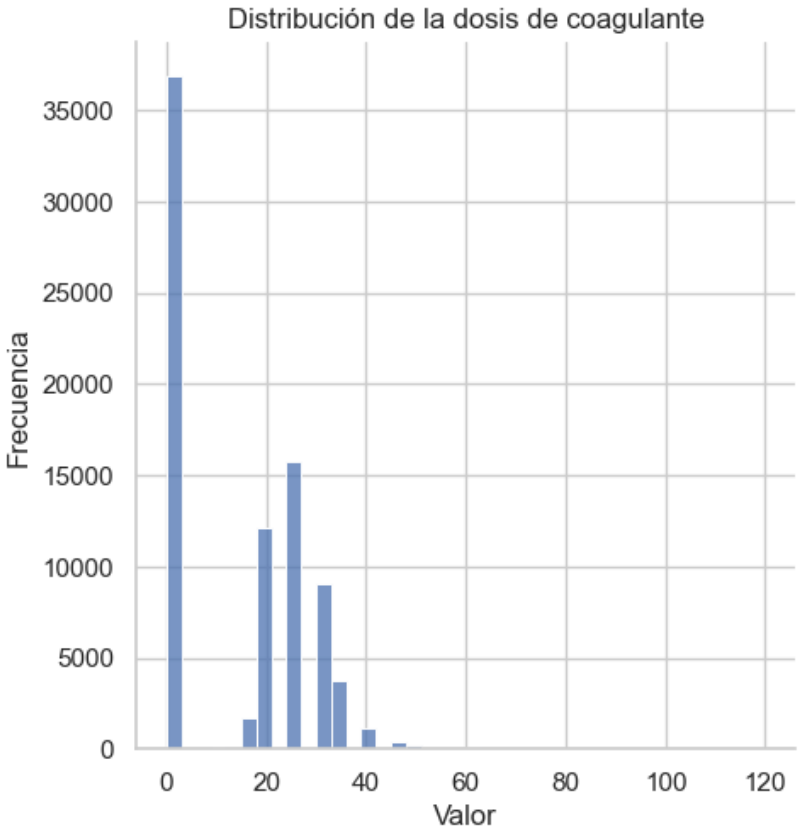

Figura 3.5: Distribución de dosis de coagulantes después de imputación de ceros. *Fuente propia.*

como se muestra en la Figura 3.6.

	Variables	% de nulos	N de nulos	N de NO nulos
0	CAL_1RIA_KG	99.966913	87619	29
1	COAGULANTE_LIQUIDO	99.887048	87549	99
2	CAL_1RIA_DOSIS	99.330276	87061	587
3	COAGULANTE_GRANULADO	93.269670	81749	5899
4	AGUA_TRATADA_ALCALINIDAD	58.748631	51492	36156
5	AGUA_CRUDA_CONDUCTIVIDAD	31.439394	27556	60092
6	AGUA_CRUDA_ALCALINIDAD	21.638828	18966	68682
7	AGUA_TRATADA_CLORO	11.745847	10295	77353
8	AGUA_TRATADA_COLOR	10.213582	8952	78696
9	COAGULANTE_DOSIS	7.037240	6168	81480
10	AGUA_CRUDA_COLOR	7.005294	6140	81508
11	AGUA_TRATADA_P.H	3.402245	2982	84666
12	AGUA_TRATADA_NTU	3.211710	2815	84833
13	AGUA_CRUDA_P.H	0.836300	733	86915
14	AGUA_CRUDA_NTU	0.563618	494	87154
15	CAUDAL	0.300064	263	87385
16	HORA	0.000000	0	87648
17	FECHA	0.000000	0	87648

Figura 3.6: Porcentaje de nulos de variables del conjunto de datos con imputación de ceros. *Fuente propia.*

Segundo, en el análisis univariable se realizó una descripción estadística para los dos conjuntos de datos, tal como se ilustra en la Figura 3.7. Este análisis evidenció que el número de registros en la columna COAGULANTE_DOSIS experimentó un aumento notable tras el proceso de imputación, creciendo de 44390 a 81480 entradas. La media de dicha columna descendió de 26.02 a 14.22 tras este proceso. La inserción de una gran cantidad de ceros ha causado que la distribución de la variable se incline hacia valores menores, generando lo que se conoce como un sesgo a la izquierda. Este sesgo indica que los valores más bajos (en este caso, los ceros) se presentan con mayor frecuencia, desplazando la media y mediana hacia el lado izquierdo de la distribución.

```
In [17]: # Resumen de estadística descriptiva de dataframe sin imputación de ceros:
         df['COAGULANTE_DOSIS'].describe()

Out[17]: count    44393.00000
         mean        26.01962
         std          6.71821
         min          0.00000
         25%         20.00000
         50%         25.00000
         75%         30.00000
         max        120.00000
         Name: COAGULANTE_DOSIS, dtype: float64

In [18]: # Resumen de estadística descriptiva de dataframe con imputación de ceros:
         df_edit['COAGULANTE_DOSIS'].describe()

Out[18]: count    81480.000000
         mean        14.215832
         std         13.848115
         min          0.000000
         25%          0.000000
         50%         20.000000
         75%         25.000000
         max        120.000000
         Name: COAGULANTE_DOSIS, dtype: float64
```

Figura 3.7: Descripción estadística de los conjuntos de datos sin imputación (arriba) y con imputación (abajo). *Fuente propia.*

Adicionalmente, se notó un aumento en la desviación estándar, pasando de 6.72 a 13.85, lo que indica una mayor dispersión en los valores de COAGULANTE_DOSIS después de la imputación. No obstante, los valores extremos (mínimo y máximo) para la dosis de coagulante se mantuvieron inalterados, siendo 0 y 120 respectivamente para ambos conjuntos.

En cuanto a los percentiles, se observaron cambios significativos tras la imputación: el percentil 25 % descendió a 0, señalando que al menos un cuarto de los datos imputados presentan un valor de 0 en la dosis de coagulante. El percentil 50 % se redujo de 25 a 20 y el percentil 75 % disminuyó levemente a 25.

En resumen, aunque la imputación incrementó el número de registros en la columna COAGULANTE_DOSIS, también alteró la distribución de sus valores.

Tercero, en el proceso de análisis multivariable, se estudiaron las interrelaciones entre las diferentes variables presentes en los conjuntos de datos. La meta era discernir las posibles discrepancias entre los conjuntos de datos imputados y los no imputados. Para

facilitar la interpretación visual de estas interrelaciones, se recurrió a mapas de calor, que se encuentran detallados en las figuras: Figura 3.8 para el conjunto de datos sin imputación de ceros y Figura 3.9 para el conjunto con los datos imputados.

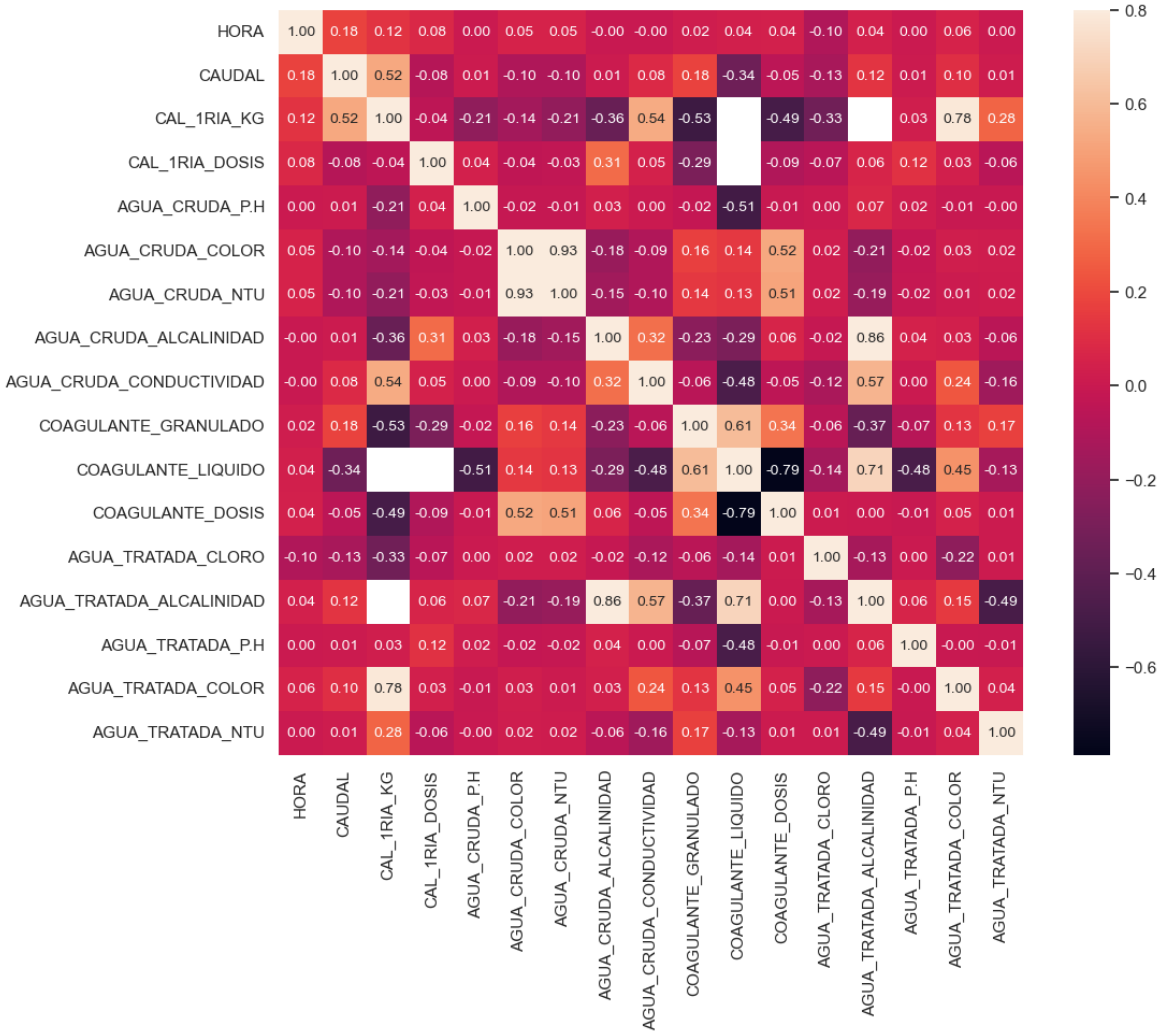

Figura 3.8: Correlación entre variables de conjunto de datos sin imputación de ceros. *Fuente propia.*

A partir de la comparación de los resultados de ambos conjuntos de datos, emerge un patrón notable. En el conjunto sin imputación, las variables `CAL_1RIA_KG`, `AGUA_CRUDA_COLOR`, `AGUA_CRUDA_NTU`, `COAGULANTE_GRANULADO` y `COAGULANTE_LIQUIDO` mostraron correlaciones significativas con la dosis de coagulante, siendo estas de -0.49, 0.52, 0.51, 0.34 y -0.79, respectivamente, como se observa en la Figura 3.8.

Este patrón sugiere que dichas variables podrían desempeñar un papel crucial en la determinación de la cantidad de coagulante necesaria. No obstante, la claridad en esta correlación parece disminuir para `CAL_1RIA_KG`, `AGUA_CRUDA_COLOR` y `AGUA_CRUDA_NTU` en el conjunto de datos con imputación de ceros. Estas variables mostraron correlaciones de -0.33, 0.37 y 0.34 respectivamente, tal como se detalla en la Figura 3.9.

Tal atenuación puede sugerir que al añadir datos imputados, es posible que se esté incorporando cierta distorsión o ruido en las interacciones entre las variables, alterando

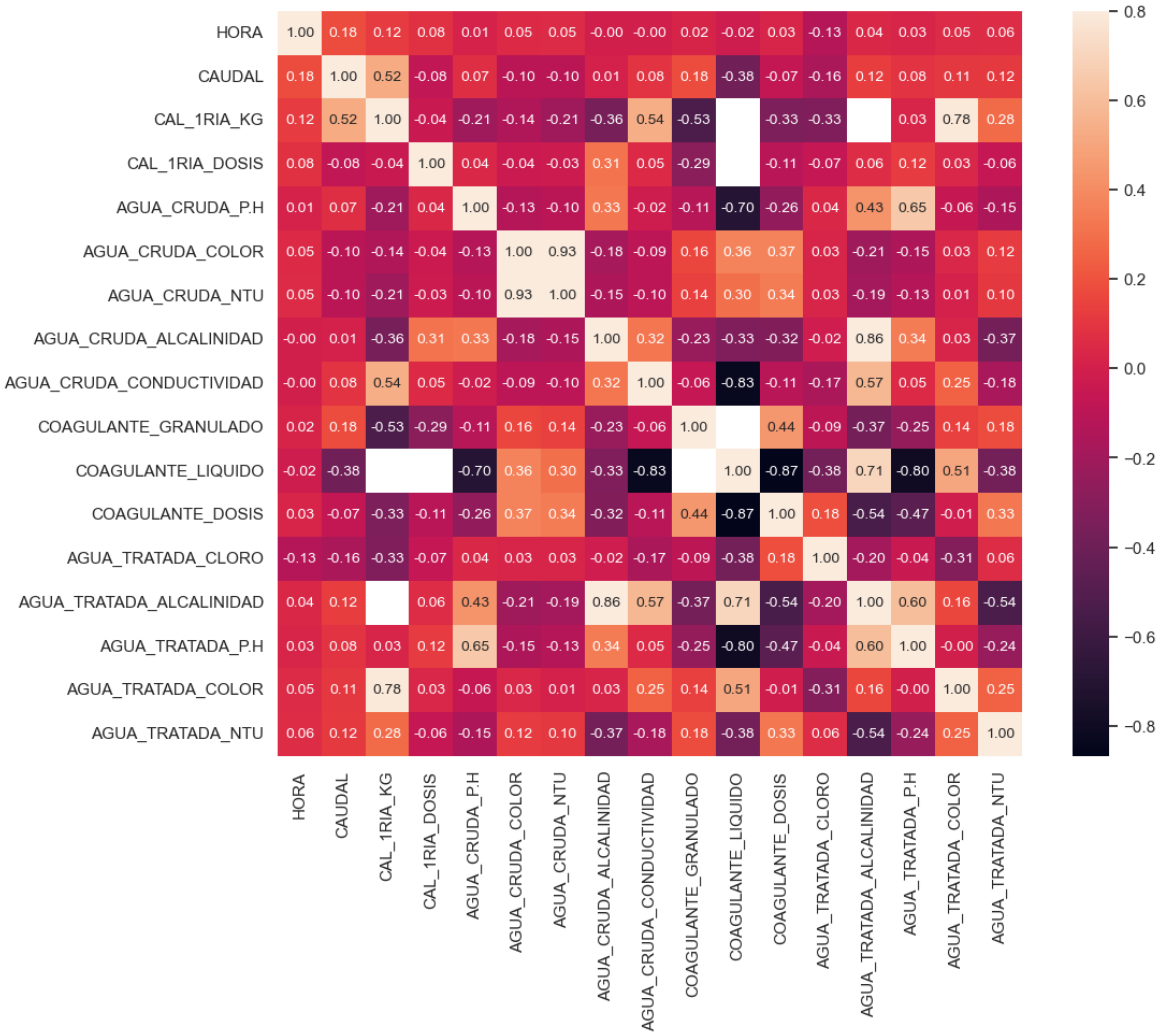

Figura 3.9: Correlación entre variables de conjunto de datos con imputación de ceros. *Fuente propia.*

las relaciones iniciales observadas en el conjunto original.

El código fuente desarrollado en este proyecto para el EDA comparativo se encuentra disponible en el repositorio de GitHub[4]

Fusión de datos hidrológicos con meteorológicos

Para este procedimiento se incorporaron los datos meteorológicos obtenidos del IDEAM al dataset descrito hasta ahora. Los datos del IDEAM incluyen medidas de precipitación, velocidad del viento, temperatura húmeda y temperatura seca. Sin embargo, hay que tener en cuenta que los datos de precipitación sólo están disponibles desde finales de 2016. Asimismo, las mediciones de las temperaturas húmeda y seca se realizaron solamente tres veces al día: a las 7:00, 13:00 y 18:00 horas.

[4]Código fuente para el EDA comparativo disponible en el enlace: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/04_EDA_Comparative.ipynb`

Estos datos meteorológicos se fusionaron tanto con el conjunto de datos que incluye las imputaciones de cero como con el que no las incluye. Cabe anotar que a pesar de la presencia de ciertos vacíos temporales en los datos resultantes de esta fusión, es factible realizar un preprocesamiento adecuado. Este proceso garantizará la obtención de un conjunto de datos de buena calidad, que proporcionará la base sólida necesaria para un posterior modelado de datos efectivo.

El código fuente desarrollado en este proyecto para la fusión de datos hidrológicos y meteorológicos se encuentra disponible en el repositorio de GitHub[5]

Análisis de valores atípicos

En el proceso de análisis, se realizó una detallada evaluación para identificar valores atípicos, considerando la calibración de los instrumentos utilizados en la recopilación de datos. En esta inspección, se encontraron anomalías, por ejemplo se encontraron valores de pH superiores a 14, lo que sugiere posibles errores de digitación. Estas inconsistencias se corrigieron manualmente en los archivos. Posteriormente, para identificar de manera efectiva otros valores atípicos, se usaron gráficos de caja que representaban cada variable hidrológica. En la Figura 3.10, se observa que todas las variables tienen datos atípicos, representados por puntos negros con forma de rombo.

En la Figura 3.10, se despliegan diagramas de caja ilustrando las variables asociadas al agua cruda. Estas mediciones fueron capturadas por los dispositivos del acueducto. Adicionalmente, el gráfico también refleja la variable relacionada con la dosis de coagulante aplicada. Según la entrevista realizada al Ingeniero Ramírez, que se puede consultar en el Anexo B, se establece que el caudal en condiciones normales no debería exceder los 800 L/s, puesto que es el valor promedio manejado en la planta de tratamiento y que caudales inferiores a 200 L/s suelen ser signos de obstrucciones en la bocatoma.

Con respecto a otras variables, como el color, turbiedad, conductividad y alcalinidad, el Ingeniero Ramírez señaló que los operadores han calibrado los equipos para obtener lecturas máximas de 500 UPC, 800 NTU, 12800 µS/cm, 100 mg/L de carbonato de calcio, respectivamente. Para el caso del pH el equipo es calibrado para el rango de 4 a 10. Los manuales técnicos de estos dispositivos están referenciados en [80, 81, 82, 83].

[5]Código fuente para la fusión de datos hidrológicos y meteorológico disponible en el enlace: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/06_Merge_With_Meteorologic.ipynb`

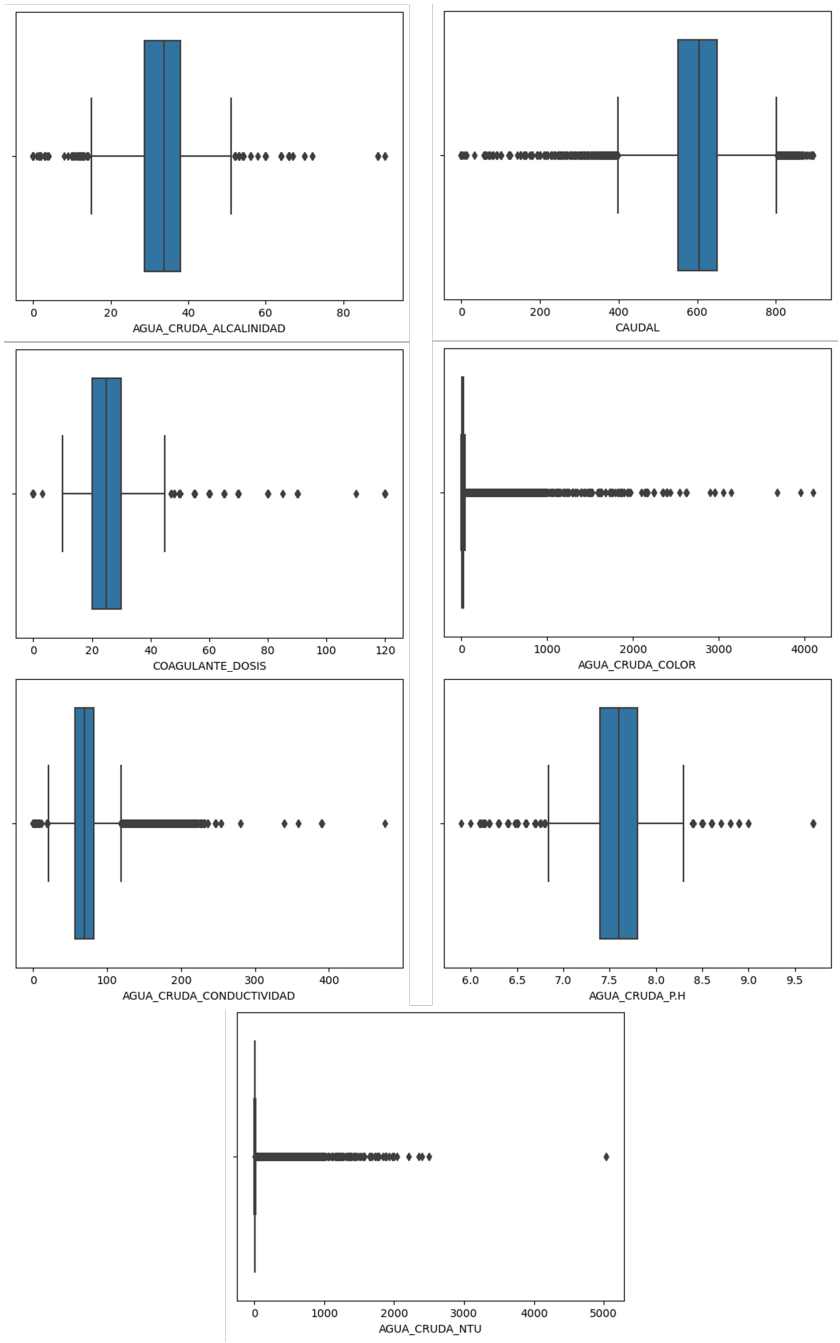

Figura 3.10: Valores atípicos en las variables. *Fuente propia.*

Por lo tanto, para garantizar la precisión e integridad de los datos, se tomó la determinación de eliminar aquellos valores que sobrepasaban los límites de calibración de los equipos o que, según la experiencia del operador, eran considerados atípicos o irreales. Específicamente, para el caudal, se filtraron valores fuera del rango de 200 a 800 L/s, y en el caso del color y la turbiedad, se descartaron valores superiores a 500 UPC y 800 NTU, respectivamente, acorde con las calibraciones de los equipos.

En cuanto a la turbiedad se encontró que existen valores elevados de color y turbiedad, los cuales pueden ser posibles. Sin embargo, se tomó la decisión de excluir aquellos datos obtenidos mediante el método de diluciones. Este método es utilizado por los

operadores cuando los valores exceden las capacidades del equipo. En estos casos, se atenúa la turbiedad o el color de la muestra combinándola con agua sin turbiedad o color. Posteriormente, las muestras diluidas se analizan conforme a procedimientos estándar, corrigiéndose con un factor de dilución. La razón detrás de esta elección radica en la potencial inexactitud de los resultados derivados de dicho método, ya que las diluciones pueden estar sujetas a errores de medición. Estos errores podrían, a su vez, afectar la precisión y confiabilidad del modelo propuesto. Por lo tanto, se tomó esta decisión tras diversas reuniones con el ingeniero Ramírez. Para mayor detalle sobre estos encuentros, se pueden consultar las transcripciones de las entrevistas en el Anexo B, donde se exponen tanto las preguntas formuladas como las respuestas proporcionadas.

En el análisis de la variable alcalinidad, se optó por eliminar aquellos valores que superaban las 100 unidades. En cuanto a las variables de conductividad y pH, los datos registrados no excedían los valores máximos de calibración de los equipos de medición utilizados, por lo que no se consideró necesario eliminar ningún valor correspondiente a estas dos variables.

A diferencia de las variables anteriores, en el análisis de la variable `COAGULANTE_DOSIS` se tuvo en cuenta la nueva política establecida para la planta de "El Tablazo", según la cual no se aplicaría coagulante bajo ciertas condiciones: específicamente, cuando la turbiedad del agua fuera menor a 3 y el horario se encontrara entre las 8 de la mañana y las 8 de la noche. Por lo tanto, se determinó que los valores de cero en la dosis de coagulante podrían introducir ruido significativo en el modelo analítico, especialmente cuando se deseara predecir la cantidad de coagulante necesaria en situaciones que no se alinearan con esta política.

Por lo tanto, en el conjunto de datos, se eliminaron los registros donde la dosis de coagulante era cero, así como aquellos registros con dosis mayores a 50, para evitar posibles datos atípicos que pudieran sesgar los resultados del análisis.

Es esencial aclarar que las decisiones de depuración de datos se llevaron a cabo solo después de un análisis detallado de los mismos. Cuando los datos se encontraban fuera de los rangos establecidos, estos eran revisados meticulosamente. Si, durante esta revisión, se identificaba un error evidente de digitación, se corregía y el dato ajustado usualmente volvía a encajar dentro de los límites aceptables. Sin embargo, en algunos casos, aunque los datos estuvieran fuera del rango, no mostraban signos claros de ser un error de registro. En tales situaciones, ante la falta de certeza y para mantener la integridad y precisión del análisis, se tomó la decisión de eliminar esos datos. El objetivo principal de estas acciones fue asegurar la calidad de la información y evitar conclusiones basadas en datos que podrían ser incorrectos o ambiguos.

El código fuente desarrollado en este proyecto para el análisis de valores atípicos se encuentra disponible en el repositorio de GitHub[6]

[6]Código fuente para el análisis de valores atípicos disponible en el enlace: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/07_EDA_Outliers.ipynb`

EDA final

Por último, en esta fase exploratoria de los datos se llevó a cabo un análisis detallado de los conjuntos de datos, que incluyó la revisión de valores nulos, tendencias temporales y correlaciones entre variables. Este análisis se realizó tanto para el conjunto de datos con imputación de ceros como para el conjunto sin dicha imputación.

Debido a la alta presencia de valores nulos en las columnas relacionadas con la variable `CAL_1RIA`, estas columnas fueron excluidas del análisis. Según lo mencionado por el ingeniero Ramírez (véase el Anexo B), la `CAL_1RIA` se utiliza específicamente en situaciones donde el agua tiene altas concentraciones de minerales, llevando a que el agua ingrese a la planta con un pH bajo. Sin embargo, este escenario ocurre raramente. Además, se optó por mantener únicamente los datos asociados al `COAGULANTE_GRANULADO`, conservando las filas que indican la dosificación con dicho coagulante. Se tomó esta decisión debido a que los registros relacionados con el `COAGULANTE_LIQUIDO` eran limitados, ya que se realizaron pocos experimentos con este coagulante, y su inclusión podría generar inconsistencias en los análisis posteriores.

Se identificó que la medición continua de parámetros como la conductividad inició en 2016, mientras que los registros de precipitación solo comenzaron a finales de ese mismo año, como se muestra en la Figura 3.11. Debido a esto, se optó por trabajar con dos conjuntos de datos adicionales que contienen información exclusivamente desde 2017, tanto para las series de datos con valores imputados como ceros como para aquellas sin esa imputación.

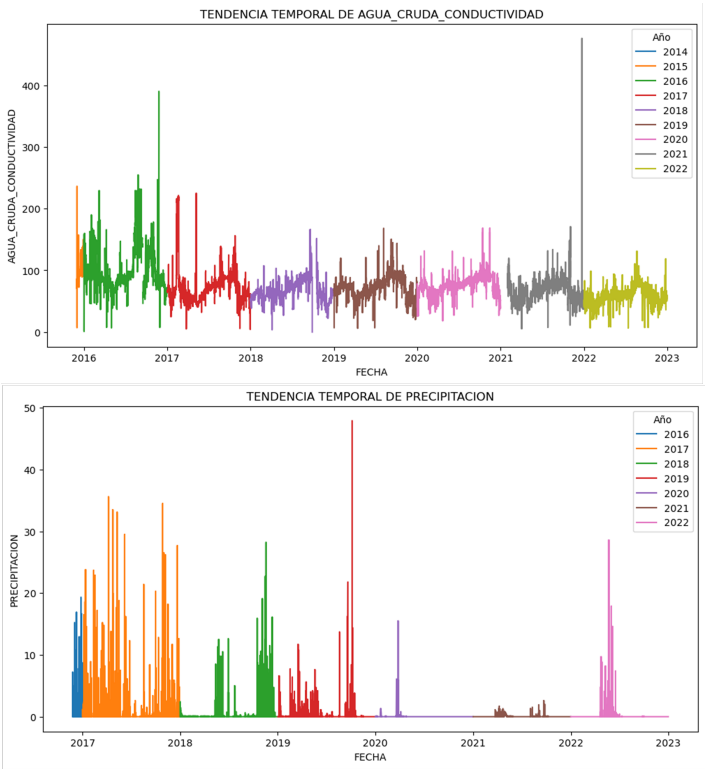

Figura 3.11: Tendencia temporal de la variable conductividad (arriba) y la variable de precipitación (abajo). *Fuente propia.*

Con el objetivo de profundizar el análisis y entender de manera precisa cómo las variables interactúan, particularmente en determinados periodos del año que reflejan la variación climática de la ciudad, se introdujeron modificaciones al conjunto de datos. Una de estas adaptaciones fue la inclusión de una nueva columna en los conjuntos de datos, la cual destaca una tendencia específica identificada en todas las variables, predominante entre los meses de junio y septiembre. Esta columna fue nombrada `CLASIFICADOR_MENSUAL`. La Figura 3.12 ilustra esta tendencia, mostrando específicamente cómo las variables de color y dosis de coagulante se comportan y se relacionan durante ese intervalo.

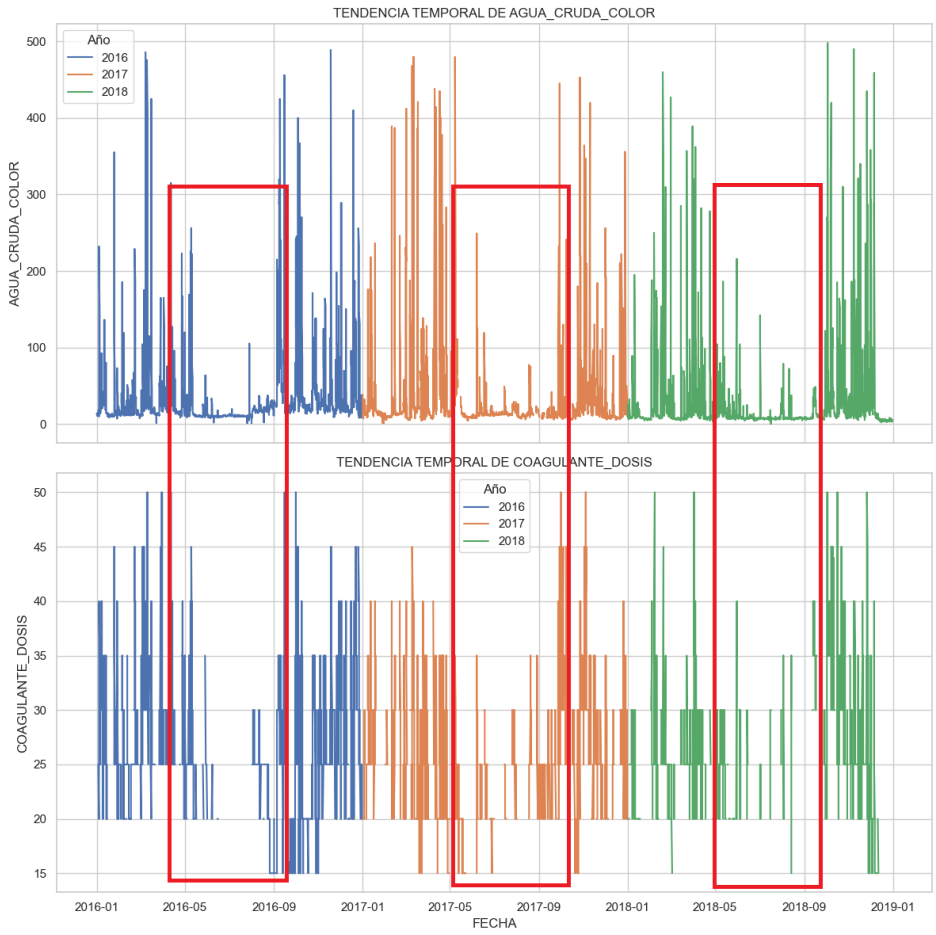

Figura 3.12: Tendencia temporal entre la variable color (arriba) y la variable de dosis de coagulante (abajo). *Fuente propia.*

Con el propósito de analizar esta tendencia de manera más estructurada, se dividió el conjunto de datos en dos segmentos. El primer segmento engloba la información recopilada específicamente entre los meses de junio y septiembre, mientras que el segundo segmento comprende los datos de todos los demás meses del año.

Adicionalmente, buscando considerar las operaciones diarias de la planta de tratamiento, se añadió otra columna al conjunto de datos. Esta nueva columna categoriza los datos según la franja horaria en la que fueron registrados. De esta manera, se establecieron dos periodos claramente diferenciados: el periodo diurno, que se extiende desde las 8 de la mañana hasta las 8 de la noche, y el periodo nocturno, que cubre las horas

restantes, incluyendo las primeras horas de la madrugada y la noche avanzada. Con estas divisiones, se busca analizar si las operaciones o eventos que ocurren en diferentes momentos del día pueden influir de manera significativa en los resultados observados. La columna fue denominada `CLASIFICADOR_HORARIO`.

Durante el análisis de correlación, se identificaron correlaciones significativas entre los parámetros de turbiedad y color, así como entre la temperatura seca y la temperatura húmeda. Estas correlaciones están representadas gráficamente en la Figura 3.13, que muestra el comportamiento en el conjunto de datos sin ediciones. A raíz de estas correlaciones, se decidió excluir los datos de temperatura seca del análisis para evitar problemas de multicolinealidad. La multicolinealidad se refiere a una situación en la cual dos o más variables predictoras en un modelo de regresión múltiple están altamente correlacionadas, dificultando así la interpretación de los efectos individuales de cada variable predictora en la variable de respuesta [84].

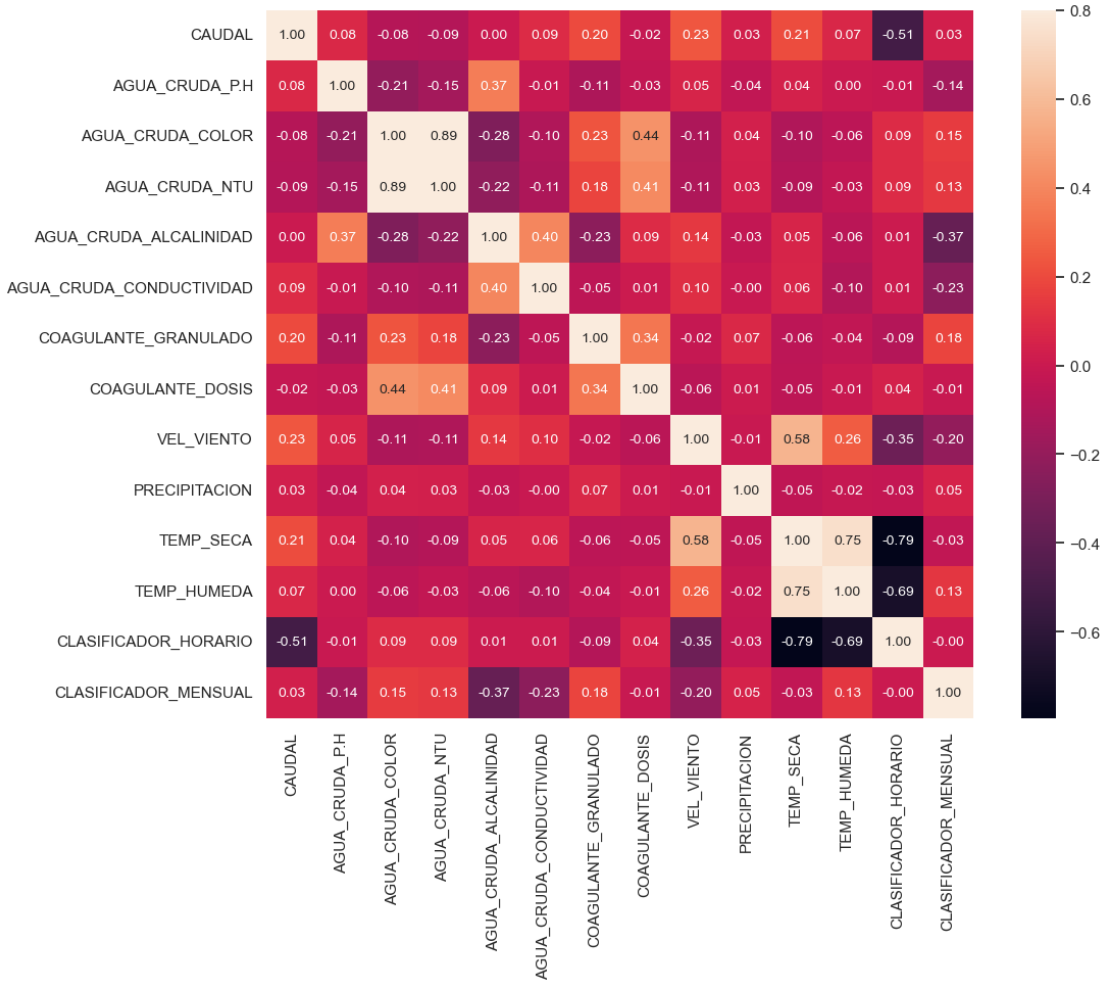

Figura 3.13: Correlaciones entre variables en conjunto de datos sin imputación *Fuente propia.*

En lo que respecta a la correlación entre los parámetros de turbiedad y color, se decidió conservar ambas variables en el análisis debido a que representan aspectos diferentes del agua, a diferencia de las temperaturas seca y húmeda, que ambas representan

mediciones de temperatura. Se optó por mantener solo la temperatura húmeda ya que se consideró más representativa para el propósito del análisis.

Como resultado de este proceso analítico exhaustivo, se generaron cuatro diferentes conjuntos de datos para su posterior análisis. Dos de ellos contienen datos con y sin imputación de ceros para el periodo comprendido entre 2013 y 2022. Los otros dos conjuntos también presentan datos con y sin imputación de ceros, pero se limitan al periodo que va desde 2017 hasta 2022. De este modo, se dispone de un total de cuatro conjuntos de datos distintos para llevar a cabo análisis adicionales. A continuación, en las Figuras 3.14, 3.15, 3.16 y 3.17 se muestra la información de cada uno de los conjuntos de datos resultantes. En adelante, se referirán respectivamente como: Conjunto de datos 2013-2022, Conjunto de datos editado 2013-2022, Conjunto de datos 2017-2022 y Conjunto de datos editado 2017-2022.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 86050 entries, 0 to 87647
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   FECHA                    86050 non-null  datetime64[ns]
 1   HORA                     86050 non-null  int64
 2   CAUDAL                   86032 non-null  float64
 3   AGUA_CRUDA_P.H           85697 non-null  float64
 4   AGUA_CRUDA_COLOR         80260 non-null  float64
 5   AGUA_CRUDA_NTU           85965 non-null  float64
 6   AGUA_CRUDA_ALCALINIDAD   67691 non-null  float64
 7   AGUA_CRUDA_CONDUCTIVIDAD 59501 non-null  float64
 8   COAGULANTE_DOSIS         43483 non-null  float64
 9   VEL_VIENTO               71055 non-null  float64
 10  PRECIPITACION            39491 non-null  float64
 11  TEMP_HUMEDA              6801 non-null   float64
 12  CLASIFICADOR_HORARIO     86050 non-null  int64
 13  CLASIFICADOR_MENSUAL     86050 non-null  int64
dtypes: datetime64[ns](1), float64(10), int64(3)
memory usage: 9.8 MB
```

Figura 3.14: Conjunto de datos 2013-2022 *Fuente propia.*

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 51573 entries, 35 to 87647
Data columns (total 14 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   FECHA                     51573 non-null  datetime64[ns]
 1   HORA                      51573 non-null  int64
 2   CAUDAL                    51562 non-null  float64
 3   AGUA_CRUDA_P.H            51282 non-null  float64
 4   AGUA_CRUDA_COLOR          48668 non-null  float64
 5   AGUA_CRUDA_NTU            51488 non-null  float64
 6   AGUA_CRUDA_ALCALINIDAD    42716 non-null  float64
 7   AGUA_CRUDA_CONDUCTIVIDAD  38055 non-null  float64
 8   COAGULANTE_DOSIS          43483 non-null  float64
 9   VEL_VIENTO                39787 non-null  float64
 10  PRECIPITACION             22310 non-null  float64
 11  TEMP_HUMEDA               4181 non-null   float64
 12  CLASIFICADOR_HORARIO      51573 non-null  int64
 13  CLASIFICADOR_MENSUAL      51573 non-null  int64
dtypes: datetime64[ns](1), float64(10), int64(3)
memory usage: 5.9 MB
```

Figura 3.15: Conjunto de datos editado 2013-2022 *Fuente propia.*

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 51808 entries, 35064 to 87647
Data columns (total 14 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   FECHA                     51808 non-null  datetime64[ns]
 1   HORA                      51808 non-null  int64
 2   CAUDAL                    51797 non-null  float64
 3   AGUA_CRUDA_P.H            51571 non-null  float64
 4   AGUA_CRUDA_COLOR          50939 non-null  float64
 5   AGUA_CRUDA_NTU            51791 non-null  float64
 6   AGUA_CRUDA_ALCALINIDAD    50800 non-null  float64
 7   AGUA_CRUDA_CONDUCTIVIDAD  50323 non-null  float64
 8   COAGULANTE_DOSIS          28844 non-null  float64
 9   VEL_VIENTO                38614 non-null  float64
 10  PRECIPITACION             38620 non-null  float64
 11  TEMP_HUMEDA               4915 non-null   float64
 12  CLASIFICADOR_HORARIO      51808 non-null  int64
 13  CLASIFICADOR_MENSUAL      51808 non-null  int64
dtypes: datetime64[ns](1), float64(10), int64(3)
memory usage: 5.9 MB
```

Figura 3.16: Conjunto de datos 2017-2022 *Fuente propia.*

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 32274 entries, 35064 to 87647
Data columns (total 14 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   FECHA                    32274 non-null  datetime64[ns]
 1   HORA                     32274 non-null  int64
 2   CAUDAL                   32266 non-null  float64
 3   AGUA_CRUDA_P.H           32065 non-null  float64
 4   AGUA_CRUDA_COLOR         31727 non-null  float64
 5   AGUA_CRUDA_NTU           32257 non-null  float64
 6   AGUA_CRUDA_ALCALINIDAD   31540 non-null  float64
 7   AGUA_CRUDA_CONDUCTIVIDAD 31276 non-null  float64
 8   COAGULANTE_DOSIS         28844 non-null  float64
 9   VEL_VIENTO               21545 non-null  float64
 10  PRECIPITACION            21548 non-null  float64
 11  TEMP_HUMEDA              3042 non-null   float64
 12  CLASIFICADOR_HORARIO     32274 non-null  int64
 13  CLASIFICADOR_MENSUAL     32274 non-null  int64
dtypes: datetime64[ns](1), float64(10), int64(3)
memory usage: 3.7 MB
```

Figura 3.17: Conjunto de datos editado 2017-2022 *Fuente propia.*

El código fuente desarrollado en este proyecto para este EDA exhaustivo se encuentra disponible en el repositorio de GitHub[7]

Imputación de datos

A partir de los cuatro conjuntos de datos obtenidos (Conjunto de datos 2013-2022, Conjunto de datos editado 2013-2022, Conjunto de datos 2017-2022 y Conjunto de datos editado 2017-2022), el siguiente paso en el análisis consistió en abordar de forma minuciosa el tratamiento de los datos faltantes. Este proceso es fundamental, dado que la presencia de valores no registrados en los conjuntos de datos tiene el potencial de afectar de manera significativa la calidad y precisión de cualquier modelo predictivo que se desarrolle a partir de estos datos.

El primer paso de este análisis consistió en identificar tanto la naturaleza como la magnitud de los datos ausentes dentro del conjunto de datos. Se formularon preguntas esenciales:

- ¿Se distribuyen estos datos faltantes de manera aleatoria o siguen un patrón determinado?
- ¿Cuál es la proporción de datos faltantes respecto al total del conjunto de datos?

[7]Código fuente para el EDA disponible en el enlace: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/10_Complet_Analysis_Without_Outliers.ipynb`

- ¿Los datos faltantes, son más prevalentes en ciertas variables que en otras?

La finalidad de este análisis radicó en adquirir una comprensión profunda sobre la naturaleza de los datos ausentes. Esta comprensión guió la estrategia adoptada para la imputación, permitiendo abordar estos valores no registrados de una forma efectiva.

Después de este estudio, se observaron ciertas relaciones. Por ejemplo, los datos perdidos de precipitación mostraban cierta correlación con los datos perdidos de velocidad del viento. También se observó una relación entre los datos perdidos de conductividad y alcalinidad. Estas relaciones se pueden visualizar en la Figura 3.18 la cual muestra los datos faltantes en blanco respecto a los datos existentes en negro. Con estas relaciones en mente, se avanzó hacia la fase de imputación.

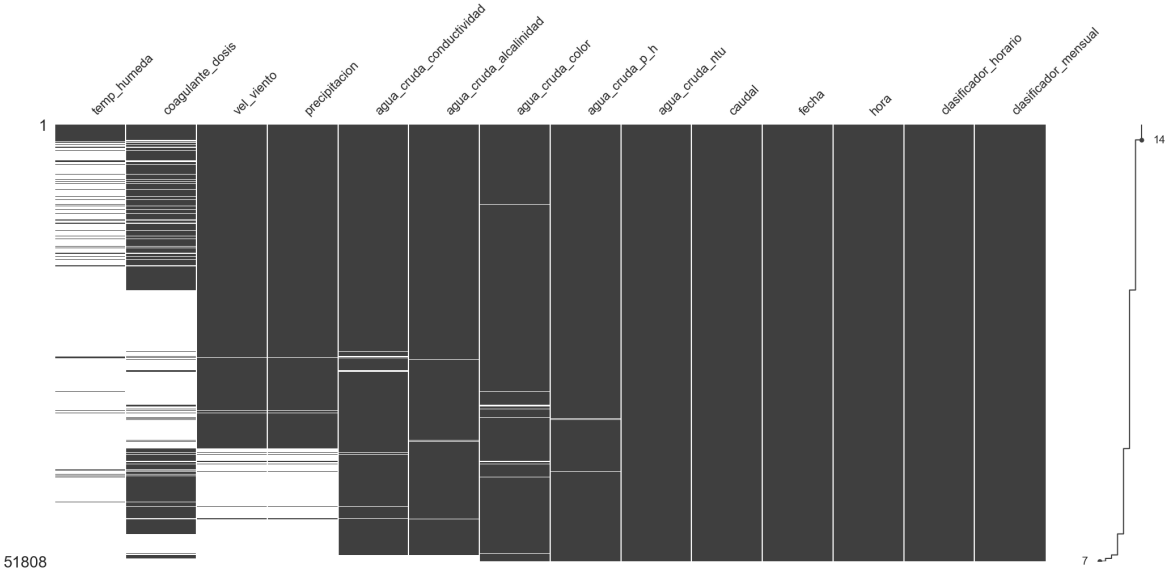

Figura 3.18: Análisis de valores faltantes. *Fuente propia.*

La imputación es un paso crítico y debe realizarse con extrema cautela para evitar la incorporación de sesgos en el conjunto de datos. Para ello, se seleccionaron métodos de imputación que estuvieran adecuadamente alineados con la naturaleza de los datos faltantes identificados.

Durante este proceso, se evaluó cada variable para determinar a cuál mecanismo de datos faltantes pertenecía. Los mecanismos de datos faltantes se clasifican en: Ausentes completamente al Azar (MCAR, *Missing Completely at Random*), Ausentes al Azar (MAR, *Missing at Random*) y Ausentes No al Azar (MNAR, *Missing Not at Random*) [85].

Los datos MCAR son aquellos cuya probabilidad de ausencia es independiente tanto de los valores observados como de los no observados. En el caso de los datos MAR, su probabilidad de estar ausentes depende solo de los valores observados, mientras que en MNAR, la probabilidad de estar ausentes depende de los valores no observados [85].

Para clasificar las variables según estos mecanismos, primero se realizó una limpieza final de los nombres de las columnas, pues la presencia de mayúsculas y puntos intermedios estaba ocasionando problemas en este análisis. Luego, se realizó una clasificación preliminar fundamentada en el porcentaje de datos faltantes. En este estudio,

se adoptó un umbral del 20 % para gestionar los datos faltantes, siguiendo prácticas recomendadas en investigación y análisis de datos. Se determinó que cuando una variable presenta más del 20 % de datos ausentes, las técnicas de imputación pueden añadir una incertidumbre considerable, al basarse en un conjunto limitado de datos reales. Estos datos imputados pueden, por ende, no reflejar adecuadamente la verdadera distribución y características de la variable, introduciendo posibles sesgos. Por esta razón, las variables que excedían el umbral del 20 % de datos faltantes no se clasificaron mediante los mecanismos convencionales.

Bajo esta premisa, solo las variables que presentaban menos del 20 % de datos faltantes fueron consideradas para evaluar a que mecanismo de datos faltantes pertenecía. Las variables analizadas en este proceso fueron: `caudal`, `agua_cruda_p_h`, `agua_cruda_color`, `agua_cruda_alcalinidad`, `agua_cruda_conductividad`. Esta evaluación se realizó sólo para los conjuntos de datos correspondientes al período desde 2017 en adelante. Para los conjuntos de datos que abarcaban el período completo, se realizó una evaluación similar, excluyendo `agua_cruda_alcalinidad` y `agua_cruda_conductividad` debido a su ausencia en los primeros años.

Con esta información, se ejecutó una prueba estadística t-test para determinar si existía una diferencia significativa en la presencia o ausencia de valores en las variables numéricas con respecto a las columnas de tendencias horarias o mensuales. En el caso de un test "two-sided", la pregunta era: ¿Existe una diferencia en la presencia o ausencia de valores de medición de la variable en evaluación? Si el valor obtenido era mayor a 0.05, se concluía que la hipótesis nula de que exista una diferencia no podía ser confirmada ni refutada. Esto implicaría que los datos de esa variable parecían estar perdidos de manera aleatoria.

No obstante, esta decisión no se basaba únicamente en los resultados del t-test, sino también en entender el porqué de las pérdidas de datos. Por ello, se clasificó cada variable para determinar si correspondían a los mecanismos MAR, MCAR o MNAR, considerando tanto los resultados de la prueba t-test como la información proporcionada en las entrevistas con el ingeniero Ramírez, jefe de la planta (Anexo B).

Tras llevar a cabo el análisis, se descubrió que las respuestas indicaban que las variables estaban asociadas con, al menos, una de las dos variables categóricas. Sumado a esto, basándose en la información proporcionada por el ingeniero Ramírez, se dedujo que dichas variables se ajustaban al mecanismo conocido como MAR. A raíz de estos hallazgos, se optó por llevar a cabo una imputación mediante interpolación con método "time". Esta decisión se fundamentó en las diversas ventajas que este método brinda, entre ellas, la facilidad de su implementación y su eficacia en el tratamiento de series temporales. Al emplear esta técnica, se minimiza el riesgo de introducir sesgos en el conjunto de datos, asegurando la calidad y precisión del análisis posterior. Para ello, se tuvo que convertir la fecha y hora en el índice de los conjuntos de datos.

Por otro lado, para las variables que no se consideraron en el análisis anterior debido a su alta proporción de datos faltantes, se optó por realizar una Imputación Múltiple por Ecuaciones Encadenadas (MICE, *Multiple Imputation by Chained Equations*), que son especialmente útiles cuando el patrón de datos faltantes es complejo y no se puede manejar adecuadamente mediante métodos más simples. Este método, basado en modelos, ayuda a imputar los datos faltantes de una manera que reduce la probabilidad de

introducir sesgos en el análisis [86].

El código fuente desarrollado en este proyecto para el análisis de datos faltantes se encuentra disponible en el repositorio de GitHub[8]

Por último, Se realizó un último de las correlaciones dentro de los cuatro conjuntos de datos ya corregidos e imputados. A partir de las correlaciones identificadas, se vio la necesidad de hacer ajustes en los datos.

En este sentido, se redujo la multicolinealidad para mejorar la precisión y fiabilidad de los análisis posteriores por medio de los siguientes procedimientos. Para los conjuntos de datos que cubren el periodo completo, se eliminaron las variables `agua_cruda_color` y `clasificador_horario`. Por otro lado, en los conjuntos de datos que abarcan información desde el año 2017 en adelante, se descartaron las variables `agua_cruda_ntu` y `clasificador_horario`.

3.2.4 Preparación de los Conjuntos de Datos:

En este punto se llevó a cabo un procedimiento de preparación de cada conjunto de datos para la fase de modelado, un aspecto clave en el marco de referencia adoptado.

Esta preparación implicó dividir cada conjunto de datos en dos subconjuntos: uno compuesto por el 80 % de los datos para entrenamiento y otro por el 20 % de los datos para pruebas como se observa en la Figura 3.19. Esta división es una práctica común en el campo del aprendizaje automático y se respalda por estudios que han demostrado su efectividad [87].

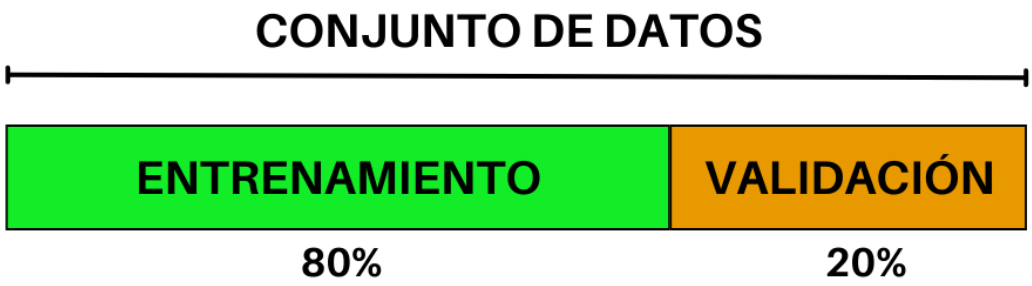

Figura 3.19: Representación de división de datos para entrenamiento de los modelos. *Fuente propia.*

El entrenamiento de los modelos con el 80 % de los datos es un paso crucial, ya que es durante esta fase que los modelos aprenden a identificar patrones existentes en los datos [88]. Este aprendizaje les permite luego hacer predicciones cuando se les presentan nuevos datos.

La etapa de pruebas, empleando el 20 % de los datos, resulta fundamental para evaluar la aplicación de los modelos a datos previamente no observados. Esta fase reviste una

[8]Código fuente para el análisis de datos faltantes disponible en el enlace: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/12_Imputation.ipynb`

importancia crítica para medir la habilidad del modelo en generalizar, es decir, en aplicar su aprendizaje a situaciones novedosas[89].

3.2.5 Modelos de predicción

Antes de adentrarse en la etapa de modelado, es esencial entender las diferencias fundamentales entre dos tipos de problemas analíticos: clasificación y regresión. Estos dos enfoques tienen objetivos y métodos diferentes. Mientras que la clasificación busca categorizar datos en grupos o clases específicas, la regresión intenta predecir valores numéricos continuos.

Dada esta diferencia clave, es crucial tratar estos problemas de manera independiente. Por lo tanto, se llevaron a cabo dos análisis distintos: uno centrado en el problema de clasificación y otro dirigido hacia el problema de regresión. Ambos análisis se realizaron con atención y detalle para garantizar resultados precisos y confiables en cada caso.

Modelo de clasificación.

En el contexto de este análisis, se llevó a cabo una transformación importante en la variable objetivo, convirtiéndola en una variable categórica. Se observó que las dosis aplicadas en la planta de tratamiento "El Tablazo" generalmente varían en incrementos de cinco unidades. Aunque hay algunas excepciones, la gran mayoría de las dosis aplicadas siguen este patrón. Con base en esta observación, se introdujo una nueva columna en el conjunto de datos para clasificar las dosis según rangos específicos. Estos rangos son: 15-19.99, 20-24.99, 25-29.99, 30-34.99, 35-39.99, 40-44.99, 45-49.99 y 50-99.99. Como resultado de esta transformación, la variable objetivo se convirtió en una variable categórica con ocho clases distintas. Esto condujo al planteamiento de un problema de clasificación multiclase. La estructura de estas clases se detalla en la Tabla 3.2.

Rango	**Nombre de la Clase**
15-19.99	`quince`
20-24.99	`veinte`
25-29.99	`veinticinco`
30-34.99	`treinta`
35-39.99	`treinta y cinco`
40-44.99	`cuarenta`
45-49.99	`cuarenta y cinco`
50-99.99	`cincuenta`

Tabla 3.2: Estructura de las clases. *Fuente Propia.*

La clasificación multiclase se refiere a tareas donde las instancias de datos deben ser asignadas a más de dos clases posibles. A diferencia de la clasificación binaria, que se centra en decidir entre dos posibles clases, la clasificación multiclase se encarga de asignar una instancia de datos a una de tres o más clases [90]. Este enfoque puede presentar varios desafíos inherentes, incluyendo el desequilibrio entre las clases, la

necesidad de diseñar estrategias de evaluación específicas y un aumento en la complejidad computacional debido a la mayor cantidad de clases involucradas [91].

En el estudio en cuestión, se observó un marcado desequilibrio entre las diferentes clases. Dicho desbalance, en el que algunas categorías tienen considerablemente más instancias que otras, puede comprometer la precisión y eficacia de los modelos predictivos que se generen. Las Figuras 3.20, 3.21, 3.22 y 3.23 reflejan visualmente esta disparidad. Particularmente, las clases `veinte` y `veinticinco` presentan un mayor número de registros, mientras que las clases `treinta y cinco`, `cuarenta`, `cuarenta y cinco` y `cincuenta` muestran una cantidad notablemente inferior, subrayando la naturaleza desequilibrada de la muestra.

Esta desigualdad entre clases debe ser meticulosamente abordada al momento de diseñar y validar modelos predictivos. No considerar este aspecto podría generar modelos con una inclinación marcada hacia las clases más abundantes, llevando a interpretaciones erradas. Es esencial, por tanto, enfrentar esta problemática para garantizar que los hallazgos del análisis sean robustos, confiables y representativos del fenómeno investigado.

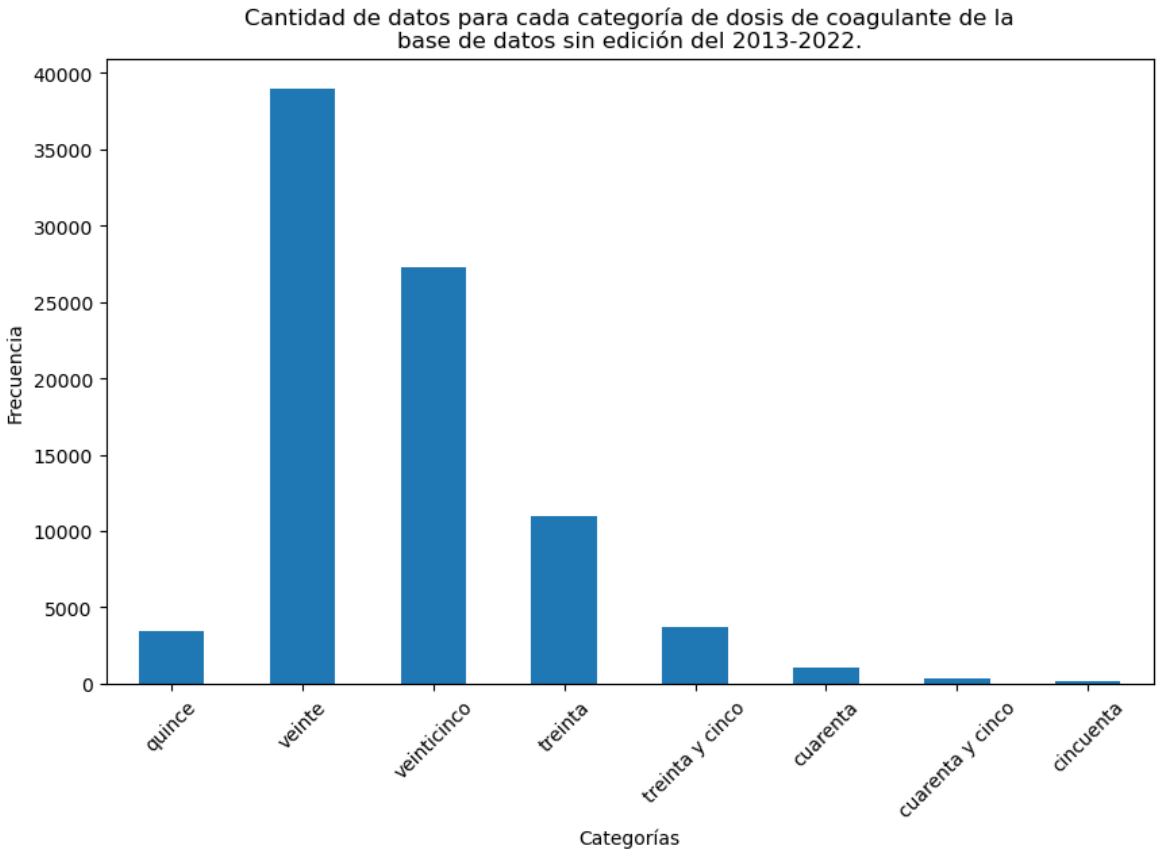

Figura 3.20: Distribución de datos entre categorías de dosis de coagulante del conjunto de datos 2013-2022. *Fuente Propia.*

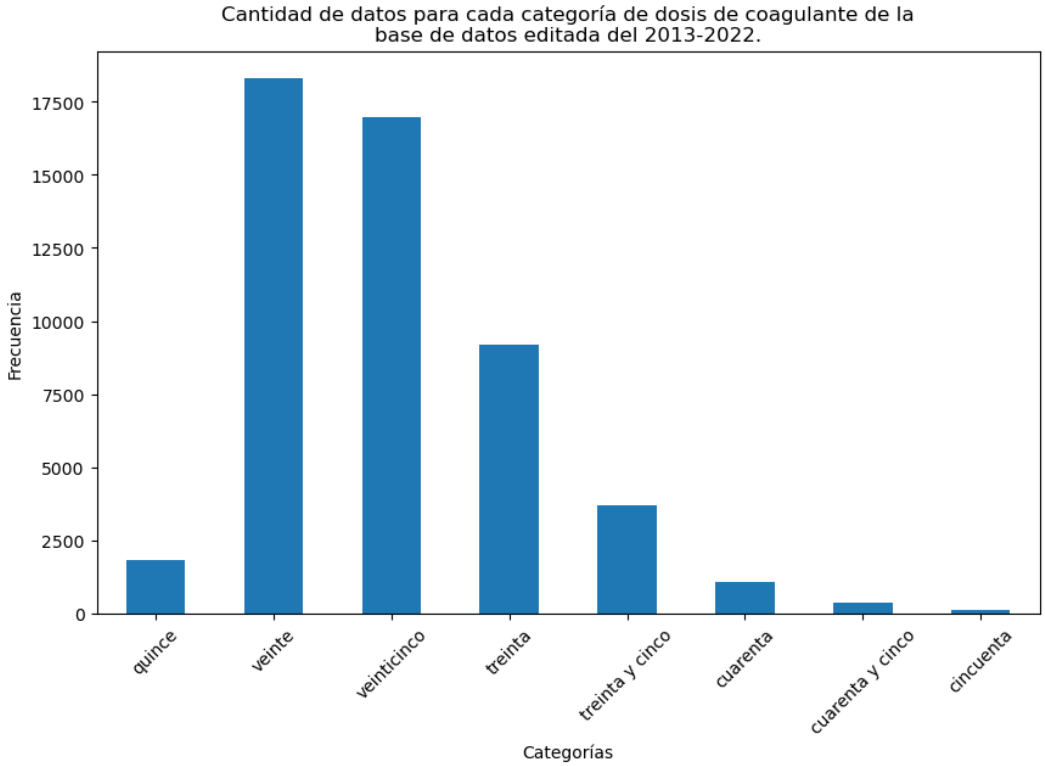

Figura 3.21: Distribución de datos entre categorías de dosis de coagulante del conjunto de datos editado del 2013-2022. *Fuente Propia.*

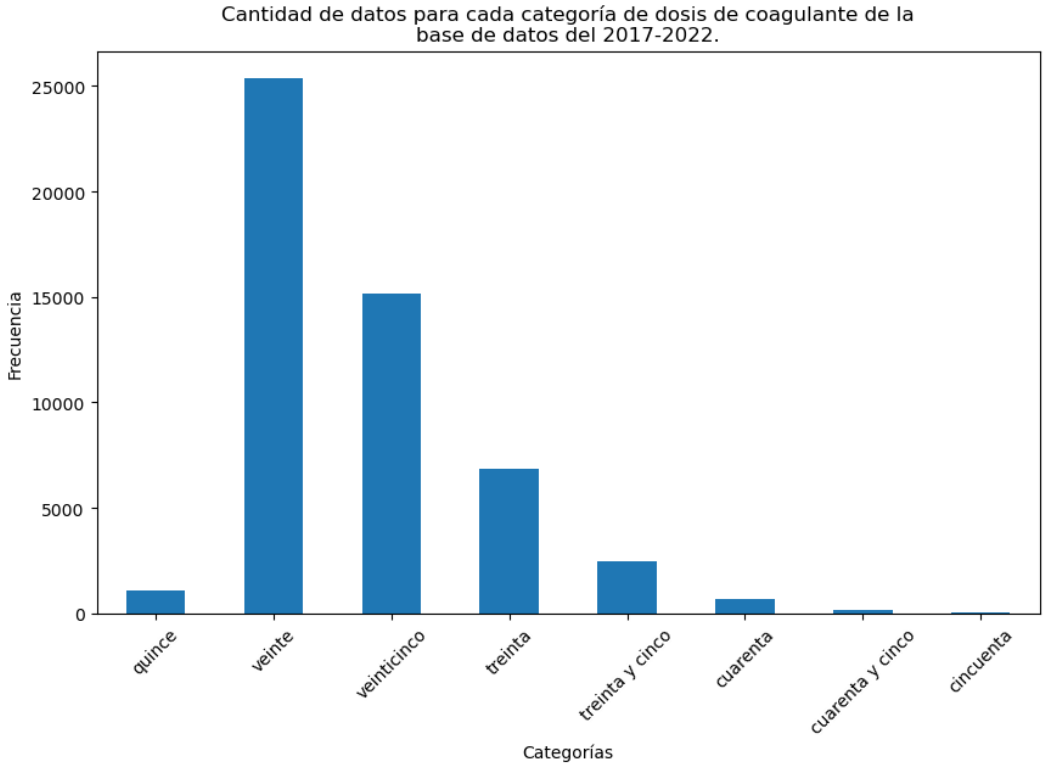

Figura 3.22: Distribución de datos entre categorías de dosis de coagulante del conjunto de datos 2017-2022. *Fuente Propia.*

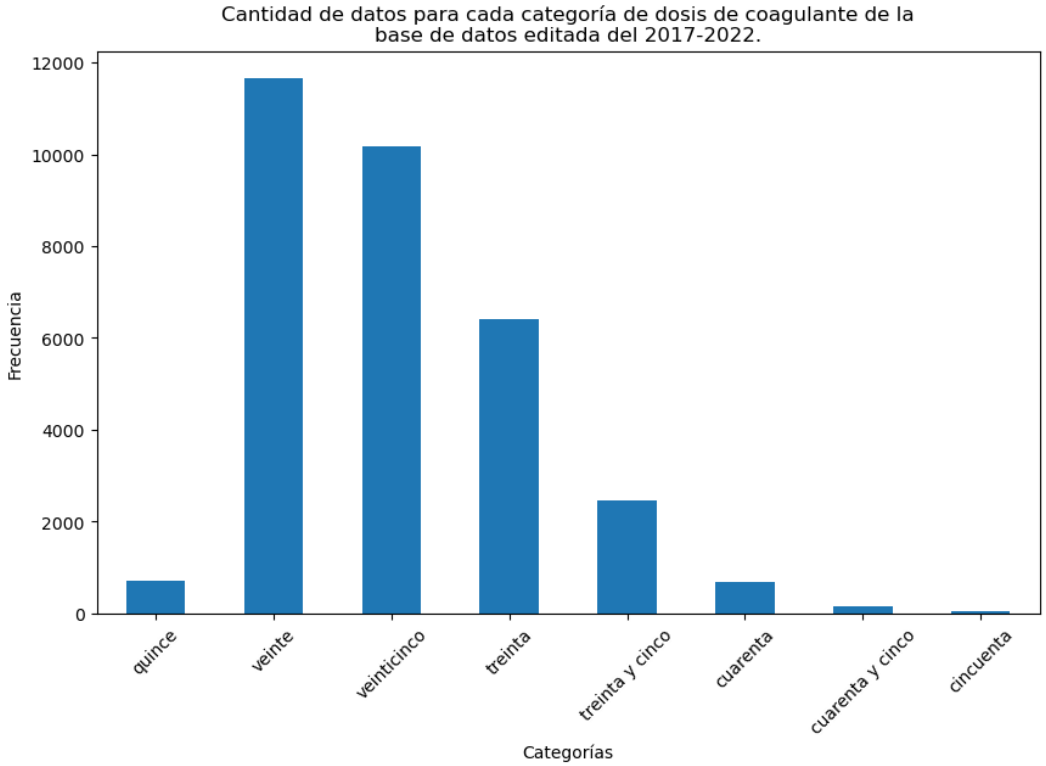

Figura 3.23: Distribución de datos entre categorías de dosis de coagulante de conjunto de datos editado del 2017-2022. *Fuente Propia.*

Previo al proceso de modelado, se consideró esencial realizar un análisis preliminar para determinar la mejor técnica de balanceo adaptada a los conjuntos de datos creados en la sección 3.2.4. Como paso inicial, se optó por una prueba utilizando regresión logística. A pesar de su denominación, la regresión logística, como se detalla en el Anexo A, no se emplea para regresión, sino para clasificación. Esta técnica es particularmente sensible al desequilibrio de clases, lo que la convierte en una herramienta útil para este tipo de evaluaciones preliminares.

- **Métricas de evaluación:** La efectividad de los modelos de clasificación diseñados en este proyecto se determinará utilizando la función `classification_report` de la biblioteca `sklearn.metrics` en Python. Esta herramienta ofrece un panorama completo al reportar varias métricas, como:
  - **Precisión (precision):** Define cuántas de las instancias que el modelo predijo como positivas eran realmente positivas. Se calcula como:

$$\text{precision} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Positivos}}$$

    [92].
  - **Exhaustividad (recall):** Mide cuántas instancias positivas reales fueron identificadas correctamente.

$$\text{recall} = \frac{\text{Verdaderos Positivos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

[92].

- **Puntaje F1 (f1-Score):** Representa una métrica compuesta que considera precisión y exhaustividad para ofrecer una visión integral de la calidad del modelo.

$$f1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

[93].

- **Support:** Indica el número de casos reales en cada clase del conjunto de datos [94].
- **Exactitud (accuracy):** La métrica refleja la proporción de predicciones correctas respecto al total de muestras evaluadas. Sin embargo, para garantizar una confiabilidad óptima en los resultados, es esencial asegurar un equilibrio en el número de muestras de cada clase dentro del conjunto de datos.

$$accuracy = \frac{\text{Cantidad de predicciones correctas}}{\text{Cantidad total de predicciones}}$$

[95]

Es esencial entender lo que estas métricas representan en el contexto de clases desequilibradas:

- **Alta precisión y alta exhaustividad:** Esta combinación indica que el modelo tiene una excelente capacidad para identificar y clasificar correctamente esa clase en particular.
- **Alta precisión y bajo exhaustividad:** Significa que, aunque el modelo clasifica con confianza cuando detecta la clase, en general, no la identifica con frecuencia.
- **Baja precisión y alto exhaustividad:** El modelo identifica frecuentemente la clase, pero a menudo se equivoca, confundiéndola con otras clases.
- **Baja precisión y bajo exhaustividad:** Esta es una señal de alarma, ya que el modelo tiene dificultades tanto para identificar como para clasificar correctamente la clase.

En situaciones de desequilibrio de clases, es común observar una alta precisión en las clases mayoritarias, lo que indica que el modelo las clasifica correctamente la mayoría de las veces. Sin embargo, esto suele ir acompañado de una baja exhaustividad en las clases minoritarias, lo que significa que el modelo tiende a pasar por alto o a clasificar erróneamente esas clases.

Los resultados obtenidos al realizar la prueba utilizando regresión logística en cada uno de los conjuntos de datos se presentan a continuación:

Clase	**Precision**	**Recall**	**F1-score**
cincuenta	0.30	0.07	0.11
cuarenta	0.00	0.00	0.00
cuarenta y cinco	0.00	0.00	0.00

quince	0.00	0.00	0.00
treinta	0.31	0.07	0.12
treinta y cinco	0.30	0.11	0.16
veinte	0.54	0.85	0.66
veinticinco	0.38	0.29	0.33

Tabla 3.3: Resultados de regresión logística en conjunto de datos 2013-2022. *Fuente Propia.*

Clase	**Precision**	**Recall**	**F1-score**
cincuenta	0.00	0.00	0.00
cuarenta	0.23	0.03	0.06
cuarenta y cinco	0.00	0.00	0.00
quince	0.00	0.00	0.00
treinta	0.34	0.07	0.11
treinta y cinco	0.31	0.13	0.19
veinte	0.47	0.70	0.56
veinticinco	0.38	0.46	0.41

Tabla 3.4: Resultados de regresión logística en conjunto de datos editado del 2013-2022. *Fuente Propia.*

Clase	**Precision**	**Recall**	**F1-score**
cincuenta	0.00	0.00	0.00
cuarenta	0.22	0.05	0.08
cuarenta y cinco	0.40	0.04	0.07
quince	0.00	0.00	0.00
treinta	0.33	0.15	0.20
treinta y cinco	0.28	0.09	0.14
veinte	0.63	0.87	0.73
veinticinco	0.40	0.34	0.36

Tabla 3.5: Resultados de regresión logística en conjunto de datos 2017-2022. *Fuente Propia.*

Clase	**Precision**	**Recall**	**F1-score**
cincuenta	0.00	0.00	0.00
cuarenta	0.26	0.02	0.04
cuarenta y cinco	0.00	0.00	0.00
quince	0.00	0.00	0.00
treinta	0.36	0.24	0.29
treinta y cinco	0.34	0.08	0.13
veinte	0.53	0.79	0.63
veinticinco	0.35	0.33	0.34

Tabla 3.6: Resultados de regresión logística en conjunto de datos editado del 2017-2022. *Fuente Propia.*

A partir de los resultados presentados en las Tablas 3.3, 3.4, 3.5 y 3.6, se pueden analizar varias características y tendencias:

- **Clases Problemáticas:** Las clases cincuenta, cuarenta, cuarenta y cinco, y quince presentan particularmente bajos valores en todas las métricas (precisión, exhaustividad y puntaje f1) en casi todos los conjuntos de datos. Esto sugiere que estas clases están siendo especialmente difíciles de clasificar por el modelo.
- **Desequilibrio Evidente:** Dado la baja exhaustividad en varias clases, es evidente que el desequilibrio de clases está afectando el rendimiento del modelo. Las clases minoritarias, como `cincuenta`, `cuarenta`, `cuarenta y cinco` y `quince`, están siendo ignoradas o mal clasificadas por el modelo en favor de clases más predominantes.
- **Mejor Rendimiento en Clases Particulares:** La clase `veinte` tiene consistentemente una precisión y exhaustividad más altos en todas los conjuntos de datos. Aquí se evidencia que esta clase es mayoritaria o que tiene características más distintivas que la hacen más fácil de clasificar.
- **Consistencia en el puntaje f1:** El puntaje f1, que es una métrica que combina precisión y exhaustividad, muestra una tendencia similar en todas los conjuntos de datos. Las clases con baja exhaustividad y precisión también tienen bajos puntajes f1, resaltando nuevamente las clases problemáticas.
- **Rendimiento General:** En general, los valores de precisión, exhaustividad y puntaje f1 son bajos para varias clases en todas los conjuntos de datos. Esto indica que el modelo de regresión logística no está manejando bien el desequilibrio de clases y que podrían ser necesarias estrategias adicionales para abordar este problema.

En resumen, los resultados ilustran claramente los desafíos asociados con el desequilibrio de clases en la tarea de clasificación. Dado este panorama, se ha valorado la importancia de explorar técnicas adicionales para optimizar el rendimiento en las categorías que presentan dificultades de clasificación. Con este propósito, se decidió implementar una variedad de técnicas de balanceo de clases con el objetivo de identificar cuál ofrecería los mejores resultados.

- **Balanceo de clases:** A continuación, se detallan las estrategias utilizadas:
  - **Penalización:** Se incorporó un parámetro adicional en el modelo de Regresión logística, especificando weight = “balanced”. Este ajuste permite que el algoritmo favorezca y equilibre automáticamente a la clase minoritaria durante el proceso de entrenamiento.
  - **Submuestreo "NearMiss"**: Esta técnica involucró la reducción de la clase mayoritaria utilizando un enfoque similar al del algoritmo KNN. La idea es

seleccionar meticulosamente cuáles observaciones de la clase dominante deben ser eliminadas, asegurando que la información esencial se preserve [96].

- **Técnica de sobremuestreo minoritario sintético (SMOTE, *Synthetic Minority Over-sampling Technique*):** Es una técnica de sobremuestreo que genera objetos sintéticos de manera aleatoria entre dos objetos de la clase minoritaria, es decir, SMOTE es una estrategia que identifica puntos vecinos cercanos en la clase minoritaria y crea nuevos puntos sintéticos entre estos nodos, siguiendo una "línea recta"[97].
- **SMOTETomek:** Es una técnica combinada que aplica tanto sobremuestreo como submuestreo al conjunto de datos. Por un lado, utiliza el método SMOTE para sobremuestrear la clase minoritaria, generando ejemplos sintéticos. Por otro lado, emplea los Tomek Links para el submuestreo de la clase mayoritaria. Los Tomek Links identifican pares de instancias que son vecinos más cercanos entre sí pero pertenecen a clases diferentes. Posteriormente, se elimina la instancia de la clase mayoritaria de cada par para conseguir un límite de decisión más claro entre las clases. Esta técnica facilita una definición más precisa del 'decision boundary' o límite de decisión entre las clases [98]."
- **Ensamble de Modelos con Balanceo:** Esta estrategia se fundamenta en la implementación de un clasificador de ensamble de modelos que aborda el desequilibrio de datos mediante el submuestreo aleatorio para equilibrar la distribución de clases en cada subconjunto [99].

Cada una de las técnicas mencionadas anteriormente fue rigurosamente aplicada a los cuatro conjuntos de datos disponibles. Esta implementación sistemática permitió obtener una visión comparativa y exhaustiva de la eficacia de cada método frente a las diferentes características y desafíos presentes en cada conjunto de datos. Con base en los resultados obtenidos de esta aplicación, se ofrece un análisis para cada conjunto de datos, destacando las particularidades observadas:

- **Conjunto de datos 2013-2022:**
  - **Técnica de Penalización:** La exactitud disminuye al 39 % en comparación al conjunto sin aplicar técnicas de balanceo, pero hay un notable aumento en la exhaustividad para clases minoritarias como `cincuenta`, `cuarenta` y `quince`. Esto indica que, aunque se está identificando a más verdaderos positivos, también se está incrementando el número de falsos positivos.
  - **Técnica de Submuestreo:** Esta técnica presenta el más bajo rendimiento con un exactitud del 5 %. Es evidente que el submuestreo extremo de las clases dominantes ha llevado a un rendimiento muy pobre en la clasificación. Los "puntaje f1s" de las clases como `veinte` y `veinticinco` son extremadamente bajos, lo que significa que esta técnica no es adecuada para este conjunto de datos.
  - **Técnica SMOTE:** Después de realizar un sobremuestreo utilizando SMO-

TE, la exactitud es del 33 %. Se observa un equilibrio más adecuado entre precisión y exhaustividad en comparación con el modelo sin balanceo, especialmente para las clases minoritarias. Aunque los números aún no son ideales, esta técnica muestra un mejoramiento respecto al modelo original.

- **Técnica SMOTETomek:** Con esta técnica se combinó sobremuestreo y submuestreo y ofreció una exactitud del 44 %. En general, la mayoría de las clases presentan un puntaje f1 más equilibrado, sugiriendo que esta técnica ofrece un balance adecuado entre identificar correctamente las clases y no marcar incorrectamente las clases opuestas.
- **Técnica de Ensamble de Modelos con Balanceo:** A pesar de que la exactitud es del 33 %, este método muestra un equilibrio notable entre precisión y exhaustividad en varias clases. Sin embargo, aún hay margen de mejora en algunas clases como `cuarenta y cinco` y `cincuenta`.
- **Conclusión:** Al considerar todas las métricas y técnicas aplicadas, la técnica SMOTETomek parece ser la más adecuada para este conjunto de datos. Esta técnica no solo presenta una mejor exactitud en comparación con los modelos de las otras técnicas, sino que también logra un balance más uniforme entre precisión y exhaustividad para las clases, lo cual es esencial cuando se trata de conjuntos de datos desequilibrados.

- **Conjunto de datos editado del 2013-2022:**
  - **Técnica de Penalización:** La exactitud se redujo al 35 %. Aquí, se observa una mejora en la detección de las clases minoritarias, aunque las métricas son bajas. Sin embargo, el compromiso se observa en categorías más grandes como `veinte` y `veinticinco`, cuyo rendimiento disminuye ligeramente.
  - **Técnica de Submuestreo:** Esta técnica llevó a un desplome drástico la exactitud al 4 %. Aunque se logró detectar la clase `quince` con una exhaustividad sorprendente, las demás clases se predijeron con métricas extremadamente bajas, lo que hace a este modelo poco práctico.
  - **Técnica SMOTE:** Con una exactitud del 30 %, SMOTE parece equilibrar un poco la detección entre clases, mostrando mejoras en la detección de clases minoritarias. Aunque las métricas no son óptimas, representan un balance en comparación con las técnicas anteriores.
  - **Técnica SMOTETomek:** Esta técnica mostró una exactitud del 36 %. Aunque la precisión y la exhaustividad no son óptimas para todas las clases, parece haber un equilibrio en el rendimiento entre las diferentes categorías.
  - **Técnica de Ensamble de Modelos con Balanceo:** Esta técnica tuvo una exactitud del 30 %. Se observa una mejora en la detección de clases minoritarias, pero a costa del rendimiento en las categorías más grandes.
  - **Conclusión:** La técnica SMOTETomek demostró ser eficiente al equili-

brar el rendimiento en todas las clases. Esta técnica ofreció valores de precisión y exhaustividad en rangos de 0.08 a 0.47 y de 0.14 a 0.66, respectivamente. A diferencia de escenarios no balanceados, este método garantiza que todas las clases sean predichas con precisión no nula. Es especialmente relevante destacar cómo SMOTETomek mejora significativamente la exhaustividad en clases menos comunes, tales como `cincuenta`, `cuarenta y cinco` y `quince`.

Aunque en las clases más prevalentes, como `veinte` y `veinticinco`, el rendimiento no fue superior al del escenario sin balanceo, es esencial considerar que un rendimiento balanceado en todas las clases puede ser crucial. Esto es particularmente cierto si al no identificar correctamente las clases menos representadas se conlleva un alto costo.

- **Conjunto de datos del 2017-2022:**
  - **Técnica de Penalización:** La técnica de penalización mejora notablemente la exhaustividad para categorías minoritarias teniendo un balanceo más uniforme, por ejemplo, `cincuenta` ahora tiene una exhaustividad de 0.60. Sin embargo, la precisión en general ha disminuido para varias categorías. La exactitud se ha reducido a 42 %.
  - **Técnica de Submuestreo:** Esta técnica ha logrado incrementar la exhaustividad de la categoría `quince` a 1.00, pero la precisión general es bastante baja para todas las categorías. La exactitud es de 0.03, lo que indica que la técnica de submuestreo no ha sido efectiva en este conjunto de datos.
  - **Técnica SMOTE:** SMOTE ha mostrado una mejora considerable en la exhaustividad de categorías como `cincuenta`, `cuarenta y cinco` y `quince`. Sin embargo, al igual que la técnica de penalización, sacrifica precisión para equilibrar la exhaustividad. Con esta técnica se obtuvo una exactitud del 40 %.
  - **Técnica SMOTETomek:** SMOTETomek ha logrado un equilibrio razonable entre precisión y exhaustividad en comparación con las técnicas anteriores. Las métricas generales indican que es una técnica prometedora para este conjunto de datos, con una exactitud de 46 %.
  - **Técnica de Ensamble de Modelos con Balanceo:** El ensamblaje de modelos con balanceo proporciona métricas similares a las de SMOTE. Mientras que se mejoró la exhaustividad de las categorías menos representadas, la precisión general es menor en comparación con el conjunto de datos sin balanceo. Obtuvo una exactitud del 39 %.
  - **Conclusión:** Teniendo en cuenta los resultados, se consideraría que SMOTETomek es el mejor método de balanceo en este caso, ya que logra un equilibrio razonable entre precisión, exhaustividad y puntaje f1 para todas las clases. Mientras que la precisión puede haber disminuido un poco para las clases minoritarias en comparación con el conjunto sin balanceo, la exhaustividad ha mejorado significativamente, lo que signi-

fica que el modelo es capaz de identificar correctamente más ejemplos de estas clases. La precisión y exhaustividad para la clase mayoritaria también sigue siendo razonablemente alta.

- **Conjunto de datos editado del 2017-2022:**
  - **Técnica de Penalización:** Con la técnica de penalización, el rendimiento general del modelo disminuyó ligeramente con una exactitud del 35 %. Sin embargo, todas las clases tienen valores distintos de cero para precisión y exhaustividad, lo que indica que el modelo ahora es capaz de reconocer todas las clases, aunque con variados niveles de éxito. Es notable que la clase `quince` tiene una exhaustividad del 0.67, lo que sugiere que la penalización permitió al modelo identificar esta clase más efectivamente en comparación con el conjunto de datos sin balanceo.
  - **Técnica de Submuestreo:** Esta técnica ha logrado incrementar el exhaustividad de la categoría `quince` a 1.00, pero la precisión general es bastante baja para todas las categorías. La exactitud es del 3 %, lo que indica que la técnica de submuestreo no ha sido efectiva en este conjunto de datos.
  - **Técnica SMOTE:** SMOTE ha mostrado una mejora considerable en la exhaustividad de categorías como `cincuenta`, `cuarenta y cinco` y `quince`. Sin embargo, al igual que la técnica de penalización, sacrifica precisión para equilibrar la exhaustividad. La exactitud obtenida fue de 30 %.
  - **Técnica SMOTETomek:** Al combinar las técnicas de sobremuestreo y limpieza, SMOTETomek ha logrado un equilibrio razonable entre precisión y exhaustividad en comparación con las técnicas anteriores. Las métricas generales indican que es una técnica prometedora para este conjunto de datos, con una exactitud del 40 %.
  - **Técnica de Ensamble de Modelos con Balanceo:** El ensamblaje de modelos con balanceo proporciona métricas similares a las de SMOTE. Mientras que se mejoró la exhaustividad de las categorías menos representadas, la exactitud es menor en comparación con el conjunto de datos sin balanceo con un 31 %.
  - **Conclusión:** Dado que el método SMOTETomek no resultó ser el más óptimo para este conjunto de datos en particular, se procedió a realizar un análisis más detallado para determinar cuál técnica de balanceo es la más apropiada. Para ello, se revisó nuevamente los métodos de balanceo utilizando un algoritmo más robusto que la regresión logística; en este caso, se optó por el algoritmo RF.

    De este nuevo análisis se obtuvieron los siguientes resultados:

    - **Penalización:** Este método tiene una precisión y exactitud similar al enfoque sin balanceo, pero aún así no logra mejorar la exhaustividad para las clases minoritarias.
    - **Submuestreo:** Aunque este método mejoró la exhaustividad para

algunas clases minoritarias, se produjo una caída significativa en la precisión y una exactitud en general.

- ◇ **SMOTE:** Este método también mejoró la exhaustividad para las clases minoritarias y tuvo una precisión y exactitud del 76 %.
- ◇ **SMOTETomek:** Este método mejoró la exhaustividad para las clases minoritarias sin comprometer demasiado la precisión. Tiene una exactitud del 75
- ◇ **Técnica de Ensamble de Modelos con Balanceo:** Este método mostró una mejora significativa en la exhaustividad para las clases minoritarias, pero a costa de una caída en la precisión.

La técnica de SMOTE proporcionó el mejor equilibrio entre la precisión y la exhaustividad. Esta técnica logró una precisión y exactitud similar al del modelo sin balanceo, pero con una mejora considerable en la exhaustividad de las clases minoritarias. Esto significa que la técnica SMOTE es capaz de mantener una alta tasa de predicciones correctas, al tiempo que mejora la capacidad del modelo para identificar correctamente las clases minoritarias.

Para visualizar el proceso de balanceo de clases realizado dirigirse al repositorio en Github [9].

En resumen, SMOTETomek demostró ser la técnica más eficaz para equilibrar el conjunto de datos 2013-2022, el conjunto de datos editado del 2013-2022 y el conjunto de datos 2017-2022. Sin embargo, para el conjunto de datos 2017-2022 la técnica SMOTE sobresalió como la mejor opción. Una vez identificado el método de balanceo óptimo, se generaron cuatro conjuntos de datos balanceados, preparados específicamente para la etapa de modelado.

- **Modelado:** Con los conjuntos de datos ya equilibrados, se llevó a cabo el proceso de modelado utilizando varios algoritmos con el objetivo de desarrollar modelos predictivos para determinar la dosis de coagulante. Los algoritmos elegidos para esta tarea incluyen RF, ERT, KNN, DT y GB. La elección de estos algoritmos se fundamentó en su popularidad y amplio uso en muchas áreas tal como se detalla en el Anexo A. Para obtener información adicional sobre estos algoritmos, se recomienda consultar dicho Anexo. Es importante destacar que, aunque las ANN han sido el algoritmo predominante en este ámbito de investigación, según se detalla en la sección 2.2, el enfoque de este estudio se inclinó hacia algoritmos menos explorados en investigaciones anteriores, con el propósito de diversificar y enriquecer el análisis.

  Con estos algoritmos en mente, se optó por la implementación de los algoritmos mediante el uso de la biblioteca `sklearn` de Python. Esta herramienta, conocida como Sci-kit Learn, está diseñada específicamente para facilitar y optimizar el análisis predictivo de datos. Una de las grandes ventajas de esta librería es que es completamente accesible para personas de cualquier nivel de experiencia en

[9] Para observar el proceso de balanceo de clases realizado, dirigirse al siguiente link: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/13_Preproccessing.ipynb`

programación. Además, está estructurada de tal manera que sus componentes se pueden adaptar y reutilizar en una variedad de contextos diferentes. Es importante resaltar que Sci-kit Learn no es una entidad aislada; en realidad, se fundamenta en otras bibliotecas reconocidas de Python: NumPy, SciPy y Matplotlib. Esta combinación le brinda una capacidad de integración armoniosa y una vasta gama de funciones.

A continuación, se detallan los módulos específicos de `sklearn` que fueron empleados en el proyecto mediante la Tabla 3.7:

Algoritmo	**Módulo en sci-kit learn**
RF	`from sklearn.ensemble import RandomForestClassifier`
ERT	`from sklearn.ensemble import ExtraTreesClassifier`
GB	`from sklearn.ensemble import GradientBoostingClassifier`
DT	`from sklearn.tree import DecisionTreeClassifier`
KNN	`from sklearn.neighbors import KNeighborsClassifier`

Tabla 3.7: Módulos de sci-kit learn utilizados para la implementación de los algoritmos de clasificación. *Fuente Propia.*

En el estudio, se identificaron aquellos modelos que mostraron un rendimiento superior, reflejado en las métricas obtenidas para cada conjunto de datos.En particular, ERT y RF demostraron ser los más eficientes. Una discusión más exhaustiva de estos hallazgos se encuentra en la sección 4.1. Una vez identificados estos modelos óptimos, se realizó un ajuste y optimización de sus hiperparámetros. Para ERT y RF, estos hiperparámetros incluyen `n_estimators`, `max_depth` y `min_samples_split`, aspectos que también se detallan en la sección 4.1. Este proceso tenía como objetivo no solo mejorar la eficacia de los modelos sino también prevenir cualquier posibilidad de sobreajuste.

Modelo de regresión.

Al abordar el modelado de regresión, la naturaleza numérica de la variable objetivo eliminaba la necesidad de una transformación inicial a variable categórica, al contrario de lo que ocurre en la clasificación. Sin embargo, un obstáculo común a ambos enfoques fue que los datos no presentaban una distribución normal. Esta observación condujo a considerar la posibilidad de transformar la variable `coagulante_dosis` para equilibrar la distribución de sus valores.

Las transformaciones de variables representan un paso esencial en muchos procedimientos analíticos, sobre todo cuando se buscan cumplir ciertos supuestos como la normalidad o la linealidad. En este proyecto, se implementó la transformación de Box-Cox sobre la variable de dosis de coagulante. Esta metodología, introducida por Box y Cox en 1964, propone una serie de transformaciones potenciales con el objetivo de converger hacia una distribución normal de los datos [100].

A pesar de la promesa teórica detrás de la transformación de Box-Cox, la evaluación preliminar indicó que la aplicación de esta técnica no proporcionó mejoras sustanciales para el conjunto de datos particular con el que se estaba trabajando. Esta conclusión se derivó a partir de pruebas utilizando un modelo de regresión lineal. Debido a esta

observación, se decidió continuar con los análisis empleando la variable en su forma original, sin aplicar transformaciones adicionales. Además, existen algoritmos que son sensibles a la escala o distribución de los datos, mientras que otros no [101]. Por lo tanto, se avanzó con la evaluación de múltiples algoritmos sin la aplicación de la transformada.

- **Modelado:** En el proceso de modelado de regresión, al igual que en el enfoque de clasificación, se llevó a cabo una serie de experimentos utilizando diversos algoritmos. El objetivo fundamental de este ejercicio fue desarrollar modelos predictivos robustos y precisos que permitan estimar de manera efectiva la dosis de coagulante necesaria. Entre los algoritmos seleccionados para esta tarea se encuentran RF, ERT, KNN, DT y GB.

  Para la implementación y entrenamiento de estos modelos, se recurrió nuevamente a la librería de Python conocida como `sklearn`. Sin embargo, a diferencia del enfoque de clasificación, en este contexto se hizo uso de los módulos específicos destinados para regresión que ofrece esta librería.

  En la Tabla 3.8 se pueden encontrar los detalles acerca de los módulos específicos que fueron empleados en el proceso de modelado de regresión.

Algoritmo	**Módulo en sci-kit learn**
RF	`from sklearn.ensemble import RandomForestRegressor`
ERT	`from sklearn.ensemble import ExtraTreesRegressor`
GB	`from sklearn.ensemble import GradientBoostingRegressor`
DT	`from sklearn.tree import DecisionTreeRegressor`
KNN	`from sklearn.neighbors import KNeighborsRegressor`

Tabla 3.8: Módulos de sci-kit learn utilizados para la implementación de los algoritmos de regresión. *Fuente Propia.*

- **Métricas de evaluación:** A fin de determinar qué modelos ofrecían un rendimiento superior, se recurrió a un conjunto específico de métricas evaluativas. En el contexto de la regresión, es común emplear métricas que midan el error de las predicciones para evaluar la calidad y precisión de los modelos desarrollados. En este proyecto, con el propósito de realizar una evaluación exhaustiva y detallada, se seleccionaron las siguientes métricas de error para la tarea:

  - **MSE:** Es una medida que representa la calidad del modelo, y se define como el promedio de los cuadrados de los errores o desviaciones, es decir, la diferencia entre el estimador y lo que se estima. El MSE mide la cantidad de error que hay entre dos conjuntos de datos, en este caso, entre los valores predichos y los valores reales. Se calcula utilizando la fórmula:

    $$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

    donde y_i es el valor real, $\hat{y}_i$ es el valor predicho y n es el número de observaciones [102].

  - **RMSE:** Esta métrica es simplemente la raíz cuadrada del MSE. El RMSE da una interpretación más directa del error del modelo en las unidades originales

de la variable de salida, siendo una medida de dispersión [102]. Se calcula como:

$$RMSE = \sqrt{MSE}$$

- **MAE:** El MAE mide la magnitud promedio de los errores entre los valores predichos y los observados, sin distinguir si los errores resultan de haber subestimado o sobreestimado los valores reales. En esencia, el MAE se centra solamente en la magnitud del error, sin considerar si la predicción fue superior o inferior al valor real. Es el promedio, en la muestra de prueba, de las diferencias absolutas entre las predicciones y las observaciones reales. Su fórmula es:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

  donde, similar al MSE, y_i es el valor real y $\hat{y}_i$ es el valor predicho [102].

Durante el proceso de evaluación, la elección de los modelos óptimos se basó en una rigurosa revisión de las métricas establecidas. Esta estrategia garantizaba que solo se seleccionaran aquellos modelos que demostraran una capacidad superior en términos de precisión y eficiencia en sus predicciones. Esta meticulosa selección se diseñó para asegurar que los modelos no solo fueran precisos en su rendimiento, sino que también maximizaran la eficiencia en el manejo de los datos y en la generación de resultados. La confiabilidad de estas métricas proporciona una base sólida para tomar decisiones informadas en la fase de selección del modelo, garantizando que los modelos escogidos sean los más adecuados para el propósito del estudio.

Adicionalmente, se efectuó una optimización de hiperparámetros. En la sección 4.2 se podrá ver a más detalle esta optimización. Esta fase de la investigación tuvo un doble objetivo: mejorar la eficacia de los modelos y reducir el riesgo de sobreajuste. De este modo, se aseguró que los modelos resultantes fueran robustos y capaces de generalizar bien frente a nuevos datos. Como complemento a esta fase, se implementó la técnica de validación cruzada, un método que divide el conjunto de datos en k grupos, utilizando alternativamente un grupo como conjunto de validación y el resto para entrenamiento. A través de este proceso, cada uno de los grupos tuvo la oportunidad de servir como conjunto de validación.

Esta estrategia, respaldada por [103, 90], es esencial para detectar si un modelo presenta variaciones significativas de rendimiento a través de los diferentes pliegues o 'folds'. Variaciones notables podrían señalar un sobreajuste a ciertas partes del conjunto de datos. La validación cruzada tiene ventajas adicionales: al entrenar el modelo en diversos subconjuntos, se adquiere un entendimiento más claro de cómo el modelo se comportará frente a datos no vistos anteriormente. Además, al no depender solo de una partición aleatoria, se minimizan sesgos en la evaluación del rendimiento del modelo, resultando en una estimación más robusta de su capacidad predictiva.

Para una exploración detallada del proceso de optimización, los resultados de la validación cruzada y los insights derivados, se invita al lector a revisar la sección 4.2, donde se explica todo con mayor profundidad.

3.3 Resumen

El capítulo inició con la presentación del marco de trabajo utilizado: CRISP ML(Q), que sirvió como guía para desarrollar el sistema de recomendación de dosis de coagulante mediante ML. Siguiendo la estructura propuesta por este marco:

- **Definición del Alcance de Aplicación de ML:** Se precisó el objetivo del trabajo de grado y se evaluó su factibilidad.
- **Fuentes de Datos:** Se describieron las dos fuentes principales de donde se extrajeron los datos para el estudio.
- **Construcción del Conjunto de Datos:** A través de un EDA inicial, se tomaron decisiones clave. Una de estas fue la creación de un conjunto adicional de datos, en el cual se imputaron valores nulos con ceros. A partir de un análisis de valores atípicos, se optó por eliminar ciertos registros basándose en criterios definidos. Además, considerando la insuficiencia de datos de algunas variables en años anteriores, se generaron dos conjuntos de datos adicionales enfocados en registros desde el año 2017 en adelante.
- **Análisis de Valores Faltantes:** Esta fase se centró en identificar los mecanismos de datos faltantes asociados a cada variable. El propósito era garantizar una imputación adecuada, evitando la introducción de sesgos.
- **Modelado:** Aquí se detallaron los enfoques adoptados, tanto de clasificación como de regresión. Para el enfoque de clasificación, se destacó la importancia de equilibrar los datos dada la presencia de desequilibrio de clases. Además, se explicaron en detalle las métricas seleccionadas y técnicas para evaluar y optimizar el rendimiento de los modelos.

Este capítulo proporcionó un recorrido exhaustivo por todo el proceso de modelado, desde la recopilación inicial de datos hasta las técnicas de evaluación y optimización.

Capítulo 4.

Resultados y Discusión

Una vez construidos los cuatro conjuntos de datos se seleccionaron algoritmos específicos para actuar como clasificadores y regresores, tal como se detalla en el capítulo 3. A partir de estos algoritmos, se generaron modelos de ML, diseñados para predecir las dosis adecuadas de coagulante. Este capítulo se dedica a un análisis exhaustivo de los resultados obtenidos por estos modelos.

En la sección 4.1, se examina el rendimiento de los modelos cuando el problema de predicción de dosis de coagulante se aborda como una tarea de clasificación. Se discuten métricas relevantes, como precisión, exhaustividad y puntaje f1, y cómo estas métricas varían entre los distintos conjuntos de datos.

Por otro lado, la sección 4.2 se enfoca en el escenario en el que la predicción de las dosis de coagulante se trata como un problema de regresión. En esta sección, se evalúan y comparan diferentes métricas de regresión, como el MSE, RMSE y MAE, para distintos modelos y conjuntos de datos.

4.1 Modelado: Enfoque de Clasificación

En el capítulo anterior, se exploraron diversas técnicas de balanceo de clases, buscando la que ofreciera el mejor equilibrio entre las distintas categorías de los conjuntos de datos. SMOTETomek emergió como la técnica más apropiada para los tres primeros conjuntos, mientras que SMOTE resultó ser óptima para el cuarto conjunto. Estas elecciones dieron como resultado conjuntos con distribuciones de clases notablemente más equilibradas.

Para observar estas transformaciones, se presentan representaciones gráficas. Las Figuras 4.1, 4.2 y 4.3 muestran cómo, gracias a SMOTETomek, clases como `cuarenta y cinco` y `cincuenta`, que antes eran escasamente representadas, ahora tienen mayor prominencia. Estas representaciones actualizadas pueden compararse con las distribuciones iniciales en las Figuras 3.20, 3.21 y 3.22. Por su parte, la Figura 4.4 destaca una distribución donde todas las clases tienen una representación equitativa.

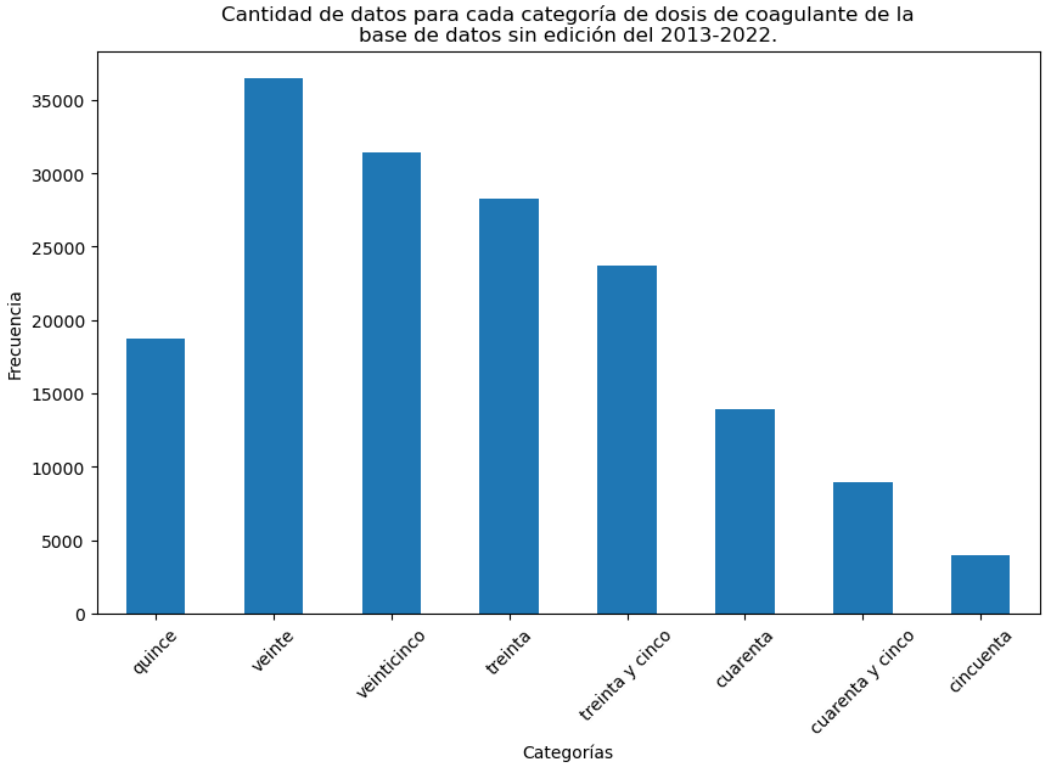

Figura 4.1: Distribución de datos de dosis de coagulante del conjunto de datos 2013-2022 balanceado. *Fuente Propia.*

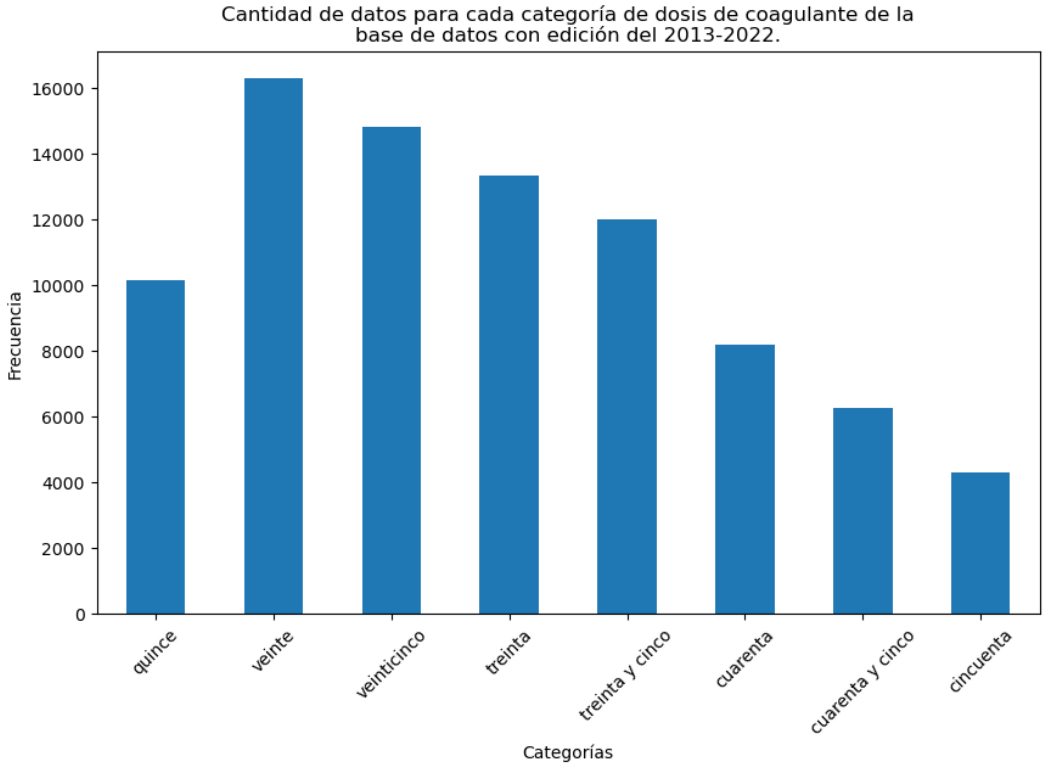

Figura 4.2: Distribución de datos de dosis de coagulante del conjunto de datos editado del 2013-2022 balanceado. *Fuente Propia.*

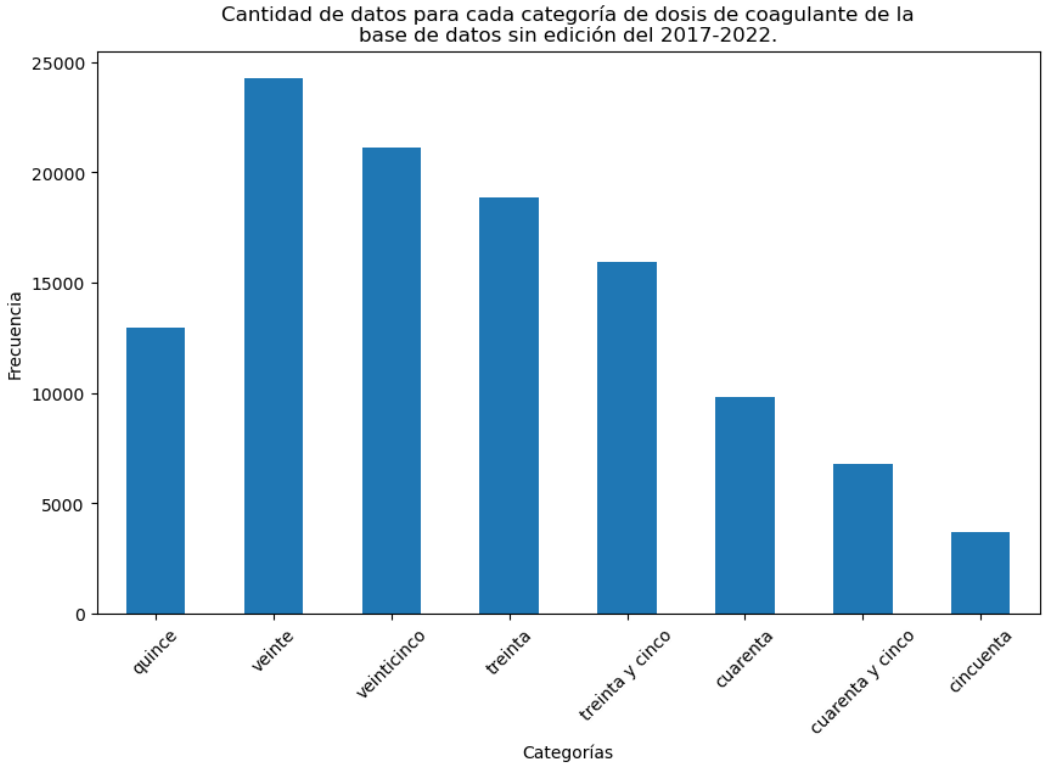

Figura 4.3: Distribución de datos de dosis de coagulante del conjunto de datos 2017-2022 balanceado. *Fuente Propia.*

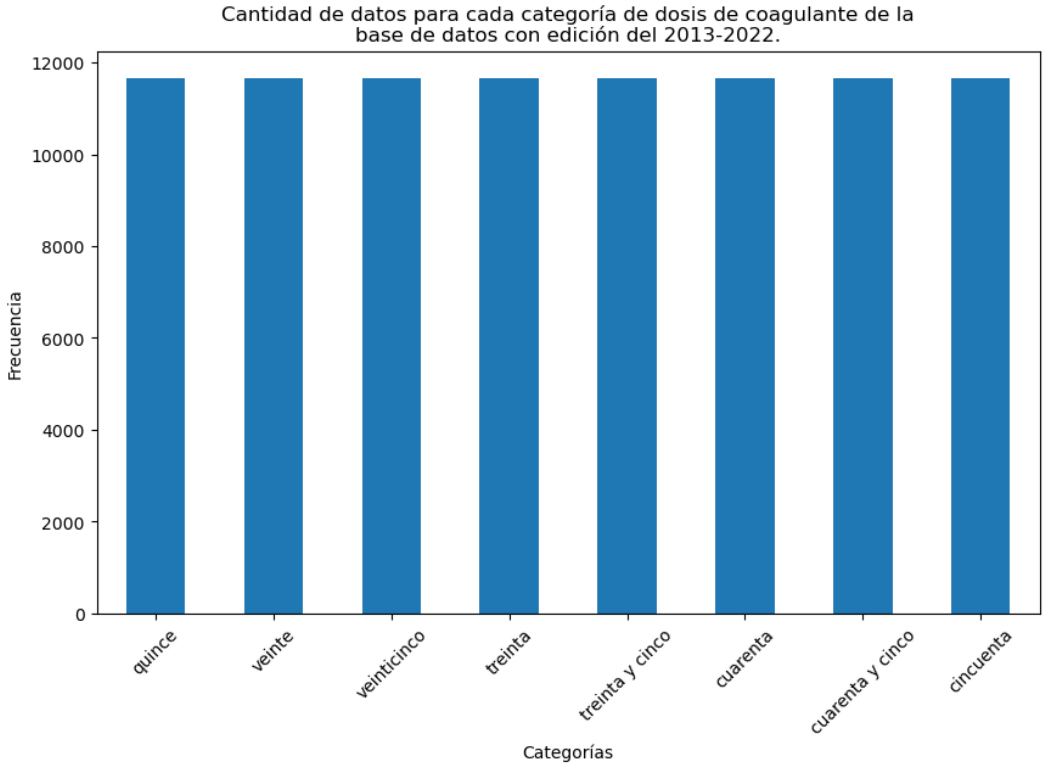

Figura 4.4: Distribución de datos de dosis de coagulante del conjunto de datos editado del 2017-2022 balanceado. *Fuente Propia.*

En el proceso de modelado de los conjuntos de datos, se utilizaron diversos algoritmos de aprendizaje automático de la biblioteca `sklearn`, tal y como se especificó en el capítulo 3. Entre los algoritmos empleados se encuentran GB, DT, KNN, RF y ERT, todos clasificadores. Estos algoritmos se implementaron con los hiperparámetros predeterminados que proporciona `sklearn.ensemble`. Para una revisión detallada de estos hiperparámetros, se puede consultar directamente la página oficial de scikit-learn[1].

Para evaluar y comparar el rendimiento de estos modelos, se emplearon diversas métricas, también explicadas en el capítulo 3. Estas métricas son: precisión, exhaustividad, puntaje f1 y exactitud. El objetivo es identificar el modelo que exhiba un rendimiento superior en términos de estas métricas.

En particular, se busca que el modelo seleccionado presente valores de precisión, exhaustividad y puntaje f1 equilibrados y cercanos a 1, lo que indicaría una alta calidad en las predicciones. Además, se aspira a que el modelo tenga un exactitud también cercano a 1, señalando que una gran proporción de las predicciones realizadas por el modelo son correctas.

4.1.1 Resultados de los modelos

- **Conjunto de datos 2013-2022.**
  - **GB**

	precision	recall	f1-score	support
cincuenta	0.71	0.61	0.66	1200
cuarenta	0.52	0.46	0.49	4198
cuarenta y cinco	0.61	0.56	0.58	2651
quince	0.75	0.77	0.76	5679
treinta	0.51	0.43	0.47	8533
treinta y cinco	0.44	0.58	0.50	6984
veinte	0.71	0.74	0.73	10887
veinticinco	0.59	0.55	0.57	9496
accuracy			0.60	49628
macro avg	0.60	0.59	0.59	49628
weighted avg	0.60	0.60	0.59	49628

Figura 4.5: Resultados de GB en clasificación Conjunto de datos 2013-2022. *Fuente Propia*

  - **DT**

[1]Para más información sobre `sklearn.ensemble`, consultar: `https://scikit-learn.org/stable/modules/ensemble.html`

	precision	recall	f1-score	support
cincuenta	0.85	0.90	0.88	1200
cuarenta	0.80	0.81	0.81	4198
cuarenta y cinco	0.86	0.88	0.87	2651
quince	0.84	0.84	0.84	5679
treinta	0.70	0.69	0.69	8533
treinta y cinco	0.74	0.76	0.75	6984
veinte	0.78	0.77	0.77	10887
veinticinco	0.68	0.67	0.67	9496
accuracy			0.75	49628
macro avg	0.78	0.79	0.78	49628
weighted avg	0.75	0.75	0.75	49628

Figura 4.6: Resultados de DT en clasificación conjunto de datos 2013-2022. *Fuente Propia*

- **KNN**

	precision	recall	f1-score	support
cincuenta	0.72	0.84	0.78	1200
cuarenta	0.73	0.82	0.77	4198
cuarenta y cinco	0.75	0.82	0.78	2651
quince	0.77	0.89	0.82	5679
treinta	0.68	0.71	0.70	8533
treinta y cinco	0.72	0.78	0.75	6984
veinte	0.79	0.72	0.75	10887
veinticinco	0.73	0.60	0.66	9496
accuracy			0.74	49628
macro avg	0.74	0.77	0.75	49628
weighted avg	0.74	0.74	0.74	49628

Figura 4.7: Resultados de KNN en clasificación conjunto de datos 2013-2022. *Fuente Propia*

- **RF**

	precision	recall	f1-score	support
cincuenta	0.96	0.96	0.96	1200
cuarenta	0.90	0.93	0.91	4198
cuarenta y cinco	0.94	0.96	0.95	2651
quince	0.92	0.93	0.92	5679
treinta	0.83	0.80	0.82	8533
treinta y cinco	0.83	0.90	0.87	6984
veinte	0.86	0.87	0.86	10887
veinticinco	0.82	0.76	0.79	9496
accuracy			0.86	49628
macro avg	0.88	0.89	0.88	49628
weighted avg	0.86	0.86	0.86	49628

Figura 4.8: Resultados de RF en clasificación conjunto de datos 2013-2022. *Fuente Propia*

- **ERT**

	precision	recall	f1-score	support
cincuenta	0.98	0.97	0.98	1200
cuarenta	0.92	0.94	0.93	4198
cuarenta y cinco	0.95	0.97	0.96	2651
quince	0.93	0.93	0.93	5679
treinta	0.85	0.81	0.83	8533
treinta y cinco	0.85	0.91	0.88	6984
veinte	0.85	0.89	0.87	10887
veinticinco	0.82	0.77	0.79	9496
accuracy			0.87	49628
macro avg	0.89	0.90	0.90	49628
weighted avg	0.87	0.87	0.87	49628

Figura 4.9: Resultados de ERT en clasificación conjunto de datos 2013-2022. *Fuente Propia*

A continuación, se presenta un análisis en base a los resultados presentados en las Figuras 4.5, 4.6, 4.7, 4.8 y 4.9, obtenidos de los modelos implementados:

- **RF**: El modelo RF mostró una exactitud general del 0.86. Entre las clases, la clase `cincuenta` obtuvo los valores más altos de precisión, exhaustividad y puntaje f1, mientras que la clase `veinticinco` presentó el valor más bajo en precisión. En general, este modelo demostró tener un buen equilibrio entre las métricas, con un promedio ponderado en todas las métricas de 0.86.
- **ERT**: ERT superó ligeramente al modelo RF con una exactitud general de 0.87. Similar al modelo RF, la clase `cincuenta` obtuvo los valores más altos en todas las métricas. Sin embargo, ERT mejoró la precisión en comparación con RF en la mayoría de las clases, lo que indica una mejora en la identificación correcta de las clases. En resumen, ERT mostró un rendimiento ligeramente superior al RF, con un promedio ponderado en todas las métricas de 0.87.
- **KNN**: Este modelo tuvo un rendimiento menor con una exactitud general de 0.74. Si bien `cincuenta` sigue siendo la clase con mejor rendimiento, se observa una disminución significativa en las métricas en comparación con los dos primeros modelos. El promedio ponderado en todas las métricas fue de 0.74.
- **DT**: DT presentó una exactitud general de 0.75. A pesar de superar a KNN en precisión general, aún está detrás de RF y ERT. La clase `cincuenta` también se destacó en este modelo, pero se observa una menor precisión en la clase `veinticinco`. El promedio ponderado en todas las métricas fue de 0.75.
- **GB**: GB tuvo el rendimiento más bajo entre todos los modelos con una exactitud general de 0.60. A pesar de que la clase `quince` mostró un buen puntaje f1, muchas de las otras clases como `cuarenta` y `treinta` tuvieron métricas significativamente más bajas. El promedio ponderado en todas las métricas fue de 0.59.

En resumen, **ERT** y **RF** son los dos modelos que presentaron los mejores resultados en general. Ambos modelos tuvieron métricas balanceadas y cercanas, con

ERT mostrando una ligera ventaja sobre RF. Por otro lado, GB tuvo el rendimiento más bajo entre los modelos analizados.

- **Conjunto de datos editado del 2013-2022.**

  - **GB**

	precision	recall	f1-score	support
cincuenta	0.68	0.72	0.70	1302
cuarenta	0.51	0.45	0.48	2486
cuarenta y cinco	0.60	0.64	0.62	1915
quince	0.70	0.72	0.71	3093
treinta	0.50	0.36	0.42	4035
treinta y cinco	0.45	0.48	0.46	3552
veinte	0.57	0.70	0.62	4778
veinticinco	0.50	0.48	0.49	4419
accuracy			0.55	25580
macro avg	0.56	0.57	0.56	25580
weighted avg	0.55	0.55	0.55	25580

Figura 4.10: Resultados de GB en clasificación conjunto de datos editada del 2013-2022. *Fuente Propia*

  - **DT**

	precision	recall	f1-score	support
cincuenta	0.88	0.88	0.88	1302
cuarenta	0.77	0.78	0.77	2486
cuarenta y cinco	0.83	0.88	0.86	1915
quince	0.82	0.81	0.81	3093
treinta	0.61	0.59	0.60	4035
treinta y cinco	0.65	0.65	0.65	3552
veinte	0.67	0.68	0.67	4778
veinticinco	0.59	0.58	0.58	4419
accuracy			0.70	25580
macro avg	0.73	0.73	0.73	25580
weighted avg	0.69	0.70	0.69	25580

Figura 4.11: Resultados de DT en clasificación conjunto de datos editada del 2013-2022. *Fuente Propia*

  - **KNN**

	precision	recall	f1-score	support
cincuenta	0.74	0.87	0.80	1302
cuarenta	0.68	0.78	0.73	2486
cuarenta y cinco	0.71	0.81	0.76	1915
quince	0.71	0.86	0.78	3093
treinta	0.57	0.58	0.58	4035
treinta y cinco	0.64	0.64	0.64	3552
veinte	0.65	0.60	0.63	4778
veinticinco	0.61	0.44	0.51	4419
accuracy			0.66	25580
macro avg	0.67	0.70	0.68	25580
weighted avg	0.65	0.66	0.65	25580

Figura 4.12: Resultados de KNN en clasificación conjunto de datos editada del 2013-2022. *Fuente Propia*

- **RF**

	precision	recall	f1-score	support
cincuenta	0.95	0.98	0.97	1302
cuarenta	0.89	0.91	0.90	2486
cuarenta y cinco	0.92	0.96	0.94	1915
quince	0.90	0.93	0.92	3093
treinta	0.78	0.71	0.75	4035
treinta y cinco	0.80	0.83	0.81	3552
veinte	0.77	0.82	0.79	4778
veinticinco	0.75	0.69	0.72	4419
accuracy			0.82	25580
macro avg	0.85	0.85	0.85	25580
weighted avg	0.82	0.82	0.82	25580

Figura 4.13: Resultados de RF en clasificación conjunto de datos editada del 2013-2022. *Fuente Propia*

- **ERT**

	precision	recall	f1-score	support
cincuenta	0.96	0.98	0.97	1302
cuarenta	0.91	0.92	0.91	2486
cuarenta y cinco	0.93	0.97	0.95	1915
quince	0.90	0.93	0.92	3093
treinta	0.82	0.73	0.77	4035
treinta y cinco	0.83	0.85	0.84	3552
veinte	0.77	0.83	0.80	4778
veinticinco	0.76	0.71	0.73	4419
accuracy			0.83	25580
macro avg	0.86	0.86	0.86	25580
weighted avg	0.83	0.83	0.83	25580

Figura 4.14: Resultados de ERT en clasificación conjunto de datos editada del 2013-2022. *Fuente Propia*

Basándose en los resultados presentados en las Figuras 4.20, 4.21, 4.22, 4.23 y 4.24 se puede analizar lo siguiente:

- **RF:** Este modelo presenta un exactitud general de 0.82. Las clases con las métricas más altas en términos de puntaje f1 son `cincuenta` y `cuarenta y cinco` con valores de 0.97 y 0.94 respectivamente. La clase `treinta` es la que tuvo el menor puntaje f1 con 0.75.
- **ERT:** Con una exactitud de 0.83, este modelo mejora ligeramente sobre el RF. Al igual que con el modelo anterior, las clases `cincuenta` y `cuarenta y cinco` tienen el puntaje f1 más alto con 0.97 y 0.95 respectivamente. Nuevamente, `treinta` es la clase con el menor puntaje f1 con 0.77.
- **KNN:** Este modelo tiene una exactitud de 0.66, siendo menor que los dos anteriores. Las clases `cincuenta` y `quince` tienen puntajes f1 de 0.80 y 0.78 respectivamente. La clase `veinticinco` tiene el menor puntaje f1 con 0.51.
- **DT:** Este modelo tiene una exactitud de 0.70. La clase `cincuenta` tiene el puntaje f1 más alto con 0.88, y "treinta" es la clase con el menor puntaje f1 con 0.60.
- **GB:** Este modelo presenta la menor exactitud de todas, con un valor de 0.55. La clase con el puntaje f1 más alto es `cincuenta` con 0.70, mientras que `treinta` tiene el puntaje f1 más bajo con 0.42.

En resumen, **RF** y **ERT** presentan los mejores resultados generales en comparación con los otros algoritmos, tanto en exactitud como en métricas individuales por clase.Por otro lado, el modelo **GB** es el que tuvo el rendimiento más bajo entre todos. Además, la clase `treinta` consistentemente tuvo el menor puntaje f1 en todos los modelos, lo que sugiere que puede haber un desafío en la clasificación de esta categoría específica en el conjunto de datos dado.

- **Conjunto de datos 2017-2022.**
  - **GB**

	precision	recall	f1-score	support
cincuenta	0.85	0.81	0.83	1118
cuarenta	0.62	0.58	0.60	2939
cuarenta y cinco	0.77	0.73	0.75	2070
quince	0.84	0.86	0.85	3996
treinta	0.55	0.50	0.52	5621
treinta y cinco	0.50	0.65	0.56	4711
veinte	0.77	0.76	0.76	7222
veinticinco	0.62	0.55	0.58	6315
accuracy			0.66	33992
macro avg	0.69	0.68	0.68	33992
weighted avg	0.66	0.66	0.66	33992

Figura 4.15: Resultados de GB en clasificación conjunto de datos 2017-2022. *Fuente Propia*

  - **DT**

	precision	recall	f1-score	support
cincuenta	0.94	0.95	0.95	1118
cuarenta	0.85	0.87	0.86	2939
cuarenta y cinco	0.91	0.93	0.92	2070
quince	0.91	0.92	0.92	3996
treinta	0.75	0.75	0.75	5621
treinta y cinco	0.80	0.80	0.80	4711
veinte	0.81	0.80	0.80	7222
veinticinco	0.71	0.70	0.71	6315
accuracy			0.81	33992
macro avg	0.84	0.84	0.84	33992
weighted avg	0.81	0.81	0.81	33992

Figura 4.16: Resultados de DT en clasificación conjunto de datos 2017-2022. *Fuente Propia*

- **KNN**

	precision	recall	f1-score	support
cincuenta	0.83	0.93	0.87	1118
cuarenta	0.76	0.84	0.80	2939
cuarenta y cinco	0.79	0.86	0.83	2070
quince	0.85	0.95	0.90	3996
treinta	0.72	0.75	0.73	5621
treinta y cinco	0.76	0.80	0.78	4711
veinte	0.83	0.74	0.78	7222
veinticinco	0.75	0.66	0.71	6315
accuracy			0.78	33992
macro avg	0.79	0.82	0.80	33992
weighted avg	0.78	0.78	0.78	33992

Figura 4.17: Resultados de KNN en clasificación conjunto de datos 2017-2022. *Fuente Propia*

- **RF**

	precision	recall	f1-score	support
cincuenta	0.99	0.99	0.99	1118
cuarenta	0.92	0.95	0.94	2939
cuarenta y cinco	0.97	0.99	0.98	2070
quince	0.95	0.97	0.96	3996
treinta	0.87	0.84	0.86	5621
treinta y cinco	0.86	0.92	0.89	4711
veinte	0.89	0.88	0.89	7222
veinticinco	0.85	0.80	0.82	6315
accuracy			0.89	33992
macro avg	0.91	0.92	0.92	33992
weighted avg	0.89	0.89	0.89	33992

Figura 4.18: Resultados de RF en clasificación conjunto de datos 2017-2022. *Fuente Propia*

- **ERT**

	precision	recall	f1-score	support
cincuenta	0.99	0.99	0.99	1118
cuarenta	0.95	0.96	0.95	2939
cuarenta y cinco	0.98	0.99	0.98	2070
quince	0.96	0.98	0.97	3996
treinta	0.88	0.85	0.86	5621
treinta y cinco	0.88	0.94	0.91	4711
veinte	0.88	0.90	0.89	7222
veinticinco	0.85	0.80	0.83	6315
accuracy			0.90	33992
macro avg	0.92	0.93	0.92	33992
weighted avg	0.90	0.90	0.90	33992

Figura 4.19: Resultados de ERT en clasificación conjunto de datos 2017-2022. *Fuente Propia*

Al basarse en los resultados presentados, se puede realizar el siguiente análisis:

- **RF:** Este algoritmo mostró un exactitud general de 0.89. A nivel macro, las métricas precisión, exhaustividad y puntaje f1 tuvieron un rendimiento similar, todas alrededor del 0.92. Es notable que el algoritmo tiene una precisión excepcional para categorías como `cincuenta`.
- **ERT:** Este modelo demostró ser ligeramente mejor que RF en términos de exactitud con un valor de 0.90. Además, tanto la precisión como el exhaustividad macro estuvieron alrededor del 0.92, indicando un rendimiento consistente a través de diferentes categorías.
- **KNN:** Este modelo tuvo una menor exactitud, siendo 0.78. Aunque el modelo tuvo buenos resultados para algunas categorías, como `cincuenta` y `quince`, en general, su rendimiento fue inferior a los modelos basados en árboles.
- **DT:** Con un exactitud de 0.81, el árbol de decisión se desempeñó mejor que KNN pero no tan bien como RF o ET. Las métricas a nivel macro, alrededor del 0.84, sugieren que el modelo tiene un rendimiento uniforme en varias categorías.
- **GB:** Este algoritmo mostró una exactitud de 0.66, lo que indica que su rendimiento es considerablemente inferior al de los modelos basados en árboles y KNN. Las métricas macro sugieren que hay un cierto desequilibrio en el rendimiento del modelo a través de las categorías, especialmente en categorías como `treinta` y `treinta y cinco`.

En resumen, los modelos **ERT** y **RF** se destacaron por su alto rendimiento general, con el modelo ERT exhibiendo una ligera superioridad en términos de exactitud. En contraste, el modelo de **GB** registró el rendimiento más bajo entre todos los algoritmos evaluados.

- **Conjunto de Datos Editado del 2017-2022.**
  - **GB**

	precision	recall	f1-score	support
cincuenta	0.68	0.72	0.70	1302
cuarenta	0.51	0.45	0.48	2486
cuarenta y cinco	0.60	0.64	0.62	1915
quince	0.70	0.72	0.71	3093
treinta	0.50	0.36	0.42	4035
treinta y cinco	0.45	0.48	0.46	3552
veinte	0.57	0.70	0.62	4778
veinticinco	0.50	0.48	0.49	4419
accuracy			0.55	25580
macro avg	0.56	0.57	0.56	25580
weighted avg	0.55	0.55	0.55	25580

Figura 4.20: Resultados de GB en clasificación conjunto de datos editado del 2017-2022. *Fuente Propia*

- **DT**

	precision	recall	f1-score	support
cincuenta	0.88	0.88	0.88	1302
cuarenta	0.77	0.78	0.77	2486
cuarenta y cinco	0.83	0.88	0.86	1915
quince	0.82	0.81	0.81	3093
treinta	0.61	0.59	0.60	4035
treinta y cinco	0.65	0.65	0.65	3552
veinte	0.67	0.68	0.67	4778
veinticinco	0.59	0.58	0.58	4419
accuracy			0.70	25580
macro avg	0.73	0.73	0.73	25580
weighted avg	0.69	0.70	0.69	25580

Figura 4.21: Resultados de DT en clasificación conjunto de datos editado del 2017-2022. *Fuente Propia*

- **KNN**

	precision	recall	f1-score	support
cincuenta	0.74	0.87	0.80	1302
cuarenta	0.68	0.78	0.73	2486
cuarenta y cinco	0.71	0.81	0.76	1915
quince	0.71	0.86	0.78	3093
treinta	0.57	0.58	0.58	4035
treinta y cinco	0.64	0.64	0.64	3552
veinte	0.65	0.60	0.63	4778
veinticinco	0.61	0.44	0.51	4419
accuracy			0.66	25580
macro avg	0.67	0.70	0.68	25580
weighted avg	0.65	0.66	0.65	25580

Figura 4.22: Resultados de KNN en clasificación conjunto de datos editado del 2017-2022. *Fuente Propia*

- **RF**

	precision	recall	f1-score	support
cincuenta	0.95	0.98	0.97	1302
cuarenta	0.89	0.91	0.90	2486
cuarenta y cinco	0.92	0.96	0.94	1915
quince	0.90	0.93	0.92	3093
treinta	0.78	0.71	0.75	4035
treinta y cinco	0.80	0.83	0.81	3552
veinte	0.77	0.82	0.79	4778
veinticinco	0.75	0.69	0.72	4419
accuracy			0.82	25580
macro avg	0.85	0.85	0.85	25580
weighted avg	0.82	0.82	0.82	25580

Figura 4.23: Resultados de RF en clasificación conjunto de datos editado del 2017-2022. *Fuente Propia*

- **ERT**

	precision	recall	f1-score	support
cincuenta	0.96	0.98	0.97	1302
cuarenta	0.91	0.92	0.91	2486
cuarenta y cinco	0.93	0.97	0.95	1915
quince	0.90	0.93	0.92	3093
treinta	0.82	0.73	0.77	4035
treinta y cinco	0.83	0.85	0.84	3552
veinte	0.77	0.83	0.80	4778
veinticinco	0.76	0.71	0.73	4419
accuracy			0.83	25580
macro avg	0.86	0.86	0.86	25580
weighted avg	0.83	0.83	0.83	25580

Figura 4.24: Resultados de ERT en clasificación conjunto de datos editado del 2017-2022. *Fuente Propia*

Basándose en las Figuras 4.20, 4.21, 4.22, 4.23 y 4.24, se puede deducir lo siguiente:

- **RF:** Este modelo presentó una exactitud de 0.90. Las métricas ponderadas (*weighted avg*) para precisión, exhaustividad y puntaje f1 también fueron 0.90, lo que indica un rendimiento homogéneo y equilibrado en la clasificación de las distintas categorías.
- **ERT:** Exhibió una exactitud de 0.91, mostrando un ligero aumento respecto a RF. Las métricas ponderadas para este modelo fueron 0.90, 0.91 y 0.91 para precisión, exhaustividad y puntaje f1, respectivamente, lo que sugiere que este modelo tuvo un mejor equilibrio en la clasificación de algunas categorías específicas.
- **KNN:** Con una exactitud de 0.79, este modelo tuvo un rendimiento inferior en comparación con RF y ERT. Las métricas ponderadas fueron consistentes con la exactitud, situándose todas alrededor de 0.79.

- **DT:** Este algoritmo alcanzó una exactitud de 0.83. Las métricas ponderadas también se encontraban en la misma línea, oscilando alrededor de 0.83.
- **GB:** Fue el modelo con el rendimiento más bajo, con una exactitud de 0.68. Las métricas ponderadas reflejaron este rendimiento, siendo 0.66, 0.68 y 0.67 para precisión, exhaustividad y puntaje f1, respectivamente.

Entre los algoritmos evaluados, **ERT** y **RF** fueron los que demostraron tener el mejor rendimiento, con exactitud de 0.91 y 0.90, respectivamente. Además, al igual que en los modelos de los anteriores conjuntos de datos **GB** fue el que presentó el rendimiento más bajo.

A partir de los resultados obtenidos, es evidente que los modelos **RF** y **ERT** sobresalieron, mostrando un rendimiento superior en todos los conjuntos de datos. Sin embargo, se observó una tendencia al sobreajuste en sus predicciones. El sobreajuste, según [104], se describe como una situación indeseable en ML. Ocurre cuando un modelo logra ajustarse perfectamente a los datos de entrenamiento pero muestra un rendimiento deficiente al enfrentarse a datos desconocidos o no vistos. Al evaluar los resultados del conjunto de pruebas y de los datos de entrenamiento, presentados en el repositorio proporcionado[2], se detectó un patrón revelador: todas las métricas de evaluación para los datos de entrenamiento alcanzaron un valor perfecto de 1. Este comportamiento es un indicativo claro del sobreajuste. Dada esta situación, se decidió llevar a cabo una evaluación más exhaustiva de los algoritmos **RF** y **ERT**, ajustando de manera específica sus hiperparámetros según cada conjunto de datos, con el propósito de reducir este fenómeno en los modelos.

En la librería `sklearn`, los algoritmos basados en árboles tienen hiperparámetros que determinan la complejidad del modelo. La complejidad se relaciona directamente con el riesgo de sobreajuste: modelos muy complejos se ajustan demasiado a los datos de entrenamiento y, por ende, tienen problemas de generalización.

Algunos hiperparámetros clave y cómo influencian el sobreajuste son:

- **n_estimators:** Aunque un mayor número de árboles puede mejorar el rendimiento, después de cierto punto, no ofrece ganancias significativas y aumenta el costo computacional.
- **max_depth:** Un árbol más profundo puede capturar más detalles, pero también corre el riesgo de aprender ruido y detalles irrelevantes de los datos, llevando a sobreajuste.
- **min_samples_split:** Un valor alto puede prevenir divisiones innecesarias, protegiendo contra el sobreajuste al requerir más datos para justificar la subdivisión de un nodo.

Para clarificar, la siguiente tabla resume los hiperparámetros ajustados y los valores seleccionados para combatir el sobreajuste:

[2]Consultar el repositorio para más detalles: `https://github.com/jvanessafdez/coagulant-dose/blob/main/src/14_Class_Modeling.ipynb`

Hiperparámetro	RF	ERT
n_estimators	200	150
max_depth	17	25
min_samples_split (solo ERT)		5

Tabla 4.1: Hiperparámetros ajustados para **RF** y **ERT** en clasificación. *Fuente Propia.*

A través de este ajuste y la validación cruzada, se buscó garantizar que los modelos no solo tuvieran un buen rendimiento en el entrenamiento, sino también una buena capacidad de generalización, reduciendo así el sobreajuste. Los resultados post-ajuste se presentan en la Tabla 4.2.

Conjunto de datos	Algoritmos	Precisión	Exhaustividad	F1-score	Accuracy
Sin edición 2013-2022	RF	0.71349	0.71438	0.70927	0.71438
	ERT	0.71066	0.71079	0.70691	0.71079
Con edición 2013-2022	RF	0.59019	0.60328	0.59080	0.60328
	ERT	0.61458	0.61450	0.60605	0.61450
Sin edición 2017-2022	RF	0.80177	0.80155	0.79698	0.80155
	ERT	0.79254	0.79472	0.78946	0.79472
Con edición 2017-2022	RF	0.80531	0.815797	0.80695	0.815797
	ERT	0.79641	0.80903	0.797698	0.80903

Tabla 4.2: Resultados de Clasificación. *Fuente Propia.*

De estos resultados se puede analizar lo siguiente:

- **Conjunto de datos sin edición 2013-2022:** En este conjunto, **RF** muestra un rendimiento ligeramente mejor en todas las métricas en comparación con **ERT**.
- **Conjunto de datos con edición 2013-2022:** En esta ocasión, **ERT** supera a **RF** en cada métrica evaluada. Sin embargo, es importante destacar que la intervención de edición disminuyó la eficiencia de ambos algoritmos en comparación con el conjunto sin edición.
- **Conjunto de datos sin edición 2017-2022: RF** mantiene su superioridad, pero con **ERT** no quedando muy atrás. Interesantemente, ambos algoritmos presentan un rendimiento más destacado en este conjunto que en el conjunto sin edición 2013-2022. Esto sugiere que los datos entre 2017 y 2022 podrían ser intrínsecamente más predictivos o presentar menor ruido.
- **Conjunto de datos con edición 2017-2022:** Aunque **RF** sigue prevaleciendo en términos de rendimiento, es en la métrica F1-score donde se evidencia una ventaja sutil sobre **ERT**. Al igual que con el conjunto de 2013-2022, la edición parece haber impactado en el rendimiento, pero esta vez mostrando una mejora en las métricas en comparación al conjunto sin edición.

En los datos del período 2013-2022, se observó que la edición tendía a disminuir el rendimiento de los modelos. Sin embargo, esta tendencia cambió notablemente con el conjunto de datos del período 2017-2022. Para este último, las ediciones no solo

mejoraron el rendimiento general, sino que fueron particularmente beneficiosas para el algoritmo **RF**. Esto indica que las modificaciones aplicadas agregaron claridad y valor a los datos del período 2017-2022.

Al analizar más a fondo, se destaca que, independientemente del conjunto de datos y su tratamiento, **RF** tiende a superar o, al menos, a igualar el rendimiento de **ERT**. Esta superioridad de **RF** alcanza su punto máximo en el conjunto "Con edición 2017-2022", donde supera todas las demás configuraciones evaluadas.

Dado lo anterior, y teniendo en cuenta todas las métricas de rendimiento, es evidente que el algoritmo **RF** entrenado con el conjunto "Con edición 2017-2022" sobresale entre las opciones, estableciéndose como el modelo más recomendado de todos los examinados.

4.2 Modelado: Enfoque de Regresión

Como se mencionó en el capítulo previo, se utilizaron varios algoritmos de ML con el propósito de construir modelos. Para ello, se empleó la biblioteca `sklearn`. Al igual que en el enfoque de clasificación, se experimentó con los algoritmos: GB, DT, KNN, RF y ERT en sus versiones de regresión.

En el proceso de evaluación y comparación de los modelos propuestos, se recurrió a un conjunto de métricas detalladas previamente en el capítulo 3. Específicamente, se consideraron las métricas MAE, MSE y RMSE. El propósito primordial de esta evaluación es determinar cuál de los modelos ostenta un rendimiento óptimo basándose en los valores de estas métricas. Es esencial recalcar que el modelo ideal debería manifestar valores de MAE, MSE y RMSE muy pequeños, y que estos no deben sobrepasar el 5 % de la media de la variable objetivo.

4.2.1 Resultados de los modelos

- **Conjunto de datos 2013-2022.**

Algoritmo	MAE	MSE	RMSE
RF	1.60557	7.29569	2.70105
ERT	1.49324	6.44024	2.53776
KNN	2.38235	12.39193	3.52022
DT	2.02447	14.96448	3.86839
GB	2.33212	11.58519	3.40370

Tabla 4.3: Resultados de implementación de algoritmos: Enfoque de regresión. conjunto de datos 2013-2022. *Fuente propia.*

De la Tabla 4.3 se puede observar que los algoritmos con mejor rendimiento para este conjunto de datos son **ERT** y **RF**. El primero exhibe el MAE, MSE y RMSE

más bajos, seguido muy de cerca por **RF**.

- **Conjunto de datos editada del 2013-2022.**

Algoritmo	MAE	MSE	RMSE
RF	2.36921	11.62499	3.40954
ERT	2.15342	10.24379	3.20059
KNN	3.18833	19.22234	4.38433
DT	2.90692	24.58249	4.95807
GB	3.20367	17.89994	4.23083

Tabla 4.4: Resultados de implementación de algoritmos: Enfoque de regresión. Conjunto de datos editada del 2013-2022 *Fuente propia.*

Al analizar la Tabla 4.4, se tiene que nuevamente **ERT** y **RF** se destacan como los algoritmos más eficientes, con **ERT** teniendo ligeramente mejores métricas en todas las categorías.

- **Conjunto de datos 2017-2022.**

Algoritmo	MAE	MSE	RMSE
RF	1.43011	5.76952	2.40198
ERT	1.33211	5.12873	2.26467
KNN	2.12789	10.18770	3.19182
DT	1.74935	11.86764	3.44494
GB	2.21396	10.04893	3.17000

Tabla 4.5: Resultados de implementación de algoritmos: Enfoque de regresión. conjunto de datos editado del 2017-2022. *Fuente propia.*

Según la Tabla 4.5 para este conjunto, **ERT** muestra el mejor rendimiento en términos de MAE, MSE y RMSE. Le sigue el algoritmo **RF** con métricas cercanas pero ligeramente superiores.

- **Conjunto de datos editado del 2017-2022.**

Algoritmo	MAE	MSE	RMSE
RF	2.11395	9.67078	3.10979
ERT	1.68697	8.27183	2.87608
KNN	2.79760	15.69079	3.94851
DT	2.40935	19.00714	4.35972
GB	2.95317	15.13011	3.83974

Tabla 4.6: Resultados de implementación de algoritmos: Enfoque de regresión. conjunto de datos editado del 2017-2022. *Fuente propia.*

Se puede analizar de la Tabla 4.6 que en este último conjunto de datos editado, **ERT** y **RF** vuelven a ser los algoritmos con el rendimiento más destacado, con **ERT** superando ligeramente a **RF** en todas las métricas.

Tras analizar los resultados de varios algoritmos en los diferentes conjuntos de datos presentados, se identifica que, con base en las métricas MAE, MSE y RMSE, los algoritmos **ERT** y **RF** destacan consistentemente por su superioridad. Su rendimiento sobresaliente en comparación con otros algoritmos sugiere que **ERT** y **RF** son las opciones más prometedoras para investigaciones adicionales.

Dado este hallazgo, se inició un riguroso proceso de optimización de hiperparámetros para ERT y RF en todos los conjuntos de datos. Esta acción se vio impulsada por la detección de signos de sobreajuste en ambos algoritmos. A través de la optimización de hiperparámetros, se busca ajustar meticulosamente el comportamiento de estos modelos, reducir el sobreajuste y, por consiguiente, determinar cuál de ellos presenta menos errores al evaluar las métricas mencionadas.

Es importante subrayar que, en la biblioteca sklearn, los hiperparámetros utilizados son consistentes tanto para tareas de clasificación como de regresión. Esta uniformidad en la configuración de hiperparámetros permite una transición fluida y coherente entre ambos tipos de tareas. En este contexto, al igual que en la fase de clasificación, la optimización de hiperparámetros se llevó a cabo principalmente con el propósito de atenuar el sobreajuste.

Los hiperparámetros específicos seleccionados para cada algoritmo, tras la optimización, son:

Hiperparámetro	**RF**	**ERT**
n_estimators	150	150
max_depth	15	25
min_samples_split	15	10

Tabla 4.7: Hiperparámetros ajustados para **RF** y **ERT** en regresión. *Fuente Propia.*

En la Tabla 4.8 se muestran los resultados derivados de la optimización. Para garantizar una evaluación más robusta de los modelos, estos resultados se obtuvieron empleando validación cruzada, siguiendo el mismo enfoque que se utilizó anteriormente en el modelado de clasificación.

Conjunto de Datos	**Algoritmo**	**Métrica**	**% que excede de la media**	**Valor**
Sin edición 2013-2022	RF	RMSE	14.24381	3.53684
		MSE	50.37812	12.50926
		MAE	9.07378	2.25308
	ERT	RMSE	14.08899	3.49840
		MSE	49.28895	12.23881
		MAE	9.04693	2.24642
Con edición 2013-2022	RF	RMSE	18.34942	4.65099
		MSE	85.34301	21.63173

		MAE	13.78873	3.49500
	ERT	RMSE	18.37800	4.65824
		MSE	85.60913	21.69919
		MAE	13.86327	3.51389
Sin edición 2017-2022	RF	RMSE	13.88625	3.44055
		MSE	47.77635	11.83768
		MAE	9.29768	2.30365
	ERT	RMSE	14.07702	3.48782
		MSE	49.09807	12.16486
		MAE	9.64315	2.38925
Con edición 2017-2022	RF	RMSE	17.18983	4.36576
		MSE	75.04673	19.05989
		MAE	13.15450	3.34089
	ERT	RMSE	17.88024	4.41412
		MSE	76.71854	19.48449
		MAE	13.37649	3.39728

Tabla 4.8: Resultados de implementación de hiperparámetros: enfoque de regresión. *Fuente propia.*

Al revisar los resultados presentados en la Tabla 4.8 se puede analizar lo siguiente:

- **Sin edición 2013-2022:** RF parece tener un rendimiento ligeramente inferior a ERT en términos de RMSE y MSE, pero mejor en MAE. Sin embargo, la diferencia en los valores de las métricas es marginal, lo que indica que ambos algoritmos son comparables en rendimiento para este conjunto de datos.
- **Con edición 2013-2022:** De nuevo, RF y ERT muestran resultados muy similares en todas las métricas. Sin embargo, los errores aquí son significativamente mayores que en el conjunto "Sin edición 2013-2022", lo que sugiere que la edición en los datos pudo haber introducido cierta variabilidad o ruido.
- **Sin edición 2017-2022:** En este conjunto, ambos algoritmos presentan errores mucho menores en comparación con los conjuntos del 2013-2022. Esto podría ser debido a la menor variabilidad en un periodo más corto o a la calidad del conjunto de datos. RF supera a ERT en todas las métricas.
- **Con edición 2017-2022:** Al igual que en el conjunto de 2013-2022, la versión editada de 2017-2022 muestra errores más altos en comparación con su versión sin edición. RF y ERT tienen un rendimiento muy similar, con RF teniendo un ligero margen en todas las métricas.

En general:

- Los algoritmos RF y ERT tienen un rendimiento muy comparable en todos los conjuntos de datos, con diferencias marginales en sus métricas.
- Independientemente del algoritmo utilizado o del conjunto de datos considerado, todos los modelos excedieron el umbral del 5 % sobre la media en sus métricas

de error. Esto indica que todos los modelos evaluados no satisfacen el criterio de éxito establecido en la sección 3.2

- Los conjuntos de datos sin edición tienden a producir errores más bajos en comparación con los conjuntos editados en ambos periodos, lo que sugiere que las ediciones pueden haber introducido ciertas complicaciones o variabilidades.
- El conjunto "Sin edición 2017-2022" con el algoritmo RF parece ser el más prometedor, dado que presenta los errores más bajos en términos de RMSE y MAE, que son métricas cruciales para la evaluación del rendimiento de modelos de regresión. Además, el porcentaje que excede de la media en todas las métricas para este conjunto y algoritmo es el menor entre todos, lo que indica una mayor precisión y menos variabilidad en comparación con otros conjuntos y algoritmos.

Por lo tanto, basado en este análisis, el mejor modelo es el generado por el algoritmo **RF** aplicado al "conjunto de datos 2017-2022".

El modelo seleccionado, junto con el modelo de clasificación identificado como el más óptimo, se proporcionaron a la planta de tratamiento de agua potable "El Tablazo". Estos modelos están destinados para pruebas y fueron implementados en un panel de control que será entregado a la planta de tratamiento el tablazo, la información sobre este tablero de control se podrá visualizar en el Anexo C.

4.3 Resumen

En este capítulo se abordó la creación de dos modelos distintos: uno para clasificación y otro para regresión a partir de la evaluación de varios modelos creados. Inicialmente, en el ámbito de clasificación, se realizó una evaluación preliminar utilizando regresión logística y RF con el fin de determinar la técnica más adecuada para el balanceo de datos. Dicha elección se basó en métricas claves como precisión, exhaustividad, puntaje f1 y exactitud. El objetivo era seleccionar una técnica que ofreciera un equilibrio en estas métricas para todas las clases y que, además, presentara los valores más elevados. A raíz de este análisis, se concluyó que SMOTETomek era la técnica óptima para los primeros tres conjuntos de datos, mientras que para el cuarto, SMOTE demostró ser la mejor opción.

Después de equilibrar los datos, se examinaron varios algoritmos: RF, ERT, GB, KNN y DT. RF y ERT destacaron, aunque evidenciaron sobreajuste, lo que llevó a la optimización de hiperparámetros. Finalmente se definió que el modelo RF para el conjunto de datos del período 2017-2022 era el más adecuado.

En cuanto al modelado de regresión, se realizó inicialmente un intento de mejorar los resultados mediante la transformación Box-Cox de la columna de dosis de coagulante. Sin embargo, esta acción no aportó mejoras significativas en las métricas de evaluación que para este caso fueron MAE, MSE y RMSE, lo que llevó a continuar con el modelado sin dicha transformación. Siguiendo un enfoque similar al análisis de clasificación, se evaluaron los mismos algoritmos, pero adaptados para regresión. Nuevamente, ERT y RF sobresalieron, pero también presentaron sobreajuste. Tras una nueva ronda de

optimización de hiperparámetros, se determinó que el modelo RF para el conjunto de datos 2017-2022 era el más eficaz en este contexto de regresión.

Capítulo 5.

Conclusiones y trabajos futuros.

En este capítulo, inicialmente se presentan las conclusiones derivadas del trabajo de grado en cuestión, las cuales se extrajeron a lo largo de su realización. A continuación, se ofrecen las recomendaciones pertinentes. Finalmente, se sugieren posibles investigaciones futuras que podrían surgir con el objetivo de mejorar este estudio o adaptarlo a otras áreas.

5.1 Conclusiones

- La inclusión de variables meteorológicas, a menudo pasadas por alto en investigaciones previas, ha demostrado ser esencial para una comprensión más profunda y una modelización precisa del proceso de coagulación. Estas variables no solo influyen en las características del agua, sino que su integración en modelos predictivos potencia la precisión en la dosificación de coagulantes. El presente estudio ha diseñado modelos avanzados que fusionan variables tanto hidrológicas como meteorológicas, ofreciendo así un sistema de recomendación innovador para la planta de tratamiento de agua .[El] Tablazo". Esta propuesta representa un avance significativo, pues no solo optimiza el tiempo de producción al minimizar la necesidad de pruebas de jarra, sino que también reduce la dependencia del criterio del operador para determinar la dosificación del coagulante. Esta integración pone de manifiesto el valor añadido de considerar datos meteorológicos en la toma de decisiones relacionadas con el tratamiento del agua.

- Tras un meticuloso análisis de tendencias temporales, se identificaron variaciones significativas en la mayoría de las variables, especialmente entre los meses de junio a septiembre. Estas fluctuaciones coinciden con el período en que Popayán experimenta condiciones climáticas más cálidas y ventosas. Esta observación respalda firmemente la hipótesis inicial del estudio, confirmando la influencia directa de las variables meteorológicas en la calidad del agua. Dicho descubrimiento no solo destaca la interdependencia entre el clima y los recursos hídricos, sino que también subraya la importancia de considerar factores ambientales en futuros estudios y estrategias de gestión del agua. Esta conclusión refuerza la necesidad de

una planificación y monitorización continua para garantizar la calidad del agua en función de las variaciones climáticas.

- Al examinar el conjunto de datos de la planta de tratamiento "El Tablazo", se identificó que la dosificación del coagulante se efectuaba mayormente en intervalos de cinco unidades. Esta observación condujo a la decisión de realizar dos análisis en paralelo: uno de clasificación y otro de regresión. Esta estrategia permitió contrastar de manera efectiva las diferencias entre ambas metodologías y las soluciones específicas que cada una aporta. En consecuencia, se logró una comprensión más profunda y completa de los datos, enriqueciendo el entendimiento global del proceso.

- El análisis de los resultados obtenidos al aplicar diferentes técnicas de balanceo de clases en varios conjuntos de datos para el enfoque de clasificación reveló la importancia y la necesidad de abordar adecuadamente el desequilibrio de clases. Las clases con menor representación, como 'cincuenta', 'cuarenta', 'cuarenta y cinco' y 'quince', a menudo se veían afectadas negativamente en su rendimiento, con bajos valores en métricas cruciales como "precision", "recall" y "f1-score". Sin embargo, al aplicar técnicas de balanceo, se pudo mejorar significativamente el rendimiento en estas clases, como era lo esperado en los criterios de exito de ML del trabajo de grado.

- Al analizar la eficacia de la transformada de Box-Cox en el contexto de ML, se observó que, para algoritmos no parametrizados como RF y ET, la transformación no influye significativamente en el rendimiento del modelo. Estos algoritmos, por su naturaleza, no dependen estrictamente de supuestos distribucionales, por lo que la normalización de datos a través de métodos como Box-Cox puede resultar redundante o incluso innecesaria cuando se busca optimizar el rendimiento.

- Tras analizar los resultados obtenidos de los distintos modelos de regresión evaluados, fue evidente que, sin importar las variaciones en el algoritmo empleado o las características específicas del conjunto de datos en cuestión, todos los modelos superaron el margen del 5 % por encima de la media en sus respectivas métricas de error. Esto lleva a la conclusión que los modelos no brindan gran confiabilidad y se ve la necesidad de evaluarlos con la integración de más datos.

- Al evaluar los distintos conjuntos de datos para tareas de clasificación y regresión, se evidenciaron patrones de comportamiento distintivos. En la clasificación, el conjunto que abarcaba el periodo 2013-2022 con edición introdujo ciertas perturbaciones en el modelo. Sin embargo, al restringir el alcance a 2017-2022, el conjunto editado demostró ser superior en términos de rendimiento. En contraste, para la regresión, las ediciones en los conjuntos mermaron el rendimiento en ambos periodos. No obstante, una constante en ambos análisis fue que los datos del periodo 2017-2022 superaron en rendimiento a sus contrapartes del periodo 2013-2022, independientemente del enfoque utilizado.

- Luego de un análisis minucioso, se observó que, a pesar de que el algoritmo ERT exhibió inicialmente un rendimiento que superaba al de RF, manifestó una tendencia notable hacia el sobreajuste. Al introducir hiperparámetros específicos para contrarrestar este comportamiento, se descubrió que aquellos que eran más

efectivos en la reducción del sobreajuste impulsaron a RF a superar a ERT en términos de métricas de rendimiento. Adicionalmente, el RF demostró una resistencia al sobreajuste más efectiva en comparación con ERT. Por ende, fue con RF donde se alcanzaron modelos con un equilibrio más adecuado entre precisión y generalización.

5.2 Trabajos Futuros

El objetivo central de esta investigación fue crear un sistema de recomendación de dosis de coagulante para la potabilización de agua en la planta de tratamiento "El Tablazo". Para ello, se diseñaron y evaluaron modelos de ML que, basándose en variables hidrológicas y meteorológicas actuales, sugieran la dosis adecuada a aplicar.

Dado el propósito anterior, y con el objetivo de mejorar y/o complementar los modelos de predicción de dosis de coagulante, se sugieren los siguientes trabajos futuros:

- Intensificar la precisión en la predicción de dosis de coagulante. Una estrategia prometedora es la optimización avanzada de los modelos a través de la integración de diversos algoritmos. En particular, la fusión del algoritmo de RF con un algoritmo genético podría marcar una diferencia sustancial en la precisión y eficacia del modelo. Esta combinación podría desencadenar un nuevo paradigma en las técnicas de predicción de dosificación.
- Para potenciar significativamente la eficacia del modelo, es crucial ampliar la base de datos con la que fue entrenado. Actualmente, el modelo se centra solo en datos de la planta "El Tablazo". Sin embargo, al incorporar información de otras plantas, como la planta de tratamiento de "Palacé" y la planta de tratamiento de "Tulcán", el modelo podría adaptarse a diversas fuentes de agua y ofrecer predicciones mucho más versátiles. Este enriquecimiento permitiría proporcionar recomendaciones más generales para el suministro de agua en toda la ciudad de Popayán, independientemente de la planta de tratamiento que lo provea. La inclusión de estas fuentes adicionales transformaría el modelo en una herramienta indispensable para toda la gestión hídrica de la ciudad.
- Considerando que la planta de tratamiento de agua cuenta con sensores para medir variables como las que sirven de entrada al sistema de recomendación de dosis de coagulante, un paso lógico y altamente beneficioso sería la implementación de un Sistema de Control y Adquisición de Datos (SCADA, *Supervisory Control And Data Acquisition*). Este sistema permitiría la automatización del proceso de recopilación y envío de datos en tiempo real al modelo de ML. Con esta integración, se eliminan las posibles ineficiencias y errores humanos asociados con la entrada manual de datos, lo que garantiza recomendaciones más precisas y oportunas. Además, un sistema SCADA proporcionaría una plataforma centralizada para monitorear, controlar y optimizar toda la operación de la planta, asegurando una gestión más eficiente y sostenible del tratamiento de agua.

Bibliografía

[1] L. Visengeriyeva, A. Kammer, I. Bär, A. Kniesz, and M. Plöd, "Crisp-ml(q). the ml lifecycle process."

[2] N. Unidas, "Decenio internacional para la acción 'el agua, fuente de vida' 2005-2015." [En línea], 2010. Disponible: https://www.un.org/spanish/waterforlifedecade/human_right_to_water.shtml.

[3] "Tratamiento de agua: Potabilización." https://www.acciona.com/es/tratamiento-de-agua/potabilizacion/, 2023. Accedido el: [tu fecha de acceso aquí].

[4] R. M. El-taweel, N. Mohamed, K. A. Alrefaey, S. Husien, A. Abdel-Aziz, A. I. Salim, N. G. Mostafa, L. A. Said, I. S. Fahim, and A. G. Radwan, "A review of coagulation explaining its definition, mechanism, coagulant types, and optimization models; rsm, and ann," *Current Research in Green and Sustainable Chemistry*, vol. 6, p. 100358, 2023.

[5] A. Mirsepassi, B. Carthers, and H. Dharmappa, "Application of artificial neural networks to the real time operation of water treatment plants," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 1, pp. 516–521, 1995.

[6] N. Valentin, T. Denoux, and F. Fotoohi, "An hybrid neural network based system for optimization of coagulant dosing in a water treatment plant," in *International Joint Conference on Neural Networks*, vol. 5, pp. 3380–3385, 1999.

[7] J. Fandiño-Piamonte and C. Camargo-Arcila, *Evaluación y optimización de la planta de tratamiento de agua potable del municipio de Purificación en el departamento de Tolima*. PhD thesis, Universidad Católica de Colombia, 2013.

[8] N. C. Chulluncuy-Camacho, "Tratamiento de agua para consumo humano," *Ing. ind. (Lima)*, no. 029, pp. 153–170, 2011.

[9] J. A. Andrade Quintero, "Automatización del proceso de potabilización de agua planta de tratamiento roldanillo valle." https://red.uao.edu.co/bitstream/10614/7573/1/T05575.pdf, 2006.

[10] M. R. G.-R. Santos, "Ajuste matemático del comportamiento de la turbiedad residual en los procesos de floculación-coagulación del agua realizados en la planta la flora del a.m.b s.a e.s.p empleando policloruro de aluminio líquido," tesis ingeniera química, Universidad Industrial de Santander, Bucaramanga, Santander, Colombia, 2011.

[11] P. de la República de Colombia, "Decreto 1575 de 2007." [En línea], May 2007. Disponible: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=30007.

[12] D. Rout, R. Verma, and K. Agarwal, "Polyelectrolyte treatment—an approach for water quality treatment," *Water Science and Technology*, vol. 40, no. 2, pp. 137–141, 1999.

[13] A. F. Santander Leyva and J. K. Ruiz Duarte, "Análisis del irca y su relación con variables meteorológicas y su ubicación geográfica para el departamento de magdalena en los años 2012 y 2013," 2017. Trabajo de grado - Pregrado, Universidad de La Salle, Bogotá, Facultad de Ingeniería, Ingeniería Ambiental y Sanitaria.

[14] A. desconocido, "Planta de tratamiento de agua potable el tablazo," 2014. Accedido el 15 de julio de 2023.

[15] R. Briceño, L. Fuentes, I. Mendoza, J. Bolaños, and Y. Caldera, "Efectividad de una suspensión gelatinosa de huesos bovinos en la clarificación de aguas con alta turbidez," *REDIELUZ*, vol. 4, no. 2, pp. 46–53, 2014.

[16] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering," in *Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering*, vol. 17, Jun 2008.

[17] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," *Engineering*, vol. 2, p. 1051, 2007.

[18] H. Yamamura, E. U. Putri, T. Kawakami, A. Suzuki, H. D. Ariesyady, and T. Ishii, "Dosage optimization of polyaluminum chloride by the application of convolutional neural network to the floc images captured in jar tests," *Separation and Purification Technology*, vol. 237, p. 116467, 2020.

[19] R. Zaque, W. Silva, and A. Santos, "Expert system for applying coagulant in water treatment: case study in nobres (brazil)," *Water Practice & Technology*, vol. 13, pp. 832–840, 12 2018.

[20] H.-J. Mälzer and S. Strugholtz, "Artificial neural networks for cost optimization of coagulation, sedimentation and filtration in drinking water treatment," *Water Supply*, vol. 8, no. 4, pp. 383–388, 2008.

[21] T. Trinh and L. Kang, "Application of response surface method as an experimental design to optimize coagulation tests," *Environmental Engineering Research*, vol. 15, 06 2010.

[22] D. Wang, J. Wu, L. Deng, *et al.*, "A real-time optimization control method for coagulation process during drinking water treatment," *Nonlinear Dyn*, vol. 105, pp. 3271–3283, 2021.

[23] Y. Wang, B. Han, and Y. Zhang, "A kind of coagulant dosing control model based on isfla-svm," in *The 27th Chinese Control and Decision Conference (2015 CCDC)*, pp. 6417–6420, 2015.

[24] S. Ghafari, H. A. Aziz, M. H. Isa, and A. A. Zinatizadeh, "Application of response surface methodology (rsm) to optimize coagulation–flocculation treatment of leachate using poly-aluminum chloride (pac) and alum," *Journal of Hazardous Materials*, vol. 163, no. 2, pp. 650–656, 2009.

[25] T. K. Trinh and L. S. Kang, "Response surface methodological approach to optimize the coagulation–flocculation process in drinking water treatment," *Chemical Engineering Research and Design*, vol. 89, no. 7, pp. 1126–1135, 2011.

[26] C. Kim and M. Parnichkun, "Prediction of settled water turbidity and optimal coagulant dosage in drinking water treatment plant using a hybrid model of k-means clustering and adaptive neuro-fuzzy inference system," *Applied Water Science*, vol. 7, pp. 3885–3902, 2017.

[27] A. Bressane, A. Goulart, C. Melo, I. Gomes, A. Loureiro, R. Negri, R. Moruzzi, A. Reis, J. Formiga, G. da Silva, and R. Thomé, "A non-hybrid data-driven fuzzy inference system for coagulant dosage in drinking water treatment plant: Machine-learning for accurate real-time prediction," *Water*, vol. 15, p. 1126, 2023.

[28] X. Xiao, S. Pang, F. Tong, S. Mao, Y. Liu, B. Tang, H. Jia, H. Wang, and Y. Du, "Intelligent dosing method for water treatment based on data mining combined with image recognition and self-learning," in *2022 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS)*, pp. 877–886, 2022.

[29] M. Achite, S. Samadianfard, N. Elshaboury, *et al.*, "Modeling and optimization of coagulant dosage in water treatment plants using hybridized random forest model with genetic algorithm optimization," *Environmental Development and Sustainability*, 2022.

[30] H. R. Maier, N. Morgan, and C. W. Chow, "Use of artificial neural networks for predicting optimal alum doses and treated water quality parameters," *Environmental Modelling & Software*, vol. 19, no. 5, pp. 485–494, 2004.

[31] C. Gagnon, B. P. Grandjean, and J. Thibault, "Modelling of coagulant dosage in a water treatment plant," *Artificial Intelligence in Engineering*, vol. 11, no. 4, pp. 401–404, 1997. Applications of Neural Networks in Process Engineering.

[32] G.-D. Wu and S.-L. Lo, "Effects of data normalization and inherent-factor on decision of optimal coagulant dosage in water treatment by artificial neural network," *Expert Systems with Applications*, vol. 37, no. 7, pp. 4974–4983, 2010.

[33] S. Haghiri, A. Daghighi, and S. Moharramzadeh, "Optimum coagulant forecasting by modeling jar test experiments using anns," *Drinking Water Engineering and Science*, vol. 11, pp. 1–8, 01 2018.

[34] B. Lamrini, A. Benhammou, M.-V. Le Lann, and A. Karama, "A neural software sensor for online prediction of coagulant dosage in a drinking water treatment plant," *Transactions of The Institute of Measurement and Control - TRANS INST MEASURE CONTROL*, vol. 27, pp. 195–213, 06 2005.

[35] K. Griffiths and R. Andrews, "The application of artificial neural networks for the optimization of coagulant dosage," *Water Science & Technology: Water Supply*, vol. 11, p. 605, 12 2011.

[36] D. Jayaweera and N. Aziz, "Development and comparison of extreme learning machine and multi-layer perceptron neural network models for predicting optimum coagulant dosage for water treatment," *Journal of Physics: Conference Series*, vol. 1123, p. 012032, 11 2018.

[37] A. Robenson, S. R. Abd Shukor, and N. Aziz, "Development of process inverse neural network model to determine the required alum dosage at segama water treatment plant sabah, malaysia," *Computer Aided Chemical Engineering*, vol. 27, 12 2009.

[38] Z. Song, Y. Zhao, X. Song, and C. Liu, "Research on prediction model of optimal coagulant dosage in water purifying plant based on nerual network," in *2009 ISECS International Colloquium on Computing, Communication, Control, and Management*, vol. 4, pp. 258–261, 2009.

[39] Z. Gormez and A. Sengul, "Prediction of optimal coagulant dosage in drinking water treatment by artificial neural network," in *1st International EWaS-MED International Conference*, 04 2013.

[40] S. Dharman, V. Chandramouli, and S. Lingireddy, "Predicting total organic carbon removal efficiency and coagulation dosage using artificial neural networks," *Environmental Engineering Science*, vol. 29, pp. 743–750, 08 2012.

[41] R. Olanrewaju, S. Muyibi, T. Salawudeen, and A. Aibinu, "An intelligent modeling of coagulant dosing system for water treatment plants based on artificial neural network," *Australian Journal of Basic and Applied Sciences*, vol. 6, pp. 93–99, 01 2012.

[42] W. Xiaojie, J. Yunzhe, and L. Xiaojing, "Research on the prediction of water treatment plant coagulant dosage based on feed-forward artificial neutral network," in *2011 International Conference on Consumer Electronics, Communications and Networks (CECNet)*, pp. 1615–1617, 2011.

[43] H. Hernandez and M.-V. Le Lann, "Development of a neural sensor for on-line prediction of coagulant dosage in a potable water treatment plant in the way of its diagnosis," in *Advances in Artificial Intelligence - IBERAMIA-SBIA 2006* (J. S. Sichman, H. Coelho, and S. O. Rezende, eds.), (Berlin, Heidelberg), pp. 249–257, Springer Berlin Heidelberg, 2006.

[44] S. Deveughele and Z. Do-Quang, "Neural networks: an efficient approach to predict on-line the optimal coagulant dose," *Water Science and Technology: Water Supply*, vol. 4, no. 5-6, pp. 87–94, 2004.

[45] L. Gomes, F. Souza, R. Pontes, T. Neto, and R. Araújo, "Coagulant dosage determination in a water treatment plant using dynamic neural network models," *International Journal of Computational Intelligence and Applications*, vol. 14, 09 2015.

[46] F. Menezes, R. Fontes, K. Oliveira-Esquerre, and R. Kalid, "Application of uncertainty analysis of artificial neural networksfor predicting coagulant and alkalizer dosages in a water treatment process," *Brazilian Journal of Chemical Engineering*, vol. 35, pp. 1369–1381, 12 2018.

[47] M. Wagh, "Application of cascade feed-forward neural network to predict coagulant dose," *Journal of Applied Water Engineering and Research*, vol. 9, 06 2021.

[48] X. Deng and L. Canguang, "Application of elm to predict the coagulant dosing in water treatment plant," *Water Science and Technology: Water Supply*, vol. 17, p. ws2016203, 12 2016.

[49] D. Wang, X. Chang, and K. Ma, "Predicting flocculant dosage in the drinking water treatment process using elman neural network," *Environmental Science and Pollution Research*, vol. 29, 01 2022.

[50] A. S. Kote and D. V. Wadkar, "Application of feed forward neural network for prediction of optimum coagulant dose in water treatment plant," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, Oct 2019.

[51] Y. Liu, Y. He, S. Li, Z. Dong, J. Zhang, and U. Kruger, "An auto-adjustable and time-consistent model for determining coagulant dosage based on operators' experience," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. PP, pp. 1–12, 12 2019.

[52] D. Jayaweera and N. Aziz, "An efficient neural network model for aiding the coagulation process of water treatment plants," *Environment, Development and Sustainability*, vol. 24, 01 2022.

[53] D.-S. Joo, D.-J. Choi, and H. Park, "Determination of optimal coagulant dosing rate using an artificial neural network," *Journal of Water Supply: Research and Technology - AQUA*, vol. 49, pp. 49–55, 02 2000.

[54] Z.-y. Song, X.-l. Song, Y.-b. Zhao, and C.-y. Liu, "Research on predictive control of coagulant dosage based on neural network," in *2011 Seventh International Conference on Natural Computation*, vol. 1, pp. 217–221, 2011.

[55] M. Baouab and S. Cherif, "Prediction of the optimal dose of coagulant for various potable water treatment processes through artificial neural network," *Journal of Hydroinformatics*, vol. 20, 07 2018.

[56] J. D. Rodríguez Vicente, "Diseño de una red neuronal artificial para la predicción de la dosis optima de policloruro de aluminio en el tratamiento de agua potable de la epmapa-sd," Master's thesis, Escuela Superior Politécnica de Chimborazo, 12 2021.

[57] H. Luo, X. Li, F. Yuan, C. Yuan, W. Huang, Q. Ji, X. Wang, B. Liu, and G. Zhu, "Application of a new architecture neural network in determination of flocculant dosing for better controlling drinking water quality," *Water*, vol. 14, no. 17, 2022.

[58] J. Zhang and D.-Y. Luo, "Multimodal control by variable-structure neural network modeling for coagulant dosing in water purification process," *Complexity*, vol. 2020, pp. 1–11, 08 2020.

[59] Aiswarya, M. Nithya Kurup, and N. Johnson, "Ann-based modeling for coagulant dosage in drinking water treatment plant," in *International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, Volume: 06 Issue: 05, May 2019*, 2019.

[60] B. Li, C. Lu, J. Zhao, J. Tian, J. Sun, and C. Hu, "Operational parameter prediction of electrocoagulation system in a rural decentralized water treatment plant by interpretable machine learning model," *Journal of Environmental Management*, vol. 333, p. 117416, 2023.

[61] S. Lin, J. Kim, C. Hua, M.-H. Park, and S. Kang, "Coagulant dosage determination using deep learning-based graph attention multivariate time series forecasting model," *Water Research*, vol. 232, p. 119665, 2023.

[62] G.-D. Wu and S.-L. Lo, "Predicting real-time coagulant dosage in water treatment by artificial neural networks and adaptive network-based fuzzy inference system," *Engineering Applications of Artificial Intelligence*, vol. 21, pp. 1189–1195, 12 2008.

[63] S. Heddam, A. Bermad, and N. Dechemi, "Anfis-based modelling for coagulant dosage in drinking water treatment plant: A case study," *Environmental monitoring and assessment*, vol. 184, pp. 1953–71, 05 2011.

[64] R. F. Yu, S.-F. Kang, S.-S. Liaw, and M.-C. Chen, "Application of artificial neural network to control the coagulant dosing in water treatment plant," *Water Science and Technology*, vol. 42, pp. 403–408, 08 2000.

[65] K. Zhang, G. Achari, H. Li, A. Zargar, and R. Sadiq, "Machine learning approaches to predict coagulant dosage in water treatment plants," *International Journal of System Assurance Engineering and Management*, vol. 4, 06 2013.

[66] C. Kim and M. Parnichkun, “Mlp, anfis, and grnn based real-time coagulant dosage determination and accuracy comparison using full-scale data of a water treatment plant,” *Journal of Water Supply: Research and Technology - Aqua*, vol. 66, 12 2016.

[67] G. Gholikandi, M. Delnavaz, and R. Riahi, “Use of artificial neural network for prediction of coagulation/flocculation process by pac in water treatment plant,” *Environmental Engineering and Management Journal*, vol. 10, pp. 1719–1725, 11 2011.

[68] A. Kote and D. Wadkar, “Modeling of chlorine and coagulant dose in a water treatment plant by artificial neural networks,” *Engineering, Technology & Applied Science Research*, vol. 9, pp. 4176–4181, 06 2019.

[69] H. Tahraoui, A.-E. Belhadj, N. Moula, S. Bouranene, and A. Amrane, “Optimisation and prediction of the coagulant dose for the elimination of organic micropollutants based on turbidity,” *Kemija u industriji/Journal of Chemists and Chemical Engineers*, vol. 70, p. 675–691, 03 2021.

[70] S. Heddam, “Extremely randomized tree: a new machines learning method for predicting coagulant dosage in drinking water treatment plant,” in *Water Engineering Modeling and Mathematic Tools* (P. Samui, H. Bonakdari, and R. Deo, eds.), pp. 475–489, Elsevier, 2021.

[71] Z. Shi, C. W. Chow, R. Fabris, J. Liu, E. Sawade, and B. Jin, “Determination of coagulant dosages for process control using online uv-vis spectra of raw water,” *Journal of Water Process Engineering*, vol. 45, p. 102526, 2022.

[72] J. Leeuwen, C. Chow, D. Bursill, and M. Drikas, “Empirical mathematical models and artificial neural networks for the determination of alum doses for treatment of southern australian surface waters,” *Aqua*, vol. 48, pp. 115 – 127, 05 2002.

[73] P. Chawakitchareon, N. Boonnao, and P. Charutragulchai, “Prediction of alum dosage in water supply by weka data mining software,” *Frontiers in Artificial Intelligence and Applications*, vol. 292, 2017.

[74] H. Wang, T. Asefa, and J. Thornburgh, “Integrating water quality and streamflow into prediction of chemical dosage in a drinking water treatment plant using machine learning algorithms,” *Water Supply*, vol. 22, 12 2021.

[75] D. V. Wadkar, R. S. Karale, and M. P. Wagh, “Application of soft computing in water treatment plant and water distribution network,” *Journal of Applied Water Engineering and Research*, vol. 10, no. 4, pp. 261–277, 2022.

[76] D. Chamanthi and D. Jayaweera, *A comprehensive study on developing neural network models for predicting the coagulant dosage and treated water qualities for a water treatment plant*. PhD thesis, Universiti Sains Malaysia, 07 2019.

[77] C. M. Kim, *Coagulant Dosage Determination Using Neural Networks and ANFIS in Drinking Water Treatment Plant*. Doctor of engineering in mechatronics, Asian Institute of Technology School of Engineering and Technology, Thailand, May 2017. Committee: Prof. Manukid Parnichkun (Chairperson), Dr. Sangam Shrestha, Dr. Mongkol Ekpanyapong, External Examiner: Prof. Marie Veronique Le Lann, Department of Genie Electrique et Informatique, INSA de Toulouse, France.

[78] S. Studer, T. B. Bui, C. Drescher, A. Hanuschkin, L. Winkler, S. Peters, and K. Müller, “Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology,” *CoRR*, vol. abs/2003.05155, 2020.

[79] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002.

[80] Hydrochem S.A.S, *User Manual, Model HCUPC01, HYDROCOLOR*, 2016.

[81] BRAND GMBH + CO KG, *Titrette® Flaschenaufsatzbürette | Bottle-Top Burette*. Otto-Schott-Str. 25, 97877 Wertheim, Germany, 2022. The original operating manual is in German. Other languages are translations of the original operating manual.

[82] HACH, *Manual HACH 2100Q y 2100Qis*, 6 ed., 2021. DOC022.92.80041.

[83] Thermo Fisher Scientific Inc., *Thermo Scientific Orion Star A111 Benchtop and Star A121 Portable pH Meters Reference Guide*, 2011. All trademarks are the property of Thermo Fisher Scientific, Inc. and its subsidiaries. The specifications, descriptions, drawings, ordering information and part numbers within this document are subject to change without notice. This publication supersedes all previous publications on this subject.

[84] D. E. Farrar and R. R. Glauber, "Multicollinearity in regression analysis: the problem revisited," *Review of Economics and Statistics*, vol. 57, no. 1, pp. 92–107, 1975.

[85] P. Mandeville, "Tema 24: Observaciones perdidas," *CIENCIA-UANL, ISSN 1405-9177, Vol. 13, Nº. 3, 2010, pags. 313-324*, 01 2010.

[86] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Tutorial in Biostatistics*, vol. ?, no. ?, p. ?, 2010. Received 3 September 2009, Accepted 14 July 2010, Published online 30 November 2010.

[87] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *International Joint Conference of Artificial Intelligence*, 1995.

[88] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *IEEE Intelligent Systems*, vol. 24, no. 2, pp. 8–12, 2009.

[89] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, Springer, second ed., 2009.

[90] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2nd ed., 2001.

[91] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[92] C. C. Aggarwal, *Outlier Analysis*. New York, NY: Springer, 1 ed., 2013. Number of Pages: XV, 446. Topics: Data Mining and Knowledge Discovery, Artificial Intelligence, Statistics and Computing, Data and Information Security, Database Management, Information Storage and Retrieval.

[93] M. Jansche, "Maximum expected f-measure training of logistic regression models," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, pp. 692–699, 2005.

[94] S. yb developers, "Classification report: Visual classification report for classifier scoring," 2023. Accedido el: [8 de agosto del 2023].

[95] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, May 2019.

[96] A. G. Huerta, "Algoritmos de clasificación para datasets equilibrados: Análisis y comparativa," Master's thesis, Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, 2018. Proyecto de Fin de Grado, Grado en Ingeniería de Computadores.

[97] F. Rodríguez Torres, *SMOTE-D, Una Versión Determinista de SMOTE*. PhD thesis, Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla, Marzo 2017. Tesis para el grado de Maestro en Ciencias en el área de Ciencias Computacionales, Supervisada por Jesús Ariel Carrasco Ochoa y José Francisco Martínez Trinidad. ©INAOE 2017 Derechos Reservados. El autor otorga al INAOE el permiso de reproducir y distribuir copias de esta tesis mencionando la fuente.

[98] Z. Wang, C. Wu, K. Zheng, X. Niu, and X. Wang, "Smotetomek-based resampling for personality recognition," *IEEE Access*, vol. 7, pp. 129678–129689, 2019.

[99] N. Mashette, "Balanced bagging classifier: Bagging for imbalanced classification." `https://medium.com/@nageshmashette32/balanced-bagging-classifier-bagging-for-imbalanced-classification-dfba66c44c14`, 2023. Accedido el: 4 de Agosto de 2023.

[100] G. E. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–252, 1964.

[101] R. B. Mayorga-Ponce, D. C. Graciano-Ventura, A. M. Hernández, P. M. Moctezuma-Jiménez, B. Pérez-Galindo, and A. Roldan-Carpio, "Cuadro comparativo de análisis paramétrico y no paramétrico: Main comparative table of parametric and non-parametric analysis," *Educación y Salud Boletín Científico Instituto de Ciencias de la Salud*, vol. 10, no. 20, pp. 90–93, 2022.

[102] L. D. S. Riveros, "Aplicación de técnicas data mining para el análisis del desempeño escolar en cundinamarca (colombia) 2015 a 2019," Master's thesis, Escuela Colombiana de Ingeniería Julio Garavito, Decanatura de Ingeniería Industrial, Decanatura de Ingeniería de Sistemas, Decanatura de Matemáticas, Maestría en Ciencia de Datos, Bogotá D.C., Colombia, 2022. Director: Wilmer Pineda Ríos, Magister en Ciencias Matemáticas. Codirector: Iván Mauricio Mendivelso Ramírez, Magister en Antropología.

[103] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York/Berlin/Heidelberg: Springer, 2nd ed., 2008.

[104] I. Amazon Web Services, "¿qué es el sobreajuste?," 2023. Accedido: 15 de agosto del 2023.

[105] R. Gil Rubio, E. A. Cruz Perez, and O. Perdomo Charry, "Modelos de machine learning para clasificar la cartera en un fondo de pensiones," Master's thesis, Facultad de Estadística, Universidad Santo Tomás, Colombia, Mar 2022.

[106] R. López, "Machine learning con python," 2015. Accedido: Julio 15, 2023.

[107] S. Fernández Villafañez, "Métodos de regresión y clasificación basados en árboles," Master's thesis, Universidad de Valladolid, Escuela de Ingenierías Industriales, Valladolid, Spain, June 2022.

[108] E. Menasalvas, A. Rodriguez, M. Guzman, S. Jimenez, and S. Duque, "Algoritmos de machine learning," *Newsletter trimestral, Catedra iDANAE*, vol. 1, no. 1, 2021.

[109] J. Faraway, *Linear Regression in R*. 2000.

[110] F. J. Barón López and F. Téllez Montiel, *Apuntes de Bioestadística*. Departamento de Matemática Aplicada, Universidad de Málaga, 2000.

[111] V. Rodriguez-Galiano, M. Sanchez-Castillo, M. Chica-Olmo, and M. Chica-Rivas, "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines," *Ore Geol. Rev.*, vol. 71, pp. 804–818, 2015.

[112] A. Navada, A. N. Ansari, S. Patil, and B. A. Sonkamble, "Overview of use of decision tree algorithms in machine learning," in *Proc. - 2011 IEEE Control Syst. Grad. Res. Colloquium, ICSGRC 2011*, pp. 37–42, 2011.

[113] C. Stergiou and D. Siganos, "Neural networks," n.d. Retrieved March 30, 2018.

[114] B. Ghimire, J. Rogan, V. Galiano, P. Panday, and N. Neeti, "An evaluation of bagging, boosting, and random forests for land-cover classification in cape cod, massachusetts, usa," *GIScience Remote Sens.*, vol. 49, no. 5, pp. 623–643, 2012.

[115] L. Breiman, "Arcing the edge," tech. rep., Technical Report 486, Statistics Department, University of California, 1997.

[116] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Annals of Statistics*, vol. 29, pp. 1189–1232, October 2001.

[117] J. H. Friedman, "Stochastic gradient boosting," *Computational statistics and data analysis*, vol. 38, no. 4, p. 367–378, 2002.

[118] L. Mason, J. Baxter, P. L. Bartlett, and M. R. Frean, "Boosting algorithms as gradient descent," in *Advances in neural information processing systems*, p. 512–518, 2000.

[119] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, pp. 3–42, 2006.

[120] M. J. Illera Bustos, D. C. Bastos Guerrero, and S. B. Sepúlveda Mora, "Adaptive neuro-fuzzy inference system (anfis) for the estimation of global solar radiation," *Investigación e Innovación en Ingenierías*, vol. 9, no. 1, pp. 34–49, 2021.

Anexo A.

Aprendizaje automático y principales algoritmos

A.1 Aprendizaje automático

ML es un campo de las ciencias informáticas relacionado con la inteligencia artificial (AI, *Artificial Intelligence*) y el análisis predictivo o aprendizaje estadístico. En esencia, hace referencia a la habilidad de un software o aparato para aprender y predecir comportamientos futuros, mediante la implementación de ciertos algoritmos en respuesta a la entrada de datos en su sistema [105].

Esta disciplina se utiliza para interpretar y clasificar grandes conjuntos de datos, ofreciendo un amplio abanico de soluciones para diferentes casos de uso, desde el mercadeo y análisis de datos, hasta la conducción autónoma de vehículos y el reconocimiento de voz e imágenes. Su implementación ha sido un catalizador para la innovación en varias industrias, mejorando procesos y generando nuevas oportunidades de negocio [105].

Los modelos creados a partir de algoritmos de ML tienen la capacidad de predecir la dosis de coagulante necesaria para reducir la turbidez del agua tratada, basándose en diversos parámetros de entrada. Estos algoritmos utilizan datos históricos para aprender y mejorar su capacidad predictiva con el tiempo. Además, la capacidad de actualizar continuamente estos modelos con nuevos datos mejora su precisión y confiabilidad a largo plazo. En la sección 2.2, se puede encontrar más información sobre el uso de estos modelos en este contexto.

Existen dos categorías principales de métodos de ML: aprendizaje no supervisado y aprendizaje supervisado. Los métodos no supervisados basan su proceso de entrenamiento en un conjunto de datos sin etiquetas o clases previamente definidas. Estos se enfocan en tareas de agrupamiento o clustering, con el objetivo de encontrar grupos homogéneos entre sí y heterogéneos entre otros [106].

Por otro lado, los problemas de aprendizaje supervisado enseñan o entrenan al algoritmo a partir de datos que ya vienen etiquetados con la respuesta correcta. Cuanto mayor sea el conjunto de datos, mayor es la posibilidad de que el algoritmo aprenda sobre el tema. Una vez concluido el entrenamiento, se le brindan nuevos datos, sin las etiquetas de las respuestas correctas, y el algoritmo utiliza la experiencia pasada que adquirió durante la etapa de entrenamiento para predecir un resultado [106].

El enfoque de este proyecto se situará en el aprendizaje supervisado. Existen dos tipos principales de problemas dentro de las técnicas de predicción de aprendizaje supervisado: los problemas de clasificación y los de regresión, distinguiéndose principalmente por el tipo de resultado que

se busca obtener. En los problemas de clasificación, la variable objetivo es cualitativa, mientras que en los problemas de regresión, la variable objetivo es cuantitativa [107]. Por ejemplo, un modelo de clasificación podría brindar un diagnóstico sobre una enfermedad basado en algunos síntomas o característica, mientras que un modelo de regresión podría prever el precio de una casa en base a sus especificaciones.

A.1.1 Principales algoritmos

De acuerdo con [108], el aprendizaje supervisado engloba una serie de algoritmos fundamentales que se pueden utilizar tanto para tareas de clasificación como de regresión. Estos métodos fundamentales que integran esta categoría son:

- **Regresión Lineal:** Dentro del campo del aprendizaje automático, la regresión lineal es comúnmente empleada para fines de previsión y predicción. Esto se debe a que el análisis de regresión tiene la capacidad de identificar relaciones causales entre variables dependientes e independientes. Conforme a los modelos de regresión, es posible predecir las variables dependientes a partir de las variables independientes [109]. Este tipo de análisis estima el valor de la variable dependiente "y" basándose en el espectro de valores de la variable independiente "x". Así, la regresión lineal establece cómo la variable dependiente se ve influenciada por sus correspondientes variables independientes.
- **Regresión Logística:** La regresión logística es un método estadístico que se utiliza cuando la variable de respuesta es dicotómica, es decir, puede tomar uno de dos posibles valores, como "Éxito.º "Fracaso". A través de este modelo, es posible estudiar cómo diversas variables independientes afectan la probabilidad de que ocurra un evento específico. Por ejemplo, podría utilizarse para determinar la probabilidad de que un estudiante tenga éxito en su primer semestre, basándose en distintas variables como sus notas previas, horas de estudio, entre otras [110].
- **K-Vecinos más Cercanos (KNN, *K-Nearest Neighbors*):** El método KNN, se encarga de conservar todos los casos existentes para posteriormente categorizar nuevos datos según ciertas medidas de similitud, como ciertas funciones de distancia. Este método categoriza los datos según la categorización prevaleciente de sus K vecinos inmediatos, basándose en la medida seleccionada. Por poner un ejemplo, con un K valorado en 1, el dato es clasificado según la categorización de su vecino más inmediato. Para datos continuos, se emplean diferentes tipos de distancias: Euclidiana, Manhattan y Minkowsky [111].
- **Árboles de Decisión (DT, *Decision Tree*):** Los árboles de decisión representan una metodología que permite dividir datos de manera recursiva en estructuras secuenciales y jerárquicas. En estas estructuras, los nodos internos o de decisión actúan como pruebas basadas en patrones conocidos derivados de los datos de entrada, mientras que los nodos de hoja señalan el valor de la clase o atributo, que puede interpretarse como categorías de dichos patrones. Al filtrar las pruebas a través de la estructura del árbol, se obtiene una salida coherente con el patrón de entrada. Los algoritmos basados en árboles de decisión tienen aplicaciones variadas y pueden reemplazar procedimientos estadísticos en tareas como la búsqueda de datos, extracción de texto y mejora de motores de búsqueda, además de tener múltiples usos en diversos ámbitos médicos [112].
- **Máquina de Vectores de Soporte (SVM, *Support Vector Machine*):** SVM forma parte de los algoritmos de aprendizaje supervisado destinados a la categorización binaria de vectores con múltiples características. Inicialmente se ideó como un clasificador lineal, pero

con el tiempo evolucionó hacia formas no lineales y se adaptó para abordar problemas de regresión [111]. Su esencia radica en llevar las características iniciales a un espacio de mayor dimensionalidad, donde las categorías se puedan distinguir linealmente mediante un hiperplano. Esta técnica toma un conjunto de datos de entrenamiento compuesto por N muestras, cada una con un vector de L características, y sus respectivos resultados conocidos V.

- **Redes Neuronales Artificiales (ANN, *Artificial Neural Network*):** Las ANN se inspiran en el funcionamiento del sistema nervioso biológico, intentando emular cómo el cerebro procesa la información. Se diseñan para aplicaciones específicas como el reconocimiento de patrones y la clasificación, y aprenden de manera similar a cómo lo hacen los seres humanos: a través de ejemplos. Durante su entrenamiento, se establecen conexiones entre neuronas y se asignan pesos a estas conexiones para lograr una configuración que cumpla con los requisitos deseados. Estas redes consisten en neuronas artificiales que imitan el comportamiento de las neuronas biológicas, procesando múltiples señales de entrada para generar una salida basada en una función de activación y los pesos asignados durante el entrenamiento. Dependiendo de la naturaleza del problema a resolver, existen diferentes tipos de ANN, siendo las supervisadas y no supervisadas las más prevalentes. Las primeras requieren una intervención activa del usuario durante el entrenamiento, mientras que las segundas generan conocimiento autónomamente a partir de los datos proporcionados. Las ANN se han aplicado ampliamente en áreas tan variadas como las ciencias naturales, salud, seguridad y arte [113].

Además, existen enfoques más sofisticados conocidos como métodos 'ensemble'. Estos métodos combinan múltiples modelos con el objetivo de mejorar la precisión y robustez de las predicciones. Entre los métodos 'ensemble' más comúnmente usados, según [108], destacan los siguientes:

- **Bosques Aleatorios (RF, *Random Forest*):** Los bosques aleatorios integra el desempeño de múltiples algoritmos de aprendizaje automático con el objetivo de clasificar o prever el valor de una determinada variable. Esta técnica requiere de una muestra de entrada compuesta por los valores de distintas características analizadas en un área de entrenamiento específica. A partir de ella, genera un conjunto de K árboles de decisión y luego promedia sus resultados. Con el propósito de reducir la correlación entre los diferentes árboles, los bosques aleatorios potencian la diversidad de estos construyéndolos a partir de variados subconjuntos de datos de entrenamiento, creados mediante una estrategia conocida como "bagging"[114].

- **Aumento de Gradiente (GB, *Gradient Boosting*):** GB es una técnica de aprendizaje automático aplicable a problemas de regresión y clasificación. Este método genera un modelo de predicción compuesto por un conjunto de modelos predictivos básicos, comúnmente árboles de decisión. El proceso de construcción del modelo se realiza de manera gradual, similar a otros métodos de boosting. Lo que distingue a GB es su capacidad para optimizar una función de pérdida diferenciable arbitraria. Breiman observó que el "boosting" podría interpretarse como un algoritmo de optimización sobre una función de coste adecuada [115]. Posteriormente, Friedman propuso algoritmos de "boosting" con enfoque en gradientes para regresión [116, 117]. Esta perspectiva fue generalizada por Mason et al., quienes describieron los algoritmos de "boosting" como algoritmos de descenso de gradientes funcionales[118]. Esta visión ha impulsado la creación de algoritmos de "boosting" en diversas áreas del aprendizaje automático y la estadística.

- **Árboles Extremadamente Aleatorios (ERT, *Extremely Randomized Trees*):** El algoritmo ERT se aplica a problemas de aprendizaje supervisado con múltiples variables numéricas de entrada. Difiere de otros métodos basados en árboles porque elige puntos de corte de forma totalmente aleatoria y utiliza todo el conjunto de datos para desarrollar los árboles. Las predicciones se combinan usando votación mayoritaria o promedio aritmético, según el problema. Esta técnica busca reducir la varianza mediante aleatorización y minimizar el sesgo utilizando toda la muestra de aprendizaje. Su eficiencia computacional es destacable, siendo similar a otros métodos pero con un procedimiento de división de nodos más simplificado. [119]

Finalmente, los estudios realizados con los algoritmos han permitido el desarrollo de modelos que combinan varias técnicas. Un ejemplo de ello es el Sistema Adaptativo de Inferencia Neuro-Difusa (ANFIS, *Adaptive Network-based Fuzzy Inference System*), que integra una ANN con Sistemas de Inferencia Difusa (FIS, *Fuzzy Inference System*, tal y como se discute en [120].

Anexo B.

Entrevista con el ingeniero Mauricio Ramirez

B.1 Preguntas después de realizar primer EDA

Al examinar la variable de dosis de coagulante, observamos que más del 50 % de los datos están ausentes. ¿Hay alguna razón específica para esta alta proporción de datos nulos?

Respuesta: Al examinar situaciones específicas en las que no hay datos sobre la dosis de coagulante, es evidente que en la mayoría de los casos no se registró el dato porque no se aplicaron coagulantes. Esto usualmente sucede cuando la turbiedad es menor a 3. A partir de 2021, establecí una norma: durante el periodo nocturno, de 8 de la noche a 8 de la mañana, se debe aplicar coagulante independientemente de la turbiedad. Sin embargo, durante el horario diurno, de 8 am a 8 pm, sólo se aplica coagulante si la turbiedad supera 3. Por ejemplo, si la turbiedad está en 2.8, 2.6 o 2.9 durante el día, no se aplica coagulante. Pero si esas mismas cifras se presentaran después de las 8 de la noche, se debería dosificar.

Se detectó una elevada cantidad de datos nulos no solo en la variable de dosis de coagulante, sino también en variables clave como cal primaria, coagulante líquido y coagulante granulado. ¿Existe algún motivo particular para esta significativa ausencia de datos?

Respuesta: En relación a la ausencia notable de datos en ciertas variables, puedo explicar que la cal primaria se aplica exclusivamente cuando el agua contiene altas concentraciones de minerales lo que ocasiona que el agua llegue a la entrada de la planta con un pH bajo, situación que se presenta en contadas ocasiones. En el caso del coagulante líquido, se realizaron solamente pruebas limitadas, lo que se refleja en la falta de registros consistentes. Por otro lado, para el coagulante granulado, su cuantificación se lleva a cabo al finalizar cada turno de los operarios, registrando los kilogramos utilizados durante ese periodo.

¿Nos podrías proporcionar una definición clara de cada una de las variables de las hojas de cálculo y sus unidades de medida?

Respuesta:

- **Caudal:** Cantidad de agua que entra a la planta y se mide en L/s
- **Cal primaria:** Es la aplicación de hidróxido de calcio en el agua cruda, medido en kilogramos. La dosis suelen ser bajas, oscilando entre 1 y 4 mg/L.
- **pH del agua cruda:** Indica la acidez o basicidad del agua. Si el pH es menor a 6.5, el agua es ácida. Si es mayor a 7.5, es básica. Un pH entre 6 y 7 indica agua neutra.
- **Color del agua cruda:** El color se refiere a la cantidad de minerales en el agua que impiden que la luz se refleje adecuadamente. Se mide en unidades de platino cobalto.
- **Turbiedad:** Está relacionada con la cantidad de partículas en suspensión en el agua que afectan cómo la luz la atraviesa y se refleja.
- **Alcalinidad:** Mide la cantidad de carbonato de calcio en el agua, lo que indica la capacidad del agua para mantener su pH. Si hay suficiente alcalinidad, el pH no tiende a bajar demasiado. Se mide en miligramos por litro de carbonato de calcio.
- **Conductividad de agua cruda:** Se mide en $\mu S/cm^2$ y se refiere a la capacidad del agua para conducir electricidad, dependiendo de su contenido de sales y otros contaminantes.
- **Coagulante granulado:** Representa la cantidad, en kilos, de coagulante que se usa en un turno de ocho horas. Generalmente, es sulfato de aluminio aplicado en forma sólida.
- **Coagulante líquido:** Se refiere al uso de sulfato de aluminio líquido tipo B o policloruro de aluminio. Su dosis es en mg/L y se aplica en mL/min. Puede intercambiarse con el coagulante granulado dependiendo de la densidad del producto.
- **Dosis de coagulante:** Se mide en mg/L. Se aplica ya sea en forma líquida o granulada, siendo la más común la granulada.
- **Cloro de agua tratada:** Se mide en mg/L y representa la cantidad de desinfectante residual en el agua para neutralizar microorganismos patógenos.
- **Alcalinidad de agua tratada:** Similar a la alcalinidad de agua cruda, pero generalmente es más baja en el agua tratada debido a la aplicación de coagulante.

Es fundamental incluir en el proceso de limpieza la eliminación de datos atípicos para cada variable presente en nuestro conjunto de datos. Para realizar esto de manera efectiva, necesitamos conocer los valores mínimos y máximos que pueden registrarse en cada variable. Estos valores se pueden encontrar en las especificaciones de los equipos utilizados para medir cada una de las variables. Si es posible, por favor proporcione el nombre y la referencia de los equipos empleados para la medición de las variables. En caso de que esté familiarizado con los rangos de medición específicos para cada variable, sería útil si pudiera describirlos para facilitar el proceso de limpieza de datos.

Respuesta:

- **Caudal:** El caudal de operación en la planta el tablazo puede variar. A veces puede bajar a 200L/s si la bocatoma se tapa. Los sensores pueden medir caudales que se mueven entre 471 y 1348 L/s. Aunque un valor real de caudal máximo debe ser de 800L/s.
- **Cal Primaria:** No hay un límite específico para la cal primaria. Las mediciones se hacen diarias o por turno.

- **pH:** El pH del agua tratada debe estar entre 6.5 y 9. En el caso del agua cruda, el pH varía entre 4 y 10. Un pH por debajo de 6.5 indica un agua muy ácida que carece de carbonatos. Un pH de 5 sería inusual y comparable con el vinagre. Los equipos pueden medir un pH de 0 a 14. Sin embargo, por calibración, el rango utilizado es de 4 a 10. En práctica, los valores máximos que se obtienen son de 9.
- **Color del Agua Cruda:** La relación entre color y turbiedad es aproximadamente de 1 a 1.8. Se puede medir un color de hasta 500 UPC con el equipo de hidrochen. Si supera este valor, se realiza una dilución para estimar el color.
- **Turbiedad:** El equipo puede medir hasta una turbiedad de 800 NTU. Si se supera este valor, se realiza una dilución para estimar la turbiedad.
- **Alcalinidad:** El rango de medición de alcalinidad es hasta 100 mg/L de carbonato de calcio.
- **Conductividad:** La conductividad puede ser medida hasta 12800 $\mu S/cm^2$
- **Cloro:** El equipo puede medir hasta 2.98 partes por millón (PPM) de cloro residual.

B.2 Preguntas después de EDA comparativo.

Tras imputar manualmente los datos conforme a las directrices previas, el número de ceros en los datos de dosis aumentó a 36880, representando el 45,3 % del total. Esta cantidad considerable podría introducir ruido en las recomendaciones. Por ello, proponemos un enfoque de recomendación en dos etapas: la primera determinará si es necesario dosificar y, si es así, la segunda etapa predecirá la cantidad de dosificación. Necesitamos su opinión sobre este enfoque y criterios para la primera etapa, como posibles reglas de coagulación según turbiedad, horarios, eventos climáticos y cierres en la planta. ¿Qué eventos específicos deberíamos considerar para coagular o abstenernos?

Respuesta: Para garantizar la calidad del agua, hemos establecido directrices específicas sobre el proceso de coagulación. En particular, cuando la turbiedad del agua es menor o igual a 3, se ha decidido no proceder con la coagulación ya que, a esos niveles de turbiedad, el proceso puede no ser necesario. Adicionalmente, hemos determinado que, independientemente de la turbiedad, es esencial coagular el agua entre las 8 de la noche y las 8 de la mañana. Esta decisión se toma con el propósito de asegurar que el agua conserve su calidad.

B.3 Preguntas tras revisar límites de calibración.

Tras revisar la variable de caudal, observamos que existen valores desde cero hasta cifras menores a 200. ¿Podrías explicar qué ocurrió con esos datos?

Respuesta: Cuando registramos datos con un valor de cero, esto indica que se suspendió el tratamiento. Sin embargo, en situaciones donde los valores son distintos de cero y menores a 200, se debe generalmente a complicaciones en la bocatoma, como obstrucciones que generan esos caudales reducidos. Dado que estos valores son atípicos, sugiero considerar la eliminación de dichos datos. De manera similar, recomendaría evaluar la exclusión de datos que superen los 800, ya que, por su inusualidad, es probable que representen errores de ingreso de información.

Tras examinar las variables de turbiedad y color, notamos que algunos valores exceden los límites de calibración. En la entrevista anterior, indicaste que ante tales situaciones se hacen diluciones para medir esos valores elevados. ¿Son confiables esos valores obtenidos tras la dilución?

Respuesta: Cuando el color y la turbiedad superan los límites establecidos, efectivamente realizamos diluciones. Sin embargo, estos valores diluidos no son totalmente confiables. Dado que la captura de estos datos se hace manualmente y que la técnica depende en gran medida de la experiencia del operario, puede introducirse un margen de error considerable. Si la proporción de valores que exceden los límites no es significativa, mi recomendación sería descartar esos datos y no considerarlos en futuros análisis.

Durante la anterior entrevista, se mencionó que el valor máximo medible de alcalinidad era 100. Sin embargo, hemos observado valores que superan ese límite. ¿Se realizan diluciones también para la alcalinidad cuando se exceden esos valores?

Respuesta: Para la alcalinidad, no se llevan a cabo diluciones. Además, el equipo no tiene la capacidad de medir valores superiores a 100. Así que, cualquier valor que exceda ese límite es erróneo y debe ser descartado.

Anexo C.

Tablero de control (Dashboard)

Se ha diseñado un dashboard que sirve como herramienta de visualización para el sistema de recomendación de coagulante en la planta de tratamiento de agua potable .[El] Tablazo.[en] la ciudad de Popayán. Este dashboard se ha desarrollado con el propósito de mejorar la eficiencia y la toma de decisiones en el proceso de tratamiento de agua, garantizando el cumplimiento de las políticas de aplicación de coagulante de la planta.

El dashboard se compone de dos partes fundamentales: el frontend y el backend. En el backend, se ha implementado una API REST desarrollada en python que incluye un bloque de validaciones y los modelos de regresión y clasificación. El bloque de validaciones es una parte crítica del sistema, ya que se adhiere estrictamente a la política de no aplicación de coagulante cuando la turbiedad del agua es menor a 3 unidades, y solo permite la aplicación de coagulante durante el período nocturno, entre las 8 p.m. y las 8 a.m., independientemente de la turbiedad. Esto garantiza un uso eficiente de los recursos y el cumplimiento de las normativas de tratamiento de agua.

Para el correcto funcionamiento del sistema de recomendación, se solicitan como entradas valores de las variables definidas para el modelo, así como la medición de turbiedad en el agua. El bloque de validación proporciona dos respuestas posibles: una recomendación de no aplicar coagulante al agua o una recomendación de aplicar coagulante. Esta decisión se basa en las validaciones mencionadas anteriormente.

Si el bloque de validaciones determina que se debe aplicar coagulante, el sistema procede a utilizar los modelos de regresión y clasificación. Estos modelos toman las entradas proporcionadas y generan recomendaciones específicas sobre la dosificación de coagulante requerida para optimizar el proceso de tratamiento de agua. Estas recomendaciones son valiosas para los operadores de la planta, ya que les permiten tomar decisiones informadas y eficientes en tiempo real, mejorando la calidad del agua tratada y reduciendo el consumo innecesario de coagulante.

El componente frontend de este dashboard fue desarrollado utilizando Node-RED, una herramienta de programación versátil diseñada para conectar dispositivos de hardware, APIs y servicios en línea de una manera creativa e innovadora.

Node-RED ofrece un entorno de desarrollo basado en el navegador que simplifica la creación de flujos de trabajo mediante la utilización de una amplia variedad de nodos disponibles en la paleta. Estos nodos se pueden conectar y configurar con facilidad, lo que permite diseñar flujos de trabajo personalizados y ejecutarlos con un simple clic.

El frontend en cuestión consta de un formulario que recopila las entradas necesarias para el modelo de recomendación. Este formulario incluye campos para ingresar los valores de las variables requeridas. Además, se integran dos botones fundamentales: uno destinado a enviar el formulario y otro que permite cancelar el envío en caso de ser necesario.

Cuando el usuario envía el formulario, el sistema procesa las entradas y muestra la recomendación generada por el sistema de recomendación. Esta recomendación es de gran importancia para el proceso de toma de decisiones en la planta de tratamiento de agua, ya que contribuye a optimizar la dosificación de coagulante y, por ende, a mejorar la calidad del agua tratada.

Para obtener una representación visual del frontend mencionado, se incluye la Figura C, que proporciona una vista completa de la interfaz de usuario y su funcionalidad para visulaizar el funcionamiento del dashboard se podrá encontrar un video de su funcionamento en youtube [1].

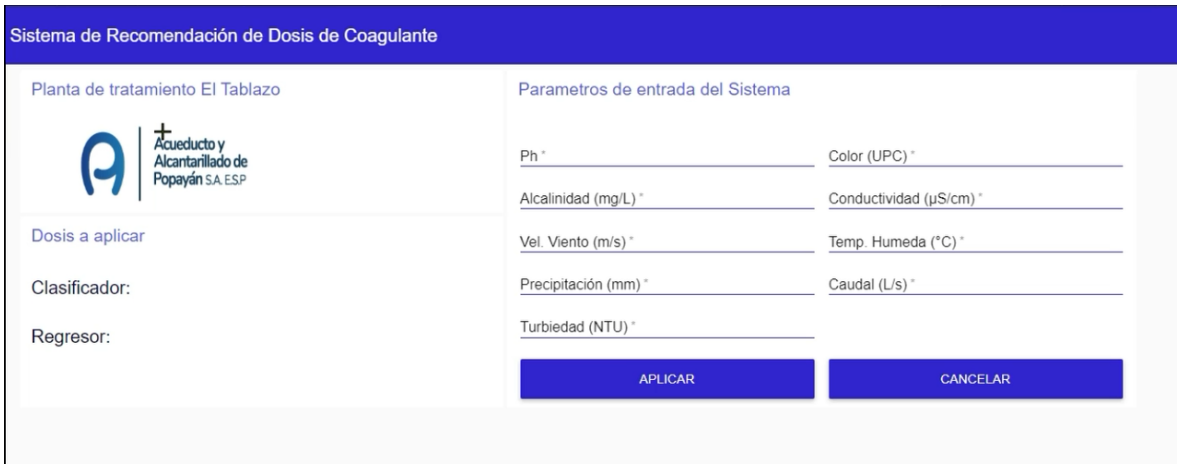

Figura C.1: Dashboard del sistema de recomendación de dosis de coagulante para la planta de tratamiento "El Tablazo" de la ciudad de Popayán.

Este dashboard no solo es una herramienta tecnológicamente avanzada, sino que también es una herramienta estratégica que contribuye significativamente a la gestión efectiva de recursos y al cumplimiento de las políticas de tratamiento de agua. Su implementación demuestra un enfoque innovador y orientado hacia la sostenibilidad en el ámbito de la potabilización del agua.

[1] Enlace de video del funcionamiento del dashboard: https://youtu.be/evf58Ty5lQY