Universidad de Oviedo
FACULTAD DE CIENCIAS

Máster en Análisis de Datos para la Inteligencia de Negocios

Trabajo de Fin de Máster

Aplicación de Modelos de Aprendizaje Automático para la Caracterización Hidrogeoquímica del Agua Subterránea en Boyacá, Colombia: Comparación de Modelos de Clasificación y Método Ensamblado Stacking

Adriana Blanco Alfonso

Dirigido por:
Norberto Corral Blanco
Carlos Carleos Artime

Julio, 2024

Resumen

El agua subterránea es un recurso fundamental para el consumo humano, la agricultura y diversas actividades industriales. Según el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), este recurso hídrico representa el 97% de las reservas de agua dulce en el planeta [1]. Esta investigación tiene como objetivo evaluar y comparar el rendimiento de distintos modelos de aprendizaje automático para la caracterización hidrogeoquímica del agua subterránea en Boyacá, Colombia. La pregunta central es: ¿Cuál es el rendimiento de diferentes modelos de aprendizaje automático en la clasificación hidrogeoquímica del agua subterránea según su fuente de captación (Aljibe, Manantial o Pozo)?

Para responder a esta pregunta, se utilizaron datos obtenidos de la Plataforma de Datos Abiertos del Estado Colombiano, que incluye información detallada sobre diversos parámetros hidrogeoquímicos. Los datos fueron analizados y preprocesados para diseñar, evaluar y comparar el rendimiento de modelos de clasificación supervisada, incluyendo: Árboles de Decisión, Bosques Aleatorios, Naive Bayes, Redes Neuronales y Métodos de Ensamblado Stacking.

Los resultados del estudio indican que los métodos de ensamblado, en particular el método Stacking, donde se combinan Bosques Aleatorios como modelo base y Naive Bayes como metamodelo, superan a los modelos individuales en términos de exactitud y capacidad discriminativa para clasificar los tipos de captación de agua subterránea. Específicamente, el modelo ensamblado Stacking alcanzó una exactitud del 87.5%, comparado con los modelos: Bosques Aleatorios (82.08%), Redes Neuronales (80.6%), Árbol de Decisión (72.64%) y Naive Bayes (70.75%).

Palabras clave: Aprendizaje Automático, Aprendizaje Supervisado, Modelos de Clasificación. Hidrogeoquímica, Captación de Agua Subterránea, Métodos ensamblados, Modelos de Clasificación, Boyacá Colombia.

Abstract

Groundwater is a fundamental resource for human consumption, agriculture, and various industrial activities. According to the Institute of Hydrology, Meteorology, and Environmental Studies (IDEAM), this water resource represents 97% of the planet's freshwater reserves. This research aims to evaluate and compare the performance of different machine learning models for the hydrogeochemical characterization of groundwater in Boyacá, Colombia. The central question is: What is the performance of different machine learning models in the hydrogeochemical classification of groundwater based on its source (Cistern, Spring, or Well)?

To answer this question, data obtained from the Colombian State Open Data Platform, which includes detailed information on various hydrogeochemical parameters, were used. The data were analyzed and preprocessed to design, evaluate, and compare the performance of supervised classification models, including: Decision Trees, Random Forests, Naive Bayes, Neural Networks, and Stacking Ensemble Methods.

The study results indicate that ensemble methods, particularly the Stacking method, which combines Random Forests as the base model and Naive Bayes as the meta-model, outperform individual models in terms of accuracy and discriminative ability to classify groundwater sources. Specifically, the Stacking ensemble model achieved an accuracy of 87.5%, compared to the individual models: Random Forests (82.08%), Neural Networks (80.6%), Decision Tree (72.64%), and Naive Bayes (70.75%).

Key words: Machine Learning, Supervised Learning, Classification Models, Hydrogeochemistry, Groundwater Catchment, Ensemble Methods, Classification Models, Boyacá Colombia.

Índice general

Índice de figuras

Índice de tablas

Capítulo I

Introducción

El agua subterránea es un recurso hídrico que se encuentra en el subsuelo, en las zonas saturadas de los acuíferos [1]. Este proporciona una fuente constante y confiable de agua para el consumo humano, la agricultura y la industria. La calidad y cantidad del agua subterránea pueden variar considerablemente dependiendo de factores como la geología del acuífero, las condiciones climáticas, la vegetación y las actividades humanas.

Cada tipo de captación puede proporcionar información valiosa sobre el estado del agua subterránea. Algunos tipos de captación son los pozos, aljibes y manantiales. Los pozos son perforaciones verticales en el suelo que alcanzan el acuífero y permiten extraer el agua subterránea. Los aljibes, por otro lado, son estructuras subterráneas diseñadas para recoger y almacenar el agua de lluvia. Los manantiales son puntos donde el agua subterránea aflora naturalmente en la superficie de la tierra.

Una forma de manejo de agua subterránea implica el estudio de la composición química del agua y su interacción con las rocas y minerales del subsuelo. Esta caracterización proporciona información sobre la calidad del agua, su origen, su evolución química y su aptitud para diferentes usos.

La importancia de las aguas subterráneas radica en su papel como fuente de agua potable, en la irrigación agrícola y en diversas actividades industriales. En muchas regiones, las aguas subterráneas son la única fuente de agua disponible, especialmente en tiempos de sequía. Además, debido a su ubicación subterránea están generalmente protegidas de la contaminación directa y son menos susceptibles a la evaporación en comparación con las fuentes de agua superficial y a medida que la situación del cambio climático se agrava,

las fuentes de agua subterránea se vuelven cada vez más esenciales para garantizar el acceso a agua segura y apta para el consumo humano [2].

Colombia, y en particular el departamento de Boyacá, presenta un gran potencial de aguas subterráneas [1] debido a su diversidad geológica y climática. Esta diversidad se refleja en la variabilidad de las características hidrogeoquímicas de sus aguas subterráneas. Además, la caracterización hidrogeoquímica puede ayudar a identificar posibles problemas de calidad del agua, como la contaminación por nitratos o metales pesados. También puede proporcionar pistas sobre la dinámica del flujo del agua subterránea, lo que es crucial para la gestión de los recursos hídricos.

Por ello, el objetivo principal de este proyecto es presentar una propuesta de diferentes modelos de aprendizaje automático para la caracterización hidrogeoquímica del agua subterránea en Boyacá, Colombia, evaluando y comparando su rendimiento en tres tipos de captación: aljibe, manantial y pozo. Para cumplir este objetivo principal, se tienen como objetivos específicos el preprocesamiento de los datos, el análisis y diseño de diferentes modelos de aprendizaje automático, y la comparación de las métricas de evaluación de los modelos, como matrices de confusión, exactitud, sensibilidad, entre otras. Estas métricas permiten entender la capacidad discriminativa de cada modelo en la clasificación de diferentes fuentes de captación de agua subterránea.

Con el fin de facilitar la comprensión del trabajo, el Capítulo II aborda los conceptos básicos sobre hidrogeoquímica y modelos supervisados. El Capítulo III presenta el problema y el contexto de la investigación. El Capítulo IV está dedicado al análisis y tratamiento del conjunto de datos, culminando con el diseño de los modelos de clasificación que son el objetivo de este estudio. En el Capítulo V se presentan los resultados y se comparan las medidas de rendimiento de los modelos. Finalmente, se exponen las conclusiones y las líneas futuras de investigación.

Capítulo II

Preliminares

Este capítulo presenta las nociones y conceptos clave para el entendimiento de la investigación. Se describen los fundamentos teóricos necesarios para contextualizar la caracterización hidrogeoquímica del agua subterránea, así como conceptos bases de las técnicas de aprendizaje supervisado utilizados.

2.1 Nociones básicas de Hidrogeoquímica

La hidrogeoquímica representa una fusión entre la hidrología y la química, que está destinada a comprender la composición química del agua subterránea y sus interacciones con los materiales geológicos que atraviesa.

2.1.1 Un poco de historia

La historia de la hidrogeoquímica se remonta a los primeros intentos por comprender la relación entre las características químicas del agua subterránea y su entorno geológico. A lo largo del siglo XIX, los científicos comenzaron a realizar mediciones sistemáticas de la composición química de las aguas subterráneas, observando variaciones significativas que atribuyeron a las diferencias en las formaciones geológicas locales. Este enfoque inicial sentó las bases para el desarrollo de la hidrogeoquímica como una disciplina que integraba principios de la hidrología, la geología y la química para entender la dinámica del agua en el subsuelo [3].

Los avances en la comprensión de la hidrogeoquímica durante el siglo XX se vieron impulsados por el desarrollo de técnicas analíticas más precisas y la aplicación de principios de termodinámica y geoquímica. Investigadores como Domenico y Schwartz [3] destacaron la importancia de considerar no solo la composición química del agua, sino

también los procesos físicos y biológicos que afectan su calidad y disponibilidad. En este contexto, la hidrogeoquímica evolucionó hacia el estudio integral de los sistemas acuíferos, incluyendo la evaluación de la vulnerabilidad a la contaminación y la gestión sostenible de los recursos hídricos.

2.1.2 Entendiendo el ciclo hidrológico y tipos de captación de aguas subterráneas

El ciclo hidrológico, fundamental para la distribución y disponibilidad del agua en la Tierra, representa un proceso continuo y dinámico que involucra múltiples etapas interconectadas [4]. Comienza con la evaporación del agua desde océanos, lagos y otros cuerpos de agua superficiales, impulsada por la energía solar. Esta agua vaporizada se eleva en la atmósfera, donde se condensa para formar nubes. Las nubes, a su vez, liberan agua en forma de precipitación, que puede caer nuevamente al océano o sobre la tierra.

Una parte de esta precipitación se infiltra en el suelo, proceso mediante el cual el agua se mueve desde la superficie hacia el subsuelo, contribuyendo a la recarga de los acuíferos subterráneos. Los acuíferos, reservorios naturales de agua subterránea almacenada en formaciones geológicas porosas, juegan un papel crucial en la regulación del suministro de agua dulce [4]. La *Figura 2.1* ilustra las etapas del ciclo hidrológico, destacando la interacción entre la precipitación, la infiltración y la recarga de acuíferos, elementos esenciales para la sostenibilidad de los recursos hídricos.

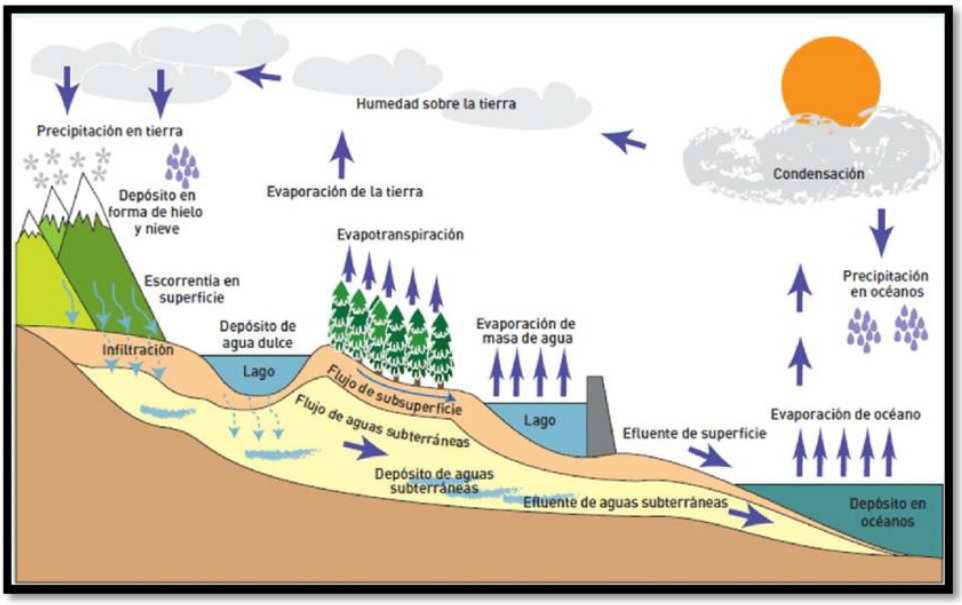


Figura 2.1 Ciclo Hidrológico del Agua

Fuente: Vega DG, Guerrero-Garcia-Rojas H, Seguí-Amórtegui L. (2016) [5].

2.1.2.1. Tipos de captación de aguas subterráneas

Los tipos de captación de aguas subterráneas, como pozos, manantiales y aljibes, representan métodos distintos para acceder y utilizar el recurso hídrico, cada uno con características únicas que afectan tanto la calidad como la cantidad del agua obtenida.

2.1.2.1.1. Manantial

Los manantiales son puntos naturales donde el agua subterránea fluye a la superficie de manera generalmente espontánea. Estos pueden variar desde pequeñas corrientes hasta grandes afloramientos, como se muestra en la *Figura 2.2*, que destaca la diversidad de formaciones de manantiales dependiendo de las condiciones geológicas locales y la estructura del acuífero. [6].

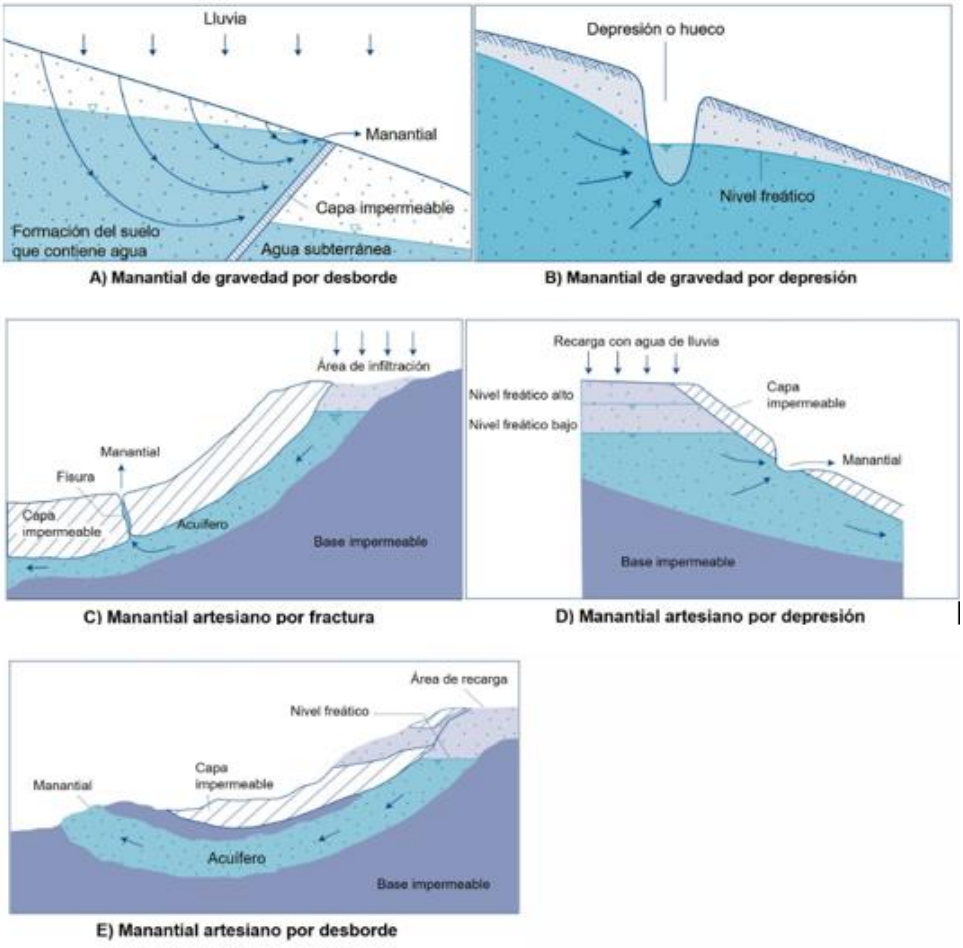


Figura 2.2 Tipos de manantial por gravedad y artesanios

Fuente: Pérez LR [7]

2.1.2.1.2. Pozo

Los pozos son perforaciones realizadas en el suelo para acceder al agua subterránea almacenada en los acuíferos. Pueden ser de varios tipos, como pozos profundos y someros, dependiendo de la profundidad y el método de construcción [6]. La construcción de un pozo incluye etapas de perforación, donde se utiliza maquinaria especializada para perforar el suelo hasta alcanzar el acuífero, luego se lleva a cabo la fase de entubación que consiste en introducir una tubería de revestimiento para prevenir el colapso del pozo y proteger el agua de contaminantes superficiales. Finalmente, en la mayoría de los casos, se instala una bomba para extraer el agua a la superficie [8].

2.1.2.1.3. Aljibe

Los aljibes son depósitos subterráneos construidos para almacenar agua, ya sea de origen pluvial o subterráneo. Tradicionalmente, se utilizaban para recolectar y almacenar agua de lluvia, pero también pueden estar conectados a un sistema de captación de aguas subterráneas. La construcción de aljibes incluye procesos de excavación, revestimiento y cubierta. Cada tipo de captación presenta variaciones significativas en la calidad y cantidad del agua recolectada, influenciadas por las características geológicas e hidrogeológicas locales. La permeabilidad del suelo, la estratigrafía del acuífero, y la presencia de contaminantes naturales o antropogénicos son factores que determinan la idoneidad y la sostenibilidad de cada método de captación.

2.1.3 Procesos hidrogeoquímicos básicos

La hidrogeoquímica se enfoca en estudiar la composición química de las aguas subterráneas y los procesos que controlan su evolución y calidad. Los principales procesos hidrogeoquímicos que afectan las aguas subterráneas incluyen la disolución y precipitación de minerales, intercambio iónico, reacción redox, adsorción y desorción.

La disolución y precipitación de minerales son procesos clave que controlan la composición química de las aguas subterráneas. La disolución ocurre cuando los minerales en las rocas y el suelo se disuelven en el agua subterránea, liberando iones que alteran su composición química. Por otro lado, la precipitación se da cuando los iones en el agua subterránea se combinan para formar minerales sólidos que se depositan fuera de la solución. Estos procesos están influenciados por factores como el pH, la temperatura y la presión parcial de dióxido de carbono [9,10]

El intercambio iónico es otro proceso crucial en la hidrogeoquímica de las aguas subterráneas. Ocurre cuando los iones en el agua intercambian lugares con iones en la superficie de los minerales arcillosos o en la materia orgánica del suelo. Este proceso puede modificar significativamente la composición química del agua subterránea y es fundamental para la eliminación de contaminantes [9].

Las reacciones de oxidación-reducción (redox) son esenciales para la química de las aguas subterráneas, ya que afectan la movilidad y la disponibilidad de muchos elementos y compuestos. Los procesos redox influyen en la concentración de oxígeno disuelto, hierro, manganeso, nitrato, sulfato y otros elementos. La presencia de materia orgánica y la actividad microbiana también juegan un papel importante en las reacciones redox [11].

La adsorción se refiere a la adhesión de iones o moléculas del agua subterránea a la superficie de sólidos como minerales o materia orgánica. La desorción es el proceso inverso, donde los iones o moléculas adheridos se liberan de la superficie de los sólidos al agua. Estos procesos son decisivos para el control de contaminantes y nutrientes en las aguas subterráneas [12].

La dilución y la mezcla son procesos físicos que afectan la concentración de solutos en las aguas subterráneas. La dilución ocurre cuando el agua subterránea se mezcla con agua de diferente composición química, reduciendo la concentración de solutos. La mezcla puede resultar de la infiltración de agua superficial o de la interacción entre diferentes cuerpos de agua subterránea [13].

2.1.4 Componentes hidrogeoquímicos analizados en esta investigación.

Con la conceptualización del ciclo hidrológico, el entendimiento de los tipos de captación de aguas subterráneas y los procesos hidrogeoquímicos básicos. Se presenta a continuación una breve descripción de los componentes hidrogeoquímicos analizados y presentes en el marco de esta investigación. Es importante destacar que, en la parte final del documento, se incluye un glosario con algunos conceptos clave para facilitar la comprensión de los términos utilizados.

- **Alcalinidad**: La alcalinidad es una medida de la capacidad del agua para neutralizar ácidos y está relacionada principalmente con la presencia de bicarbonatos, carbonatos e hidróxidos. Es un indicador importante del sistema buffer (o amortiguador) del agua, que estabiliza el pH.

- **Bicarbonato**: El bicarbonato es una forma importante de alcalinidad en el agua y juega un papel crucial en el sistema de buffer del agua, ayudando a mantener el pH estable.
- **Calcio:** El calcio es un ion importante que contribuye a la dureza del agua. Es esencial para la salud humana y el desarrollo de las plantas, pero en altas concentraciones puede causar incrustaciones en sistemas de agua.
- **Cloruro:** El cloruro es un anión común en las aguas subterráneas que puede provenir de fuentes naturales o de contaminación antropogénica, como la intrusión salina o el uso de desinfectantes.
- **Color:** El color del agua es un parámetro físico que puede indicar la presencia de sustancias disueltas o en suspensión. En el contexto de la hidrogeoquímica subterránea, el color del agua es un indicador clave de la calidad del agua y puede reflejar la presencia de compuestos orgánicos e inorgánicos que influyen en los procesos geoquímicos. El color del agua puede ser causado por varias sustancias naturales y antropogénicas. Siguiendo a Fetter CW, los compuestos inorgánicos, como los metales (por ejemplo, hierro y manganeso), también pueden contribuir al color del agua.
- **Conductividad Eléctrica a 25°C:** La conductividad eléctrica mide la capacidad del agua para conducir electricidad, lo cual está relacionado con la concentración de iones disueltos en el agua. Se reporta a 25°C para estandarizar las mediciones.
- **Dureza:** La dureza del agua es causada principalmente por la presencia de calcio y magnesio. Es un parámetro importante para evaluar la calidad del agua, ya que afecta el sabor, la formación de incrustaciones y la eficiencia de los detergentes.
- **Fosfato:** Sales o ésteres del ácido fosfórico.
- **Hierro Total:** El hierro total incluye todas las formas de hierro en el agua. El hierro puede provenir de la disolución de minerales en el suelo y rocas y puede causar problemas de sabor, color y precipitación en las redes de distribución de agua.
- **Magnesio:** El magnesio, junto con el calcio, contribuye a la dureza del agua. Es un nutriente esencial para los seres humanos y las plantas, pero en altas concentraciones puede causar efectos laxantes y problemas de incrustación.
- **Manganeso:** El manganeso es un oligoelemento que puede estar presente en el agua subterránea debido a la disolución de minerales. En altas concentraciones,

puede causar problemas de sabor, manchas en la ropa y efectos adversos en la salud.

- **Nitrato:** Sales formadas por la combinación del ácido nítrico HNO3 con una base.
- **pH:** El pH es una medida de la acidez o alcalinidad de una solución. Este parámetro es fundamental en el estudio de la calidad del agua, tanto superficial como subterránea, ya que influye en numerosas reacciones químicas y biológicas que ocurren en el medio acuático.
- **Potasio:** El potasio es un nutriente esencial para las plantas y está presente en el agua subterránea en concentraciones variables. Es menos móvil que otros iones, pero su presencia puede influir en la calidad del agua para usos agrícolas.
- **Sodio:** El sodio es un catión común en el agua subterránea. Sus concentraciones pueden ser indicativas de procesos naturales o de contaminación antropogénica. El sodio es importante en la evaluación de la calidad del agua para consumo humano y riego.
- **Sólidos Disueltos Totales:** Los sólidos disueltos totales representan la cantidad de minerales y sales disueltas en el agua.
- **Sólidos Totales:** Los sólidos totales incluyen todos los sólidos disueltos y en suspensión en el agua. Este parámetro es crucial para evaluar la calidad del agua en términos de potabilidad y uso agrícola.
- **Sulfato:** El sulfato es un anión común en el agua subterránea que puede provenir de la disolución de minerales como el yeso. En altas concentraciones, puede causar efectos laxantes y problemas de corrosión en sistemas de agua.
- **Temperatura:** La temperatura del agua subterránea afecta a las reacciones químicas y biológicas, la solubilidad de gases y la densidad del agua. Además, la temperatura es un indicador importante de la interacción entre el agua subterránea y superficial.
- **Turbiedad:** Mide la claridad del agua mediante la cantidad de partículas en suspensión en el agua. Algas, materia orgánica, contaminantes, entre otros, pueden contribuir a enturbiar el agua

2.2 Nociones de Aprendizaje Supervisado

El aprendizaje supervisado es una modalidad del aprendizaje automático donde se entrena un modelo usando un conjunto de datos etiquetado, como se observa en la *Figura 2.3*. Este conjunto consiste en pares de entrada-salida, donde cada entrada está asociada con una salida conocida. El objetivo del modelo es aprender una función que relacione las entradas con las salidas correctas. Este proceso incluye varios elementos fundamentales que permiten al modelo aprender a partir de los datos y realizar predicciones precisas. Estos elementos comprenden el conjunto de datos, el modelo en sí y la función de pérdida, la cual evalúa la diferencia entre las predicciones del modelo y las salidas reales.

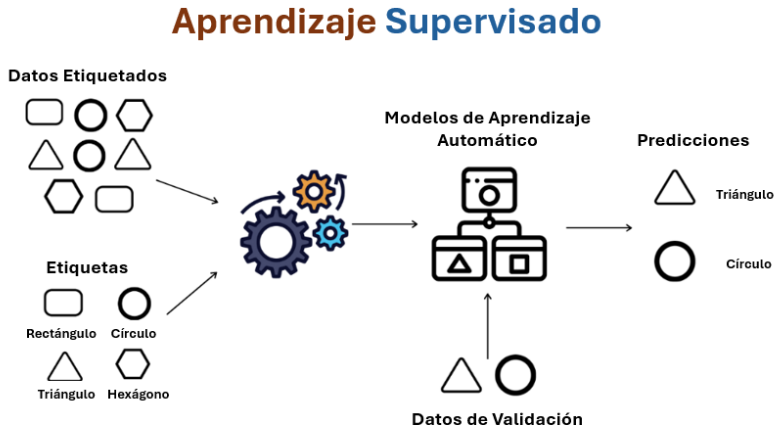


Figura 2.3 Aprendizaje supervisado

Fuente: Gupta N. [17]. Imagen adaptada.

El aprendizaje supervisado tiene sus raíces en la teoría estadística. A principios del siglo XX, la teoría de la probabilidad y la estadística proporcionaron las bases para el desarrollo de modelos predictivos. Entre los eventos clave en el desarrollo de estos algoritmos se destacan:

- **Años 1950-1970:** El perceptrón, desarrollado por el estadounidense Frank Rosenblatt en 1958, es una forma simple de red neuronal capaz de realizar tareas de clasificación, no obstante, este modelo presentó limitaciones en la identificación de patrones visuales [14]. Este período también vio el desarrollo de la prueba Turing y la primera conferencia sobre Inteligencia Artificial en *Darmouth College*.

- **Años 1970-1990:** Durante este tiempo, se avanzó en el desarrollo de algoritmos de aprendizaje automático, incluyendo el algoritmo de los vecinos más cercanos (k-NN) y los árboles de decisión. Estos métodos proporcionaron formas más flexibles y robustas de manejar datos complejos y no lineales [15].
- **Años 1990-2000:** El incremento de datos y del poder computacional permitieron el desarrollo de algoritmos más sofisticados. El auge de los métodos de ensamble, como los bosques aleatorios [16] y los métodos de boosting, mejoró significativamente la precisión y la estabilidad de los modelos predictivos.
- **Años 2010-Presente:** Con el gran volumen de datos y el avance en hardware, especialmente de las GPU, han permitido el desarrollo de redes neuronales profundas (deep learning). Estos modelos han transformado campos como el reconocimiento de imágenes y el procesamiento del lenguaje natural, alcanzando niveles de rendimiento antes inalcanzables.

2.2.1 Modelos de clasificación

Se presenta a continuación la descripción de los modelos de clasificación implementados en este estudio. Es importante señalar que, estos modelos de clasificación se entrenan utilizando datos específicos para identificar patrones que les permitan prever resultados en nuevos casos. Es así como, para estimar de manera imparcial las probabilidades de error se emplean conjuntos de datos diferenciados: uno se utiliza para entrenar el modelo y otro distinto para medir su efectividad o calidad.

2.2.1.1 Árboles de Decisión:

El árbol de decisión es una de las técnicas más populares y comprensibles en el campo del aprendizaje supervisado. Su desarrollo se remonta al año 1984, con trabajos pioneros de autores como Breiman, Friedman, Olshen y Stone, quienes introdujeron el concepto en su libro "Classification and Regression Trees" (CART) [18]. Los árboles de decisión han evolucionado significativamente desde entonces, y su simplicidad y efectividad los han convertido en herramientas ampliamente utilizadas en diversas aplicaciones.

Un árbol de decisión puede catalogarse como un árbol de clasificación (cuando la variable objetivo es cualitativa) o como un árbol de regresión (cuando la variable objetivo es cuantitativa). En esta investigación, nos centraremos en los árboles de clasificación. Un árbol de clasificación es un modelo de predicción empleado cuando la variable respuesta u objetivo es cualitativa. Este modelo utiliza un conjunto de reglas de decisión en forma de árbol para predecir el valor de la variable objetivo. La estructura del árbol consta de nodos, ramas y hojas, como se ilustra en la *Figura 2.4*:

- Nodo Raíz: El nodo inicial del árbol que representa toda la población o el conjunto de datos.

- Nodos de Decisión (t): Los nodos de decisión son puntos en los que se toman decisiones basadas en una prueba condicional sobre una característica específica. El conjunto de todos los nodos de decisión se denota como T. En un árbol binario, un nodo padre específico se representa por t_p, mientras que para los nodos hijos se emplean las etiquetas t_L y t_R para los nodos izquierdo y derecho respectivamente.

- Ramas: Conectan nodos y representan el resultado de una prueba condicional.

- Hojas: Nodos terminales que representan una clase o valor de la variable objetivo.

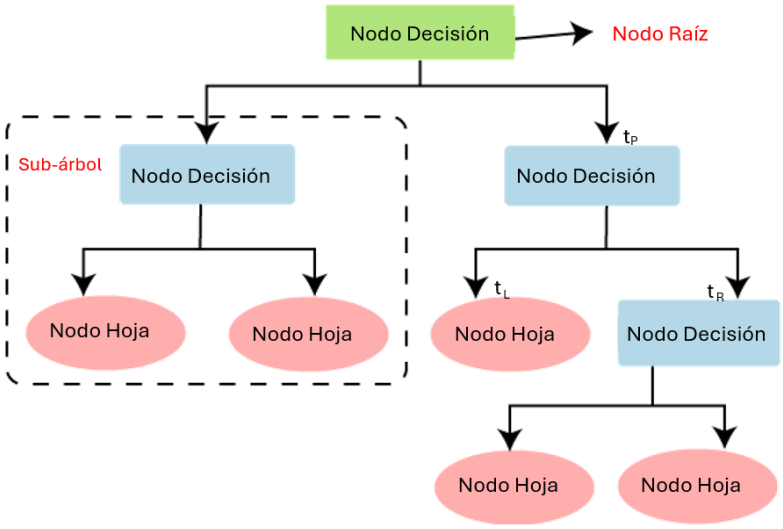


Figura 2.4 Estructura de árbol de decisión

Fuente: Jauregui AF [19]. Imagen adaptada

El proceso de construcción de un árbol de clasificación implica seleccionar la característica más relevante en cada nodo para dividir el conjunto de datos de manera que las clases se separen lo mejor posible. Esto se realiza utilizando criterios de división como el Índice de Gini, la Entropía o la ganancia de información.

Criterio de división:

Índice Gini: El índice de Gini mide la impureza de un nodo. Se define como la probabilidad de que un punto seleccionado al azar sea clasificado incorrectamente si se clasificara aleatoriamente de acuerdo con la distribución de las clases en el nodo.

$$Gini(D) = \sum_{j=1}^{k} p_j(1 - p_j) = 1 - \sum_{i=1}^{k} p_j^2$$

Donde p_j es la proporción de observaciones que pertenecen a la clase j en el conjunto de datos D. Si la clasificación fuera perfecta, las p_j sería unos o ceros y el índice de Gini valdría cero.

Entropía: La entropía mide la cantidad de incertidumbre o desorden en un conjunto de datos. Se define como:

$$Entropía(D) = -\sum_{j=1}^{k} p_j \, log\,(p_j), con\ la\ condición\ 0\log(0) = 0$$

Al igual que en el caso anterior, si todos los p_j fueran unos o ceros la entropía valdría cero.

Probabilidades de error:

Para estimar la probabilidad de error en la predicción, se emplea el criterio de restitución, que consiste en calcular la probabilidad de error condicionada al nodo-hoja.

$$r(t) = 1 - max\ p(j|t)$$

La probabilidad de cometer un error dado que se pertenece al $nodo - hoja\ t$ es:

$$R(t) = p(t)\, r(t)$$

La probabilidad global de error de un árbol T se estima sumando las probabilidades de error de todos los nodos:

$$R(T) = \sum_{t \in T} R(t)$$

Por último, cabe resaltar que, existen varios criterios para decidir cuándo dejar de dividir un nodo. Algunos de estos incluyen situaciones en las que el número de sujetos en el nodo está por debajo de un umbral específico. Comúnmente, se permite que el árbol crezca considerablemente y luego se podan las ramas que no aportan una mejora significativa en la reducción de la probabilidad de error.

Entre los criterios más utilizados para este propósito se encuentra el criterio de coste computacional, que tiene en cuenta la complejidad del árbol. Este criterio incorpora una penalización, denotada como cp (por omisión, $cp = 0.01$), por el incremento en complejidad, es decir, en el número de hojas. Sea $R(T)$ la tasa de clasificación errónea (riesgo) en el árbol T, y sea $| T |$ su número de hojas. El costo de un árbol se calcula como:

$$Rcp\ (T) = R(T) + cp \cdot | T | \cdot R(T_\infty)$$

Donde T_∞es el árbol sin divisiones. Para cada valor de cp se calcula el árbol con el costo mínimo.

2.2.1.2 Bosques Aleatorios

Los Bosques Aleatorios (Random Forest) son un algoritmo de aprendizaje automático basado en la teoría de los árboles de decisión y la técnica de ensamblaje presentado por Breiman [16], este método es ampliamente utilizado en tareas de clasificación y regresión debido a su robustez y capacidad para manejar grandes volúmenes de datos y características.

En el apartado 2.1.1.1 se explicó la construcción del árbol de decisión basada en el principio de reducción de impureza. Los bosques aleatorios, como se ilustra en la *Figura 2.5*, combinan múltiples árboles de decisión. De manera que, dada una muestra de N individuos con p variables independientes se construyen arboles de decisión siguiendo los pasos:

- Se elige al azar, con reemplazamiento, una muestra de tamaño N de los datos originales. Este método de muestreo se conoce como "*bootstrap*".
- Una vez seleccionada la muestra *bootstrap*, se procede a construir un árbol de decisión con los siguientes criterios:
    - Se elige al azar un número pequeño de variables independientes m, donde generalmente $m < p$.

- Se realiza una división binaria del conjunto de entrenamiento utilizando la variable que mejor predice la variable dependiente. Este proceso se repite en cada nodo hijo hasta que cada árbol se desarrolle tanto como sea posible sin realizar ninguna poda, asegurando que los nodos finales contengan muy pocos individuos.
- Los individuos que no fueron incluidos en la muestra bootstrap se denominan "Out of Bag" (OOB) y se utilizan para la validación del modelo.

Este proceso de los bosques aleatorios introduce una mayor diversidad entre los árboles y reduce la correlación entre ellos, mejorando la capacidad de generalización del modelo. Breiman [20] formalizó esta idea en el algoritmo de bosques aleatorios, donde en cada división de un árbol, en lugar de considerar todas las p variables predictoras, se selecciona al azar un subconjunto de m variables ($m < p$) como candidatos para la división.

El parámetro principal en los bosques aleatorios es el número de variables m seleccionadas aleatoriamente en cada división. Este valor puede ajustarse utilizando técnicas de validación cruzada. Los valores usuales de m son $\frac{p}{3}$ para la clasificación y $\sqrt{p}$ para problemas de regresión. Otro parámetro es el número de árboles en el bosque, aunque en la práctica esto se maneja más como un problema de convergencia, asegurándose de que el bosque tenga suficientes árboles para estabilizar las predicciones.

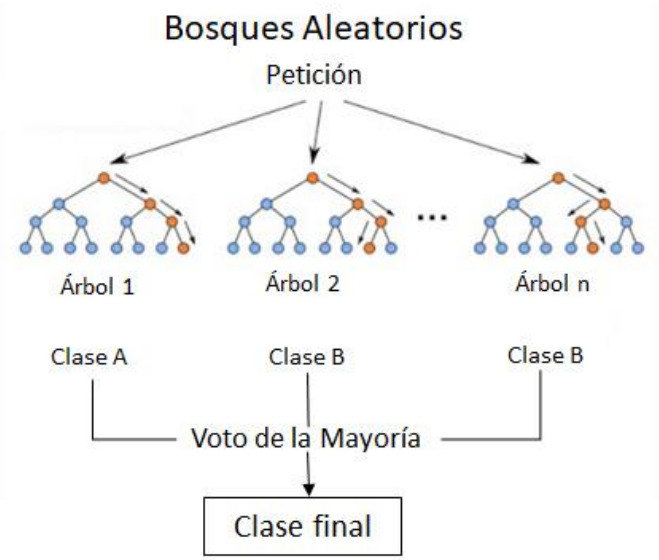


Figura 2.5 Estructura de Bosques Aleatorios

Fuente: Huacasi HYP [21]

La calidad de la predicción para un individuo i se mide mediante el margen de clasificación, definido como:

$$Margin\ (i) = P\ (i\ en\ la\ clase\ real) - \max(p\ (i\ en\ otras\ clases))$$

Por otra parte, la importancia de una variable x_m se calcula siguiendo los pasos:

- Usar la muestra OOB para el árbol k.
- Calcular el margen M_o para muestras OOB.
- Permutar los valores de la variable x_m en la muestra.
- Aplicar el árbol k a la muestra OOB con los valores permutados.
- Calcular el nuevo margen M con la variable x_m permutada.
- Calcular la diferencia entre $M_o\ y\ M$.

La importancia de la variable x_m se define como:

$$I(x_m) = M_o - M$$

Los bosques aleatorios también se pueden emplear para medir la proximidad entre individuos. Para ello, se aplica cada árbol a todos los individuos de la muestra y se determina en qué nodo final se encuentran. La proximidad entre los individuos i y j se mide como la proporción de veces que se encuentran en el mismo nodo final. Esta medida es invariante frente a transformaciones monótonas.

2.2.1.3 Naive Bayes

El clasificador Naive Bayes es uno de los más efectivos modelos de clasificación debido a su simplicidad, resistencia al ruido, poco tiempo para el procesamiento y alto poder predictivo (tasa de acierto) [22]. Este algoritmo de clasificación está basado en el teorema de Bayes, con una suposición fundamental de independencia condicional entre los atributos. Este clasificador es ampliamente utilizado en problemas de clasificación debido a su simplicidad y efectividad, especialmente en escenarios de alta dimensionalidad.

2.2.1.3.1 Teorema de Bayes:

El teorema de Bayes proporciona una forma de calcular la probabilidad posterior de una hipótesis H dado un conjunto de datos D:

$$P(H \mid D) = \frac{P(D \mid H)\, P(H)}{P(D)}$$

donde:

- $P(H \mid D)$ es la probabilidad posterior de la hipótesis H dado el conjunto de datos D
- $P(D \mid H)$ es la probabilidad de observar los datos D dado que la hipótesis H es verdadera.
- $P(H)$ es la probabilidad a priori de la hipótesis H.
- $P(D)$ es la probabilidad de los datos D.

2.2.1.3.2 Clasificador Naive Bayes:

El clasificador Naive Bayes aplica el teorema de Bayes con la asunción de independencia condicional entre los atributos. Dado un conjunto de atributos $X = \{x_1, x_2, \dots, x_n\}$ y una clase C, la probabilidad de C dado X se calcula como [23]:

$$P(C \mid x) = \frac{P(x \mid C)\, P(C)}{P(x)}$$

donde:

- $P(x \mid C)$ se denomina verosimilitud *(likelihood).*
- $P(C \mid x)$ se denomina la probabilidad a posteriori de la clase (posterior).
- $P(C)$ se denomina la probabilidad a priori de la clase.
- $P(x)$ se denomina a la verosimilitud marginal (o probabilidad a priori del vector de atributos).

Se selecciona la clase con la mayor probabilidad a posteriori. Es decir, aquella clase C cuya $P(C|x)$ sea máxima, expresado como:

$$clase(x) = \underset{C \in \{1,\dots,m\}}{\arg\max}\ P(C \mid x) = \underset{C \in \{1,\dots,m\}}{\arg\max} \frac{P(x|C)\, P(C)}{P(x)}$$

Dado que $P(X)$ no depende de C basta con tomar el valor máximo de:

$$clase(x) = \underset{C \in \{1,\dots,m\}}{\arg\max}\ P(C \mid x) \cdot P(C)$$

El clasificador Naive Bayes asume que los atributos son independientes entre sí, una vez sabemos que la instancia pertenece a la clase C. Es decir, si el vector x está compuesto por n atributos $(x_1, x_2, \dots, x_n)$, la probabilidad condicionada $P(x \mid C)$ viene dada por:

$$P(x \mid C) = P(x_1 \mid C) P(x_2 \mid C), \dots, P(x_n \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$$

Tipos de clasificadores Naive Bayes:

Existen varios tipos de clasificadores Naive Bayes, dependiendo de cómo se modela $P(x_i \mid C)$:

- *Gaussiano*: El clasificador Naive Bayes gaussiano asume que los atributos siguen una distribución normal:

  $$x_i \mid C \sim N(\mu_i^c, (\sigma_i^c)^2)$$

  Para un atributo x_i dado una clase C, la probabilidad condicional se modela utilizando la función de densidad gaussiana:

$$P(x_i \mid C) = \frac{1}{\sqrt{2\,\pi}\ \sigma_i^c} \exp\left(-\frac{(x_i - \mu_i^c)^2}{2(\sigma_i^c)^2}\right)$$

donde μ_C y σ_C son la media y la desviación estándar de los atributos de la clase C.

- *Multinomial:* El clasificador Naive Bayes multinomial asume que las características provienen de distribuciones multinomiales [24].

$$P(X_1 = x_1, \ldots, X_d = x_d | Y = C_k) = \frac{n!}{\prod_{i=1}^{d} x_i!} \cdot \prod_{i=1}^{d} p_{ik}^{x_i}$$

  donde $P_{ik} = P(categoría\ i\ |C_k)$

- *Bernoulli:* El clasificador Naive Bayes de Bernoulli es adecuado para atributos binarios (presencia/ausencia, verdadero/falso). La probabilidad condicional se modela como:

$$P(x_i \mid C_k)^{x_i} \cdot [1 - p\,(x_i|C_k)]^{1-x_i} \qquad \begin{matrix} si\ \ x_i = 1 \\ si\ x_i = 0 \end{matrix}$$

2.2.1.4 Redes Neuronales Perceptrón Multicapa (Multilayer Perceptrons, MLPs):

El perceptrón multicapa (MLP, por sus siglas en inglés) es una red neuronal artificial que consiste en múltiples capas de nodos, o neuronas, que procesan las entradas y generan salidas. Esta estructura es una extensión del perceptrón simple y permite a la red aprender y modelar relaciones no lineales complejas. Componentes del perceptrón multicapa [25,26]:

- Capa de entrada (I): Representa los datos de entrada que se alimentan a la red neuronal. Cada entrada x_i corresponde a una característica del conjunto de datos.
- Salida (o): Es el resultado final que produce la red neuronal después de procesar las entradas a través de las capas ocultas y de salida.
- Capas ocultas (h): Consiste en H neuronas ocultas. Las salidas de las neuronas en la capa oculta se calculan como:

$$y_k = \phi_o \left(\alpha_k + \sum_{h=1}^{H} w_{hk} \phi_h \left(\alpha_h + \sum_{i=1}^{I} w_{ih} x_i \right) \right)$$

$$y_k = \phi o \left(\sum_{h=0}^{H} w_{hk} \phi_h \left(\sum_{i=0}^{I} w_{ih} x_i \right) \right)$$

donde:

- w es el peso, coeficiente de sinapsis.
- α es el sesgo (bias).
- Φ es la función de activación.

- Función de activación (ϕ): Es una función matemática aplicada a la salida de cada neurona. Su propósito es permitir que el MLP aprenda y modele relaciones no lineales. Comúnmente, la función de activación en la capa oculta es la función logística (sigmoide):

$$\phi_h(x) = \ell(x) = \frac{\exp(x)}{1 + \exp(x)}$$

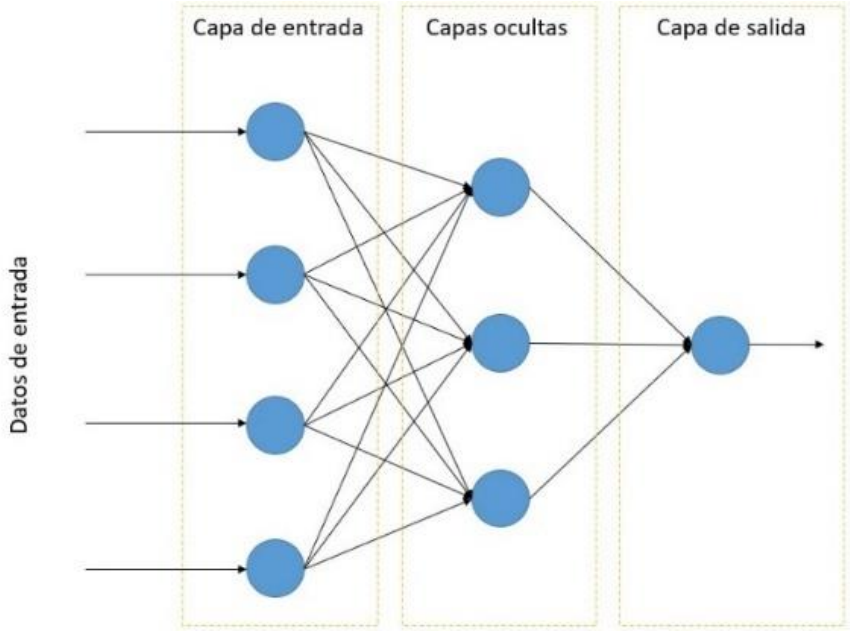


Figura 2.6 Arquitectura de redes neuronales perceptrón multicapa

Fuente: Interactive Chaos [27]

2.2.1.5 Ensamblado de Modelos Stacking

El ensamblado de modelos es una técnica fundamental en el aprendizaje automático que busca combinar múltiples modelos individuales para mejorar la precisión y la capacidad de generalización de las predicciones. Un ejemplo prominente de modelos ensamblados es el de los Bosques Aleatorios detallados en el apartado 2.2.1.2, los cuales combinan múltiples árboles de decisión para formar un modelo robusto. Otra estrategia efectiva de ensamblado es el método Stacking, o apilamiento, que consiste en entrenar un metamodelo capaz de combinar las predicciones de uno o varios modelos base [28].

Como se ilustra en la *Figura 2.7*, el modelo ensamblado Stacking es una técnica que involucra dos niveles de modelos: los modelos base (*first-level learners*) y el metamodelo (*second-level learner*). La idea central es entrenar varios modelos base utilizando el conjunto de datos original y luego entrenar un metamodelo utilizando las predicciones de estos modelos base como características de entrada.

Siguiendo a Zhi-Hua Zhou [29] los modelos stacking involucran los niveles:

- **Modelos Base:** Los modelos base son una serie de clasificadores o regresores que se entrenan sobre el conjunto de datos original $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$. Cada modelo base h_t $(donde\ t \in \{1,2, \dots, T\})$ produce una predicción para cada instancia en el conjunto de datos.

- **Metamodelo**: El metamodelo se entrena utilizando un nuevo conjunto de datos generado a partir de las predicciones de los modelos base. Este conjunto de datos tiene como características las predicciones de los modelos base y como etiquetas las etiquetas originales del conjunto de datos.

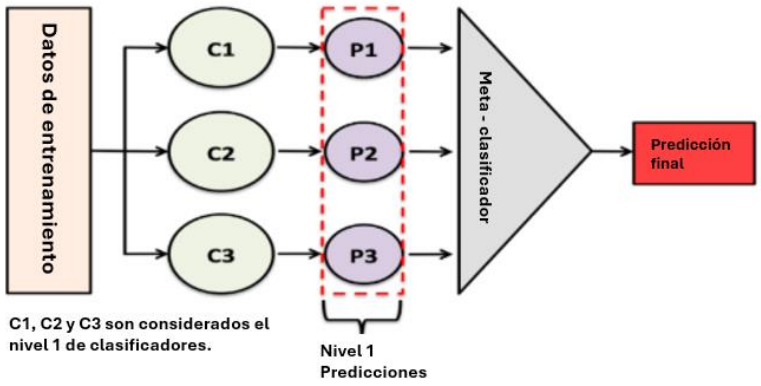


Figura 2.7 Estructura de modelo ensamblado Stacking

Fuente: Ali M, Haider MN, Lashari SA, Sharif W, Khan A, Ramli DA [28]. Imagen adaptada.

2.2.2 Evaluación y rendimiento de los modelos

La evaluación y el rendimiento de los modelos de clasificación son cruciales para determinar la efectividad de un modelo predictivo. Estas métricas son fundamentales para obtener una evaluación precisa y detallada del rendimiento del modelo, independientemente del número de clases o categorías de la variable de respuesta.

2.2.2.1 Matriz de confusión

Una matriz de confusión es una herramienta fundamental para la evaluación de modelos de clasificación. En este estudio, donde la variable de respuesta tiene tres niveles distintos (C_1, C_2, C_3), la matriz de confusión es una matriz 3×3 que resume las predicciones del modelo.

La matriz de confusión para tres clases se define como:

Tabla 2.1. Matriz de confusión

	$\boldsymbol{Predicho\ C_1}$	$\boldsymbol{Predicho\ C_2}$	$\boldsymbol{Predicho\ C_3}$
$\boldsymbol{Real\ C_1}$	TP_1	FP_{12}	FP_{13}
$\boldsymbol{Real\ C_2}$	FN_{21}	TP_2	FP_{23}
$\boldsymbol{Real\ C_3}$	FN_{31}	FN_{32}	TP_3

donde:

- TP_i (Verdadero Positivo): Número de ejemplos correctamente clasificados como C_i.
- FP_{ij} (Falso Positivo): Número de ejemplos de clase C_i incorrectamente como C_j.
- FN_{ij} (Falso Negativo): Número de ejemplos de clase C_j incorrectamente como C_i.

2.2.2.2 Métricas de evaluación

Dado que, este estudio responde a problemas de clasificación se utilizan las siguientes métricas para evaluar el rendimiento de los modelos de aprendizaje automático: $exactitud\ (accuracy), precisión,\ \ sensibilidad\ (recall),\ especificidad$ y $F1 - Score$ [28].

- **Exactitud (Accuracy)**: Mide con qué frecuencia un modelo de aprendizaje automático predice correctamente el resultado.

$$Exactitud = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones}$$

- **Precisión:** Proporción de instancias predichas correctamente como entre todas las instancias predichas como:

$$Precisión = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ Positivos}$$

- **Sensibilidad (*Recall*)**: Proporción de instancias correctamente clasificadas entre todas las instancias que realmente pertenecen a esa clase.

$$Sensibilidad\ (Recall) = \frac{Verdaderos\ positivos}{Verdaderos\ positivos\ + Falsos\ Negativos}$$

- **Especificidad:** proporción de instancias negativas correctamente clasificadas entre todas las instancias que realmente son negativas.

$$Especificidad = \frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos}$$

- **F1 -Score:** Media armónica de la precisión y la sensibilidad:.

$$F1 = 2 \cdot \frac{Precisión \cdot Recall}{Recall + Precisión}$$

Capítulo III

Descripción del problema.

3.1 Planteamiento

El agua subterránea es un recurso vital a nivel global, desempeñando un papel fundamental en el suministro de agua potable, la agricultura y la industria. Según el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), constituye el 97% de los recursos de agua dulce del planeta, excluyendo el hielo polar, y suministra agua potable a al menos 1.500 millones de personas [1]. Además, representa una reserva estratégica protegida contra eventos catastróficos.

El agua subterránea podría ser una fuente alternativa segura de abastecimiento para las poblaciones vulnerables a eventos hidroclimatológicos, aquellas con alta demanda de consumo o en riesgo de contaminación. Sin embargo, existe un bajo nivel de conocimiento, monitoreo, y preparación técnica y académica sobre el tema, esta situación limita el manejo integral del recurso hídrico subterráneo (Ministerio de Ambiente y Desarrollo Sostenible, MADS, 2014) [1].

Según el Banco Mundial, en su documento "The Hidden Wealth of Nations: The Economics of Groundwater in Times of Climate Change", las aguas subterráneas pueden proteger la seguridad alimentaria y fomentar el crecimiento económico y la creación de empleo. No obstante, este recurso ha sido infravalorado y sobreexplotado, sin considerar adecuadamente su sostenibilidad a largo plazo, debido en parte a la falta de investigaciones sistemáticas sobre su importancia económica [30].

Colombia, con una extensión de 1.141.748 km² de territorio continental, está dividida en 32 departamentos, 1.096 municipios, 5 distritos y 20 corregimientos departamentales. La gestión del recurso hídrico en Colombia está bajo la supervisión del Ministerio de Medio Ambiente, la Agencia Nacional de Licencias Ambientales (ANLA), el Servicio Geológico

Colombiano, el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM), las Corporaciones Autónomas Regionales (CAR), las Secretarías municipales, la Policía ambiental, los acueductos municipales y las Juntas de Acción Comunal (JAC). En 2014, el Estudio Nacional de Agua (ENA 2014) permitió reconocer el estado y la dinámica del agua en el país, identificando las cuencas y subzonas hidrográficas prioritarias para mejorar la gestión del recurso hídrico. Se determinó que el 62% de la precipitación pluvial se convierte en recurso hídrico superficial y se identificaron 61 sistemas acuíferos con una oferta hídrica potencial de agua subterránea de 5.848 km³, distribuidos en 16 provincias hidrogeológicas [31].

Figura 3.1 Mapa hidrográfico de Colombia 2002

Fuente: IGAC - Instituto Geográfico Agustin Codazzi [31]

Boyacá, situado en el centro de Colombia, es una región caracterizada por una compleja geología que incluye formaciones sedimentarias, volcánicas e intrusivas. Esta diversidad

geológica influye directamente en la hidrogeología de la región, determinando las características de los acuíferos y la disponibilidad de agua subterránea. Según la Corporación Autónoma Regional de Boyacá, esta región ha enfrentado retos de racionamiento de agua, que se han manejado con el uso integrado del agua superficial y subterránea, garantizando un suministro de agua potable durante todo el año [33].

En el marco del proyecto de modelación hidrogeológica de la zona centro del Departamento de Boyacá, se han identificado y evaluado diversas unidades geológicas con el objetivo de caracterizar el recurso hídrico subterráneo disponible. Según el informe del proyecto, se han inventariado puntos de agua como pozos, aljibes y manantiales, recolectando datos fisicoquímicos y evaluando las condiciones de captación y uso (Proyecto Modelo Hidrogeológico de la Zona Centro del Departamento de Boyacá, 2020) [34].

La *Figura 3.2* ilustra las tres principales fuentes de captación de agua en el municipio de Boyacá: el manantial de la Vereda Escalones, el aljibe de la Vereda Concepción en Cómbita y el pozo profundo de la Vereda Runta en Tunja. Estas fuentes representan los tipos de captación evaluados en el estudio para caracterizar el recurso hídrico subterráneo en la región.

Figura 3.2 Diversas fuentes de captación de agua en Boyacá, Colombia

Fuente: CORPOBOYACÁ [35]

La caracterización hidrogeoquímica del agua subterránea es crucial para su gestión sostenible. En este contexto, surge la necesidad de explorar cómo los modelos de aprendizaje automático pueden contribuir a esta caracterización en Boyacá, Colombia.

Específicamente, se requiere evaluar cómo estos modelos pueden clasificar de manera precisa los tipos de captación (Aljibes, Manantiales y Pozos) y determinar el rendimiento comparativo entre diferentes modelos de clasificación en este contexto. La *Figura 3.3* ilustra los puntos de captación donde se inventariaron o tomaron las muestras de agua en el departamento de Boyacá para el presente estudio.

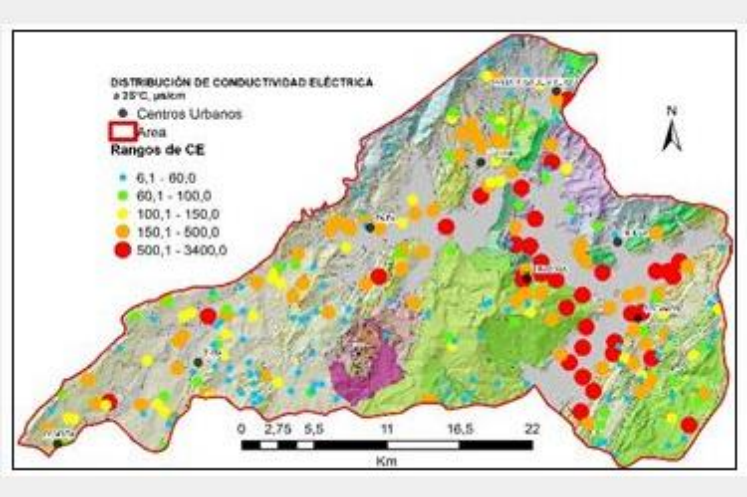


Figura 3.3 Ubicaciones de puntos de captación de muestras del estudio

Fuente: Servicio Geológico Colombiano [34]

La pregunta central que guía esta investigación es: ¿Cuál es el rendimiento de diferentes modelos de aprendizaje automático en la clasificación hidrogeoquímica del agua subterránea según su fuente de captación (Aljibe, Manantial o Pozo)?

Capítulo IV

Análisis y Diseño

En este capítulo se presenta la base de datos utilizada en la investigación. Se lleva a cabo un análisis exploratorio unidimensional y multivariante con el objetivo de comprender los datos, realizar transformaciones cuando sea necesario y explorar las relaciones entre las variables. Posteriormente, se procede con el diseño de los modelos de clasificación.

4.1 Conjunto de datos

El conjunto de datos utilizado en esta investigación proviene de la plataforma de datos públicos de Colombia, datosabiertos.gov.co, específicamente del conjunto titulado "Hidrogeoquímica zona centro Boyacá" [8]. Este conjunto de datos se desarrolló en el marco del proyecto del modelo hidrogeológico de la zona centro del Departamento de Boyacá e incluye información detallada sobre el control geológico para hidrogeología, el inventario de puntos de agua (pozos, aljibes y manantiales), datos físico-químicos (pH, conductividad eléctrica, temperatura), propiedades organolépticas (turbiedad, olor y color), condiciones de captación y uso, prospección geoeléctrica, evaluación hidroclimatológica e hidrogeoquímica y perforaciones de pozos exploratorios.

El conjunto de datos consta de 295 observaciones y 29 variables, cada una representando diferentes parámetros hidrogeoquímicos y características de las fuentes de agua subterránea. Las concentraciones de las variables analizadas están expresadas en diferentes unidades de medidas. La estructura del conjunto se presenta en la *Tabla 4.1*.

Tabla 4.1 Descripción de variables del conjunto de datos del estudio

Variable	Descripción	Unidad de medida
fid	Identificadores únicos para cada observación	No aplica
objectid	Identificadores únicos para cada observación	No aplica
no	Identificadores únicos para cada observación	No aplica
identifica	Código de identificación de la muestra	No aplica
x	Coordenadas Geográficas	No aplica
y	Coordenadas Geográficas	No aplica
z	Altitud	Metros sobre el nivel del mar (msnm)
tipo_de_ca	Tipo de captación: 1 = Aljibe 2 = Manantial 3 = Pozo	No aplica
p_h	Ph del agua	
t	Temperatura	Grados Celsius (°C)
ce_25c	Conductividad eléctica a 25ºC	Microsiemens por centímetro (µS/cm)
alcalinidad	Alcalinidad	Miligramos por litro (mg/L)
color	Color	Unidades de color verdadero (TCU)
st	Sólidos totales	Miligramos por litro (mg/L)
sdt	Sólidos Disueltos Totales	Miligramos por litro (mg/L)
hco3	Bicarbonato	Miligramos por litro (mg/L)
cl	Cloruro	Miligramos por litro (mg/L)
po4	Fosfato	Miligramos por litro (mg/L)
k	Potasio	Miligramos por litro (mg/L)
fe_t	Hierro Total	Miligramos por litro (mg/L)
ca	Calcio	Miligramos por litro (mg/L)
so4	Sulfato	Miligramos por litro (mg/L)

no3	Nitrato	Miligramos por litro (mg/L)
mg	Magnesio	Miligramos por litro (mg/L)
na	Sodio	Miligramos por litro (mg/L)
turbiedad	Turbidez	Unidades Nefelométricas de Turbidez (UNT)
dureza	Dureza	Miligramos por litro (mg/L)
mn	Manganeso	Miligramos por litro (mg/L)

4.2 Análisis exploratorio y preprocesamiento de datos

Se lleva a cabo un análisis exploratorio de datos utilizando el software estadístico RStudio. La base de datos objeto de estudio, como se describe en el apartado 4.1, consta de 295 observaciones distribuidas en 29 variables. El proceso inicial incluye el preprocesamiento y la transformación de las variables para facilitar su manipulación y análisis. Se realiza una selección manual de las variables relevantes para la investigación, depurando la base de datos mediante la exclusión de identificadores únicos para cada observación, códigos de muestra, así como coordenadas geográficas y altitud.

Tabla 4.2 *Estructura del conjunto de datos (variable cualitativa)*

Variable cualitativa	**Número de observaciones por cada tipo**
tipo	Aljibe: 105 Manantial: 156 Pozo: 31

Tabla 4.3 *Estructura del conjunto de datos (variables numéricas)*

Variables numéricas	**Mínimo**	**1er. Cuartil**	**Media**	**Desv. Típica**	**Mediana**	**3er Cuartil**	**Máximo**
pH	4.60	5.90	6.41	0.70	6.40	7.0	8.50
t	10.40	15.50	16.84	2.25	16.80	18.10	26.70
ce_25c	6.09	44.90	274.85	431.07	95.30	248.50	3400
alcalinidad	0.30	6.50	64.13	106.75	20.50	72.50	945.00
color	0.00	2.0	22.61	42.90	7.0	22.0	380.00

st	13.0	58.0	261.3	701.52	107.0	210.5	10729
sdt	6.00	32.75	172.98	288.11	76.0	164.0	2078
hco3	0.00	7.90	77.63	130.19	25.0	86.25	1153
cl	0.30	1.990	22.16	49.30	5.0	21.95	506
po4	-0.10	0.10	0.14	0.24	0.10	0.10	2.8
k	0.00	0.70	4.01	10.46	1.45	3.52	138
fe_t	0.00	0.10	0.90	4.61	0.10	0.20	55.8
ca	0.20	2.80	24.99	42.50	7.85	28.70	366
so4	0.00	1.40	38.62	105.36	4.45	20.20	1040
no3	0.00	0.20	5.00	12.27	0.95	4.30	105
mg	0.10	0.90	5.75	11.26	2.05	5.35	92.60
na	0.10	1.57	13.37	54.16	3.80	13.37	508
turbiedad	0.30	1.50	39.40	190.33	4.85	20.25	2302
dureza	0.90	12.47	86.34	137.52	32.0	101.75	1022
mn	0.00	0.05	0.21	0.53	0.05	0.077	4

Como se puede apreciar en la *Tabla 4.2 y 4.3*, la base de datos seleccionada para el estudio contiene únicamente variables numéricas relevantes para el modelado, con la excepción de la variable respuesta denominada "tipo", que indica los tipos de captación de agua.

4.2.1 Análisis univariante:

4.2.1.1 Variable categórica

Inicialmente, se analiza la variable tipo de captación, representada en la *Figura 4.1*, donde se observa que esta variable respuesta presenta cinco niveles, con un 0.3% de registros correspondientes a agua superficial y un 0.7% sin etiqueta. Estos valores resultan no significativos para los objetivos del estudio, por lo que se procede a eliminarlos. De esta manera, se filtra la base de datos para trabajar únicamente con los tipos de captación de agua principales: aljibe, manantial y pozo.

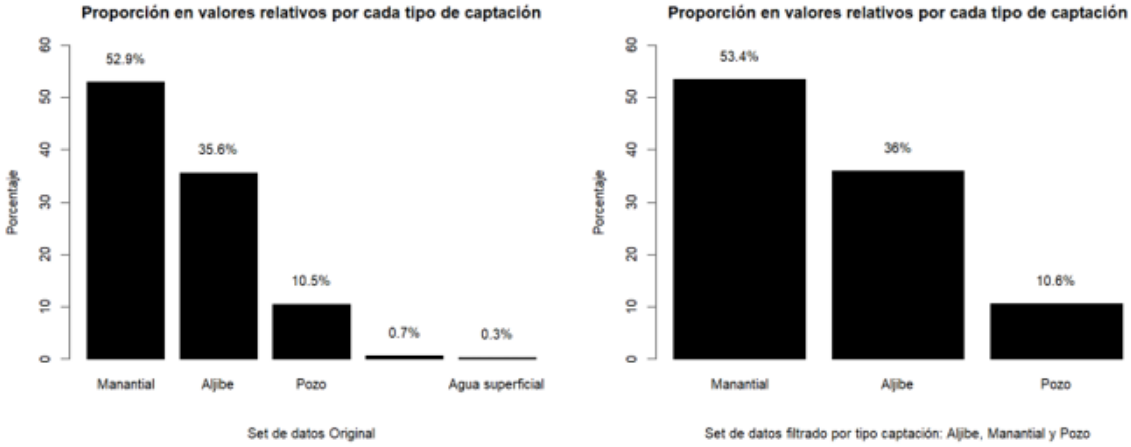


Figura 4.1 Proporción relativa de los tipos de captación de agua subterránea

Descripción. Esta figura muestra la proporción de registros para cada tipo de captación del conjunto de datos original. Se presenta también la proporción después de filtrar los tipos de captación específicos para este estudio, que incluyen únicamente Manantial, Pozo y Aljibe.

Cabe resaltar que los datos no están balanceados para los tres tipos de captación, ya que, en valores absolutos, se tienen 156 registros para el tipo de captación Manantial, 105 para Aljibe y 31 para Pozo. La cantidad de registros para Pozo es notablemente inferior en comparación con las otras categorías. Por este motivo, en la sección 4.3 se detalla el proceso de manejo del desequilibrio de clases mediante *oversampling*, específicamente utilizando la técnica SMOTE (Synthetic Minority Over-sampling Technique).

4.2.1.1 Variables numéricas

En el análisis univariante de las variables numéricas, la Figura 4.3 presenta los histogramas correspondientes. Estos gráficos permiten explorar las distribuciones de las propiedades del agua estudiadas.

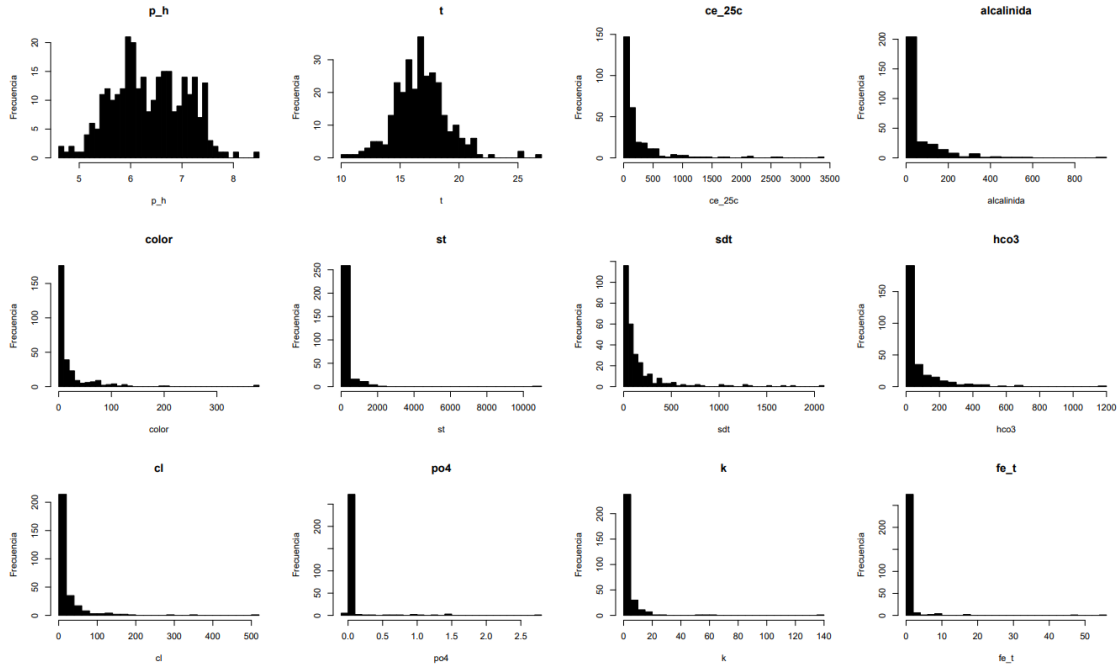


Figura 4.2 Histogramas variables numéricas

Mediante el análisis gráfico de la *Figura 4.2* y numérico (ver anexo 1), se observa que la distribución del pH presenta una media de 6.41 y una mediana de 6.4, lo que indica una ligera asimetría hacia valores ligeramente ácidos. Los datos oscilan entre 4.60 y 8.50, con una desviación estándar de 0.91, sugiriendo una dispersión moderada alrededor de la media.

Por otro lado, la temperatura promedio es de 16.84°C, con una mediana de 16.80°C. Los valores varían considerablemente entre 10.40°C y 26.70°C, mostrando una variabilidad notable. La distribución parece simétrica, con una desviación estándar de 2.4, lo que indica una dispersión moderada.

Es importante destacar que, como se observa en la *Figura 4.2* las variables numéricas (excepto el pH y la temperatura) muestran distribuciones sesgadas hacia la derecha. Esta característica indica una concentración de valores más bajos, acompañada de algunos valores extremadamente altos que pueden influir en la media y en la dispersión de los datos.

Con base en esta observación, se han eliminado los registros que contenían los valores máximos de las variables alcalinidad, color, sólidos totales, bicarbonato y turbiedad, ya que, como se evidencia en la *Figura 4.2*, estos valores extremos se desvían significativamente del resto de los registros. Tras esta eliminación, la base de datos queda con 287 observaciones y 21 variables.

A pesar de la eliminación de estos registros, se ha decidido mantener los valores atípicos restantes debido a la naturaleza de las variables con distribución sesgada a la derecha y colas largas. Esto sugiere que los valores atípicos podrían representar variaciones reales y significativas en las concentraciones de las propiedades del agua.

4.2.2 Análisis multivariante:

En esta sección se aborda el análisis multivariante, comenzando con la estandarización de las variables numéricas para normalizar sus escalas. A continuación, se realizan las pruebas de contraste de hipótesis no paramétricos para identificar las variables con diferencias significativas. Finalmente, se lleva a cabo el análisis bivariante y de correlación, evaluando las relaciones entre pares de variables tanto globalmente como dentro de cada grupo definido por la variable categórica.

4.2.2.1 Estandarización de variables

La estandarización de una variable X implica transformar sus valores para que la nueva variable Z tenga una media (μ) igual a 0 y una desviación estándar (σ) igual a 1. La fórmula para estandarizar cada variable:

$$Z = \frac{X - \mu_x}{\sigma_x}$$

donde:

- Z es el valor estandarizado.
- X es el valor original de la variable.
- μ_x es la media de la variable X.

- σ_x es la desviación estándar de la variable X.

4.2.1.1 Contraste de hipótesis y test no paramétricos

Una vez estandarizadas las variables, se evalúa la normalidad de las distribuciones utilizando la prueba de Shapiro-Wilk. Esta prueba contrasta la hipótesis nula de que una muestra proviene de una población normalmente distribuida H_0: $X \sim N(\mu, \sigma^2)$ o que, en caso contrario sigue otra distribución H_1 : $\nsim N(\mu, \sigma^2)$. La prueba de Shapiro-Wilk se basa en la correlación entre los datos y los valores esperados correspondientes en una distribución normal. Un valor de $p < 0.05$ indica que podemos rechazar la hipótesis nula y, por lo tanto, los datos no siguen una distribución normal.

Tabla 4.4 Resultados de la prueba Shapiro-Wilk de normalidad en todo el conjunto de datos

Variable	**Shapiro-Wilk p- valor**
pH	0.01174
t	0.00002733
ce_25c	< 2.2e-16
alcalinidad	< 2.2e-16
color	< 2.2e-16
st	< 2.2e-16
sdt	< 2.2e-16
hco3	< 2.2e-16
cl	< 2.2e-16
po4	< 2.2e-16
k	< 2.2e-16
fe_t	< 2.2e-16
ca	< 2.2e-16
so4	< 2.2e-16
no3	< 2.2e-16
mg	< 2.2e-16
na	< 2.2e-16
turbiedad	< 2.2e-16
dureza	< 2.2e-16
mn	< 2.2e-16

Como se observa en la *Tabla 4.4*, los resultados de la prueba de Shapiro-Wilk evidencian que todas las variables presentan valores de *p* inferiores a 0.05, por lo que no se tiene

evidencia significativa para aceptar la hipótesis nula de que las variables siguen una distribución normal.

Se realiza la prueba Shapiro-Wilk para cada variable dentro de los grupos (Aljibe, Manantial, Pozo). Los resultados presentados en la *Tabla 4.5*, sugieren que, para la mayoría de las variables los *p-valores* son menores a 0.05, lo que sugiere que no se tiene evidencia significativa para aceptar la hipótesis nula de normalidad en las variables. Por lo tanto, se concluye que la mayoría de las variables no siguen una distribución normal dentro de cada grupo, con excepción de las variables pH en pozos y temperatura en aljibes y pozos, las cuales registran *p-valores* mayores a 0.05.

Tabla 4.5 *Resultado de prueba Shapiro-Wilk para cada variable dentro de los grupos*

Variable	Aljibe	Manantial	Pozo
pH	0.02247	0.0167	0.2
t	0.5465	0.00001902	0.3405
ce_25c	1.16e-14	< 2.2e-16	0.00000592
alcalinidad	1,84E-07	< 2.2e-16	0.000002366
color	1,42E-12	< 2.2e-16	0.000000001829
st	7,34E-11	1.23e-15	0.000005779
sdt	3,59E-11	9,34E-11	0.000001408
hco3	1.81e-10	< 2.2e-16	0.000002361
cl	< 2.2e-16	< 2.2e-16	0.0000002722
po4	< 2.2e-16	< 2.2e-16	8,07E-09
k	< 2.2e-16	< 2.2e-16	0.000001974
fe_t	< 2.2e-16	< 2.2e-16	0.000000004779
ca	1,35E-10	< 2.2e-16	0.000002581
so4	< 2.2e-16	< 2.2e-16	0.00000001755
no3	< 2.2e-16	< 2.2e-16	1,15E-07
mg	7,51E-13	< 2.2e-16	0.0000004485

na	< 2.2e-16	< 2.2e-16	0.00000006967
turbiedad	< 2.2e-16	< 2.2e-16	0.000000003523
dureza	1,01E-09	< 2.2e-16	0.000005396
mn	< 2.2e-16	< 2.2e-16	0.00001529

Dado que, no se cumplen los supuestos de normalidad, se aplica la prueba no paramétrica de Kruskal-Wallis. Este es un método que se utiliza para probar la hipótesis de igualdad de distribución de una variable en k grupos, siendo obligatorio únicamente que las observaciones sean independientes.

La prueba de Kruskal-Wallis contrasta la hipótesis nula de que todas las muestras provienen de la misma población (o de poblaciones con la misma mediana). Esta prueba se basa en los rangos de los datos en lugar de sus valores originales. Un valor de $p < 0.05$ indica que al menos uno de los grupos difiere significativamente de los otros.

Tabla 4.6 Resultados de la prueba no paramétrica Kruskal Wallis

Variable	**Kruskal Wallis p-valor**	**Significación**
pH	0.00000447	Significativa
t	1.302e-10	Significativa
ce_25c	1.214e-14	Significativa
alcalinidad	1.794e-13	Significativa
color	0.003077	Significativa
st	4.46e-13	Significativa
sdt	2.64e-15	Significativa
hco3	5.965e-14	Significativa
cl	0.000000002587	Significativa
po4	0.7278	No significativa
k	0.00001789	Significativa
fe_t	4.134e-10	Significativa
ca	< 2.2e-16	Significativa
so4	5.706e-13	Significativa
no3	0.009921	Significativa
mg	6.238e-12	Significativa
na	4.677e-12	Significativa

turbiedad	0.453	No significativa
dureza	< 2.2e-16	Significativa
mn	0.00000007614	Significativa

En la *Tabla 4.6* se presentan los resultados obtenidos de la prueba no paramétrica de Kruskal-Wallis, la cual es utilizada para determinar si existen diferencias significativas en las distribuciones de varias muestras independientes. Los resultados muestran que las variables pH, temperatura, conductividad eléctrica a 25ºC, alcalinidad, color, sólidos totales, sólidos disueltos totales, bicarbonato, cloruro, potasio, hierro total, calcio, sulfato, nitrato, magnesio, sodio, dureza y manganeso son estadísticamente significativas al nivel del 5% de significancia. Esto implica que existe suficiente evidencia estadística para rechazar la hipótesis nula, sugiriendo que estas variables presentan diferencias significativas en sus medianas entre los grupos comparados. Por otro lado, las variables fosfato y turbiedad no son significativas, indicando que no hay evidencia suficiente para afirmar que sus distribuciones difieren entre los grupos estudiados.

Por lo anterior, se continúa el estudio únicamente con las variables significativas estadísticamente mediante la prueba de Kruskal Wallis, ya que estas variables han demostrado tener diferencias entre los grupos comparados.

4.2.1.2 Análisis bivariante

Los resultados obtenidos a partir del análisis bivariante y gráfico presentado en la *Figura 4.3*, se destacan varias observaciones significativas sobre los componentes hidrogeoquímicos en función de los tres tipos de captación: aljibe, manantial y pozo.

En primer lugar, se observa que la variable hierro total presenta concentraciones más elevadas cuando el agua proviene de pozos en comparación con manantial y aljibe. Este fenómeno es notable y tiene sentido si consideramos la metodología de extracción o punto de captación del agua. Los pozos están diseñados para extraer agua de capas geológicas más profundas, las cuales están en contacto con formaciones ricas en hierro, lo que puede resultar en mayores concentraciones de este elemento. En contraste, los manantiales son corrientes de agua subterránea que emergen naturalmente a la superficie, mientras que los

aljibes son estructuras artesanales que alcanzan el nivel del agua y se profundizan ligeramente por debajo de este, ubicándose en lugares donde el acuífero se encuentra cerca de la superficie [35], por lo que su profundidad es menor en comparación con los pozos profundos.

En cuanto al bicarbonato, se evidencia que tanto los pozos como los aljibes presentan valores más elevados de bicarbonato en comparación con el agua proveniente de manantial. Este hallazgo es relevante, ya que el bicarbonato actúa como un amortiguador químico que ayuda a mantener el pH del agua dentro de un rango adecuado. El bicarbonato neutraliza los ácidos presentes, haciendo que el agua sea más alcalina. Esto se observa en la *Figura 4.3*, donde los puntos de captación de pozo y aljibe registran una media de pH más alto, sugiriendo que el agua de estos puntos es menos ácida que el agua de manantial. Un pH más alcalino generalmente indica que el agua es menos corrosiva y sugiere la presencia de formaciones geológicas ricas en minerales como calcio, magnesio y bicarbonato. Como se ilustra en la *Figura 4.3*, estos compuestos presentan medias mayores en los puntos de captación de pozo y aljibe en comparación con los de manantial, lo que apoya la relación entre la alcalinidad y la composición mineral del agua.

Por último, se destaca que el color del agua de manantial registra valores medios mayores en comparación con el color del agua de pozos y aljibes.

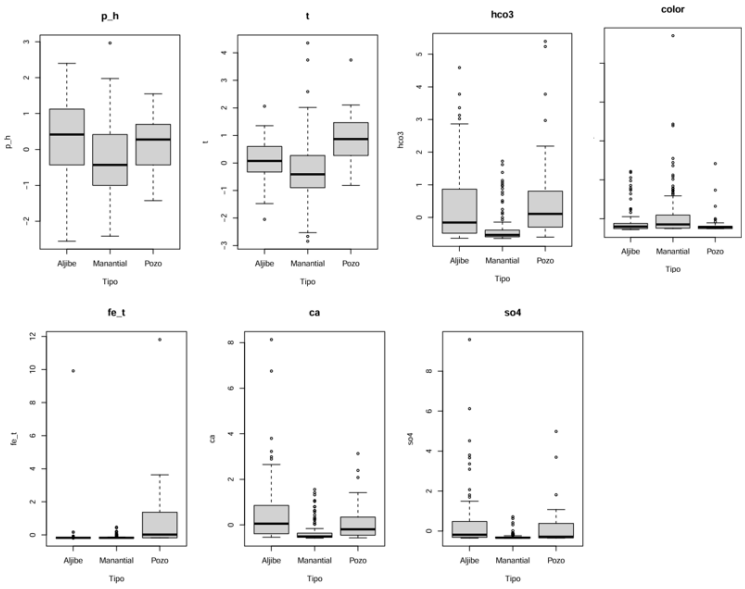


Figura 4.3 Comparación bivariante de parámetros hidrogeoquímicos en aguas subterráneas según el tipo de captación (Aljibe, Manantial y Pozo) en Boyacá, Colombia

4.2.1.3 Correlación

El análisis de correlación realizado sobre las variables hidrogeoquímicas del agua subterránea en Boyacá, Colombia, revela interrelaciones significativas entre diversas variables. Cabe resaltar que un valor de correlación cercano a 1 indica una correlación positiva fuerte entre dos variables, lo que significa que cuando una variable aumenta, la otra también tiende a aumentar, por el contrario, un valor de correlación cercano a -1 indica una correlación negativa fuerte entre dos variables, lo que significa que cuando una variable aumenta, la otra tiende a disminuir.

En la *Figura 4.4* se presentan los principales hallazgos del análisis de correlación de las variables hidrogeoquímicas estandarizadas de las aguas subterráneas en Boyacá [ver anexo 3]:

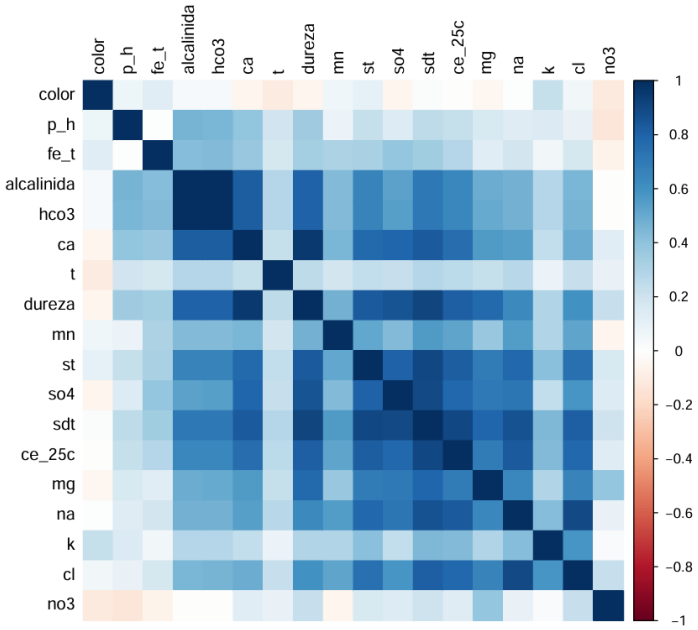


Figura 4.4 Correlación de variables numéricas

El análisis de correlación sugiere que la conductividad eléctrica a 25°C muestra una correlación alta y positiva con varios componentes. Específicamente, la conductividad eléctrica tiene correlaciones de 0.90 con sólidos disueltos totales, 0.82 con sólidos totales, 0.75 con calcio y 0.80 con dureza. Estas relaciones indican una interdependencia significativa entre la conductividad y la presencia de estos componentes en el agua subterránea de la región.

Además, se observa una correlación extremadamente alta de 0.996 entre la alcalinidad y la concentración de bicarbonato, lo cual es consistente, ya que el bicarbonato es un componente principal de la alcalinidad en aguas naturales. También se observa una alta correlación positiva entre la alcalinidad del agua, el calcio, dureza, y sólidos totales disueltos de 0.82, 0.80 y 0.71 respectivamente. Esto sugiere que, a medida que aumenta la alcalinidad del agua, también tienden a aumentar las concentraciones de calcio, dureza y sólidos totales disueltos en el agua.

Por otro lado, el color del agua muestra una correlación baja con la mayoría de las variables analizadas. Cabe destacar la existencia de una débil correlación negativa con la

temperatura del agua, con un coeficiente de -0.1. Este hallazgo sugiere que, a medida que la temperatura del agua aumenta, la intensidad del color tiende a disminuir, y viceversa.

4.2.1.3.1 Correlación por tipo de captación

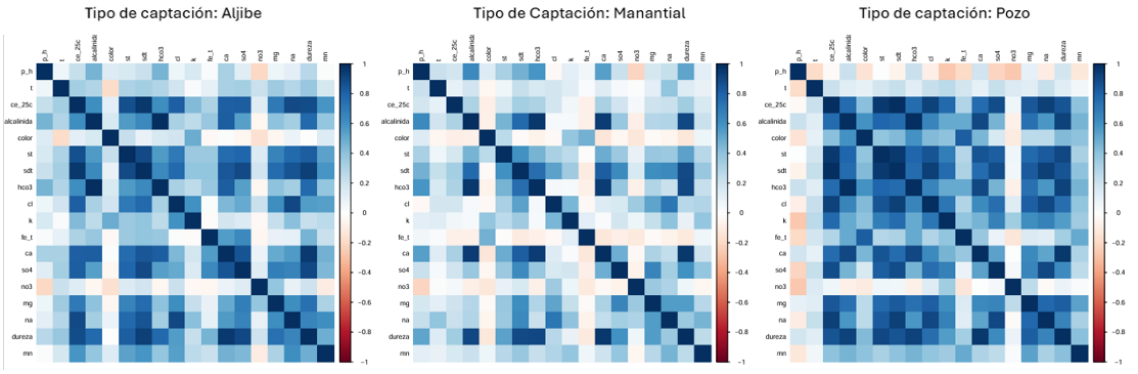


Figura 4.5 Correlación de variables estandarizadas, por tipo de captación de agua: Aljibe, Manantial y Pozo

Al comparar las correlaciones entre los componentes hidrogeoquímicos de las tres fuentes de captación de agua, se observan algunas diferencias notables. Por ejemplo, el pH presenta una correlación positiva débil con los componentes potasio, hierro, sulfato y la temperatura en aljibes, con valores de 0.16, 0.023, 0.12 y 0.06, respectivamente. De manera similar, en manantiales, estas correlaciones son de 0.11, 0.51, 0.47 y 0.21. Sin embargo, en el agua captada en pozos, las correlaciones con el pH son negativas y débiles, con valores de -0.26, -0.19, -0.21 y -0.18.

Estas correlaciones sugieren que, en los manantiales, un aumento en la concentración de hierro tiende a estar asociado con un aumento en el pH del agua. Esto indica que, en estas fuentes, a medida que el agua se vuelve más alcalina, la concentración de hierro tiende a ser mayor. Por el contrario, en los pozos, un aumento en el pH está asociado con una disminución en la concentración de hierro. Es decir, en las aguas de pozo, a medida que el agua se vuelve más alcalina, la concentración de hierro tiende a disminuir.

Otro caso notable es la correlación entre hierro y calcio, donde se observa que, en el agua captada en manantiales, esta correlación es negativa con un valor de -0.12. En contraste,

los aljibes muestran una correlación positiva de 0.55, y los pozos también presentan una correlación positiva de 0.37.

Finalmente, se observa una relación significativa y positiva entre la dureza del agua y la alcalinidad en los tres tipos de captación analizados. En el caso de los pozos, el coeficiente de correlación es de 0.92, lo que indica una fuerte asociación entre estas dos variables. En los manantiales, la correlación es de 0.90, y en los aljibes, aunque ligeramente menor, sigue siendo notable con un valor de 0.77. Estos resultados sugieren que, independientemente del tipo de captación, existe una tendencia consistente en que a medida que aumenta la dureza del agua, también lo hace la alcalinidad.

4.3 Manejo del Desbalance de Datos:

Como se señaló en el apartado 4.2.1.1, en el conjunto de datos se registra desbalance en la variable respuesta “tipo de captación”, con 156 registros para manantial, 105 para aljibe y solo 31 para pozo. Esta disparidad entre clases puede sesgar los modelos de aprendizaje automático hacia las clases mayoritarias, comprometiendo la precisión y generalización de las predicciones [37,38].

Para abordar este problema, se aplica la técnica de sobre muestreo (oversampling) utilizada en problemas de clasificación desbalanceada. En este caso, se utiliza el método SMOTE (Synthetic Minority Over-sampling Technique), el cual introduce ejemplos sintéticos nuevos en lugar de duplicar observaciones existentes. SMOTE genera estos ejemplos creando puntos de datos nuevos que están en el espacio entre observaciones de la clase minoritaria.

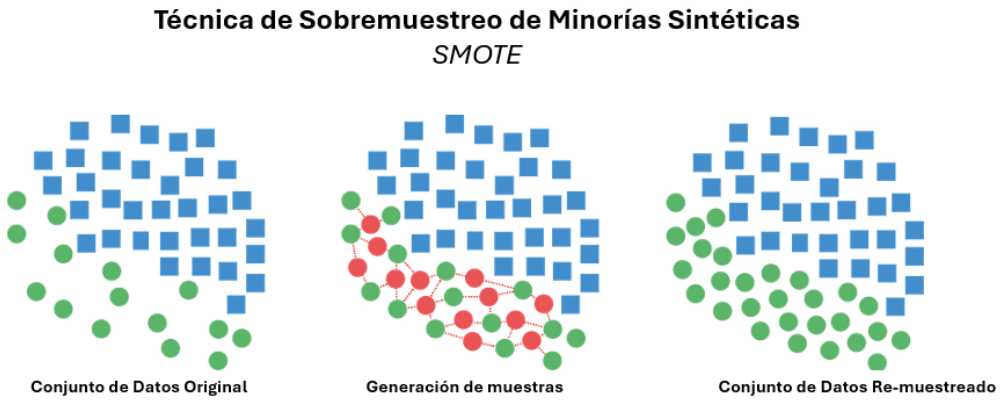


Figura 4.6 Manejo desbalance de datos, método SMOTE

Fuente: Dholakiya P. [39]. Imagen Adaptada.

El proceso que se lleva a cabo es:

- Se selecciona una observación aleatoria x_i de la clase minoritaria.
- Se encuentran los k vecinos más cercanos $x_{i1}, x_{i2}, \dots, x_{ik}$ de x_i dentro de la clase minoritaria utilizando una medida de distancia como por ejemplo la distancia Euclidiana.
- Finalmente se generan observaciones sintéticas, donde para cada vecino cercano x_{ik} (donde $k \in \{1,2, \dots, K\}$), se genera una nueva observación sintética x', usando la fórmula:

$$x' = x_i + \delta * (x_{ik} - x_i)$$

donde:

- δ es un valor aleatorio en el intervalo [0,1]. Este proceso genera una observación que está en la línea entre x_{ik} y x_i.

La implementación llevada a cabo en el presente estudio, inicia con la recodificación de la variable "tipo" a valores numéricos (1=Aljibe, 2=Manantial y 3= Pozo) para evitar posibles advertencias del software Rstudio al aplicar la técnica de manejo de desbalance de datos. Luego, se aplica el algoritmo SMOTE donde para cada punto x_i del tipo de

captación pozo, y para cada uno de sus $k = 5$ vecinos más cercanos x_{ik}, se generan $dup\_size = 2$ ejemplos sintéticos.

Estos parámetros se seleccionan para controlar el número de vecinos considerados en la generación de muestras sintéticas (k) y la cantidad de muestras sintéticas creadas por cada muestra minoritaria (dup_size), hasta alcanzar el tamaño deseado, aumentando así su representación en el conjunto de datos.

En la *Tabla 4.7* se muestran los valores absolutos de cada tipo de captación antes y después del proceso de manejo de desbalance de la clase minoritaria “Pozo”. Se observa que inicialmente había 31 registros para la variable "Pozo", y después del proceso de sobre muestreo se cuenta con 93 registros, mientras que las observaciones de las otras variables (Aljibe y Manantial) se mantuvieron sin cambios.

Tabla 4.7 *Número de observaciones antes y después de aplicar técnica de desbalance para la clase minoritaria*

	Aljibe	**Manantial**	**Pozo**
Número de registros sin aplicar técnica SMOTE	105	156	31
Número de registros aplicando técnica SMOTE	105	156	93

4.4 Diseño y análisis de modelos de clasificación

En este apartado, se aborda el diseño y análisis de modelos de clasificación multiclase. Este estudio se centra específicamente en la comparación del rendimiento de los siguientes modelos: Árbol de Clasificación, Random Forest, Naive Bayes, Red Neuronal Artificial Perceptrón Multicapa, así como un modelo ensamblado Stacking que combina como modelo base Bosques Aleatorios y como metamodelo Naive Bayes.

4.4.1 Conjunto de entrenamiento y validación

Inicialmente, el conjunto de datos se divide en sets de entrenamiento y validación. El set de entrenamiento está compuesto por el 70% de los registros de la base de datos, utilizado para ajustar los parámetros del modelo. El 30% restante forma el set de validación, destinado a evaluar el desempeño y la capacidad de generalización del modelo.

4.4.2 Árbol de Clasificación

4.4.2.1 Diseño de árbol de clasificación

Para el Árbol de Clasificación, el modelo se entrena con la variable respuesta tipo, en función de todas las variables predictoras significativas. El modelo se configura con los siguientes parámetros:

- **Número mínimo de observaciones:** un nodo N debe contener al menos 15 observaciones antes de que se considere la partición $|N| \geq 15$. Este ajuste previene la sobre fragmentación del árbol, asegurando que las divisiones se realicen solo cuando haya suficiente información para justificar una partición adicional.
- **Profundidad máxima del árbol:** Se establece una profundidad máxima $d(N) \leq 8$. Está profundidad máxima limita el número de niveles que el árbol puede tener, controlando así la complejidad del modelo.
- **Coste computacional:** Este parámetro controla la complejidad del árbol mediante la configuración de un valor de cp (complexity parameter). Un cp de 0.01 indica que una partición de cualquier nodo solo se lleva a cabo si disminuye la falta de ajuste (cost-complexity) en al menos un 1%.

La estructura del Árbol de Clasificación con estos parámetros se diseña para lograr un equilibrio entre la precisión del modelo y su capacidad de generalización. El proceso de división se repite para cada nodo hijo hasta que se cumpla alguna de las condiciones de parada: el nodo contiene menos de 15 observaciones o la profundidad del árbol alcanza 8 niveles. La configuración específica busca maximizar la pureza de las clases en cada nodo terminal.

4.4.2.2 Análisis de árbol de clasificación

En la *Figura 4.7*, se observa la arquitectura del árbol de clasificación para las características hidrogeoquímicas. El árbol de clasificación comenzó con 244 observaciones en el nodo raíz. En la primera división del nodo raíz, se crean dos ramas principales: la rama izquierda, con 211 observaciones, clasifica predominantemente como tipo de captación manantial, y la rama derecha, con 33 observaciones, clasifica principalmente como tipo de captación pozo. En el nodo 2, se realiza una nueva división basada en la variable calcio resultando en dos nodos secundarios: el nodo 4 y el nodo 5.

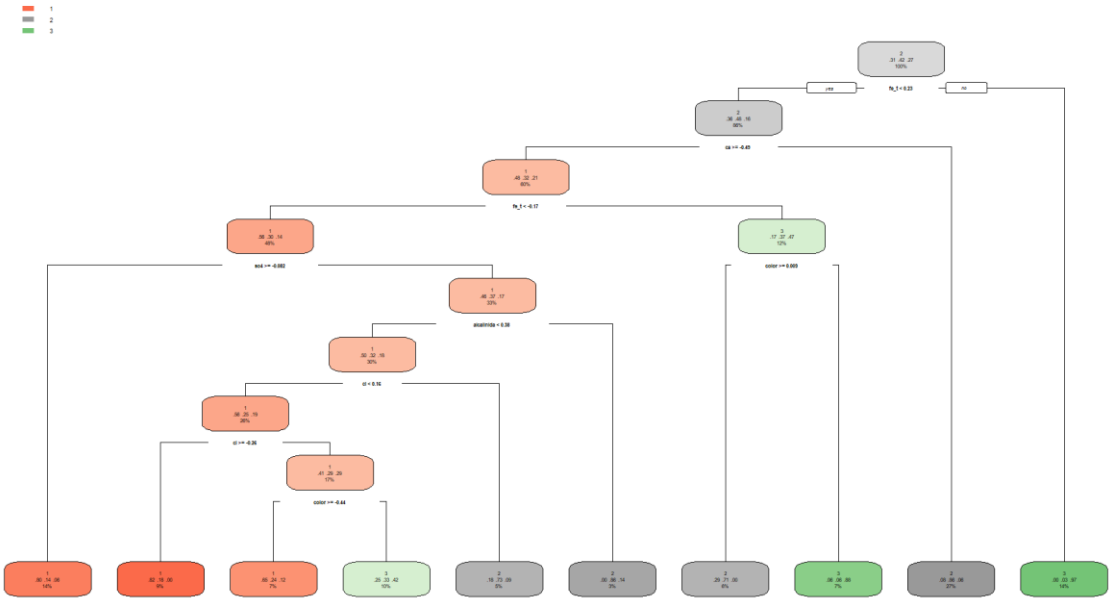


Figura 4.7 Árbol de Clasificación para características hidrogeoquímicas de fuentes de agua subterránea en Boyacá

El análisis de la importancia de las variables revela que la concentración de hierro total es la variable más relevante, con una puntuación de 16, destacándose en la clasificación de los tres tipos de captación de agua, seguido por la variable dureza, con una importancia de 14, y calcio, con una puntuación de 13.

De manera que, la variable hierro identificada como la más influyente en el modelo descrito, destaca por su capacidad para discriminar efectivamente entre diferentes clases de fuentes de agua, ya sean aljibes, manantiales o pozos. Esta observación se fundamenta en la magnitud de la reducción de la impureza que esta variable proporciona al dividir los nodos del árbol, reflejando así su significativa contribución a la precisión clasificatoria

del modelo. implica que variaciones en los niveles de hierro total tienen un impacto sustancial en la asignación de nuevas observaciones a categorías específicas de fuentes de agua. Por lo tanto, la detección de niveles elevados de hierro total en una muestra de agua probablemente conducirá a su clasificación correspondiente según los criterios establecidos por el modelo.

Por el contrario, variables como pH y potasio, aunque presentes en el modelo y contribuyendo modestamente a la precisión global, muestran una importancia relativa menor. Esto indica que los cambios en estos parámetros ejercen una influencia menos significativa en las predicciones del modelo comparados con variables como el hierro, dureza o calcio.

4.4.3 Bosques Aleatorios

4.4.3.1 Diseño

La configuración del modelo Bosques Aleatorios, incluye los parámetros:

- **Número de árboles: 800.** Este parámetro define el número de árboles de decisión que componen el bosque. En este caso, se construyen 800 árboles. Un mayor número de árboles generalmente mejora la precisión del modelo y su capacidad de generalización, ya que las predicciones se basan en la agregación de los resultados de todos los árboles. Sin embargo, también incrementa el tiempo de cómputo.

- **Complejidad de cada árbol: rpart.control(cp = 0.01**). Este parámetro controla la complejidad de cada árbol individual dentro del bosque mediante la configuración de un valor de cp (complexity parameter). Un *cp* de 0.01 indica que una partición de cualquier nodo solo se lleva a cabo si disminuye la falta de ajuste (cost-complexity) en al menos un 1%.

Este ajuste previene la sobre fragmentación de los árboles, asegurando que cada partición realizada sea significativa y contribuya a la reducción del error del modelo**.**

4.4.3.2 Análisis

El análisis de importancia de las variables en el modelo resalta varias características clave que influyen en la caracterización hidrogeoquímica del agua subterránea en Boyacá, Colombia. Entre las variables evaluadas, el hierro se destaca como la más importante, con un valor de reducción de Gini de 19.03. Esto sugiere que las variaciones en los niveles de hierro total tienen una influencia considerable en la capacidad del modelo para clasificar las fuentes de agua correctamente. En otras palabras, el hierro total es un predictor clave en la determinación de la clase de las fuentes de agua.

Otras variables con alta importancia incluyen el calcio, dureza, temperatura y alcalinidad con valores de 12.56, 12.55, 10.92 y 9.79 respectivamente. Estas variables también tienen un papel significativo en la clasificación de los tipos de captación de agua en Boyacá, Colombia. La alta importancia de estos parámetros indica que su variabilidad es crítica para el modelo, contribuyendo de manera notable a la precisión en la clasificación.

Por otro lado, variables como potasio y ph con valores de 5.64 y 4.88 respectivamente, tienen una importancia menor en comparación con las anteriores. Sin embargo, siguen siendo relevantes en el contexto del modelo, aunque su impacto en la clasificación es relativamente menor en comparación con las variables de mayor importancia.

4.4.4 Naive Bayes

4.4.4.1 Diseño

Para el modelo Naive Bayes, se utiliza la estimación de densidad del kernel (KDE) para las variables continuas, dado que estas no siguen una distribución normal. La configuración del modelo incluye los siguientes parámetros clave:

- **Estimación de densidad del Kernel (usekernel = TRUE)**: Este parámetro indica que se utilizará la estimación de densidad del kernel (KDE) para las variables continuas en lugar de asumir una distribución normal.

4.4.4.2 Análisis

El modelo Naive Bayes, proporciona las probabilidades a priori de cada clase de la variable tipo. Los valores observados para aljibe, manantial y pozo son 75, 103 y 66 respectivamente. Estos valores indican el número de instancias de cada clase en el conjunto de datos. El tipo de captación manantial tiene la mayor frecuencia, seguida por aljibes y pozos, lo que sugiere que el modelo, en ausencia de otras variables, tiene una mayor probabilidad de predecir el tipo de captación manantial.

4.4.5 Red Neuronal Artificial Perceptrón Multicapa

4.4.5.1 Diseño

La configuración de la Red Neuronal Artificial Perceptrón Multicapa incluye los siguientes parámetros y configuraciones:

- **Tamaño de la capa oculta:** Se ha definido una capa oculta de tamaño 20. Esta elección permite que la red capture patrones complejos en los datos hidrogeoquímicos, mejorando su capacidad de modelado sin incurrir en una complejidad excesiva que pueda conducir al sobreajuste del modelo.

- **Decaimiento de peso:** Se aplica un decaimiento de peso con un valor de 0.1. Este parámetro es crucial para regularizar la red, ayudando a prevenir el sobreajuste al penalizar los pesos de las conexiones más grandes.

- **Número máximo de iteraciones:** Se establece un límite de 200 iteraciones para el entrenamiento del modelo. Este límite controla el tiempo computacional y asegura que el modelo converja adecuadamente sin prolongar innecesariamente el proceso de entrenamiento

4.4.5.2 Análisis

El modelo de red neuronal perceptrón multicapa (MLP) presenta una configuración con 18 neuronas en la capa de entrada, 20 neuronas en la capa oculta y 3 neuronas en la capa de salida, totalizando 443 pesos. Esta red utiliza una función de *softmax* para la modelización, con un parámetro de decaimiento establecido en 0.1. El análisis de los pesos muestra la contribución de cada neurona de entrada a las neuronas de la capa oculta, lo que permite identificar la importancia relativa de cada variable de entrada en la generación de la salida.

4.4.6 Modelo Ensamblado Stacking

4.4.6.1 Diseño

En el enfoque de ensamblado Stacking, se combinan múltiples modelos de aprendizaje para mejorar el rendimiento de la predicción. En este caso, se utilizó un modelo de Bosques Aleatorios como base para generar predicciones iniciales. Posteriormente, se creó un conjunto de datos que contiene las predicciones del modelo de Bosques Aleatorios junto con la variable objetivo, formando así los datos de entrada para el metamodelo. El metamodelo, en este caso, se diseñó con un clasificador Naive Bayes.

Para la validación cruzada del metamodelo, los datos se dividieron en un conjunto de entrenamiento (70%) y un conjunto de prueba (30%). El clasificador Naive Bayes fue elegido metamodelo debido a su capacidad para manejar distribuciones de probabilidad, su simplicidad y potencia.

4.4.6.2 Análisis

El análisis del metamodelo Naive Bayes en el enfoque de ensamblado Stacking revela importantes aspectos de su funcionamiento y rendimiento. La distribución a priori muestra un balance relativamente uniforme entre las clases con 21, 35 y 18 observaciones en cada una de las tres categorías respectivamente. Esta distribución equitativa es esencial para el entrenamiento efectivo del modelo, asegurando que todas las clases estén representadas adecuadamente y evitando sesgos hacia clases específicas.

Capítulo V

Resultados y Discusión

En esta sección se presentan los hallazgos obtenidos en el estudio. Se aborda un proceso detallado de comparación de las medidas de rendimiento de los modelos. Finalmente, se profundiza en la discusión acerca de estos hallazgos.

5.1 Comparación de resultados

Para cada modelo evaluado, se comienzan presentando las matrices de confusión en la *Tabla 5.1*. Estas matrices son herramientas fundamentales para visualizar y comprender el rendimiento de la clasificación, ya que ofrecen una descripción detallada de los resultados del proceso. En ellas se muestra el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos. Estas métricas son esenciales para evaluar la precisión de las predicciones y la capacidad del modelo para discriminar entre diferentes clases.

Tabla 5.1 *Matrices de confusión sobre los tipos de captación de agua subterránea*

Árboles Clasificación	**Aljibe**	**Manantial**	**Pozo**
Aljibe	56.6 %	36.6%	6.6%
Manantial	16.3%	77.55%	6.12%
Pozo	7.41%	11.11%	81.48%

Bosques Aleatorios	**Aljibe**	**Manantial**	**Pozo**
Aljibe	75%	25%	0.0%
Manantial	16.32%	84.31%	7.84%
Pozo	7.41 %	7.41%	85.18%

Naive Bayes	**Aljibe**	**Manantial**	**Pozo**
Aljibe	65.21%	26.08%	8.69%
Manantial	17.39%	66.66%	15.94%
Pozo	0.0%	0.0%	100%

Red Neuronal Perceptrón Multicapa	Aljibe	Manantial	Pozo
Aljibe	70.83%	29.16%	0.0%
Manantial	10.86%	89.13%	0.0%
Pozo	13.88%	11.11%	75%

Modelo Ensamblado Stacking	Aljibe	Manantial	Pozo
Aljibe	83.33%	16.6%	0.0%
Manantial	5.26%	84.21%	10.52%
Pozo	0.0%	0.0%	100%

El modelo de Árboles de Clasificación en la categoría de manantial clasifica correctamente el 77.55% de las muestras. Sin embargo, confunde el 16.3% de las muestras de manantial, clasificándolas incorrectamente como aljibe. De manera similar, en el caso de las muestras de aljibe, el modelo identifica correctamente el 56.6% de las muestras, pero clasifica erróneamente el 36.6% de estas, asignándolas a la categoría de manantial, lo que sugiere que el modelo tiene dificultades para distinguir entre muestras de aljibe y manantial.

En comparación, el modelo de Bosques Aleatorios clasifica correctamente el 84.31% de las muestras de manantial, el 85.18% de pozo y el 75% de aljibe. Este modelo presenta dificultades principalmente en la distinción entre aljibe y manantial. Por otro lado, se observa que las muestras de pozo son clasificadas erróneamente como aljibe y manantial en un 7.41%. Estos errores de clasificación de muestras de agua de pozo clasificadas como manantial son menores en comparación con los errores en los Árboles de Clasificación, donde se registró una tasa de clasificación incorrecta del 11%.

El modelo Naive Bayes presenta una clasificación correcta del 100% de las muestras de la categoría pozo, lo que sugiere un excelente rendimiento en esta categoría. Sin embargo, su rendimiento es menor en la clasificación de aljibe y manantial, con clasificaciones correctas del 65.21% y 66.66%, respectivamente.

La red neuronal perceptrón multicapa muestra una clasificación correcta del 75% en las muestras de pozo, sin embargo, se observa que los errores de clasificación son mayores en este tipo de captación respecto a los modelos anteriores, ya que, la red neuronal clasifica erróneamente un 13.88% de las muestras de pozo como aljibe y un 11.11% como manantial. Por otra parte, este modelo presenta una alta tasa de aciertos en la clasificación de manantiales, alcanzando un 89.13%.

Finalmente, el modelo ensamblado Stacking entrenado con las predicciones de Bosques Aleatorios, clasifica correctamente el 83.33% de las muestras de aljibe, el 84.21% de manantial y el 100% de pozo. Este modelo registra el mayor porcentaje de aciertos en la clasificación de aljibes en comparación con los demás modelos, y al igual que el modelo Naive Bayes, logra una precisión del 100% en la clasificación de pozos.

En conclusión, el mayor porcentaje de error se observa cuando los modelos clasifican las muestras de aljibe incorrectamente como manantial. Esto sugiere que los modelos presentan una notable dificultad para distinguir entre estas dos fuentes de captación de agua subterránea.

5.1.1 Comparación de otras medidas de rendimiento de los modelos:

Tabla 5.2 *Exactitud (Accuracy) de los modelos de clasificación*

	Árbol de clasificación	**Bosques Aleatorios**	**Naive Bayes**	**Red Neuronal**	**Modelo Ensamblado**
Exactitud	72.64%	82.08%	70.75%	80.6%	87.5%

A partir de los resultados de las métricas de rendimiento de los modelos, en la *Tabla 5.2* se observa que el modelo ensamblado Stacking mostró el mejor desempeño con una exactitud (*accuracy)* de 87.5%, seguido por Bosques Aleatorios con 82.08%, la Red Neuronal Perceptrón Multicapa con 80.06% y el Árbol de Clasificación con 72.64%. Naive Bayes presento un desempeño menor, con 70.75%

Los modelos también fueron evaluados por su desempeño en cada tipo de fuente, utilizando métricas de Sensibilidad, Especificidad, Precisión y F1-Score y los resultados se presentan en la *Tabla 5.3.*

Tabla 5.3 *Métricas de desempeño de los modelos de clasificación*

Árbol de decisión			
	Aljibe	Manantial	Pozo
Sensibilidad	62.96%	73.07%	81.48%
Especificidad	83.54%	79.62%	93.67%
Precisión	56.66%	77.44%	81.48%
F1-Score	59.64%	75.24%	81.48%

Bosques aleatorios			
	Aljibe	Manantial	Pozo
Sensibilidad	77.77%	82.69%	85.18%
Especificidad	89.87%	85.18%	96.20%
Precisión	72.41%	84.31%	88.46%
F1-Score	75%	83.49%	86.79%

Naive Bayes			
	Aljibe	Manantial	Pozo
Sensibilidad	55.55%	88.46%	51.85%
Especificidad	89.87%	57.40%	100%
Precisión	65.21%	66.66%	100%
F1-Score	60%	76.03%	68.29%

Red Neuronal Multicapa			
	Aljibe	Manantial	Pozo
Sensibilidad	82.35%	80.98%	79.23%
Especificidad	83.44%	89.54%	100%
Precisión	63%	78.8%	100%
F1-Score	71.38%	79.87%	88.41%

El análisis de las métricas específicas por tipo de captación sugiere que los modelos de Bosques Aleatorios y el ensamblado Stacking ofrecen las mejores precisiones globales, especialmente en la identificación de muestras de pozo, donde ambos modelos alcanzan una sensibilidad del 100%. Este rendimiento superior indica que estos modelos son efectivos en la identificación esta categoría, lo que sugiere un alto nivel de confianza en

sus predicciones para pozos. Además, el modelo ensamblado Stacking presenta una sensibilidad de 92.30 % para manantial y 60% para muestras de agua de aljibes.

En cuanto a la clasificación de aljibes y manantiales, todos los modelos presentan una mayor proporción de errores, confirmando los resultados previamente observados en las matrices de confusión. Los modelos tienden a confundir estas dos categorías, donde algunas muestras de aljibe fueron clasificadas incorrectamente como manantial y viceversa. Este patrón de error es consistente entre los modelos y sugiere una mayor dificultad en distinguir entre estas dos fuentes de captación de agua subterránea, en comparación con la clasificación de pozos.

5.2 Discusión

El presente estudio ha aplicado modelos de aprendizaje automático para distinguir el tipo de captación del agua según sus características hidrogeoquímicas. El objetivo general de esta investigación ha sido evaluar y comparar la eficiencia de diferentes modelos de aprendizaje automático para la caracterización hidrogeoquímica del agua subterránea en Boyacá, Colombia. Los resultados obtenidos han confirmado que se ha cumplido este objetivo, destacando el mejor rendimiento del modelo ensamblado Stacking (modelo base: Bosques Aleatorios y metamodelo: Naive Bayes) en términos de precisión y robustez.

Los resultados obtenidos en esta investigación han evidenciado una clara variación en la eficiencia de los diferentes modelos de clasificación aplicados a la caracterización hidrogeoquímica del agua subterránea en Boyacá, Colombia. El modelo ensamblado Stacking se ha destacado con el mejor rendimiento, alcanzando una exactitud del 87.5%, seguido por Bosques Aleatorios y la Red Neuronal Multicapa con 82.08% y 80.6%, respectivamente. Sin embargo, los modelos Árbol de Clasificación y Naive Bayes han obtenido un menor rendimiento, con una exactitud de 72.64% y 70.75% respectivamente.

El modelo ensamblado ha sobresalido debido a su capacidad para integrar las fortalezas del modelo base Bosques Aleatorios y el metamodelo Naive Bayes. La combinación de

Bosques Aleatorios y Naive Bayes ha permitido manejar la complejidad y diversidad de los datos hidrogeoquímicos, superando las limitaciones individuales de cada modelo. Sin embargo, su implementación conlleva mayor complejidad y requerimientos de validación comparado con modelos como Árboles de Clasificación y Bosques Aleatorios, que, aunque ligeramente menos precisos, ofrecen ventajas en interpretabilidad y eficiencia computacional.

El análisis de la importancia de las variables en los modelos de Árbol de Clasificación y Bosques Aleatorios ha revelado que el hierro es la variable más influyente para discriminar entre los tipos de captación de agua. Este hallazgo es consistente con el análisis bivariante, que ha mostrado que los pozos presentan mayores concentraciones de hierro en comparación con aljibes y manantiales. La importancia del hierro en la clasificación se refuerza al observar que las muestras de pozo son clasificadas con mayor precisión que las de aljibe y manantial, como lo evidencian las menores tasas de error en las matrices de confusión. Los modelos presentan un mayor porcentaje de clasificaciones erróneas entre las muestras de manantial y aljibe, lo que sugiere una mayor similitud hidrogeoquímica entre estas dos fuentes en comparación con los pozos.

El análisis de correlación ha mostrado que el hierro tiene una correlación negativa con el pH en el agua de pozo, indicando que mayores concentraciones de hierro se asocian con un pH más bajo, haciendo el agua más ácida. Esta relación inversa no se observa en el agua de aljibes y manantiales, donde las correlaciones son positivas y débiles.

En conclusión, esta investigación ha demostrado que los modelos de aprendizaje automático, en particular los modelos ensamblados Stacking y Bosques Aleatorios, son técnicas adecuadas para la clasificación hidrogeoquímica del agua, proporcionando resultados satisfactorios.

Conclusiones y líneas futuras

En el estudio de la caracterización hidrogeoquímica del agua subterránea en Boyacá, Colombia, se han identificado componentes significativos, como el hierro, calcio, dureza y color, que han permitido discriminar entre diferentes tipos de fuentes de captación (aljibe, manantial y pozo). La implementación de modelos de aprendizaje automático ha mostrado que los modelos ensamblados Stacking y Bosques Aleatorios han alcanzado un mayor porcentaje de exactitud en la clasificación de las muestras de agua según su tipo de captación.

Se han empleado técnicas estadísticas como la prueba no paramétrica de Kruskal-Wallis para identificar variables significativas al nivel del 5%. Entre las variables estadísticamente significativas en el estudio se han encontrado pH, temperatura, conductividad eléctrica a 25ºC, alcalinidad, color, bicarbonato, cloruro, potasio, hierro total, calcio, entre otras. El uso de técnicas de manejo del desbalance de datos ha sido fundamental para estabilizar el rendimiento de los modelos debido a la menor representatividad de la clase de pozos en el conjunto de datos original.

En este estudio, los modelos de aprendizaje automático evaluados han demostrado resultados satisfactorios para la clasificación hidrogeoquímica del agua subterránea en Boyacá, Colombia. El modelo ensamblado Stacking ha destacado con el mejor rendimiento, logrando una exactitud del 87.5%, seguido por los Bosques Aleatorios con 82.08%, la Red Neuronal Perceptrón Multicapa con 80.6%. Los modelos Árbol de Clasificación y Naive Bayes han mostrado un rendimiento inferior, con una exactitud de 72.64% y 70.75% respectivamente.

Estos resultados son satisfactorios porque indican la capacidad de los modelos para manejar la complejidad y variabilidad inherentes a los datos hidrogeoquímicos. Los modelos ensamblados Stacking, Bosques Aleatorios y la Red Neuronal Perceptrón Multicapa, con tasas de predicción correctas superiores al 80%, sugieren una notable capacidad de generalización sobre datos de validación, evitando el sobreajuste.

La técnica de ensamblado proporciona una mayor precisión en las predicciones. El modelo ensamblado Stacking, en particular, posee la ventaja de combinar las fortalezas de diferentes modelos, reduciendo así la tasa de error. Sin embargo, este modelo sacrifica interpretabilidad en comparación con los Bosques Aleatorios. Estos últimos, además de demostrar un rendimiento satisfactorio, combinan una alta capacidad predictiva con una mayor interpretabilidad. Por tanto, desde las perspectivas de interpretabilidad y rendimiento, los Bosques Aleatorios se presentan como una opción destacada.

Finalmente, es importante subrayar que el método Naive Bayes, cuando se ha utilizado de manera independiente, ha exhibido una menor exactitud en la predicción del tipo de captación de agua en comparación con su uso como metamodelo en el ensamblado Stacking. Es así como, al integrarse este modelo en un ensamblado con otros modelos, como los Bosques Aleatorios, el rendimiento de Naive Bayes mejora significativamente, aprovechando sus características inherentes de simplicidad y eficacia.

Como líneas futuras de investigación, se propone aplicar estos modelos a otras regiones con condiciones geológicas distintas, con el fin de evaluar la robustez de los modelos en contextos diversos. Adicionalmente, se sugiere la implementación y comparación del rendimiento del método ensamblado Stacking con otros modelos ensamblados, como Bagging y Boosting, comparando la exactitud y el error de las predicciones. Asimismo, se sugiere analizar el rendimiento del método ensamblado Stacking mediante la inclusión de otros modelos base, con el objetivo de evaluar si se produce alguna mejora significativa en las predicciones.

Glosario

Las expresiones utilizadas en el presente documento deben ser entendidas con el significado que a continuación se indica. Los términos han sido tomados de la Real Academia de la Lengua Española y la plataforma web quimica.es donde se presentan conceptos, investigaciones y noticias de la industria química.

A

Adsorción: Proceso mediante el cual átomos, iones o moléculas de una sustancia se adhieren a la superficie de otra sustancia.

Anión: Ion con carga negativa.

Acuífero: Capa o vena subterránea que contiene agua.

B

Buffer: Tampón o amortiguador es una o varias sustancias químicas que afectan a la concentración de los iones de hidrógeno (o hidronios) en el agua. Siendo que pH no significa otra cosa que potencial de hidrogeniones (o peso de hidrógeno), un "buffer" mantiene estable el pH mediante la adición de ácidos o bases.

C

Catión: Ion con carga positiva.

Conductividad: Propiedad que tienen los cuerpos de transmitir el calor o la electricidad.

D

Desorción: Emisión de un fluido previamente adsorbido por un material.

Disolución: Mezcla homogénea, a nivel molecular, de una o más especies químicas que no reaccionan entre sí.

E

Electrolito: Sustancia que se descompone en iones (partículas cargadas de electricidad) cuando se disuelve en los líquidos del cuerpo o el agua, permitiendo que la energía eléctrica pase a través de ellos.

Éster: Compuesto orgánico que resulta de sustituir un átomo de hidrógeno de un ácido por un radical alcohólico.

I

Ion: Átomo o agrupación de átomos que por pérdida o ganancia de uno o más electrones adquiere carga eléctrica.

R

Redox: Dicho de una reacción química: Que se caracteriza por la oxidación de un reactante y la reducción del otro.

S

Solubilidad: Medida de la capacidad de una determinada sustancia para disolverse en otra. Puede expresarse en moles por litro, en gramos por litro, o en porcentaje de soluto; en algunas condiciones se puede sobrepasarla, denominándose a estas soluciones sobresaturadas. La sustancia que se disuelve se denomina soluto y la sustancia donde se disuelve el soluto se llama disolvente.

Solutos: Cuerpo que está disuelto.

Referencias

[1] IDEAM. ESTADO DE LAS AGUAS SUBTERRÁNEAS EN COLOMBIA. IDEAM. 2020 jul.

[2] WWF. ¿Qué es el agua subterránea y por qué es tan importante? World Wild Life. 2020. Disponible en: https://www.worldwildlife.org/descubre-wwf/historias/que-es-el-agua-subterranea-y-por-que-es-tan-importante

[3] Domenico PA, Schwartz FW. Physical and Chemical Hydrogeology. John Wiley & Sons; 1997.

[4] Chow VT, Maidment DR, Mays LW. Applied Hydrology. McGraw-Hill Science, Engineering & Mathematics; 1988.

[5] Vega DG, Guerrero-Garcia-Rojas H, Seguí-Amórtegui L. Huella hídrica: análisis como instrumento estratégico de gestión para el aprovechamiento eficiente de los... ResearchGate. 1 de diciembre de 2016; Disponible en: https://www.researchgate.net/publication/326623110_Huella_hidrica_analisis_como_instrumento_estrategico_de_gestion_para_el_aprovechamiento_eficiente_de_los_recursos_hidricos

[6] Foster SSD, Chilton PJ. Groundwater: the processes and global significance of aquifer degradation. Philosophical Transactions - Royal Society Biological Sciences [Internet]. 358(1440):1957-72. Disponible en: https://doi.org/10.1098/rstb.2003.1380

[7] Pérez LR. Manantiales. SSWM - Find Tools For Sustainable Sanitation And Water Management! Disponible en: https://sswm.info/es/gass-perspective-es/tecnologias-de/tecnologias-de-abastecimiento-de-agua/manantiales

[8] Alley WM, Healy RW, LaBaugh JW, Reilly TE. Flow and Storage in Groundwater Systems. Science. 296(5575). Disponible en: https://doi.org/10.1126/science.1067123

[9] Appelo CAJ, Postma D. Geochemistry, Groundwater and Pollution. CRC Press; 2004.

[10] Drever JI. The Geochemistry of Natural Waters: Surface and Groundwater Environments. 1997.

[11] Stumm W, Morgan JJ. Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters. John Wiley & Sons; 2012.

[12] Harter RD, Naidu R. An Assessment of Environmental and Solution Parameter Impact on Trace-Metal Sorption by Soils. Soil Science Society Of America Journal [Internet]. 1 de mayo de 2001;65(3):597-612. Disponible en: https://doi.org/10.2136/sssaj2001.653597x

[13] Fetter CW. Applied Hydrogeology, 2e. 2000.

[14] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review [Internet]. 1 de enero de 1958;65(6):386-408. Disponible en: https://doi.org/10.1037/h0042519

[15] Nearest neighbor pattern classification. IEEE Journals & Magazine | IEEE Xplore. 1967. Disponible en: https://ieeexplore.ieee.org/document/1053964

[16] Rubén Fernández Casal (ruben.fcasal@udc.es), Julián Costa Bouzas (julian.costa@udc.es), Manuel Oviedo de la Fuente (manuel.oviedo@udc.es). 3.2 Bosques aleatorios | Aprendizaje Estadístico [Internet]. Disponible en: https://rubenfcasal.github.io/aprendizaje_estadistico/bosques-aleatorios.html

[17] Gupta N. A guide to Supervised Learning - Neha Gupta - Medium. Medium. 16 de abril de 2024; Disponible en: https://medium.com/@ngneha090/a-guide-to-supervised-learning-f2ddf1018ee0

[18] Merino RFM, Chacón CIÑ. Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Dialnet. 2017. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=6230447

[19] Jauregui AF. Cómo programar un árbol de decisión en Python desde 0. Ander Fernández. 2023. Disponible en: https://anderfernandez.com/blog/programar-arbol-decision-python-desde-0/

[20] Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statistical Science. 1 de agosto de 2001;16(3). Disponible en: https://doi.org/10.1214/ss/1009213726

[21] Huacasi HYP. Bosques aleatorios - Hebert Yuri Puma Huacasi - Medium. Medium. 15 de diciembre de 2021; Disponible en: https://medium.com/@hpumah/bosques-aleatorios-482163ace92e

[22] Alarcón C. Optimización del clasificador “Naive Bayes” usando árbol de decisión C4.5. Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matem·ticas, Unidad de Posgrado; 2015.

[23] 02.1 Clasificación naive-Bayes — Introducción al Aprendizaje Automático. Disponible en: https://dcain.etsin.upm.es/~carlos/bookAA/02.1_MetodosdeClasificacion-Naive-Bayes.html

[24] IBM. ¿Qué son los clasificadores Naive Bayes? IBM. Disponible en: https://www.ibm.com/es-es/topics/naive-bayes

[25] Ripley, B. D. (1996). Pattern recognition and neural networks. https://doi.org/10.1017/cbo9780511812651

[26] Venables WN, Ripley BD. Modern Applied Statistics with S. Statisctics and computing/Statistics and computing. 2002. Disponible en: https://doi.org/10.1007/978-0-387-21706-2

[27] Interactive Chaos. El perceptrón multicapa. Disponible en: https://interactivechaos.com/es/manual/tutorial-de-deep-learning/el-perceptron-multicapa

[28] Ali M, Haider MN, Lashari SA, Sharif W, Khan A, Ramli DA. Stacking Classifier with Random Forest functioning as a Meta Classifier for Diabetes Diseases Classification. Procedia Computer Science. 1 de enero de 2022;207:3459-68. Disponible en: https://doi.org/10.1016/j.procs.2022.09.404

[29] Zhi-Hua Zhou. Ensemble Methods: Foundations and Algorithms. Available from: https://tjzhifei.github.io/links/EMFA.pdf.

[30] World Bank Group. The Hidden Wealth of Nations: Groundwater in Times of Climate Change. World Bank. 2024 ene. Disponible en: https://www.worldbank.org/en/topic/water/publication/the-hidden-wealth-of-nations-groundwater-in-times-of-climate-change

[31] Embajada de España en Guatemala, Cooperación Española. La importancia de las aguas subterráneas en la gestión integrada de los recursos hídricos: aplicaciones prácticas en proyectos de cooperación internacional para el desarrollo. 2017 nov.

Disponible en:
https://intercoonecta.aecid.es/Gestin%20del%20conocimiento/relatoria_igme_web.pdf

[32] IGAC-Instituto Geográfico Agustin Codazzi. Regiones hidrogeológicas de Colombia. 2002. Disponible en: https://www.gifex.com/fullsize/2011-08-26-14559/Regiones_hidrogeologicas_de_Colombia.html

[33] CORPOBOYACÁ. Corpoboyacá está desarrollando el estudio del “Plan de manejo ambiental del sistema acuífero de Tunja”. Corpoboyacá. 2016. Disponible en: https://www.corpoboyaca.gov.co/noticias/corpoboyaca-esta-desarrollando-el-estudio-del-plan-de-manejo-ambiental-del-sistema-acuifero-de-tunja/

[34] SERVICIO GEOLÓGICO COLOMBIANO. Conductividad Eléctrica zona centro de Boyacá. 2020. Disponible en: https://datos-sgcolombiano.opendata.arcgis.com/datasets/9f89136c92944b63bea40c3eb2e0c45a_1/about

[35] CORPOBOYACÁ. Cartilla Acuífero Tunja.

Plan de Manejo Ambiental del Sistema Acuífero de Tunja

[36] Hidrogeoquimica zona centro Boyaca | Datos Abiertos Colombia. 2024. Disponible en: https://www.datos.gov.co/dataset/Hidrogeoquimica-zona-centro-Boyaca/bs3u-igmb/about_data

[37] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling technique. Journal Of Artificial Intelligence Research/˜The œJournal Of Artificial Intelligence Research. 1 de junio de 2002;16:321-57. Disponible en: https://doi.org/10.1613/jair.953

[38] SMOTE function - RDocumentation. Disponible en: https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE

[39] Dholakiya P. SMOTE(Synthetic Minority Over-sampling Technique) - parth dholakiya - Medium. Medium [Internet]. 26 de abril de 2023; Disponible en: https://medium.com/@parthdholakiya180/smote-synthetic-minority-over-sampling-technique-4d5a5d69d720

Anexos

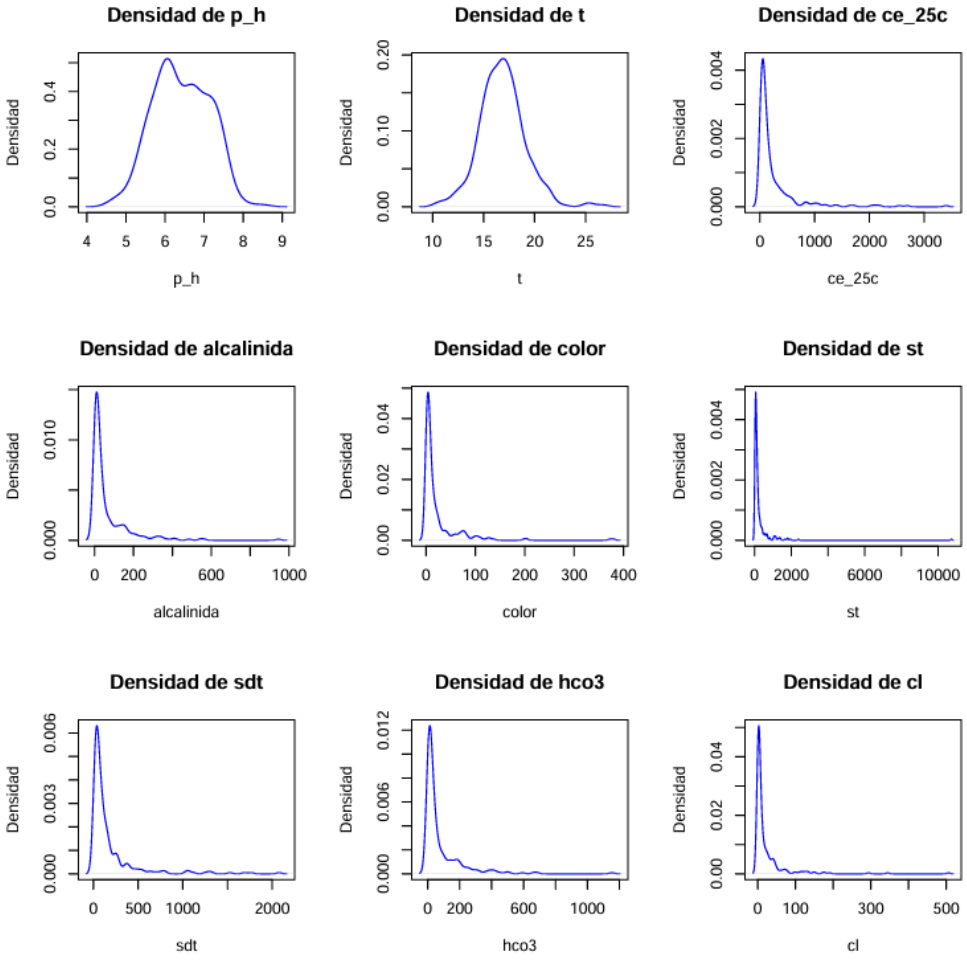
Anexo 1. Análisis Univariante de las Variables Numéricas en el Estudio Hidrogeoquímico de la Zona Centro de Boyacá.

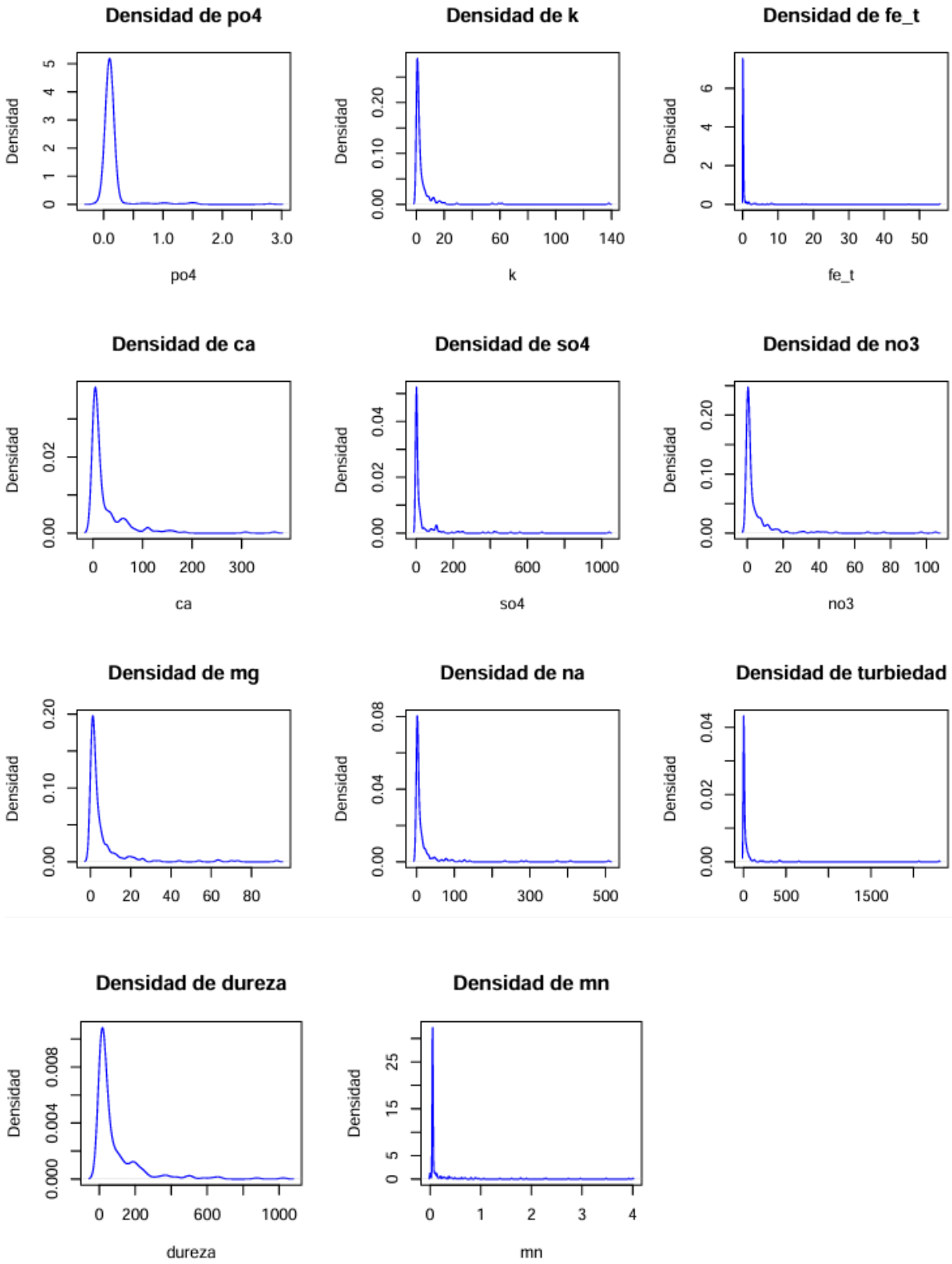
```
      p_h              t             ce_25c          alcalinida        color             st              sdt        
 Min.   :4.600   Min.   :10.40   Min.   :   6.09   Min.   :  0.30   Min.   :  0.00   Min.   :   13.0   Min.   :   6.00  
 1st Qu.:5.900   1st Qu.:15.50   1st Qu.:  44.90   1st Qu.:  6.50   1st Qu.:  2.00   1st Qu.:   58.0   1st Qu.:  32.75  
 Median :6.400   Median :16.80   Median :  95.30   Median : 20.50   Median :  7.00   Median :  107.0   Median :  76.00  
 Mean   :6.416   Mean   :16.84   Mean   : 247.85   Mean   : 64.13   Mean   : 22.61   Mean   :  261.3   Mean   : 172.98  
 3rd Qu.:7.000   3rd Qu.:18.10   3rd Qu.: 248.50   3rd Qu.: 72.50   3rd Qu.: 22.00   3rd Qu.:  210.5   3rd Qu.: 164.00  
 Max.   :8.500   Max.   :26.70   Max.   :3400.00   Max.   :945.00   Max.   :380.00   Max.   :10729.0   Max.   :2078.00  

      hco3              cl              po4                k               fe_t              ca              so4        
 Min.   :   0.00   Min.   :  0.30   Min.   :-0.1000   Min.   :  0.000   Min.   : 0.0000   Min.   :  0.20   Min.   :   0.00  
 1st Qu.:   7.90   1st Qu.:  1.90   1st Qu.: 0.1000   1st Qu.:  0.700   1st Qu.: 0.1000   1st Qu.:  2.80   1st Qu.:   1.40  
 Median :  25.00   Median :  5.00   Median : 0.1000   Median :  1.450   Median : 0.1000   Median :  7.85   Median :   4.45  
 Mean   :  77.63   Mean   : 22.16   Mean   : 0.1435   Mean   :  4.010   Mean   : 0.9077   Mean   : 24.99   Mean   :  38.62  
 3rd Qu.:  86.25   3rd Qu.: 21.95   3rd Qu.: 0.1000   3rd Qu.:  3.525   3rd Qu.: 0.2000   3rd Qu.: 28.70   3rd Qu.:  20.20  
 Max.   :1153.00   Max.   :506.00   Max.   : 2.8000   Max.   :138.000   Max.   :55.8000   Max.   :366.00   Max.   :1040.00  

      no3               mg               na            turbiedad          dureza             mn          
 Min.   :  0.000   Min.   :  0.100   Min.   :  0.100   Min.   :   0.30   Min.   :   0.90   Min.   :0.00000  
 1st Qu.:  0.200   1st Qu.:  0.900   1st Qu.:  1.575   1st Qu.:   1.50   1st Qu.:  12.47   1st Qu.:0.05000  
 Median :  0.950   Median :  2.050   Median :  3.800   Median :   4.85   Median :  32.00   Median :0.05000  
 Mean   :  5.006   Mean   :  5.759   Mean   : 19.416   Mean   :  39.40   Mean   :  86.34   Mean   :0.21178  
 3rd Qu.:  4.300   3rd Qu.:  5.350   3rd Qu.: 13.375   3rd Qu.:  20.25   3rd Qu.: 101.75   3rd Qu.:0.07775  
 Max.   :105.000   Max.   : 92.600   Max.   :508.000   Max.   :2302.00   Max.   :1022.00   Max.   :4.00000  
```

Anexo 2. Gráficos de Densidad de las Variables Numéricas en el Estudio Hidrogeoquímico de la Zona Centro de Boyacá




Densidad de po4
Densidad
po4
Densidad de k
k
Densidad de fe_t
fe_t
Densidad de ca
ca
Densidad de so4
so4
Densidad de no3
no3
Densidad de mg
mg
Densidad de na
na
Densidad de turbiedad
Densidad de dureza
dureza
Densidad de mn
mn

Anexo 3. Correlación de componentes hidrogeoquímicas del agua subterránea (Variables estandarizadas)

	p_h	t	ce_25c	alcalinida	color	st	sdt	hco3	cl	k
p_h	1.000000000	0.19692131	0.239995538	0.468448801	0.079291918	0.2303226	0.25613978	0.453112274	0.09696865	0.15448157
t	0.196921312	1.00000000	0.262498012	0.288348530	-0.100704491	0.2481106	0.28530999	0.280808724	0.22731966	0.08732845
ce_25c	0.239995538	0.26249801	1.000000000	0.648830284	-0.007808433	0.8262005	0.90680969	0.649185563	0.78466563	0.43016881
alcalinida	0.468448801	0.28834853	0.648830284	1.000000000	0.030474902	0.6603544	0.71067466	0.996155207	0.45872488	0.28332969
color	0.079291918	-0.10070449	-0.007808433	0.030474902	1.000000000	0.1031779	0.01165622	0.032501933	0.05818336	0.23743532
st	0.230322639	0.24811062	0.826200520	0.660354367	0.103177946	1.0000000	0.90761478	0.661403914	0.74528264	0.41264774
sdt	0.256139783	0.28530999	0.906809691	0.710674662	0.011656219	0.9076148	1.00000000	0.710998267	0.81988156	0.44525169
hco3	0.453112274	0.28080872	0.649185563	0.996155207	0.032501933	0.6614039	0.71099827	1.000000000	0.46116615	0.28489954
cl	0.096968649	0.22731966	0.784665631	0.458724879	0.058183363	0.7452826	0.81988156	0.461166155	1.00000000	0.58416802
k	0.154481569	0.08732845	0.430168808	0.283329694	0.237435317	0.4126477	0.44525169	0.284899542	0.58416802	1.00000000
fe_t	0.005559715	0.17982579	0.286404499	0.428956346	0.127935501	0.3234433	0.34754831	0.430167183	0.17262559	0.05833381
ca	0.392223609	0.23380135	0.754901511	0.828234406	-0.052580533	0.7758247	0.83984691	0.828078484	0.49059037	0.24896771
so4	0.145507433	0.22980648	0.785114321	0.539596200	-0.053163884	0.8087482	0.89509007	0.541356886	0.58006145	0.24364826
no3	-0.130821422	0.09409927	0.138860545	-0.006853283	-0.117480824	0.1685042	0.20234273	-0.004710633	0.22023754	0.02148885
mg	0.163108256	0.23597694	0.692660837	0.499550367	-0.048613952	0.6967602	0.79059775	0.501450361	0.66088201	0.29515690
na	0.131730122	0.27334336	0.834156107	0.475479592	0.008993627	0.7867271	0.86744618	0.474243502	0.89851083	0.42132027
dureza	0.357180803	0.25988393	0.815418357	0.806556167	-0.057002467	0.8329507	0.91399486	0.807073787	0.60118568	0.29140889
mn	0.083554397	0.18643486	0.528388092	0.432086157	0.064227737	0.5107636	0.56506931	0.433962679	0.52535179	0.29814838

	fe_t	ca	so4	no3	mg	na	dureza	mn
p_h	0.005559715	0.39222361	0.14550743	-0.130821422	0.16310826	0.131730122	0.35718080	0.08355440
t	0.179825791	0.23380135	0.22980648	0.094099269	0.23597694	0.273343364	0.25988393	0.18643486
ce_25c	0.286404499	0.75490151	0.78511432	0.138860545	0.69266084	0.834156107	0.81541836	0.52838809
alcalinida	0.428956346	0.82823441	0.53959620	-0.006853283	0.49955037	0.475479592	0.80655617	0.43208616
color	0.127935501	-0.05258053	-0.05316388	-0.117480824	-0.04861395	0.008993627	-0.05700247	0.06422774
st	0.323443275	0.77582471	0.80874825	0.168504160	0.69676017	0.786727139	0.83295070	0.51076360
sdt	0.347548307	0.83984691	0.89509007	0.202342734	0.79059775	0.867446185	0.91399486	0.56506931
hco3	0.430167183	0.82807848	0.54135689	-0.004710633	0.50145036	0.474243502	0.80707379	0.43396268
cl	0.172625591	0.49059037	0.58006145	0.220237537	0.66088201	0.898510832	0.60118568	0.52535179
k	0.058333807	0.24896771	0.24364826	0.021488854	0.29515690	0.421320266	0.29140889	0.29814838
fe_t	1.000000000	0.37898817	0.38684621	-0.065730347	0.12172930	0.180363635	0.33281570	0.31427323
ca	0.378988172	1.00000000	0.79790358	0.120568241	0.56236391	0.547147922	0.95997118	0.45650195
so4	0.386846213	0.79790358	1.00000000	0.148357125	0.70841153	0.727116161	0.85394274	0.43687309
no3	-0.065730347	0.12056824	0.14835712	1.000000000	0.38298912	0.096290634	0.22246148	-0.05704871
mg	0.121729301	0.56236391	0.70841153	0.382989119	1.00000000	0.645429087	0.77146114	0.37901317
na	0.180363635	0.54714792	0.72711616	0.096290634	0.64542909	1.000000000	0.63959919	0.55267347
dureza	0.332815705	0.95997118	0.85394274	0.222461480	0.77146114	0.639599186	1.00000000	0.47951123
mn	0.314273229	0.45650195	0.43687309	-0.057048706	0.37901317	0.552673470	0.47951123	1.00000000

Anexo 4. Script Rstudio

```
# Análisis hidrogeoquímico:
# Base de datos::
# https://www.datos.gov.co/dataset/Hidrogeoquimica-zona-centro-
Boyaca/bs3u-igmb/about_data

# Librerías: ----
library(Rcmdr)
library(dplyr)
library(ggplot2)
library(caret)
library(pROC)
library(corrplot)
library(rpart)
library(rpart.plot)
library(MASS)
library(randomForest)
library(nnet)
library(missForest)
library(smotefamily)
library(zoo)
library(adabag)
library(e1071)
library(car)

 # Carga de datos: ----
df <-
read.csv("C:/Users/ablan/OneDrive/Escritorio/TFM/conductividad_boyaca.
csv")%>%
  janitor::clean_names()
df<-df[,-5]
str(df)
```

```
summary(df)

#-----------------------------------------------------------------
# Análisis exploratorio::
#-----------------------------------------------------------------
# Transformación y tratamiento de variables: -----
# Convertimos a tipo factor las variables categóricas::
df <- df %>%
  mutate(
    fid = as.factor(fid),
    objectid = as.factor(objectid),
    no = as.factor(no),
    identifica = as.factor(identifica),
    x = as.factor(x),
    y = as.factor(y),
    z = as.factor(z),
    tipo_de_ca = as.factor(tipo_de_ca)
  )

df <- df%>%
  rename(tipo=tipo_de_ca)

proporciones <- prop.table(table(df$tipo)) * 100
proporciones <- proporciones[order(proporciones, decreasing = TRUE)]
barplot(proporciones,
        main = "Proporción en valores relativos por cada tipo de
captación",
        sub= "Set de datos Original",
        ylab = "Porcentaje",
        col = "black",
        ylim = c(0, 60))  # Ajusta el límite del eje y de acuerdo a
tus datos
text(x = barplot(proporciones, plot = FALSE),
     y = proporciones + 2,
     labels = paste0(round(proporciones, 1), "%"),
     pos = 3)

# Filtramos y solo seleccionamos el dataframe para incluir los niveles
"Aljibe", "Manantial" y "Pozo
df_filtrado <- df %>%
  filter(tipo %in% c("Aljibe", "Manantial", "Pozo"))

df <- df_filtrado %>%
  mutate(tipo = factor(tipo))

proporciones <- prop.table(table(df$tipo)) * 100
proporciones <- proporciones[order(proporciones, decreasing = TRUE)]
barplot(proporciones,
        main = "Proporción en valores relativos por cada tipo de
captación",
        sub="Set de datos filtrado por tipo captación: Aljibe,
Manantial y Pozo",
        ylab = "Porcentaje",
        col = "black",
        ylim = c(0, 60))
text(x = barplot(proporciones, plot = FALSE),
     y = proporciones + 2,
     labels = paste0(round(proporciones, 1), "%"),
     pos = 3)
```

```
#-----------------------------------------------------------------
# Análisis univariante:
#-----------------------------------------------------------------
# Histogramas
pdf("C:/Users/ablan/OneDrive/Escritorio/TFM/Histogramas.pdf", width =
14, height = 8.5)
par(mfrow=c(3, 4))
for (i in 9:28) {
  hist(df[, i],
       main = colnames(df)[i],
       col = "black",
       breaks = 30,
       xlab = colnames(df)[i],
       ylab = "Frecuencia")

}
dev.off()

# Densidad:
pdf("C:/Users/ablan/OneDrive/Escritorio/TFM/Densidades.pdf")
par(mfrow=c(3, 2))
for (i in 9:28) {
  densidad_variable <- density(df[, i])
  plot(densidad_variable,
       main = paste("Densidad de", colnames(df)[i]),
       xlab = colnames(df)[i],
       ylab = "Densidad",
       col = "blue")
}
dev.off()

#-----------------------------------------------------------------
# Manejo de valores extremos:
#-----------------------------------------------------------------
df <- df[, 8:28]
valores_extremos <- df %>%
  summarise(across(c(alcalinidad, color, st, hco3, turbiedad),
which.max)) %>%
  unlist()
df <- df[valores_extremos, ]

#-----------------------------------------------------------------
# Estandarización:
#-----------------------------------------------------------------
df_scale <- scale(df[,9:28])
df_scale <- as.data.frame(df_scale)
df <- cbind(df[, "tipo", drop = FALSE], df_scale)
str(df)

#-----------------------------------------------------------------
# Prueba de normalidad:
#-----------------------------------------------------------------

# Prueba en el conjunto de datos completo
for (i in 1:20) {
  z = df[,2:21]
  print(names(z[i]))
  print(shapiro.test(z[,i]))
```

```
}

# Prueba por grupo: Aljibe, Manantial y Pozo

columnas <- 2:21
for (col in columnas) {
  var_nombre <- colnames(df)[col]
  cat("Test de Shapiro-Wilk para", var_nombre, "\n")
  print(shapiro.test(df[df$tipo == "Aljibe", col]))
  print(shapiro.test(df[df$tipo == "Manantial", col]))
  print(shapiro.test(df[df$tipo == "Pozo", col]))
  cat("\n")
}

#-----------------------------------------------------------------
# Prueba Kruskall Wallis:
# Variables significativas aquellas variables con p-valor<0.05
#-----------------------------------------------------------------

for (i in 1:20) {
  z <- df[,2:21]
  print(names(z)[i])
  print(kruskal.test(z[,i]~df$tipo))
}

# Guardamos las variables significativas para crear un df solo con las
variables significativas::
variables_significativas <- c()
for (i in 1:20) {
  z <- df[,2:21]
  print(names(z)[i])
  resultado <- kruskal.test(z[,i]~df$tipo)
  print(resultado)

  if (resultado$p.value < 0.05) { # Si el valor p es menor que 0.05,
añadir el nombre de la variable al vector
    variables_significativas <- c(variables_significativas,
names(z)[i])
  }
}

print(variables_significativas) # variables significativas. == 17
var_sig <- df[,variables_significativas]
df <- cbind(df[, "tipo", drop = FALSE], var_sig)
str(df)

#-----------------------------------------------------------------
# Análisis bivariante
#-----------------------------------------------------------------
# Boxplots::
pdf("C:/Users/ablan/OneDrive/Escritorio/TFM/boxplot_bivariante_var_sig
nificativ.pdf", width = 14, height = 8.5)
par(mfrow=c(2,5))
for(i in 2:19){
  formula <- as.formula(paste(colnames(df)[i], "~ tipo"))
  boxplot(formula, data = df, main=colnames(df[i]), xlab="Tipo")
}
dev.off()
#-----------------------------------------------------------------
```

```
# Análisis de correlación:
#-----------------------------------------------------------------
cor <- cor(df[, 2:19])
corrplot(cor, method="circle", addCoef.col="black",
         number.cex=0.7, tl.col="black")
pdf("C:/Users/ablan/OneDrive/Escritorio/TFM/correlacion_numericas.pdf"
)
corrplot(cor, method="color", order="AOE",tl.col="black", title="\n
Correlación variables númericas estandarizadas")
dev.off()

# Correlación por tipo de captación:
aljibe <- subset(df, subset = df$tipo == "Aljibe")
cor_aljibe <- cor(aljibe[,2:19], use = "complete");cor_aljibe
manantial <- subset(df, subset = df$tipo == "Manantial")
cor_manantial <- cor(manantial[,2:19], use = "complete");cor_manantial
pozo <- subset(df, subset = df$tipo == "Pozo")
cor_pozo <- cor(pozo[,2:19], use = "complete");cor_pozo

# Gráficos de correlación por tipo de captación:
pdf("C:/Users/ablan/OneDrive/Escritorio/TFM/correlacion_tipo.pdf",
width = 14, height = 8.5)
par(mfrow=c(1,3))
corrplot(cor_aljibe, method = "color", tl.cex = 0.7, tl.col =
"black",number.cex = 0.7,
         title ="Tipo captación: Aljibe")
corrplot(cor_manantial, method = "color", tl.cex = 0.7, tl.col =
"black", number.cex = 0.7,
         title ="Tipo captación: Manantial")
corrplot(cor_pozo, method = "color", tl.cex = 0.7, tl.col = "black",
number.cex = 0.7,
         title ="Tipo captación: Pozo")
dev.off()

######################################################################
# Antes de crear los modelos, aplicamos técnica SMOTE,para tratamiento
desbalance de clases:
######################################################################
data <- df # guardamos df como data

set.seed(1806)
data$tipo <- as.integer(factor(data$tipo, levels = c("Aljibe",
"Manantial", "Pozo"),
                               labels = c(1, 2, 3)))# convertimos la
variable tipo a numerica
str(data)
data_smote <- SMOTE(X = data,
                    target = data$tipo,
                    K = 5, dup_size = 3)
names(data_smote)
data_balanceado <- data.frame(data_smote$data)
colnames(data_balanceado)[ncol(data_balanceado)] <- "tipo"
data_balanceado$tipo <- as.factor(data_balanceado$tipo)
print(table(data_balanceado$tipo))
data<-data_balanceado[,-20]

pdf("C:/Users/ablan/OneDrive/Escritorio/TFM/hist_oversampling.pdf")
proporciones_df <- table(data_balanceado$tipo)
barplot(proporciones_df,
```

```
        main = "Valores absolutos de observaciones por tipo de
captación\nDespués de aplicar Oversampling",
        ylab = "Número de observaciones",
        col = "black",
        ylim = c(0, 200))
axis_labels <- c("Aljibe", "Manantial", "Pozo")
text(x = barplot(proporciones_df, plot = FALSE),
     y = proporciones_df + 2,
     labels = paste0(round(proporciones_df, 1)),
     pos = 3)
axis(side = 1, at = 1:3, labels = axis_labels)
dev.off()

######################################################################
# Modelos de clasificación:::
######################################################################

# 1. Árbol de decisión
# 2. Bosuqes aleatorios
# 3. Red neuronal perceptrón multicapa
# 4. Naive bayes
# 5. Ensamble de modelos Stacking.
# 6. Comparación de modelos.

#---------------------------------------------------------------------
# Conjunto de entrenamiento y validación (test y train):
#---------------------------------------------------------------------
set.seed(1806)
n<-dim(data)[1];n
data.train<-sample(1:n, 0.70*n); data.train
data.test<-setdiff(1:n, data.train);data.test
length(data.train)
length(data.test)

#---------------------------------------------------------------------
# Modelo 1: Árbol de decisión:
#---------------------------------------------------------------------
model_arbol <- rpart(tipo ~ .,
                     data = data[data.train, ],
                     method = "class",
                     minsplit = 15,
                     maxdepth = 8,
                     cp=0.01)

summary(model_arbol)
rpart.plot(model_arbol)
predict_arbol <- predict(model_arbol, data[data.test, ], type =
"class")
conf_arbol <- table(predict_arbol, data[data.test, ]$tipo)
accuracy_arbol <- sum(diag(prop.table(conf_arbol))) * 100
model_arbol$variable.importance
paste("Accuracy: ", round(accuracy_arbol, 2), "%", sep = "")

# Otras medidas de rendimiento:  sensibilidad, especificidad, F1 Score
medidas_arbol<-confusionMatrix(predict_arbol, data[data.test, ]$tipo)
precision_arbol<-medidas_arbol$byClass[,'Pos Pred Value']
sensibilidad_arbol <- medidas_arbol$byClass[,'Sensitivity']
especificidad_arbol<- medidas_arbol$byClass[,'Specificity']
```

```
f1_score_arbol <-
2*((sensibilidad_arbol*precision_arbol)/(sensibilidad_arbol+precision_
arbol))

#-----------------------------------------------------------------
# Modelo 2: Bosques Aleatorios:
#-----------------------------------------------------------------
model_randomf <- randomForest(tipo ~ ., data = data[data.train,],
ntree = 800,
                              control = rpart.control(cp=0.01))
summary(model_randomf)
print(model_randomf)
names(model_randomf)
head(model_randomf$importance) # Variables más importantes para la
clasificación.
model_randomf$ntree
model_randomf$type
margen <- margin(model_randomf, data[data.train,]) # Margen de cada
punto
plot(margen) ## Gráfica de los márgenes


predic_randomf <- predict(model_randomf, newdata=data[-data.train,])
confusion_randomf <- table(predic_randomf, data[-data.train,]$tipo)
prop.table(confusion_randomf)*100

accuracy_randomf<-sum(diag(confusion_randomf)) /
sum(confusion_randomf)*100;accuracy_randomf
paste("Accuracy Bosques aleatorios (Random Forest): ",
round(accuracy_randomf, 2), "%", sep = "")

# Otras medidas de rendimiento del modelo:
medidas_random<-confusionMatrix(predic_randomf, data[data.test,
]$tipo)
precision_random<-medidas_random$byClass[,'Pos Pred Value']
sensibilidad_random <- medidas_random$byClass[,'Sensitivity']
especificidad_random<- medidas_random$byClass[,'Specificity']
f1_score_random <-
2*((sensibilidad_random*precision_random)/(sensibilidad_random+precisi
on_random))

#-----------------------------------------------------------
#Modelo 3: Redes neuronales perceptrón multicapa
#-----------------------------------------------------------
model_mlp <- nnet(tipo ~ ., data = data[data.train,],
                  size =20, decay = 0.1, maxit = 200)

predict_mlp <- predict(model_mlp, data[-data.train,], type = "class")
confusion_mlp <- table( predict_mlp, data[data.test,]$tipo)
confusion_mlp <- round(prop.table(confusion_mlp, 2)*100,
1);confusion_mlp
accuracy_mlp <- sum(diag(confusion_mlp)) / sum(confusion_mlp) * 100
paste("Accuracy red neuronal perceptrón multicapa: ",
round(accuracy_mlp, 2), "%", sep = "")

# Otras medidas de rendimiento del modelo:
verdaderos_positivos <- diag(confusion_mlp)
falsos_positivos <- colSums(confusion_mlp) - verdaderos_positivos
falsos_negativos <- rowSums(confusion_mlp) - verdaderos_positivos
```

```
verdaderos_negativos <- sum(confusion_mlp) - verdaderos_positivos -
falsos_positivos - falsos_negativos

precision_mlp <- verdaderos_positivos / (verdaderos_positivos +
falsos_positivos)
sensibilidad_mlp <- verdaderos_positivos / (verdaderos_positivos +
falsos_negativos)
especificidad_mlp <- verdaderos_negativos / (verdaderos_negativos +
falsos_positivos)
f1_score_mlp <- 2 * ((sensibilidad_mlp * precision_mlp) /
(sensibilidad_mlp + precision_mlp))

#-------------------------------------------------------
#Modelo 4: Naive Bayes
#-------------------------------------------------------
model_naive_bayes <- naiveBayes(tipo ~ ., data = data[data.train,],
                                usekernel = TRUE)
predic_naive_bayes <- predict(model_naive_bayes, newdata =
data[data.test,])
confusion_naive_bayes <- table(predic_naive_bayes,
data[data.test,]$tipo)
accuracy_naive_bayes <- sum(diag(confusion_naive_bayes)) /
sum(confusion_naive_bayes) * 100
paste("Accuracy Naive Bayes: ", round(accuracy_naive_bayes, 2), "%",
sep = "")

# Otras medidas:
medidas_naive <- confusionMatrix(predic_naive_bayes, data[data.test,
]$tipo)
precision_naive <- medidas_naive$byClass[,'Pos Pred Value']
sensibilidad_naive <- medidas_naive$byClass[,'Sensitivity']
especificidad_naive<- medidas_naive$byClass[,'Specificity']
f1_score_naive <-
2*((sensibilidad_naive*precision_naive)/(sensibilidad_naive+precision_
naive))

#-----------------------------------------------------------------
# Modelo 5: Ensamblado stacking (Modelo base Bosques Aleatorios y
metamodelo Naive Bayes
#-----------------------------------------------------------------
# Tomamos como base los modelos que hicimos de  Bosques Aleatorios.
# creamos el df para el metamodelo::
meta_data <- data.frame(
  predict_randomf = predic_randomf, # base modelo bosques aleatorios
  tipo = data[-data.train, "tipo"]
)

# Train y test(entrenamiento y validación) del meta modelo para hacer
validación cruzada::
set.seed(2105) #== 0504
n_metadata <- dim(meta_data)[1];n_metadata
metadata.train <- sample(1:n_metadata, 0.70*n_metadata);
metadata.train
metadata.test <- setdiff(1:n_metadata, metadata.train);metadata.test
length(metadata.train)
length(metadata.test)

# Metamodelo Naive Bayes
metamodelo <- naiveBayes(tipo ~ ., data = meta_data[metadata.train,],
```

```
                          usekernel = TRUE)
predict_metamodelo <- predict(metamodelo, newdata = meta_data[-
metadata.train,])

conf_matriz_ensamblado <- table(predict_metamodelo,
meta_data[metadata.test,]$tipo)
accuracy_ensamblado <- sum(diag(prop.table(conf_matriz_ensamblado))) *
100; accuracy_ensamblado

# Otras medidas de rendimiento::
medidas_ensamblado <- confusionMatrix(predict_metamodelo,
meta_data[metadata.test,]$tipo)
precision_ensamblado<-medidas_ensamblado$byClass[,'Pos Pred Value']
sensibilidad_ensamblado <- medidas_ensamblado$byClass[,'Sensitivity']
especificidad_ensamblado<- medidas_ensamblado$byClass[,'Specificity']
f1_score_ensamblado <-
2*((sensibilidad_ensamblado*precision_ensamblado)/(sensibilidad_ensamb
lado+precision_ensamblado))

#-----------------------------------------------------------------
# Comparacion modelos:
#-----------------------------------------------------------------
df_comparacion <- data.frame(
  accuracy_arbol = round(accuracy_arbol,2),
  accuracy_randomf = round(accuracy_randomf,2),
  accuracy_red = round(accuracy_mlp,2),
  accuracy_naive_bayes = round(accuracy_naive_bayes,2),
  accuracy_ensamblado = round(accuracy_ensamblado,2)
)


which.max(df_comparacion)
paste("El Accuracy más alto es del ensamblado stacking con:",
df_comparacion$accuracy_ensamblado, "%")

#-----------------------------------------------------------------
--
# Comparación matrices de confusion:
#-----------------------------------------------------------------
--

matriz_arbol <- prop.table(table(predict_arbol, data[data.test,
]$tipo),1)*100
matriz_bosque <- prop.table(table(predic_randomf,
data[data.test,]$tipo),1)*100
matriz_red <- prop.table(table( predict_mlp,
data[data.test,]$tipo),1)*100
matriz_naive <- prop.table(table(predic_naive_bayes,
data[data.test,]$tipo),1)*100
matriz_ensamblado <- prop.table(table(predict_metamodelo,
meta_data[metadata.test,]$tipo),1)*100
```