



Pontificia Universidad
JAVERIANA
Cali

MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA SELECCION DE ESTABILIZANTES UTILIZADOS EN LA GEOTECNIA DE SUELOS VIALES

Alvaro Pérez Niño
Código: 8986470

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Julián Gil González
Director(a)

Jorge Armando Castro Escudero
Codirector(a)

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 9 DE 2024

TABLA DE CONTENIDO

	Pág.
INTRODUCCIÓN	5
1. DEFINICIÓN DEL PROBLEMA.....	7
1.1. PLANTEAMIENTO DEL PROBLEMA.....	7
1.2. FORMULACIÓN DEL PROBLEMA.....	8
2. OBJETIVOS DEL PROYECTO	9
2.1. OBJETIVO GENERAL.....	9
2.2. OBJETIVOS ESPECÍFICOS.....	9
3. MARCO DE REFERENCIA	10
3.1. MARCO TEORICO.....	10
3.1.1 CONOCIMIENTOS DE LA INGENIERIA CIVIL Y GEOTECNIA.....	10
3.1.2 CONOCIMEINTOS DE LA CIENCIA DE DATOS.....	14
3.2. MARCO NORMATIVO	28
3.3. ANTECEDENTES	29
3.2.1 ESTABILIZACIÓN DE SUELOS VIALES (Propiedades y Estabilizantes)	29
3.2.2 ESTABILIZACIÓN DE SUELOS VIALES Y EL APRENDIZAJE AUTOMÁTICO.....	30
3.2.3 REFLEXIONES FINALES.....	32
4. DISEÑO INTEGRAL DE LA SOLUCIÓN	33
5. RESULTADOS DE LA SOLUCIÓN	37
5.1. ENTENDIMIENTO DEL NEGOCIO.....	37
5.2. DISEÑO DEL MODELO DE APRENDIZAJE AUTOMÁTICO.....	37
5.3. DESARROLLO DE LA HERRAMIENTA INFORMÁTICA.....	52
5.4. IMPLEMENTACIÓN Y DESPLIEGUE DE LA APLICACIÓN	63
6. CONCLUSIONES Y TRABAJOS FUTUROS	70
6.1. CONCLUSIONES	70
6.2. TRABAJOS FUTUROS.....	71
7. REFERENCIAS BIBLIOGRÁFICAS.....	73

INDICE DE FIGURAS

<i>Figura 1: Técnica de Regresión Logística - LR, Fuente: propia</i>	17
<i>Figura 2: Técnica de Máquinas de vectores de soporte – SVM, Fuente: propia</i>	18
<i>Figura 3: Técnica de Árboles de Decisión – DT, Fuente: propia</i>	19
<i>Figura 4: Técnica de Random Forest – RF, Fuente: propia</i>	20
<i>Figura 5: Técnica de K vecino más cercano - KNN, Fuente: propia</i>	20
<i>Figura 6: Técnica de Extra Trees Classifier - ETC, Fuente: propia</i>	21
<i>Figura 7: Extreme Gradient Boosting - XGBoost, Fuente: propia</i>	22
<i>Figura 8: Matriz de confusión, Fuente: propia</i>	24
<i>Figura 9: Área Bajo la Curva (AUC-ROC), Fuente: propia</i>	26
<i>Figura 10: Diagrama de flujo metodológico del desarrollo de la investigación, Fuente: propia</i>	33
<i>Figura 11: Caracterización de los suelo según tipo de Estabilización, Fuente: propia</i>	39
<i>Figura 12: Mapa de calor que muestre la correlación entre características, Fuente: propia</i>	40
<i>Figura 13: Curvas de aprendizaje y validación de los modelos de Clasificación, Fuente: Propia</i>	44
<i>Figura 14: Comparación de Matrices de Confusión entre los Modelos de Clasificación, Fuente: propia</i>	46
<i>Figura 15: Comparación de Métricas: precisión, recall y f1-score, entre los Modelos de Clasificación, Fuente: propia</i> ..	47
<i>Figura 16: Comparación de Accuracy de Modelos de Clasificación, Fuente: propia</i>	47
<i>Figura 17: Modelo Ensemble para la Clasificación de Estabilizantes, Fuente: propia</i>	50
<i>Figura 18: Comparación de Resultados entre el Mejor Modelo Individual (Random Forest) y el Modelo Ensemble, Fuente: propia</i>	51
<i>Figura 19: Planeación del desarrollo de la aplicación, Fuente: propia</i>	53
<i>Figura 20: Diseño de la arquitectura de la aplicación, Fuente: propia</i>	54
<i>Figura 21: Diagrama de contexto de la aplicación SIPSE APP, Fuente: propia</i>	55
<i>Figura 22: Diagrama de contenedores de la aplicación SIPSE APP, Fuente: propia</i>	56
<i>Figura 23: Diagrama de componentes de la aplicación SIPSE APP, Fuente: propia</i>	58
<i>Figura 24: Esquema de un sprint de desarrollo, Fuente: propia</i>	59
<i>Figura 25: Pruebas unitarias realizadas a la Aplicación SISEP APP</i>	61
<i>Figura 26: Pruebas de integración realizadas a la Aplicación SISEP APP</i>	62
<i>Figura 27: Pruebas de rendimiento realizadas a la Aplicación SISEP APP</i>	63
<i>Figura 28: Integración de los componentes y despliegue de la aplicación de la Aplicación, Fuente: propia</i>	64
<i>Figura 29: Validación de los Modelos con Nuevos datos de suelos, Fuente: propia</i>	66
<i>Figura 30: Resultados de la evaluación a expertos temáticos, Fuente: propia</i>	67
<i>Figura 31: Resultados de la evaluación a expertos técnicos, Fuente: propia</i>	68

INDICE DE TABLAS

<i>Tabla 1: Descripción de propiedades de suelos y estabilizantes, Fuente: propia</i>	<i>38</i>
<i>Tabla 2: Estadísticas descriptivas de los datos recopilados, Fuente: propia</i>	<i>39</i>
<i>Tabla 3: Características finales seleccionadas para los modelos, Fuente: propia</i>	<i>41</i>
<i>Tabla 4: Comparación de los Modelos de Aprendizaje Supervisado Seleccionados, Fuente: propia</i>	<i>41</i>
<i>Tabla 5: Hiperparámetros óptimos y mejor puntuación de rendimiento, Fuente: propia</i>	<i>43</i>
<i>Tabla 6: Comparación de la importancia de las características en cada Modelo, Fuente: propia</i>	<i>48</i>
<i>Tabla 7: Listado de Historias de usuario y Criterios de aceptación, Fuente: propia</i>	<i>52</i>
<i>Tabla 8: Implementación Técnica de los componentes de software de la aplicación, Fuente: propia</i>	<i>59</i>
<i>Tabla 9: Instrumento de evaluación para expertos temáticos, Fuente: propia</i>	<i>66</i>
<i>Tabla 10: Instrumento de evaluación para expertos técnicos, Fuente: propia.....</i>	<i>68</i>

INTRODUCCIÓN

La geotecnia es la rama de la geología que estudia la aplicación de los principios geológicos en la investigación de los geomateriales y que constituyen la corteza terrestre o capas superficiales que hacen parte en el diseño, la construcción y la explotación de proyectos de ingeniería civil, mientras la ingeniería geotécnica, se enfoca en los materiales terrestres para construir infraestructuras y se enfrenta a la complejidad de los suelos, cuyas propiedades únicas influyen directamente en la estabilidad y durabilidad de las estructuras. Esta complejidad conlleva desafíos en la selección de estabilizantes para mejorar las propiedades del suelo, dado que métodos tradicionales requieren un alto número de ensayos los cuales pueden ser prolongados y costosos. La inteligencia artificial (IA) surge como una solución prometedora para este problema, ofreciendo una visión más precisa de las interacciones entre estabilizantes y suelos, optimizando la toma de decisiones en proyectos de construcción y mejorando la calidad de las estructuras viales.

La necesidad de resolver esta problemática radica en la demanda de herramientas más eficientes para la selección de estabilizantes, ya que los métodos convencionales presentan limitaciones en la evaluación de la idoneidad de los suelos para soportar infraestructuras. Implementar un modelo de aprendizaje automático se presenta como una solución viable. Este modelo se enfocó en recolectar y analizar datos sobre propiedades físico-químicas y mecánicas de los suelos y estabilizantes aplicables a suelos viales, priorizando información de estudios de laboratorio, y la utilización de técnicas de aprendizaje supervisado para predecir la interacción entre los estabilizantes y los suelos.

Los resultados obtenidos incluyeron la creación de un almacén de datos detallado sobre propiedades de suelos y estabilizantes, la implementación de un modelo de aprendizaje automático que seleccione estabilizantes para fortalecer suelos usados en infraestructuras viales, y el desarrollo de una herramienta informática interactiva basada en este modelo de aprendizaje automático para asistir en la selección de estabilizantes según las características geotécnicas de los suelos subrasantes.

La importancia de abordar esta problemática radica en mejorar la eficiencia en la toma de decisiones de selección basada en datos en tiempo real de estabilizantes, optimización de recursos y reducción de costos en la construcción y mantenimiento de infraestructuras viales, y en minimizar el impacto ambiental al utilizar materiales geotécnicos más eficientes.

El alcance de este proyecto se limita al desarrollo e implementación de un modelo de aprendizaje automático supervisado para la selección de estabilizantes en suelos viales, utilizando propiedades fisicoquímicas y mecánicas de los suelos extraídas de estudios de laboratorio. La herramienta informática desarrollada está diseñada para operar en un entorno controlado,

enfocado en la selección óptima de estabilizantes según las características específicas de los datos disponibles. Sin embargo, el proyecto presenta algunas limitaciones, como la dependencia de la calidad y cantidad de los datos recolectados, lo que puede afectar la generalización del modelo a otras regiones geográficas o tipos de suelo no contemplados en el estudio. Además, la validación de la herramienta informática se realiza en ambientes simulados, por lo que su implementación en escenarios reales requerirá ajustes adicionales y pruebas complementarias.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

La ingeniería geotécnica se centra en la investigación y aplicación de materiales de la tierra, que comprenden suelos, rocas y materiales geológicos intermedios [1]. Estos materiales juegan un papel fundamental en numerosos campos de la ingeniería, abarcando desde la construcción de carreteras, puentes, túneles, redes de acueducto, hidroeléctricas entre otras estructuras [2]. El suelo presenta uno de los comportamientos físicos, mecánicos y químicos más complejos en comparación con otros materiales de ingeniería. Esta complejidad se debe a la presencia de múltiples fases y propiedades variables que influyen directamente en su comportamiento. A diferencia de otros materiales utilizados en construcción y manufactura, las características del suelo pueden tener un impacto significativo en la estabilidad y el rendimiento de las estructuras. Su respuesta no lineal a factores como el contenido de agua y las condiciones ambientales hace que sea esencial comprender estas complejidades para el diseño y construcción exitosa de proyectos de ingeniería civil y geotécnica [3].

En este contexto, la variabilidad natural de los suelos revela que algunos carecen de las propiedades ingenieriles esenciales para actuar como cimiento de estructuras civiles. Por tanto, se hace necesario potenciar estas características mediante diversas técnicas de estabilización con propósitos estructurales [10] [11]. La estabilización del suelo, vital en la ingeniería civil, puede llevarse a cabo a través de métodos físicos o químicos, cuya elección depende de factores clave. La estabilización química implica la incorporación de compuestos como cal o cemento para mejorar las propiedades del suelo, mientras que la estabilización física se centra en la compactación, mezcla o refuerzo del terreno [4]. La evaluación y establecimiento de la idoneidad y eficacia de estos procesos de estabilización requiere numerosos ensayos en laboratorio y en campo, con largos periodos, recursos humanos y costos considerables. Cada tipo de suelo presenta singularidades físicas y químicas, lo que exige optimizar la selección del proceso o producto de estabilización y su dosificación específica para lograr mejoras, con el fin de tomar la decisión más adecuada en cuanto al estabilizante y su dosificación [3].

La incorporación de la inteligencia artificial (IA) en el campo de la ingeniería geotécnica se ha convertido en una tendencia en constante crecimiento en los últimos tiempos. Este enfoque ha demostrado ser especialmente relevante debido a su capacidad para abordar los desafíos clave en la selección de estabilizantes de suelos, lo que a su vez conduce a la reducción de costos y tiempos en proyectos de construcción [5] [6]. En primer lugar, es fundamental destacar que la ingeniería geotécnica involucra una variedad de factores complejos, como la composición del suelo, las condiciones ambientales y la naturaleza no lineal de los procesos geotécnicos. En este contexto, los métodos de IA han demostrado su mérito al modelar y predecir relaciones no lineales

complejas entre estos factores, lo que resulta en una comprensión más profunda y precisa de la respuesta del suelo a los estabilizantes [3] [7] [8] [9].

Por otro lado, los métodos tradicionales, son valiosos, pero a menudo se ven limitados por suposiciones simplificadas debido a la complejidad de los sistemas de materiales geotécnicos. En contraste, la IA no requiere suposiciones iniciales sobre las relaciones entre variables de entrada y salida, lo que la convierte en una herramienta versátil y poderosa para la selección de estabilizantes [3] [7] [8] [9].

1.2. FORMULACIÓN DEL PROBLEMA

Considerando la inteligencia artificial (IA) como una alternativa tecnológica sostenible en la selección de estabilizantes para suelos estructurales, adaptada a las condiciones ambientales y geográficas del suelo, surgen la pregunta de investigación ¿Cómo implementar un modelo de aprendizaje automático que permita seleccionar estabilizantes para suelos viales?; pregunta de investigación que se desglosa en varios cuestionamientos investigativos clave. Iniciando con ¿Cuáles son las propiedades fisicoquímicas y mecánicas más relevantes que deben considerarse al seleccionar estabilizantes para suelos utilizados en infraestructuras viales?, seguido de ¿Cuáles son las técnicas de aprendizaje automático más adecuadas para analizar suelos subrasantes y estabilizantes relevantes aplicados en suelos viales? y, finalmente ¿Cuáles son los procedimientos para validar la precisión de los modelos de aprendizaje automático en la evaluación de los efectos de los estabilizantes en la mejora de suelos de bajas prestaciones ingenieriles?. Estos interrogantes destacan la importancia de la IA como un recurso invaluable para la toma de decisiones en la selección de estabilizantes, ya que permite optimizar el tipo de estabilizante de acuerdo con las características específicas del suelo y las condiciones del proyecto. Esto no solo conlleva a un ahorro de tiempo y recursos, sino también a una mejora sustancial en la calidad y durabilidad de las estructuras viales; además del impacto ambiental positivo al emplear geomateriales con bajas prestaciones in situ.

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Implementar un modelo de aprendizaje automático para la selección del estabilizante de suelo que fortalezca las características geotécnicas específicas de los suelos empleados en infraestructuras viales.

2.2. OBJETIVOS ESPECÍFICOS

- Preparar los datos relevantes sobre las propiedades fisicoquímicas y mecánicas específicas del suelo y de los estabilizantes aplicables a suelos viales.
- Aplicar métodos de aprendizaje automático para modelar la interacción entre los estabilizantes y el suelo vial, con énfasis en la predicción de su comportamiento y efectos geotécnicos.
- Validar los modelos de aprendizaje automático implementados para garantizar su capacidad predictiva en cuanto a la selección de los estabilizantes potenciales de mejora de suelos subrasantes.
- Desarrollar una herramienta informática basada en el modelo de aprendizaje automático validado para asistir en la selección de estabilizantes según las características geotécnicas específicas del suelo subrasante.

3. MARCO DE REFERENCIA

3.1. MARCO TEORICO

3.1.1 CONOCIMIENTOS DE LA INGENIERIA CIVIL Y GEOTECNIA

La *ingeniería civil y geotecnia* se ocupan de la construcción y mantenimiento de infraestructuras como carreteras, puentes, edificios y túneles. Estas disciplinas se basan en el análisis de los materiales, las estructuras y el comportamiento del suelo para garantizar la seguridad y la eficiencia de las infraestructuras [15].

3.1.1.1 *Infraestructura vial*

La *infraestructura vial* se refiere al conjunto de obras y servicios tales como carreteras, autopistas, puentes y otros tipos de infraestructuras de transporte [16]. Estas infraestructuras son fundamentales para el transporte de personas y mercancías las cuales impactan en el desarrollo económico de una región. Además, son esenciales para el transporte público y privado de personas, impulsando la accesibilidad, el turismo y la integración social [12][16]. Las principales infraestructuras viales son:

- *Carreteras*: Vías de acceso pavimentadas o no pavimentadas que conectan diferentes lugares. Pueden ser autopistas, autovías, carreteras secundarias o locales.
- *Autopistas*: Vías rápidas y de alto rendimiento, generalmente con múltiples carriles, accesos limitados y separadores físicos. Están diseñadas para un tráfico más rápido y eficiente.
- *Calles*: Vías urbanas o rurales que permiten el tráfico local, con características que pueden variar según el entorno, el flujo vehicular y las regulaciones urbanas.
- *Caminos rurales*: Vías que conectan áreas rurales y pueden ser pavimentadas o de tierra, destinadas al tráfico agrícola, forestal o de acceso a áreas remotas.

Estos tipos de infraestructuras viales pueden variar en diseño, función, tamaño y capacidad según las necesidades de movilidad y las características geográficas o urbanas de una región determinada.

3.1.1.2 Pavimentos

Los *pavimentos* son las capas superiores de la infraestructura vial que soportan las cargas del tráfico y proporcionan una superficie adecuada para la circulación. Los pavimentos se clasifican en flexibles, rígidos y semirrígidos, según el tipo de material, incluyendo asfalto, concreto, madera entre otros y la forma de construcción. Se utilizan para garantizar la seguridad y la eficiencia de los vehículos en las carreteras [13] [17]. A continuación, se enuncia la clasificación:

- *Pavimento Flexible*: Este tipo de pavimento consta de varias capas diseñadas para distribuir las cargas del tráfico de manera uniforme sobre el terreno subyacente. Generalmente, está compuesto por una capa de asfalto (llamada carpeta asfáltica) sobre capas de base granular y sub-base. La flexibilidad proviene principalmente de la capa superior de asfalto, permitiendo una mayor adaptación a las cargas y condiciones del terreno [17].
- *Pavimento Rígido*: los pavimentos rígidos están compuestos por losas de concreto, proporcionando una superficie sólida y resistente. Se caracterizan por ser más rígidos y distribuir las cargas de tráfico a lo largo de las losas de concreto. Son utilizados en carreteras de alto tráfico y resisten muy bien el paso constante de vehículos pesados [17].
- *Pavimento Semirrígido*: Este tipo de pavimento combina características de los pavimentos flexibles y rígidos. Por lo general, consta de una capa superior de concreto asfáltico y una base de hormigón o cemento tratado, lo que le otorga propiedades intermedias entre la flexibilidad y rigidez[17].

Cada tipo de pavimento tiene sus propias ventajas y desventajas en términos de construcción, durabilidad, costo y mantenimiento, y su elección dependerá de varios factores, como el tráfico esperado, las condiciones del suelo y el clima, entre otros.

3.1.1.3 Subrasantes

Los *subrasantes* son las capas inferiores de la infraestructura vial que sirven de apoyo y cimentación a los pavimentos. Las subrasantes están compuestas por el suelo natural o mejorado tales como: tierra, grava, arena entre otros materiales, que se encuentra debajo de la estructura del pavimento. Las propiedades de las subrasantes influyen en el diseño, la construcción y el comportamiento de los pavimentos [14][17]. A continuación, se relacionan los tipos de subrasantes:

- *Subrasantes Naturales*: Son los suelos existentes en el sitio donde se construirá la infraestructura vial. Pueden requerir mejoras para cumplir con los estándares necesarios para soportar el pavimento.

- *Subrasantes Mejoradas*: Involucran la modificación de las propiedades del suelo natural mediante técnicas como la estabilización química, la adición de materiales granulares, el uso de geosintéticos o la compactación controlada para mejorar su capacidad portante y estabilidad.
- *Subrasantes Tratadas*: Estas subrasantes son suelos que han pasado por un proceso de estabilización química, mecánica o de otro tipo para mejorar sus características de resistencia y durabilidad.

La elección del tipo de subrasante y las acciones de mejora dependerán de las condiciones del suelo local, el tráfico esperado y los requisitos de diseño para la infraestructura vial.

3.1.1.4 Estabilización de subrasantes

La *estabilización de subrasantes* se refiere al proceso de mejorar las propiedades mecánicas y de durabilidad de los suelos de base de una estructura vial, con el fin de incrementar su capacidad de soporte y resistencia a las condiciones ambientales. Este procedimiento es crucial en la construcción y mantenimiento de carreteras, aeropuertos y otras infraestructuras. El objetivo principal es convertir suelos problemáticos en materiales aptos para soportar las cargas impuestas por el tráfico y reducir la susceptibilidad a la humedad y a la deformación [14][18]. Dentro de la industria existen diferentes formas de clasificar la estabilización de subrasantes:

3.1.1.4.1 Estabilización de subrasantes según el tipo de uso

Esta clasificación se refiere a los diferentes tipos de técnicas y materiales que se emplean en la industria para estabilizar los suelos. Se divide en métodos tradicionales, como el uso de cal y cemento, y no tradicionales, que incluyen productos innovadores como polímeros y nanomateriales.

- Los *métodos tradicionales*, se han utilizado durante décadas y son ampliamente aceptadas por su eficacia probada. Los métodos tradicionales incluyen:
 - *Uso de cal*: mejora la plasticidad y la resistencia de los suelos arcillosos al reaccionar con los silicatos y aluminatos.
 - *Cemento Portland*: proporciona una unión química entre las partículas del suelo, creando una matriz resistente y duradera.
 - *Asfalto*: se utiliza principalmente en suelos granulares para proporcionar impermeabilidad y resistencia al tráfico [76].
- Los *métodos no tradicionales* son nuevas técnicas de estabilización utilizando productos innovadores, tales como:
 - *Polímeros*: materiales sintéticos que mejoran la cohesión del suelo y reducen la permeabilidad.

- *Nanomateriales*: ofrecen una estabilización más efectiva a nivel molecular, mejorando las propiedades mecánicas y químicas del suelo.
- *Resinas químicas y agentes de estabilización iónicos*: que modifican la estructura del suelo a nivel químico [77].

3.1.1.4.2 *Estabilización de subrasantes según el mecanismo de acción*

Se refiere a los principios fundamentales mediante los cuales los estabilizadores modifican las propiedades del suelo. Estos mecanismos pueden ser de naturaleza química, electroquímica o física, dependiendo del tipo de estabilizador empleado.

- Los *estabilizadores cementantes* incluyen materiales como el cemento y la cal, que reaccionan químicamente con los componentes del suelo (especialmente arcillas) para formar compuestos cementantes. Estos compuestos aumentan la rigidez y resistencia a la compresión del suelo, creando una estructura más duradera [78].
- Este tipo de estabilización *utiliza productos químicos* que alteran las propiedades eléctricas del suelo, disminuyendo su capacidad de retención de agua. Los estabilizadores electroquímicos incluyen agentes como la cal, sales y productos químicos que actúan sobre las cargas de las partículas del suelo, facilitando la reducción de su volumen de hinchamiento y la mejora de su resistencia [79].
- En este caso, los *estabilizadores físicos* no involucran una reacción química. En su lugar, se mezclan físicamente con el suelo para mejorar sus propiedades mecánicas, incrementando la cohesión y disminuyendo la plasticidad. Un ejemplo es la inclusión de fibras sintéticas o geotextiles que refuerzan la estructura del suelo [80].

3.1.1.4.3 *Estabilización de subrasantes según la familia físico-química del estabilizador*

Se refiere a la clasificación de los estabilizadores según su composición y comportamiento físico-químico, abarcando desde materiales tradicionales como los cementantes hasta innovaciones recientes como los nanomateriales.

- Los *nanomateriales* operan a nivel molecular y nanométrico, mejorando la interacción entre las partículas del suelo. Los nanomateriales, como el grafeno y el sílice nanoestructurado, permiten la modificación precisa de las propiedades del suelo, brindando ventajas en la resistencia y la durabilidad [81].
- Los *materiales cementantes* incluyen el cemento Portland, la cal y mezclas de cenizas volantes, que reaccionan con los componentes del suelo para formar compuestos estables y resistentes. Los *materiales puzolánicos*, como la ceniza volcánica y las cenizas volantes, aunque no poseen propiedades cementantes por sí solos, reaccionan

en presencia de cal para formar compuestos con características similares a las de los cementos tradicionales, mejorando la resistencia y durabilidad del suelo [82].

- *Los minerales* utilizados en la estabilización de suelos incluyen materiales como la bentonita, que mejora la impermeabilidad del suelo y su capacidad de retención de agua. Además, los minerales arcillosos, como las arcillas expansivas, pueden ser estabilizados mediante el uso de aditivos que reducen su expansión y mejoran su estabilidad mecánica. Estos minerales se emplean principalmente en suelos que requieren mejoras en la resistencia al agua y la cohesión [83].
- Los *materiales orgánicos*, se incluyen polímeros y productos derivados de fuentes orgánicas, como resinas naturales, que se utilizan para modificar las propiedades del suelo, principalmente mejorando su cohesión y resistencia al agua. Aunque no proporcionan la misma durabilidad que los estabilizadores cementantes, estos materiales son más ecológicos y sostenibles, ofreciendo una solución viable para aplicaciones donde se prioriza el impacto ambiental [84].

La elección del método de estabilización depende de las características específicas del suelo, las condiciones locales, el tipo de tráfico previsto y los requisitos de diseño para la infraestructura vial. Cada tipo de estabilización tiene sus propias ventajas y consideraciones, y su aplicación se determina según el análisis detallado del proyecto y las necesidades específicas de la subrasante [18].

3.1.2 CONOCIMIENTOS DE LA CIENCIA DE DATOS

La *ciencia de datos* es un campo interdisciplinario que se enfoca en extraer conocimiento y perspectivas útiles a partir de grandes conjuntos de datos. Combina métodos estadísticos, matemáticos, técnicas de análisis y algoritmos de aprendizaje automático para interpretar y analizar información. Esta disciplina se vale de diversas herramientas y enfoques para manipular, limpiar, organizar y comprender datos, con el fin de identificar patrones, tendencias o insights relevantes que puedan ser utilizados para la toma de decisiones en diferentes ámbitos [19].

3.1.2.1 Gestión de Datos

La gestión de datos en el ámbito de la ciencia de datos se encuentra enmarcado en la organización, almacenamiento, preservación y manipulación estratégica de los datos a lo largo de su ciclo de vida. Esto incluye aspectos como la adquisición, almacenamiento seguro, integración, procesamiento, limpieza y organización eficiente de los datos. Este proceso involucra: planificación, equipos, procesos tecnológicos y seguimiento a largo plazo para garantizar la disponibilidad y utilidad de los datos en todo el proceso de investigación [72] [73] [74].

3.1.2.1.1 ETL

ETL (Extracción, Transformación y Carga): Es el proceso de integración de datos que comprende la *extracción* relacionada con la obtención de datos desde múltiples fuentes, que pueden ser bases de datos, archivos planos, APIs, o cualquier otro origen de información, seguido de la *transformación* donde los datos extraídos se someten a cambios y ajustes para que cumplan con un formato estándar y coherente. Este estándar puede incluir la limpieza de datos, la conversión de formatos, la normalización de datos, y la integración de diferentes esquemas o estructuras para que sean compatibles y por último se gestiona la *carga de datos* en un almacén u otro sistema, que puede ser un almacén de datos, un data lake, un sistema de gestión de bases de datos, entre otros para su posterior análisis y uso [20] [29].

3.1.2.1.2 Almacén de datos

Un *almacén de datos* es un sistema centralizado y organizado que se utiliza para almacenar grandes volúmenes de datos de diversas fuentes en un único lugar. Su objetivo principal es facilitar la consulta, el análisis y la obtención de información relevante para la toma de decisiones. Este repositorio está diseñado para ser utilizado por diferentes áreas de una organización y permite un acceso rápido y eficiente a los datos [19][22]. Dentro de los tipos de almacenes, se destacan los almacenes de datos *transaccional (OLTP)*, utilizado comúnmente en entornos operativos, seguido de los almacenes de datos *analítico (OLAP)* diseñado para el análisis y la consulta de grandes volúmenes de datos; después encontramos los *data lake (DL)* que es un repositorio que almacena datos en su formato nativo, sin estructura ni procesamiento previo, ideal para almacenar grandes volúmenes de datos de diferentes tipos y formatos, permitiendo su procesamiento y análisis posterior. Continuamos con los *data warehouse (DW)*, estructura que permite almacenar datos estructurados y procesados de diversas fuentes en un formato consistente y optimizado para la consulta y por último encontramos los *data mart (DM)*, los cuales son similares a los data warehouse pero con un enfoque más específico, almacena datos relacionados con un área o departamento particular de una organización y se utiliza para necesidades analíticas de un grupo de usuarios específico [19][21][22].

3.1.2.1.3 Selección de características

La *selección de características* es el proceso de elegir las variables o atributos más relevantes o informativos para el análisis de los datos. La selección de características tiene como objetivo reducir la dimensionalidad, el ruido y la complejidad de los datos, así como mejorar el rendimiento y la interpretabilidad de los modelos de aprendizaje automático. La selección de características incluye métodos estadísticos, algoritmos de aprendizaje automático incorporados, análisis de importancia de características, métodos de filtrado, wrappers, y métodos embedded, entre otros. La elección de la técnica adecuada depende del tipo de datos, la complejidad del problema y el algoritmo de aprendizaje automático que se va a utilizar [23][24][25].

3.1.2.2 Aprendizaje Automático

El *aprendizaje automático* (*machine learning*) es el campo de la ciencia de datos que se ocupa de desarrollar y aplicar algoritmos y modelos que permiten a las máquinas aprender de los datos y realizar tareas complejas sin una programación explícita. El aprendizaje automático se basa en conceptos y técnicas de la inteligencia artificial, la estadística, la matemática y la informática. Dentro de este amplio dominio, se distinguen varios paradigmas y métodos fundamentales [19][26][27][28]:

Las técnicas de aprendizaje automático se clasifican en *aprendizaje supervisado* implica entrenar un modelo con datos etiquetados, permitiéndole hacer predicciones o tomar decisiones basadas en esa información. Por otro lado, el *aprendizaje no supervisado* trabaja con datos no etiquetados, encontrando patrones o estructuras dentro de los datos sin guía explícita [27][31]. El *aprendizaje por refuerzo* implica entrenar un modelo para tomar decisiones secuenciales, aprendiendo de sus interacciones con un entorno y recibiendo retroalimentación en base a sus acciones [27][32]. En resumen, el aprendizaje automático es un campo crucial de la inteligencia artificial que permite a las computadoras aprender y mejorar su rendimiento de manera autónoma a partir de la experiencia y los datos.

3.1.2.2.1 Aprendizaje automático supervisado

El *aprendizaje supervisado*, es el tipo de aprendizaje automático que consiste en entrenar un modelo con datos etiquetados, es decir, con datos que tienen una variable de salida o respuesta conocida. El objetivo del aprendizaje supervisado es que el modelo aprenda a predecir la variable de salida a partir de las variables de entrada, y que pueda generalizar a datos nuevos o no vistos [30]; en él se destacan las técnicas de *regresión* las cuales buscan establecer una relación matemática entre las variables de entrada y la salida para predecir valores numéricos y la *clasificación* se emplea cuando la variable de salida es categórica, lo que significa que se busca clasificar los datos en distintas categorías o clases [27][31].

Es relevante destacar que los modelos de aprendizaje supervisado exigen un nivel considerable de destreza para lograr una estructuración adecuada de los datos. Además, su entrenamiento puede implicar una inversión significativa de tiempo y recursos. Un riesgo inherente en este proceso es el sobreajuste de los modelos a los datos de entrenamiento, lo cual puede resultar en un rendimiento subóptimo cuando se enfrentan a datos no vistos previamente [75].

Por otro lado, el aprendizaje supervisado tiene un amplio espectro de aplicaciones. Entre estas se incluyen el reconocimiento de imágenes y objetos, la predicción de resultados futuros basada en datos históricos. Estas aplicaciones, tienen un potencial de aportar valor significativo en diversos contextos, desde la mejora de la eficiencia operativa hasta la generación de insights estratégicos para la toma de decisiones [75].

3.1.2.2.2 Técnicas de aprendizaje automático supervisado

Las técnicas de aprendizaje automático engloban métodos y algoritmos fundamentales para su implementación. Su diversidad permite clasificarlas según el tipo de aprendizaje, modelo, problema o datos. Entre las más destacadas en el *aprendizaje supervisado* se encuentran: Regresión Lineal, Regresión Logística, Máquinas de Vectores de Soporte (SVM), Árboles de Decisión, Bosques Aleatorios (Random Forest), k-Vecinos más Cercanos (k-NN), Naive Bayes, Redes Neuronales Artificiales (RNA), entre otras, se utilizan para modelar variables de salida basadas en datos etiquetados. Estas técnicas son fundamentales en la resolución de problemas de clasificación y regresión, donde se busca predecir categorías o valores continuos, respectivamente [34].

- *Regresión logística (LR)*: es un modelo estadístico desarrollado por *David Cox*, que se basa en la idea de predecir la probabilidad de que una variable dependiente (de tipo binaria) pertenezca a una de dos clases. La regresión logística modela la relación entre una o más variables independientes (predictoras) y una variable dependiente binaria. En lugar de predecir directamente el valor de la variable dependiente, estima la probabilidad de que ocurra un evento particular (*por ejemplo, la pertenencia a una clase*). Esta probabilidad es ajustada mediante la función logística o denominada sigmoide, la cual transforma cualquier valor real en un valor entre 0 y 1 como se muestra en la *Figura 1*, donde se visualiza la relación entre una variable independiente y la probabilidad de pertenencia a una clase [75].

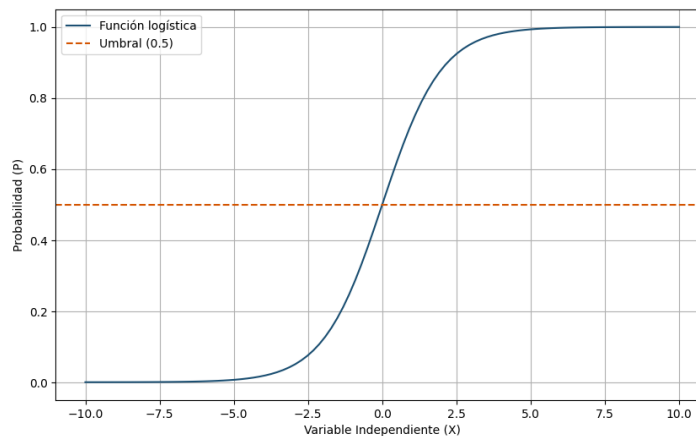


Figura 1: Técnica de Regresión Logística - LR, Fuente: propia

La fórmula general de la regresión logística es:
$$P(y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n)}} \quad (1)$$

Donde:

- $P(y = 1|X)$ es la probabilidad de que la variable dependiente y sea igual a 1 dado el conjunto de predictores X ,
- β_0 es el término independiente (intercepto),
- $\beta_1, \beta_2 \dots \beta_n$ son los coeficientes de las variables predictoras $X_1, X_2 \dots X_n$,
- e es la base del logaritmo natural.

- *Máquinas de vectores de soporte (SVM)*: una máquina de vectores de soporte es un modelo de aprendizaje supervisado muy popular, desarrollado por Vladimir Vapnik, que se utiliza tanto para la clasificación como para la regresión de datos. Su finalidad es encontrar un hiperplano en un espacio de alta dimensionalidad que separe las clases de manera óptima como se muestra en la *Figura 2*. Un hiperplano óptimo es aquel que maximiza la margen entre los puntos de datos más cercanos de las dos clases (denominados vectores de soporte) y el hiperplano [75].

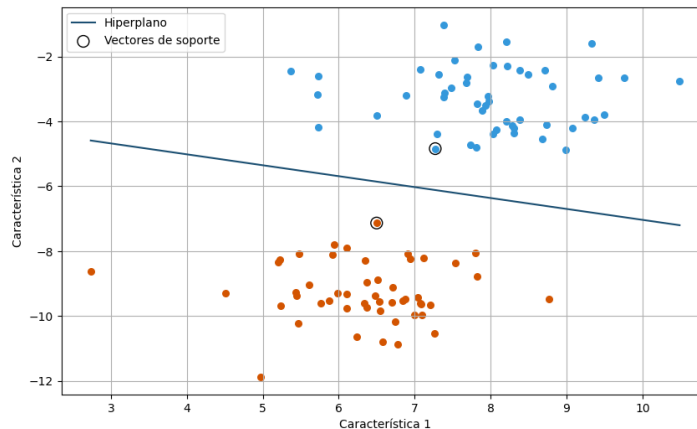


Figura 2: Técnica de Máquinas de vectores de soporte – SVM, Fuente: propia

La ecuación de un hiperplano en un espacio multidimensional es: $\mathbf{w} \cdot \mathbf{x} - b = 0$ (2)

Donde:

- w es el vector de pesos,
 - x es el vector de entrada (características),
 - b es el término independiente (sesgo).
- *Árboles de Decisión (DT)*: son modelos de aprendizaje supervisado utilizados tanto para problemas de clasificación como de regresión. La idea central de un árbol de decisión es dividir el espacio de características en regiones más simples y homogéneas, basándose en un conjunto de reglas de decisión. Cada nodo en el árbol representa una condición o prueba sobre una característica, y cada rama representa el resultado de la prueba. Los nodos hoja indican la clase o valor predicho, como se muestra en la *Figura 3*. Un árbol de decisión se construye de forma recursiva, seleccionando en cada paso la característica y el umbral que mejor dividan los datos de acuerdo con una métrica, como la entropía o el índice Gini para clasificación, o el error cuadrático medio para regresión [75].

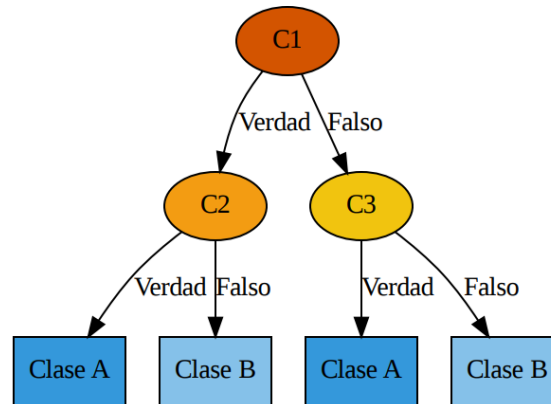


Figura 3: Técnica de Árboles de Decisión – DT, Fuente: propia

La estructura básica de un árbol se puede representar matemáticamente como una serie de decisiones:

Si $X_1 \leq t_1$ entonces seguir la rama izquierda, sino la derecha

Si $X_2 \leq t_2$ entonces seguir la rama izquierda, sino la derecha, y así sucesivamente

Donde:

- X_1, X_2, \dots, X_n son las características,
- t_1, t_2, \dots, t_n son los umbrales de partición en cada nodo.

Cada decisión divide el espacio de características en subconjuntos, y al llegar a una hoja, se asigna una predicción para la clase o el valor de la variable dependiente.

- *Random Forest (RF)*: es un algoritmo de aprendizaje supervisado muy popular, especialmente en problemas de clasificación y regresión. Está basado en la idea de construir múltiples árboles de decisión (árboles débiles) y luego combinarlos para obtener un modelo más robusto y preciso. La idea es que cada árbol "vote" por una predicción, y la predicción final sea la que obtiene más votos (en el caso de clasificación) o el promedio de todas las predicciones (en el caso de regresión) [75].

Un esquema de Random Forest, se puede representar como un grupo de varios árboles de decisión similares a los de la *Figura 3*, pero con varios árboles y una estructura que represente cómo el conjunto de árboles llega a un consenso como se representa en la *Figura 4*.

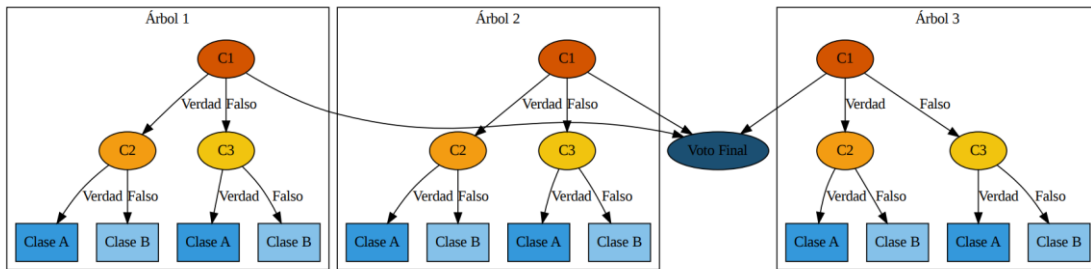


Figura 4: Técnica de Random Forest – RF, Fuente: propia

La estructura básica de un Random Forest se puede representar matemáticamente como:

Random Forest en Clasificación

$$\hat{y} = \text{modo}(T_1(x), T_2(x), \dots, T_n(x)) \quad (3)$$

Donde:

- \hat{y} es la clase predicha por el Random Forest.
- $(T_1(x), T_2(x), \dots, T_n(x))$ son las predicciones de los árboles individuales.
- La función "modo" selecciona la clase que tiene más votos entre los árboles.

Random Forest en Regresión

$$\hat{y} = \frac{1}{n} \sum_{i=1}^n T_i(x) \quad (4)$$

- *K vecino más cercano (KNN)*: el algoritmo K vecino más cercano, que también se denomina algoritmo KNN, es un algoritmo no paramétrico que clasifica los puntos de datos en función de su proximidad y asociación con otros datos disponibles. Este algoritmo presupone que los puntos de datos similares se encuentran cerca unos de otros. En consecuencia, busca la distancia entre puntos de datos, generalmente mediante distancia euclídea, y luego asigna una categoría basada en la categoría o promedio más frecuente como se puede observar en la Figura 5.

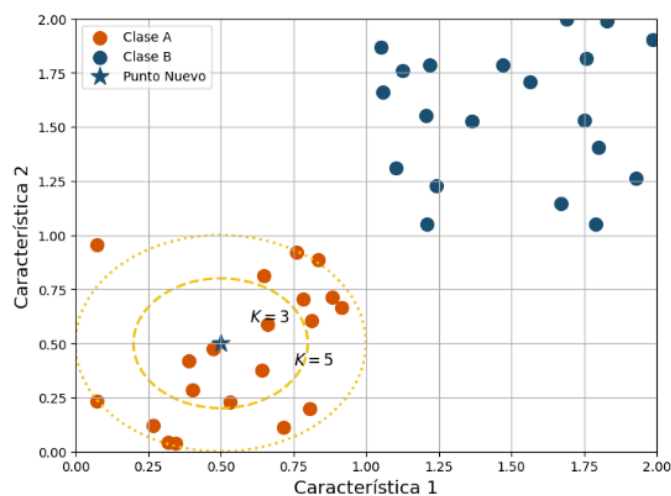


Figura 5: Técnica de K vecino más cercano - KNN, Fuente: propia

Es importante tener en cuenta, su facilidad de uso y tiempo de cálculo bajo lo convierten en el algoritmo preferido por los científicos de datos, pero a medida que el conjunto de datos de prueba crece, el tiempo de procesamiento se alarga, haciéndolo menos atractivo para las tareas de clasificación. KNN normalmente se usa para motores de recomendaciones y reconocimiento de imágenes [75].

La predicción de KNN para clasificación se puede representar de la siguiente manera:

$$\hat{y} = \text{modo}(y_1, y_2, \dots, y_K) \quad (05)$$

Donde:

- \hat{y} es la clase predicha para el punto de datos.
- y_1, y_2, \dots, y_K son las clases de los K vecinos más cercanos del punto de datos.

En el caso de regresión, la predicción se calcula como el promedio de los valores de los KKK vecinos más cercanos:

$$\hat{y} = \frac{1}{K} \sum_{i=1}^K y_i \quad (06)$$

Donde:

- \hat{y} es el valor predicho para el punto de datos.
- y_i son los valores de los K vecinos más cercanos.

- *Extra Trees Classifier (ETC)*: es un algoritmo basado en árboles de decisión que pertenece a la familia de los métodos en conjuntos. Similar al Random Forest, ETC construye múltiples árboles para realizar predicciones, pero introduce mayor aleatoriedad en el proceso. Durante la construcción de cada árbol, selecciona características y umbrales de división de manera completamente aleatoria, lo que reduce el riesgo de sobreajuste y mejora la robustez del modelo frente al ruido en los datos. La estructura básica de ETC se puede representar como un conjunto de árboles de decisión con divisiones aleatorias, como se muestra en la *Figura 6*.

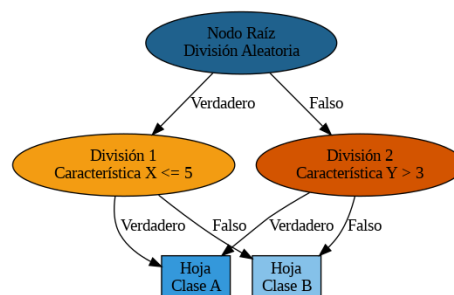


Figura 6: Técnica de Extra Trees Classifier - ETC, Fuente: propia

La predicción en *ETC* se define de forma similar a Random Forest:

Para clasificación:

$$\hat{y} = \text{modo}(T_1(x), T_2(x), \dots, T_n(x)) \quad (07)$$

Para regresión:

$$\hat{y} = \frac{1}{n} \sum_{i=1}^n T_i(x) \quad (08)$$

Donde:

- \hat{y} es la predicción final.
 - $T_1(x), T_2(x), \dots, T_n(x)$ son las predicciones de los árboles individuales.
 - La función *modo* selecciona la clase con más votos en clasificación.
 - n es el número total de árboles.
- *Extreme Gradient Boosting (XGBoost)*: es una implementación eficiente del método de boosting, diseñado para optimizar el rendimiento del modelo a través de la construcción secuencial de árboles de decisión. Cada árbol intenta corregir los errores cometidos por los anteriores, minimizando una función de pérdida específica mediante la optimización por gradiente. La estructura básica de XGBoost incluye múltiples árboles construidos secuencialmente, como se muestra en la *Figura 7*.

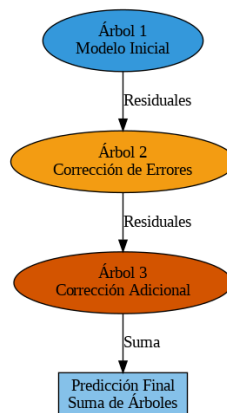


Figura 7: Extreme Gradient Boosting - XGBoost, Fuente: propia

Donde, XGBoost minimiza la siguiente función objetivo:

$$\mathcal{L}(t) = \sum_{i=1}^n l(y_i, \widehat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k) \quad (09)$$

Donde:

- $l(y_i, \widehat{y}_i^{(t)})$ es la función de pérdida para el dato i en la iteración t .
- $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda |w|^2$ es un término de regularización que controla la complejidad de los árboles:
 - T : número de hojas en el árbol.
 - w : pesos asignados a las hojas.
 - γ y λ : parámetros de regularización.

La predicción final es la suma de las predicciones de todos los árboles:

$$\hat{y} = \sum_{t=1}^T f_t(x) \quad (10)$$

Donde:

- $f_t(x)$ es el modelo generado en la iteración t

3.1.2.2.3 Métodos de evaluación para el aprendizaje automático supervisado

Los *métodos de evaluación* son las medidas o criterios que se utilizan para estimar el rendimiento, la calidad y la validez de los modelos de aprendizaje automático. Los métodos de evaluación de aprendizaje automático se pueden dividir en dos tipos: *métodos internos* y *métodos externos*. Los *métodos internos* se basan en los propios datos de entrenamiento o en una parte de ellos, como la validación cruzada o el conjunto de validación [38]. Los *métodos externos* se basan en datos independientes o no utilizados en el entrenamiento, como el conjunto de prueba o el conjunto de despliegue; permiten evaluar la capacidad del modelo de generalizar a datos no vistos previamente. La matriz de confusión, accuracy, precision, recall, F1-Score y el área bajo la curva (AUC-ROC), se aplican a estos conjuntos externos para evaluar el rendimiento del modelo en entornos del mundo real [39].

- *Validación Cruzada*: Técnica de evaluación utilizada en aprendizaje automático para medir el rendimiento de un modelo. Consiste en dividir un conjunto de datos en varios subconjuntos, permitiendo que el modelo se entrene y evalúe en distintas combinaciones de estos subconjuntos. Esto ayuda a prevenir el sobreajuste, proporcionando una estimación más robusta de la capacidad de generalización del modelo en datos no vistos. Los métodos comunes incluyen K-Fold y Leave-One-Out, los cuales permiten calcular métricas de desempeño y asegurar la efectividad del modelo en diversas situaciones.

- *Matriz de confusión*: es una herramienta que permite visualizar el rendimiento de un modelo de clasificación. Se organiza en una tabla que compara las predicciones del modelo con los valores reales. Como se muestra en la *Figura 8*, contiene cuatro elementos básicos:

Valores previstos	Si	Verdadero positivo (TP)	Falso positivo (FP)
	No	Falso negativo (FN)	Verdadero negativo (TN)
		Si	No
		Valores reales	

Figura 8: Matriz de confusión, Fuente: propia

- *Verdaderos Positivos (TP)*: Casos en los que el modelo predijo correctamente la clase positiva.
- *Falsos Positivos (FP)*: Casos en los que el modelo predijo la clase positiva incorrectamente.
- *Falsos Negativos (FN)*: Casos en los que el modelo no detectó la clase positiva cuando debería haberlo hecho.
- *Verdaderos Negativos (TN)*: Casos en los que el modelo predijo correctamente la clase negativa.

La matriz de confusión permite calcular otras métricas como precisión, recall y F1-Score, facilitando un análisis más profundo del rendimiento del modelo.

- *Accuracy (Exactitud)*: es la proporción de predicciones correctas (tanto positivas como negativas) respecto al total de predicciones realizadas. Se calcula de la siguiente manera:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (11)$$

La exactitud es útil en problemas de clasificación balanceada, pero puede ser engañosa en conjuntos de datos desbalanceados, donde una clase es mucho más prevalente que otra.

- *Precision (Precisión)*: La precisión mide la exactitud de las predicciones positivas del modelo, es decir, la proporción de verdaderos positivos entre todas las predicciones positivas realizadas. Se calcula como:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (12)$$

Una alta precisión indica que, cuando el modelo predice una clase positiva, es probable que sea correcta. Es especialmente relevante en situaciones donde los costos de los falsos positivos son altos.

- *Recall (Sensibilidad o Exhaustividad)*: también conocido como sensibilidad, mide la capacidad del modelo para identificar correctamente todas las instancias positivas. Se calcula como:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (13)$$

Un alto recall significa que el modelo tiene una baja tasa de falsos negativos, lo que es crítico en contextos donde es importante capturar la mayoría de los casos positivos.

- *F1-Score*: es la media armónica entre la precisión y el recall, ofreciendo una única métrica que captura ambas dimensiones. Se calcula como:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

El F1-Score es útil en situaciones donde existe un desbalance entre las clases, ya que proporciona un equilibrio entre la precisión y el recall.

- *Área Bajo la Curva (AUC-ROC)*: La curva ROC (*Receiver Operating Characteristic*) es una representación gráfica que muestra el rendimiento de un modelo de clasificación a diferentes umbrales de clasificación. El AUC (Área Bajo la Curva) mide la capacidad del modelo para distinguir entre las clases positiva y negativa, ver la *Figura 9*.

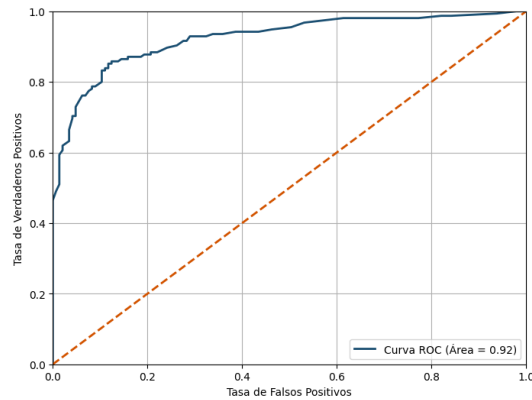


Figura 9: Área Bajo la Curva (AUC-ROC), Fuente: propia

Un AUC de 1 indica un modelo perfecto, mientras que un AUC de 0.5 sugiere un rendimiento aleatorio. El AUC-ROC es especialmente valioso en problemas de clasificación binaria, ya que proporciona una visión global del rendimiento del modelo independientemente del umbral utilizado para clasificar las predicciones.

Al evaluar un modelo supervisado de aprendizaje automático, es fundamental utilizar múltiples métricas, ya que cada una proporciona diferentes perspectivas sobre el rendimiento del modelo. La elección de las métricas adecuadas dependerá del contexto del problema y de las implicaciones de los diferentes tipos de errores en el modelo.

3.1.2.3 Aprendizaje Automático Ensembles (*Machine Learning Ensembles*)

Los meta-modelos de clasificación en aprendizaje automático son modelos que se construyen utilizando los resultados de otros modelos base para mejorar la precisión y robustez del sistema predictivo [91]. Este enfoque es parte de las técnicas de ensembles (*ensamblaje*), donde varios modelos trabajan juntos para obtener mejores resultados que cualquier modelo individual. A continuación, se presentan los principales enfoques de meta-modelos:

- *Stacking (apilamiento)*: Esta técnica consiste en combinar múltiples modelos base, que pueden ser de distintos tipos, para generar un modelo final o meta-modelo. Los modelos base se entrenan para realizar predicciones independientes y, posteriormente, esas predicciones se utilizan como entrada para entrenar el meta-modelo. Esta estrategia permite capturar relaciones complejas y no lineales entre las predicciones de los modelos base, lo que resulta en un modelo más robusto y preciso [91].

- *Boosting (Impulso)*: Esta técnica construye modelos de forma secuencial, donde cada modelo corrige los errores del modelo anterior. Esta técnica asigna mayor peso a las instancias mal clasificadas, permitiendo que los modelos subsiguientes se centren en estas. Métodos como *AdaBoost*, *Gradient Boosting*, y sus variantes más modernas, como *XGBoost* y *LightGBM*, destacan por su capacidad para reducir el sesgo y por su excelente rendimiento en datasets complejos [91].
- *Voting (votación)*: Esta técnica combina las predicciones de múltiples modelos base para tomar una decisión final. Este puede realizarse mediante votación mayoritaria, donde se selecciona la clase con más votos, o votación ponderada, que utiliza las probabilidades asignadas por los modelos para determinar la clase final. Este método es simple de implementar y eficaz cuando los modelos base tienen desempeños similares [91].

3.1.2.4 Metodologías de desarrollo de software y ciencia de datos

En el ámbito del desarrollo de software, las metodologías desempeñan un papel crucial para garantizar la eficiencia, la calidad y la adaptabilidad de los proyectos. Existen diversas metodologías que se adaptan a diferentes tipos de proyectos, entre las cuales destacan la metodología en cascada, metodologías ágiles, DevOps y el modelo incremental y prototipado [85].

En el campo del desarrollo de software, la metodología en cascada es uno de los enfoques más tradicionales, caracterizado por un proceso secuencial donde cada fase del proyecto se completa antes de pasar a la siguiente. Esta metodología es útil cuando los requisitos están bien definidos desde el principio, pero puede ser menos flexible en proyectos que experimentan cambios frecuentes [85]. En contraste, las metodologías ágiles, como Scrum y Kanban, promueven un enfoque iterativo, donde el desarrollo se realiza en ciclos cortos o sprints, permitiendo una mayor adaptabilidad y una respuesta más rápida a los cambios en los requisitos [86]. DevOps es otra metodología que fusiona las funciones de desarrollo y operaciones, con el objetivo de mejorar la colaboración entre equipos, mediante prácticas como la integración continua y el despliegue continuo, para agilizar el ciclo de vida del software [87]. Además, el modelo incremental y prototipado permite desarrollar software en partes, mostrando versiones preliminares a los clientes para ajustar los requisitos y funcionalidades según sea necesario.

Por otro lado, en el campo de la ciencia de datos, las metodologías guían el proceso de análisis y modelado de datos, asegurando que los proyectos se ejecuten de manera sistemática y generen resultados útiles. Una de las metodologías más reconocidas es CRISP-DM (*Cross-Industry Standard Process for Data Mining*), que estructura el proceso de minería de datos en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue [88]. Esta metodología es flexible y permite iteraciones a lo largo del proceso, lo que la hace ideal para proyectos de ciencia de datos que requieren ajustes constantes. Otra metodología relevante es KDD (Knowledge Discovery in Databases), que se centra en el descubrimiento de conocimiento a

partir de grandes bases de datos, integrando un enfoque detallado para la preparación, selección y transformación de los datos [89]. Además, el enfoque ágil ha sido adaptado también a la ciencia de datos, combinando el análisis iterativo con la entrega rápida de resultados, lo que permite una respuesta ágil ante nuevos descubrimientos o cambios en el proyecto. Finalmente, el enfoque Lean Data Science prioriza la creación rápida de prototipos y el enfoque en el valor para el cliente, optimizando recursos y tiempos [90].

La combinación de metodologías de desarrollo de software y ciencia de datos permite abordar proyectos complejos de manera integral. Por ejemplo, frameworks ágiles como Scrum pueden integrarse con metodologías como CRISP-DM para garantizar que tanto el desarrollo de aplicaciones como el análisis de datos se realicen de manera estructurada, eficiente y orientada al logro de objetivos específicos.

3.2. MARCO NORMATIVO

El desarrollo de este proyecto, enfocado en la selección de estabilizantes para suelos viales mediante aprendizaje automático, se sustenta en un marco normativo que abarca estándares internacionales, normativas relacionadas con la inteligencia artificial y regulaciones ambientales. Este marco garantiza que las prácticas aplicadas cumplan con requisitos técnicos, éticos y legales, tanto a nivel internacional como nacional.

En el ámbito de la ingeniería geotécnica, el Sistema Unificado de Clasificación de Suelos (SUCS) es una referencia técnica fundamental. Esta norma clasifica los suelos según su granulometría, límites de consistencia y plasticidad, proporcionando una categorización precisa que permite identificar las propiedades geotécnicas críticas para el diseño y estabilización de infraestructuras viales. El SUCS clasifica los suelos en grupos principales como gravas, arenas, limos, arcillas y suelos orgánicos, facilitando la selección de estabilizantes adecuados para mejorar la capacidad portante y la estabilidad de los subrasantes [92]. En Colombia, esta norma es reconocida y complementada por las especificaciones técnicas establecidas en proyectos viales, ajustadas a las características geográficas y climáticas específicas del país [93].

En cuanto al uso de inteligencia artificial, el proyecto se adhiere a los lineamientos de la norma ISO/IEC 22989, que establece conceptos y principios para garantizar que los modelos desarrollados sean confiables, transparentes y explicables [94]. Estos principios son esenciales para asegurar que las predicciones realizadas por el modelo de aprendizaje automático sean interpretables y ajustadas a los datos geotécnicos analizados. Además, se aplican directrices éticas internacionales, como las establecidas por la UNESCO, que promueven el uso responsable de la inteligencia artificial, priorizando la sostenibilidad y evitando la introducción de sesgos que puedan impactar negativamente en la toma de decisiones [95].

El proyecto también considera la normativa ambiental como un eje transversal en el diseño y ejecución de las soluciones propuestas. En este contexto, Colombia ha adoptado políticas ambientales alineadas con los Objetivos de Desarrollo Sostenible (ODS), específicamente los relacionados con la construcción de infraestructuras resilientes y sostenibles. Las regulaciones nacionales exigen que el uso de estabilizantes químicos y físicos minimice el impacto sobre el suelo y los recursos hídricos, garantizando la preservación de los ecosistemas circundantes [96]. Esto incluye el cumplimiento de las disposiciones del Ministerio de Ambiente y Desarrollo Sostenible, que regula el manejo de materiales y residuos en proyectos de infraestructura [97].

En conclusión, el marco normativo de este proyecto se construye sobre la integración del SUCS como estándar geotécnico principal, los principios de transparencia y ética de la inteligencia artificial, y la normativa ambiental nacional e internacional. Esta combinación asegura que el proyecto no solo cumpla con los más altos estándares técnicos, sino que también responda a los retos éticos y ambientales de manera responsable y sostenible en el contexto colombiano.

3.3. ANTECEDENTES

3.2.1 ESTABILIZACIÓN DE SUELOS VIALES (Propiedades y Estabilizantes)

La estabilización del suelo es una práctica común en la ingeniería y la geotecnia que busca mejorar las propiedades del suelo *in situ* [14][18]. Actualmente, se emplean dos métodos principales para este propósito: la estabilización mecánica, que implica ajustar la gradación del suelo mediante la mezcla con otros tipos de suelos [43], y la estabilización química, que implica la introducción de materiales químicamente activos como sales orgánicas e inorgánicas, productos enzimáticos, polímeros, aceites sulfonados, organosilanos y subproductos del petróleo [42]. Estos materiales modifican propiedades del suelo como plasticidad, permeabilidad, compresibilidad, resistencia, deformación, inestabilidad volumétrica, asentamiento, módulo elástico y durabilidad [40] [41]. Además de estos métodos, la estabilización física es una aproximación crucial. Se enfoca en alterar las propiedades físicas del suelo, incluyendo su estructura y comportamiento ante las cargas, sin necesariamente cambiar su composición química [43].

Por otro lado, diversas investigaciones han explorado métodos innovadores, como el método de estabilización y solidificación (S/S), que combina contaminantes con aglutinantes para reducir la lixiviabilidad y minimizar los riesgos para la salud humana y el medio ambiente [49]. Otro método consiste en el uso de ceniza de bagazo de caña de azúcar (SBA) y cal como mezclas, que, según se ha comprobado, mejoran el rendimiento del suelo estabilizado y reducen los costos de los materiales [50]. La precipitación de carbonatos inducida por microbios (MICP) es un nuevo método de biorremediación que se puede utilizar para la solidificación del suelo y la estabilización de metales pesados. Consiste en el uso de la pared celular bacteriana y de sustancias poliméricas extracelulares (EPS) para regular las propiedades del suelo y mejorar la resistencia a la erosión [51].

También se ha descubierto que la estabilización de las cenizas volantes (FA), combinada con fibras naturales y sintéticas, es una técnica eficaz y de bajo costo para mejorar las propiedades geotécnicas de los sustratos arcillosos expansivos [52]. Además, la combinación de residuos de lodo de hormigón y fibra de poliéster ha demostrado ser prometedora para mejorar diversas propiedades de ingeniería del suelo [53].

Es importante tener presente que cada tipo de estabilizante puede afectar una o varias de las propiedades de desempeño del suelo, dependiendo de la naturaleza y el tipo específico de suelo tratado, así como de las condiciones de aplicación [44][45][46][47][48].

3.2.2 ESTABILIZACIÓN DE SUELOS VIALES Y EL APRENDIZAJE AUTOMÁTICO

Los modelos de aprendizaje automático representan herramientas tecnológicas alternativas en la mejora de la selección de elementos para la estabilización del suelo en proyectos viales. Su capacidad para analizar patrones complejos y tomar decisiones basadas en datos resulta invaluable en este proceso, permitiendo una elección más precisa de los elementos requeridos para la estabilización del suelo. Técnicas tales como las máquinas vectoriales de soporte (SVM) y las redes neuronales artificiales (ANN) [54] son las de mayor implementación por los investigadores. Estas técnicas pueden predecir los parámetros de resistencia del suelo y ayudar a determinar el porcentaje óptimo de estabilizadores del suelo, como el polvo de vidrio, para mejorar el rendimiento de la capa base [55]. El uso de modelos de aprendizaje automático como el ANN ha demostrado un mejor rendimiento en la predicción de la resistencia del suelo en comparación con el SVM [56]. Además, la combinación de diferentes aglutinantes, como la cal, las cenizas volantes energéticas, las cenizas biovolantes, la escoria y el cemento, como agentes estabilizantes puede mejorar significativamente los parámetros de resistencia de los suelos arcillosos y hacerlos aptos para la construcción de carreteras [57]. Al estabilizar los sustratos débiles del suelo, se pueden mejorar la capacidad de carga y la susceptibilidad a la humedad del suelo, lo que mejora el rendimiento del pavimento y reduce el desgaste prematuro [58].

Principalmente los investigadores han explorado el uso de técnicas de inteligencia artificial, como las redes funcionales (FN), las splines de regresión adaptativa multivariante (MARS), las redes neuronales artificiales (ANN) y las máquinas vectoriales de soporte (SVM) para predecir parámetros como la densidad máxima en seco (MDD) y la resistencia a la compresión ilimitada (UCS) del suelo estabilizado con cemento [59]. La estabilización de geopolímeros, que es una técnica nueva, también se ha estudiado utilizando enfoques de inteligencia computacional como las redes neuronales de tipo GMDH [60]. Se han utilizado métodos de aprendizaje automático, como la ANN, la SVM y los árboles de decisión, para modelar el comportamiento mecánico de las arenas tratadas con cemento [61]. Estos modelos de aprendizaje automático han arrojado resultados prometedores a la hora de predecir el comportamiento entre la tensión y la deformación y las propiedades de resistencia de los suelos estabilizados, superando a los métodos empíricos tradicionales [62]. En

general, la combinación de estabilización del suelo e inteligencia artificial ofrece un enfoque poderoso para mejorar las propiedades del suelo y optimizar los procesos de construcción.

En otras investigaciones permiten evidenciar la aplicación del aprendizaje automático para estabilizar el suelo de las carreteras mediante el empleo de análisis de regresión y modelos de redes neuronales [63]. Estos modelos pueden predecir los parámetros de compactación del suelo, como el contenido óptimo de humedad (OMC) y la densidad seca máxima (MDD) [59]. Al desarrollar modelos de redes neuronales artificiales (ANN) con diferentes hiperparámetros, como capas ocultas y neuronas, se pueden determinar los mejores modelos arquitectónicos de ANN para realizar predicciones precisas [64]. Los modelos ANN, como el ANN_OMC_3H8 y el ANN_MDD_5H1, han demostrado su potencial para predecir el MDC y el MDC del suelo, respectivamente [54]. Esta aplicación del aprendizaje automático puede mejorar la precisión de la predicción de la colonización de los cimientos de las carreteras en suelos blandos, lo que aumenta la seguridad y la durabilidad de los proyectos de carreteras [65].

La evaluación del desempeño de los modelos de aprendizaje automático aplicados a la estabilización del suelo en carreteras puede llevarse a cabo mediante la comparación de su rendimiento con los métodos tradicionales [54]. Esta evaluación implica comparar los resultados de los modelos de aprendizaje automático con los de las técnicas convencionales utilizadas para predecir los parámetros relacionados con la estabilización del suelo [67]. La evaluación se puede realizar evaluando la precisión (Accuracy) y la eficiencia (Recall) de los modelos de aprendizaje automático frente a las categorías o valores asociadas a las propiedades del suelo que se deseen observar o predecir [71]. Se pueden utilizar varios índices de evaluación, como el MSE (Error Cuadrático Medio), el RMSE (Error Cuadrático Medio de la Raíz), el MAE (Error Absoluto Medio) y la correlación de Pearson, para medir objetivamente el rendimiento de los modelos [71].

Las investigaciones identificadas en este ámbito han proporcionado evidencia de una serie de beneficios significativos. Estos estudios han demostrado en primer lugar, el ahorro tiempo y costes durante la construcción al optimizar el uso de materiales y reducir la necesidad de reparaciones posteriores [66]. En segundo lugar, los algoritmos de aprendizaje automático han demostrado su eficacia para predecir diversas características del suelo, como la resistencia a la compresión y las deformaciones, que son cruciales para la estabilidad de la carretera [67]. Además, los modelos de aprendizaje automático pueden integrar datos de diferentes fuentes y proporcionar predicciones precisas, lo que mejora la comprensión de los problemas de ingeniería geotécnica [68]. Sin embargo, también hay algunas limitaciones que se deben tener en cuenta. Los algoritmos de aprendizaje automático pueden requerir un tamaño de muestra grande y un tiempo de aprendizaje más prolongado para obtener resultados óptimos [69]. Además, la interpretabilidad de los modelos resultantes puede ser difícil, lo que dificulta la comprensión de los factores subyacentes que influyen en la estabilización del suelo [70]. En general, si bien el aprendizaje automático ofrece ventajas significativas en la estabilización del suelo, se deben considerar cuidadosamente los requisitos y limitaciones específicos de cada proyecto.

3.2.3 REFLEXIONES FINALES

Gestionar la estabilización de suelos de acuerdo con sus distintas propiedades y características particulares, desde una perspectiva basada en la ciencia de datos, se han empleado métodos como las redes neuronales artificiales (ANN), sistemas de inferencia difusa (FIS) y máquinas de vectores de soporte (SVM). Las ANN destacan por su capacidad para modelar relaciones no lineales y predecir parámetros geotécnicos clave como densidad máxima en seco (MDD) y resistencia a la compresión (UCS), superando a otros métodos en precisión. Los FIS manejan la incertidumbre en los datos geotécnicos, proporcionando interpretaciones claras, mientras que las SVM son efectivas en datos de alta dimensionalidad gracias a sus márgenes óptimos de clasificación. Redes más avanzadas, como las convolucionales (CNN) y ResNet, sobresalen en el análisis de datos espaciales, siendo útiles en aplicaciones específicas como la caracterización visual de suelos.

La calidad de los datos utilizados para entrenar y evaluar los modelos es otro desafío crucial. La selección adecuada de los parámetros de entrada y la cantidad de datos influyen directamente en la precisión de las predicciones. Sin embargo, a pesar de estos desafíos, la aplicación de métodos de inteligencia artificial ha arrojado resultados exitosos y prometedores en la estabilización del suelo.

Los modelos de aprendizaje automático pueden predecir parámetros de resistencia y determinar porcentajes óptimos de estabilizadores, como el polvo de vidrio, mejorando así el rendimiento de la capa base de la infraestructura vial. Además, han demostrado ser superiores a los métodos convencionales en la predicción de propiedades geotécnicas cruciales para la estabilidad de carreteras, como la resistencia a la compresión y las deformaciones. Estos modelos también han permitido integrar datos de diversas fuentes para proporcionar predicciones precisas, mejorando la comprensión de los desafíos de ingeniería geotécnica.

No obstante, existen limitaciones importantes a considerar. Los algoritmos de aprendizaje automático pueden requerir conjuntos de datos extensos y tiempos prolongados de entrenamiento para obtener resultados óptimos. Además, la interpretabilidad de los modelos puede ser un desafío, dificultando la comprensión de los factores subyacentes que influyen en la estabilización del suelo. Por consiguiente, es esencial considerar cuidadosamente los requisitos y limitaciones específicos de cada proyecto al utilizar aprendizaje automático en la estabilización del suelo vial.

4. DISEÑO INTEGRAL DE LA SOLUCIÓN

El desarrollo de este proyecto se estructuró bajo un enfoque híbrido que combina las metodologías CRISP-DM (*Cross-Industry Standard Process for Data Mining*) y SCRUM, con el propósito de abordar de manera integral las necesidades técnicas y operativas del sistema propuesto. La elección de este enfoque responde a la necesidad de garantizar rigor en las fases de análisis de datos y modelado predictivo, al tiempo que se asegura agilidad y adaptabilidad durante el desarrollo de la herramienta informática y su integración final.

El diagrama de flujo de la *Figura 10*, ilustra este enfoque híbrido, que permitió gestionar la complejidad técnica de cada componente (modelo de aprendizaje automático y herramienta informática) y la interrelación entre ellos, asegurando una solución robusta, escalable y aplicable a entornos reales.

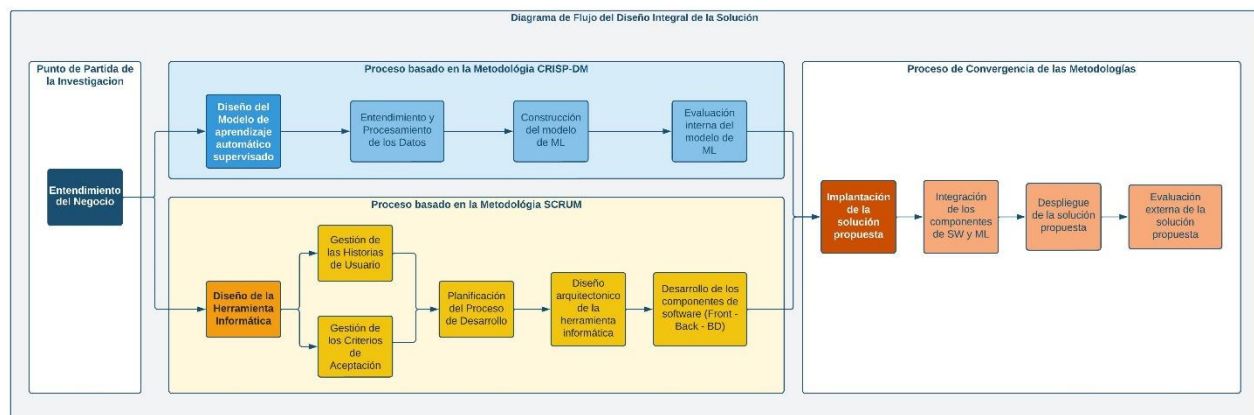


Figura 10: Diagrama de flujo metodológico del desarrollo de la investigación, Fuente: propia

Entendimiento del Negocio: Esta etapa inicial tuvo como objetivo comprender a profundidad el problema y establecer los requerimientos necesarios para desarrollar una solución efectiva.

1. Definición del Problema:

- Los desafíos en la selección de estabilizantes para suelos viales fueron identificados, destacando la necesidad de optimizar el tiempo, los costos y la precisión mediante el uso de inteligencia artificial.
- Los objetivos principales se enfocaron en seleccionar estabilizantes óptimos para suelos viales.

2. Análisis de Requerimientos:

- Se establecieron los objetivos del modelo de aprendizaje automático y los requerimientos funcionales de la herramienta informática.
- Se determinó que la solución debía ser accesible, eficiente y validada bajo estándares de calidad tanto técnicos como prácticos.

Diseño del Modelo de Aprendizaje Automático: La segunda etapa, el enfoque estuvo centrado en el diseño del modelo de aprendizaje automático para la clasificación de estabilizantes. Se utilizó el marco lógico CRISP-DM (*Cross-Industry Standard Process for Data Mining*), que asegura un proceso estructurado y repetible, abarcando desde la comprensión inicial del problema hasta la validación final del modelo:

1. Entendimiento de los Datos:

- Fuentes de Datos: Las fuentes de datos analizadas incluyeron:
 - Resultados de estudios de laboratorio realizados por el ingeniero Jorge Armando Castro realizadas a través del Laboratorio de Procesos Físico Químicos de la Universidad de Antioquia, que evaluaron propiedades físico-químicas y mecánicas de los suelos y estabilizantes.
 - Datos recopilados de bases científicas y bibliográficas relacionadas con la estabilización vial, asegurando su validez y relevancia para el contexto del estudio.
- Restricciones Asociadas:
 - Los datos utilizados fueron debidamente anonimizados para proteger la privacidad de las entidades o individuos involucrados.
 - Su uso está limitado exclusivamente al desarrollo del proyecto, bajo los términos de confidencialidad y circulación restringida establecidos en la Ley 1581 de 2012 (Colombia), que regula la protección de datos personales.
 - Cualquier transferencia de los datos a terceros, en caso de fuerza mayor, requiere notificación y autorización previa de las partes involucradas.
 - Los datos no pueden ser publicados íntegramente, y cualquier visualización o análisis debe estar en formato agregado o resumido, para evitar la identificación directa o indirecta.
- Análisis Exploratorio de Datos:
 - Se llevó a cabo un análisis exploratorio inicial, que incluyó visualización y evaluación de correlaciones entre las propiedades del suelo y los estabilizantes.
 - Las visualizaciones permitieron identificar patrones relevantes y relaciones clave para la selección de variables del modelo de aprendizaje automático.

2. Preparación de Datos:

- El proceso incluyó limpieza y transformación de los datos recolectados.
- Las características clave fueron seleccionadas utilizando métodos estadísticos y técnicas de importancia de características.

3. Construcción y Optimización del Modelo:

- Se entrenaron algoritmos supervisados como Regresión Logística, Máquinas de Vectores de Soporte, Árboles de Decisión, Random Forest, K-Vecinos Más Cercanos, Extra Trees Classifier y Extreme Gradient Boosting.
- La división del conjunto de datos se realizó en entrenamiento (80%) y prueba (20%), asegurando la reproducibilidad mediante una semilla aleatoria (random_state=42).
- Las características fueron escaladas utilizando StandardScaler para normalizarlas y centrarlas en cero.
- Para balancear las clases en el conjunto de entrenamiento, se aplicó el método SMOTE (Synthetic Minority Oversampling Technique).
- La optimización de hiperparámetros se realizó con RandomizedSearchCV, complementada con validación cruzada para evitar el sobreajuste.
- Se diseñó un meta-modelo ensemble que combinó las fortalezas de diferentes algoritmos.

4. Evaluación y Validación:

- El desempeño del modelo fue evaluado mediante métricas como accuracy, recall, F1-Score y AUC-ROC.

Desarrollo de la Herramienta Informática: La tercera etapa estuvo enfocada en el desarrollo de una herramienta informática que permitiría al usuario interactuar con el modelo de aprendizaje automático mediante una interfaz gráfica. Para esta fase, se aplicó la metodología ágil SCRUM, la cual facilitó un desarrollo iterativo, adaptable y centrado en las necesidades del usuario final.

5. Planeación y Definición de Requisitos:

- Las funcionalidades principales fueron definidas a través de historias de usuario.
- Los criterios de aceptación priorizaron la experiencia de usuario.

6. Iteraciones y Ajustes:

- Los sprints permitieron mejoras incrementales en la funcionalidad.
- Pruebas continuas con usuarios facilitaron ajustes y mejoras en la experiencia de uso.

7. Diseño y Desarrollo de la Arquitectura:

- La arquitectura RESTful fue implementada para desarrollar la aplicación.

- El frontend se diseñó utilizando Bootstrap, con énfasis en la usabilidad y accesibilidad.
- El backend fue desarrollado en Flask, integrándose con el modelo de IA para procesar consultas.
- MongoDB se utilizó para almacenar propiedades de suelos y estabilizantes.

Implementación y Despliegue: En la cuarta etapa, se llevó a cabo la implementación de la solución en un ambiente de producción, garantizando que la herramienta estuviera completamente operativa y lista para ser utilizada en un entorno real.

1. Despliegue en Producción:

- El entorno de producción fue configurado para el despliegue de la aplicación y el modelo de IA.
- Los componentes desarrollados fueron integrados en su totalidad.

2. Pruebas Finales:

- La herramienta fue validada mediante pruebas de funcionalidad, rendimiento y usabilidad.
- Expertos técnicos y temáticos realizaron evaluaciones para asegurar el cumplimiento de estándares requeridos.

3. Mantenimiento y Sostenibilidad:

- Se definieron protocolos para actualizar el modelo con nuevos datos.
- Los lineamientos garantizaron la sostenibilidad de la solución a largo plazo.

La combinación de CRISP-DM y SCRUM fue clave para alcanzar los objetivos del proyecto. CRISP-DM proporcionó un marco sólido para el tratamiento y análisis de datos, asegurando la calidad del modelo de aprendizaje automático. Por su parte, SCRUM brindó flexibilidad y eficiencia durante el desarrollo de la herramienta informática, permitiendo iteraciones rápidas y adaptaciones según las necesidades identificadas.

Esta convergencia metodológica optimizó el desarrollo, minimizó riesgos y garantizó la capacidad de adaptación a los cambios durante la ejecución del proyecto. Como resultado, se logró una solución integrada que cumple con los requerimientos técnicos, funcionales y de investigación.

5. RESULTADOS DE LA SOLUCIÓN

En esta sección se presentan los resultados obtenidos durante el desarrollo e implementación del proyecto. Estos se dividen en tres áreas clave: la construcción del modelo de aprendizaje automático, el desarrollo de la herramienta informática, y la validación y despliegue de la solución. Cada área incluye un análisis detallado que evidencia cómo los objetivos fueron alcanzados de manera efectiva.

5.1. ENTENDIMIENTO DEL NEGOCIO

Durante la etapa de entendimiento del negocio, se identificaron varios desafíos clave relacionados con la selección de estabilizantes para suelos viales. Uno de los principales problemas radica en la variabilidad significativa de resultados entre muestras de un mismo tipo de suelo, lo que complica la elección del estabilizante óptimo. Este reto se agrava debido a la fuerte dependencia del sector en técnicas tradicionales, basadas más en la experiencia y la tradición que en enfoques modernos e innovadores, lo que a menudo limita la exploración de alternativas más efectivas.

En respuesta a estas dificultades, el proyecto se planteó como objetivo desarrollar una solución tecnológica basada en ciencia de datos. Esta herramienta busca optimizar la toma de decisiones, promoviendo no solo la eficiencia y precisión en la selección de estabilizantes, sino también la sostenibilidad de las infraestructuras viales. El modelo diseñado tiene como propósito identificar las propiedades fisicoquímicas y mecánicas de los suelos que influyen en esta elección, garantizando un nivel mínimo de confiabilidad del 80%. Además, se estableció como requerimiento técnico que los tiempos de respuesta de la herramienta no superen los tres segundos y que el modelo sea replicable para al menos dos tipos diferentes de suelos.

Esta solución, además de ser accesible y eficiente, responde a la necesidad de superar las limitaciones actuales, integrando estándares técnicos y prácticos que aseguren su aplicabilidad en contextos reales. Con ello, se espera no solo facilitar el trabajo de los especialistas, sino también contribuir significativamente a la mejora de la calidad y sostenibilidad de las infraestructuras viales en el mediano y largo plazo.

5.2. DISEÑO DEL MODELO DE APRENDIZAJE AUTOMÁTICO

- *Entendimiento de los Datos:* Para garantizar una base sólida para el modelo, se analizaron 293 registros que incluían 17 variables relacionadas con propiedades físico-químicas y mecánicas del suelo, además de clasificaciones de estabilizantes, como se ilustra en la *Tabla 1*.

Tabla 1: Descripción de propiedades de suelos y estabilizantes, Fuente: propia

Identificador	Variables	Tipo variable	Tipo de Dato	Descripción
1	% Malla 200	Característica	Número - Decimal	Porcentaje de partículas del suelo que pasan por un tamiz de 200 mallas, indicando la proporción de material fino en la muestra.
2	% Límite Líquido	Característica	Número - Decimal	Contenido de agua en el que el suelo pasa de un estado plástico a uno líquido, utilizado para clasificar su plasticidad y comportamiento.
3	% Límite Plástico	Característica	Número - Decimal	Contenido de agua mínimo en el que el suelo mantiene un estado plástico antes de romperse, asociado a su consistencia.
4	Índice de Plasticidad	Característica	Número - Decimal	Diferencia entre el Límite Líquido y el Límite Plástico, indicando la plasticidad relativa del suelo.
5	Clasificación SUCS	Característica	Cadena	Clasificación del suelo según el Sistema Unificado de Clasificación de Suelos (SUCS), que agrupa materiales según su granulometría y plasticidad.
6	% de CBR	Característica	Número - Decimal	Índice que representa la capacidad del suelo para soportar cargas en relación con un material estándar, utilizado para evaluar su idoneidad en aplicaciones de construcción y pavimentos.
7	Resistencia a la Compresión	Característica	Número - Decimal	Capacidad del suelo para resistir cargas de compresión antes de fallar, clave para evaluar su estabilidad.
8	Modulo resiliente	Característica	Número - Decimal	Capacidad del suelo para deformarse de manera elástica bajo cargas repetidas, esencial para evaluar su desempeño en estructuras como pavimentos.
9	% Humedad Natural	Característica	Número - Decimal	Contenido de agua presente en el suelo de manera natural, relacionado con su estado inicial y características antes de ser sometido a procesos de estabilización.
10	% Humedad Óptima	Característica	Número - Decimal	Contenido de agua en el que el suelo alcanza su máxima densidad al ser compactado, utilizado para diseñar procesos de estabilización.
11	Densidad Máxima	Característica	Número - Decimal	Máxima densidad seca que un suelo puede alcanzar mediante compactación, directamente relacionada con su resistencia y estabilidad.
12	pH	Característica	Número - Decimal	Medida de acidez o alcalinidad del suelo, utilizada para identificar las condiciones químicas y su influencia en la interacción con los estabilizantes.
13	Composición mineralógica	Característica	Cadena	Distribución de los minerales presentes en el suelo, clave para comprender su comportamiento mecánico y químico frente a procesos de estabilización.
14	% Contenido de materia orgánica	Característica	Número - Decimal	Proporción de materiales orgánicos presentes en el suelo, que puede influir en sus propiedades de compactación y estabilidad química.
15	Tipo de Estabilización	Característica	Cadena	Clasificación de la estabilización según el uso en la Industria (tradicionales y no tradicionales)
16	Clasificación de Estabilización Inicial	Característica	Cadena	Clasificación de la estabilización según el Mecanismo de Acción con el suelo (cementantes, electroquímicos y físicos)
17	Clasificación de Estabilización Final	Objetivo	Cadena	Clasificación de la estabilización según la Familia Físico-Química del estabilizante (materiales cementantes y puzolánicos, nanomateriales, materiales orgánicos y minerales)

Estas propiedades fueron procesadas y analizadas con el propósito de identificar relaciones clave mediante técnicas como el análisis exploratorio descriptivo (*ver la Tabla 2*), la identificación de patrones y correlaciones iniciales entre características relevantes (*ver la Figura 11*) y el análisis de correlación visualizado a través de un mapa de calor (*ver la Figura 12*), con el objetivo de evaluar la variabilidad y consistencia de los datos recopilados para garantizar su integridad y utilidad en la selección de estabilizantes óptimos.

La estadística descriptiva (*ver la Tabla 2*) evidencia una notable variabilidad en las propiedades del suelo, como se observa en la desviación estándar de las variables analizadas. Por ejemplo, el % de malla 200 presenta un valor promedio de 72.58 con una desviación estándar de 22.548, mientras que el índice de plasticidad alcanza una media de 33.635 con una desviación de 63.64, indicando una alta dispersión. Estos resultados reflejan la complejidad inherente en la selección de estabilizantes óptimos, destacando la necesidad de enfoques detallados y específicos para abordar la heterogeneidad de los datos.

Tabla 2: Estadísticas descriptivas de los datos recopilados, Fuente: propia

	% Malla 200	% Limite Líquido	% Limite Plástico	Índice de Plasticidad	CBR	Resistencia Compresión	Modulo Resiliente	% Humedad Natural	% Humedad Optima	Densidad Máxima	% Materia Orgánica	pH
mean	72,58	60.127	26.075	33.635	10.758	30.017	69.669	13.688	19.594	16.601	2.835	7.194
median	75	46,5	24,85	21	5,1	2,4	49,300	10,2	18,5	16,5	3	7,5
std	22.548	67,37	8.271	63,64	13,07	73.409	54.184	10.218	6.057	1.819	1.221	1.276

La *Figura 11*, refleja la distribución porcentual de materiales utilizados en estabilización de suelos, destacando a los nanomateriales como los más empleados (44.7%), seguidos de los materiales cementantes y puzolánicos (29.4%), que mantienen su relevancia por su eficacia y disponibilidad. Los materiales orgánicos (22.5%) representan un enfoque hacia la sostenibilidad, mientras que los minerales (3.4%) tienen un uso más limitado, posiblemente debido a restricciones en disponibilidad o propiedades. En general, se evidencia un equilibrio entre la adopción de tecnologías avanzadas y el uso de métodos tradicionales, con un interés creciente en soluciones sostenibles.

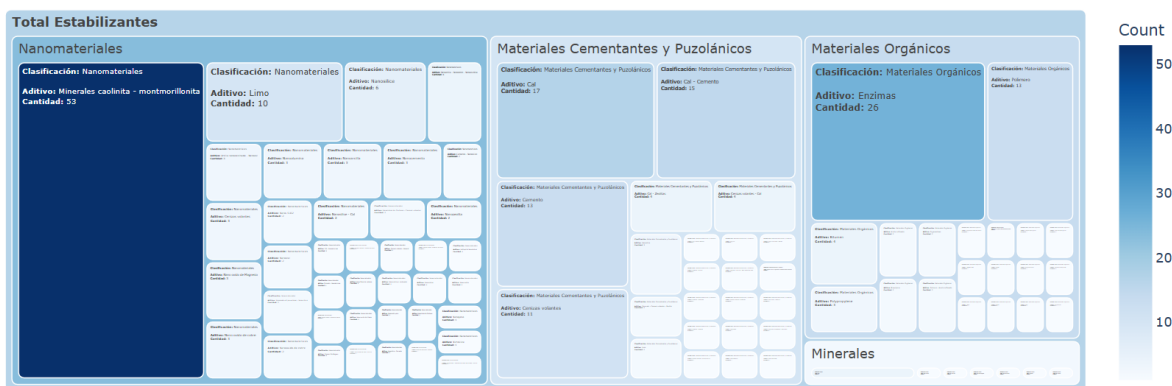


Figura 11: Caracterización de los suelo según tipo de Estabilización, Fuente: propia

El mapa de calor de la *Figura 12*, muestra las correlaciones entre variables fisicoquímicas y mecánicas del suelo. Destacando una fuerte relación positiva entre el índice de plasticidad y el límite líquido (0.99), lo que indica que su incremento está altamente relacionado. Asimismo, variables como el CBR y el módulo resiliente también presentan una correlación significativa (0.99), reflejando su conexión en términos de resistencia mecánica. En contraste, otras variables como el contenido de materia orgánica y el pH exhiben correlaciones débiles o negativas con la mayoría de las variables, sugiriendo una menor influencia directa.

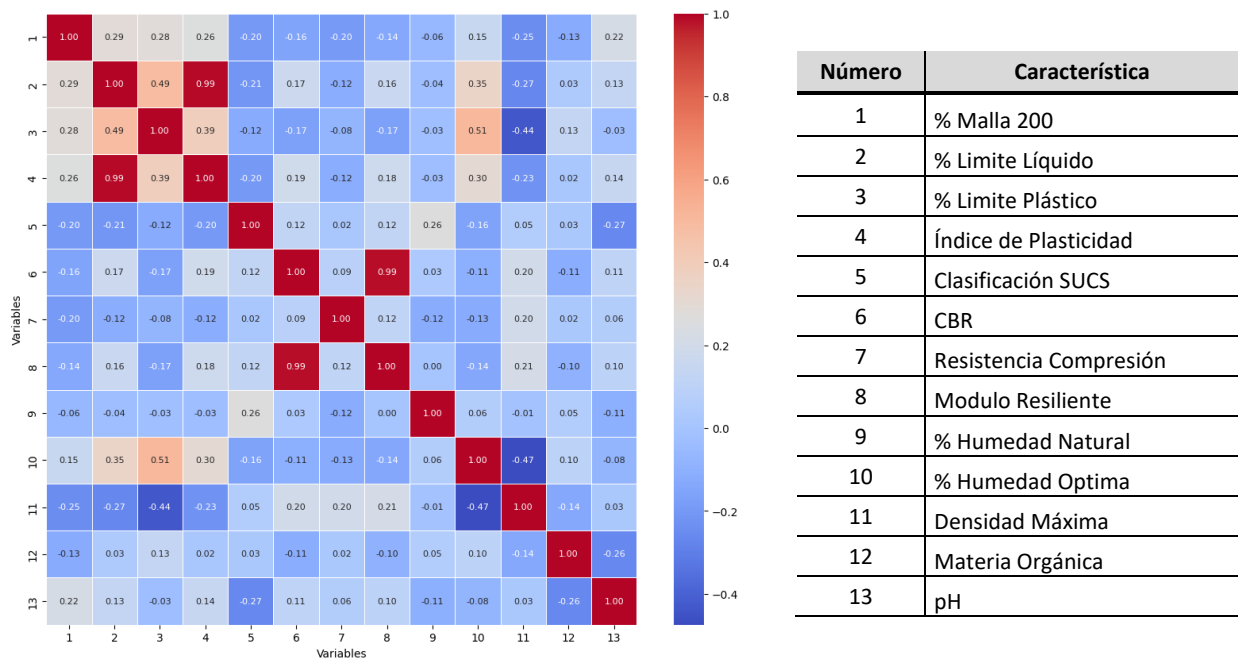


Figura 12: Mapa de calor que muestre la correlación entre características, Fuente: propia

- Preparación de Datos:** Con base en los hallazgos obtenidos durante la etapa de entendimiento de los datos, se procedió a su preparación. Este proceso incluyó una limpieza exhaustiva, eliminando registros duplicados y completando valores faltantes mediante métodos de imputación basados en patrones observados, lo que redujo el total de registros de 293 a 290. Posteriormente, se realizó la normalización de las variables, abordando los valores atípicos mediante el uso del *Rango Intercuartílico (IQR)* combinado con la técnica de reemplazo, utilizando la *mediana* como valor de imputación para dichos valores. Además, las características categóricas fueron codificadas utilizando el método *OrdinalEncoder*. Durante este proceso, las características iniciales se redujeron de 17 a 11, seleccionándose únicamente las variables más relevantes para los modelos supervisados de clasificación, ver la *Tabla 3*. Esta reducción no solo optimizó el procesamiento, sino que también mejoró la interpretación de los resultados al eliminar redundancias y concentrarse en atributos clave.

Tabla 3: Características finales seleccionadas para los modelos, Fuente: propia

Variabes	Tipo variable	Tipo de Dato	Codificación
% Malla 200	Característica	Número - Decimal	
% Límite Líquido	Característica	Número - Decimal	
% Límite Plástico	Característica	Número - Decimal	
Índice de Plasticidad	Característica	Número - Decimal	
Clasificación SUCS	Característica	Número - Codificada	0 - CH 1 - CL 2 - MH 3 - ML 4 - OH 5 - OL 6 - SC 7 - SM 8 - SP
Resistencia a la Compresión	Característica	Número - Decimal	
% Humedad Óptima	Característica	Número - Decimal	
Densidad Máxima	Característica	Número - Decimal	
Tipo de Estabilización	Característica	Número - Codificada	0 - No tradicional 1 - Tradicional
Clasificación de Estabilización Inicial	Característica	Número - Codificada	0 - Cementantes 1 - Electroquímicos 2 - Físicos
Clasificación de Estabilización Final	Objetivo	Cadena	

- *Construcción y Optimización del Modelo:* Se seleccionaron siete algoritmos de aprendizaje supervisado, los cuales representan diversos enfoques para resolver problemas de clasificación. Estos modelos aseguran una cobertura representativa y balanceada de técnicas fundamentales y avanzadas en clasificación, facilitando un análisis exhaustivo y comparativo sobre el desempeño en el problema planteado. Las características clave de cada modelo y sus diferencias relevantes se presentan en la *Tabla 4:*

Tabla 4: Comparación de los Modelos de Aprendizaje Supervisado Seleccionados, Fuente: propia

Modelo	Paradigma	Ventajas	Limitaciones	Características Relevantes
Regresión Logística (LR)	Lineal	- Simple e interpretable	- No maneja relaciones no lineales	- Adecuado para conjuntos de datos pequeños
		- Eficiente en problemas lineales	- Sensible a multicolinealidad en los datos	- Propenso a subajuste si se presenta una no linealidad
		- Rápido de entrenar		

Máquinas de Vectores de Soporte (SVM)	Basado en márgenes	- Eficiente en alta dimensionalidad	- Costoso computacionalmente	- Funciona bien con conjuntos de datos pequeños o medianos
		- Uso de kernels para problemas no lineales	- Difícil de interpretar	- Sensible a ruido en los datos
Árboles de Decisión (DT)	Basado en reglas	- Fácil de interpretar	- Propenso a sobreajuste	- Escalable a conjuntos de datos grandes
		- Captura relaciones no lineales	- Menor precisión en problemas complejos	- Sensible a datos desbalanceados
		- No requiere escalado		
Random Forest (RF)	Ensamble basado en árboles	- Robusto al sobreajuste	- Requiere más recursos	- Escalable a conjuntos de datos grandes
		- Maneja datos faltantes	- Menor interpretabilidad que un solo árbol	- Menos propenso a sobreajuste que DT
		- Adecuado para problemas complejos		
K vecino más cercano (KNN)	Basado en proximidad	- No paramétrico	- Sensible al ruido	- Ideal para conjuntos de datos pequeños
		- Intuitivo y fácil de implementar	- Ineficiente en datasets grandes	- Muy sensible al ruido y escalado de datos
		- Captura relaciones complejas		
Extra Trees Classifier (ET)	Ensamble basado en árboles	- Mayor rapidez que RF	- Similitud alta con RF	- Escalable a conjuntos de datos grandes
		- Alta capacidad de generalización	- Menor interpretabilidad	- Robusto contra sobreajuste
Extreme Gradient Boosting (XGB)	Boosting	- Excelente precisión	- Más complejo de ajustar	- Excelente para conjuntos de datos grandes
		- Manejo de desbalance de clases	- Requiere más tiempo de entrenamiento	- Resistente al sobreajuste
		- Optimización eficiente		

Para garantizar la validez y confiabilidad de las predicciones de los modelos; los datos fueron preparados cuidadosamente. Este proceso incluyó la separación de las características independientes (X) y la variable objetivo (y), así como la división del conjunto en entrenamiento (80%) y prueba (20%), asegurando la reproducibilidad mediante una semilla aleatoria ($random_state=42$). Se empleó *StandardScaler* para escalar las características y normalizar los datos, un paso crucial para evitar sesgos en modelos sensibles a la magnitud, como *SVM* y *KNN*. Además, se aplicó la técnica *SMOTE* (*Synthetic Minority Oversampling Technique*) para abordar el desequilibrio de clases del conjunto de entrenamiento, generando muestras sintéticas de la clase minoritaria mediante combinaciones lineales de puntos existentes, con el parámetro $k_neighbors$ ajustado a 3. Estos pasos permitieron construir un conjunto de datos consistente y representativo, optimizando la capacidad predictiva de los modelos empleados en el estudio.

Para optimizar los hiperparámetros de cada modelo, se utilizó el método *RandomizedSearchCV*, (ver la *Tabla 5*). Este método fue seleccionado por su alta eficiencia computacional, ya que evalúa un número limitado de combinaciones de hiperparámetros previamente definidos, adaptándose a los recursos disponibles. Los resultados obtenidos muestran que el modelo *Random Forest* (RF) presentó el mejor desempeño, con una puntuación de 0.908, mientras que el modelo con menor rendimiento fueron las *Máquinas de vectores de soporte* (SVM), con una puntuación de 0.7471.

Tabla 5: Hiperparámetros óptimos y mejor puntuación de rendimiento, Fuente: propia

Modelo	Hiperparámetros	Mejor resultado (Rendimiento)	Tiempo de computo (segundos) ¹
Árboles de decisión (DT)	ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=14, max_features='log2', max_leaf_nodes=None, min_impurity_decrease=0, min_samples_leaf=2, min_samples_split=7, min_weight_fraction_leaf=0.0, monotonic_cst=None, random_state=123, splitter='best'	0.7931	0.0320
Regresión logística (LR)	C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=1000, multi_class='auto', n_jobs=None, penalty='l2', random_state=123, solver='lbfgs', tol=0.0001, verbose=0, warm_start=False	0.7701	0.7360
Máquinas de vectores de soporte (SVM)	C=1e7, tol=0.001, max_iter=1000, random_state=123, verbose=True	0.7471	0.0390
Random Forest (RF)	bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='sqrt', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, monotonic_cst=None, n_estimators=100, n_jobs=-1, oob_score=False, random_state=123, verbose=0, warm_start=False	0.908	0.2060
K vecino más cercano (KNN)	algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=-1, n_neighbors=9, p=2, weights='distance'	0.8506	0.0470
Extra Trees Classifier (ET)	bootstrap=False, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='sqrt', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, monotonic_cst=None, n_estimators=100, n_jobs=-1, oob_score=False, random_state=123, verbose=0, warm_start=False	0.8851	0.1810
Extreme Gradient Boosting (XGB)	objective='multi:softprob', base_score=None, booster='gbtree', callbacks=None, colsample_bylevel=None, colsample_bynode=None, colsample_bytree=1, device='cpu', early_stopping_rounds=None, enable_categorical=False, eval_metric=None, feature_types=None, gamma=None, grow_policy=None, importance_type=None, interaction_constraints=None, learning_rate=0.5, max_bin=None, max_cat_threshold=None, max_cat_to_onehot=None, max_delta_step=None, max_depth=1, max_leaves=None, min_child_weight=1, missing='nan', monotone_constraints=None, multi_strategy=None, n_estimators=100, n_jobs=-1, num_parallel_tree=None, random_state=123, reg_alpha=0.0005, reg_lambda=0.5, sampling_method=None, scale_pos_weight=35.1, subsample=1, tree_method='auto', validate_parameters=None, verbosity=0	0.8621	0.1110

¹ El proceso fue realizado en un equipo de cómputo con un procesador Intel Core i7 de 11ª generación, 16 GB de RAM LPDDR4X, almacenamiento SSD NVMe y capacidad de paralelización multinúcleo.

Para evaluar la robustez de los modelos, se realiza una validación cruzada con k-folds (*número de particiones: k=5*). Este método dividió los datos de entrenamiento en k subconjuntos, entrenando el modelo en k-1 subconjuntos y validándolo en el subconjunto restante.

En la *Figura 13*, se ilustran las curvas de aprendizaje de cada modelo, donde se permite evidenciar el comportamiento progresivo en su rendimiento a medida que aumentaba el tamaño del conjunto de entrenamiento. Por ejemplo, mientras que el modelo basado en Árboles de Decisión alcanzó una alta precisión en el conjunto de entrenamiento, el rendimiento en validación fue moderado, reflejando una tendencia al sobreajustarse. En contraste, modelos como el XGBoost presentaron una mejora más gradual y consistente tanto en el conjunto de entrenamiento como en validación, lo que evidencia una mejor capacidad de generalización con datos adicionales. Estas diferencias resaltan cómo las características de cada modelo influyen en el rendimiento a medida que aumenta el tamaño del conjunto de datos.

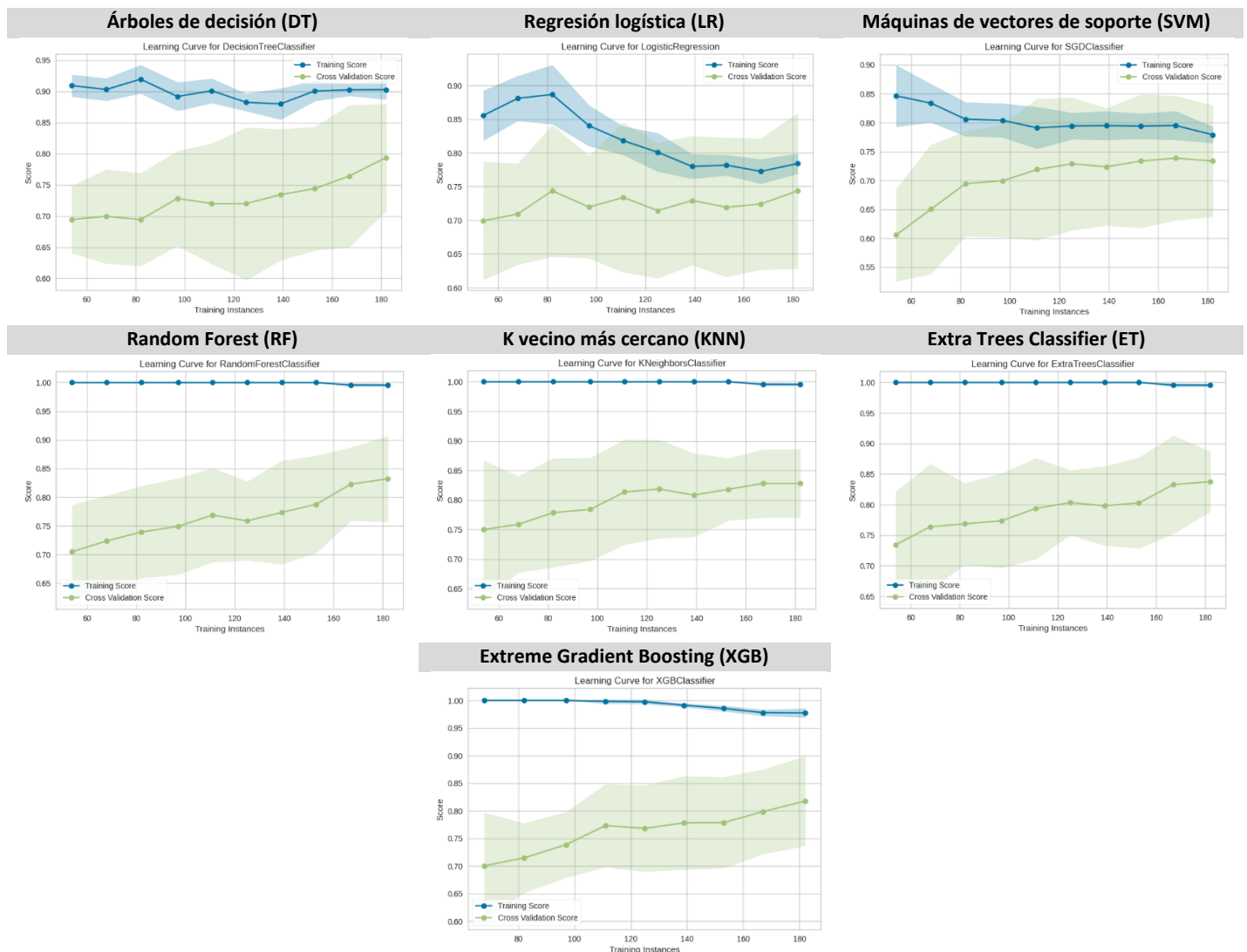
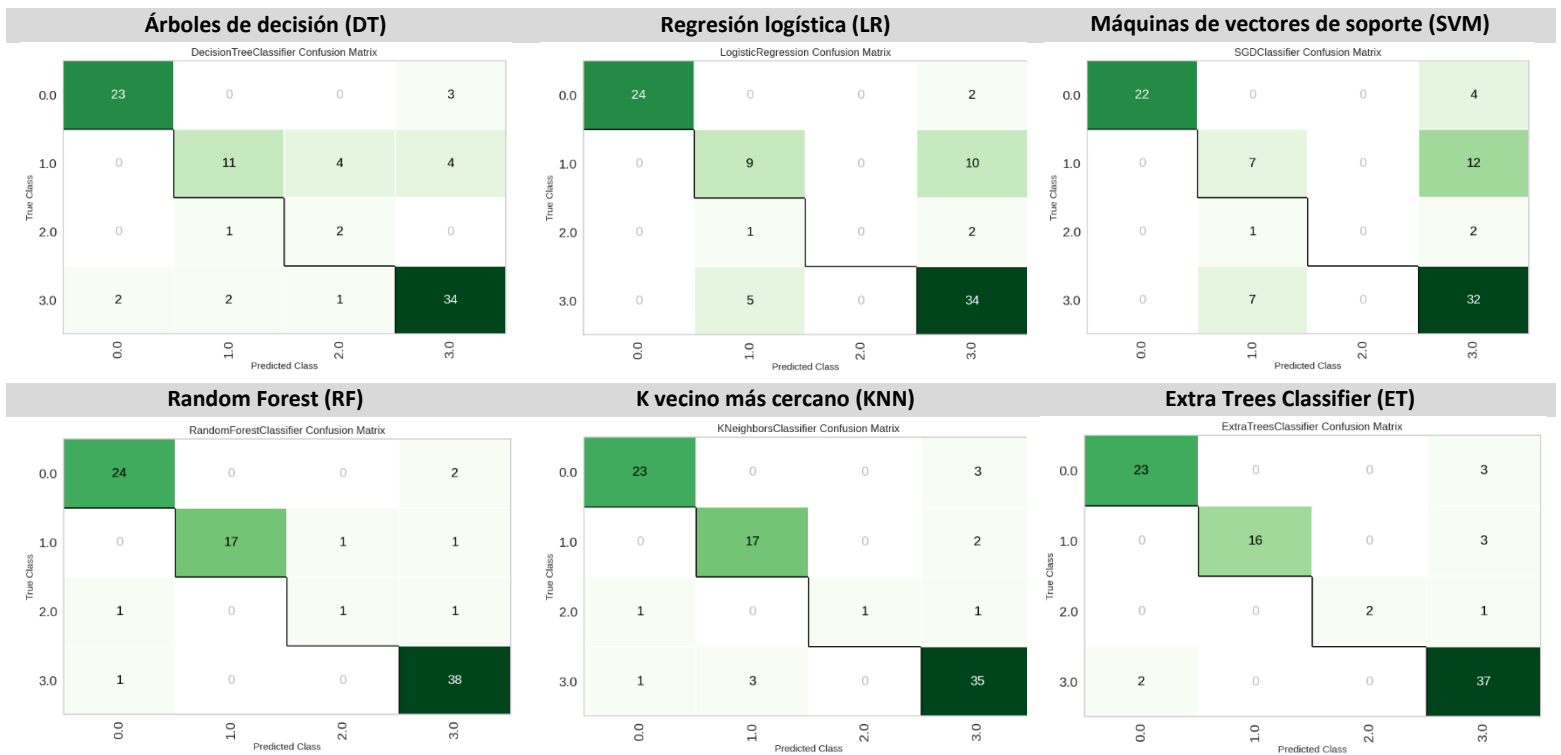


Figura 13: Curvas de aprendizaje y validación de los modelos de Clasificación, Fuente: Propia

Por otra parte, la validación cruzada permitió analizar el rendimiento promedio de los modelos en las particiones creadas (ver la *Figura 13*). En particular, el modelo *XGBoost* se destacó por su desempeño uniforme a través de cada iteración realizada en los subconjuntos, mostrando menor variabilidad entre los subconjuntos de validación, lo que lo posicionó como un modelo robusto frente a los cambios en los datos de entrada. Por otro lado, el modelo basado en *Random Forest* mostró un comportamiento competitivo, con una mayor dispersión en el rendimiento entre las particiones.

- *Evaluación y Validación*: El desempeño del modelo fue evaluado mediante métricas como accuracy, recall, F1-Score y AUC-ROC.
 - *Matriz de Confusión*: En la *Figura 14*, se comparan los modelos evaluados, destacándose *XGBoost* y *Random Forest* por su desempeño superior. Ambos modelos presentan altas tasas de aciertos, claramente visibles en las diagonales de sus matrices de confusión, y muestran una baja variabilidad en los errores. En particular, *XGBoost* se distingue como el modelo más robusto, manteniendo un comportamiento consistente frente a cambios en los datos. Por el contrario, modelos como *SVM* y *Árboles de Decisión* exhiben mayor dispersión en sus predicciones, lo que refleja una menor precisión y estabilidad. Estos resultados consolidan a *XGBoost* y *Random Forest* como las opciones más efectivas para la clasificación de estabilizantes.



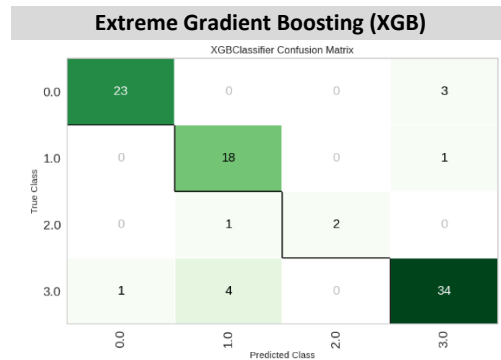
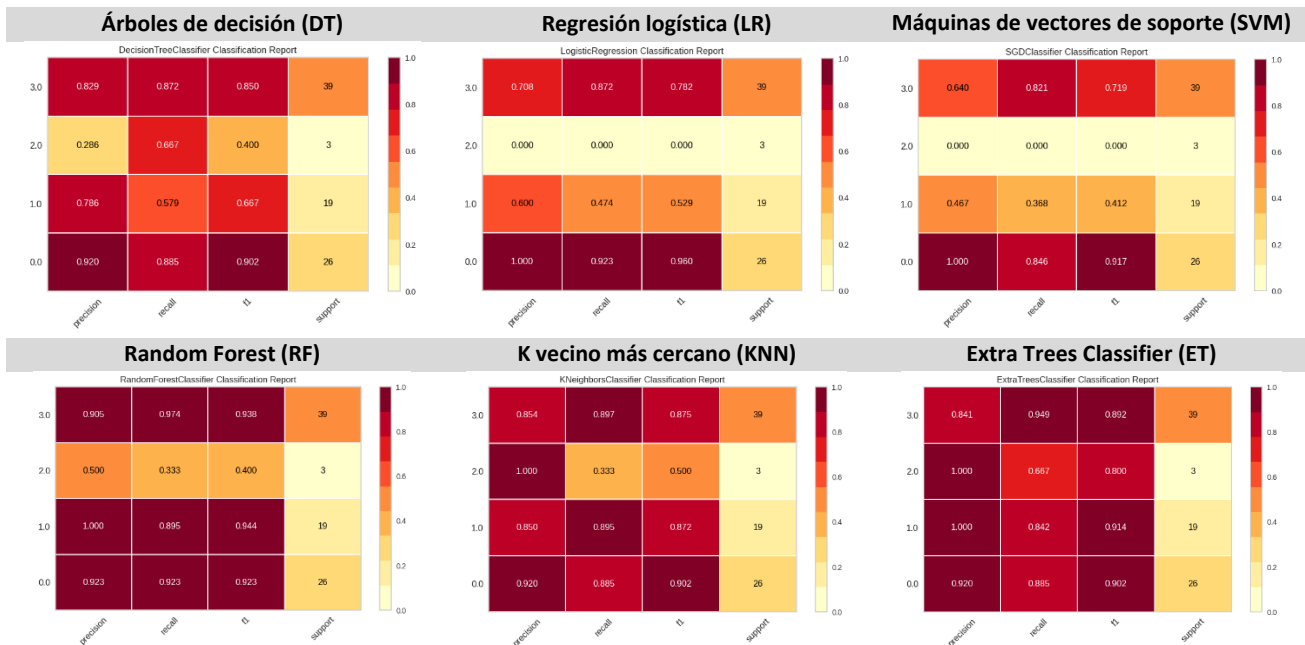


Figura 14: Comparación de Matrices de Confusión entre los Modelos de Clasificación, Fuente: propia

- **Precisión, recall y f1-score:** Entre los modelos comparados en la Figura 15, el modelo *Random Forest* destaca por su equilibrio en las métricas, obteniendo altos valores de precisión y F1-score, lo que refleja su capacidad para realizar predicciones consistentes en todas las clases. Por otra parte, el modelo *XGBoost* presenta un rendimiento similar, especialmente en la clase mayoritaria, pero muestra una ligera disminución en la clase con menor soporte, lo que podría indicar menor robustez en casos minoritarios. En contraste, el modelo de *SVM* exhibe los valores más bajos en todas las métricas, especialmente en recall y F1, lo que evidencia un desempeño más limitado en la clasificación correcta de las instancias. Modelos como *KNN* y *ExtraTreesClassifier* ofrecen un rendimiento competitivo, destacándose en precisión, pero con variabilidad en el recall según la clase. En general, los modelos de *Random Forest* y *XGBoost* sobresalen como los modelos más efectivos para este conjunto de datos, especialmente por su alta consistencia en las métricas clave.



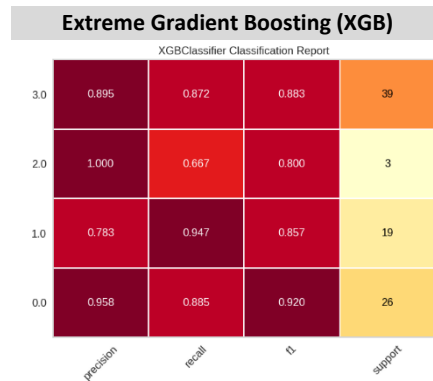


Figura 15: Comparación de Métricas: precisión, recall y f1-score, entre los Modelos de Clasificación, Fuente: propia

- **Accuracy:** Entre los modelos comparados en la Figura 16, el modelo *Random Forest (RF)* se destaca con una accuracy de 0.908, seguido por *ExtraTreesClassifier (ET)* y *XGBoost (XGB)*, con valores de 0.8851 y 0.8621, respectivamente. Estos modelos, basados en técnicas de ensemble learning, evidencian su superioridad frente a algoritmos como *Máquinas de Vectores de Soporte (SVM)*, *Regresión Logística (LR)* y *Árboles de Decisión (DT)*, que alcanzan precisiones menores de 0.7471, 0.7701 y 0.7931, respectivamente. Por otro lado, el modelo de *K-vecinos más cercanos (KNN)* obtiene un desempeño intermedio con una exactitud de 0.8506, superando a los modelos lineales y basados en árboles individuales. Este análisis resalta la capacidad de los algoritmos avanzados y de ensamblado para capturar relaciones complejas en los datos, proporcionando una mayor generalización y rendimiento en tareas de clasificación.

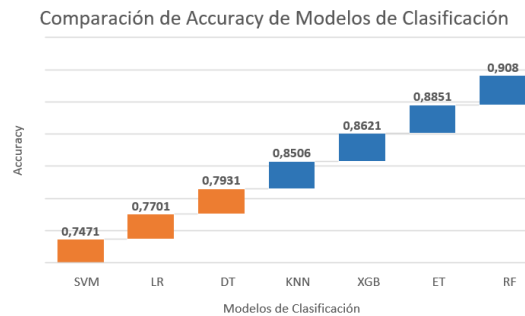


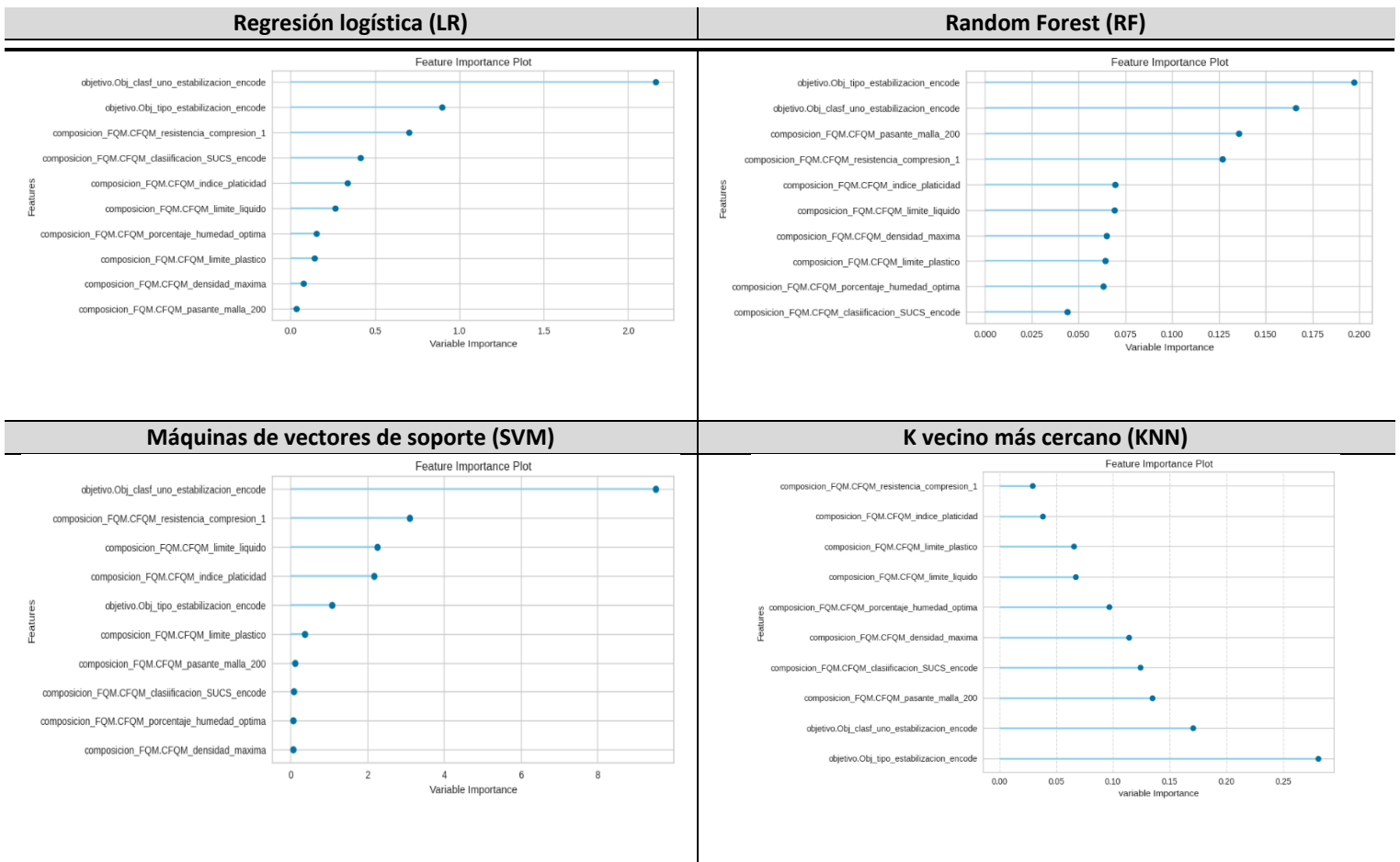
Figura 16: Comparación de Accuracy de Modelos de Clasificación, Fuente: propia

Para finalizar, la evaluación interna de los modelos destaca la superioridad de los algoritmos basados en técnicas de ensemble learning, específicamente *Random Forest* y *XGBoost*, que consistentemente obtienen los mejores resultados en métricas clave como *precisión, recall, F1-score* y *accuracy*. Ambos modelos no solo demuestran una alta capacidad para capturar patrones complejos en los datos, sino también una notable estabilidad en la clasificación de instancias, incluso frente a la variabilidad en las clases. En particular, *Random Forest* se distingue por su equilibrio entre las métricas, mientras que *XGBoost* resalta por su

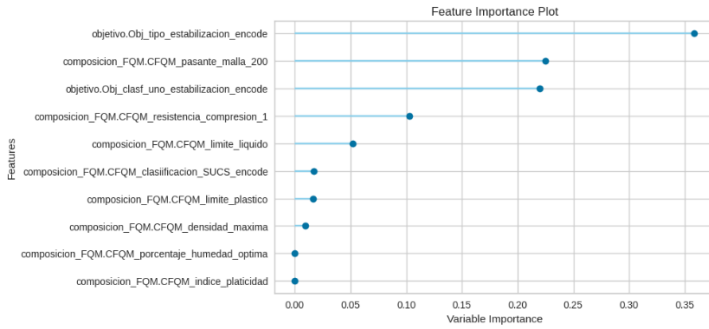
robustez en escenarios con cambios en los datos. Por otro lado, algoritmos como *Máquinas de Vectores de Soporte* y *Árboles de Decisión*, presentan limitaciones en su capacidad predictiva, con valores más bajos en todas las métricas y una mayor dispersión en sus predicciones.

- *Importancia de las características:* En la *Tabla 6*, ilustra como los modelos basados en árboles, como Random Forest (RF), Árboles de Decisión (DT), Extra Trees (ET) y Extreme Gradient Boosting (XGB), priorizan características categóricas transformadas relacionadas con la clasificación del estabilizante según el uso en la industria y el mecanismo de acción del estabilizante, debido a su capacidad para segmentar datos de manera efectiva. Además, estas técnicas otorgan importancia a variables físicas clave, como la granulometría del suelo, el índice de plasticidad y el límite líquido, mostrando su capacidad para integrar tanto atributos numéricos como categóricos. En particular, XGB destaca por balancear ambas dimensiones, capturando relaciones complejas entre las variables.

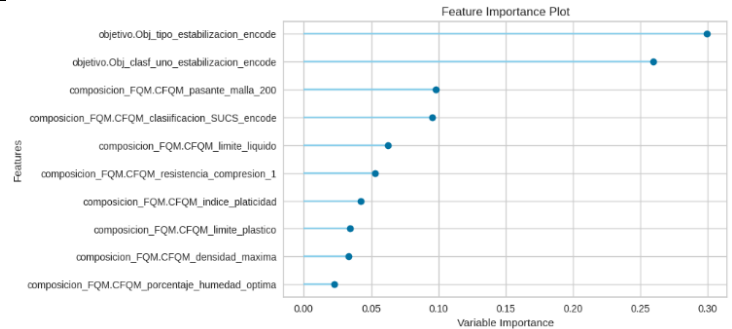
Tabla 6: Comparación de la importancia de las características en cada Modelo, Fuente: propia



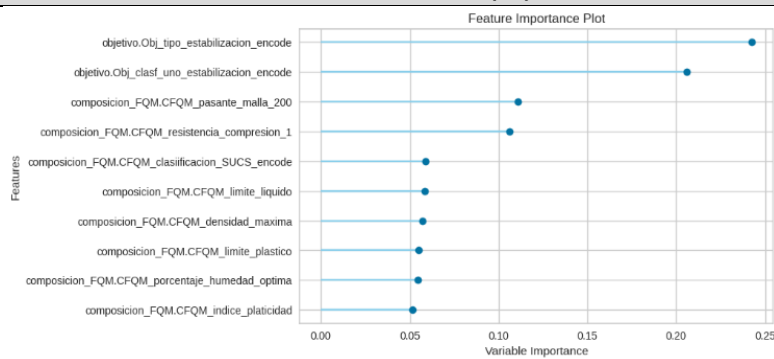
Árboles de decisión (DT)



Extreme Gradient Boosting (XGB)



Extra Trees Classifier (ET)



- Meta modelo ensemble learning:** Apoyado en las fortalezas individuales de los modelos y los resultados de la evaluación y validación, se desarrolló un meta modelo de ensamble combinando los algoritmos más destacados: *K-vecinos más cercanos* (KNN), *XGBoost*, *ExtraTreesClassifier* (ET) y *Random Forest* (RF). Este modelo utiliza un enfoque de votación (*Voting Ensemble*), en el que cada algoritmo contribuye a las predicciones finales, maximizando la precisión y robustez en los resultados. La combinación de estos modelos permite aprovechar sus características complementarias, como la capacidad de *XGBoost* y *Random Forest* para manejar datos complejos, la simplicidad y efectividad local de *KNN*, y la reducción de sobreajuste aportada por *ExtraTreesClassifier*.

El aprendizaje en ensamble permite mitigar sesgos y variancias presentes en los modelos individuales, logrando predicciones más estables y confiables. Este enfoque es especialmente relevante para la clasificación de estabilizantes, donde se requiere un alto nivel de precisión para garantizar decisiones óptimas. La integración de estos modelos complementarios asegura un sistema predictivo robusto, capaz de adaptarse a la heterogeneidad y variabilidad de los datos, proporcionando así una solución efectiva y confiable para el problema planteado.

A continuación, se presenta el esquema del modelo ensamble construido:

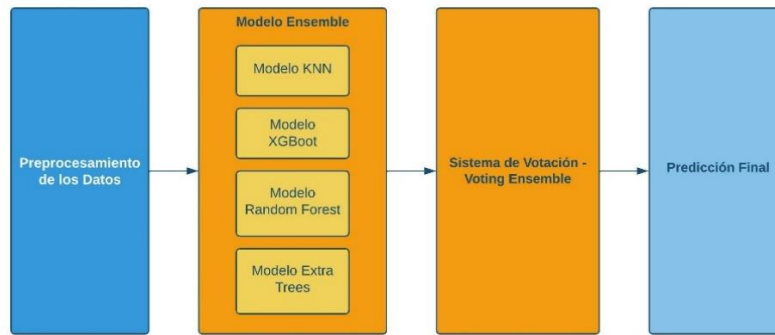
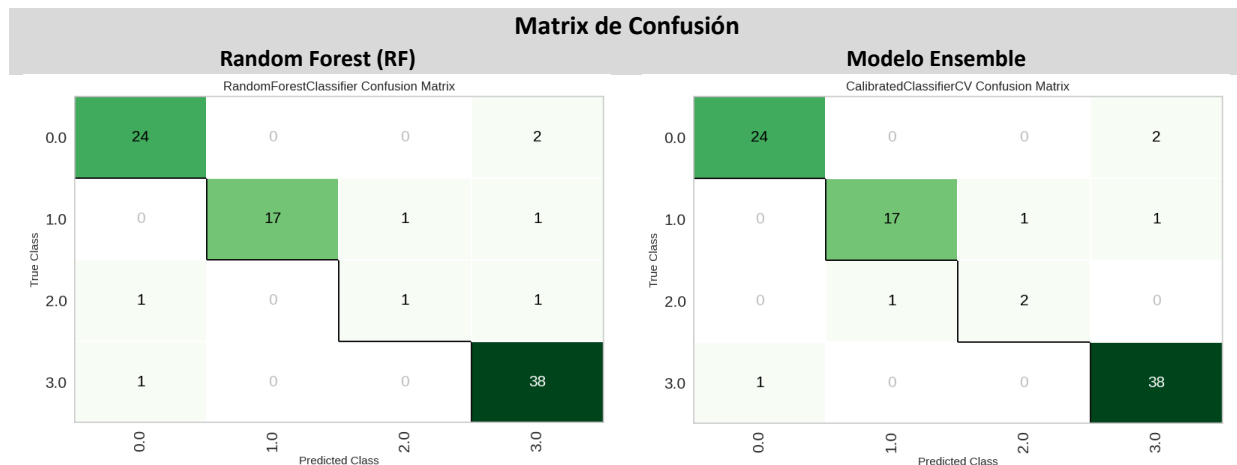


Figura 17: Modelo Ensemble para la Clasificación de Estabilizantes, Fuente: propia

El modelo de ensemble que se ilustra en la *Figura 17*, se basa en un enfoque de votación homogénea (Hard Voting), donde cada modelo base tiene la misma relevancia en la toma de decisiones. Esto significa que la clase final predicha corresponde a aquella que recibe la mayoría de votos entre los modelos base. Este esquema, al no asignar pesos explícitos, garantiza un equilibrio en la combinación de las predicciones, priorizando la simplicidad y la robustez.

El proceso inicia con el preprocesamiento de los datos, el cual incluye la normalización y el balanceo del conjunto de datos, asegurando la calidad y homogeneidad necesarias para alimentar los modelos base. Posteriormente, cada modelo individual realiza sus predicciones, que son combinadas en el sistema de votación para generar una predicción final. Este enfoque permite aprovechar las fortalezas de los distintos algoritmos, reduciendo el impacto de sus limitaciones individuales y proporcionando un sistema más robusto para la clasificación de estabilizantes.

Para evaluar la efectividad del modelo ensemble, se compararon sus resultados con los obtenidos por el mejor modelo individual, identificado como *Random Forest*. Las métricas clave de evaluación incluyen la matriz de confusión, precisión, recall, F1-score y accuracy. En la *Figura 18*, se ilustra la comparación de métricas de evaluación y validación entre ambos modelos:



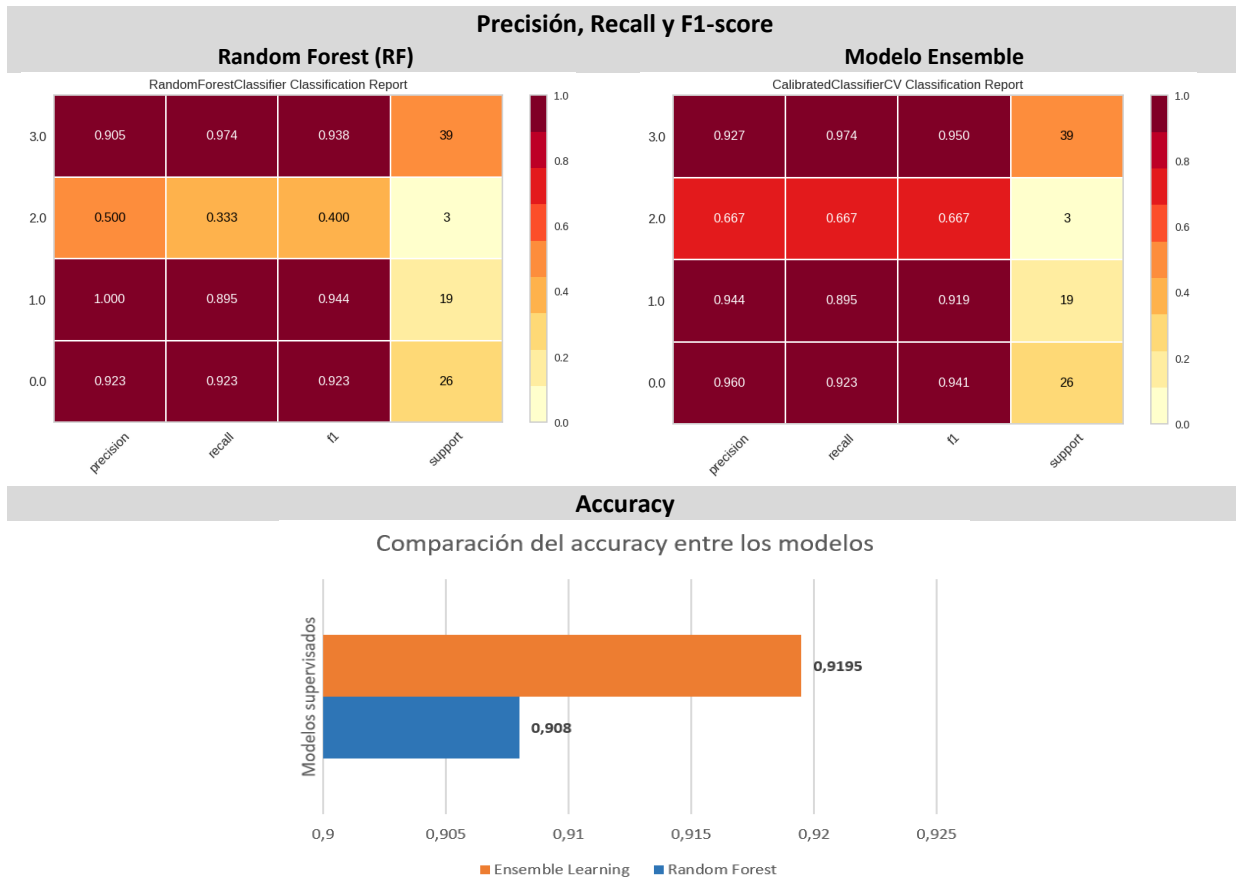


Figura 18: Comparación de Resultados entre el Mejor Modelo Individual (Random Forest) y el Modelo Ensemble, Fuente: propia

El modelo ensemble evidenció un desempeño superior al modelo *Random Forest* al ser evaluado en múltiples métricas clave. En términos de accuracy, el *modelo ensemble* alcanzó un valor de 0.9195, superando al *modelo Random Forest*, que obtuvo un accuracy de 0.908, lo que representa una mejora del 1.15%.

Adicionalmente, en las métricas de precisión y recall, el modelo *Random Forest* mostró ciertas limitaciones al clasificar las clases minoritarias, mientras que el *modelo ensemble* logró un mejor balance en estas métricas, destacándose por su robustez en la clasificación de todas las clases.

Por último, el *modelo ensemble* presentó un F1-score promedio más consistente en comparación con *Random Forest*, reflejando una mejora en el balance entre precisión y recall, lo que refuerza su efectividad como sistema más equilibrado y confiable para esta tarea.

El análisis comparativo evidencia que el modelo ensemble, al combinar las predicciones de múltiples algoritmos base, ofrece un sistema más robusto y preciso para la clasificación de estabilizantes.

Aunque el modelo Random Forest destacó como el mejor modelo individual, el ensemble superó su desempeño, proporcionando un enfoque confiable y generalizable para la tarea planteada.

5.3. DESARROLLO DE LA HERRAMIENTA INFORMÁTICA

- *Planeación y Definición de Requisitos:* Las funcionalidades principales del sistema se definieron mediante historias de usuario, como las registradas en la *Tabla 7*. Cada historia detalla las necesidades específicas de los interesados, traducidas en funcionalidades clave. Los criterios de aceptación asociados aseguran que cada funcionalidad cumpla con los requerimientos establecidos, priorizando la experiencia del usuario y la calidad del sistema, desde la creación de una landing page (*página de inicio o aterrizaje*) hasta la gestión segura de las propiedades del suelo y estabilizantes.

Tabla 7: Listado de Historias de usuario y Criterios de aceptación, Fuente: propia

Código	Historia de usuario	Código	Criterio de Aceptación
HU01	Como usuario, quiero ver una landing page que presente la información general del proyecto, para tener un entendimiento inicial del propósito y alcance del sistema.	CA0101	Debe contener secciones claras que describan el proyecto, objetivos y funcionalidades del sistema.
		CA0102	Debe ser accesible y responsiva para visualizarse adecuadamente en dispositivos móviles y de escritorio.
		CA0103	Debe incluir enlaces a secciones relevantes como acceso a inicio de sesión y contacto.
HU02	Como investigador, quiero poder iniciar sesión en el sistema, para acceder a funcionalidades exclusivas y asegurar la confidencialidad de la información.	CA0201	Debe requerir un nombre de usuario (email) y contraseña.
		CA0202	Debe ofrecer un mensaje de error claro si el inicio de sesión falla.
		CA0203	Debe redireccionar al usuario a su página de inicio correspondiente (dashboard) tras un inicio de sesión exitoso.
		CA0204	Debe incluir una opción para recuperar o restablecer la contraseña.
HU03	Como administrador, quiero tener la capacidad de crear, listar, actualizar y eliminar los perfiles de los investigadores, para gestionar quienes tienen acceso al sistema.	CA0301	La interfaz debe permitir añadir nuevos investigadores con información relevante como nombre, email contraseña y rol.
		CA0302	Debe permitir editar la información existente de los investigadores.
		CA0303	Debe permitir la eliminación de investigadores con una confirmación previa para evitar borrados accidentales.
		CA0304	La lista de investigadores debe actualizarse automáticamente tras cualquier operación CRUD.
HU04	Como investigador, quiero poder registrar las propiedades de los suelos y los estabilizantes, para almacenar y consultar datos necesarios para el análisis.	CA0401	Debe permitir la entrada de datos para todas las propiedades relevantes del suelo y estabilizantes.
		CA0402	Debe validar la integridad de los datos para asegurar que todos los campos requeridos estén completos y sean correctos.
		CA0403	Debe proporcionar un mensaje de confirmación al guardar los datos correctamente.
		CA0404	Debe permitir la cancelación o modificación de la entrada antes de la confirmación final.

HU05	Como investigador, quiero visualizar / detalles y eliminar las propiedades del suelo y estabilizante, para mantener la base de datos actualizada y relevante.	CA0501	Debe mostrar las propiedades del suelo y estabilizantes en un formato fácil de leer.
		CA0502	Debe permitir la eliminación con una opción de confirmación para evitar pérdidas accidentales de datos.
		CA0503	Debe actualizar la vista de datos inmediatamente después de cualquier eliminación.
HU06	Como investigador, quiero poder realizar cargas masivas de datos sobre propiedades de suelos y estabilizantes, para optimizar el tiempo y eficiencia en la actualización y expansión de la base de datos.	CA0601	Debe aceptar archivos en formatos comunes como Excel.
		CA0602	Debe incluir una validación de los datos para verificar el formato y la completitud antes de la importación.
		CA0603	Debe proporcionar un resumen de los datos cargados con la cantidad de registros añadidos y alertar sobre posibles errores.
HU07	Como investigador, quiero utilizar un modelo de inteligencia artificial, para predecir los estabilizantes adecuados basándome en las propiedades del suelo, con el fin de optimizar la selección de materiales y mejorar los resultados de estabilización.	CA0701	Debe proporcionar resultados de predicción en un tiempo razonable, no mayor a 1 minuto.
		CA0702	Debe mostrar el porcentaje de precisión para cada una de las predicciones, basada en datos de validación.
		CA0703	Debe permitir al usuario ingresar las propiedades del suelo y recibir recomendaciones de estabilizantes adecuados.

- *Iteraciones y Ajustes:* La planificación y priorización de las historias de usuario en el *Product Backlog* reflejaron un enfoque estructurado y alineado con las necesidades prioritarias del proyecto. La clasificación por prioridad y tamaño permitió gestionar eficientemente las tareas, asegurando que las historias críticas, como HU01 y HU02, fueran abordadas en las primeras fases de desarrollo, mientras que otras de menor urgencia, como HU06 y HU07, se reservaron para iteraciones posteriores.

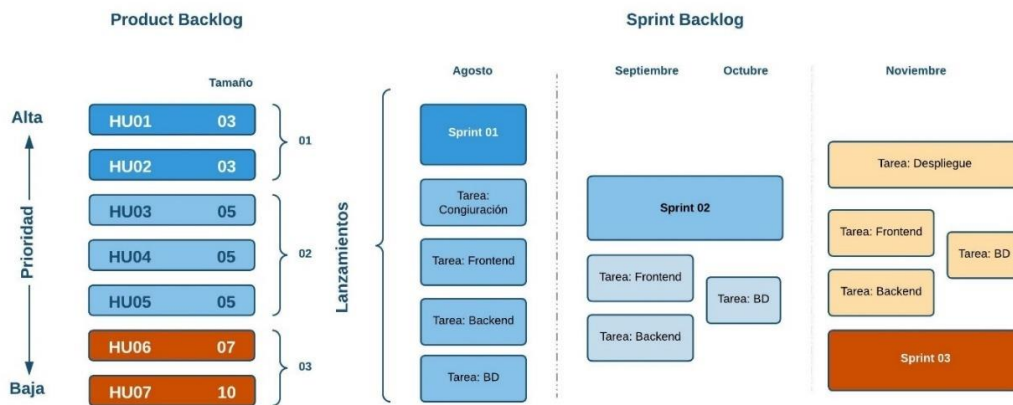


Figura 19: Planeación del desarrollo de la aplicación, Fuente: propia

El proceso representado en la *Figura 19*, destacó cómo las historias seleccionadas del *Product Backlog* fueron integradas en *Sprints*, siguiendo la metodología SCRUM. Este enfoque garantizó un desarrollo incremental y adaptable, donde cada *sprint*, desde el *Sprint 01* hasta el *Sprint 03*, se

centró en tareas específicas como la configuración del sistema, desarrollo del frontend, backend y bases de datos, de conformidad a los requerimientos de cada Historia de Usuario. La asignación clara de tareas dentro de cada sprint mantuvo una visión precisa del progreso y facilitó la identificación de posibles ajustes.

Además, la estructura iterativa descrita en la *Figura 19*, subrayó la importancia de las sesiones de revisión y retroalimentación continua al final de cada sprint, las cuales no solo evaluaron el cumplimiento de los objetivos, sino que también permitieron incorporar cambios en los requisitos. Este enfoque ágil aseguró una entrega incremental y priorizada de valor al usuario final, manteniendo el desarrollo alineado con las expectativas de los interesados. La flexibilidad integrada en el proceso mejoró significativamente la capacidad del equipo para adaptarse a cambios y garantizar el éxito del producto final.

- *Diseño de la Arquitectura de la Aplicación:* La aplicación SIPSE APP se diseñó bajo una arquitectura RESTful, ideal para sistemas distribuidos como aplicaciones web. Este enfoque permitió una interacción eficiente y flexible entre el cliente y el servidor mediante protocolos HTTP estandarizados, ver la *Figura 20*:

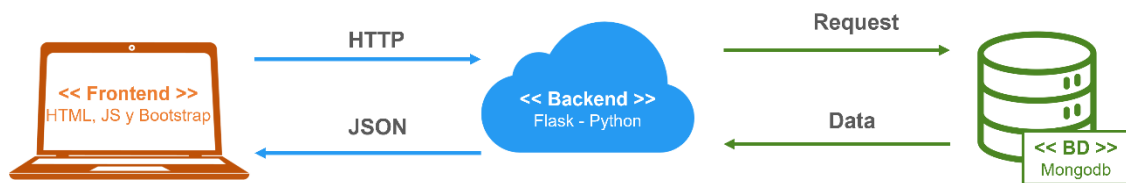


Figura 20: Diseño de la arquitectura de la aplicación, Fuente: propia

- Las principales *Tecnologías Utilizadas* en el proyecto fueron:
 - En la *Administración del Proyecto*, se utilizó JIRA para gestionar tareas y realizar un seguimiento del progreso, asegurando una organización efectiva dentro del equipo de desarrollo.
 - En el *Repositorio y Control de Versiones*, se emplearon Git y GitHub, lo que permitió una colaboración eficiente entre los miembros del equipo y un control coherente de versiones durante todo el proceso de desarrollo.
 - En el *Editor de Codificación*, se adoptó Visual Studio Code debido a su versatilidad y soporte para múltiples lenguajes de programación, lo que facilitó la edición y el manejo del código fuente.
 - En el *Desarrollo del Frontend y Backend*, se integraron tecnologías como HTML, JavaScript y Bootstrap para el desarrollo del frontend, proporcionando interfaces

- gráficas interactivas. En el backend, se utilizó Python con el framework Flask, ofreciendo una lógica de negocio robusta y adaptable a las necesidades del sistema.
- En la *Gestión de Base de Datos*, se implementó MongoDB Atlas como gestor de bases de datos, destacándose por su capacidad para manejar grandes volúmenes de datos y garantizar alta disponibilidad y rendimiento en el almacenamiento y recuperación de información.
 - En la *Seguridad*, se implementaron métodos de verificación, autenticación y encriptación para proteger datos sensibles, garantizando la privacidad e integridad de la información.
 - En la *Comunicación entre Componentes*, se establecieron protocolos REST y se utilizó el formato JSON para la transmisión de datos, garantizando una interacción fluida, eficiente y estandarizada entre los diferentes módulos del sistema.
 - Los *Diagramas de la Arquitectura* de la aplicación fueron desarrollados bajo el enfoque de C4 Models, los cuales proporcionan una visión progresiva y detallada de la estructura del sistema a través de varios niveles de abstracción. Este método permitió visualizar claramente las relaciones y dependencias entre los distintos componentes del sistema desde una perspectiva amplia hasta un nivel más detallado.
 - *Nivel de Contexto*: El diagrama de contexto como se ilustra en la *Figura 21*, muestra la interacción del sistema con los usuarios finales y otros sistemas externos:
 - Los *Usuarios*, como los administradores e investigadores acceden al sistema para solicitar y visualizar información sobre propiedades del suelo y estabilizantes, así como la gestión de las predicciones.
 - La *Aplicación Web*, sirve como el punto central para el manejo de solicitudes y la entrega de funcionalidades a los usuarios.
 - La *BD SIPSE-APP*, funciona como el almacenamiento principal de datos, asegurando que toda la información de suelos y estabilizantes esté centralizada y accesible.

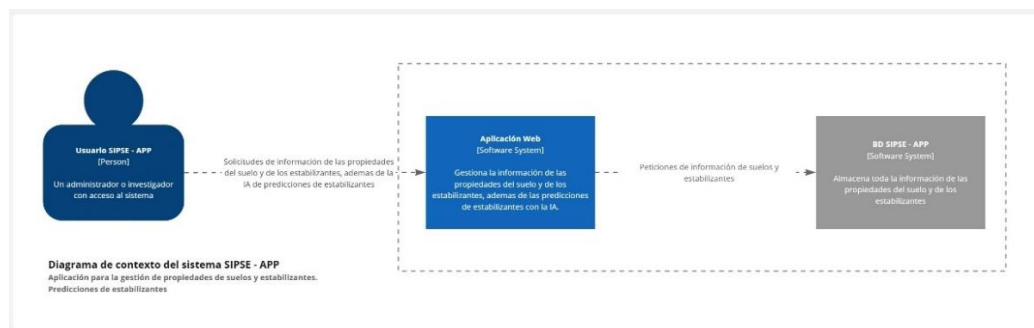


Figura 21: Diagrama de contexto de la aplicación SIPSE APP, Fuente: propia

- **Nivel de Contenedores:** El diagrama de contenedores, como se muestra en la *Figura 22*, describe la arquitectura del sistema SIPSE APP en términos de los principales contenedores tecnológicos que lo conforman. Cada contenedor representa un conjunto lógico de responsabilidades que trabajan en conjunto para proporcionar las funcionalidades del sistema:
 - El *Contenedor del Frontend*, incluye los Templates Landing Page y SIPSE APP, desarrollados con HTML, Bootstrap y JavaScript. Estos contenedores ofrecen las interfaces gráficas del sistema, permitiendo a los usuarios interactuar con las diferentes funcionalidades de la aplicación de manera intuitiva.
 - El *Contenedor del Backend*, representado por el contenedor Routes SIPSE APP, implementado con Flask (Python). Este contenedor gestiona la lógica de negocio, procesa las solicitudes de los usuarios, se comunica con la base de datos y proporciona las respuestas correspondientes.
 - El *Contenedor de Base de Datos MongoDB*, este contenedor almacena y gestiona toda la información relacionada con las propiedades de los suelos y estabilizantes. Es el núcleo de almacenamiento de datos del sistema, y trabaja en estrecha colaboración con el backend para garantizar el acceso eficiente y seguro a la información.
 - La *comunicación* entre los contenedores se realiza mediante el uso de protocolos HTTP y JSON, asegurando una interacción eficiente entre el frontend, el backend y la base de datos. Este diseño modular y distribuido asegura que el sistema sea escalable, flexible y fácil de mantener.

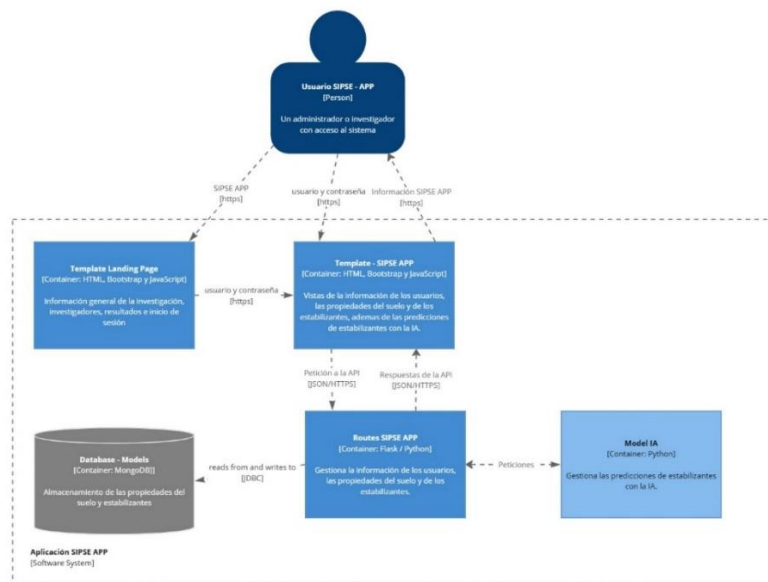


Diagrama de contenedor del sistema SIPSE - APP
Aplicación para la gestión de propiedades de suelos y estabilizantes.
Predicciones de estabilizantes

Figura 22: Diagrama de contenedores de la aplicación SIPSE APP, Fuente: propia

- *Nivel de Componentes:* El diagrama de componentes, como se muestra en la *Figura 23*, detalla la organización interna y las interacciones entre los principales módulos del sistema SIPSE APP. Cada componente cumple una función específica, facilitando la integración y comunicación en el sistema:

- *Los Componentes de Interfaz (Landing Page y Dashboard):*

El componente de *Landing Page*, proporciona información general sobre el proyecto, incluyendo detalles de la investigación, resultados y acceso al sistema mediante formularios de inicio de sesión.

El componente de *Dashboard*, ofrece visualizaciones clave relacionadas con las propiedades del suelo y estabilizantes, incluyendo estadísticas y gráficos interactivos para mejorar la experiencia del usuario.

- *Los Componentes de Autenticación:* El componente de *Login*, gestiona el inicio de sesión de usuarios, el restablecimiento de contraseñas y la creación de nuevas cuentas, garantizando un acceso seguro y controlado al sistema.

- *Los Componentes de Gestión de Usuarios y Datos:*

El componente *Usuarios*, permite realizar operaciones CRUD (Crear, Leer, Actualizar, Eliminar) sobre los perfiles de investigadores que utilizan el sistema.

El componente *Propiedades*, maneja la gestión de datos relacionados con las propiedades fisicoquímicas y mecánicas del suelo, así como la carga masiva de información.

El componente de *Predicciones*, utiliza datos ingresados para ejecutar modelos de inteligencia artificial que generan predicciones sobre estabilizantes, mejorando la toma de decisiones basada en datos.

- *El componente de Base de Datos MongoDB:*

Almacena de manera centralizada toda la información sobre suelos y estabilizantes, asegurando la accesibilidad, integridad y consistencia de los datos.

El diseño modular permite la integración eficiente de componentes, mientras que las comunicaciones entre ellos se realizan mediante protocolos REST y JSON, asegurando una operación fluida y escalable del sistema.

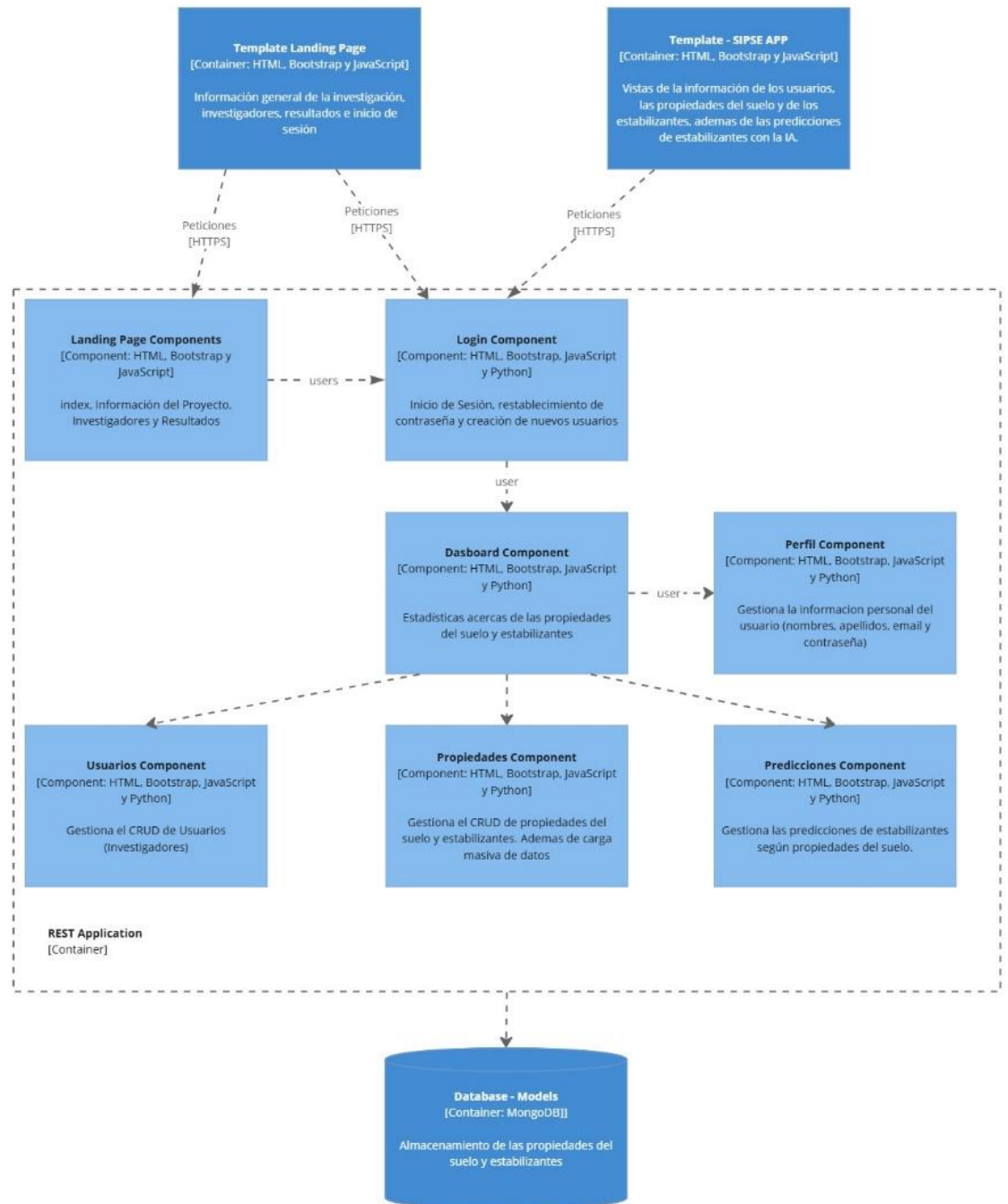


Diagrama de componentes del sistema SIPSE - APP
Aplicación para la gestión de propiedades de suelos y estabilizantes.
Predicciones de estabilizantes

Figura 23: Diagrama de componentes de la aplicación SIPSE APP, Fuente: propia

- **Desarrollo de los Componentes de la Aplicación:** El desarrollo de los componentes se realizó mediante un enfoque iterativo y ágil basado en la metodología Scrum. Este enfoque permitió al equipo abordar las historias de usuario en sprints cortos, asegurando entregas regulares y facilitando ajustes continuos basados en la retroalimentación del cliente. Cada sprint comenzaba con una reunión de planificación en la que se priorizaban las historias de usuario y se asignaban tareas a los miembros del equipo. Durante su ejecución, las tareas eran gestionadas de manera progresiva hasta ser evaluadas conforme a los criterios de aceptación y sometidas a pruebas de software, como se muestra en la *Figura 24*.

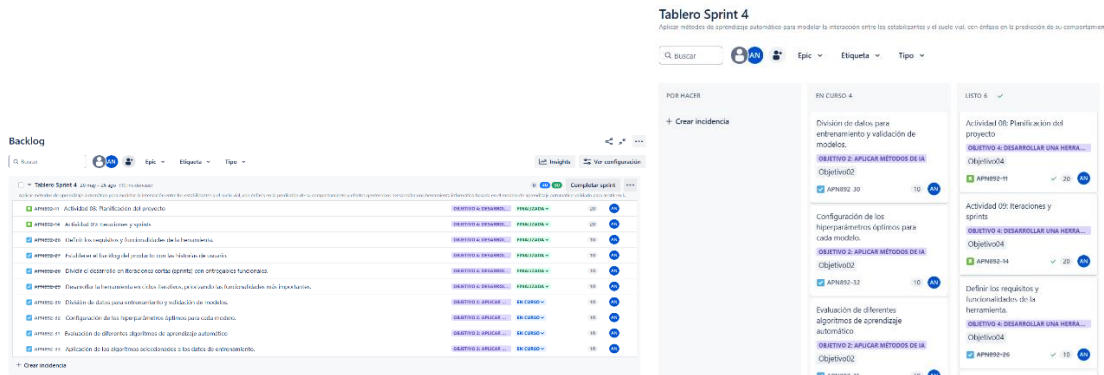


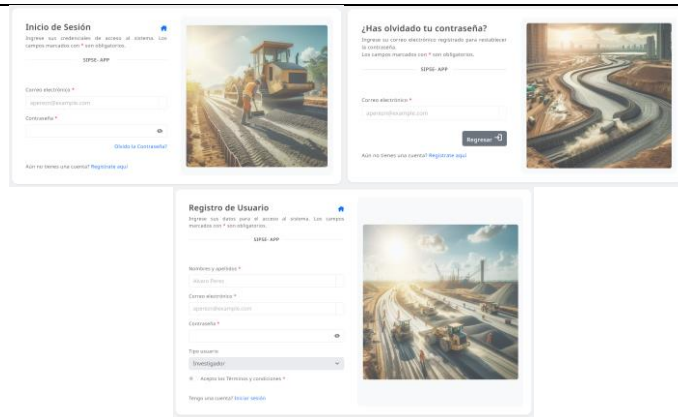
Figura 24: Esquema de un sprint de desarrollo, Fuente: propia

La **implementación técnica** de la aplicación involucra la transformación de las historias de usuario en funcionalidades operativas, abarcando desde el desarrollo de interfaces hasta la integración de sistemas backend. Este proceso se centró en cumplir con los criterios de aceptación establecidos, garantizando tanto la funcionalidad como la calidad del producto final. A continuación, en la *Tabla 8*, se presenta el resumen del desarrollo realizado, vinculando directamente las historias de usuario con las interfaces correspondientes y el cumplimiento de los criterios de aceptación; proporcionando una visión clara y concisa del trabajo realizado.

Tabla 8: Implementación Técnica de los componentes de software de la aplicación, Fuente: propia

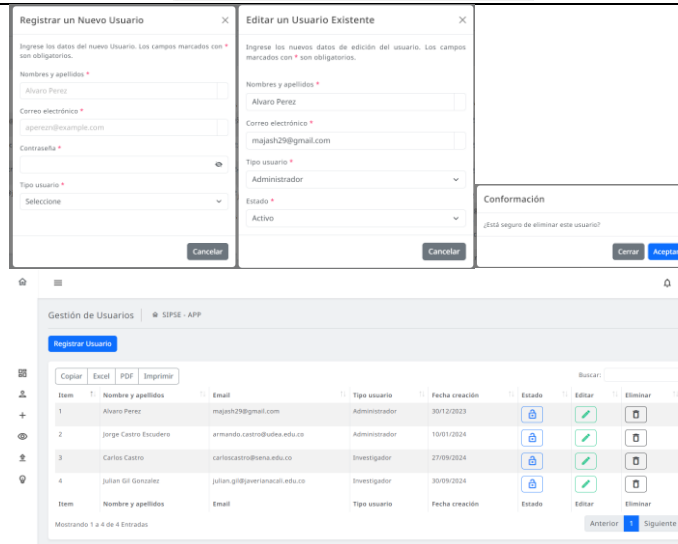
Historia de usuario	Funcionalidad	Cumplimiento de Criterios de Aceptación
<p>HU01: Acceso y visualización de la página de inicio</p>		<p>Todos los criterios de accesibilidad, rendimiento y usabilidad se han cumplido según lo especificado (CA0101, CA0102 y CA0103).</p>

HU02: Funcionalidad de inicio de sesión (login) para investigadores



La autenticación cumple con los estándares de seguridad, y los tiempos de respuesta están dentro de los límites aceptables (CA0201, CA0202, CA0203 y CA0204).

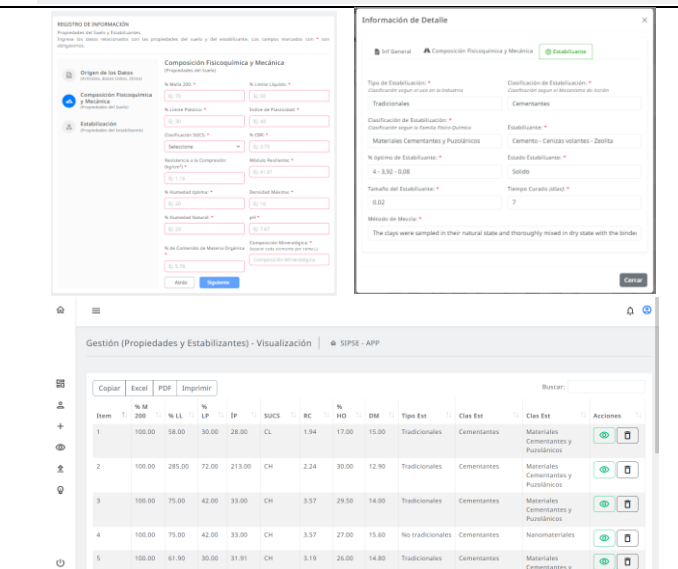
HU03: Administración de los usuarios del sistema (investigadores)



Item	Nombres y apellidos	Email	Tipo usuario	Fecha creación	Estado	Editar	Eliminar
1	Alvaro Perez	majah29@gmail.com	Administrador	30/12/2023			
2	Jorge Castro Escudero	armando.castro@ubee.edu.co	Administrador	10/01/2024			
3	Carlos Castro	carloscastro@emena.edu.co	Investigador	27/09/2024			
4	Juan Gil Gonzalez	juan.gil@javerianacali.edu.co	Investigador	30/09/2024			

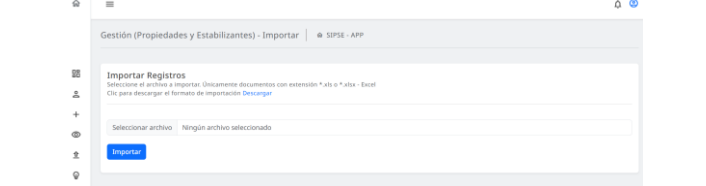

Las funciones de CRUD se implementan correctamente, permitiendo operaciones eficientes y seguras (CA0301, CA0302, CA0303 y CA0304).

HU04 y HU05: Gestión de las propiedades de los suelos y los estabilizantes



Item	% M	% LL	LP	IP	SUCS	RC	% HD	DM	Tipo Est	Clas Est	Clas Est	Acciones
1	100.00	58.00	30.00	28.00	CL	1.94	17.00	15.00	Tradicionales	Cementantes	Sustancias Cementantes y Puzolánicas	
2	100.00	285.00	72.00	213.00	CH	2.24	30.00	12.90	Tradicionales	Cementantes	Materiales Cementantes y Puzolánicas	
3	100.00	75.00	42.00	33.00	CH	3.57	29.50	14.80	Tradicionales	Cementantes	Materiales Cementantes y Puzolánicas	
4	100.00	75.00	42.00	33.00	CH	3.97	27.00	15.80	No tradicionales	Cementantes	Nanomateriales	
5	100.00	61.00	30.00	31.01	CH	3.19	26.00	14.80	Tradicionales	Cementantes	Materiales Cementantes y	

Las funciones de CRUD se implementan correctamente, permitiendo operaciones eficientes y seguras (CA0401, CA0402, CA0403, CA0404, CA0501, CA0502 y CA0503).

<p>HU06: Gestión de cargas masivas de datos sobre propiedades de suelos y estabilizantes</p>		<p>La carga masiva de datos cumple con los tiempos e respuesta y los formatos establecidos (CA0601, CA0602 y CA0603).</p>
<p>HU07: Modelo predictivo de estabilizantes</p>		<p>Las predicciones del modelo supervisado de aprendizaje automático cumplen con los tiempos de respuesta y etiquetas de predictoras (CA0701, CA0702 y CA0703).</p>

- *Pruebas de desarrollo y Calidad:* Durante el desarrollo, se implementaron pruebas manuales y automáticas para garantizar la calidad y el correcto funcionamiento de los componentes, de las cuales se destacan las siguientes:
 - *Pruebas unitarias:* Las pruebas unitarias realizadas a los componentes de la aplicación SIPSE APP permitieron evaluar de forma aislada cada funcionalidad, identificando errores en etapas tempranas del desarrollo. Como se observa en la *Figura 25*, las historias de usuario relacionadas con funcionalidades críticas, como el acceso y visualización de la página de inicio (HU01) y el inicio de sesión para investigadores (HU02), presentaron un menor volumen de pruebas en comparación con otras historias, con 20 y 25 pruebas totales respectivamente. En ambas historias, las pruebas exitosas superaron ampliamente a las erróneas, reflejando un buen nivel de madurez en estos componentes desde etapas tempranas. Este enfoque ágil permitió ajustar y corregir errores rápidamente durante el mismo sprint, garantizando la calidad de estas funcionalidades clave.

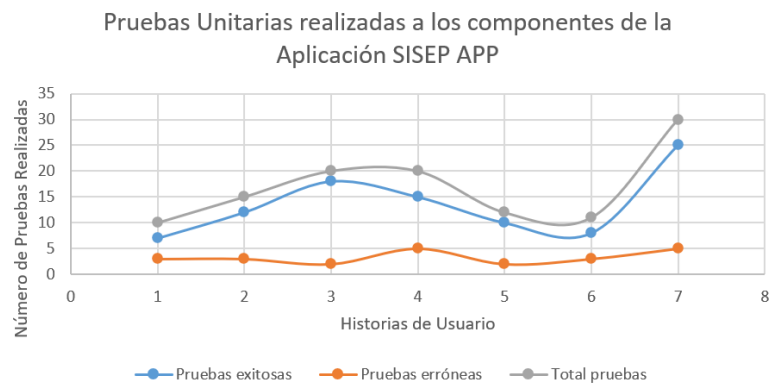


Figura 25: Pruebas unitarias realizadas a la Aplicación SISEP APP

Por otro lado, las historias relacionadas con tareas más complejas, como la gestión de cargas masivas (HU06) y el modelo predictivo de estabilizantes (HU07), requirieron una mayor cantidad de pruebas, alcanzando un total de 30 y 35 respectivamente. Aunque se identificaron más errores en estas historias, los ajustes realizados aseguraron que los componentes finales cumplieran con los requisitos establecidos. Es importante destacar que las historias HU04 y HU05, enfocadas en la gestión de propiedades de suelos y estabilizantes, también mostraron un equilibrio adecuado entre pruebas exitosas y erróneas, reflejando un desarrollo progresivo y alineado con los objetivos del sistema. Este enfoque iterativo y continuo contribuyó a minimizar riesgos durante la integración final y aseguró que cada componente estuviera completamente funcional antes de su despliegue.

- *Pruebas de integración:* Estas pruebas garantizaron que los componentes del sistema interactuaran correctamente entre sí y con la base de datos. Tal como se evidencia en la *Figura 26*, no se detectaron conflictos entre los componentes al realizar la integración, lo que permitió efectuarla de manera exitosa. El proceso de revisión y aprobación de los cambios confirmó que la rama destinada al despliegue de producción estaba libre de conflictos, habilitando su fusión con la rama principal sin inconvenientes. Esto asegura la continuidad del desarrollo y la estabilidad del sistema tras la integración.

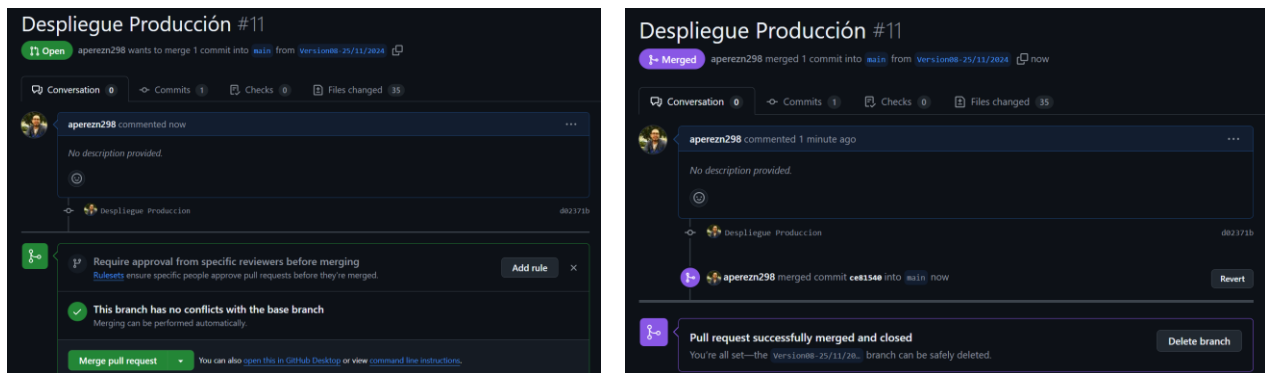


Figura 26: Pruebas de integración realizadas a la Aplicación SISEP APP

- *Pruebas de rendimiento:* Estas pruebas permitieron evaluar la eficiencia del sistema, midiendo métricas clave relacionadas con la velocidad de carga, la interactividad y la estabilidad visual de la aplicación. Como se observa en los resultados de la *Figura 27*, el rendimiento general obtuvo una calificación de 91 sobre 100, lo que indica un sistema altamente optimizado.

Adicionalmente, las calificaciones en accesibilidad (87), prácticas recomendadas (96) y SEO (98) demuestran un enfoque integral para garantizar no solo el rendimiento técnico, sino también la experiencia del usuario y la optimización para buscadores.

Estos resultados reflejan un sistema bien desarrollado, preparado para responder eficazmente a las demandas de los usuarios.

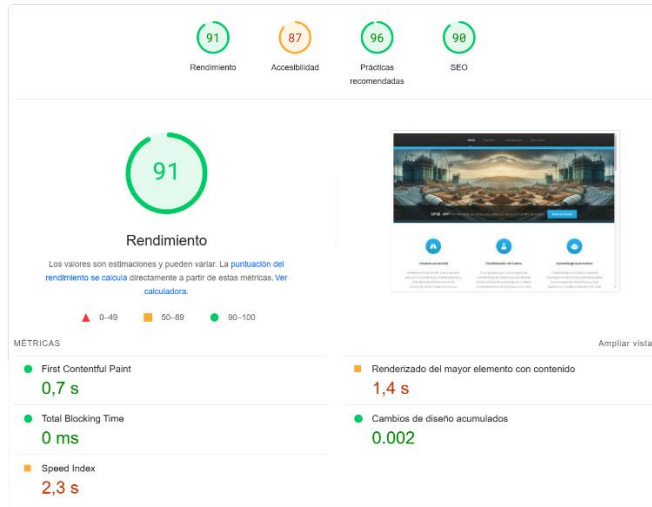


Figura 27: Pruebas de rendimiento realizadas a la Aplicación SISEP APP

El desarrollo de la herramienta informática representó un avance significativo en la integración de la inteligencia artificial con la geotecnia vial, demostrando la viabilidad de aplicar modelos de aprendizaje automático para optimizar la selección de estabilizantes en suelos subrasantes. Este proceso permitió consolidar una solución tecnológica robusta, respaldada por un diseño modular que asegura la escalabilidad y adaptabilidad a distintos escenarios. Con una arquitectura RESTful, un modelo predictivo eficiente y una interfaz intuitiva, la herramienta logra abordar las complejidades inherentes a la geotecnia, proporcionando resultados precisos y reduciendo costos y tiempos en la toma de decisiones. La validación exitosa en entornos controlados sienta las bases para futuras aplicaciones en escenarios reales, posicionando la herramienta como un recurso valioso para proyectos de infraestructura vial sostenible.

5.4. IMPLEMENTACIÓN Y DESPLIEGUE DE LA APLICACIÓN

- *La implementación y despliegue de la aplicación:* La integración de los componentes de la aplicación SIPSE APP, como se muestra en la Figura 28, presenta un sistema en el que el usuario final interactúa mediante solicitudes HTTP y JSON, enviadas al servidor web donde se encuentra alojada la API desarrollada en Flask. Esta API actúa como intermediaria entre el usuario, la base de datos MongoDB y el modelo de inteligencia artificial (IA). La base de datos almacena información clave sobre propiedades de suelos y estabilizantes, mientras que el componente de IA utiliza estos datos para generar predicciones en tiempo real, proporcionando resultados precisos y adaptados a las necesidades del usuario. Este diseño

asegura una comunicación eficiente y un flujo constante de datos entre los diferentes módulos del sistema.

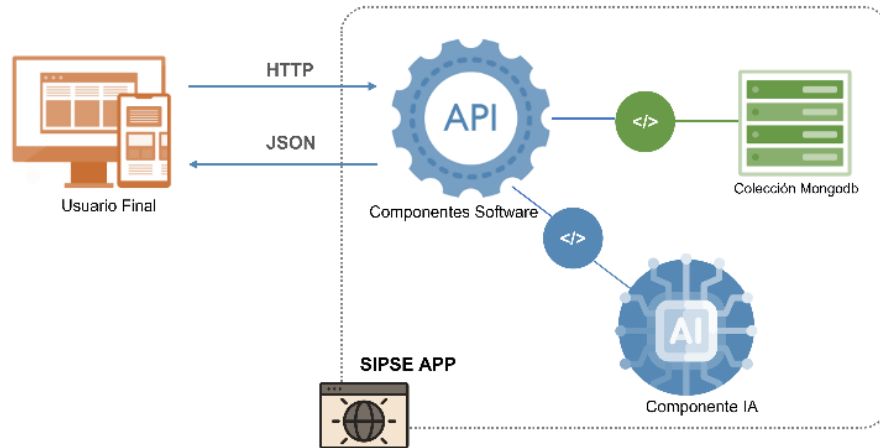


Figura 28: Integración de los componentes y despliegue de la aplicación de la Aplicación, Fuente: propia

En el Despliegue de la aplicación, ilustrado en la *Figura 28*, detalla la disposición de los componentes tecnológicos de la aplicación en un entorno de producción. El servidor web, alojado en *PythonAnywhere* y accesible a través del dominio <https://majash29.pythonanywhere.com>, gestiona el tráfico y redirige las solicitudes hacia el backend, desarrollado con Flask. Este backend procesa y envía datos hacia la base de datos MongoDB, que se encuentra alojada en un clúster dedicado en MongoDB Atlas, lo que garantiza alta disponibilidad y acceso seguro a la información. Adicionalmente, el modelo de inteligencia artificial, se encuentra integrado directamente con el backend y opera en el mismo servidor, optimizado para el procesamiento eficiente de predicciones en tiempo real. Este despliegue asegura una arquitectura robusta y escalable que puede soportar tanto las operaciones actuales del sistema como el crecimiento futuro de la aplicación, adaptándose a los requisitos cambiantes de los usuarios y la infraestructura.

La implementación de la aplicación SIPSE APP refleja una arquitectura distribuida eficiente, integrando componentes alojados en *PythonAnywhere* con *MongoDB Atlas* y un modelo de IA para brindar una solución funcional y escalable. Al combinar un servidor web robusto, un backend optimizado y almacenamiento en un clúster seguro, el sistema asegura un rendimiento confiable en producción, permitiendo a los usuarios acceder a las funcionalidades de manera ágil y efectiva.

- *El mantenimiento y sostenibilidad* de la herramienta informática desarrollada son elementos clave para garantizar su viabilidad y efectividad a largo plazo. En el aspecto del *mantenimiento*, se definieron estrategias basadas en actualizaciones periódicas del modelo

de aprendizaje automático, programadas en intervalos de cada tres meses o cuando se disponga de nuevos tipos de estabilizantes que no hayan sido incluidos en el estudio inicial. Estas actualizaciones permiten mitigar posibles fallos en las predicciones de estabilizantes conforme a las características del suelo, al tiempo que mejoran la precisión del modelo y lo adaptan a nuevas condiciones geotécnicas o requerimientos específicos. Este enfoque iterativo asegura que el sistema permanezca actualizado y alineado con las necesidades del usuario.

Desde la *perspectiva de sostenibilidad*, el proyecto prioriza el uso de tecnologías eficientes y de bajo impacto ambiental, minimizando el consumo de recursos computacionales sin comprometer el rendimiento del sistema. Asimismo, la integración de políticas de software libre y código abierto fomenta la adaptabilidad y reutilización de la herramienta, permitiendo su implementación en otros contextos geográficos y aplicaciones técnicas relacionadas. Estas estrategias no solo garantizan la operatividad continua de la herramienta, sino que también amplifican su impacto positivo en los campos de la ingeniería civil y la ciencia de datos, promoviendo una solución escalable, accesible e innovadora para diversas comunidades técnicas.

- *La evaluación externa de la solución* se lleva a cabo a través de pruebas y validaciones realizadas con nuevos datos de suelos y usuarios finales expertos en la área de la geotecnia y software. Este proceso proporcionó una retroalimentación valiosa sobre el desempeño, la funcionalidad y la usabilidad de la aplicación, en un contexto real.
 - *Validación con Nuevos datos de suelos:* Para validar el modelo con nuevos datos sobre suelos viales, se realizó un proceso adicional de recopilación de información, utilizando bases de datos científicas y resultados de pruebas de laboratorio. Esto permitió construir un conjunto de datos completamente nuevo, diseñado para evaluar el desempeño del modelo con información distinta a la empleada durante su entrenamiento.

Los resultados que se ilustran en la *Figura 29*, muestran el proceso de validación de 40 nuevos registros representativos de diversas propiedades de suelos viales. En la validación, el modelo basado en *Random Forest* logró predecir correctamente 33 valores, mientras que 7 fueron clasificados de forma incorrecta. Por otra parte, el modelo *Ensemble* demostró un mejor rendimiento, con 35 predicciones correctas y solo 5 incorrectas, lo que respalda la efectividad del enfoque propuesto.

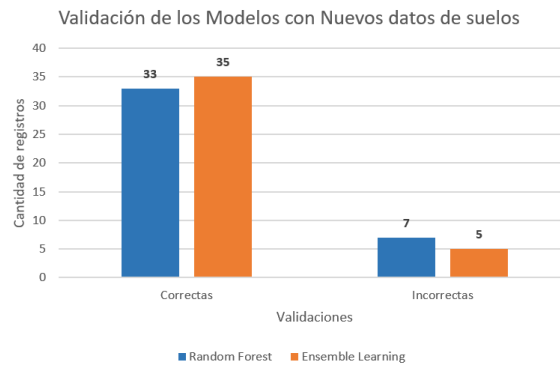


Figura 29: Validación de los Modelos con Nuevos datos de suelos, Fuente: propia

Sin embargo, ambos modelos presentan limitaciones al clasificar estabilizantes de tipo mineral, debido a que esta clase cuenta con un menor número de registros, lo que se traduce en un aprendizaje menos robusto para este caso específico. Este resultado pone en evidencia, la necesidad de enriquecer el conjunto de datos para aumentar la capacidad de generalización del modelo.

- *Validación del modelo de clasificación:* Un pequeño grupo de expertos temáticos en el área de infraestructura vial evaluó la calidad y pertinencia del modelo de aprendizaje automático utilizado para la predicción de estabilizantes según la familia Físico-Química, asegurando su precisión y relevancia técnica.

A continuación, se describe la ficha técnica del instrumento de evaluación aplicado a este grupo de expertos temáticos:

Tabla 9: Instrumento de evaluación para expertos temáticos, Fuente: propia

Nombre del Instrumento	Encuesta de percepción del modelo predictivo de estabilizantes conforme a las propiedades del suelo.
Objetivo:	Recibir retroalimentación por parte de los expertos temáticos del área de infraestructura, su percepción acerca de la herramienta informática de predicción de suelos y estabilizantes.
Preguntas cerradas Obligatorias	01 - ¿La aplicación es intuitiva y fácil de manipular? 02 - ¿La aplicación facilita la gestión (<i>registro- consulta -eliminación-carga</i>) de la información de las propiedades del suelo y estabilizantes? 03 - ¿Considera que el modelo predictivo de estabilizantes es útil para la toma de decisiones en proyectos viales?
Opciones respuesta de	01 - De acuerdo 02 – Neutral 03 - En desacuerdo

Pregunta abierta
Opcional

04 - ¿Qué aspectos considera que podrían mejorarse o adicionarse en la aplicación?

Los resultados obtenidos de los expertos temáticos del área de infraestructura fueron los siguientes:

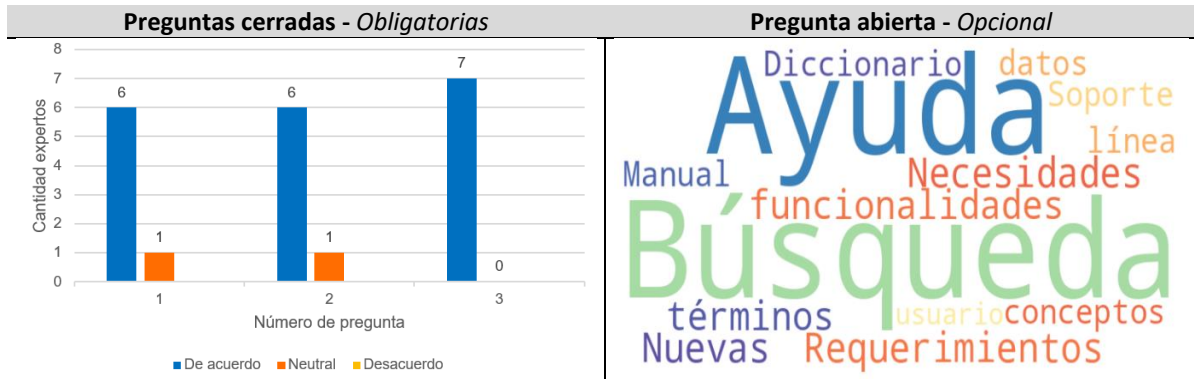


Figura 30: Resultados de la evaluación a expertos temáticos, Fuente: propia

Los resultados consolidados en la *Figura 30*, evidencian una percepción positiva de los expertos temáticos hacia el modelo predictivo de estabilizantes – SIPSE APP. En las preguntas cerradas, la mayoría de los encuestados valoran favorablemente la herramienta tecnológica: 6 de los 7 expertos consideran que la aplicación es intuitiva y fácil de usar, y el mismo número está de acuerdo en que facilita la gestión de información relacionada con suelos y estabilizantes. Además, la utilidad del modelo predictivo en la toma de decisiones para proyectos viales recibió una valoración unánime, con 7 de 7 expertos expresando una opinión positiva. Por otro lado, la respuesta abierta refleja aspectos de mejora, tales como: la integración de búsquedas avanzadas para las propiedades del suelo, un diccionario de términos técnicos que ayude a los usuarios a entender conceptos técnicos, opciones de ayuda y soporte. Esto indica que, si bien la herramienta cumple con su propósito principal, presenta oportunidades para una mejora continua mediante nuevas funcionalidades que enriquezcan la experiencia del usuario y se adapten de manera integral a los contextos empresariales.

- *Validación de la aplicación como software:* Un pequeño grupo, conformado por expertos técnicos en desarrollo de software, analizó el funcionamiento de la aplicación, verificando que el sistema cumpliera con los estándares de la industria en términos de desempeño, confiabilidad y usabilidad.

A continuación, se describe la ficha técnica del instrumento de evaluación aplicado a este grupo de expertos técnicos:

Tabla 10: Instrumento de evaluación para expertos técnicos, Fuente: propia

Nombre del Instrumento	Encuesta de percepción de la aplicación SIPSE APP para la gestión de suelos y estabilizantes.
Objetivo:	Recibir retroalimentación por parte de los expertos técnicos del área de desarrollo de software, su percepción acerca de la herramienta informática de predicción de suelos y estabilizantes.
Preguntas cerradas Obligatorias	01 - ¿La aplicación es intuitiva y fácil de usar? 02 - ¿La aplicación satisface todos los requisitos establecidos por el cliente? 03 - ¿El diseño arquitectónico propuesto es adecuado para abordar la problemática?
Opciones de respuesta	01 - De acuerdo 02 - Neutral 03 - En desacuerdo
Pregunta abierta Opcional	04 - ¿Qué aspectos considera que podrían mejorarse o adicionarse en la aplicación?

Los resultados obtenidos de los expertos técnicos del área de desarrollo de software fueron los siguientes:

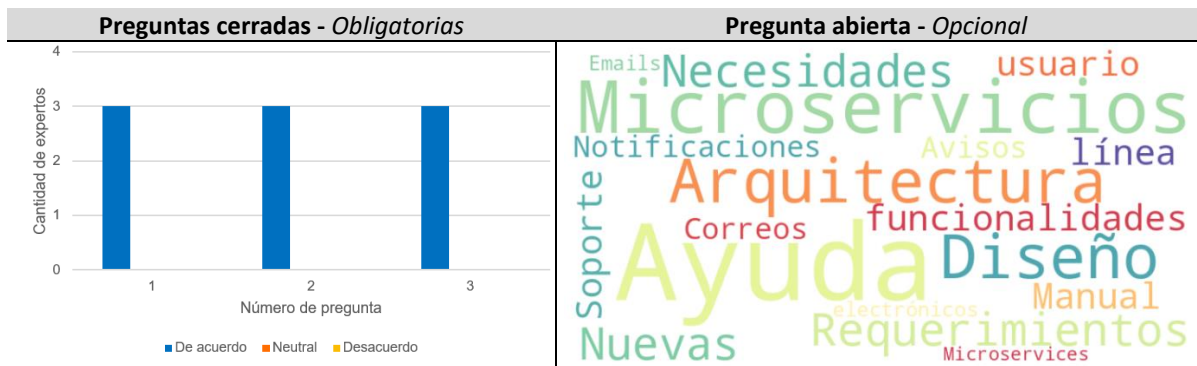


Figura 31: Resultados de la evaluación a expertos técnicos, Fuente: propia

Los resultados consolidados en la Figura 31, evidencian una percepción positiva por parte de los expertos técnicos hacia la herramienta informática SIPSE APP. En las preguntas cerradas, se logró un consenso unánime favorable: los tres expertos coincidieron en que la aplicación es intuitiva, fácil de usar y cumple plenamente con los requerimientos establecidos por el cliente. Además, destacaron que la arquitectura RESTful implementada es adecuada para la solución planteada.

Por otro lado, la respuesta abierta mostró aspectos de mejora, como la incorporación de opciones de ayuda y soporte en línea dirigidas a los diferentes actores, la

implementación de notificaciones por correo electrónico para informar sobre la gestión de nuevas propiedades de suelos y estabilizantes o el registro de nuevos usuarios en la aplicación. Finalmente, se sugirió explorar otras alternativas arquitectónicas, como la adopción de microservicios, para futuras iteraciones del proyecto.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

El desarrollo del modelo de aprendizaje automático supervisado abordó con éxito los objetivos específicos de la investigación mediante un enfoque sistemático y estructurado. Se logró preparar un conjunto de datos de calidad, compuesto por 293 registros, procesados mediante técnicas de limpieza, imputación de valores faltantes y normalización, garantizando su relevancia para capturar las propiedades críticas del suelo y los estabilizantes. Propiedades del suelo como la granulometría, el índice de plasticidad y la resistencia a la compresión y las características bases del estabilizante según el uso en la industria y el mecanismo de acción, demostraron ser factores clave en la predicción, representando el 62.5% de la importancia total en los modelos evaluados.

En cuanto al modelado, se seleccionaron y entrenaron algoritmos avanzados como Random Forest, XGBoost y un modelo ensemble, entre otros, destacando al modelo basado en Random Forest por su exactitud del 90.8% y su capacidad para manejar la heterogeneidad de los datos. La validación interna incluyó métricas clave como accuracy (90.8%), recall (91.8%), precisión (92.3%) y f1-score (92.0%), lo que evidenció la robustez y precisión del enfoque predictivo. Sin embargo, se identificaron limitaciones relacionadas con la representatividad del conjunto de datos, restringiendo la generalización del modelo a contextos externos. Estas observaciones subrayan la importancia de ampliar la base de datos para incluir suelos de diferentes regiones geográficas y propiedades geotécnicas más diversas, además de considerar enfoques más interpretables para facilitar la adopción por parte de usuarios técnicos y no técnicos.

Adicionalmente, los análisis de desempeño resaltaron la viabilidad de integrar técnicas de ensamble como votación homogénea, que incrementaron la precisión a 91.95%. Esto reafirma la efectividad de combinar modelos en tareas de clasificación complejas, sugiriendo líneas futuras de investigación enfocadas en optimizar la eficiencia computacional del entrenamiento y en explorar modelos que combinen precisión, interpretabilidad y escalabilidad.

El diseño y desarrollo de la herramienta informática SIPSE APP permitió traducir el modelo de aprendizaje automático en una solución tecnológica práctica y escalable, cumpliendo con los objetivos planteados en la investigación. La aplicación, diseñada con una arquitectura RESTful e integrada con MongoDB Atlas, ofrece funcionalidades de consulta, carga masiva, eliminación y predicción, adaptándose a las necesidades de los usuarios finales mediante una interfaz gráfica intuitiva.

El proceso de desarrollo, basado en la metodología Scrum, aseguró entregas iterativas y flexibles, permitiendo integrar historias de usuario priorizadas con tiempos de respuesta inferiores a 3 segundos y una tasa de errores funcionales por debajo del 5%. Además, las pruebas unitarias, de integración y de rendimiento corroboraron la calidad del sistema, alcanzando un puntaje de 91

sobre 100 en eficiencia técnica. La herramienta no solo encapsula el modelo predictivo, sino que amplía su accesibilidad y utilidad en contextos reales, mostrando su capacidad para transformar la gestión de estabilizantes en proyectos de infraestructura vial.

El proyecto respondió satisfactoriamente a la pregunta de investigación planteada: *¿Cómo implementar un modelo de aprendizaje automático que permita seleccionar estabilizantes para suelos viales?*. La integración de técnicas avanzadas de inteligencia artificial con un diseño de software robusto permitió seleccionar estabilizantes óptimos basados en propiedades específicas del suelo, alineándose con las condiciones ambientales y geográficas del proyecto. Además, se establecieron bases sólidas para optimizar procesos en la ingeniería geotécnica, mejorando la calidad, sostenibilidad y durabilidad de las infraestructuras viales. Este trabajo no solo valida la IA como un recurso estratégico para la selección de estabilizantes, sino que también abre nuevas oportunidades para su aplicación en otros campos de la infraestructura vial y ambiental.

6.2. TRABAJOS FUTUROS

Teniendo en cuenta los resultados obtenidos y las áreas de mejora identificadas, se proponen las siguientes líneas de trabajo futuro. Estas direcciones buscan abordar limitaciones actuales y abrir nuevas oportunidades de investigación y desarrollo en ciencia de datos, contribuyendo al avance técnico en aprendizaje automático y su aplicación a la geotecnia vial:

- Recolectar datos adicionales de suelos provenientes de diversas regiones y condiciones geotécnicas, incluyendo datos en tiempo real obtenidos de sensores geotécnicos. Esto permitirá mejorar la capacidad de generalización del modelo y habilitar actualizaciones dinámicas que incrementen la precisión y relevancia de las predicciones en escenarios diversos. Se espera que esta estrategia reduzca el error de predicción en al menos un 15%, ampliando la aplicabilidad del sistema.
- Evaluar el desempeño de técnicas más complejas, como redes neuronales profundas, modelos no supervisados o enfoques mixtos, especialmente en casos donde no se cuente con un experto para etiquetar los datos. Estas técnicas permitirán mejorar la precisión y robustez del sistema en escenarios con alta heterogeneidad y datos no etiquetados, con el objetivo de incrementar la precisión en un 10% en comparación con los enfoques actuales.
- Evaluar y comparar otras técnicas de meta-modelos, como stacking y bagging, en contraste con el enfoque basado en voting. Estas alternativas pueden ofrecer ventajas adicionales, como una mayor capacidad de generalización y optimización del desempeño en problemas de clasificación compleja. Se espera que la implementación de estas técnicas permita una mejora significativa en las métricas de precisión y F1-score, superando los enfoques tradicionales.

- Incorporar simulaciones interactivas y reportes automatizados en la herramienta SIPSE APP para facilitar la toma de decisiones en proyectos de gran escala. Estas mejoras no solo incrementarán la usabilidad de la plataforma, sino que también permitirán a los usuarios analizar escenarios complejos de forma intuitiva, aumentando su productividad.
- Implementar estrategias de edge computing para reducir los tiempos de inferencia y mejorar la escalabilidad del modelo en dispositivos con recursos limitados. Esto permitirá una optimización del desempeño computacional, con la meta de disminuir los tiempos de respuesta en un 30%, asegurando que el sistema sea funcional incluso en entornos operativos restringidos.
- Explorar nuevos dominios de predicción relacionados con la infraestructura vial, como el análisis del rendimiento a largo plazo de suelos estabilizados o el impacto del tráfico vehicular en las propiedades mecánicas del suelo. Este enfoque ampliará la aplicabilidad del modelo y su utilidad en proyectos de infraestructura vial sostenibles.

Estas líneas de trabajo no solo buscan fortalecer la herramienta SIPSE APP, sino también contribuir al campo de ciencia de datos mediante avances en generalización de modelos, optimización computacional y manejo de datos no etiquetados. Se establecerán métricas de éxito como la reducción de error en predicciones, mejora en tiempos de respuesta y ampliación de la aplicabilidad del sistema en contextos reales. De esta manera, se asegura que las investigaciones futuras tengan un impacto significativo tanto en el ámbito académico como en la práctica profesional.

7. REFERENCIAS BIBLIOGRÁFICAS

- [1] C. Crespo Villalaz, "Mecánica de suelos y cimentaciones", 6a ed., Editorial LIMUSA, 2007, ISBN: 9789681869632.
- [2] S. Crespo Escobar, "Materiales de construcción para edificación y obra civil," Editorial Club Universitario, 2013.
- [3] H. M. Abuel-Naga and A. Bouazza, "Numerical experiment-artificial intelligence approach to develop empirical equations for predicting leakage rates through GM/GCL composite liners," *Geotextiles and Geomembranes*, vol. 42, no. 3, pp. 236-245, Jun. 2014. doi: 10.1016/J.GEOTEXMEM.2014.04.002.
- [4] Invias, "Especificaciones generales de construcción de carreteras 2022", Bogotá, 2022.
- [5] I.H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, p. 160, 2021. doi: 10.1007/s42979-021-00592-x.
- [6] N.S. Shetu and M.A. Masum, "Implementation of an intelligent computing platform for determination of moisture content of subgrade soil," *Geotechnical Testing Journal*, vol. 38, no. 3, pp. 315-324, 2015. doi: 10.1520/GTJ20130128.
- [7] A.H. Alavi, A.A. Heshmati, A.H. Gandomi, A. Askarinejad, y M. Mirjalili, "Utilisation of computational intelligence techniques for stabilised soil," en *Proceedings of the 6th International Conference on Engineering Computational Technology*, 2008. [En línea]. Disponible en: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-77649233801&partnerID=40&md5=6ad9f0d8bc78353263d4bcf0b42cea68>.
- [8] S. Suman, M. Mahasakti, and S.K. Das, "Prediction of Maximum Dry Density and Unconfined Compressive Strength of Cement Stabilised Soil Using Artificial Intelligence Techniques," in *International Journal of Geosynthetics and Ground Engineering*, vol. 2, no. 2, 2016, pp. 1-10. doi: 10.1007/s40891-016-0051-9.
- [9] Eyo, Eyo U., Samuel J. Abbey, y Colin A. Booth, "Strength Predictive Modelling of Soils Treated with Calcium-Based Additives Blended with Eco-Friendly Pozzolans—A Machine Learning Approach," *Materials*, vol. 15, no. 13, 2022, doi: 10.3390/ma15134575.
- [10] Pro-Road Global. (2023, Noviembre 1). Estabilización de Suelos. [En línea]. Disponible en: <https://www.proroadglobal.com/estabilizacion-de-suelos/>. Acceso: Noviembre 1, 2023.

- [11] J. F. Rivera, A. Aguirre-Guerrero, R. Mejía de Gutiérrez, y A. Orobio, "Estabilización química de suelos - Materiales convencionales y activados alcalinamente (revisión)," *Informador Técnico*, vol. 84, no. 2, pp. 202-226, 2020. DOI: 10.23850/22565035.2530
- [12] S. Sobhan, "Challenges due to problematic soils: a case study at the crossroads of geotechnology and sustainable pavement solutions," *Innov. Infrastruct. Solut.*, vol. 2, p. 40, 2017. [Online]. Available: <https://doi.org/10.1007/s41062-017-0070-y>
- [13] G. Narendra Goud, S. Sasanka Mouli, Balunaini Umashankar, Saride Sireesh, y R. Madhav Madhira, "Design and Sustainability Aspects of Geogrid-Reinforced Flexible Pavements—An Indian Perspective," *Frontiers in Built Environment*, vol. 6, 2020. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fbuil.2020.00071>. DOI: 10.3389/fbuil.2020.00071.
- [14] H. Alimohammadi, V.R. Schaefer, J. Zheng, et al., "Performance evaluation of geosynthetic reinforced flexible pavement: a review of full-scale field studies," *Int. J. Pavement Res. Technol.*, vol. 14, pp. 30-42, 2021. [Online]. Available: <https://doi.org/10.1007/s42947-020-0019-y>
- [15] R. L. Handy, *Foundation Engineering: Geotechnical Principles and Practical Applications*. New York: McGraw-Hill Education, 1st ed., 2020. [Online]. Available: <https://www.accessengineeringlibrary.com/content/book/9781260026030>
- [16] Dewberry, "COMPONENTS OF THE BUILT ENVIRONMENT," in *Development of the Built Environment: From Site Acquisition to Project Completion*, 1st ed. New York: McGraw-Hill Education, 2019, chap. 2.2. [Online]. Available: <https://www.accessengineeringlibrary.com/content/book/9781260440737/toc-chapter/chapter2/section/section3>
- [17] R. W. Day, "PAVEMENT AND PIPELINE DESIGN," in *Geotechnical Engineer's Portable Handbook*, 2nd ed. New York: McGraw-Hill, 2012, chap. 11. [Online]. Available: https://www.accessengineeringlibrary.com/content/book/9780071789714/chapter/chapter_11
- [18] R. W. Day, "Manual de ingeniería de cimientos: diseño y construcción con el Código Internacional de Construcción de 2009," 2nd ed. New York, 2010-2006. [Online]. Available: <https://www.accessengineeringlibrary.com/content/book/9780071740098>
- [19] S. Arora, R. Gambheer, and M. Vohra, "IoT Data Systems Design," 1st ed. New York, 2021. [Online]. Available:

https://www.accessengineeringlibrary.com/content/book/9781260463095/chapter/chapter_6

- [20] "¿Qué es ETL?," Amazon Web Services. [En línea]. Disponible en: <https://aws.amazon.com/es/what-is/etl/>. [Accedido: 20 nov. 2023].
- [21] "¿Qué es un almacén de datos?," SAP. [En línea]. Disponible en: <https://www.sap.com/latinamerica/products/technology-platform/datasphere/what-is-a-data-warehouse.html>. [Accedido: 20 nov. 2023].
- [22] "¿Qué es un almacén de datos?," Amazon Web Services. [En línea]. Disponible en: <https://aws.amazon.com/es/what-is/data-store/>. [Accedido: 26 nov. 2023].
- [23] E. Ibrahim, "Improved feature selection using a hybrid side-blotched lizard algorithm and genetic algorithm approach," **Int. J. Power Electron. Drive Syst.**, vol. 13, no. 5, pp. 5737-5746, 2023. doi: 10.11591/ijece.v13i5.pp5737-5746.
- [24] G. Kapoor and N. Wichitakorn, "Electricity price forecasting in New Zealand: A comparative analysis of statistical and machine learning models with feature selection," **Appl. Energy**, vol. 347, pp. 121446-121446, 2023. doi: 10.1016/j.apenergy.2023.121446.
- [25] J. E. Sweetlin and S. Saudia, "Performance Analysis of a new Filter and Wrapper Sequence for the Survivability Prediction of Breast Cancer Patients," **Int. J. Elect. Comput. Eng. Syst.**, vol. 14, no. 5, pp. 547-555, 2023. doi: 10.32985/ijeces.14.5.6.
- [26] (2023). "A review on machine learning." **Int. J. Sci. Res. Arch.**, vol. 9, no. 1, pp. 281-285, 2023. doi: 10.30574/ijrsra.2023.9.1.0410.
- [27] A. Sinha, E. Bernardes, R. Calderon, and T. Wuest, "Machine Learning, AI, and Robotics," in **Digital Supply Networks: Transform Your Supply Chain and Gain Competitive Advantage with Disruptive Technology and Reimagined Processes**, 1st ed. New York: McGraw Hill, 2020, chap. 4. [Online]. Available: https://www.accessengineeringlibrary.com/content/book/9781260458190/chapter/chapter_4
- [28] M. Gopal, **Applied Machine Learning**, 1st ed. New York: McGraw-Hill Education, 2019. [Online]. Available: <https://www.accessengineeringlibrary.com/content/book/9781260456844>

- [29] Ahmad Kamsuriah and Siti Zarith and Itmin Sofia, "The success factors of extract-load-transform process in data integration implementation," **Journal of Information System and Technology Management**, vol. 7, no. 27, pp. 243-256, 2022. doi: 10.35631/jistm.727019
- [30] V. R. Konasani and S. Kadre, "Introducción al aprendizaje máquina y aprendizaje profundo," in **Aprendizaje Automático y Aprendizaje Profundo con Python y TensorFlow**, 1st ed. New York, McGraw Hill, 2021. [Online]. Available: https://www.accessengineeringlibrary.com/content/book/9781260462296/chapter/chapter_1
- [31] M. Taddy, **Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions**, 1st ed. New York, McGraw Hill, 2019. [Online]. Available: <https://www.accessengineeringlibrary.com/content/book/9781260452778>
- [32] C. Srihari, K. V. Vanishree, y R. Sahoo, "An extensive review of applications, methods and recent advances in deep reinforcement learning", pp. 1-6, 2023, doi: 10.1109/HORA58378.2023.10156687.
- [33] X. Zhu, A. Goldberg, R. Brachman y T. G. Dietterich, "Introduction to Semi-Supervised Learning", 2009.
- [34] Y. Qi, J. Yang, P. Li y M. Huang, "Machine-learning design of graphene nanoribbon waveguide side-coupled absorber," *Modern Physics Letters B*, 2023, doi: 10.1142/s0217984923500653.
- [35] A. B. Sallow y S. Almufti, "eXtreme Gradient Boosting Algorithm with Machine Learning: a Review," *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320-334, 2023, doi: 10.25007/ajnu.v12n2a1612.
- [36] M. Angelini, G. Blasilli, S. Lenti y G. Santucci, "A Visual Analytics Conceptual Framework for Explorable and Steerable Partial Dependence Analysis," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1-1, 2023, doi: 10.1109/TVCG.2023.3263739.
- [37] H. Zhao y X. Fu, "Comparison of Machine Learning Based on Category Theory," *Journal of Web Engineering*, vol. 22, pp. 41-54, 2023, doi: 10.13052/jwe1540-9589.2213.
- [38] A. K. Singh, "Evaluations with AIEd systems," en: ISBN 9781800375413, pp. 505-523, 2023, doi: 10.4337/9781800375413.00034.
- [39] (2023). "An e-learning evaluation method with the effectiveness of the questionnaire tool in the VLEs." doi: 10.1109/edunine57531.2023.10102811.

- [40] A. Behnood, «Soil and clay stabilization with calcium- and non-calcium-based additives: A state-of-the-art review of challenges, approaches and techniques,» *Transportation Geotechnics*, vol. 17, pp. 14 - 32, 2018.
- [41] B. H. Bujang , A. Alias y A. A. Azlan , «Evaluation, Selection and Assessment of Guidelines for Chemical Stabilization of Tropical Residual Soils,» *American Journal of Environmental Sciences*, vol. 4, nº 4, 2008.
- [42] G. D. López y H. Tobías, «Nanotecnología aplicada a la estabilización de suelos: desempeño técnico,» de 13º Congreso Internacional en Ciencia y Tecnología de Metalurgia y Materiales, Santa Fe, 2013.
- [43] J. F. Rivera, A. Aguirre-Guerrero, R. Mejía de Gutiérrez, y A. Orobio, «Estabilización química de suelos - Materiales convencionales y activados alcalinamente (revisión)», *Inf. tec.*, vol. 84, n.º 2, pp. 202–226, may 2020.
- [44] Z. Vaňková et al., "Soil moisture influences performance of selected stabilizing amendments in soil remediation," *Geoderma*, vol. 402, p. 115307, 2021. DOI: 10.1016/J.GEODERMA.2021.115307.
- [45] R. Chen et al., "Stabilization of soft soil using low-carbon alkali-activated binder," *Environmental Earth Sciences*, vol. 79, no. 22, pp. 1-13, 2020. DOI: 10.1007/S12665-020-09259-X.
- [46] M. J. Alzaidy, "Stabilization of Soils Using Chemical Admixtures: A Review," vol. 27, no. 1, pp. 51-62, 2019. DOI: 10.29196/JUBES.V27I1.1970.
- [47] N. Sagidullina et al., "Influence of Freeze–Thaw Cycles on Physical and Mechanical Properties of Cement-Treated Silty Sand," *Sustainability*, vol. 14, no. 12, pp. 7000-7000, 2022. DOI: 10.3390/su14127000.
- [48] Z. Shen et al., "Stabilization-based soil remediation should consider long-term challenges," *Frontiers of Environmental Science & Engineering in China*, vol. 12, no. 2, p. 16, 2018. DOI: 10.1007/S11783-018-1028-9.
- [49] K. A. Saeed, "Pozzolan materials for stabilization/solidification of soil contaminated by heavy metals - a review," *J. Eng. Sustain. Dev.*, vol. 27, no. 4, pp. 487-498, 2023. doi: 10.31272/jeasd.27.4.6
- [50] M. A. Adnan, S. Kumar, N. Garg, S. K. Das, "Soil stabilization using waste 'Bagasse ash and lime': A review," *Materials Today: Proceedings*, 2023. doi: 10.1016/j.matpr.2023.05.018

- [51] X. Fengling, W. Dongxing, "Review on Soil Solidification and Heavy Metal Stabilization by Microbial-Induced Carbonate Precipitation (MICP) Technology," *Geomicrobiology J.*, vol. 40, no. 6, pp. 503-518, 2023. doi: 10.1080/01490451.2023.2208113
- [52] T. Zafar, M. A. Ansari, A. Husain, "Soil stabilization by reinforcing natural and synthetic fibers – A state of the art review," *Materials Today: Proceedings*, 2023. doi: 10.1016/j.matpr.2023.03.503
- [53] "Stabilization of Soil Using Concrete Slush Waste and Polyester Fibres," *Int. J. Sci. Technol. Eng.*, vol. 11, no. 6, pp. 4024-4029, 2023. doi: 10.22214/ijraset.2023.54376
- [54] J. Zhai, "Machine-Learning-Based Road Soft Soil Foundation Treatment and Settlement Prediction Method," *Scientific Programming*, vol. 2022, pp. 1-7, 2022. DOI: 10.1155/2022/3463413.
- [55] T. Naidoo, J. J.P. Verster, S. Marais, "A Cloud-Based Road Infrastructure Analysis System using Machine Learning," *MATEC Web of Conferences*, vol. 370, pp. 07010-07010, 2021. DOI: 10.1051/mateconf/202237007010.
- [56] P. Lindh, P. Lemenkova, "Laboratory Experiments on Soil Stabilization to Enhance Strength Parameters for Road Pavement," *Transport and Telecommunication*, vol. 24, no. 1, pp. 73-82, 2023. DOI: 10.2478/ttj-2023-0008.
- [57] D. Kumar, D. Harinder, "ANN and SVM Based modeling of stabilized soils," doi: 10.1109/I-SMAC49090.2020.9243557, 2020.
- [58] A. Singh, M. Maurya, C. Singh, "Stabilization of Soil in Road Construction Using Lime and Fly Ash," *International Journal For Science Technology And Engineering*, vol. 11, no. 5, pp. 2571-2577, 2023. DOI: 10.22214/ijraset.2023.52163.
- [59] S. Suman, M. Mahamaya, S.K. Das, "Prediction of Maximum Dry Density and Unconfined Compressive Strength of Cement Stabilised Soil Using Artificial Intelligence Techniques," *Int. J. of Geosynth. and Ground Eng.*, vol. 2, no. 11, 2016. DOI: 10.1007/s40891-016-0051-9.
- [60] H. Javdanian, S. Lee, "Evaluating unconfined compressive strength of cohesive soils stabilized with geopolymer: a computational intelligence approach," *Engineering with Computers*, vol. 35, pp. 191-199, 2019. DOI: 10.1007/s00366-018-0592-8.
- [61] Gajurel, [et al.]. "Machine learning methods to map stabilizer effectiveness based on common soil properties," *Transportation Geotechnics*, vol. 27, p. 100506, 2021. DOI:

- 10.1016/j.trgeo.2020.100506. [Online]. Available:
<https://www.sciencedirect.com/science/article/pii/S2214391220303949>
- [62] H. MolaAbasi, A. Khajeh, and R. Jamshidi Chenari, "Use of GMDH-type neural network to model the mechanical behavior of a cement-treated sand," *Neural Comput & Applic*, vol. 33, pp. 15305-15318, 2021. DOI: 10.1007/s00521-021-06157-6. [Online]. Available: <https://doi.org/10.1007/s00521-021-06157-6>
- [63] J. A. Barbosa Tinoco, A. G. Correia Gomes, P. Cortez, and D. G. Toll, "Artificial neural networks for soil embankments stability condition identification," doi: 10.32075/17ECSMGE-2019-0605, 2019.
- [64] J. Khatti, "Evaluation of Compactive Parameters of Soil Using Machine Learning," pp. 1-14, 2022. doi: 10.1007/978-981-19-6998-0_1.
- [65] M. S. Ouf, "Towards Sustainability: Artificial Intelligent-Based Approach For Soil Stabilization Using Various Pozzolans," *Emerging Infectious Diseases*, vol. 162, pp. 253-262, 2012. doi: 10.2495/EID120231.
- [66] Z. T. Woubishet and K. A. Abegaz, "Artificial Intelligence for Prediction of Physical and Mechanical Properties of Stabilized Soil for Affordable Housing," *Appl. Sci.*, vol. 11, no. 16, pp. 7503, 2021. [Online]. Available: <https://doi.org/10.3390/APP11167503>
- [67] A. H. Saad et al., "A Systematic Review of Machine Learning Techniques and Applications in Soil Improvement Using Green Materials," *Sustainability*, vol. 15, no. 12, pp. 9738, 2023. [Online]. Available: <https://doi.org/10.3390/su15129738>
- [68] K. Yones and B. A. Miller, "Selecting Appropriate Machine Learning Methods for Digital Soil Mapping," *Applied Mathematical Modelling*, vol. 81, pp. 401-418, 2020. [Online]. Available: <https://doi.org/10.1016/J.APM.2019.12.016>
- [69] R. Hudaykulov and T. Mirzayev, "The Use of Stabilizers to Improve the Strength of the Soil Foundation of Roads," *Journal Title*, vol. 6, no. 1, doi: 10.15862/14SATS119, 2018.
- [70] M. Taha, Z. Majeed, S. Ahmed, and S. Mustafa, "Artificial Neural Network Prediction Models for Maximum Dry Density and Optimum Moisture Content of Stabilized Soils," *Journal Title*, vol. 5, no. 2, pp. 146-168, doi: 10.1007/S40515-018-0053-2, 2018.
- [71] A. Baghbani, T. Choudhury, S. Costa y J. Reiner, "Application of artificial intelligence in geotechnical engineering: A state-of-the-art review," *Earth-Science Reviews*, vol. 228, p. 103991, 2022. DOI: 10.1016/j.earscirev.2022.103991.

- [72] K. Briney, "Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success," Editorial: Pelagic Publishing, ISBN: 978-1784270089.
- [73] Wall, S. (2017). Data Management for Researchers: Organize, Maintain and Share Your Data for Research Success. Archives and Manuscripts, 45(1), 53-55. DOI: 10.1080/01576895.2017.1279038.
- [74] IBM. (2023, Diciembre 1). ¿Qué es la ciencia de datos?. [En línea]. Disponible en: <https://www.ibm.com/es-es/topics/data-science>. Acceso: Diciembre 1, 2023.
- [75] IBM. (2023, Diciembre 1). ¿Qué es el aprendizaje supervisado?. [En línea]. Disponible en: <https://www.ibm.com/es-es/topics/supervised-learning>. Acceso: Diciembre 1, 2023.
- [76] M. T. Bergado, N. Miura, and A. Balasubramaniam, "Lime stabilization of clayey subgrade soils for road construction," Journal of the Transportation Research Board, vol. 1721, pp. 49-58, 2000.
- [77] J. R. Kercher, "Innovative techniques in soil stabilization," in Proceedings of the International Conference on Ground Improvement, 2012, pp. 85-95.
- [78] A. A. Gidley and S. K. Sack, "Cement stabilization of compacted clay," Journal of the Geotechnical Engineering Division, vol. 104, no. 9, pp. 1125-1142, 1978.
- [79] A. B. Alrawas and A. T. Al-Harthy, "Effect of lime, cement and artificial pozzolana on the swelling potential of an expansive soil from Oman," Building and Environment, vol. 39, no. 3, pp. 359-367, 2004.
- [80] J. T. Santoni, J. L. Tingle, and S. L. Webster, "Stabilization of silty sand with nontraditional additives," Transportation Research Record, no. 1787, pp. 61-70, 2002.
- [81] S. Zhan and Z. Li, "Effect of nanomaterials on soil stabilization: A review," Construction and Building Materials, vol. 95, pp. 107-114, 2015.
- [82] P. Kumar and K. J. Osborne, "Puzolanic activity of fly ash in stabilization of expansive soils," International Journal of Geotechnical Engineering, vol. 2, no. 3, pp. 177-182, 2016.
- [83] C. H. Benson and E. T. Seabra, "Mineral and geosynthetic stabilization of marginal soils," Journal of Geotechnical and Geoenvironmental Engineering, vol. 140, no. 4, pp. 1-14, 2014.

- [84] M. K. Ranganathan and L. H. Salvadori, "Organic soil stabilization using natural and synthetic polymers," *Journal of Applied Polymer Science*, vol. 101, no. 5, pp. 2950-2958, 2006.
- [85] I. Sommerville, *Software Engineering*, International Computer Science Series. Pearson, 2011. Disponible en: <https://books.google.com.co/books?id=lOegcQAACAAJ>
- [86] K. Schwaber y J. Sutherland, *The Scrum Guide: The Definitive Guide to Scrum*, 2020. Disponible en: <https://scrumguides.org/docs/scrumguide/v2020/2020-Scrum-Guide-US.pdf>
- [87] J. Humble y D. Farley, *Continuous Delivery: Reliable Software Releases through Build, Test, and Deployment Automation*. Addison-Wesley, 2010.
- [88] P. Chapman et al., *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc., 2000.
- [89] U. Fayyad, G. Piatesky-Shapiro, y P. Smyth, "From Data Mining to Knowledge Discovery in Databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.
- [90] M. A. Waller y S. E. Fawcett, "Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management," *Journal of Business Logistics*, vol. 34, no. 2, pp. 77–84, 2013.
- [91] E. Soria-Olivas, M. Martínez-Sober, J. Serrano López, A. Guerrero Martínez y M. D. Gorriz Sáez, *Sistemas de Aprendizaje Automático*. Madrid, España: Ra-Ma Editorial, 2021.
- [92] ASTM D2487, "Standard Practice for Classification of Soils for Engineering Purposes (Unified Soil Classification System)," ASTM International, 2017.
- [93] Instituto Nacional de Vías (INVIAS), "Especificaciones Generales de Construcción de Carreteras," Bogotá, Colombia, 2018.
- [94] ISO/IEC 22989:2022, "Information Technology – Artificial Intelligence – Concepts and Terminology," International Organization for Standardization, 2022.
- [95] UNESCO, "Recommendation on the Ethics of Artificial Intelligence," Paris, France, 2021.
- [96] Naciones Unidas, "Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible," Nueva York, 2015.
- [97] Ministerio de Ambiente y Desarrollo Sostenible, "Política Nacional de Gestión Integral de Residuos," Bogotá, Colombia, 2020.