ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Predicción de resistencia a la compresión en geopolímeros por medio de técnicas de aprendizaje de máquina.

PROYECTO INTEGRADOR

Previo la obtención del Título de:

Ingeniero en Computación

Presentado por:
Jhossias Pawhell Calderón Lopez
Eddy Ronaldo Calderón García

GUAYAQUIL - ECUADOR
Año: 2022-2023

DEDICATORIA

Dedico este proyecto en primer lugar a Dios, que me ha dado la vida, la fuerza e iluminado mi mente siempre, a las personas que me acompañaron en todo momento como mis padres que siempre estuvieron apoyándome con todo, a mis hermanos que han estado siempre, a mi esposa que en los últimos años ha sido un pilar fundamental en mi vida y a todas las personas que confían en mi día a día.

Jhossias Calderón.

Dedico esta tesis a Dios por permitirme concluir mi carrera, a mis padres por todo su sacrificio y esfuerzo. A toda mi familia, que siempre estuvieron prestos a apoyarme, a la memoria de mis tíos Agustín y Francisco que fueron parte fundamental de este logro. A mis amigos y compañeros en especial a Nicole, Carlos y Ginger gracias por acompañarme a lo largo de la carrera, por enseñarme a perseverar y creer en mí.

Eddy Calderón.

AGRADECIMIENTOS

Agradecemos a las personas que hicieron posible la consecución de este proyecto, a nuestro tutor Miguel Realpe, que, con su guía y sus enseñanzas, sacamos adelante la elaboración del mismo, a Haci Baykara, que, con su amplia experiencia y conocimientos nos ayudó a entender el mundo de los materiales y que también con su guía sacamos el producto final, y a todos los profesores a lo largo de nuestras carreras que hicieron posible que apliquemos sus enseñanzas.

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; Jhossias Calderón López, Eddy Calderón García y damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Jhossias Calderón

Eddy Calderón

EVALUADORES

MIGUEL ANDRES REALPE ROBALINO

Digitally signed by MIGUEL ANDRES REALPE ROBALINO
DN: c=EC, o=BANCO CENTRAL DEL ECUADOR, ou=ENTIDAD DE CERTIFICACION DE INFORMACION-ECIBCE, l=QUITO, serialNumber=0000793814, cn=MIGUEL ANDRES REALPE ROBALINO
Date: 2023.02.11 14:18:25 -05'00'

.....

Miguel Realpe Robalino

PROFESOR DE LA MATERIA/TUTOR

RESUMEN

Determinar un modelo que realice la predicción para conseguir la mejor resistencia a la compresión en un geopolímero es primordial para que éste, junto a todas sus demás propiedades, y el ser amigable con el medioambiente, termine siendo la mejor opción contra el cemento portland el cual está en desventaja comparando sus propiedades fisicoquímicas como su resistencia y su producción que termina siendo una de las principales detonadoras del calentamiento global. Para ello se realizó un análisis exploratorio de datos con la información obtenida en base a las pruebas experimentales con variaciones en las mezclas hechas por la Facultad de Ingeniería Mecánica y Ciencias de la Producción cuya correlación de variables nos permitieron escoger el mejor modelo de aprendizaje de máquina para la predicción con un menor error. Dentro de las variables con mayor correlación con la resistencia a la compresión se encuentran el contenido de NaOH y Na2SiO3, lo que permitió determinar que CatBoost es el mejor modelo, tomando en cuenta que se realizaron también mezclas de diferentes modelos, pero no se obtuvieron mejores resultados. Además, con el modelo, se implementó una interfaz sencilla y amigable para su uso y conseguir las predicciones deseadas. Lo obtenido se asemeja a la literatura en la cual los mejores modelos se basan en algoritmos bajo árboles de decisión y boosting debido a la importancia que les dan a las variables que tienen mayor correlación con la resistencia.

Palabras Clave: Aprendizaje Máquina, Boosting, Geopolímeros, Predicción

ABSTRACT

Determining a model that performs the prediction to achieve the best compressive strength in a geopolymer is essential so that this, along with all its other properties, and being environmentally friendly, ends up being the best option against portland cement which is at a disadvantage comparing its physicochemical properties such as its strength and its production which ends up being one of the main triggers of global warming. For this purpose, an exploratory data analysis was carried out with the information obtained based on the experimental tests with variations in the mixtures made by the Faculty of Mechanical Engineering and Production Sciences whose correlation of variables allowed us to choose the best machine learning model for the prediction with a lower error. Among the variables with the highest correlation with compressive strength are NaOH and Na2SiO3 content, which allowed us to determine that CatBoost is the best model, considering that mixtures of different models were also made, but no better results were obtained. In addition, with the model, a simple and friendly interface was implemented for its use and to achieve the desired predictions. The obtained results are like the literature in which the best models are based on algorithms under decision trees and boosting due to the importance given to the variables that have a higher correlation with resistance.

Keywords: *Machine Learning, Boosting, Geopolymers, Prediction*

ÍNDICE GENERAL

ABREVIATURAS

ESPOL Escuela Superior Politécnica del Litoral

FIMCP Facultad de Ingeniería en Mecánica y Ciencias de la Producción

LEMAT Laboratorio de Evaluación de Materiales

SIMBOLOGÍA

ml	Mililitro
uL	Microlitro
M	Molar
ºC	Celsius
MPa	Megapascal
Na2SiO3	Silicato de Sodio
NaOH	Hidróxido de Sodio

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

La incursión en el campo de los geopolímeros a nivel mundial crece a medida que pasan los años y aumenta el grado de priorización de la conservación del medio ambiente, administración de recursos, optimización de costos, etc. Por lo tanto, el enfoque a crear un nuevo material de construcción es verdaderamente viable conociendo las bondades de este. Sin embargo, para llevar a cabo estos estudios, es necesario un correcto análisis exploratorio de datos obtenidos con ciertas pruebas realizadas y la inclusión del aprendizaje de máquinas con el fin de obtener modelos que arrojen predicciones confiables y con un bajo margen de error de métricas que, de no ser con esta implementación, los costos para pruebas experimentales acabarían con el interés de llevar a cabo el proyecto. En este capítulo se describe la problemática a nivel ambiental y estructural que existe a nivel de la industria de la construcción, así como la alternativa para lidiar con el problema, en base las revisiones bibliográficas que se realizaron en la investigación.

1.1 Descripción del problema

Estudios en geopolímeros representan un campo en crecimiento constante en referencia a su usabilidad como materiales de construcción amigables con el medioambiente. En las últimas décadas, la preocupación por la emisión de dióxido de carbono al producir cemento portland común que representa un 8% de emisiones globales anuales de dióxido de carbono, ha desencadenado una proliferación de estudios de diferentes materias primas para la síntesis de geopolímeros o geopolimerización tales como ceniza volante, desechos industriales, metacaolín, entre otras. Otros de los beneficios de distintas soluciones alternativas, además de ser amigables con el medio ambiente, es la resistencia muy por encima de otros materiales, y el poder soportar temperaturas extremas sin problemas.

Ecuador está incursionando en el campo de geopolímeros, en la actualidad, existe un material de construcción cuyas propiedades ofrecen ventajas competitivas con respecto a materiales tradicionales. ESPOL, y su facultad FIMCP (Facultad de Ingeniería en Mecánica y Ciencias de la Producción), con la ayuda del docente e

investigador Haci Baykara, son los precursores de esta innovación. Este material, compuesto por aluminosilicatos cristalinos o zeolitas naturales, que son materia prima local en abundancia debido a su geología y su ubicación en el cinturón de fuego, en adición a arena de río que se utiliza para la síntesis del geopolímero y las soluciones activadoras alcalínicas está en un período de pruebas tomando distintas proporciones de sus componentes. Por tal motivo, se desea encontrar la proporción exacta para obtener los mejores resultados posibles.

Es necesario predecir la resistencia a la compresión de morteros geopoliméricos utilizando bases de datos experimentales tabulados en Excel donde varían las proporciones de los materiales y parámetros de elaboración. Con ello, se desea reducir considerablemente el número de pruebas a realizar y utilizar la cantidad de materiales de forma óptima para lograr una resistencia mayor a las pruebas experimentales, y por supuesto, a la del cemento portland común (21 MPa).

1.2 Justificación del problema

Un modelo de predicción de resistencia a la compresión en geopolímeros trae consigo una reducción en la cantidad de pruebas experimentales que se deben realizar y el poder determinar cuál es la cantidad exacta de los componentes a mezclarse para lograr obtener un geopolímero con la mayor resistencia posible, siendo una de las características positivas, que junto a otros beneficios, concluyen en una alternativa óptima para la sustitución parcial de materiales ordinarios tal como el cemento portland.

La implementación del modelo implica la obtención de un material con los mayores estándares en resistencia y altamente viable que proporcionan una alta competitividad con respecto a otras alternativas, beneficiando a la industria de la construcción con sus bondades y el impacto positivo que tiene con respecto al cuidado del medio ambiente.

1.3 Objetivos

1.3.1 Objetivo General

Implementar un modelo de predicción de la resistencia a la compresión en geopolímeros por medio de técnicas de aprendizaje de máquina con la mayor precisión de predicción para optimizar la mezcla.

1.3.2 Objetivos Específicos

1. Realizar un análisis exploratorio de datos experimentales para encontrar, a través de métodos de visualización de datos, relaciones entre las variables.
2. Entrenar diversos modelos de aprendizaje de máquina para obtener variabilidad de los resultados.
3. Determinar el modelo o la combinación de modelos con mayor precisión para predecir los datos de compresión y optimizar la mezcla del geopolímero.

1.4 Marco teórico

1.4.1 Geopolímeros

La producción de cemento portland consume una gran cantidad de energía y materia prima, al mismo tiempo que emite entre un 5-7% de dióxido de carbono, el cual es responsable del calentamiento global. Por tal motivo, se está reemplazando el cemento tradicional con materiales cementantes suplementarios (Singh & Middendorf, 2020).

En las últimas décadas, el estudio de geopolímeros ha tenido un repunte importante debido a los beneficios que se obtienen en comparación con materiales tradicionales. Un geopolímero es un polímero inorgánico que se forma a través de la reacción de policondensación de materiales que contienen aluminosilicatos en condiciones alcalinas, por ello el uso de activadores de silicatos alcalinos y siguiendo condiciones de curado específicas (Singh & Middendorf, 2020).

Figura 1.1 Diagrama de flujo esquemático para la producción de geopolímeros [Shehata et al., 2021].

1.4.1.1 Estudios previos en geopolímeros

Debido al auge de los estudios en geopolímeros como alternativa amigable con el medio ambiente, más resistente y duradera, ESPOL realizó un estudio utilizando zeolita natural, que es un aluminosilicato cristalino, arena de río y soluciones activadoras alcalínicas (Hidróxido de sodio y Silicato de Sodio) para obtener la resistencia a la compresión máxima, que es uno de los indicadores relevantes al comparar materiales de construcción en base a las proporciones de la mezcla y las características específicas de curado (Ulloa et al., 2022).

Con las pruebas experimentales realizadas en la investigación de (Ulloa et al., 2022), se obtuvo la mejor resistencia a la compresión, alcanzando los 17MPa, con un curado a 60°C y una mezcla con las siguientes proporciones:

- NAOH: 14M
- Na2SiO3/NaOH: 3
- Activador/Zeolita: 0.5
- Arena/Zeolita: 1.5

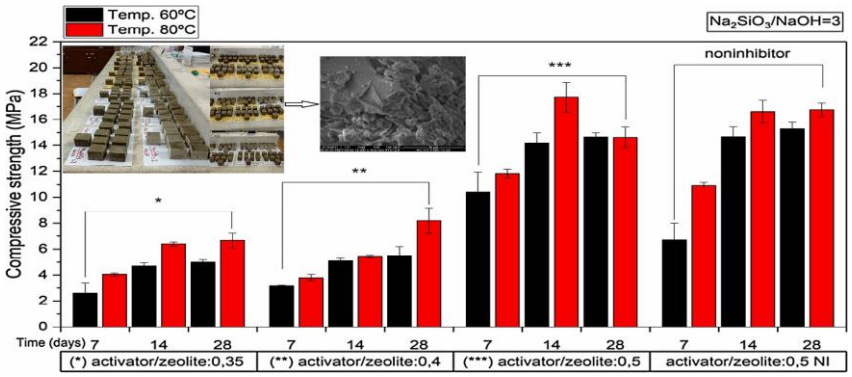

Figura 1.2 Resistencia a la compresión con proporción Na2SiO3/NaOH: 3 [Ulloa et al., 2022].

1.4.2 Análisis exploratorio de datos

El Análisis Exploratorio de Datos (AED) es uno de los primeros pasos para el descubrimiento de nuevos conocimientos de cualquier índole, en el que los científicos de datos exploran de una manera interactiva distintos conjuntos de datos en base a una serie de operaciones de estadística descriptiva encontrando promedios, medianas, modas, etc. Para luego realizar operaciones de análisis como filtros, agregaciones y visualización de datos, que es una técnica donde los datos se presentan en formatos gráficos para obtener nueva información (Milo & Somech, 2020).

Mientras los datos o la información son aquellos resultados que se obtienen al realizar pruebas, experimentos, operaciones; el análisis de datos es un proceso en el que se lleva a cabo una limpieza, transformación y modelamiento para poder extraer información útil para su uso en negocios, industrias, organizaciones, etc. Para la realización de este proceso existen actualmente librerías tales como: Numpy, Sklearn, Seaborn, Pandas, etc.

Sklearn: Scikit-learn es una de las librerías más útiles para aprendizaje de máquina en el lenguaje de programación Python, incluyendo métodos de modelado de datos y un sinnúmero de herramientas para regresiones, agrupamientos, reducciones de dimensionalidad, clasificación, etc. (Nongthombam & Sharma, 2021).

Seaborn: Seaborn es una librería de visualización de datos en Python para realizar gráficos estadísticos basada en matplotlib que también es una librería de visualización de datos e integrada con las estructuras de datos de Pandas que es usada para convertir datos en forma tabular, lo que ayuda a su entendimiento. Cuando recibe un conjunto de datos, Seaborn asigna los valores a atributos visuales tales como colores, tamaños, estilos y decora el gráfico con etiquetas de ejes, leyendas, entre otras. (Waskom, 2021).

1.4.2.1 Pasos del análisis exploratorio de datos

1. Recopilación y exploración de datos: Los datos se obtienen a través de distintas fuentes para su exploración que se basa en identificar el contenido, las características, el tamaño y la posible relación entre los datos visualizándose mediante datos estructurados.
2. Limpieza de datos: Consiste en escanear los datos inconsistentes, duplicados, irrelevantes, atípicos, los cuales nos impiden tener los resultados deseados, para removerlos y obtener datos de calidad, esenciales para un correcto análisis. (Sahoo* et al., 2019).
3. Análisis exploratorio de datos: Según explican (Nongthombam & Sharma, 2021), en este punto los conjuntos de datos están libres de errores y pueden ser analizados usando Python y sus librerías que generan gráficos estadísticos multidimensionales.
   a. Identificación de las variables: Se identifican y ordenan las variables para su análisis. (Rao et al., 2021).
      i. Tipos de datos: Los tipos de datos más comunes en Python son enteros, objetos, flotantes, los cuales se imprimen a través de la función .dtypes según explican (Nongthombam & Sharma, 2021).
      ii. Describiendo el conjunto de datos: Se realiza un resumen obteniendo media, conteo, máximo, mínimo, etc. a través de la función describe(). (Nongthombam & Sharma, 2021).
   b. Análisis de variables univariadas, bivariadas y multivariadas: Se identifican las correlaciones que existen entre los atributos (Rao et al., 2021).
   c. Manejo de datos faltantes y datos aberrantes: Al tener datos faltantes o aberrantes se pueden producir resultados no deseados o inexactos, por tal motivo la forma de solucionar este problema es eliminándolos (Rao et al., 2021) y, de ser posible, utilizando media, mediana y moda para rellenar esos campos vacíos.
   d. Visualización de datos
      i. Univariada: Proporciona un resumen para cada variable del conjunto de datos.

1. Histograma: Un histograma es un gráfico de barras que es utilizado para ilustrar conteo, porcentaje, distribuciones de los datos, etc. a lo largo de sus ejes X y Y. (Nongthombam & Sharma, 2021).
2. Diagrama de tallo: En este diagrama los datos se agrupan tallos y hojas, donde el tallo es el dígito mayor y la hoja el dígito menor. Según (Nongthombam & Sharma, 2021) este es utilizado comúnmente para comparar datos y resalta la moda.
3. Diagrama de caja: El diagrama de caja muestra una comparación de los grupos de datos representando el centro, tendencia, sesgo, simetría y valores atípicos, utilizando mediana, cuartiles, mínimo y máximo (Nongthombam & Sharma, 2021).

ii. Bivariada: Comprensión de las conexiones entre dos variables del conjunto. Entre los ejemplos están el diagrama de violín y de caja (Sahoo et al., 2019).

iii. Multivariada: Relaciones entre dos o más variables. Se tienen gráficos de pares y de dispersión 3D.
    1. Gráfico de dispersión: En este gráfico se utilizan puntos que indican valores para las diferentes variables ubicados en los ejes horizontal y vertical (Nongthombam & Sharma, 2021).
    2. Mapa de calor: Utiliza tonos de colores en una tabla bidimensional para representar las variables y su relación.

1.4.3 Modelos de Aprendizaje de Máquina

El Aprendizaje de Máquina sirve para crear modelos de regresión y clasificación a partir de conjuntos de datos etiquetados. Proporciona una forma fácil de modelar un sistema para investigar fenómenos en diversas áreas científicas. Por lo expuesto, conocer de las técnicas de Aprendizaje de Máquina abre un abanico de posibilidades para resolver y entender estos fenómenos (Hancock & Khoshgoftaar, 2020).

1.4.3.1 CatBoost Regressor

En investigaciones de Khoshgoftaar. CatBoost es un algoritmo de código abierto basado en árboles de decisión con salida a finales de 2018, con éxito en aprendizaje automático en Big Data, es muy aplicado en el aprendizaje automático de datos con heterogéneos y categóricos. En la literatura recomiendan un ajuste a los hiperparámetros ya que se muestra con una sensibilidad a los mismos también se ve que es usado en campos interdisciplinarios como Finanzas, Medicina, Astronomía, Biología, Bioquímica, Ciber Seguridad, Meteorología entre otros para cada una de las áreas detalla un estudio en específico (2020). Por lo tanto, CatBoost es una buena opción para el trabajo interdisciplinario.

1.4.3.2 Gradient Boosting Regressor

Introducido en 2011 con el objetivo de desarrollar un modelo que pueda minimizar la función de pérdida. El algoritmo cambia los pesos de las muestras para entrenar múltiples clasificadores, combinando los clasificadores linealmente para mejorar el rendimiento de la clasificación. (Gong et al., 2020) Es un algoritmo basado en árboles, la ventaja principal del GMB es utilizar menos recursos computacionales y evitar el sobreajuste a través de su función objetivo (Kaloop et al., 2020).

1.4.3.3 Decision Tree Regressor

Según Patil y Kuljarni, este algoritmo construye una estructura de árbol donde cada nodo no-hoja representa una evaluación de atributo y las hojas representan etiquetas de clase. Se analizan los datos de entrenamiento y se identifican los atributos con más información con respecto al resto, ya que tal atributo puede clasificar o categorizar de manera efectiva considerándose como nodo raíz (2019). Un árbol puede verse como una aproximación por partes constante y similar a un diagrama de flujo donde cada nodo interno denota una prueba en un atributo, cada rama representa un resultado y cada nodo de la hoja contiene una etiqueta de clase.

1.4.3.4 Random Forest Regressor

Random Forest es un algoritmo que selecciona un subgrupo de características al azar para crear un arreglo de árboles de decisión. (Rathakrishnan et al., 2022). En investigaciones de Farooq et al. Sirve para clasificación y regresión, a diferencia del Decisión Tree donde se construye un solo árbol, en Random Forest se utilizan varios árboles asemejándose a un bosque, se eligen arbitrariamente datos diferentes que se destinan a todos los árboles. Cada árbol proporciona una regresión, para estimar el error se combina los errores de cada árbol (2021).

1.4.3.5 Support Vector Regressor

Support Vector Machine aparece en 1992 por la necesidad de herramientas de regresión y clasificación basadas en predicciones. En este modelo se busca una línea de decisión denominado hiperplano, que separa los conjuntos de datos o clases evitando sobreajustes hacia una u otra clase, por medio de la introducción de un margen que mantiene al hiperplano lo más alejado posible de todas las clases (Somvanshi et al., 2016).

1.4.4 Recopilación de datos

El dataset utilizado en el presente proyecto contiene un total de 54 registros con muestras de mortero de geopolímero basado en zeolita junto con dos activadores alcalinos (NaOH 14M y Na2Si03) variando proporciones, arena de rio, agua añadida, temperatura de secado y tiempo de curado. Todas estas muestras fueron tomadas en el Laboratorio de Ensayos Metrológicos y de Materiales (LEMAT) en ESPOL (Ulloa et al., 2022). Un geopolímero es un sustituto del cemento donde se usan materiales puzolánicos ricos en silicios y aluminios junto con activadores alcalinos que sirven como aglutinantes según explican (Chithambar Ganesh & M. Muthukannan, 2018)

CAPÍTULO 2

2. METODOLOGÍA

El cumplimiento de los objetivos propuestos requiere una serie de pasos que se efectuaron con el fin de determinar cuál es el modelo o la fusión de modelos que utilizar para obtener una predicción confiable en términos de aproximación a la realidad. Dentro de los pasos que se siguieron está una preparación previa, que fue el levantamiento de requerimientos, los riesgos y cómo mitigarlos, y el diseño de cómo se va a solucionar el problema. Dentro de este diseño consta un análisis exploratorio de datos con un informe detallado de la relación entre las variables y su importancia en el estudio, además del entrenamiento de varios modelos de predicción de resistencia a la compresión para posteriormente implementar un API con una interfaz amigable que permita la interacción con los modelos. En este capítulo se describe la solución, la metodología para llevarla a cabo en base a los requerimientos propuestos, tomando en cuenta un análisis exploratorio y entrenamientos de modelos de aprendizaje de máquina.

2.1 Análisis

Realizar innumerables pruebas experimentales con distintas mezclas del geopolímero trae consigo un incremento sustancial en el uso de materiales, mala administración de recursos, demoras y otros aspectos que determinaron que no debe ser la única opción para determinar una resistencia a la compresión. Por tal motivo, se propuso la implementación de un modelo de aprendizaje máquina tradicional para predecir la resistencia a la compresión del material de construcción que, en base a una entrada de datos, se obtenga una resistencia aproximada a una prueba experimental, lo que ayudó a la optimización de recursos y su aportación a los estudios en el país.

2.1.1 Requerimientos

Dentro de los requerimientos funcionales y no funcionales que el cliente solicitó se encuentran los siguientes:

Tabla 2.1 Requerimiento Funcional 1

RF-1	Análisis exploratorio de datos
Versión	1.0
Tipo	Desarrollo
Dependencias	
Descripción	Se podrá visualizar un análisis exploratorio de datos que aporten información relevante al estudio
Comentarios	Ninguno
Criterio de validación	**Visualización** de análisis exploratorio de datos a través de un reporte

Tabla 2.2 Requerimiento Funcional 2

RF-2	Modelo de aprendizaje máquina óptimo para predicción
Versión	1.0
Tipo	Desarrollo
Dependencias	
Descripción	Se podrá predecir la resistencia a la compresión del geopolímero con un modelo de aprendizaje máquina óptimo
Comentarios	Ninguno
Criterio de validación	**Casos de prueba** donde se valide que los resultados que se obtengan del modelo en cuestión se acerquen a los datos experimentales.

Tabla 2.3 Requerimiento Funcional 3

RF-3	Promediador de modelos
Versión	1.0
Tipo	Desarrollo
Dependencias	
Descripción	Todo público interesado podrá acceder al proyecto alojado en la nube.
Comentarios	Ninguno
Criterio de validación	**Validación** de la existencia del proyecto en la nube y con acceso a todo el público en general

Tabla 2.4 Requerimiento No Funcional 1

RNF-1	Proyecto disponible en la nube
Versión	1.0
Tipo	Desarrollo
Dependencias	
Descripción	Todo público interesado podrá acceder al proyecto alojado en la nube.
Comentarios	Ninguno
Criterio de validación	**Validación** de la existencia del proyecto en la nube y con acceso a todo el público en general

2.1.2 Alcance y limitaciones de la solución

Se propuso como alcance de la solución un análisis exploratorio para detectar información relevante, identificar relaciones entre las variables y la creación de modelos entrenados con un dataset de pruebas experimentales que realicen una predicción de resistencia a la compresión en geopolímeros.

Por otra parte, la limitación que se encontró es la limitada cantidad de pruebas experimentales, lo que compone un pequeño dataset, que es el que se utilizó para los entrenamientos a los modelos, además de los elementos químicos que quedan fijados sin oportunidad de agregar más o cambiar su composición. Por lo tanto, estas limitaciones, podrían tener como consecuencia un modelo sobreajustado o subajustado que determinaría una predicción inexacta.

2.1.3 Riesgos y beneficios de la solución

Dentro de los riesgos que se encuentran con la solución planteada, se presenta la posibilidad de que el modelo cargado en la nube pueda eliminarse por algún error, por tal motivo es importante poseer un respaldo de este para volver a cargarlo. Además, se encuentra el riesgo de predicciones erróneas si en los datos de entrada se consideran extremos máximos o mínimos que se alejan de los datos de entrenamiento del modelo, generando una incertidumbre epistémica que se da por la falta de datos de entrenamiento, por lo que, para prevenirlo, es necesaria una base de datos con los datos de entrada aproximados que se pueden ingresar.

Por otro lado, dentro de los beneficios de la solución tenemos que, dado que se entrenan varios modelos, se escoge al modelo o la fusión de modelos que provean una predicción con el menor error, lo que permite obtener un geopolímero altamente calificado para usarlo en la industria. En adición, existe el acceso de forma remota desde cualquier dispositivo.

2.2 Diseño de la solución

El diseño de la solución propuesta contiene una serie de pasos definidos con el principal objetivo de obtener la mejor predicción con un bajo margen de error posible. Por tal motivo, dentro de los pasos que se detallarán más adelante, se presenta la recopilación y exploración del conjunto de datos brindados por los investigadores, a los mismos se les realiza una limpieza de datos con el fin de que no existan resultados de baja calidad, para posteriormente, realizar un análisis exploratorio tomando en cuenta todas las variables del conjunto de datos y obtener información relevante para el estudio. Finalmente, obtener los mejores modelos para entrenarlos, realizar pruebas y concluir en el modelo o la fusión de modelos que devuelvan la resistencia a la compresión en geopolímeros más cercana a la realidad para su despliegue en la nube.

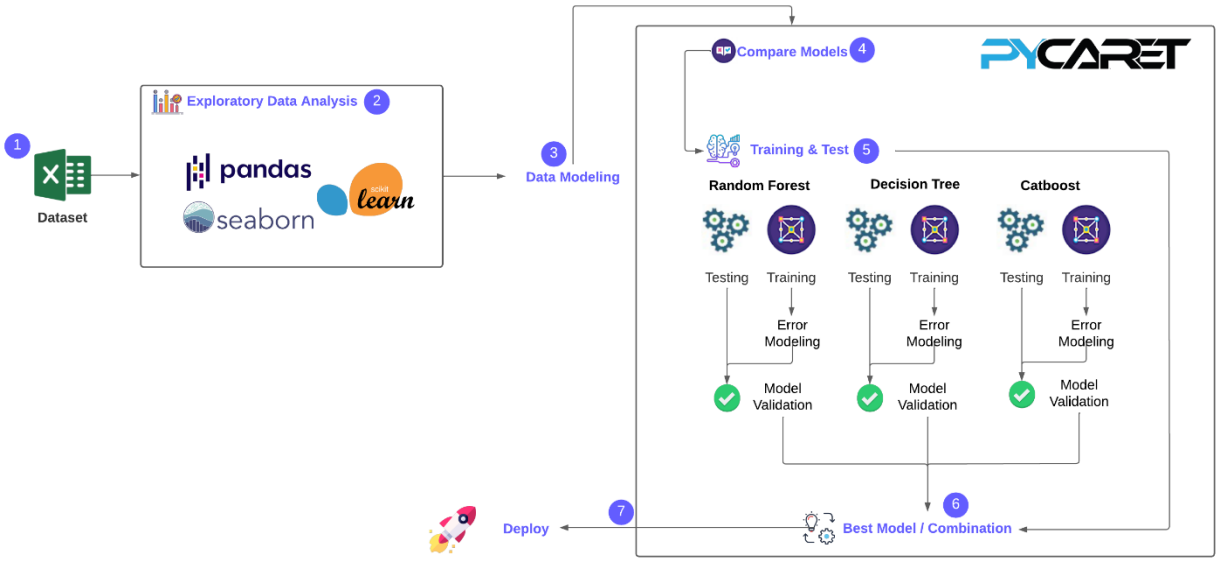

Figura 2.1 Diseño de la solución

2.2.1 Recopilación y exploración de datos

En primer lugar, se obtuvo el conjunto de datos que detalla las constantes, variables, sus valores y resultados de las pruebas experimentales realizadas a lo largo del estudio. Estas constantes y variables albergadas en el documento Excel son:

Tabla 2.5 Variables y constantes del conjunto de datos

Variable/Constante	Detalle
Mezcla	Identificador de la prueba experimental, contiene la relación del activador/zeolita y la temperatura a la que fue sometida.
Concentración de NaOH líquido en Molaridad M	Constante, en este caso es de 14 M
Relación Na2SiO3 NaOH	
Contenido de Na2SiO3 en mL	
Contenido de NaOH en mL	
Contenido de extra agua en mL	
Contenido de Arena en g	Constante, 135g.
Contenido de Zeolita en g	Constante, 90g,
Temperatura de Curado en Celsius	Tiempo de curado al que se sometió el geopolímero, tiene 2 posibles valores: 60° y 80°.
Contenido de inhibidor de corrosión en uL	Constante, 125 uL.
Relación Total líquido binder de zeolita	
Agua total en mL	Constante, 43,7 mL.
Relación agua total zeolita	Constante, 0,485 mL.
Relación Arena zeolita	Constante, 1,5
Relación activador zeolita	
Número de cubos	Cantidad de cubos tomados en la prueba experimental.
Total días de curado	Tiempo que tomó de curado, con posibilidades de 7, 14 y 28 días.
Resistencia Compresión en MPa	Resistencia obtenida en base a las pruebas.

2.2.2 Limpieza de datos

Una vez revisado el conjunto de datos que se obtuvo basados en pruebas experimentales, se realiza una limpieza de datos, en el cual se escanean los datos que tengan inconsistencias, datos duplicados, atípicos que impiden obtener resultados de calidad con un bajo margen de error para un correcto análisis.

2.2.3 Análisis exploratorio de Datos

Se realizó a cabalidad el análisis exploratorio de datos con el fin de obtener un contexto de las variables, su relación, su relevancia con respecto a lo que se desea predecir y realizar una representación visual de lo que se obtuvo.

Dentro de las herramientas que se utilizaron para realizar el análisis exploratorio de los datos se encuentra pandas profiling, el cual genera informes estadísticos de resumen y metadatos para un determinado conjunto de datos, proporcionando detalle de cada variable, correlaciones, valores duplicados y faltantes, etc. Se escogió esta herramienta por su facilidad de implementación, resultados de calidad para el entendimiento, análisis de datos, visualización gráfica y por su accesibilidad. Además de todas las funciones para estadística que nos incluye la librería pandas siendo nuestra base de todo el análisis exploratorio de datos.

2.2.3.1 Identificación de variables y descripción del conjunto de datos

En este punto, se obtuvieron los tipos de datos de las variables encontradas en el conjunto de datos, que podían ser categóricos o numéricos y se realizó un resumen de estas, con su media, varianza, desviación estándar, mínimo, máximo, cuantiles, etc. con la función describe(), la cual es parte de la librería pandas.

2.2.3.2 Manejo de datos faltantes y aberrantes

Para el manejo de datos faltantes y aberrantes, los cuales pueden producir resultados con baja calidad, inexactos o no deseados, se utilizó pandas profiling, que arroja cuáles son las constantes del conjunto de datos ayudándose de histogramas y matrices para su visualización y posterior eliminación.

2.2.3.3 Visualización de datos

Dentro de los análisis que se realizan para visualizar los datos se encuentran en análisis univariado, bivariado y multivariado, todo realizado a través de las funciones de la librería pandas y su herramienta, pandas profiling.

Para en análisis univariado se revisó la distribución de valores para cada variable del conjunto de datos, a través de histogramas, diagramas de tallos y de cajas.

Una vez realizado el análisis univariado de los datos, se realizó el análisis bivariado, obteniendo diagramas de dispersión para cada par de variables, lo cual, por medio de parejas ordenadas ayudó a saber si existe una correlación entre las variables o la falta de relación entre las mismas. También se utilizó la función jointplot de la librería Seaborn para medir cada una de las variables vs el objetivo que es la resistencia a la compresión.

Por último, se realizó el análisis multivariado de los datos a través de los gráficos de correlación que se generó en el reporte de pandas profiling, en los cuales se puede visualizar correlaciones fuertes negativas con valores cercanos a -1 y correlaciones fuertes positivas con valores cercanos a +1.

2.2.4 Entrenamiento de modelos de aprendizaje de máquina

Teniendo la base, que es el análisis exploratorio de datos, se dio paso al entrenamiento de varios modelos de aprendizaje de máquina para la predicción de la resistencia a la compresión.

Las herramientas que se usaron para el entrenamiento de los modelos fueron las librerías pandas y pycaret, las cuales ayudaron a decidir cuáles son los mejores modelos para su posterior entrenamiento e implementación. Se utilizó la herramienta pycaret como base ya que sirvió como apoyo en todo el proceso proporcionando una forma fácil de entrenar modelos hasta levantarlos en una aplicación con la opción de acelerar el entrenamiento de los modelos en GPU ayudando con el entorno de entrenamiento, especificando los datos y la variable a predecir. Las herramientas fueron escogidas debido a su extensa documentación y facilidad de uso.

2.2.4.1 Pasos para obtener los mejores modelos para la predicción

Obtener cuáles son los modelos a entrenar para este estudio en específico es parte esencial para lograr resultados satisfactorios, por lo que, fue de suma importancia seguir los siguientes pasos:

1. Se eliminó las variables que no deseamos, las cuales se obtuvieron al momento de utilizar la herramienta pandas profiling enlistando las constantes, para su limpieza en este punto utilizando la función dataframe.drop(constants) de la librería pandas.
2. Se estableció en el conjunto de datos la variable a predecir que es resistencia a la compresión en MPa y las demás variables a través de la librería pycaret y su función setup(df, target, session_id, normalize, normalize_method, numeric_features, use_gpu) que establece el ambiente con los datos a trabajar preparando tareas como la normalización de datos, la división de los datos en grupos de entrenamiento y pruebas (por defecto 70% y 30% respectivamente), además de proveer la opción de acelerar el procesamiento usando la GPU.
3. Por último, se aplicó la función compare_models(fold) de la librería pycaret, la cual comparó entre varios modelos de aprendizaje de máquina usando cross-validation que entrena varios modelos, con un fold de 10 subconjuntos de datos del grupo de entrenamiento y arrojó una tabla con los mejores modelos tomando en cuenta varias métricas de evaluación las cuales son:
    i. MAE(Error absoluto medio): Esta métrica calcula la diferencia absoluta entre los valores reales y los que arroja la predicción.
    ii. MSE(Error cuadrático medio): Es una de las métricas más comunes, la diferencia de MAE debido a que calcula la diferencia al cuadrado entre los valores reales y los predichos.
    iii. RMSE(Raíz del error cuadrático medio): Esta métrica no es más que obtener la raíz cuadrada del MSE.
    iv. RMSLE(Error de registro de la raíz cuadrática media): Es muy usada cuando se desarrolla un modelo al cual no se le indican valores de entrada, por lo que la salida varía considerablemente.
    v. R2(R al cuadrado): Esta métrica indica el rendimiento del modelo, independientemente del contexto.

2.2.4.2 Creación de los modelos

Con los modelos elegidos, el siguiente paso fue crear los modelos para su posterior entrenamiento y evaluación completa. La forma de crearlo resultó sencilla debido a que pycaret posee una función para crear modelos la cual es create_model(model,fold,round). De aquí en adelante, estos pasos se los realizó para todos los modelos para, posteriormente, iniciar con la búsqueda del mejor o la fusión de los mejores modelos con el fin de obtener la predicción de la resistencia a la compresión con menor error.

2.2.4.3 Ajuste de hiperparámetros y evaluación de modelos

Se realizó un ajuste de hiperparámetros con la función tune_model(model) la cual ayuda a obtener valores óptimos para que el rendimiento del modelo se maximice. Una vez ajustado el modelo, se lo evaluó a través de la función evaluate_model(model), la cual permitió observar varios tipos de gráficos para analizar, por ejemplo, la importancia de las variables del conjunto de datos, por medio del gráfico de error de predicción, curva de validación, árbol de decisión, curva de aprendizaje, residuos interactivos, etc.

2.2.4.4 Predicción de modelos con datos de prueba

Para realizar una predicción, pycaret posee la función predict_model(model) la cual arroja una etiqueta y puntuación usando el modelo entrenado, así mismo, se visualizó el error a través de distintas métricas como el error promedio, el error cuadrático medio, etc., con los datos de prueba que se determinaron.

2.2.5 Implementación de API

Con el fin de realizar la implementación del API, se usó la función create_app(model) que provee pycaret. Esta función crea una aplicación básica para la predicción, con cajas de texto para llenar según el parámetro que corresponda y con ello, obtener la predicción de los valores de entrada enviados, proporcionando una interfaz amigable, sencilla y fácil de usar. Una vez creada el api, esta se carga en la nube para su posterior visualización de los interesados.

CAPÍTULO 3

3. RESULTADOS Y ANÁLISIS

Conocer la importancia de cada una de las variables que se toman en consideración para la producción de geopolímeros, así como la influencia de cada una de ellas en la resistencia a la compresión es fundamental en el estudio para, en base a aquello, determinar cuál es el modelo óptimo y con menor error a usar para las predicciones.
En este capítulo se describen los resultados obtenidos a través del análisis exploratorio de datos univariado, bivariado y multivariado donde se conocen las relaciones entre variables y su relevancia. Además, se identifican cuáles son los modelos por entrenar y su proceso, para, posteriormente seleccionar el modelo o la fusión de modelos que permitan obtener la predicción con menor error. Por último, se muestra la implementación del API generado para la interacción con el usuario.

3.1 Análisis exploratorio de datos

En base al análisis exploratorio de datos realizado, se obtuvieron las relaciones y la importancia de las variables tales como el silicato de sodio, hidróxido de sodio, el tiempo de curado que, dependiendo de su cantidad, ocasionan un incremento o decremento de la resistencia a la compresión. Además de otras variables que no influyen en mayor medida para lograr un fortalecimiento de las propiedades del geopolímero.

3.1.1 Análisis univariado

Para el análisis univariado se realizaron gráficos de distribución para cada una de las variables que se presentan en el conjunto de datos con sus distintos sesgos y 1 o más picos tal como se observan en la Figura 3.1.

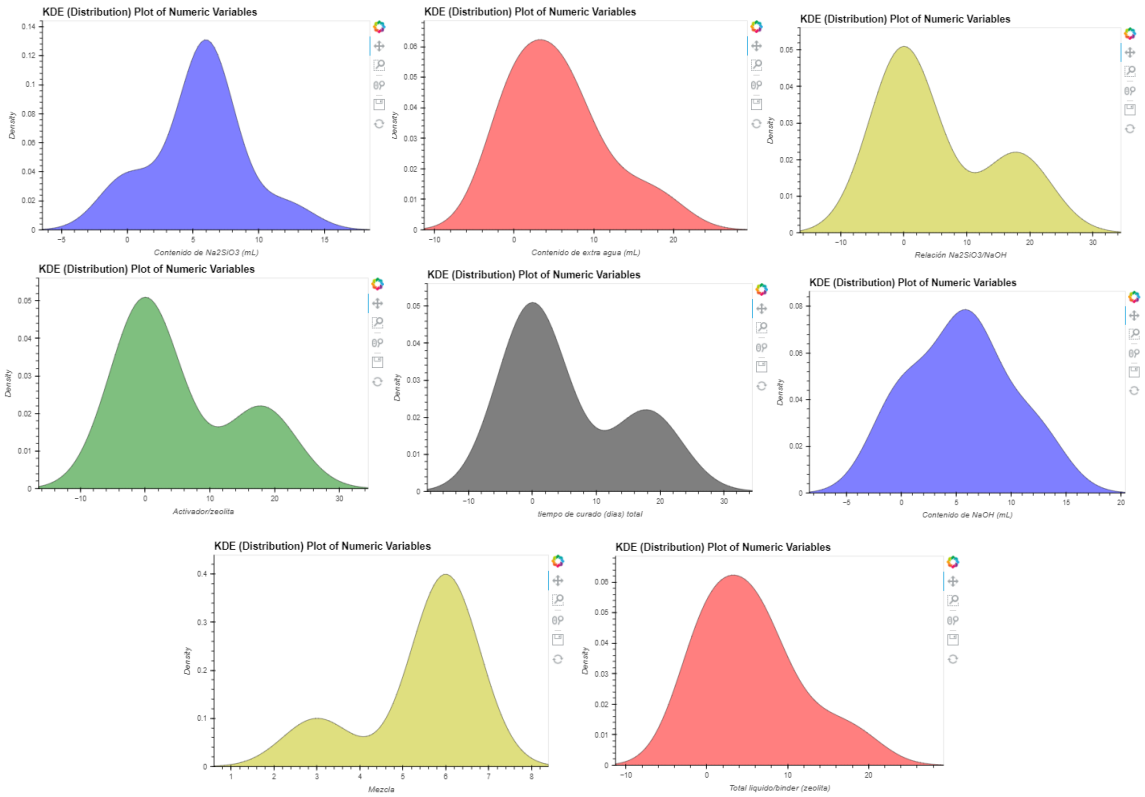

Figura 3.1 Distribución de cada una de las variables.

3.1.2 Análisis bivariado

El análisis bivariado se realizó en base a la obtención de gráficos de dispersión e histogramas para cada una de las variables considerando como variable independiente la resistencia a la compresión en cada uno.

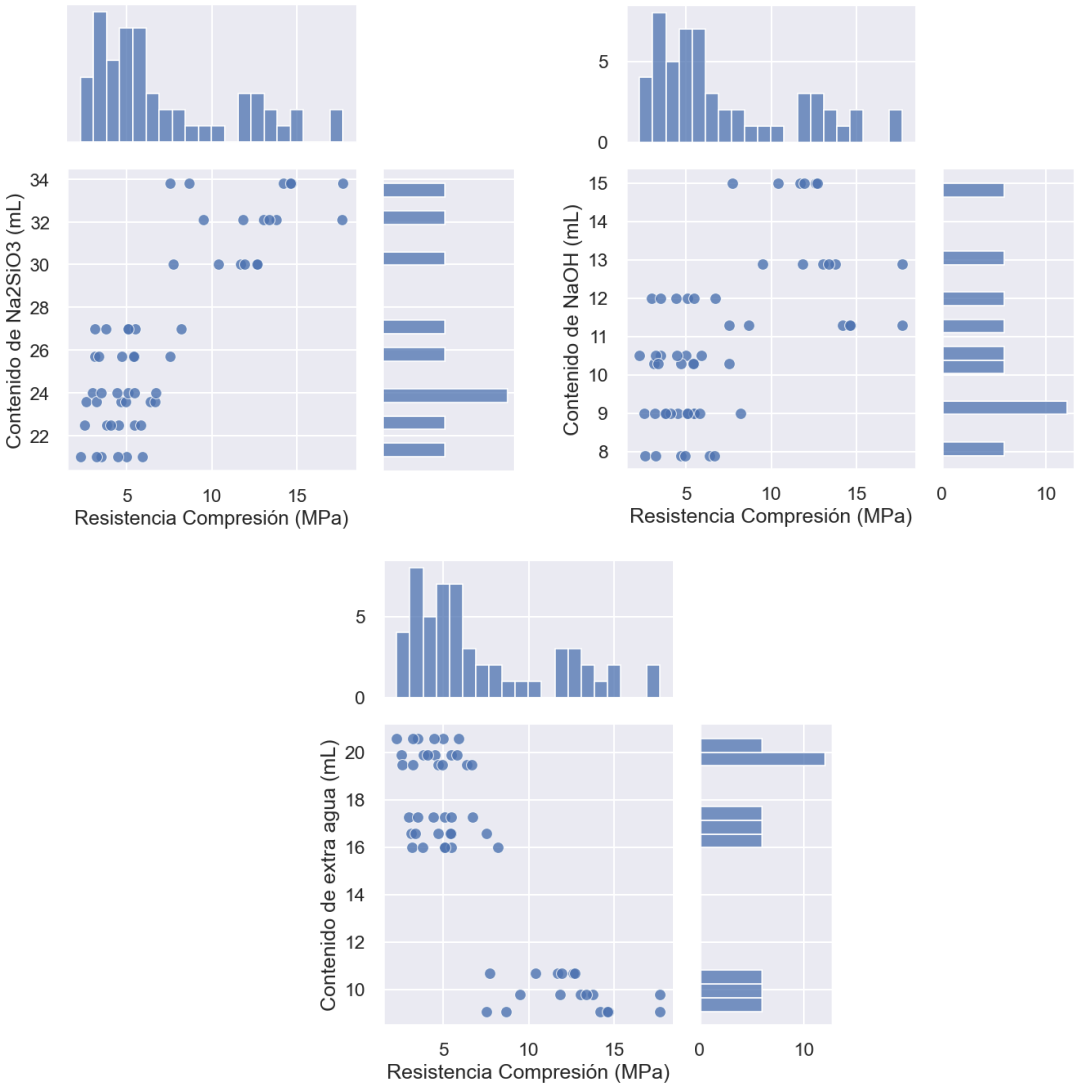

Figura 3.2 Correlación entre contenido de Na2SiO3, contenido de NaOH, contenido de extra agua, temperatura de curado (ºC) contra Resistencia a la compresión (MPa)

Con respecto a la Figura 3.2, se puede observar la relación que existe entre el contenido de NA2SiO3 y la resistencia a la compresión, dando como resultado que mientras mayor es la cantidad de NA2SiO3 agregada a la mezcla, la resistencia a la compresión será mayor debido al efecto como agente alcalinizante que, una vez disuelto en agua, da como resultado una solución alcalina activando la reacción química entre los componentes del geopolímero y acelerando el proceso de curado.

Por otra parte, si se considera el contenido de NAOH, que sirve como activador o gelificante, y que ayuda a que se pueda mezclar el agua con el aluminosilicato para formar el geopolímero, se observa que la resistencia a la compresión deja de crecer una vez superados los 13 mL, por tal motivo, es necesaria una cantidad que no exceda de los valores óptimos para que la resistencia no disminuya obteniendo una estructura microcristalina densa y ordenada.

En el caso del agua, su principal función es la trabajabilidad con la mezcla debido a su espesor, sin embargo, de acuerdo con el gráfico, se observa que mientras más extra agua se añade, menor resistencia a la compresión se obtiene.

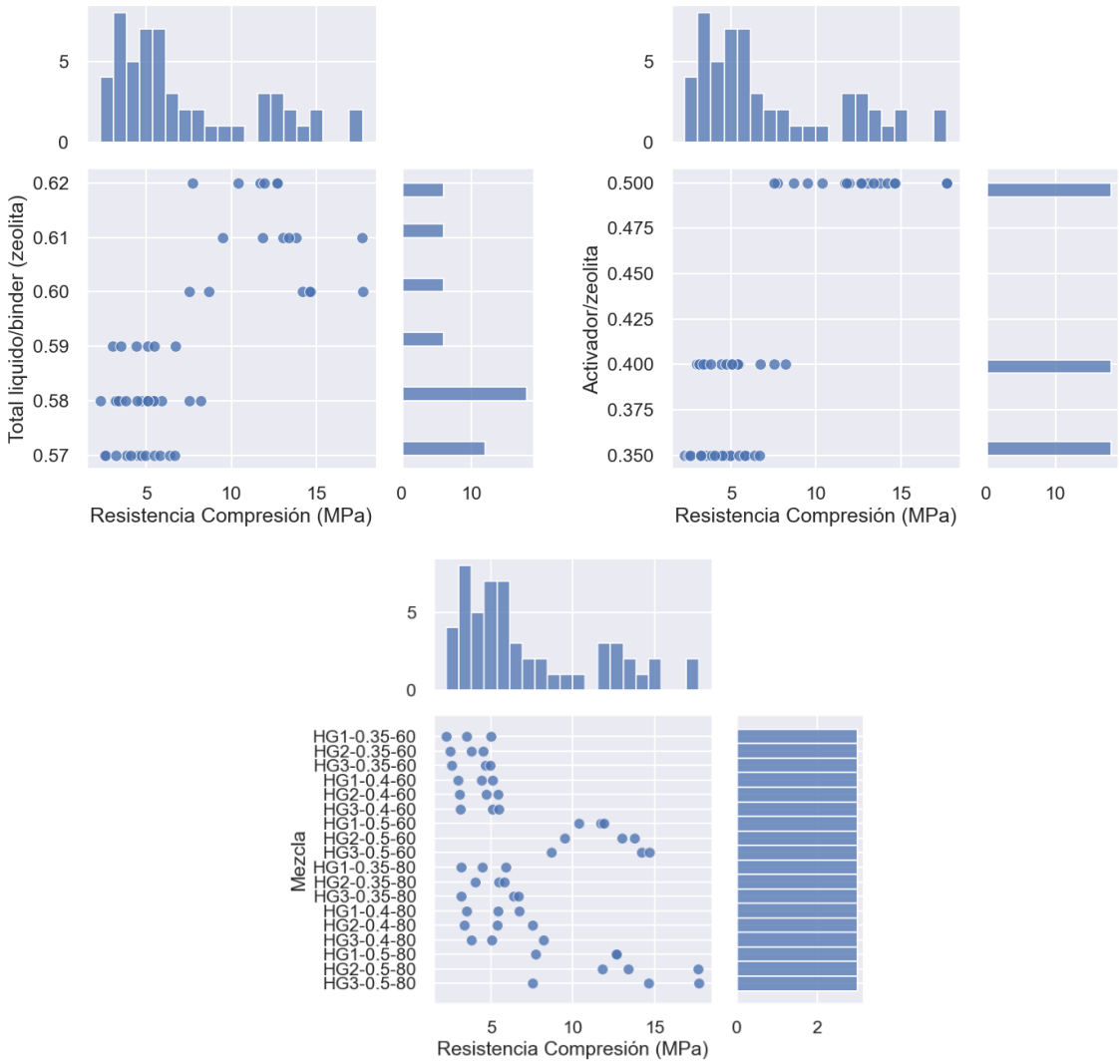

Figura 3.3 Correlación entre Contenido de zeolita, activador/zeolita, Mezcla y Resistencia a la compresión (MPa)

La relación activador/zeolita de 0,5 demuestra considerablemente una mayor resistencia a la compresión. Por lo tanto, mientras mayor zeolita se incluya, la mezcla tendrá una mejor propiedad de resistencia.

Por otro lado, en el caso de total líquido/binder, los mejores resultados se obtienen con una relación de 0.60 a 0.61, disminuyendo al aumentar o menorar las proporciones

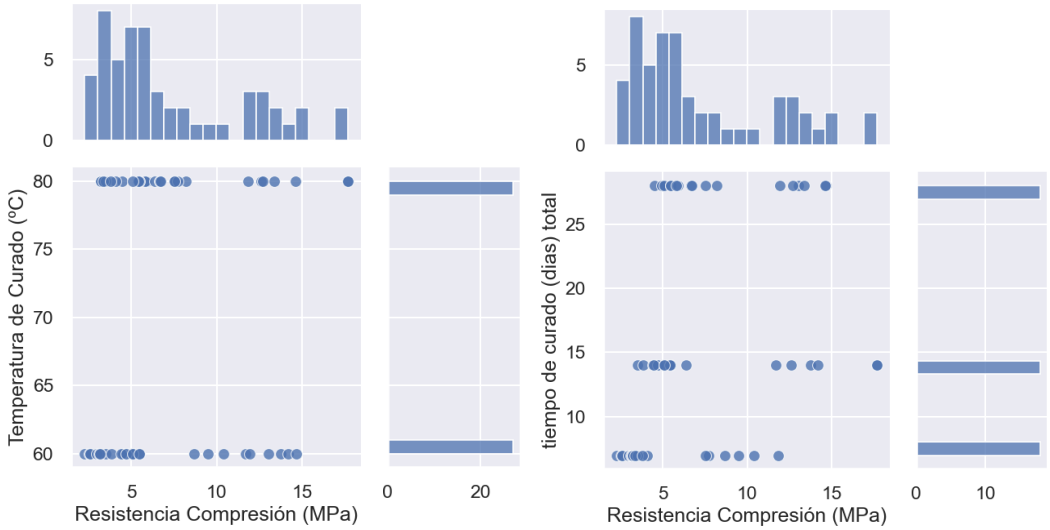

Figura 3.4 Correlación entre temperatura de curado (°C), tiempo de curado (días), y Resistencia a la compresión (MPa)

La temperatura de curado, que influye afectando la velocidad de la reacción química de los componentes del geopolímero, tiene un máximo en resistencia a la compresión en 80º, sin embargo, no determina una relación considerable con la resistencia debido a que influyen en mayor medida otros factores que acompañan a la temperatura. Por otro lado, si se analizan los días de curado, existe un máximo en resistencia con día 14 días, sin embargo, en general se demuestra un incremento constante cuando se consideran 28 días de curado. Es necesario tener un equilibrio entre la temperatura de curado y los días para obtener resultados óptimos.

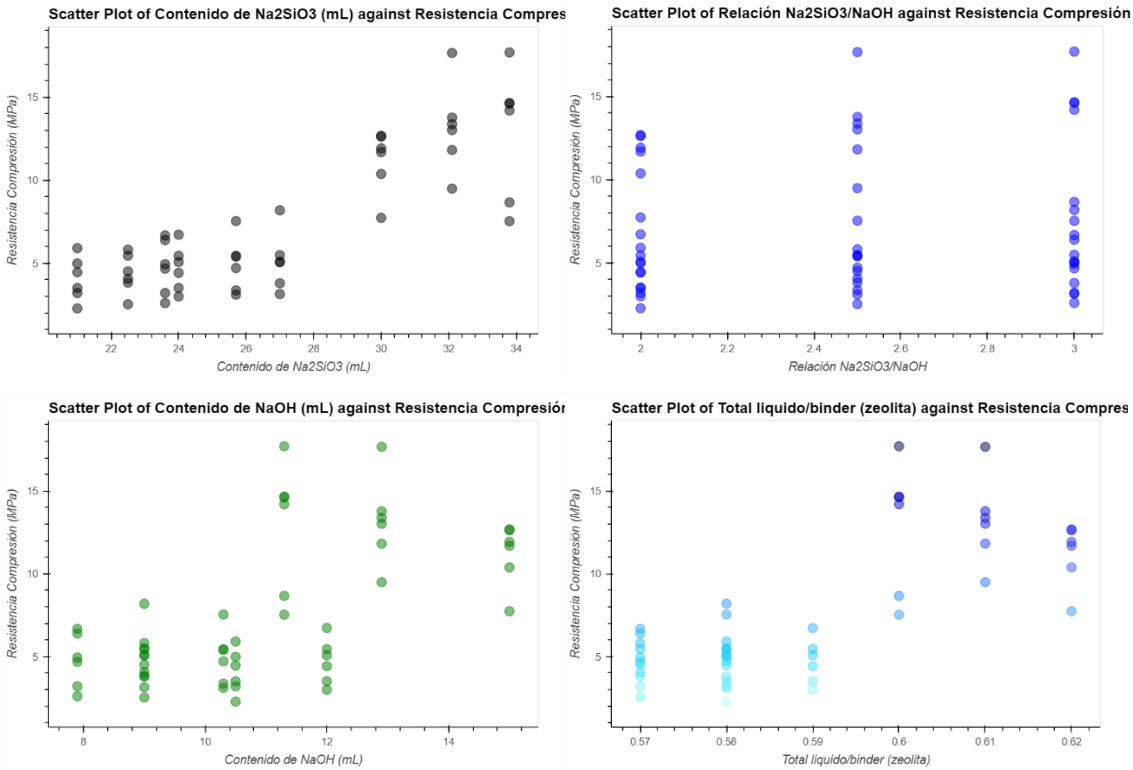

Figura 3.5 Dispersión entre parámetros de entrada y Resistencia a la compresión (MPa)

Los gráficos de dispersión de la Figura 3.5 demuestran que, si el contenido de Na2SiO3 aumenta, se incrementa la resistencia a la compresión, caso contrario al contenido de NaOH y el total líquido/binder que, una vez llegado a un máximo de 13 mL y de 0.6-0.61 respectivamente, a medida que se incrementa la proporción, la resistencia a la compresión del geopolímero disminuye.

Finalmente, la relación Na2SiO3/NaOH, corresponden a un mejor desempeño si se proporciona de 2.5 a 3, disminuyéndose con valores inferiores a 2.5.

3.1.3 Análisis multivariado

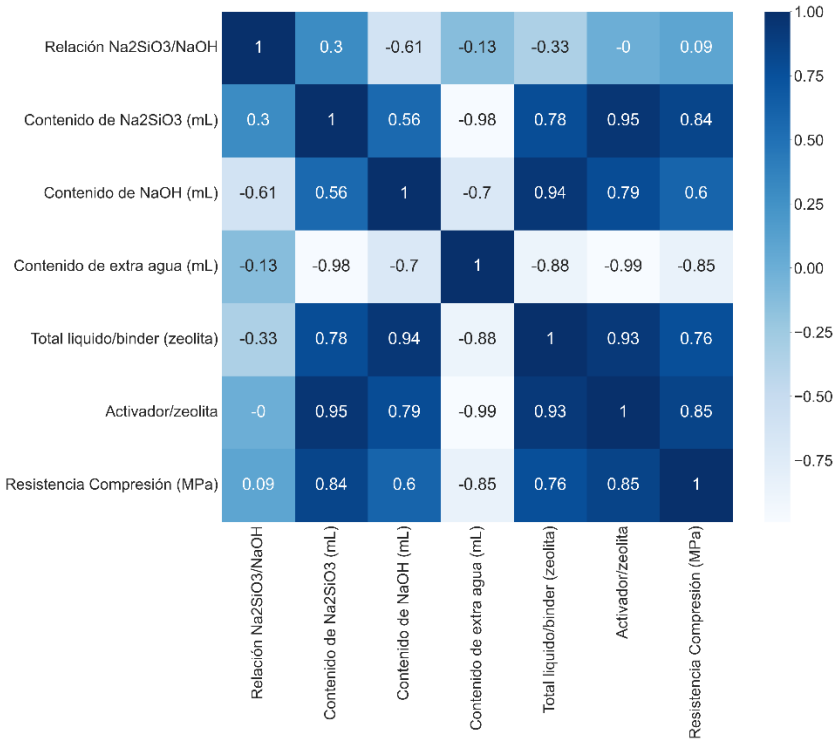

Figura 3.6 Correlación entre variables

Analizando la correlación y la influencia de las variables con respecto a la resistencia a la compresión, se determina que existen unas que influyen más que otras, por ejemplo, la relación activador/zeolita la cual está estrechamente relacionada a la resistencia a la compresión con un valor de 0.85. Así mismo, el contenido de Na2SiO3 y el total líquido/binder los cuales siguen con un valor de 0.84 y 0.76 debido a las propiedades que contienen para aumentar la resistencia y, en el caso del total líquido, se considera también el NaOH.

Caso contrario, con el contenido de extra agua y la relación Na2SiO3/NAOH las cuales arrojan valores menores a 0.1 expresando la poca relación o correlación negativa.

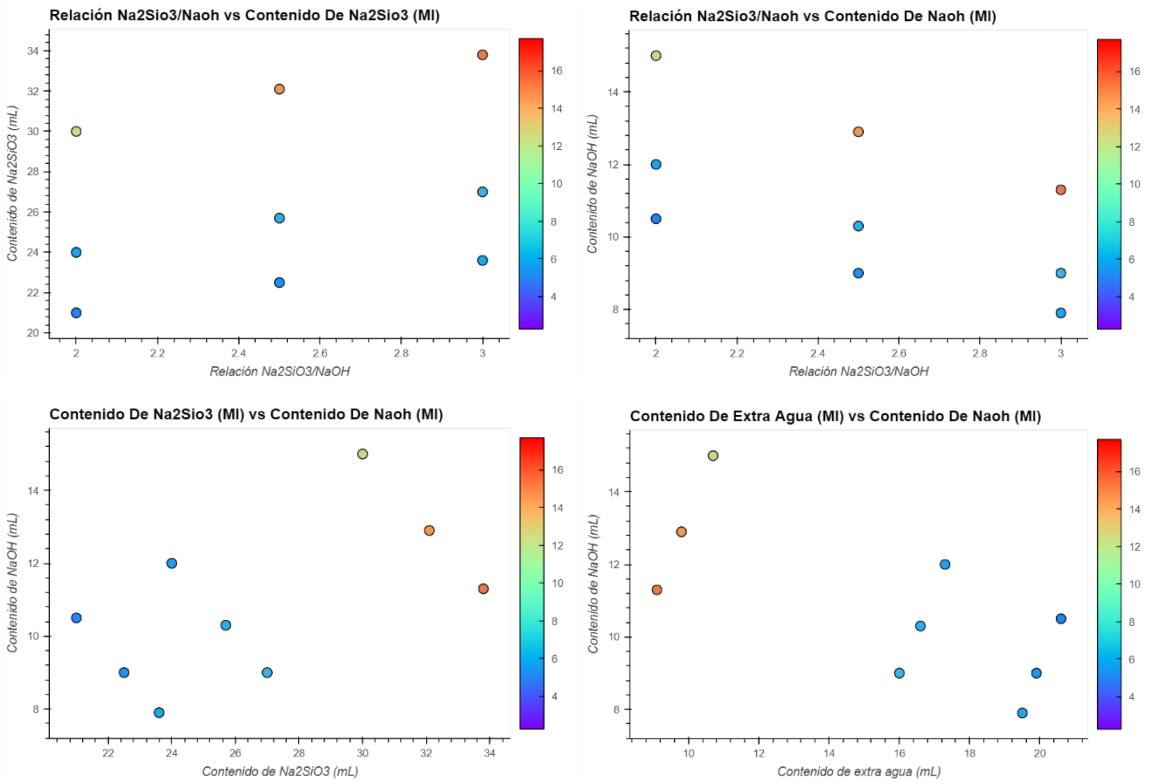

Figura 3.7 Dispersión entre parámetros de entrada y Resistencia a la compresión (MPa)

Una vez realizado el análisis entre pares, y determinando cuánto influyen las relaciones en la resistencia a la compresión, se obtiene, de acuerdo con la Figura 3.7, mientras la proporción de la relación Na2SiO3/NaOH aumenta, el geopolímero mantiene o crece en pequeñas cantidades su resistencia, sin embargo, si el contenido en mL de Na2SiO3 aumenta, su resistencia se puede disparar a valores máximos dependiendo de la cantidad agregada debido a su influencia en el proceso de curado. Por otro lado, a mayor proporción de la relación Na2SiO3/NaOH disminuyendo el NaOH, las propiedades mejoran debido a la relación que se debe obtener. Una relación inadecuada produce una reacción excesiva o incompleta, afectando negativamente a la resistencia.

Además, la Figura 3.7 demuestra que mientras mayor cantidad de Na2SiO3, y una cantidad no exagerada de NaOH, que no sobrepase los 13 mL, se tiene una relación lineal positiva, cuyo comportamiento inverso se observa en la relación

del contenido de extra agua y NaOH, que, mientras menos agua se agregue y exista una cantidad moderada de NaOH, se consiguen los valores óptimos.

3.2 Modelos de Aprendizaje de Máquinas

Dado que a través de librerías de pycaret se obtuvieron los mejores modelos para este modelo de predicción en base a varias métricas de evaluación, se realizaron los entrenamientos de los 5 mejores modelos, obteniendo, además, gráficos de error de predicción y grado de importancia de los parámetros de entrada para cada uno de los modelos, lo cual ayudó a decidir cuál es el mejor como se muestra a continuación.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
dt	Decision Tree Regressor	0.7533	1.1996	1.0039	0.6653	0.1206	0.1152	0.0700
rf	Random Forest Regressor	0.6749	1.2949	0.9188	0.6077	0.1093	0.1064	0.3760
catboost	CatBoost Regressor	0.9166	1.8888	1.1354	0.6028	0.1377	0.1532	1.1990
gbr	Gradient Boosting Regressor	0.8485	1.7004	1.1986	0.5529	0.1260	0.1230	0.1290
ada	AdaBoost Regressor	0.7433	1.2376	0.9889	0.4463	0.1180	0.1186	0.1480
et	Extra Trees Regressor	0.8310	1.9023	1.1910	0.4299	0.1208	0.1162	0.3380
knn	K Neighbors Regressor	1.1787	3.8131	1.6478	-0.3189	0.1739	0.1789	0.1150
en	Elastic Net	1.8376	5.5615	2.2177	-1.0258	0.2682	0.3032	0.0690
lasso	Lasso Regression	2.0662	6.6537	2.4288	-1.1409	0.2976	0.3494	0.0690
huber	Huber Regressor	1.6660	5.4597	2.1716	-1.5098	0.2401	0.2465	0.1220
br	Bayesian Ridge	1.6712	4.6304	2.0318	-1.6146	0.2427	0.2659	0.0690
omp	Orthogonal Matching Pursuit	1.9103	5.8072	2.3464	-1.6366	0.2999	0.3274	0.0690
ridge	Ridge Regression	1.6955	4.6869	2.0372	-1.7593	0.2453	0.2719	0.0700
lr	Linear Regression	1.6755	5.4003	2.1111	-2.0124	0.2380	0.2462	0.1200
par	Passive Aggressive Regressor	2.3899	8.9499	2.8290	-2.3456	0.4035	0.3768	0.0690
llar	Lasso Least Angle Regression	3.1760	15.9814	3.6463	-2.5022	0.4525	0.5537	0.0680
lightgbm	Light Gradient Boosting Machine	3.1760	15.9814	3.6463	-2.5022	0.4525	0.5537	0.1690
dummy	Dummy Regressor	3.1760	15.9814	3.6463	-2.5022	0.4525	0.5537	0.0630
lar	Least Angle Regression	207.0004	89876.0756	258.5859	-45436.0479	3.0003	33.4082	0.0740

Figura 3.8 Tabla comparativa de modelos entrenados y evaluados con validación cruzada

En una comparativa de los 3 mejores modelos, estos coinciden en que son basados en árboles de decisión, y los siguientes que siguen en la lista son modelos boosting que crean varios modelos más pequeños de manera secuencial para formar un modelo robusto con el objetivo de minimizar la función de pérdida.

En investigaciones de Kaloop et al. demuestra un mejor rendimiento prediciendo la compresión utilizando modelos gradient tree boosting machine. (2020)
Según una nueva investigación, los algoritmos boosting presentan el mejor rendimiento para predecir la resistencia a la compresión. (Rathakrishnan et al., 2022)

Por lo tanto, según la literatura los mejores modelos coinciden con trabajos anteriores de otros investigadores mostrando un mejor rendimiento y adaptabilidad para predecir la resistencia a la compresión.

3.2.1 Entrenamiento de modelos

A continuación, se describen los modelos con mayor eficiencia para predecir la resistencia a la compresión. Cada uno de ellos con su respectivo gráfico de importancia de características y error de predicción el cual determina si es el mejor modelo para el ejercicio. Basándose en R^2 o coeficiente de determinación el cual demuestra que, si es cercano a 1, el modelo tiene una mayor precisión en la predicción.

3.2.1.1 CatBoost Regressor

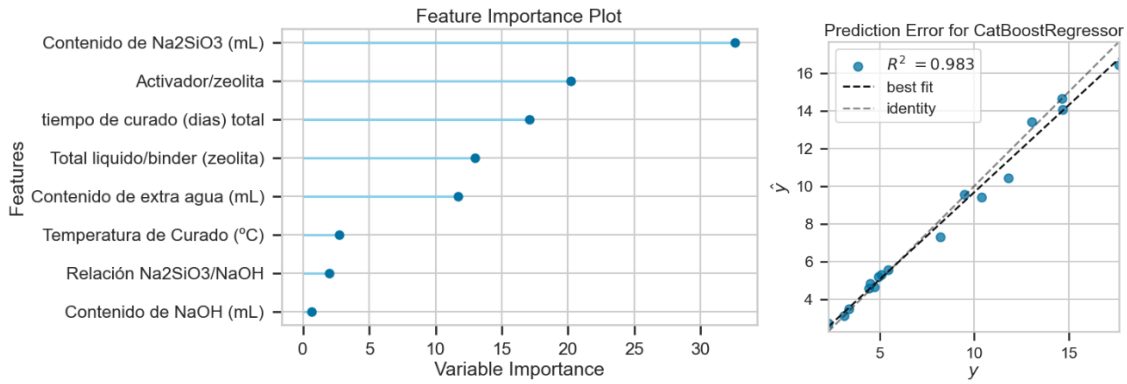

Figura 3.9 Importancia de variables y predicción de Error para Catboost Regressor

Catboost, un algoritmo basado en árboles de decisión es el mejor modelo para la predicción de la resistencia a la compresión en geopolímeros. En la Figura

3.9 se muestra el gráfico de importancia de variables. Donde la mayor importancia le da al contenido del agente alcalinizante Na2SiO3, seguido por la relación activador/zeolita. Por otra parte, el contenido de NaOH se expresa como la variable de menor importancia.

Además, se muestra la gráfica del error de predicción para este modelo con un coeficiente de determinación de 0.983, el cual, comparando con los demás modelos, se aproxima en mayor medida a 1. Por lo que se concluye que,
Catboost es el algoritmo cuyas características se adaptan a lo que se requiere para lograr una predicción con el menor error. Dentro de las características se incluyen su uso para problemas de clasificación y regresión, así como la flexibilidad en el manejo de variables categóricas. Además, se logran buenos resultados cuando el conjunto de datos es pequeño, como el que se utilizó para el entrenamiento y realiza un ajuste de hiperparámetros mínimo para que no ocurra un sobreajuste.

3.2.1.2 Random Forest

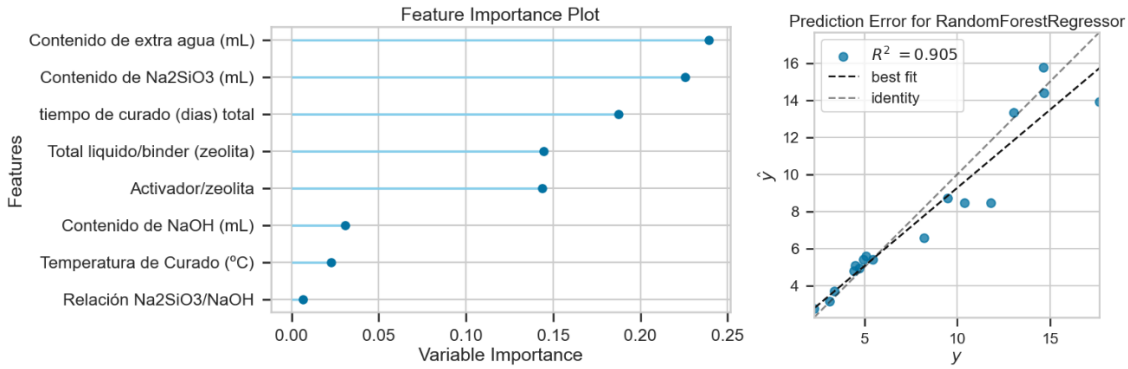

Figura 3.10 Importancia de variables y predicción de Error para Random Forest

Random Forest, un algoritmo que ha resultado ser el mejor en muchos problemas de regresión, arroja un orden de importancia de variables tomando al contenido de extra agua como la variable de mayor importancia, seguida del contenido de Na2SiO3. Por otra parte, coloca a la relación de Na2SiO3/NaOH

como variable de menor importancia, dando un error de predicción con un coeficiente de determinación de 0.905 como se visualiza en la Figura 3.10, colocándolo como un modelo sugerido, pero no mejor que Catboost para este problema.

3.2.1.3 Decision Tree

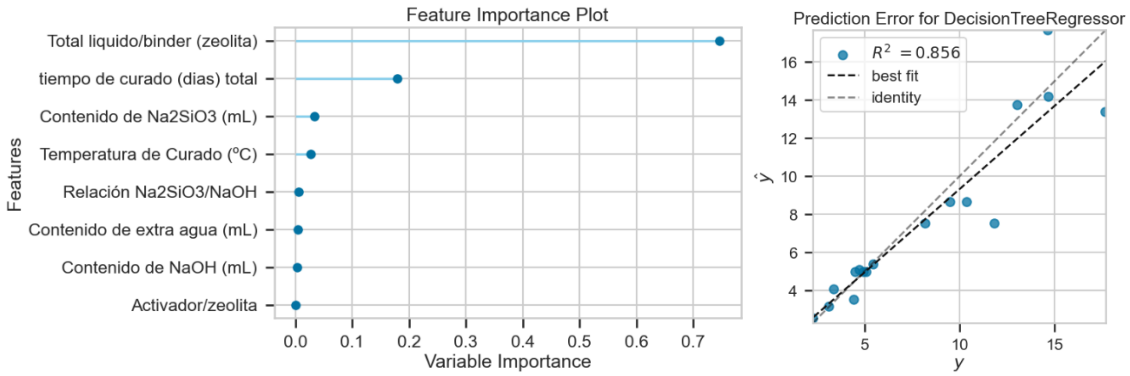

Figura 3.11 Importancia de variables y predicción de Error para Decision Tree

En la Figura 3.11 se observa la importancia de variables para el algoritmo Decision Tree, el cual toma como variable de mayor importancia al Total líquido/binder (zeolita), dejando muy por debajo al resto de variables que, considerando el contexto del problema, deberían tener un mayor grado de importancia. Se observa también un coeficiente de determinación de 0.856, inferior a los algoritmos de Catboost y Random Forest.

3.2.1.4 Gradient Boost

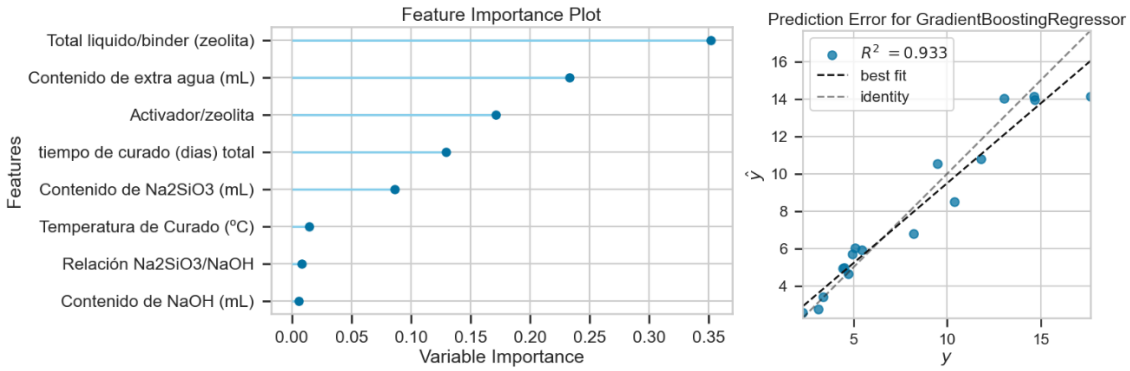

Figura 3.12 Importancia de variables y predicción de Error para Gradient Boost

Observando la Figura 3.12, de acuerdo con el gráfico de importancia de variables, Gradient Boost toma como variable de mayor importancia al total líquido/binder (zeolita), seguido del contenido de extra agua y teniendo como variable de menor importancia al contenido de NaOH. Esta priorización de variables da como resultado un coeficiente de determinación de 0.933, siendo uno de los mejores modelos, sin embargo, aún por debajo de Catboost.

3.2.1.5 AdaBoost

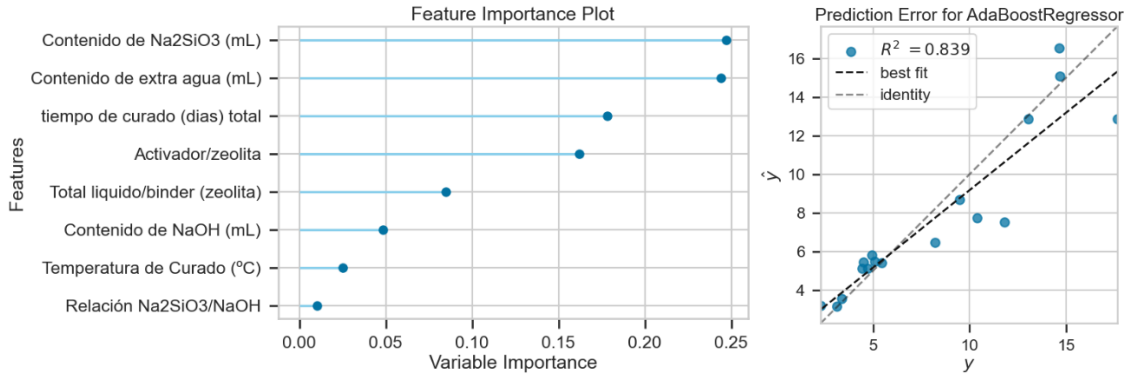

Figura 3.13 Importancia de variables y predicción de Error para AdaBoost

AdaBoost, que es un algoritmo de aprendizaje en serie, considera en el gráfico de importancia de variables al contenido de Na2SiO3 y contenido de extra agua como variables de mayor importancia, y a la relación de Na2SiO3/NaOH como

de menor importancia, obteniendo un coeficiente de determinación en el gráfico del error de predicción de 0.839.

3.2.2 Métricas de error de algoritmos individuales

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	CatBoost Regressor	0.4225	0.3634	0.6028	0.9832	0.0603	0.0542

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Random Forest Regressor	0.9515	2.0584	1.4347	0.9048	0.1277	0.1094

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Decision Tree Regressor	1.1176	3.1110	1.7638	0.8562	0.1499	0.1174

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Gradient Boosting Regressor	0.8854	1.4401	1.2000	0.9334	0.1088	0.1078

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	AdaBoost Regressor	1.2551	3.4933	1.8690	0.8385	0.1750	0.1508

Figura 3.14 Métricas de Error de los diferentes modelos al predecirlos con datos de prueba

En la Figura 3.14, se muestran los distintos modelos junto a sus métricas de error luego de realizar la predicción con datos de prueba. Dando a CatBoost como el modelo con el menor error de todos, resaltando un coeficiente de determinación o R2 con el valor más cercano a 1, específicamente de 0.98 significando un ajuste lineal casi perfecto.

3.2.3 Combinación de modelos de aprendizaje de máquinas

Una vez revisados los modelos individualmente para conocer cuál es el mejor de ellos, se realizaron 3 combinaciones de modelos para obtener, de ser posible, una combinación con métricas de error que reflejen una mayor efectividad contra los algoritmos individuales.

Las combinaciones fueron las siguientes:

- Blender 1: Decision tree, Random Forest, CatBoost.

- Blender 2: Gradient boost, Adaboost, CatBoost.
- Blender 3: Gradient boost, Adaboost, Decision tree.

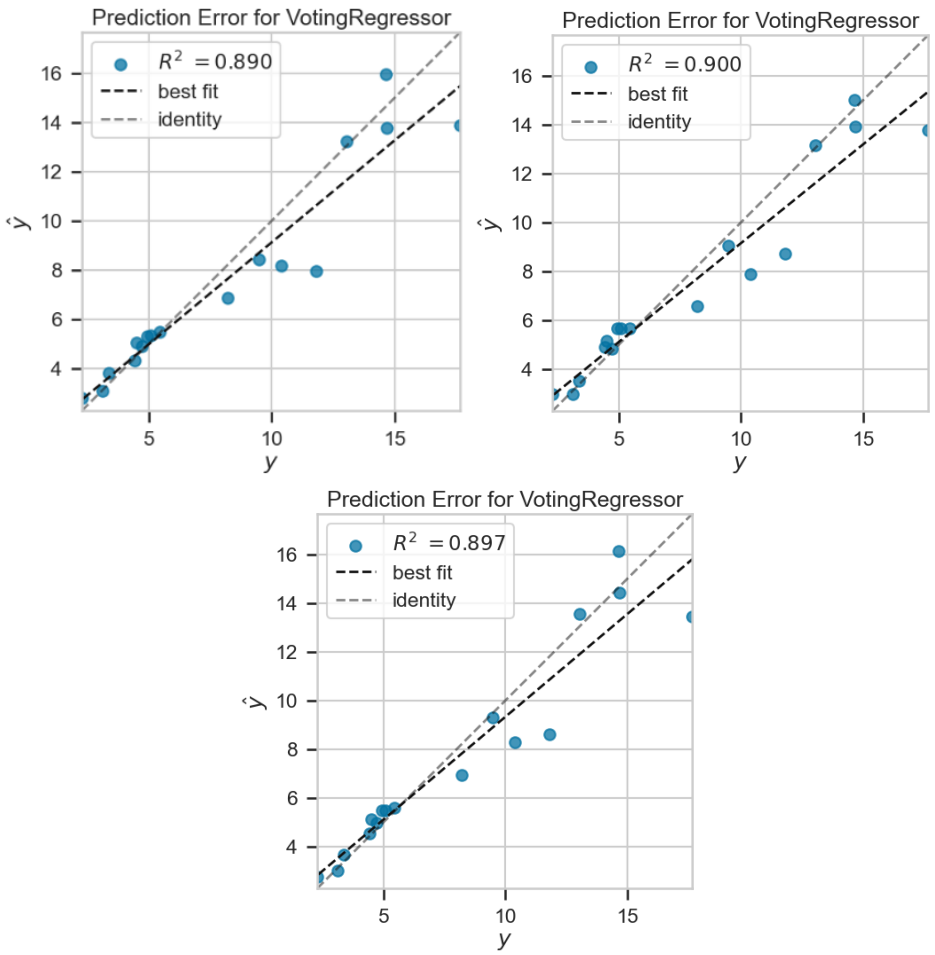

Figura 3.15 Predicción del error de las 3 combinaciones

Las 3 combinaciones de modelos que se realizaron arrojaron coeficientes de determinación por debajo de los algoritmos vistos individualmente, por tal motivo, la combinación no es considerada como una mejora para lograr resultados con mayor efectividad y precisión.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Voting Regressor	1.0062	2.3898	1.5459	0.8895	0.1370	0.1102

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Voting Regressor	0.9754	2.1555	1.4682	0.9004	0.1368	0.1198

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Voting Regressor	0.9541	2.2304	1.4935	0.8969	0.1267	0.1068

Figura 3.16 Métricas de error de las 3 combinaciones

Las métricas de error tomando en consideración en este caso las tres combinaciones de algoritmos no son mejores que los algoritmos vistos individualmente, por lo que la conclusión de que Catboost es el mejor algoritmo para este ejercicio, se mantiene.

3.3 API

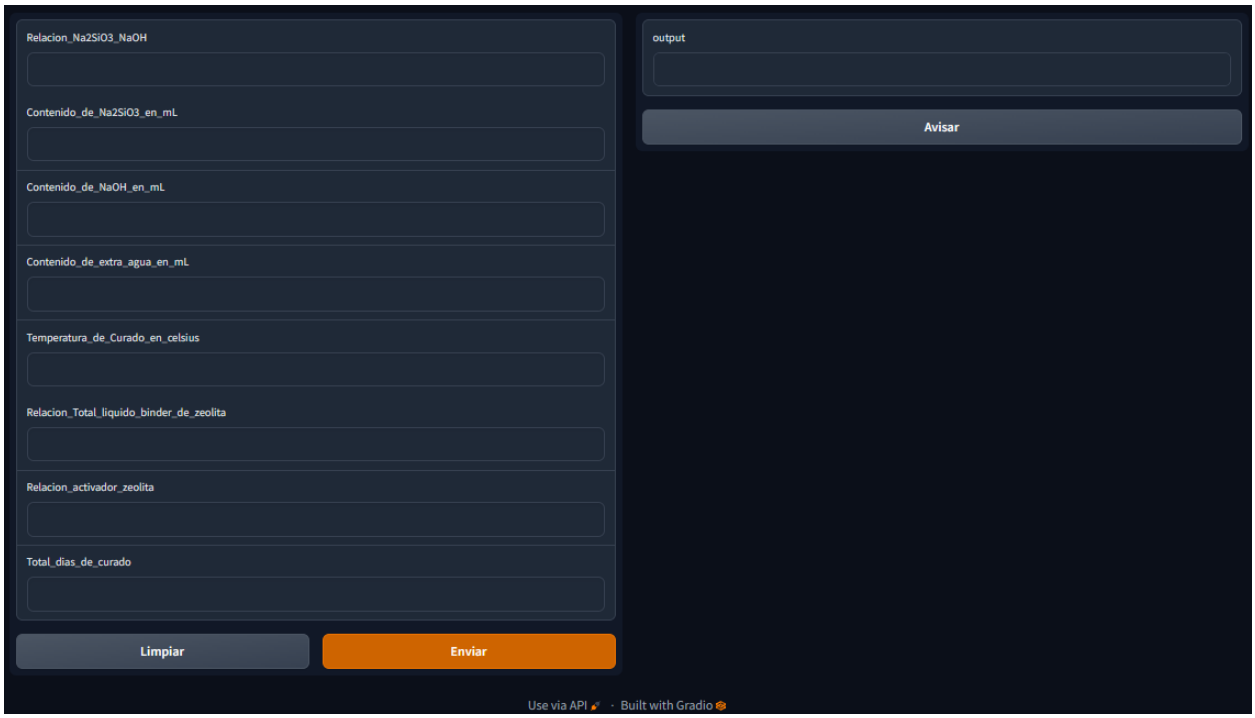

Figura 3.17 Vista de API

Usando la librería Gradio, la cual ayuda a construir una interfaz web para modelos de aprendizaje de máquinas, se crea una API que corre localmente la cual

permite compartirse con otros usuarios funcionando hasta 72 horas para su uso y posterior retroalimentación. En el apartado izquierdo, se muestran los campos donde se deben colocar los valores de entrada del modelo para la predicción que se desea realizar.

Una vez insertado los valores de entrada, en la parte inferior, se muestran dos botones: Uno para limpiar la caja de texto y otro para enviar los datos a procesar. A partir del envío de los datos, en el apartado derecho se muestran los resultados de la predicción en base a los valores de entrada ingresados.

Link Google Colab : https://drive.google.com/file/d/1foGbmKdbWWuL-BuG4mXcu1aZEJTZU-nT/view?usp=sharing

CAPÍTULO 4

4. CONCLUSIONES Y RECOMENDACIONES

Implementar un modelo de aprendizaje de máquinas para la predicción de la resistencia a la compresión en geopolímeros que, si bien es cierto, aporta a la preservación del medio ambiente debido a la cantidad muy por debajo de CO2 que emite con respecto a la fabricación de materiales tradicionales como cemento portland, además de sus propiedades físico-químicas que reflejan, sin lugar a dudas, ventajas contra materiales de construcción tradicionales, como es el caso de su resistencia a la compresión, de sus propiedades anticorrosivas, resistencia a altas temperaturas, etc. Resultó en un trabajo de investigación minuciosa que, con la ayuda de un correcto análisis exploratorio de datos, y entrenamientos basados en el conjunto de datos que se otorgaron, con un ajuste de hiperparámetros adecuado, determinaron los puntos que se describen en este capítulo.

4.1 Conclusiones

1. Dentro del análisis exploratorio de datos, se obtuvieron las variables que, por sus propiedades físicas y químicas, influyen o aportan en mayor o menor medida a un incremento en la resistencia a la compresión en geopolímeros, como es el caso del contenido de Na2SiO3, como agente alcalinizante, la relación activador/zeolita, el contenido de NaOH, los cuales, en una medida equilibrada, arrojan máximos en la resistencia.
2. El balance que debe existir entre los elementos que conforman la mezcla para la fabricación del geopolímero es de suma importancia para una resistencia considerable. Las variables, en su mayoría, responden negativamente a una insuficiente o excesiva cantidad del compuesto, como es el caso del contenido de extra agua, que humedecería de forma negativa a la mezcla o la temperatura de curado, que, afecta a la velocidad de reacción de los componentes.

3. Los algoritmos basados en árboles de decisión responden de una mejor manera para este tipo de problemas, seguidos por los algoritmos boosting que también arrojan buenos resultados debido a sus características como la construcción de un modelo robusto en base a modelos secuenciales.
4. Entre los modelos entrenados, se obtuvo que el mejor modelo para la realización del modelo de predicción de resistencia a la compresión es CatBoost, debido a sus características como el fácil manejo de variables categóricas, la flexibilidad cuando existen conjuntos de datos pequeños, la importancia que le da a cada una de las variables y su coeficiente de determinación de 0.98, el cuál es el más alto con respecto a los demás como Random Forest, AdaBoost, Gradient boosting, Decision Tree, que, si bien es cierto, son considerados también como los más recomendados, no superan en las métricas de error a Catboost.
5. Una vez realizados los entrenamientos a los modelos de manera individual, se consideró realizar varias combinaciones de estos modelos para verificar si existe una mejora en la predicción, sin embargo, una vez obtenidos los coeficientes de determinación, las métricas de error y la importancia a las variables que le dan las 3 combinaciones posibles, no arrojaron valores que estén por encima de los modelos de manera individual, por tal motivo, se descartó la posibilidad de realizar la implementación con combinaciones de modelos.
6. Según la literatura los mejores modelos coinciden con trabajos anteriores de otros investigadores mostrando un mejor rendimiento y adaptabilidad para predecir la resistencia a la compresión.

4.2 Recomendaciones

1. Se recomienda la realización del análisis exploratorio de los datos, para problemas de este tipo siguiendo al pie de la letra los pasos tales como la limpieza de los datos, la eliminación de las constantes, etc. que darán como resultado la correcta correlación de las variables, y ayudarán a determinar el modelo, si se desea buscar otro, que considere como prioridad a esas variables con una mayor correlación.

2. Hay que considerar que la combinación de modelos que arrojen los mejores resultados no siempre va a superar a esos modelos vistos de manera individual, por tal motivo, es necesario revisar si existe una posibilidad de mejora, sin descartar los estudios previos en los modelos individuales.
3. Se recomienda que se investigue en mayor medida la posibilidad de implementar el modelo de aprendizaje de máquinas con librerías u otro tipo de herramientas que permitan una mayor flexibilidad en temas de interfaz y tiempo de acceso al API al compartir con otros usuarios.

BIBLIOGRAFÍA

Singh, N. & Middendorf, B. (2020). Geopolymers as an alternative to Portland cement: An overview. Construction and Building Materials, 237, 117455. https://doi.org/10.1016/j.conbuildmat.2019.117455

Shehata, N., Sayed, E. T. & Abdelkareem, M. A. (2021). Recent progress in environmentally friendly geopolymers: A review. Science of The Total Environment, 762, 143166. https://doi.org/10.1016/j.scitotenv.2020.143166

Ulloa, Néstor, Jiménez, Mirian, Serrano, Byron, & Serrano, Carlos. (2022). Geopolímeros basados en zeolitas naturales como una alternativa de materiales de construcción. Revista ingeniería de construcción, 37(1), 5-13. https://dx.doi.org/10.7764/ric.00011.21

Ganesh, C., & Muthukannan, M. (2018). A review of recent developments in geopolymer concrete. International Journal of Engineering and Technology(UAE), 7, 696–699. doi:10.14419/ijet.v7i4.5.25061

Milo, T., & Somech, A. (2020). Automating Exploratory Data Analysis via Machine Learning: An Overview. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. https://doi.org/10.1145/3318464.3383126

Nongthombam, K. & Sharma, D. (2021) IJERT-Data Analysis using Python. International Journal of Engineering Research and Technology (IJERT), vol. 10, pp. 1-5.

Waskom, M. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

Sahoo, K., Samal, A. K., Pramanik, J., & Pani, S. K. (2019). Exploratory Data Analysis using Python. International Journal of Innovative Technology and Exploring Engineering, 8(12), 4727–4735. https://doi.org/10.35940/ijitee.l3591.1081219

Rao, A. S., Vardhan, B. V. & Shaik, H. (2021). Role of Exploratory Data Analysis in Data Science. 2021 6th International Conference on Communication and Electronics Systems (ICCES), pp. 1457-1461. https://doi.org/10.1109/icces51350.2021.9488986

Rathakrishnan, V., Bt. Beddu, S. & Ahmed, A. N. (2022). Predicting compressive strength of high-performance concrete with high volume ground granulated blast-furnace slag replacement using boosting machine learning algorithms. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-12890-2

Kaloop, M. R., Kumar, D., Samui, P., Hu, J. W. & Kim, D. (2020). Compressive strength prediction of high-performance concrete using gradient tree boosting machine. Construction and Building Materials, 264, 120198. https://doi.org/10.1016/j.conbuildmat.2020.120198

Gong, M., Bai, Y., Qin, J., Wang, J., Yang, P. & Wang, S. (2020). Gradient boosting machine for predicting return temperature of district heating system: A case study for residential buildings in Tianjin. Journal of Building Engineering, 27, 100950. https://doi.org/10.1016/j.jobe.2019.100950

Farooq, F., Ahmed, W., Akbar, A., Aslam, F., & Alyousef, R. (2021). Predictive modeling for sustainable high-performance concrete from industrial wastes: A comparison and optimization of models using ensemble learners. Journal of Cleaner Production, 292, 126032. doi:10.1016/j.jclepro.2021.126032

Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. Journal of Big Data, 7(1). https://doi.org/10.1186/s40537-020-00369-8

Patil, S. & Kulkarni, U. (2019). Accuracy Prediction for Distributed Decision Tree using Machine Learning approach. 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). https://doi.org/10.1109/icoei.2019.8862580

Somvanshi, M., Chavan, P., Tambade, S. & Shinde, S. V. (2016). A review of machine learning techniques using decision tree and support vector machine. 2016 International Conference on Computing Communication Control and automation (ICCUBEA). https://doi.org/10.1109/iccubea.2016.7860040