Análisis de la influencia del suelo en la relación capacidad-torque (k_t) en la instalación de los pilotes helicoidales

Presentado por

Andrés Stiven Jaramillo Cataño

Informe de práctica presentado para optar al título de Ingeniero Civil

Asesores

Edwin Fabián García Aristizábal, Ph.D.

Derly Estefany Gómez García, MSc

Universidad de Antioquia

Facultad de Ingeniería

Ingeniería Civil

Medellín, Antioquia, Colombia

2024

Dedicatoria

Le dedico este logro a mis padres, abuelos y mi pareja Ledis Miranda por el apoyo incondicional, también a mis amigos por siempre darme ánimos en los momentos más difíciles.

Agradecimientos

Un agradecimiento a todos los profesores que hicieron parte de mi proceso y darme los conceptos y herramientas necesarias para el desarrollo de mi carrera profesional. Al profesor Edwin Fabian García y la profesora Derly Estefany Gómez por el asesoramiento en este proceso de práctica y también a la empresa Axiatech S.A.S por permitirme desarrollar mi práctica en sus instalaciones, principalmente al Ingeniero Sebastián Bedoya por el acompañamiento.

Tabla de contenido

Resumen 8

Abstract 9

Introducción 10

1 Objetivos 13

2 Marco teórico 14

3 Metodología 18

3.1 Creación de la base de datos 18

3.2 Limpieza de los datos 19

3.3 Desarrollo del Código 19

3.4 Análisis de los datos 19

4 Resultados y Análisis 21

4.1 Creación y limpieza de la base de datos 21

4.2 Creación del código 23

4.2.1 Análisis exploratorio 24

4.2.3 Preprocesamiento 29

4.2.3 Entrenamiento 31

4.2.4 Validación 32

4.2.5 Predicción 34

5 Conclusiones 35

6 Recomendaciones 36

Referencias 37

Lista de tablas

Tabla 1 Librerías utilizadas 18

Tabla 2 Matriz 22

Tabla 3 Tipo de variable de las columnas 24

Tabla 4 Partición de entrenamiento 29

Tabla 5 Partición de prueba 29

Tabla 6 Matriz transformada 30

Tabla 7 Media de las métricas de validación cruzada 32

Tabla 8 Predicciones del modelo 34

Lista de figuras

Figura 1 Partes de un pilote helicoidal 14

Figura 2 Distribución de la variable respuesta 25

Figura 3 Distribución de las variables numéricas 26

Figura 4 Correlaciones con Kt 26

Figura 5 Matriz de correlaciones 27

Figura 6 Distribución del error 32

Figura 7 Diagnóstico de los residuos 33

Lista de ecuaciones

Ecuación 1:Capacidad ultima de los pilotes helicoidales. 11

Siglas, acrónimos y abreviaturas

m	Metros
cm	Centímetros
ft	Pies
lb	Libras
kN	Kilonewtons
MSc	Magister Scientiae
PhD	Philosophiae Doctor
UdeA	Universidad de Antioquia

Resumen

En el presente trabajo se expone el desarrollo de un código en el lenguaje de programación Python para determinar la relación entre las propiedades índice del suelo y la relación capacidad - torque de los pilotes helicoidales. El programa se entrenó y se comprobó utilizando las bases de datos de la empresa Axiatech S.A.S. Estas bases de datos contienen información detallada, incluyendo registros de torque, tipo de suelo, propiedades índices y resultados de pruebas de carga realizadas en diversos proyectos desarrollados por la empresa; sin embargo, la matriz de datos utilizada en el proyecto se caracteriza por la presencia de numerosas variables sin valores registrados a la profundidad específica en la que se llevó a cabo la prueba de carga. Ante esta situación, se adoptó la estrategia de suponer el último valor registrado de cada variable hasta la profundidad donde se disponía de información de las pruebas de carga.

El código consiste principalmente de un análisis exploratorio de las variables, un entrenamiento del código, la predicción y el análisis de los residuos del análisis; el programa terminado permite aproximar el comportamiento del factor Kt en distintas tipologías de suelo y diferentes propiedades índice. El modelo predictivo utilizado fue una regresión lineal.

Palabras clave: factor Kt, capacidad última, torque de instalación, pruebas de carga, Python, machine learning, bases de datos.

Abstract

This work describes the development of a code in the Python programming language to determine the relationship between soil index properties and the torque-capacity ratio of helical piles. The program was trained and validated using the databases of the company Axiatech S.A.S. These databases contain detailed information, including torque records, soil type, index properties, and results of load tests conducted on various projects developed by the company. However, the data matrix used in the project is characterized by the presence of numerous variables without recorded values at the specific depth where the load test was performed. In response to this situation, the strategy was adopted to assume the last recorded value of each variable up to the depth where information on load tests was available.

The code primarily consists of an exploratory analysis of the variables, code training, prediction, and residual analysis. The completed program allows to approximate the behavior of the Kt factor in different soil typologies and various index properties. The predictive model used was a linear regression.

*Keywords***:** Kt factor, ultimate capacity, installation torque, load testing, python, machine learning, databases.

Introducción

Los pilotes helicoidales representan una alternativa destacada en el ámbito de las cimentaciones. Estos están conformados por un eje principal de acero constituidos de hélices o platos helicoidales, estos pilotes se extienden mediante fustes adicionales según el diseño establecido. Su aplicación abarca diversas industrias, incluyendo petróleo, gas, energía, ferrocarriles, inmobiliarias, comercio e infraestructuras eléctricas, donde se desempeñan como fundaciones para torres, torrecillas y otras estructuras, como faros y edificaciones variadas. En la industria eléctrica, los pilotes helicoidales se han convertido en una opción fundamental para el soporte de torres y torrecillas, siendo su uso una práctica común en el diseño y construcción. Axiatech S.A.S cuenta con amplia experiencia en el diseño, suministro e instalación de este tipo de pilotes, focalizando su atención especialmente en el sector eléctrico. No obstante, su experiencia se extiende a proyectos diversificados que abarcan distintas áreas de construcción. La empresa ha llevado a cabo numerosos proyectos de ingeniería en distintos países de Sudamérica, incluyendo Colombia, Chile y Perú.

El enfoque central de este trabajo se dedica al análisis del parámetro capacidad-torque (Kt) y su interrelación con distintos tipos de suelos y sus propiedades índices. El propósito principal se focaliza en obtener una medida más precisa de la capacidad última de los pilotes helicoidales, permitiendo así determinar con mayor certeza la carga máxima que puede soportar el pilote sin llegar a la falla. Esto enfoque busca optimizar el resultado de los análisis, proporcionando las bases necesarias para llegar a mejores conclusiones y, en última instancia, lograr un diseño óptimo de los pilotes helicoidales. En los proyectos que desarrolla la empresa, se ha evidenciado que la capacidad última de los pilotes calculada con la **Ecuación 1** subestima la capacidad real de los pilotes al tomar un valor Kt teórico, el cual depende únicamente del diámetro del fuste del pilote, esta conclusión se apoya mediante la ejecución pruebas de carga, las cuales permiten, en algunos casos, obtener la capacidad última real del pilote, y encontrar una desviación entre el valor obtenido con la medición directa y el valor calculado con la **Ecuación 1**. El factor Kt tiene una gran importancia en la instalación de los pilotes helicoidales, puesto según (Perko, 2009) la capacidad última de un pilote depende de este factor. A continuación, se presenta la ecuación para hallar la capacidad última:

$$Pu = Kt * T$$

Ecuación 1*:Capacidad ultima de los pilotes helicoidales.*

Donde:

Pu: Capacidad última del pilote
Kt: Factor de relación capacidad-torque.
T: Torque de instalación

El método de Davison se utiliza para calcular la capacidad de carga del pilote helicoidal (Pu). Este método ofrece una evaluación eficaz y precisa de la capacidad última del pilote en función de las condiciones específicas del suelo, este método utiliza el concepto de deformaciones admisibles en el rango elástico para aproximar la capacidad del pilote.

Al poder determinar la influencia del tipo de suelo y las propiedades índices de estos en el factor Kt, se podrá conocer con mayor certeza los valores de la capacidad real que puede soportar el pilote helicoidal y tener un mayor umbral en la carga que soportar el pilote comparada con la capacidad que entrega la **Ecuación 1** con Kt teórico. Esta investigación posibilitará la justificación técnica para minimizar, disminuir u optimizar la ejecución de pruebas de carga en determinados lugares donde no se alcance el torque objetivo. Al contar con una comprensión más precisa del comportamiento de la relación capacidad-torque en suelos con características geológicas y comportamiento de torque similares, se establece un fundamento sólido para respaldar esta elección. Esta información permitirá tomar decisiones fundamentadas al anticipar el rendimiento del pilote en contextos geotécnicos similares, contribuyendo así a la eficiencia y optimización de los procesos de diseño y ejecución de proyectos de cimentación. Axiatech S.A.S cuenta con una gran cantidad de datos obtenidos en los diferentes proyectos dentro de los cuales se destacan los registros de torque, propiedades índices de los suelos, geología regional y en algunos casos pruebas de carga. Con los datos mencionados anteriormente y el programa Python, se construirán códigos que permitan, el tratamiento y el análisis de los datos; encontrando matrices de correlación, las cuales permitan determinar la relación entre las variables dos a dos y adicionalmente encontrar patrones de comportamiento dentro de los datos; lo que posibilita hacer predicciones del comportamiento del factor Kt para condiciones similares en la estratigrafía y registro de torque. Se asumirá que el suelo tendrá el mismo comportamiento de la capacidad cuando se tenga una

estratigrafía y un registro de torque parecido, esto con el fin de extrapolar el análisis a los sitios donde no se tenga pruebas de carga.

1 Objetivos

1.1 Objetivo general

Aproximar la influencia del suelo y sus propiedades índice en la relación capacidad-torque (Kt) y a su vez en la capacidad última de los pilotes helicoidales mediante el procesamiento de datos con el lenguaje de programación Python.

1.2 Objetivos específicos

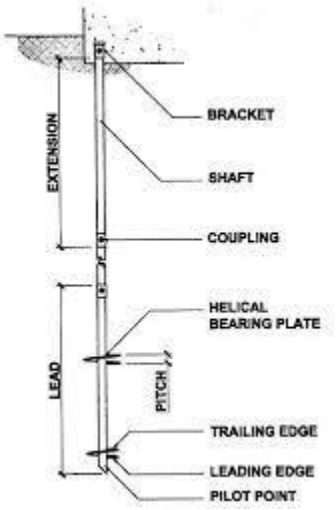
- Obtener una base de datos única integrando los diferentes bancos de información de los diferentes proyectos.
- Implementar código en el lenguaje Python que permita la organización de las bases de datos y el llenado de vacíos de información en las mismas.
- Determinar una matriz de correlaciones, donde se pueda determinar el grado de asociación entre las diferentes variables dos a dos.
- Analizar el comportamiento de la relación capacidad-torque en los diferentes tipos de suelos.
- Implementar el uso del Machine Learning para hallar patrones de comportamiento entre las variables analizadas

2 Marco teórico

(Perko, 2009) menciona que los pilotes helicoidales son cimentaciones de acero, estos son introducidos en el suelo mediante torsión. Los pilotes helicoidales están compuestos por un eje principal de acero con un juego de hélices o platos helicoidales, seguidos por extensiones del fuste de acuerdo con el diseño, como se observa en la **Figura 1.**

Figura 1

Partes de un pilote helicoidal



Nota: Fuente Helical piles

Dice (Perko, 2009) que los pilotes helicoidales soportan con una gran capacidad de carga tanto a compresión como a tracción, la cual los hace útiles en diferentes estructuras como torres de energía, estructuras comunes, hasta ser utilizados como anclajes en el sostenimiento de muros de concreto armado. Debido a la constante demanda por reducir tiempos en la construcción y utilizar materiales más amigables con el medio ambiente, la comunidad científica ha centrado esfuerzos en reducir los tiempos de construcción en cualquiera de sus fases, invirtiendo tiempo y dinero en la investigación de materiales que se enmarquen en los objetivos de desarrollo sostenible más específicamente en el objetivo número 9 el cual expresa “Construir infraestructuras resilientes, promover la industrialización sostenible y fomentar la innovación”(ONU, 2015). Es por lo anterior

que los pilotes helicoidales juegan un papel muy importante en las construcciones hoy día, primero son una práctica amigable con el medio ambiente al ser construidos, en algunos casos, con material reutilizado en la industria petrolera y segundo al reducir los tiempos en las obras debido a su rápida instalación, al no necesitar excavaciones ni requerir de tiempo de fraguado, esto en contraste con los pilotes convencionales de concreto, los cuales demandan largos tiempos de instalación al necesitar de una posible excavación y un tiempo para el fraguado del concreto.

El método más utilizado para predecir la capacidad última del pilote, denominado relación capacidad-par, es una correlación empírica entre el torque de instalación, en cual según se define como el promedio de los últimos 3 registros de torque antes de alcanzar su posición final, y la relación torque-Capacidad (Kt) (Perko, 2009). El factor de relación capacidad-torque (Kt), que se nombrará simplemente factor Kt, se ha utilizado en el diseño de pilotes helicoidales y anclajes durante más de medio siglo.

Se han realizado numerosas investigaciones para predecir con exactitud este factor. Sin embargo, casi todos estos factores Kt divulgados son en función de la geometría del fuste y se desprecia otras variables tales como el tipo de suelo y propiedades índice del suelo (Hoyt R & Clemence S, 1989). Según (Lutenegger et al., 2015) pruebas recientes a escala real (compresión axial y tracción) en arcilla, arena y lecho de roca han demostrado que el Kt tradicional utilizado, basado únicamente en el tamaño del eje puede mejorarse. Esta tendencia de que el factor Kt es subestimado solamente calculado en función del diámetro del pilote se ha evidenciado en los proyectos que ha participado la empresa Axiatech S.A.S, ya que en desarrollo de los proyectos se ha evidenciado que los torques previstos, basados en que el factor solamente depende del diámetro del pilote, no han sido posible alcanzarse en su mayoría y que presentan desviaciones significativas. No obstante, realizando pruebas de carga en algunos lugares, se ha validado que el desempeño de los pilotes satisface las solicitaciones para las que se diseñaron. La relación entre capacidad y par parece depender del tamaño del eje, la geometría del eje, la configuración de la hélice, la dirección de la carga y el tipo de suelo (Lutenegger et al., 2015). Las pruebas de carga mencionadas anteriormente según (ASTM, 2013) permite determinar la capacidad de un pilote in situ, mediante las deformaciones que se presentan en el pilote ante la aplicación de una carga.

(Cárdenas, 2018) menciona que la analítica de datos es fundamental para la toma de decisiones y que las herramientas de análisis de datos predictivos buscan mejorar el análisis de datos al predecir comportamientos futuros. Estas herramientas utilizan algoritmos de aprendizaje automático para identificar patrones en los datos existentes y luego usar esos patrones para predecir lo que sucederá en nuevos eventos. Es por lo anterior que se busca generar un programa en Python que permita analizar diferentes clases de datos que se tienen los cuales son: datos cualitativos que según (Peña, 2017) "se denominan datos cualitativos a todos aquellos que buscan caracterizar o resaltar atributos de un hecho, persona, comunidad, organización o situación no medible o sujeta a representación numérica." Además, de datos cuantitativos que se definen como "aquellos susceptibles a la medición y representación numérica" (Peña, 2017). Menciona (Raschka & Mirjalili, 2017) que Python es un lenguaje de programación de alto nivel y de propósito general que se ha vuelto muy popular en diversos campos, incluyendo desarrollo web, análisis de datos, inteligencia artificial, machine learning y más. Además, (Raschka & Mirjalili, 2017) hace énfasis en que es conocido por su sintaxis clara y legible y que Python tiene una amplia comunidad de desarrolladores y una gran cantidad de bibliotecas lo que facilitan el desarrollo en diversas áreas.

Para (Raschka & Mirjalili, 2017) el machine Learning, o Aprendizaje Automático en español, es un campo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones a partir de datos y tomar decisiones. Para efectos de este trabajo requiere el uso de regresiones para predecir el comportamiento de las variables involucradas. Dice (Montgomey Douglas et al., 2012) los modelos de regresión son un tipo de modelo estadístico que se utiliza para modelar la relación entre una variable dependiente (o de respuesta) y una o más variables independientes (o predictoras). El objetivo principal es entender cómo las variables independientes afectan a la variable dependiente y predecir sus valores. Hay varios tipos de regresión adaptados a diferentes situaciones y tipos de datos. Aquí hay algunos de los tipos de regresión más comunes:

- Regresión Lineal Simple: Modela la relación lineal entre una variable independiente y una variable dependiente.
- Regresión Lineal Múltiple: Extiende la regresión lineal simple para incluir múltiples variables independientes.

- Regresión Polinómica: Modela la relación entre variables utilizando un polinomio de grado superior.
- Regresión Logística: Utilizada cuando la variable dependiente es binaria (sí/no, 1/0). Estima la probabilidad de que ocurra un evento.
- Regresión Ridge (Regularización L2): Introduce una penalización en los coeficientes para evitar el sobreajuste (overfitting).
- Regresión Lasso (Regularización L1): Similar a Ridge, pero puede conducir a la selección de características al forzar algunos coeficientes a cero.
- Regresión Elastic Net: Combina los términos de penalización de Ridge y Lasso.
- Regresión No Lineal: Utilizada cuando la relación entre variables no es lineal. Puede incluir funciones no lineales como exponenciales, logaritmos, etc.
- Regresión de Poisson: Utilizada para variables de respuesta que siguen una distribución de Poisson, comúnmente en modelos de conteo.
- Regresión de Cox: Utilizada para datos de supervivencia en análisis de supervivencia.
- Regresión Robusta: Menos sensible a los valores atípicos que la regresión ordinaria.

Estos son solo algunos ejemplos, y existen muchas otras variantes y métodos específicos para diferentes casos de uso. La elección del tipo de regresión depende de la naturaleza de los datos y los objetivos del análisis.

3 Metodología

La estructura del trabajo se divide en las siguientes etapas claramente definidas, cada una desempeñando un papel crucial en el desarrollo general del proyecto:

3.1 Creación de la base de datos

En primera instancia se realiza un análisis en las carpetas de proyectos de la empresa para identificar todas las fuentes de datos relevantes para el proyecto, esto implica no solo identificar fuentes de información relevantes existentes en la empresa sino necesario realizar tareas de limpieza de datos para abordar posibles inconsistencias o valores faltantes necesarios para la creación de la matriz. Al tratarse de información distribuida en varios bancos de datos, es necesario llevar a cabo procesos de integración de datos. En la **Tabla 1** se presentan las librerías utilizadas para la generación del código y una breve descripción de las funcionalidades de cada una.

Tabla 1

Librerías utilizadas

Librería	Descripción
NumPy	Operaciones matriciales y numéricas eficientes en Python.
pandas	Estructuras de datos y herramientas de análisis de datos.
tabulate	Crea tablas ASCII elegantes a partir de listas de datos.
matplotlib	Biblioteca de visualización para gráficos estáticos.
seaborn	Basada en Matplotlib, proporciona una interfaz de alto nivel para gráficos estadísticos atractivos.
statsmodels	Modelos estadísticos para realizar análisis de datos.
sklearn	Herramientas simples y eficientes para minería y análisis de datos.
multiprocessing	Facilita la creación y gestión de procesos en paralelo.
random	Generación de números aleatorios en Python.
itertools	Proporciona funciones para trabajar con iterables de manera eficiente.

3.2 Limpieza de los datos

Posterior a la integración de las bases de datos, constituye una etapa muy importante preparar los datos antes del análisis. Durante este proceso, se detectan y corrigen errores, se normalizan y estandarizan los datos, se aborda la gestión de valores faltantes y se verifica la consistencia. Además, se manejan valores atípicos y se valida la conformidad de los datos con las bases de datos de origen. Este proceso asegura que los datos estén en un estado óptimo, mitigando posibles impactos negativos en la precisión de los modelos y contribuyendo a una toma de decisiones basada en datos confiable y efectiva.

3.3 Desarrollo del Código

Esta fase se centra en la generación de cadenas de código destinadas al análisis de los datos, y se lleva a cabo mediante el lenguaje de programación Python. Aquí, se desarrollan algoritmos que permitirán explorar, visualizar y modelar los datos de manera eficiente. La elección de Python como lenguaje de programación es estratégica, ya que ofrece una amplia gama de bibliotecas y herramientas especializadas en ciencia de datos, como pandas, NumPy y scikit-learn. Este enfoque no solo facilita la implementación de análisis avanzados, sino que también permite la integración de técnicas de machine learning para la generación de modelos predictivos. La fase de desarrollo de código no solo implica la creación de funcionalidades específicas, sino también la implementación de buenas prácticas de programación y documentación para garantizar la comprensión y mantenimiento a lo largo del tiempo.

3.4 Análisis de los datos

La etapa de análisis de datos constituye un paso fundamental en el proceso, desglosándose en dos fases complementarias. Inicialmente, el análisis descriptivo se enfoca en describir las características clave del conjunto de datos, haciendo un análisis exploratorio de los mismos para identificar patrones de comportamiento, este análisis básicamente arroja luz sobre tendencias. Este enfoque aplicado sobre los datos es bastante importante, ya que sienta las bases para inferencias y conclusiones fundamentadas. La segunda fase es el análisis predictivo, en esta se proyecta el

comportamiento futuro de los datos, siendo esencial la aplicación de algoritmos de Machine Learning para modelar y realizar pronósticos precisos. La combinación de estas subetapas no solo enriquece la comprensión de eventos pasados, sino que también capacita para tomar decisiones informadas basadas en la anticipación de tendencias y comportamientos futuros, otorgando un valor estratégico y práctico a la información analizada.

4 Resultados y Análisis

4.1 Creación y limpieza de la base de datos

Para la recolección de datos, se llevó a cabo un análisis de las diversas bases de datos de la empresa Axiatech obtenidas de los proyectos realizados. Después de examinar cada base de datos, se realizó una selección de los datos más relevantes, priorizando aquellos que ofrecían una mayor cantidad de información. Después, se integraron las bases de datos, que contenían información detallada, los datos proporcionados se presentan en forma tabular, donde cada columna representa una variable específica. La primera columna indica la "Profundidad", seguida por las columnas "T" (Torque), "Kt" (Factor de correlación Torque-Capacidad), "Qult" (Capacidad última), "N" (número de golpes/pie necesarios para atravesar le suelos con un tomamuestras), "USCS" (Clasificación del suelo según Sistema Unificado de Clasificación de Suelos), "Humedad", "LL" (Límite Líquido), "LP" (Límite Plástico), "IP" (Índice de Plasticidad), "Grava" (Porcentaje de grava), "Arena" (Porcentaje de arena), "Finos" (Porcentaje de finos), y "Tefect" (Torque efectivo) (Ver **Tabla 2**). Estos términos representan diversas propiedades y características geotécnicas asociadas a un determinado contexto, como pruebas de carga en proyectos de ingeniería. No obstante, la matriz de datos al final de este proceso se caracterizó por la presencia de numerosas variables con valores no registrados a profundidades donde se encontraba información de las pruebas de carga. Ante esta situación, se adoptó la siguiente estrategia, asumiendo el último valor registrado de cada variable hasta la profundidad donde se disponía de información de las pruebas de carga. Esta aproximación permitió abordar la falta de datos en ciertas profundidades y posibilitó la continuación del análisis y modelado con una matriz de datos más completa.

Aunque esta solución podría generar ciertas incertidumbres al asumir el último valor registrado para las variables sin datos a profundidades específicas, es crucial destacar que el enfoque adoptado se centra en la obtención del programa inicial. Este primer paso es esencial para avanzar en el análisis y modelado, permitiendo la identificación de patrones y tendencias iniciales. Además, se reconoce la provisionalidad de los resultados dados los supuestos realizados. Se planifica ejecutar nuevamente el programa en el futuro, una vez se disponga de información más completa y detallada. Este enfoque iterativo garantiza que, a medida que se recopile y se integre

más información, el programa se ajustará y refinará, proporcionando resultados más precisos en las iteraciones subsiguientes.

En la **Tabla 2** se presenta la matriz final con la que se hizo el entrenamiento, predicción y validación del programa.

Tabla 2

Matriz

Profundidad [m]	T [lb*ft]	Kt [ft^-1]	Qult [kN]	N	USCS	Humedad [%]	LL	LP	IP	Grava [%]	Arena [%]	Finos [%]	Tefect [lb*ft]
6	8370	15	491	95	CL	8	38	13	25	0	4	96	7559
10	9850	16	592	10	ML	25	NL	NP	NP	0	8	92	8160
6	13115	13	691	95	CL	8	38	13	25	0	4	96	11765
8	6467	12	327	15	SC	9	NL	NP	NP	30	55	15	6352
6	11327	11	500	95	CL	9	38	13	25	0	17	83	9928
12	10483	13	556	4	CL	18	30	11	19	0	16	84	9973
7	5790	14	327	18	SC	2	NL	NP	NP	0	74	26	5440
7	7328	24	650	43	CL	7	32	11	21	0	21	79	6132
6	7097	21	525	15	CL	13	27	15	12	0	33	67	5662
6	7428	15	429	43	ML	16	NL	NP	NP	2	36	62	6294
11	11179	8	359	29	SW-SC	3	NL	NP	NP	7	83	10	9891
8	11516	11	465	21	SW-SC	11	NL	NP	NP	2	86	12	9883
12	14091	10	623	37	SC	20	NL	NP	NP	0	54	46	13927
7	12520	12	525	48	CL	16	34	15	19	0	38	62	10210
7	5390	23	486	19	CL	17	36	18	18	0	18	82	4672
9	7547	15	475	46	SW-SC	12	NL	NP	NP	14	80	6	6950
6	4110	22	379	27	CL	13	25	10	15	1	48	51	3868
3	15595	10	476	32	SC	3	NL	NP	NP	0	60	40	11203
5	12250	10	524	69	CL	10	45	15	30	0	6	94	11478
4	7723	18	562	95	ML	6	NL	NP	NP	0	5	95	7019
10	10636	10	450	13	ML	11	NL	NP	NP	0	43	57	9775
7	4463	24	450	59	SC	1	NL	NP	NP	13	75	12	4281
4	12928	11	535	95	CL	10	42	17	25	0	1	99	10782
5	3953	28	435	31	SC	4	NL	NP	NP	2	78	20	3536
6	5400	23	548	28	SC	20	NL	NP	NP	0	66	34	5298
6	9390	15	586	30	SW-SC	11	NL	NP	NP	0	90	10	8665
8	3600	33	631	25	CL	14	41	14	27	0	35	65	4317
5	3670	28	450	22	SC	24	NL	NP	NP	0	88	12	3563
5	5425	16	465	79	SC	17	NL	NP	NP	0	86	14	6428

Profundidad [m]	T [lb*ft]	Kt [ft^-1]	Qult [kN]	N	USCS	Humedad [%]	LL	LP	IP	Grava [%]	Arena [%]	Finos [%]	Tefect [lb*ft]
6	11250	13	582	15	SC	12	28	16	12	0	56	44	10218
5	3220	39	535	22	SW	14	NL	NP	NP	11	85	4	3067
6	5829	22	580	39	SH	16	50	19	31	1	6	93	5878
6	4525	29	487	19	SW-SC	12	NL	NP	NP	16	74	10	3795
7	9375	17	673	27	CL	17	30	13	17	1	35	64	8735
12	11382	12	580	15	ML	14	NL	NP	NP	0	44	56	10828
6	12805	10	536	8	SC	11	NL	NP	NP	1	67	32	12326
7	5015	31	558	31	SC	16	NL	NP	NP	0	55	45	4064
5	7663	22	735	55	SC	7	NL	NP	NP	10	75	15	7431
5	3500	29	385	37	GC	5	53	34	13	53	34	13	2962

Como se evidencia en la **Tabla 2**, la matriz resultante de los procesos previos presenta un total de 40 filas y 14 columnas, con 14 variables en total, de las cuales 13 son independientes y 1 es la variable respuesta, que para este caso en particular es el parámetro Kt. Dentro de las variables más destacadas se tiene que la profundidad varía en un rango comprendido entre 3 y 12 metros, con un promedio de 7 metros. Respectivamente, el torque exhibe una variación entre 3220 y 15595 lb*ft, con una media de 8287 lb*ft. La variable respuesta muestra una variación entre 8 y 39 ft^{-1}, con un promedio de 18 ft^{-1}; mientras que la capacidad última se sitúa entre 735 kN y 327 kN, con un promedio de 323 kN. Estos detalles proporcionan un panorama detallado de la diversidad y distribución de las variables en la matriz, permitiendo una comprensión más profunda de la información recopilada y allanando el camino para análisis y modelado subsiguientes.

Dentro del proceso de limpieza de datos, se realizó una verificación de la consistencia de estos. Este análisis se centró en asegurar que los valores registrados fueran congruentes y estuvieran dentro de los rangos esperados. Se realizaron comparaciones y validaciones para identificar posibles discrepancias, anomalías o valores atípicos que pudieran afectar la integridad de los datos.

4.2 Creación del código

En los subcapítulos siguientes se expone detalladamente el contenido en el código.

4.2.1 Análisis exploratorio

Para iniciar con el análisis exploratorio, se va a indagar sobre el tipo de variable de cada columna de la **Tabla 2**. En la **Tabla 3** se puede observar el tipo de las variables que intervienen en el problema de predicción, se tiene valores tanto numéricos como de cualitativas. Dentro de las variables numéricas encontramos enteros (int64) y valores reales (float64) y para las cualitativas tenemos cadenas de texto o string (object).

Tabla 3

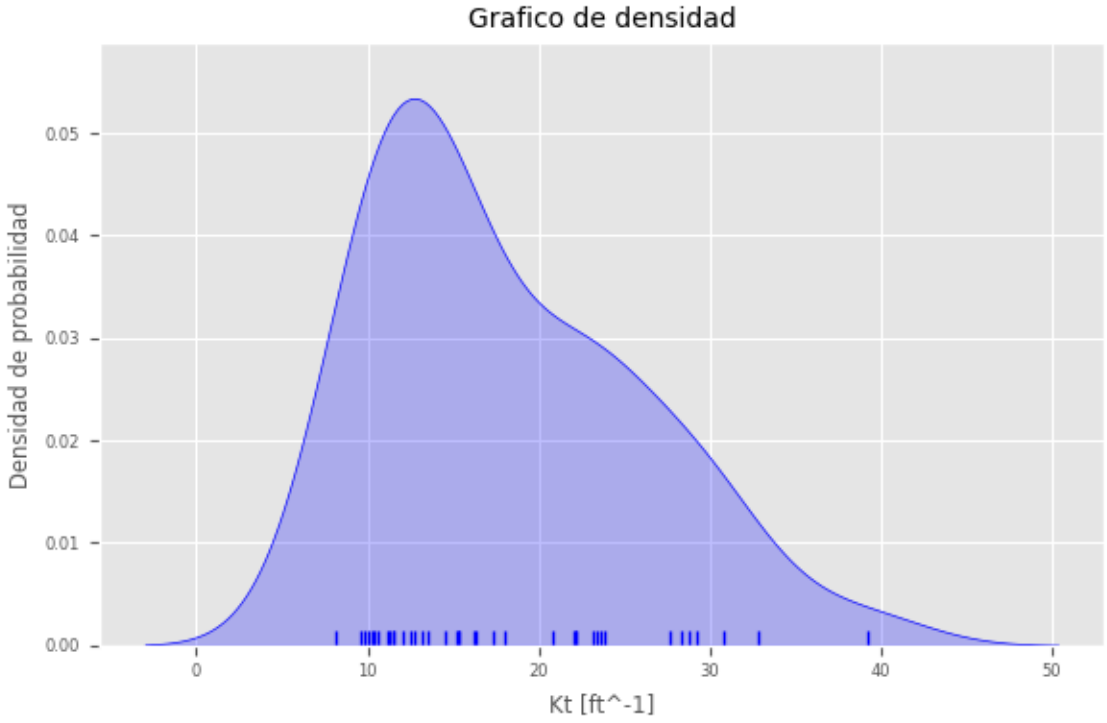
Tipo de variable de las columnas

Variable	Tipo de variable
Profundidad	float64
T	float64
Kt	float64
Qult	int64
N	object
USCS	float64
Humedad	object
LL	int64 and object
LP	int64 and object
IP	int64 and object
Grava	int64
Arena	int64
Finos	int64
Tefect	float64

Cuando se desarrolla un modelo, resulta importante examinar la distribución de la variable objetivo, ya que, en última instancia, es lo que se busca predecir.

Figura 2

Distribución de la variable respuesta

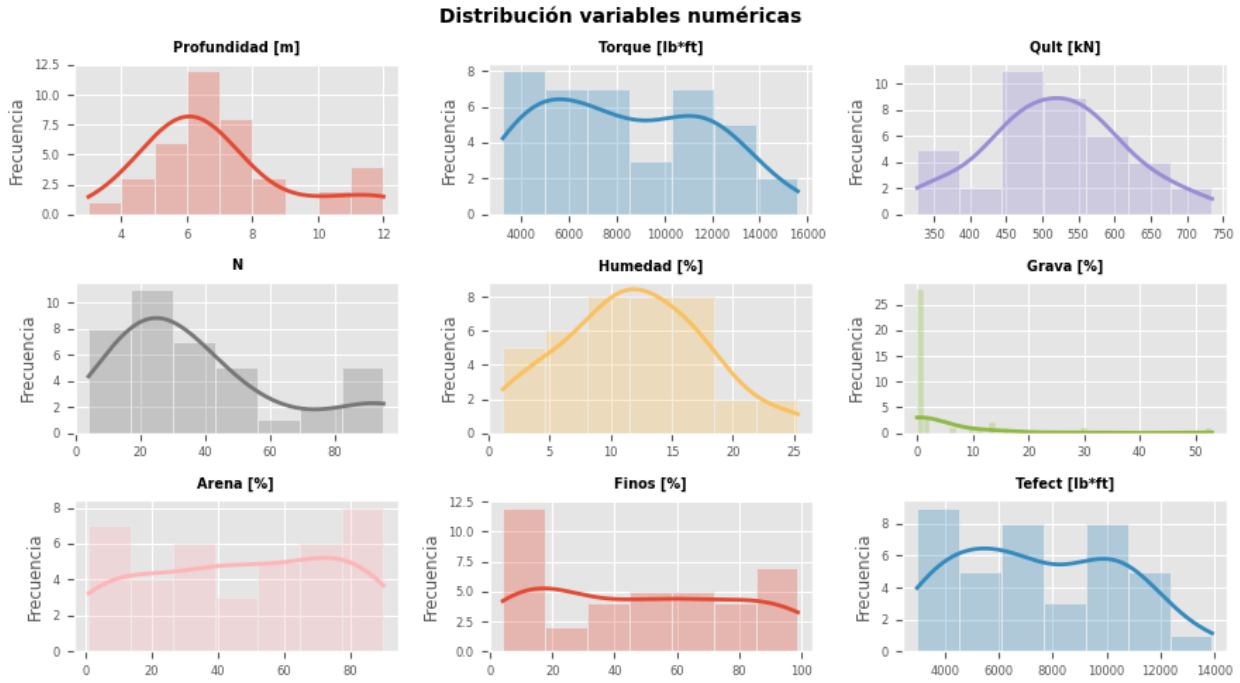


En la **Figura 2** se presenta el gráfico de densidad para la variable 'Kt', este proporciona una representación visual de la distribución de probabilidad de los datos, los cuales presentan una asimetría positiva, pues los datos se acumulan a la izquierda del gráfico.

Para proseguir con el análisis exploratorio, se procederá con la clasificación de las variables en cuantitativas y cualitativas. En el caso de las variables cuantitativas, se generarán histogramas individuales para cada una. Este enfoque permitirá visualizar la distribución de frecuencias y obtener percepciones más detalladas sobre la forma y la dispersión de cada variable cuantitativa en el conjunto de datos. La representación gráfica mediante histogramas facilitará la identificación de patrones, tendencias y posibles valores atípicos. En la **Figura 3,** se logra determinar los diferentes comportamientos de las variables independientes.

Figura 3

Distribución de las variables numéricas



Luego, se procede a generar un gráfico de dispersión para cada variable en relación con la variable respuesta. Esto proporciona una representación gráfica detallada de la relación entre cada variable independiente y la variable de interés, permitiendo una evaluación visual de posibles patrones, tendencias y correlaciones. La información revelada en estos gráficos puede ser crucial para comprender la influencia relativa de cada variable en la variable respuesta. La observación de la dispersión de los puntos en el gráfico facilita la identificación de cómo se relacionan las variables, así como la detección de posibles valores atípicos que podrían afectar la calidad del modelo. Este análisis es muy importante en el proceso de exploración de datos, proporcionando tendencias valiosas para orientar análisis y la toma de decisiones en etapas posteriores del análisis de datos. El resultado de la ejecución se presenta en la **Figura 4**. Es importante mencionar que la variable respuesta Kt está representada en el eje x, mientras que las demás variables cuantitativas están representadas en el eje y.

Figura 4

Correlaciones con Kt

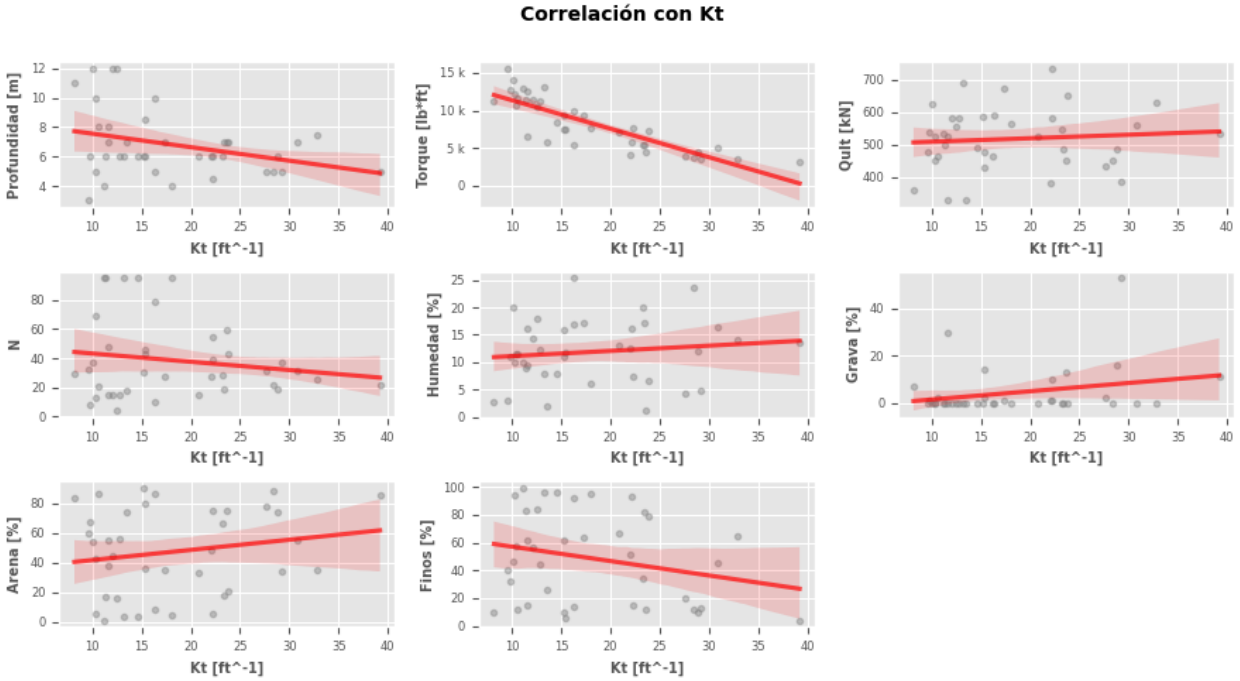
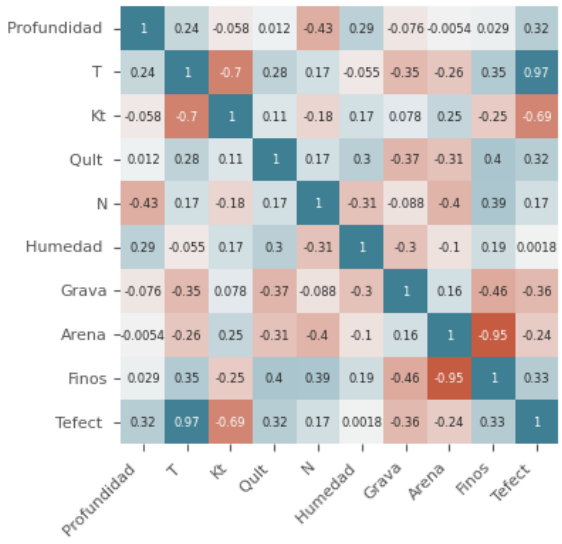


Figura 5

Matriz de correlaciones



El resultado en términos de coeficientes de correlación se muestra en la **Figura 5**. La matriz de correlaciones contiene la representación numérica del ajuste de las variables en el problema. Este

coeficiente se calcula dos a dos para cada par de variables, las variables se ajustan a la línea recta si el valor es cercano a uno (1) y entre más se acerca a cero (0) la correlación es más baja.

4.2.2 División entrenamiento y prueba

La evaluación de la capacidad predictiva de un modelo implica verificar la proximidad entre sus predicciones y los valores reales de la variable respuesta. Para cuantificar este desempeño de manera precisa, es esencial contar con un conjunto de observaciones no visto durante el ajuste del modelo, conocido como conjunto de prueba. Para lograr esto, se divide el conjunto de datos disponible en un conjunto de entrenamiento y un conjunto de prueba. La selección adecuada del tamaño de estas divisiones depende de la cantidad de datos disponible y del nivel de confianza necesario en la estimación del error; comúnmente, una proporción del 80%-20% ha demostrado ser efectiva. La asignación de las particiones debe realizarse de manera aleatoria o aleatoria-estratificada para garantizar una representación justa de la variabilidad de los datos en ambos conjuntos.

Es crucial verificar que la distribución de la variable respuesta sea similar tanto en el conjunto de entrenamiento como en el de prueba al realizar la partición de los datos. Para garantizar esto, la función train_test_split de scikit-learn ofrece la posibilidad de conservar las características en los dos conjuntos extraídos. En la ***Tabla 4*** y **Tabla 5** se logra determinar que las características de los conjuntos de entrenamiento y prueba son parecidos en cuanto a las medidas de tendencia central y desviación.

Tabla 4

Partición de entrenamiento

Partición de entrenamiento	
Numero de valores	31.00
media	21.70
Desviación estándar	6.55
máximo	32.86
mínimo	10.05

Tabla 5

Partición de prueba

Partición de prueba	
Numero de valores	8.00
media	19.57
Desviación estándar	5.03
máximo	26.33
mínimo	9.55

4.2.3 Preprocesamiento

Cuando los predictores numéricos tienen diferentes escalas o variabilidades, pueden influir significativamente en modelos de machine learning. Algoritmos como SVM, redes neuronales y Lasso son sensibles a estas disparidades, lo que puede llevar a que predictores con mayores escalas o varianzas dominen el modelo injustamente. Para abordar esto, dos estrategias comunes son la normalización, que ajusta las escalas de los predictores, y las transformaciones robustas, que son útiles en presencia de valores atípicos o distribuciones no lineales. En este caso se hizo uso de la normalización de los datos para llevarlos.

La binarización, o one-hot-encoding, implica la creación de variables dummy para cada nivel de variables cualitativas. Al utilizar OneHotEncoder con drop='first', se eliminan redundancias y se evitan problemas en modelos sensibles a la correlación perfecta entre predictores,

como los lineales sin regularización o las redes neuronales. El preprocesamiento en scikitlearn se realiza eficientemente mediante ColumnTransformer y pipelines, y el módulo sklearn.preprocessing ofrece diversas opciones adicionales para la transformación de datos. Básicamente lo que hace el programa es transformar las variables cualitativas en números, para que el modelo lo pueda leer y hacer predicciones más eficientes.

La matriz resultado luego de aplicar estas transformaciones, tanto a las variables cuantitativas como cualitativas, a la matriz original se obtiene la siguiente:

Tabla 6

Matriz transformada

Profundidad	T	Qult	N	Humedad	LL	LP	IP	Grava	Arena	Finos	Tefect
0.07	1.29	0.03	0.29	0.82	0.08	1.45	0.68	0.53	0.49	0.56	0.88
0.07	1.11	0.71	0.70	01.94	0.08	0.13	0.04	1.40	0.79	1.01	1.09
0.39	0.05	0.30	2.04	00.69	1.01	0.13	1.23	0.53	1.67	1.63	0.00
0.07	0.26	1.28	0.10	0.94	0.62	1.70	0.04	0.53	1.08	1.09	0.47
0.52	0.51	1.94	0.94	0.42	0.08	0.13	0.04	3.91	0.09	0.92	00.40
1.90	0.89	1.62	0.42	1.67	0.08	0.13	0.04	0.51	1.06	1.08	0.78
0.07	0.71	1.94	0.83	1.79	0.08	0.13	0.04	0.53	0.75	0.57	0.70
0.39	1.46	1.69	2.04	0.69	1.01	0.13	1.23	0.53	1.67	1.63	1.40
2.35	0.95	0.58	0.94	0.47	0.08	0.13	0.04	0.53	0.29	0.37	1.09
0.39	0.91	0.60	0.94	0.09	1.72	2.23	2.91	0.53	0.13	0.01	0.89
0.85	1.35	0.71	0.68	2.21	0.08	0.13	0.04	0.53	1.24	1.01	1.32
0.85	1.21	0.02	1.07	0.32	2.92	1.45	2.83	0.53	1.60	1.57	1.30
0.39	1.37	0.14	1.20	00.13	0.08	0.13	0.04	0.38	0.51	0.38	1.59
0.39	0.23	0.92	0.10	0.75	0.08	0.13	0.04	0.23	0.56	0.56	0.42
0.39	0.36	0.64	0.38	0.15	0.08	0.13	0.04	0.53	1.30	0.08	0.37
2.35	0.68	0.34	0.35	1.16	0.17	0.70	0.68	0.53	0.25	1.25	0.80
0.39	0.21	0.42	0.49	0.17	0.53	0.48	0.95	0.38	0.15	0.21	0.22
0.07	0.95	0.36	0.34	0.87	0.08	0.13	0.04	0.53	0.09	0.03	0.16
0.31	1.41	0.13	2.04	0.33	2.10	3.02	1.23	053	01.77	1.72	1.07

Profundidad	T	Qult	N	Humedad	LL	LP	IP	Grava	Arena	Finos	Tefect
0.07	0.35	1.51	0.49	1.00	0.17	0.13	0.32	0.38	0.60	0.62	0.39
0.85	0.82	0.56	1.44	0.97	0.08	0.13	0.04	0.53	1.17	0.95	0.37
0.08	0.16	2.13	0.55	0.81	0.08	0.13	0.04	0.95	0.79	0.92	004
0.39	1.09	0.34	0.79	0.05	0.08	0.13	0.04	1.84	0.75	1.08	01.25
0.52	0.99	0.56	0.71	0.06	0.08	0.13	0.04	0.23	1.17	1.01	0.77
0.85	1.48	0.13	0.68	0.33	0.08	0.13	0.04	0.23	1.17	1.01	1.49
0.30	1.37	1.09	0.57	0.42	1.83	0.66	1.87	0.53	0.60	0.65	01.07
0.85	1.26	0.86	0.34	1.37	0.08	0.13	0.04	0.23	0.89	0.76	01.33
0.75	0.19	0.47	0.22	0.00	0.08	0.13	0.04	1.55	0.96	1.20	0.20
0.39	0.83	0.26	0.45	1.54	0.08	0.13	0.04	0.53	0.48	0.32	0.75
2.35	1.75	1.01	0.12	1.52	0.08	0.13	0.04	0.53	0.06	0.06	2.12
1.31	0.14	0.40	2.04	1.04	0.08	0.13	0.04	0.53	1.63	1.60	0.18

De aquí en adelante la **Tabla 6** servirá de referencia para realizar los procedimientos restantes.

4.2.3 Entrenamiento

Se procede a ajustar el modelo de regresión lineal con regularización ridge para predecir el comportamiento de la variable respuesta (Kt) utilizando todos los predictores disponibles. Se emplea la implementación Ridge de scikit-learn con los argumentos predeterminados. Es fundamental recalcar que, de aquí en adelante se utilizarán los datos de la **Tabla 6**. Para garantizar que el preprocesamiento se realice exclusivamente con los datos de entrenamiento y no se filtre información del conjunto de prueba, se emplea un pipeline que combina las transformaciones necesarias (como la estandarización) con el entrenamiento del modelo. Esta práctica asegura una aplicación coherente del preprocesamiento en ambos conjuntos de datos, contribuyendo a una evaluación precisa y justa del rendimiento del modelo en datos no vistos.

4.2.4 Validación

El error mostrado por defecto después de entrenar un modelo es el error en el conjunto de entrenamiento, que mide la diferencia en las predicciones para las observaciones ya conocidas, en la **Tabla 7** se presenta el error obtenido. Para obtener una estimación más precisa antes de recurrir al conjunto de prueba, se pueden emplearon estrategias de validación basadas en resampling. Scikit-learn. Estas estrategias, como cross_val_score(), toman como primer argumento un estimador, que puede ser directamente un modelo o un pipeline. Es crucial destacar que las métricas de error de regresión se devuelven en negativo para que un valor más cercano a 0 representa un ajuste mejor.

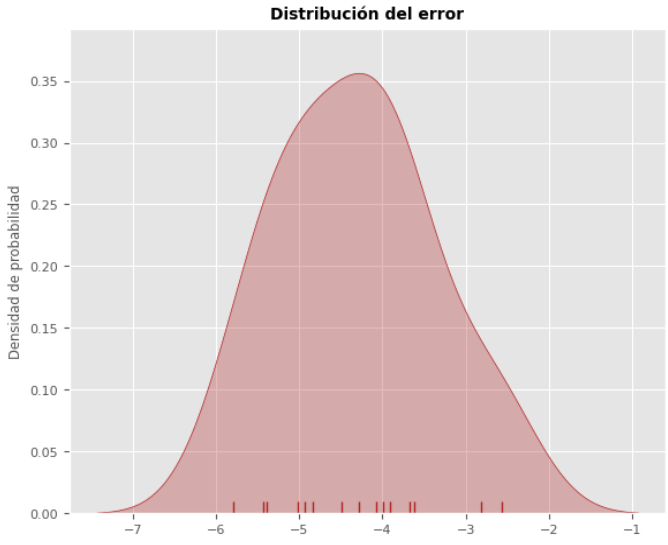
Tabla 7

Media de las métricas de validación cruzada

Media de las métricas de validación cruzada	-4.95

Figura 6

Distribución del error

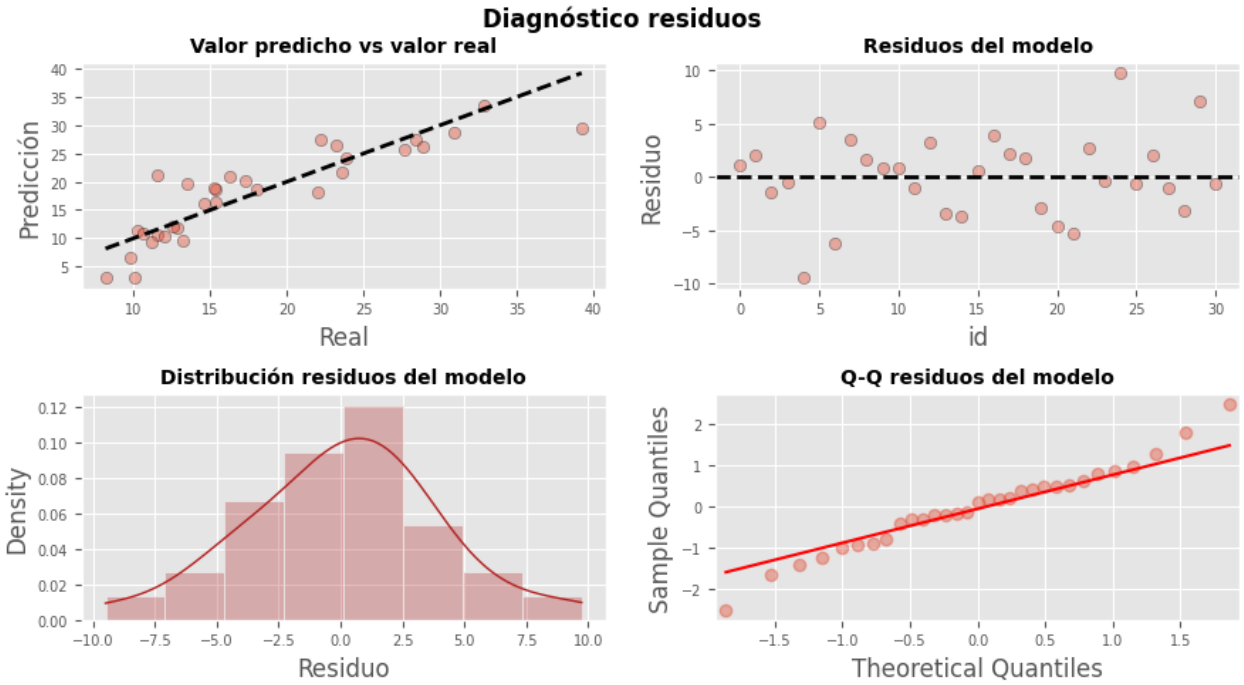


La **Figura 6** representa la distribución de los errores derivados del procedimiento en el conjunto de entrenamiento revela una tendencia hacia la media de los datos. Este patrón de comportamiento sugiere un ajuste normal en los datos presentados, indicando que las predicciones del modelo

tienden a agruparse alrededor de la media de los valores reales. Este comportamiento es un indicativo favorable de la capacidad del modelo para capturar la variabilidad de los datos de entrenamiento y generar predicciones cercanas a los valores observados. En la **Figura 7** se presentan 4 gráficos, los cuales se describen a continuación. El gráfico "valor predicho vs valor real" es un gráfico de dispersión de los valores reales vs los predichos el modelo, el resto de los gráficos son representaciones de los residuos que arrojan información de los residuos que arroja el modelo.

Figura 7

Diagnóstico de los residuos



La gráfica nombrada "valor predicho vs valor real" incluida en la **Figura 7** exhibe las predicciones realizadas por el modelo de entrenamiento evidenciando un notable ajuste a los datos, sugiriendo una alta confiabilidad en las predicciones dentro de este conjunto de valores. Este fuerte ajuste indica que el modelo ha aprendido de manera efectiva las relaciones entre los predictores y la variable respuesta en el conjunto de entrenamiento. No obstante, es crucial destacar que la evaluación de la confiabilidad del modelo debe extenderse al conjunto de prueba para garantizar su capacidad de generalización a nuevos datos. A continuación, se incluirán los valores arrojados

por el modelo para el conjunto de prueba, proporcionando una evaluación exhaustiva de su desempeño en un conjunto de datos independiente.

4.2.5 Predicción

Finalmente, se presentan las predicciones del modelo una vez que ha sido entrenado. Estas predicciones representan las estimaciones del modelo para la variable respuesta (Kt) basadas en las características de nuevas observaciones de datos no vistos durante el entrenamiento. Esta etapa es crucial para evaluar el rendimiento del modelo en situaciones del mundo real y para entender cómo se generaliza a datos que no formaron parte del proceso de aprendizaje. La **Tabla 8** presentan tanto los valores reales, columna denominada "Kt", como las predicciones del modelo, columna "Predicción"; como se puede observar los valores predichos por el modelo tienen una baja dispersión si tenemos presente que el número de datos utilizados dentro del desarrollo del proyecto es bajo.

Tabla 8

Predicciones del modelo

kt	predicción
20.844538	21.205380
29.219431	29.316511
23.387298	25.706241
16.300000	20.010436
11.321480	9.493852
22.182178	29.969536
10.349579	11.050441
9.551582	5.524458

5 Conclusiones

1. Se llevó a cabo la depuración de las bases de datos existentes en la empresa, consolidando la información en una única matriz utilizada en los códigos de entrenamiento, validación y predicción.
2. Se realizó un análisis exploratorio de las variables, identificando relaciones entre ellas mediante un análisis de dos a dos y examinando los factores de correlación.
3. Se implementó un código en Python utilizando metodología de aprendizaje automático para predecir la variable Kt en función de las propiedades índices del suelo, el registro de torque y la capacidad última del suelo. El modelo seleccionado fue una regresión lineal.
4. El código resultante demostró una capacidad de predicción notable y se ajustó eficazmente a los datos fuente, indicando un rendimiento satisfactorio del modelo en la estimación de la variable Kt basada en las características del suelo y otros factores considerados.
5. Los errores entregados por el programa en el proceso de predicción fueron relativamente pequeños.
6. Es importante consolidad una base de datos más robusta, con el fin de alimentar mejor el modelo y así obtener predicciones cada vez más precisas.

6 Recomendaciones

Como se dijo en un principio la actividad de recolección de la información fue de las más difícil de completar, motivo por el cual se decidió suponer los datos hasta la profundidad donde se tenía la información de la prueba de carga. Es por lo anterior que se harán una serie de recomendaciones para construir la matriz y poder alimentar de una mejor manera el código realizado.

- Tratar de gestionar información del suelo de los proyectos pasados y complementar la matriz con esta información.
- Obtener datos de propiedades índices a la profundidad de las pruebas de carga en futuros proyectos.
- Consolidar información que se encuentra en la red de pruebas de carga en diferentes tipos de suelos, esta recopilación de datos se puede obtener de diversas fuentes, como bases de datos geotécnicas e informes de proyectos anteriores.
- Mantener la matriz de trabajo actualizada a medida que se obtiene nueva información.

Es importante resaltar que, si bien en principio la matriz se definió con el conjunto de variables que se presentaron en el documento, estas pueden variar y hacer una selección diferente, incluso el programa se puede enfocar de otra manera a otro proceso dentro de la empresa para resolver necesidades puntuales.

Referencias

ASTM. (2013). *Métodos Normativos Para Pruebas de Pilotes Bajo Carga Estática de Compresión Axial.*

Cárdenas, J. M. (2018). *El machine learning a través de los tiempos, y los aportes a la humanidad denniye hinestroza ramírez*.

Hoyt R, & Clemence S. (1989). *Uplift Capacity of Helical anchors in Soil*.

Lutenegger, A. J., Erikson, J., & Williams, N. (2015). *Evaluating installation disturbance of helical anchors in clay from field vane tests*.

Montgomey Douglas, Peck Elizabeth, & Vinning Geoffrey. (2012). *Introduction to Linear Regression Analysis*.

ONU. (2015). *Objetivos y metas de desarrollo sostenible*.

Peña, S. (2017). *Análisis de Datos*.

Perko, H. A. (2009). *Helical piles: a practical guide to design and installation*.

Raschka, Sebastian., & Mirjalili, Vahid. (2017). *Python machine learning: machine learning and deep learning with Python, scikit-learn, and TensorFlow*.