UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos

Obtención del perfil penetrométrico del suelo aplicando *Machine Learning* al registro de parámetros de perforación

TESIS DOCTORAL

Presentada para optar al título de Doctor por:

Eduardo Martínez García
Ingeniero de Caminos, Canales y Puertos

Madrid, 2025

UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos

Doctorado en Ingeniería de Estructuras, Cimentaciones y Materiales

Obtención del perfil penetrométrico del suelo aplicando *Machine Learning* al registro de parámetros de perforación

TESIS DOCTORAL

Presentada para optar al título de Doctor por:

Eduardo Martínez García
Ingeniero de Caminos, Canales y Puertos

Bajo la dirección de:
Dr. Marcos García Alberti
Dr. Antonio Alfonso Arcos Álvarez

Madrid, 2025

Título: *Machine Learning* aplicado al registro de parámetros de perforación. Obtención del perfil penetrométrico del suelo

Autor: Eduardo Martínez García

Programa de Doctorado: Doctorado en Ingeniería de Estructuras, Cimentaciones y Materiales

Dirección de tesis:

Dr. Marcos García Alberti, Departamento de Ingeniería Civil: Construcción, Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos, Universidad Politécnica de Madrid (Director)

Dr. Antonio Alfonso Arcos Álvarez, director Departamento de Ingeniería y Morfología del Terreno, Escuela Técnica Superior de Ingenieros de Caminos, Canales y Puertos, Universidad Politécnica de Madrid (Codirector)

Revisores externos:

Tribunal de tesis:

Fecha de defensa:

A mi padre.

Agradecimientos

Quiero empezar estos agradecimientos dejando claro que son todos los que están, pero no están todos los que son. A todas aquellas personas que han contribuido a hacer posible este trabajo, gracias.

Primero quiero agradecer a mis directores, Marcos y Antonio, por su ayuda, apoyo y, sobre todo, por su genuino interés en este trabajo desde la primera vez que nos reunimos. Su entusiasmo ha sido parte imprescindible de esta tesis.

Por otro lado, a mis compañeros de Menard España, por ser unas personas con las que da gusto trabajar. Tengo que citar especialmente a mis superiores, José Antonio y Juan Carlos, por apreciar el proyecto y darme el visto bueno desde el inicio. También por ayudarme con la elección de las obras analizadas. Y a Joaquín por su asesoramiento en los apartados más relacionados con la maquinaria.

Sin salir de la empresa, a Menard R+D y a Piotr, también por su apoyo desde el momento en que esta investigación llegó a su conocimiento y por su confianza en mí para llevarla a la práctica. La ayuda de Menard ha permitido la presentación de este trabajo en varios de los congresos en los que se ha publicado.

A Rocío, mi pareja, la gran damnificada de estos años de trabajo, por hacer mi vida más amable.

A mi madre, por su labor de corrección, y por todo lo demás.

Abstract

Geotechnics is a branch of engineering that deals with large uncertainties. Soils and rocks are materials subjected to natural, uncontrolled processes that are, in many cases, not fully understood. Before working with these materials, survey campaigns are commonly conducted, which face budgetary and access limitations. When the works are being done, there is an absence of certainty about the actual nature of the soil.

In the case of soil drilling, one way to reduce uncertainty is by using a parameter recording system. This enables tracking how the drill's parameters varied during the process, providing an idea of the characteristics of the materials encountered. However, these parameters do not have a direct relationship with the strength and deformation parameters commonly used in geotechnics.

On another note, machine learning is a branch of artificial intelligence aimed at extracting information from large datasets, which has revolutionized countless fields in recent years.

This thesis analyses the application of machine learning to drilling data with the objective of reducing uncertainty about the soil being drilled. The data comes from the recorded execution parameters of rigid inclusions created through ground displacement. The goal is to correlate this data with cone penetration resistance, whether static or dynamic.

To achieve this, all the necessary steps for applying machine learning to a problem of this nature have been studied in depth: nature and format of the input data, anomaly detection, dataset construction, feature engineering, selecting the best algorithms, explainability of the resulting models, etc.

According to the results, ensemble trees algorithms deliver the best results and are capable of accurately reproducing the soil's penetrometric profile. It was also found that the proposed additional features improve the models. However, oversampling does not appear to be useful in this case, and using depth as a feature proves problematic.

Results for estimating the penetrometric profile of one site using data from another site are poor. However, analyzing data from a group of sites yields more promising results.

Ultimately, machine learning algorithms can be a valuable tool for obtaining the penetrometric profile of the traversed soil based on drilling data from rigid inclusions. This allows for additional information about the soil to be obtained at each drilling point, in the same terms that would be provided by conducting a geotechnical investigation. In this way, the uncertainty associated with the executed rigid inclusions is reduced.

Resumen

La geotecnia es una rama de la ingeniería que se enfrenta a unas grandes incertidumbres. Los suelos y rocas son materiales sometidos a procesos naturales no controlados y, en muchos casos, no completamente conocidos. Antes de trabajar sobre estos materiales se realizan campañas de ensayos que se encuentran con limitaciones presupuestarias y de acceso. Cuando finalmente se ejecutan los trabajos, no se tiene una total certeza de la naturaleza real del suelo.

En el caso de trabajos de perforación del suelo, una manera de reducir la incertidumbre es contar con un sistema de registro de parámetros. Esto permite conocer como han variado los parámetros de perforación durante el proceso y tener una idea de las características de los materiales atravesados. Sin embargo, estos parámetros no tienen una relación directa con los parámetros de resistencia y deformación que se utilizan normalmente en geotecnia.

Por otro lado, el aprendizaje automático es una rama de la inteligencia artificial que busca extraer información de grandes conjuntos de datos y que ha revolucionado infinidad de campos en los últimos años.

En esta tesis se analiza la aplicación del aprendizaje automático a datos de perforación con el objetivo de reducir la incertidumbre sobre el suelo atravesado. Los datos provienen del registro de parámetros de ejecución de inclusiones rígidas realizadas mediante desplazamiento del terreno. Se busca relacionar estos datos con la resistencia a la penetración en el ensayo penetrométrico, ya sea estático o dinámico.

Para ello se han estudiado en profundidad todos los pasos necesarios para la aplicación del aprendizaje automático en un problema de este tipo: la naturaleza y formato de los datos de entrada, la detección de anomalías, la constitución de los conjuntos de datos, la ingeniería de características, la elección de los mejores algoritmos, la explicabilidad de los modelos obtenidos, etc.

Según los resultados obtenidos, los algoritmos de árboles ensamblados son los que presentan mejores resultados y son capaces de reproducir de manera fiel el perfil penetrométrico del suelo. También se ha comprobado que las características adicionales propuestas mejoran los modelos. Por otro lado, el sobremuestreo no parece ser útil en este caso, y el uso de la profundidad como características resulta problemático.

Los resultados a la hora de estimar el perfil penetrómetrico de una obra a partir de datos de otra son pobres. Pero analizando datos de un conjunto de obras los valores obtenidos sí son prometedores.

En definitiva, los algoritmos de aprendizaje automático pueden ser una interesante herramienta para obtener el perfil penetrométrico del suelo atravesado a partir de los datos de perforación de inclusiones rígidas. Esto permite obtener información adicional del suelo en cada punto en que se ejecuta una perforación en los mismos términos que nos la proporcionaría la ejecución de un reconocimiento geotécnico. De esta manera se disminuye la incertidumbre asociada a las inclusiones rígidas ejecutadas.

Tabla de Contenido

Agradecimientos ..... *v*
Abstract ..... *vii*
Resumen ..... *ix*
Lista de Figuras ..... *xiv*
Lista de Tablas ..... *xx*
Abreviaturas, Acrónimos y Símbolos ..... *xxi*
Glosario ..... *xxvii*

1. Introducción y motivación **1**

1.1. Alcance de la investigación 4
1.2. Estrategia de investigación 6
1.3. Esquema de la tesis 6

2. Estado del arte **9**

2.1. Perforación de suelos mediante desplazamiento a rotación 9
2.2. Funcionamiento de las máquinas perforadoras 10
2.2.1. Componentes principales de una máquina de perforación 11
2.2.2. Sistema hidráulico 12
2.2.3. Transmisión de la fuerza a la herramienta de perforación 13
2.3. Registro de parámetros de perforación 15
2.3.1. Normas aplicables al registro de parámetros 15
2.3.2. Parámetros de perforación 15
2.3.3. Sistemas de medición 16
2.3.4. Factores que influyen en los resultados del registro de parámetros de perforación 18
2.3.5. Obtención de los parámetros de perforación 18
2.3.6. Parámetros compuestos 20
2.4. Ensayos de penetración 22
2.4.1. Penetrómetro dinámico 22
2.4.2. Penetrómetro estático 27
2.5. Conceptos de *Machine Learning* 34
2.5.1. Aprendizaje supervisado y no supervisado 34
2.5.2. Clasificación y regresión 35
2.5.3. Entrenamiento y evaluación 35
2.5.4. Sesgo y varianza. Sobreajuste y subajuste 35
2.5.5. Parámetros e hiperparámetros 36
2.6. Algoritmos de ML 40
2.6.1. Modelos bayesianos 40

2.6.2. Modelos lineales ... 41
2.6.3. Modelos de árboles de decisión ... 44
2.6.4. Modelos ensamblados ... 46
2.6.5. Modelos de vecinos ... 47
2.6.6. Modelos de máquinas de vectores de soporte ... 48
2.6.7. Redes neuronales ... 51
2.7. Preprocesamiento ... 53
2.7.1. Binarización ... 53
2.7.2. Análisis de componentes independientes ... 54
2.7.3. Análisis de componentes principales ... 54
2.7.4. Aglomeración de características ... 55
2.7.5. Escalar las características ... 56
2.7.6. Normalización ... 57
2.7.7. Método Nyström ... 57
2.7.8. Muestreo RBF ... 57
2.7.9. Características polinómicas ... 58
2.7.10. Codificación one-hot ... 58
2.7.11. Selectores ... 58
2.7.12. AutoML ... 59
2.8. Revisión bibliográfica ... 62
2.8.1. Revisiones sobre el uso de ML en el campo de la geotecnia ... 63
2.8.2. Aprendizaje automático con datos provenientes del registro de perforación ... 69
3. Metodología **76**
3.1. Obras estudiadas ... 76
3.1.1. Obra 1 ... 77
3.1.2. Obra 2 ... 81
3.1.3. Obra 3 ... 85
3.1.4. Obra 4 ... 88
3.1.5. Obra 5 ... 91
3.2. Herramientas utilizadas ... 112
3.3. Análisis inicial de los datos ... 116
3.4. Formato de archivo ... 118
3.5. Anomalías en los penetrómetros ... 119
3.6. Anomalías en los parámetros de perforación ... 122
3.7. Etiquetado de las muestras ... 126
3.8. Constitución del conjunto de datos ... 127
3.8.1. Método de variabilidad ... 129
3.8.2. Método continuo ... 131
3.9. Ingeniería de características ... 134
3.10. Elección de *pipelines* ... 137
3.11. Métricas de evaluación ... 142

3.12. Explicabilidad 151
3.13. Procedimiento aplicado 154
4. Resultados y discusión 159
4.1. Introducción 159
4.2. Detección de anomalías 165
4.3. Construcción del conjunto de datos en función de la distancia 168
4.4. Resultados obtenidos con TPOT 170
4.4.1. Explicabilidad de los resultados 179
4.4.2. Efecto de la distancia 182
4.4.3. Efecto del tamaño del conjunto de datos 184
4.5. Resultados con *pipeline* simplificada 185
4.6. Comparación con los parámetros compuestos 188
4.7. Predicción de datos ajenos al entrenamiento 189
4.8. Predicción de un conjunto de obras 192
5. Conclusiones 195
6. Publicaciones derivadas de esta tesis 200
7. Futuras líneas de investigación 201
Bibliografía 203

Lista de Figuras

Figura 1.1: Sección transversal del terreno con posible interpretación de las prospecciones P1 y P2 1

Figura 1.2: Sección transversal del terreno con otra posible interpretación de las prospecciones P1 y P2 2

Figura 1.3: Registro de parámetros y golpeos. Se muestra el golpeo calculado (línea continua azul) y el golpeo del penetrómetro más cercano (línea discontinua roja) 5

Figura 1.4: Esquema de la tesis. 8

Figura 2.1: Barrenas de pilotes de desplazamiento a rotación (Gwizdała & Więcławski, 2020) 9

Figura 2.2: Esquema simplificado de fuerzas en la barrena durante la perforación (Krasinski, 2015) 10

Figura 2.3: Esquema de funcionamiento de la maquinaria de perforación. 14

Figura 2.4: Tabla 1 de la *UNE-EN ISO 22476-2* 23

Figura 2.5: Máquina GRIZZLY capaz de ejecutar penetrómetros tipo DPSH 25

Figura 2.6: Ejemplo de resultados obtenidos de la ejecución de un penetrómetro dinámico 26

Figura 2.7: Posiciones de los filtros de presión de poro *UNE-EN ISO 22476-1* (AEN/CTN 103 Geotecnia, 2012) 29

Figura 2.8: Corrección de la resistencia del cono y del rozamiento por fuste debido al efecto del área final desigual *UNE-EN ISO 22476-1* (AEN/CTN 103 Geotecnia, 2012) 31

Figura 2.9: Máquina de CPT de ConeTec 33

Figura 2.10: Representación de valores obtenidos durante la ejecución de un CPT 33

Figura 2.11: Ejemplo de sobreajuste (a), subajuste (b) y buen ajuste (c) 36

Figura 2.12: Ejemplo de optimización bayesiana con dos hiperparámetros 39

Figura 2.13: Aproximación gaussiana de la distribución de dos clases para una característica 41

Figura 2.14**:** Ejemplo de regresión lineal 42

Figura 2.15: Sigmoide logística 43

Figura 2.16: Ejemplo de árbol de decisión limitado a 4 nodos hoja 44

Figura 2.17: Funcionamiento del algoritmo KNN 48

Figura 2.18: Posibles hiperplanos que separan dos clases 49

Figura 2.19: Vectores de soporte 50

Figura 2.20: Uso de kernel radial. A la izquierda se muestran dos clases que no serían separables en el espacio bidimensional. A la derecha estas dos clases son fácilmente separables en el espacio tridimensional tras el uso de un kernel radial 51

Figura 2.21: Esquema de red neuronal con capa de entrada con P variables, una única capa oculta con K neuronas y capa de salida con M variables a predecir 52

Figura 2.22: Ejemplo de funcionamiento de PCA 55

Figura 2.23: Dendrograma de aglomeración de las capitales europeas en función de la distancia entre ellas. Dependiendo de que distancia observemos encontraremos un distinto números de agrupaciones 56

Figura 2.24: Ejemplo de *pipeline* utilizado en TPOT (Olson et al., 2016) 60

Figura 2.25: Esquema de GP 62

Figura 2.26: Número de artículos publicados sobre IA en diferentes ámbitos de la ingeniería geotécnica (Baghbani et al., 2022) 63

Figura 3.1: Esquema del procedimiento de ejecución de una inclusión rígida (imagen cortesía de Menard) 76

Figura 3.2: Distribución de inclusiones y penetrómetros para la Obra 1 78

Figura 3.3: Golpeos en función de la cota para los penetrómetros de la Obra 1 80

Figura 3.4: Distribución de inclusiones y penetrómetros para la Obra 2 82

Figura 3.5: Golpeos en función de la cota para los penetrómetros de la Obra 2 84

Figura 3.6: Distribución de inclusiones y penetrómetros para la Obra 3 86

Figura 3.7: Golpeos en función de la cota para los penetrómetros de la Obra 3 88

Figura 3.8: Distribución de inclusiones y penetrómetros para la Obra 4 89

Figura 3.9: Golpeos en función de la cota para los penetrómetros de la Obra 4 91

Figura 3.10: Distribución de inclusiones y penetrómetros para la Obra 5 92

Figura 3.11**:** Resistencia por punta (izquierda) y por fuste (derecha) en función de la cota para los penetrómetros de la Obra 5 112

Figura 3.12**:** Pestaña inicial de Anaconda Navigator (superior). Interfaz de un Jupyter Notebook (inferior) 114

Figura 3.13**:** Representación de los datos para la Obra 5 117

Figura 3.14: Presencia de anomalías en un penetrómetro DPSH perteneciente en la Obra 1..... 121

Figura 3.15**:** Penetrómetro original y perfil suavizado mediante media móvil 122

Figura 3.16: Aislamiento mediante divisiones aleatorias 124

Figura 3.17: Gráfico por pares con características de la Obra 1 125

Figura 3.18: Detección de anomalías en la Obra 1..... 126

Figura 3.19: Interpolación lineal para obtener valores con cotas diferentes a las originales 127

Figura 3.20: Perfil geotécnico. Se observa cómo cambia el terreno según aumenta la distancia a los penetrómetros 128

Figura 3.21: Distancia entre las inclusiones y su prospección más cercana para la Obra 1. 130

Figura 3.22: Representación de la distancia máxima según el método de variabilidad para la Obra 1 131

Figura 3.23: Decrecimiento de la información individual li en función de la distancia di a la prospección y de distintos valores de α 133

Figura 3.24: Información L a lo largo de la Obra 1 para $\alpha = 0.2$ y $Lmin = 10$ 134

Figura 3.25: Ejemplo de matriz de correlación..... 136

Figura 3.26: Distribución de los diferentes conjuntos de la docena del datasaurio 143

Figura 3.27: Representación mediante diagrama de caja de la distribución real de los golpeos de los penetrómetros de la Obra 1 junto con la distribución de los golpeos predichos para uno de los algoritmos analizados..... 145

Figura 3.28: Diagrama de Taylor para los pipelines de TPOT de regresión con método de variabilidad para la Obra 1 146

Figura 3.29: Error residual en la Obra 1 para uno de los algoritmos analizados 147

Figura 3.30: Error residual en función de la profundidad para la Obra 1 según uno de los algoritmos analizados 147

Figura 3.31: Matriz de costos de alineación 148

Figura 3.32: Conversión utilizando DTW 149

Figura 3.33: Perfil original, perfil sobrestimado en 10 puntos y perfil desplazado 150

Figura 3.34: Gráfico de expectativa individual condicionada 152

Figura 3.35. Gráfico de dependencia parcial 153

Figura 3.36: Diagrama de enjambre 154

Figura 3.37: Esquema del procedimiento aplicado 157

Figura 4.1: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 1 para ETR con características originales, medias y vector de correlación. 160

Figura 4.2: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 2 para ETR con características originales, medias y vector de correlación. 161

Figura 4.3: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 3 para ETR con características originales, medias y vector de correlación. 162

Figura 4.4: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 4 para ETR con características originales, medias y vector de correlación. 163

Figura 4.5: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 5 para ETR con características originales, medias y vector de correlación. 164

Figura 4.6: Gráfico por pares con características de la Obra 1 165

Figura 4.7: Detección de anomalías para la Obra 1 166

Figura 4.8: Gráfico por pares de la Obra 4 tras la eliminación de anomalías 167

Figura 4.9**:** Representación de la distancia máxima según el método de variabilidad para la Obra 4 169

Figura 4.10: Información L a lo largo de la Obra 4 para $\alpha = 0.2$ y $Lmin = 10$ 170

Figura 4.11: Número de veces que se escoge un determinado algoritmo en TPOT, para enfoque de clasificación en la Obra 1 171

Figura 4.12: Número de veces que se escoge un determinado algoritmo en TPOT, para enfoque de regresión en la Obra 1 172

Figura 4.13: Evolución de las métricas del *pipeline* de TPOT con regresión, método continuo y sobremuestreo en función de las características consideradas para la Obra 1 174

Figura 4.14: Representación mediante diagrama de caja de la distribución real de los golpeos de los penetrómetros de la Obra 1 junto con la distribución de los golpeos predichos para regresión, método continuo y sin sobremuestreo 175

Figura 4.15: Representación mediante diagrama de caja de la distribución real de los golpeos de los penetrómetros de la Obra 1 junto con la distribución de los golpeos predichos para regresión, método continuo y sin sobremuestreo. Se añaden como características el vector y la media 175

Figura 4.16: Error residual para regresión y método continuo con las características originales y para la Obra 1 176

Figura 4.17: Error residual para regresión y método continuo con media y vector de correlación para la Obra 1 176

Figura 4.18: Error residual para regresión y método continuo con profundidad, media, vector de correlación y sobremuestreo 177

Figura 4.19: Error residual en función de la profundidad para regresión y método continuo para la Obra 1 178

Figura 4.20: Error residual en función de la profundidad para regresión y método continuo con media y vector de correlación para la Obra 1 178

Figura 4.21: gráfico ICE para regresión y método continuo con profundidad para la Obra 1 179

Figura 4.22: diagrama de enjambre para regresión y método continuo con profundidad para la Obra 1 179

Figura 4.23: gráfico ICE para regresión y método continuo con media y vector de correlación para la Obra 1 181

Figura 4.24:diagrama de enjambre para regresión y método continuo con media y vector de correlación para la Obra 1 182

Figura 4.25: diagrama de enjambre para regresión y método continuo media y vector de correlación para la Obra 3 182

Figura 4.26. Evolución del coeficiente de correlación, CUSUM y número de muestras con la distancia para los algoritmos de regresión, método continuo y sin sobremuestreo para la Obra 1 183

Figura 4.27. Evolución del coeficiente de correlación, CUSUM y número de muestras con la distancia para los algoritmos de clasificación, método continuo y sin sobremuestreo para la Obra 4 184

Figura 4.28: Evolución del coeficiente de correlación y CUSUM con el tamaño de la muestra para los algoritmos de regresión, método continuo y sin sobremuestreo 185

Figura 4.29: Diagramas de Taylor para los conjuntos con método continuo y considerando regresión para la Obra 1. Arriba, *pipelines* de TPOT. Abajo ETR 187

Figura 4.30: Diagrama de Taylor de la Obra 2 con algoritmo entrenado en la Obra 1 para los conjuntos de regresión y con método continuo..... 190

Figura 4.31: Diagrama de Taylor de la Obra 3 con algoritmo entrenado en la Obra 1 para los conjuntos de regresión y con método continuo..... 191

Figura 4.32: Diagrama de Taylor de la Obra 3 con algoritmo entrenado en las obras 1 y 2 para los conjuntos de regresión y con método continuo 191

Figura 4.33: Diagrama de Taylor del conjunto de las obras 1, 2 y 3 para los conjuntos de regresión y con método continuo 192

Figura 4.34: diagrama de enjambre para regresión y método continuo con profundidad para el conjunto de las obras 1, 2 y 3 193

Figura 4.35: gráfico ICE para regresión y método continuo con profundidad. Conjunto de las Obras 1, 2 y 3 194

Lista de Tablas

Tabla 2.1: Parámetros compuestos 21

Tabla 2.2: Ejemplo de codificación *one-hot* 58

Tabla 2.3: Resumen de los artículos que utilizan ML con datos de MWD 71

Tabla 3.1**:** características de la maquinaria utilizada 77

Tabla 3.2: Ejemplo de dos filas de la matriz de muestras con los valores correspondientes de la variable objetivo 115

Tabla 3.3: Vector de correlación resultante. 136

Tabla 3.4: Diferentes conjuntos de características utilizados. 137

Tabla 3.5: Preprocesos considerados con TPOT 138

Tabla 3.6: Algoritmos de regresión considerados en TPOT 138

Tabla 3.7: Selectores considerados con TPOT para regresión 139

Tabla 3.8: Algoritmos de clasificación considerados en TPOT 140

Tabla 3.9: Selectores considerados con TPOT para clasificación 141

Tabla 3.10: Métricas de los diferentes conjuntos de la docena del datasaurio 144

Tabla 3.11: Diferentes combinaciones analizadas 156

Tabla 4.1: Número de muestras según los métodos de variabilidad y continuo para las diferentes obras 168

Tabla 4.2: Métricas para los pipelines obtenidos con TPOT de regresión y con el método continuo para la Obra 1 173

Tabla 4.3: Resultados obtenidos en la comparativa entre los parámetros compuestos, y ETR con método continuo, media y vector de correlación 189

Abreviaturas, Acrónimos y Símbolos

Abreviaturas y acrónimos

UPM	Universidad Politécnica de Madrid
1D-CNN	1D Convulational Neural Network. Red neuronal convolucional unidimensional
AAPE	Average Absolute Percentage Error. Error porcentual absoluto promedio
AGS	Association of Geotechnical & Geoenvironmental Specialists. Asociación de Especialistas Geotécnicos y Geoambientales
ANN	Artificial Neural Networks. Redes neuronales artificiales
ANN-ELM	Artificial Neural Network - Extreme Learning Machines. Red neuronal de aprendizaje extremo
ANN-FN	Artificial Neural Network-Funtional Network. Red neuronal artificial
ANN-GR	Artificial Neural Network-General Regression. Red neuronal artificial de regresión general
ANN-PSO	Artificial Neural Network-Particle Swarm Optimization. Red neuronal con optimización de enjambre de partículas
ANN-SaDE	Artificial Neural Network-Self-adaptive Differential Evolution. Red neuronal artificial con Evolución Diferencial Auto-adaptativa
ANN-Wavenet	Artificial Neural Network-Wavenet. Red neuronal artificial con funciones de activación wavelet
ANFIS	Adaptive Neuro-Fuzzy Inference System. Sistema de inferencia neuro-difuso adaptativo
API	Application Programming Interface. Interfaz de programación de aplicaciones
AutoML	Auto Machine Learning. Aprendizaje automático automatizado
COA	Cuckoo Optimization Algorithm. Algoritmo de optimización cuco

CNN	Convulational Neural Network. Red neuronal convulacional
CPT	Cone Penetration Test. Ensayo de penetración con cono
DL	Deep learning. Aprendizaje profundo
DPH	Dynamic Probing Heavy. Ensayo de penetración dinámica pesada
DPL	Dynamic Probing Light. Ensayo de penetración dinámica ligera
DPM	Dynamic Probing Medium. Ensayo de penetración dinámica mediana
DPSH	Dynamic Probing Super Heavy. Ensayo de penetración dinámica superpesada
DT	Decission Trees. Árboles de decisión
DTW	Dynamic Time Warping. Deformación dinámica temporal
EPB	Earth Pressure Balance. Balance de presión de tierras
FA	Factorial Analysis. Analisis factorial
FNN	Feedforward Neural Network. Red neuronal con alimentación hacia delante
GAN	Generative Adversarial Network. Redes generativas antagónicas
GEP	Gene Expression Programming. Programación de expresión genética
GNB	Gaussian Naïve-Bayes. Naïve-Bayes gausiano
GP	Genetic Programing. Programación genética
IA	Inteligencia Artificial
ICA	Independent Component Analysis. Análisis de componentes independientes
ICE	Individual Conditional Expectation. Expectativa individual condicionada
KNN	K-Nearest Neighbors. K-vecinos más cercanos
LiR	Linear Regression. Regresión linear
L-IRL	Linguistic Iterative Rule Learning. Aprendizaje de reglas iterativo linguistico

LoR	Logistic Regression. Regresión logística
LSTM	Long Short-Term Memory. Red neuronal de memoria larga a corto plazo
MAE	Mean Average Error. Error promedio
MARS	Multivariate Adaptive Regression Spline. Regresión spline adaptativa multivariante
ML	Machine Learning. Aprendizaje automático
MWD	Measurement While Drilling. Registro de parámetros durante la perforación
NB	Naïve-Bayes
PCA	Principal Component Analysis. Análisis de componentes principales
PDF	Portable Document Format. Formato de documento portátil
PDP	Partial Dependency Plot. Gráfico de dependencia parcial
RBF	Radial Basis Function. Función de base radial
RF	Random Forest. Bosques aleatorios
RFC	Random Forest Classifier. Bosques aleatorios de clasificación
RFE	Recursive Feature Elimination. Eliminación recursiva de características
RMR	Rock Mass Rating
RMSE	Root of Mean Squared Error. Raíz del error cuadrático medio
RNN	Recursive Neural Network. Red neuronal recursiva
ROP	Rate of Penetration. Velocidad de penetración
RQD	Rock Quality Designation
S-IRL	Scatter Iterative Rule Learning. Aprendizaje de reglas iterativo disperso
SVM	Support Vector Machines. Máquinas de vectores de soporte
SVM-LS	Least Square Support Vector Machines. Máquinas de vectores de soporte de mínimos cuadrados

SVR	Support Vector Regression. Vectores de soporte de regresión
TBM	Tunnel Boring Machine. Tuneladora
XGBoost	eXtreme Gradient Boosting. Aumento extremo de gradiente
XML	eXtensible Markup Language. Lenguaje de Marcado Extensible

Símbolos

Símbolo	Nombre	Unidad
α	Coeficiente de eficiencia de la perforación	-
β	Coeficiente de eficiencia del trabajo de empuje	-
γ	Coeficiente de eficiencia del trabajo de par de rotación	-
ω	Coeficientes del modelo de regresión lineal	-
a	Longitud de penetración medida	m
a	Relación de área neta	-
A_c	Área proyectada de la sección transversal del cono	mm2
A_n	Área de la sección transversal de la célula de carga o el vástago	mm2
A_s	Área de la superficie del manguito de fricción	mm2
A_{sb}	Área de la sección transversal de la parte inferior del manguito de fricción	mm2
A_{st}	Área de la sección transversal de la parte superior del manguito de fricción	mm2
C_R	Par de rotación medido en cabeza	kN·m
D_s	Diámetro máximo del tornillo de perforación	m
D_{s0}	Diámetro mínimo del tornillo de perforación	m
E_s	Energía específica	
E	Energía normalizada	
f	Frecuencia del martillo	Hz
F_s	Fuerza del manguito de fricción medida axialmente	kN

f_s	Rozamiento por fuste medido	MPa
f_t	Rozamiento por fuste corregido	MPa
F_0	Peso total que ejerce la herramienta de perforación cuando no se ejerce ningún empuje adicional	kN
F_E	Fuerza de empuje aplicada sobre la herramienta	kN
$F_{máx.}$	Máxima fuerza de empuje	kN
h_s	Longitud del taladro de perforación	m
$h_{s1}, h_{s2}, h_{s3}, h_{s4}$	Longitudes de las diferentes partes del taladro de perforación	m
k_c	Relación entre la presión hidráulica del motor de par, p_{CR}, y el par sobre la herramienta de perforación	
k_E	Relación entre la presión hidráulica del motor del cabestrante, p, y el empuje adicional sobre la herramienta de perforación	-
M_T	Par de rotación total aplicado sobre la barrena de perforación	kN·m
M_{Tb}	Componente del par de rotación correspondiente a la punta de la barrena	kN·m
M_{Ts}	Componente del par de rotación correspondiente al fuste de la barrena	kN·m
n	Número de varillas	-
p	Presión hidráulica medida en el motor del cabestrante	MPa
p_{CR}	Presión hidráulica medida en la mesa de rotación	MPa
p_E	Presión calculada de empuje neto sobre la herramienta de perforación	MPa
p_M	Presión de golpeo medida	MPa
$p_{máx.}$	Presión de empuje máxima	MPa
p_O	Empuje bruto aplicado a la herramienta de perforación	MPa
P_R	Resistencia a la penetración	s/0,2 m
p_X	Presión de retención medida	MPa

Q_c	Fuerza medida axialmente en el cono	kN
q_c	Resistencia del cono medida	MPa
q_t	Resistencia del cono corregida	MPa
Q_{Tv}	Fuerza vertical aplicada en el taladro	kN
R_{SR}	Resistencia suelo-roca	MPa/(m/s)
s	Espaciamiento de la hélice del taladro de perforación	m
S_d	Índice de Somerton calculado	-
s_O	Área de perforación medida (= superficie de la herramienta de perforación)	m2
t	Tiempo medido	s
t_{Tb}	Fricción unitaria en la punta de la barrena	kPa
t_{Ts}	Fricción unitaria a lo largo del fuste de la barrena	kPa
u_1, u_2, u_3	Presiones de poro medidas en las posiciones 1, 2 y 3 respectivamente	MPa
v_A	Velocidad de penetración	m/h
v_R	Velocidad de rotación de la corona de perforación medida	r/min
W_H	Peso de la corona giratoria	kN
W_R	Peso de la barra de perforación	kN
X	Matriz de muestras	-
y	Vector de variable objetivo	-

Glosario

Algoritmo de aprendizaje automático: conjunto de instrucciones para construir un programa en función de los datos de partida.

Anomalía: valor registrado que no está asociado a la realidad del problema.

Aprendizaje automático: rama de la inteligencia artificial que busca extraer información de grandes conjuntos de datos.

Aprendizaje supervisado: aprendizaje automático en el que se conoce la respuesta esperada.

Aprendizaje no supervisado: aprendizaje automático en el que no existe una «respuesta correcta» si no que se buscan similitudes en los datos de cara a poder agruparlos en función de sus características.

Barrena: herramienta de perforación capaz de atravesar diversas capas del terreno.

Características: variables de entrada de los algoritmos.

Clasificación: enfoque de un problema de aprendizaje automático en el que las muestras se colocan en diferentes categorías.

Conjunto de datos: conjunto de muestras en este caso obtenidas del proceso de perforación.

Empuje: fuerza aplicada sobre la herramienta de perforación.

Entrenamiento: proceso por el que se eligen los parámetros internos de un algoritmo de aprendizaje automático, obteniendo un modelo.

Evaluación: proceso en el que se comprueban los resultados de un modelo con datos ajenos al entrenamiento.

Explicabilidad: grado en que las personas podemos comprender la causa de una decisión tomada por un algoritmo.

Hiperparámetros: parámetros que limitan las opciones del modelo resultante de entrenar un algoritmo.

Inclusiones rígidas: tratamiento del terreno en el que se incorporan al terreno unas columnas al terreno con una rigidez mucho más alta que este.

Inteligencia artificial: disciplina científica, que se ocupa de crear programas informáticos que ejecutan operaciones comparables a las que realiza la mente humana, como el aprendizaje o el razonamiento lógico

Machine learning: ver «aprendizaje automático».

Modelo: conjunto de fórmulas matemáticas que son el resultado del entrenamiento de un algoritmo de aprendizaje automático.

Parámetros compuestos: fórmulas que combinan varios parámetros de perforación para obtener valores con un mayor significado físico.

Par de rotación: fuerza aplicada a un determinado brazo de palanca que genera un giro.

Penetrómetro: investigación geotécnica que mide la resistencia del suelo a la introducción de una punta cónica.

Perfil penetrométrico: perfil en función de la profundidad obtenido en la ejecución de un penetrómetro.

Perforadora: máquina con potencia suficiente con lay capacidad de introducir una barrena en el terreno.

Pipeline: conjunto de pasos seguidos a la hora de aplicar un algoritmo de aprendizaje automático.

Preprocesamiento: fase de preparación de los datos de entrada para que los algoritmos tengan un comportamiento óptimo.

Prospección: ensayo geotécnico basado en introducir algún tipo de elemento en el suelo, como los penetrómetros y los sondeos.

Registro de parámetros: recogida de datos relevantes al proceso de perforación sin afectar al mismo.

Regresión: enfoque de un problema de aprendizaje automático en el que se busca predecir los valores de una variable continua.

Sesgo: error introducido al intentar aproximar un problema real mediante un modelo mucho más simple.

Sobreajuste: fenómeno por el que un modelo se ajusta en exceso a los datos de entrenamiento.

Sobremuestreo: proceso porque el que se crean muestran sintéticas cuando existen clases muy poco representadas.

Subajuste: fenómeno en el que un modelo demasiado simple no es capaz de aproximar la forma real del problema.

Varianza: magnitud en la que cambiaría nuestra predicción de haberla realizado en un conjunto de entrenamiento diferente.

Velocidad de penetración: distancia que avanza la barrena en el terreno para una determinada unidad de tiempo.

Velocidad de rotación: número de giros que realiza la barrena en una determinada unidad de tiempo.

1. Introducción y motivación

Imaginemos que tenemos una parcela en la que queremos construir un edificio. Este necesita una cimentación que transmita las cargas al suelo. Para conocer las características de dicho suelo necesitamos hacer prospecciones que nos proporcionen información sobre éste (su naturaleza y resistencia). Como nuestro presupuesto es limitado, realizamos tan solo un puñado de estas prospecciones, cumpliendo los requisitos de la normativa vigente.

Tras realizar nuestras prospecciones ya tenemos información sobre el terreno en los puntos donde hemos realizados los ensayos. Sin embargo ¿cuáles son las características del terreno en el resto de la parcela?

No lo sabemos.

La realidad es que entre puntos ensayados desconocemos las características del terreno. Imaginemos que para nuestra parcela hemos realizados las prospecciones P1 y P2 y hemos detectado las unidades geotécnicas 1. Blanda, 2. Media y 3. Dura. Una posible interpretación de estos penetrómetros se muestra en la Figura 1.1.

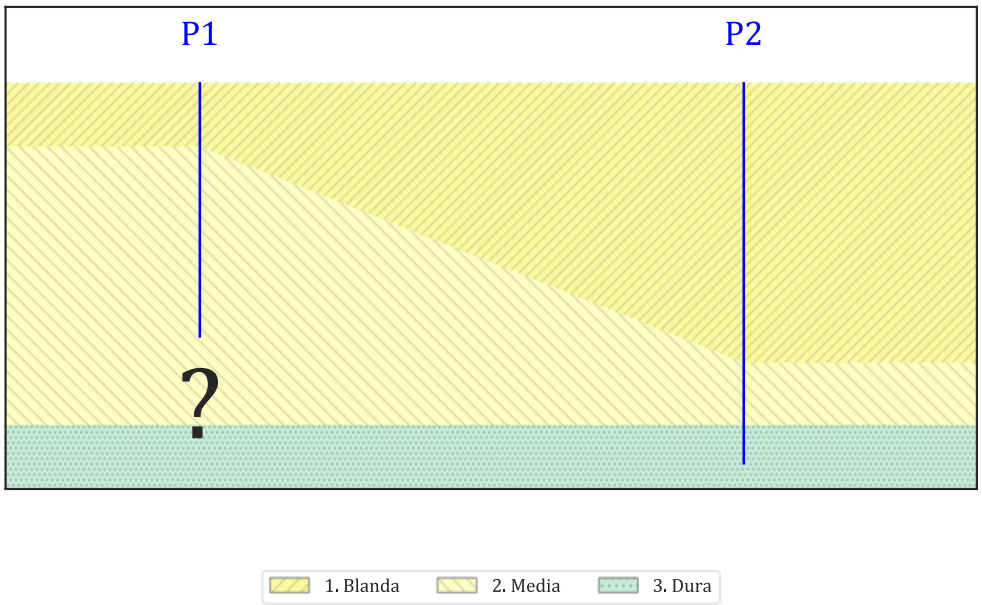


Figura 1.1: Sección transversal del terreno con posible interpretación de las prospecciones P1 y P2

Pero, con la misma información, el terreno puede perfectamente tener la distribución mostrada en la Figura 1.2.

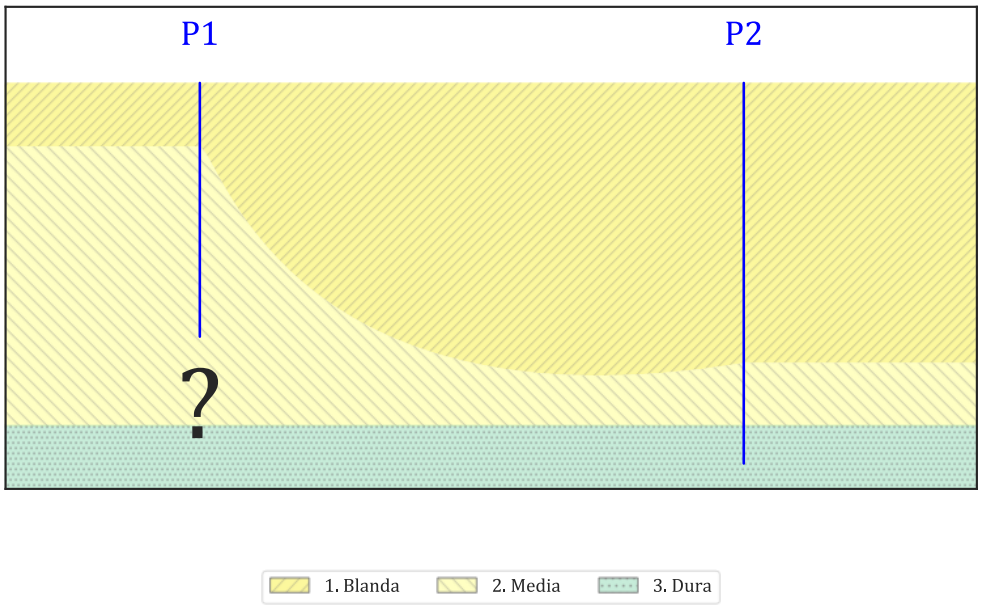


Figura 1.2: Sección transversal del terreno con otra posible interpretación de las prospecciones P1 y P2

El terreno en superficie es blando, por lo que la cimentación de nuestro edificio debe alcanzar, al menos, la unidad Media. Pero la longitud de esa cimentación cambia mucho entre lo mostrado en las figuras 1.1 y 1.2.

¿Cómo podemos asegurar entonces que la cimentación alcanza el terreno adecuado? El objetivo de esta tesis es, precisamente, obtener una mayor seguridad sobre el suelo que atravesamos durante la perforación de tratamientos del terreno o cimentaciones profundas.

Para ello, utilizamos los parámetros de perforación y aprendizaje automático. El aprendizaje automático es una rama de la inteligencia artificial que busca obtener información de grandes conjuntos de datos y es muy útil cuando las relaciones entre las variables que afectan al problema no son bien conocidas, como en el caso que nos ocupa.

En este caso, buscamos que el aprendizaje automático nos genere los valores correspondientes a los ensayos geotécnicos, convirtiendo cada perforación en su propio ensayo. Esto aumenta enormemente la cantidad de información disponible sobre el terreno ya que los ensayos que se realizan son pocos y dispersos, mientras

que los elementos de cimentación, pilotes o tratamientos del terreno, tenderán a cubrir toda la huella de la edificación. En obras de cierta envergadura podríamos tener información, una vez acabados los trabajos, **en cien veces más puntos** que con los datos originales. Un ejemplo lo encontramos en la Figura 1.3 donde se realizan 607 perforaciones frente a 7 penetrómetros.

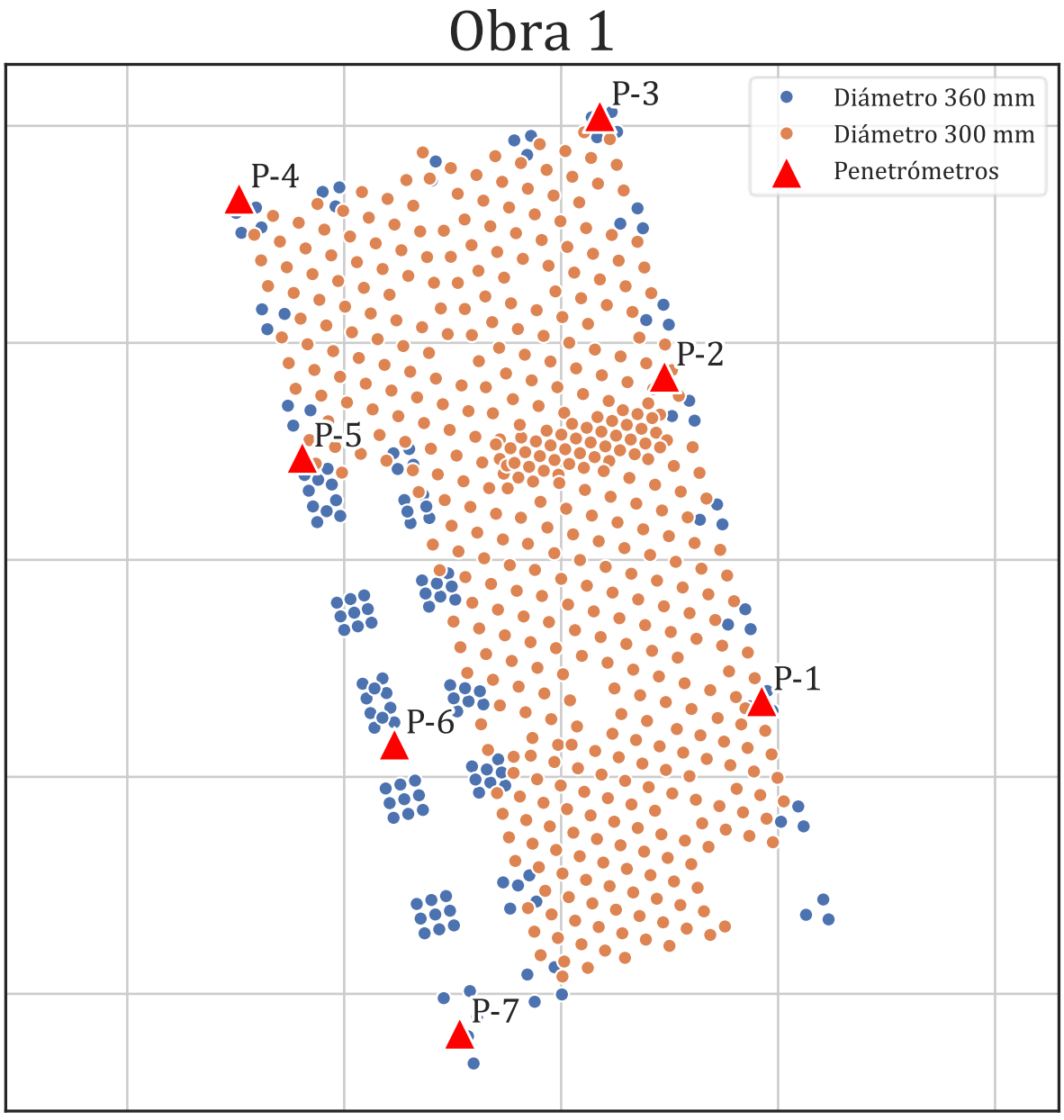


Figura 1.3: Distribución de inclusiones y penetrómetros para la Obra 1

Esto nos permite aumentar mucho la seguridad en nuestra cimentación y optimizar el diseño durante la ejecución.

1.1. Alcance de la investigación

El objetivo buscado en esta tesis es reproducir el perfil penetrométrico del suelo a partir del registro de parámetros de perforación de inclusiones rígidas. Se busca obtener un valor de golpeo o de resistencia a la penetración a lo largo de la longitud de las inclusiones realizadas. De esta manera, a cada cota donde tenemos valores de par de rotación, velocidad de rotación, velocidad de penetración y empuje aplicado, obtendríamos además un valor de golpeo del penetrómetro que debe ser lo más cercano posible a los datos reales. Esto puede representarse gráficamente como se muestra en la Figura 1.4.

Para alcanzar el objetivo principal se abordan una serie de objetivos secundarios, principalmente de carácter técnico, fundamentales para alcanzar una predicción consistente:

- Evaluación y comparación de algoritmos de aprendizaje automático, para determinar cuál es el que mejor se adapta en este caso.
- Comprobación de cuáles son las mejores estrategias en el tratamiento de los datos para problemas geotécnicos.
- Llevar a cabo un proceso de ingeniería de características que permita extraer la mayor cantidad de información posible de los datos disponibles.
- Estudiar la capacidad de generalización de los modelos entre diferentes contextos geotécnicos (diferentes obras).

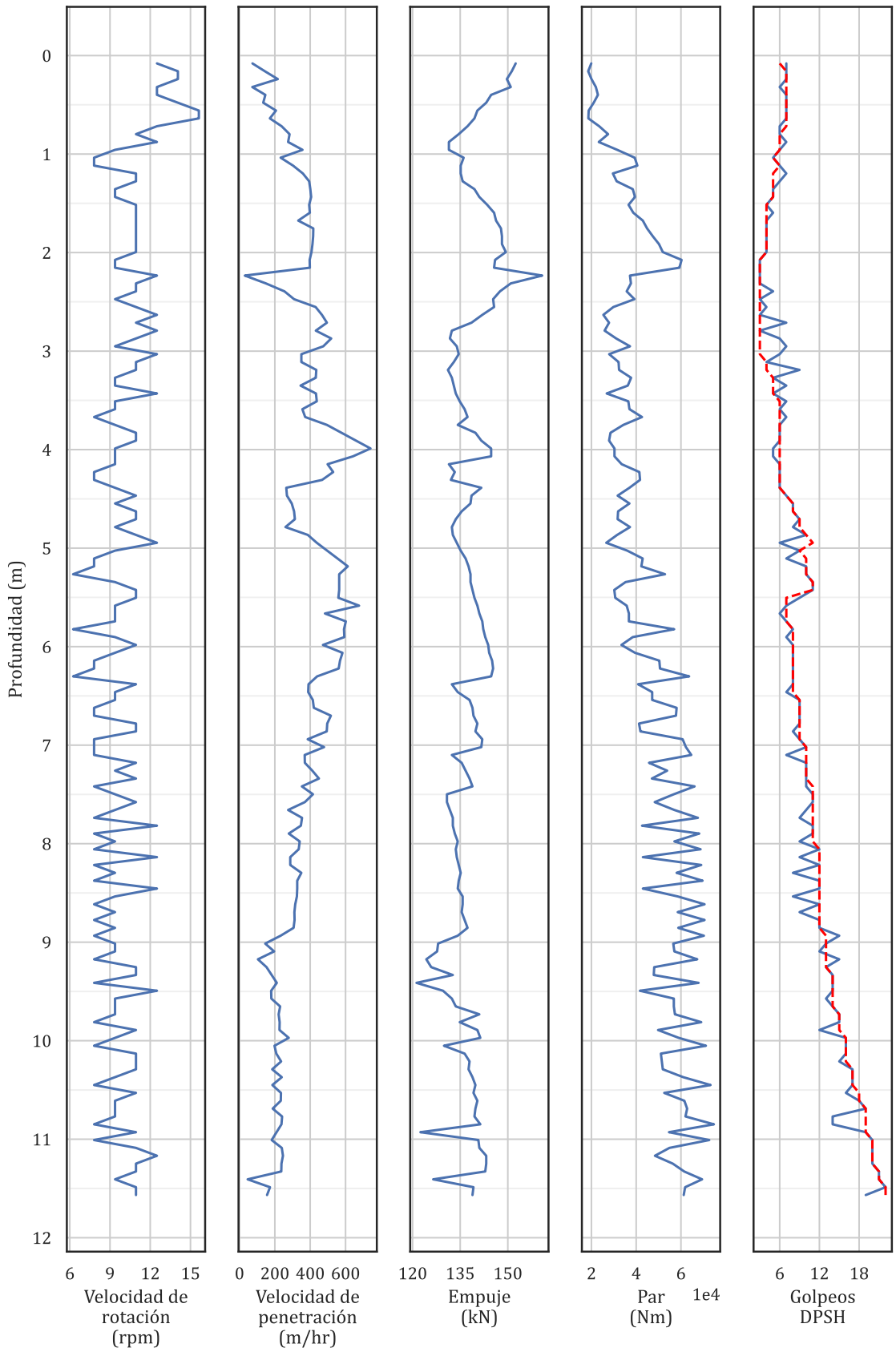


Figura 1.4: Registro de parámetros y golpeos. Se muestra el golpeo calculado (línea continua azul) y el golpeo del penetrómetro más cercano (línea discontinua roja)

1.2. Estrategia de investigación

Para alcanzar el objetivo de predecir el perfil penetrométrico del suelo se ha seguido una estrategia estructurada en varias fases. En primer lugar, se recopilaron datos reales de distintas obras, que incluían tanto registros de parámetros de perforación como resultados de ensayos penetrométricos. A partir de estos datos se realizó un proceso de depuración y detección de anomalías para asegurar su calidad y fiabilidad.

Posteriormente, se asignaron valores de los ensayos de penetración a los datos de perforación. Para ello se utilizaron métodos basados en la variabilidad (método de variabilidad) y en la cantidad de información (método continuo). Con ello se construyeron los conjuntos de datos para poder entrenar los modelos.

A continuación, se aplicó un proceso de ingeniería de características, creando nuevas variables derivadas de los datos de perforación y teniendo en cuenta la forma de la herramienta de perforación. Para comprobar cuales eran los algoritmos más apropiados de entre todos los posibles se han utilizado herramientas de AutoML (*Auto Machine Learning*).

Finalmente, se evaluó la capacidad predictiva de los modelos mediante diferentes métricas y, también, visualmente, comprobando la predicción del perfil penetrométrico tanto en los valores como en la forma. Se analizó también su comportamiento al aplicarse sobre obras no incluidas en el entrenamiento, explorando así los límites y posibilidades de generalización.

Esta estrategia ha permitido validar la utilidad del aprendizaje automático como herramienta para reducir la incertidumbre en geotecnia a partir de datos de perforación.

1.3. Esquema de la tesis

El presente documento se divide en 4 partes además de la actual:

Estado del arte: describe los métodos de perforación considerados, los ensayos penetrométricos y el registro de parámetros. Introduce conceptos claves relacionados con el *Machine Learning* y repasa la bibliografía relevante sobre la aplicación de estas técnicas en problemas geotécnicos.

Metodología: presenta las características de las obras estudiadas y los datos disponibles. Describe asimismo los métodos seguidos para el tratamiento de los datos (detección de anomalías, etiquetado de muestras, ingeniería de características), la construcción de los conjuntos utilizados para el entrenamiento y evaluación, y las técnicas de aprendizaje automático empleadas.

Resultados y discusión: Expone los principales resultados obtenidos con los diferentes enfoques analizados, incluyendo la identificación de anomalías, la evaluación del desempeño de distintos algoritmos de aprendizaje automático, el efecto de la distancia y del tamaño de muestra sobre la calidad de la predicción, y la generalización entre obras.

Discusión final: Analiza críticamente los resultados obtenidos, resalta las principales contribuciones del trabajo, y formula conclusiones en torno a la utilidad práctica y las limitaciones del método propuesto. También presenta posibles líneas futuras de investigación derivadas de esta tesis.

En la Figura 1.5 se muestran estos apartados en forma de esquema.

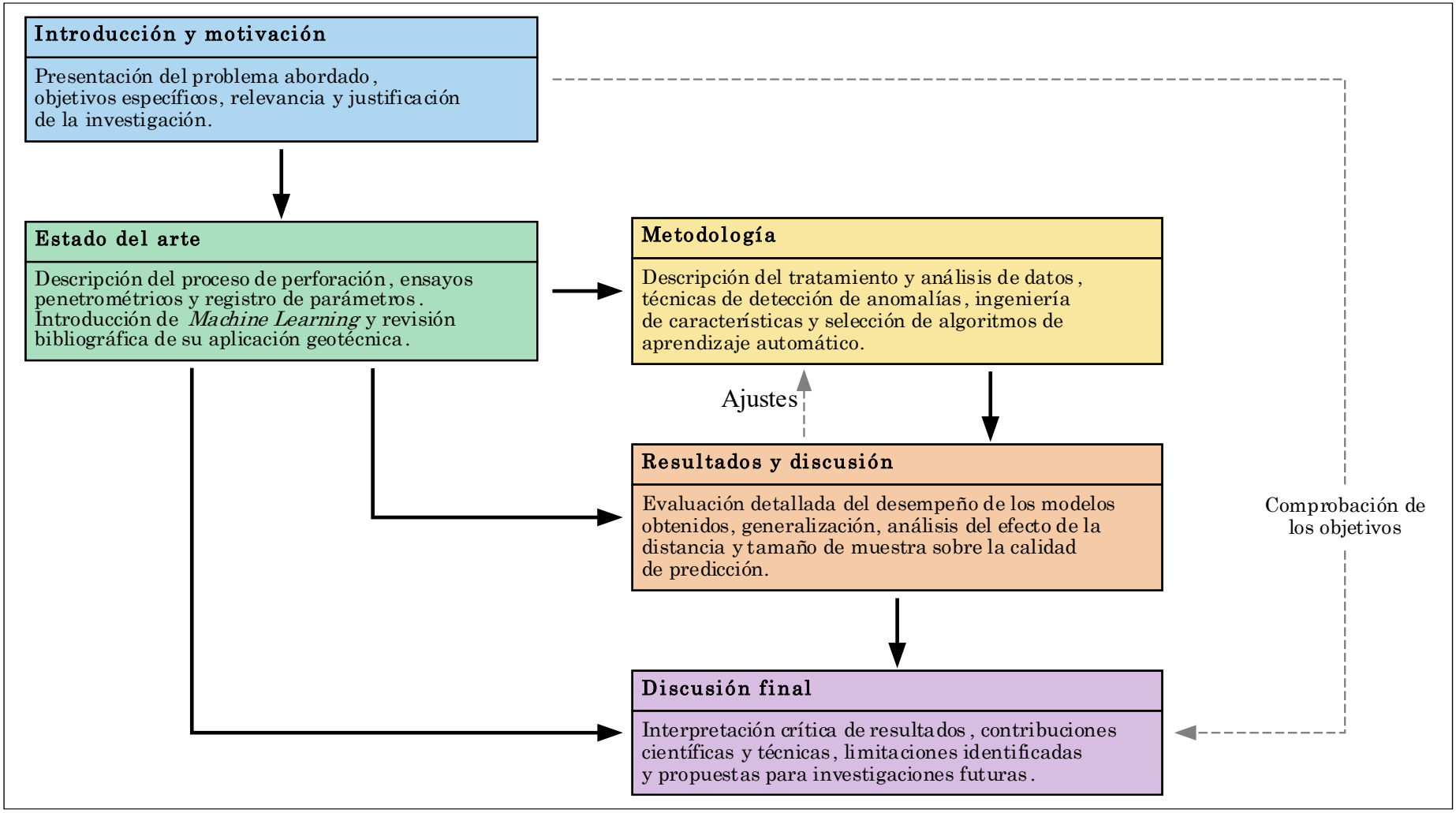


Figura 1.5: Esquema de la tesis.

2. Estado del arte

2.1. Perforación de suelos mediante desplazamiento a rotación

La perforación de suelos puede realizarse mediante diferentes métodos. En esta tesis, los datos recogidos provienen de inclusiones rígidas, perforaciones similares a un pilote realizadas, en este caso, mediante desplazamiento a rotación. Para este tipo de perforaciones podemos acudir a la norma *EN 12699:2016. Ejecución de trabajos geotécnicos especiales. Pilotes de desplazamiento* (AEN/CTN 103 Geotecnia, 2016b).

Existen diferentes denominaciones para estos pilotes en función del tipo de herramienta y de la empresa que los ejecute: Atlas, Omega, SPD, etc. (Figura 2.1); pero todos ellos funcionan de manera similar. La parte inferior de la barrena tiene una forma helicoidal yel diámetro de esta aumenta hasta llegar a una zona central donde, en la mayoría de los casos, se produce el cambio de dirección de la hélice. Gracias a esta forma, el terreno perforado es desplazado radialmente, sin extraerlo a la superficie.

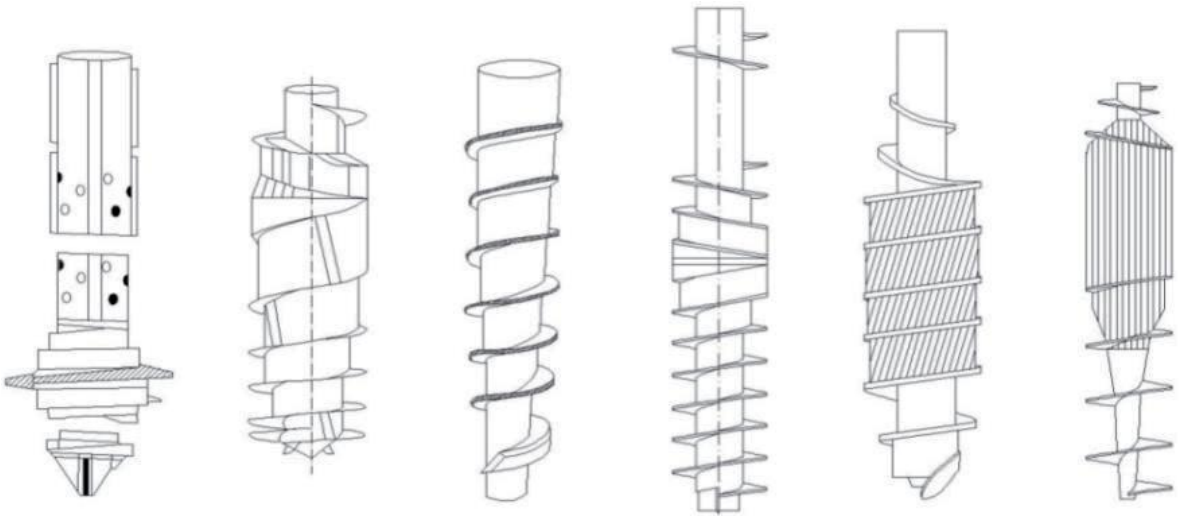


Figura 2.1: Barrenas de pilotes de desplazamiento a rotación (Gwizdała & Więcławski, 2020)

Para realizar la perforación se aplica un par de rotación y un empuje. Este proceso de introducción de la herramienta en el suelo es complejo, si bien puede analizarse de manera simplificada según se recoge en la Figura 2.2.

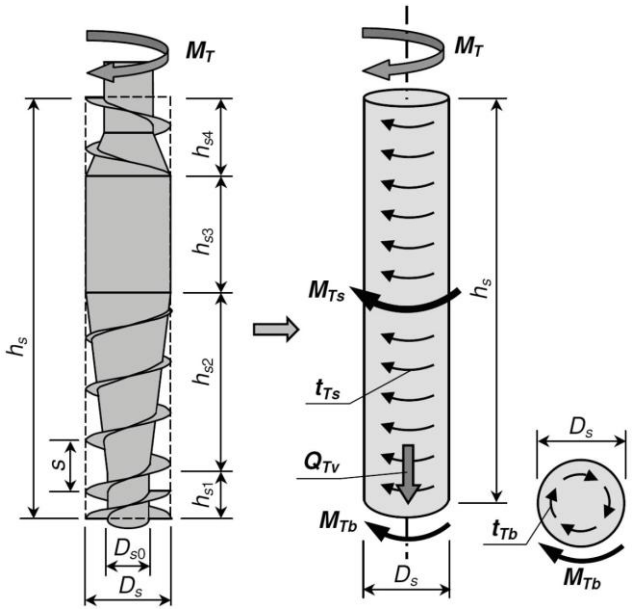


Figura 2.2: Esquema simplificado de fuerzas en la barrena durante la perforación (Krasinski, 2015)

Según este esquema, el par de rotación total aplicado M_T puede dividirse en dos componentes: M_{Ts} que se relaciona con la fricción a lo largo del fuste de la barrena t_{Ts}; y M_{Tb} relacionado con la resistencia en la punta de la barrena t_{Tb}. Esta distribución del par es distinta a otros procesos de perforación, donde este es principalmente aplicado al nivel de la punta de la herramienta de perforación.

2.2. Funcionamiento de las máquinas perforadoras

Las plataformas de perforación de pilotes o tratamientos del terreno a rotación son maquinaria de gran porte diseñadas para atravesar el terreno mediante un proceso de perforación rotatoria. Utilizan una broca que atraviesa el terreno mediante la aplicación de una fuerza de empuje y un par de rotación.

Estas máquinas cuentan con un motor diésel y, mediante sistemas hidráulicos, transmiten la potencia necesaria de manera eficiente a los diferentes componentes del útil de perforación.

2.2.1. Componentes principales de una máquina de perforación

Una máquina de perforación consta, de manera general, de los siguientes componentes:

Chasis y tren de orugas: base portante, generalmente sobre orugas, que otorga estabilidad y movilidad a la plataforma. Sobre el chasis se montan el resto de los subsistemas y aloja el motor diésel, componentes hidráulicos y los contrapesos para equilibrar el equipo.

Mástil: estructura vertical de alta resistencia que guía y sostiene el conjunto de perforación en posición alineada. El mástil soporta la mesa de rotación y el *Kelly*, manteniéndolos verticales durante la perforación. Mediante cilindros hidráulicos puede inclinarse o plegarse para el transporte. Debe resistir las fuerzas de reacción de la perforación.

Mesa de rotación: mecanismo motriz ubicado en el mástil que genera el movimiento de rotación de la herramienta de perforación. Consiste en un motor (o motores) de accionamiento habitualmente hidráulico. A la mesa de rotación se conecta la barra Kelly a la que, al girar, transmitir un par elevado para atravesar el terreno. Se desplaza por el mástil ascendiendo y descendiendo.

Barra *Kelly*: es una barra de acero hueca de sección suficiente y diseñada para transferir tanto el par como la fuerza vertical desde la mesa de rotación hasta la herramienta de perforación y en su interior conducir el hormigón a presión. El Kelly puede tener varios tramos permitiendo perforar profundidades mayores que la longitud del mástil.

Herramienta de perforación: acoplada al extremo del *Kelly*, es el implemento que permite la perforación, y en el caso de las obras de esta tesis, el desplazamiento del terreno. En el apartado anterior (2.1) ya se ha tratado la forma de esta herramienta en el caso de la perforación mediante desplazamiento a rotación. La rotación de la herramienta, combinada con peso y empuje, permite atravesar el terreno.

Cabrestantes: la máquina suele poseer al menos un cabrestante (*winche*) principal, con cables de acero generalmente, utilizado para izar o descender la maniobra de perforación (herramienta, *Kelly* y mesa de rotación). En muchas perforadoras modernas, el cabrestante principal cumple la función de empuje, aplicando fuerza descendente adicional a la maniobra de perforación.

En conjunto, estos componentes permiten posicionar la máquina, colocar la maniobra en el punto de perforación, mantener la alineación y transmitir el par y

el empuje necesarios para atravesar el terreno. Todo este sistema mecánico trabaja de forma coordinada con la hidráulica de la máquina.

2.2.2. Sistema hidráulico

El sistema hidráulico es el encargado de accionar y controlar la mayoría de las funciones de la plataforma, proporcionando la potencia y precisión requeridas. Los principales componentes hidráulicos serían los siguientes:

Motor diésel y bombas hidráulicas: un motor diésel transmite, mediante el eje del cigüeñal, la energía para accionar una o varias bombas hidráulicas. Suele haber bombas principales de caudal variable dedicadas a las funciones de gran potencia (rotación, empuje, elevación) y bombas secundarias para servicios auxiliares. Estas bombas convierten la potencia mecánica del motor en flujo de aceite a alta presión, que alimenta los actuadores de la máquina.

Motores hidráulicos rotativos: ubicados en la mesa de rotación y en los cabrestantes (y también en el sistema de translación de orugas), son motores hidráulicos (generalmente de tipo pistones axiales o radiales de alto par) que convierten la energía hidráulica de presión nuevamente en movimiento giratorio mecánico. El motor del cabezal convierte el flujo de aceite en par para girar la barra *Kelly* y la herramienta. Los motores de los *winches* permiten enrollar o desenrollar los cables para izar la carga. La velocidad y par se controlan modulando el flujo y presión de aceite que reciben.

Cilindros hidráulicos: actúan para producir movimientos lineales. En una plataforma de perforación encontramos cilindros de gran tamaño en funciones como: elevación/descenso del mástil (para colocar la máquina en posición de trabajo o transporte), inclinación del mástil, sistema de empuje (en algunos casos), etc.

Válvulas de control y circuito hidráulico: el flujo de aceite hacia cada actuador es gobernado por un conjunto de válvulas hidráulicas de control (distribuidores, electroválvulas proporcionales, etc.) que responden a las órdenes del operador en la cabina.

Unidad de control y sensores: aunque no son componentes puramente hidráulicos, las perforadoras modernas cuentan con sistemas electrónicos que monitorizan y gestionan el sistema hidráulico. Sensores de presión, caudal y posición informan a un controlador que ajusta válvulas proporcionales y bombas de caudal variable.

Tanque, filtro y radiador: el fluido hidráulico circula desde un tanque o depósito, atravesando filtros que eliminan impurezas, hacia las bombas y actuadores, retornando luego al tanque. Dada el alta potenciada manejada, el aceite tiende a calentarse, por lo que se incluyen radiadores de aceite para disipar calor y mantener la temperatura de operación.

2.2.3. Transmisión de la fuerza a la herramienta de perforación

La transmisión de la fuerza desde la fuente de potencia hasta la herramienta de perforación sigue una cadena mecánico-hidráulica: comienza en el motor diésel, pasa por las bombas hidráulicas, luego por motores hidráulicos, y finalmente llega a los elementos mecánicos (engranajes, barra *Kelly*, herramienta).

Primero, el par del motor diésel acciona la bomba hidráulica la cual recibe fluido a baja presión proveniente del tanque. Entonces la bomba suministra aceite hidráulico a alta presión a los diferentes motores. En el caso de la mesa de rotación el motor está diseñado para entregar torques elevados y girar la barra *Kelly* que a su vez transmite el par a la herramienta de perforación. Este par aplicado se suma a la fuerza de empuje, adicional al peso de la herramienta y el *K*elly, suministrada por el motor del cabrestante para atravesar el terreno.

Esta aplicación de fuerzas se maneja mediante el control hidráulico. Una mayor presión hidráulica equivale a un mayor par o empuje aplicado. Medir estas presiones permite llevar a cabo un registro de parámetros.

Este proceso de transmisión se esquematiza en la Figura 2.3.

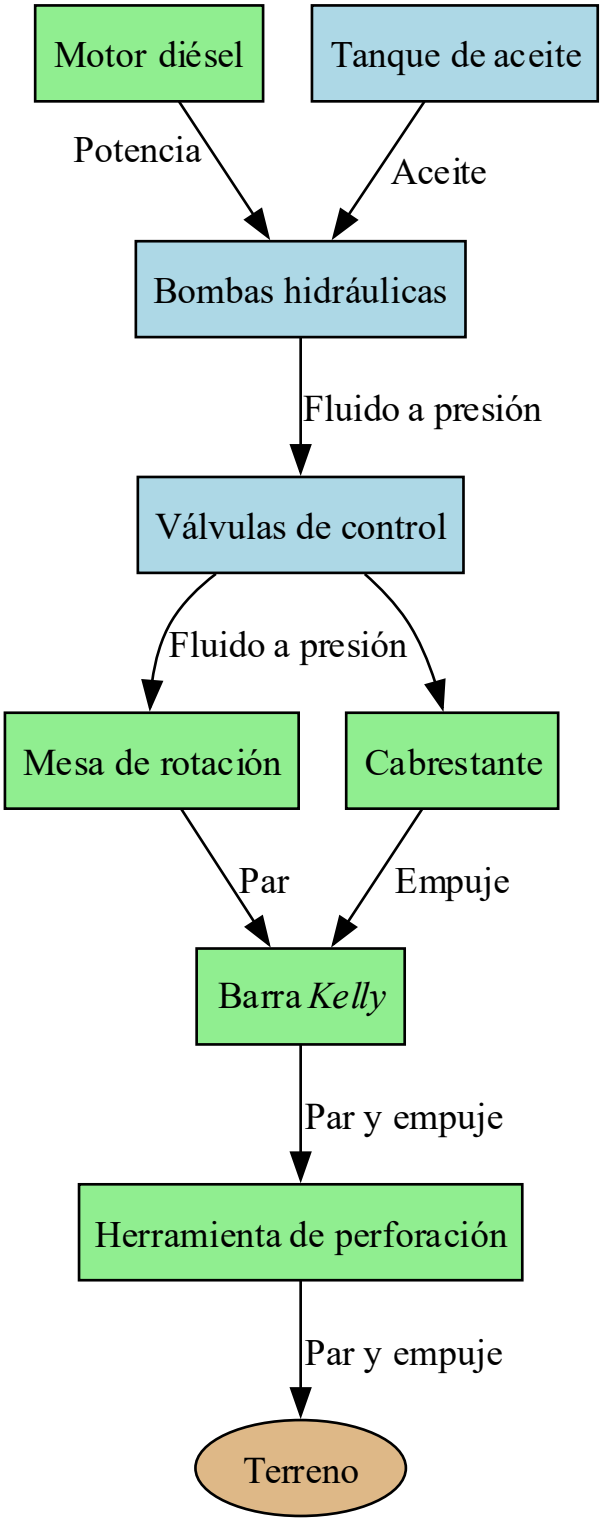


Figura 2.3: Esquema de funcionamiento de la maquinaria de perforación.

2.3. Registro de parámetros de perforación

El registro de parámetros de perforación se refiere a la recogida de datos relevantes del proceso de perforación sin afectar al mismo. Si bien, en algunos casos, puede hacerse manualmente, lo más usual es que se realice mediante sistemas automatizados que controlan una serie de sensores instalados en el equipo de perforación. Estos sensores proporcionan, de manera continua y en tiempo real, diferentes aspectos de la perforación. Los datos quedan registrados para su consulta y análisis posterior.

2.3.1. Normas aplicables al registro de parámetros

En el apartado 4 del artículo 7.9 del *Eurocódigo 7* (AEN/CTN 140 Eurocódigos estructurales, 2016), se hace referencia a los aspectos que debería incluir el registro de cada pilote. Ciñéndonos exclusivamente a los parámetros relacionados con la perforación a rotación tendríamos el diámetro, la longitud, la velocidad de penetración o el par de rotación.

En la norma *UNE-EN 1536:2011* para pilotes perforados (AEN/CTN 103 Geotecnia, 2016a), en su Tabla B.2, se recoge que se pueden controlar en el proceso de perforación la velocidad de rotación, la velocidad de penetración, la longitud y el par.

Por otro lado, la norma *UNE-EN 12699:2016* para pilotes de desplazamiento (AEN/CTN 103 Geotecnia, 2016b), en su apartado 9.2.8 , recoge que, para pilotes atornillados, los registros deberían incluir el par y la presión aplicada.

Ninguna de estas normas entra en profundidad en el tema del registro de parámetros. Existe, no obstante, la norma *UNE-EN ISO 22476-15:2016* (AEN/CTN 103 Geotecnia, 2016c), que trata precisamente de ello. Sin embargo, en principio, no es aplicable a la ejecución de pilotes, sino que se centra en el registro de parámetros de sondeos de investigación geotécnica. A falta de un documento específico de registro de parámetros para pilotes o tratamientos del terreno, podemos acudir a esta norma, ya que la mayor parte de sus principios son igualmente aplicables.

2.3.2. Parámetros de perforación

Esta norma *UNE-EN ISO 22476-15* define los parámetros de perforación como «Parámetros medidos y registrados en la maquinaria de perforación durante la

perforación (principalmente presiones hidráulicas, profundidad, velocidad de penetración, velocidad de rotación, presión y flujo del fluido, etc.).»

Estos parámetros, recoge la norma, pueden clasificarse en tres categorías:

- Parámetros impuestos por el método (por ejemplo, tipo y diámetro de la herramienta de perforación) y parámetros no controlados (por ejemplo, desgaste de las herramientas).
- Parámetros establecidos por el operador (por ejemplo p_O presión de empuje y v_R velocidad de rotación).
- Parámetros dependientes de la respuesta del terreno (por ejemplo v_A velocidad de penetración y C_R par de rotación)[1].

Algunos de los parámetros pueden obtenerse mediante medición directa, mientras que otros deben calcularse usando relaciones de calibración, como veremos más adelante.

2.3.3. Sistemas de medición

De manera general, los sistemas de medición automática incluyen:

- Equipos para el registro y visualización de los datos.
- Un sistema de adquisición de datos que incluye un convertidor de señal.
- Transductores fijados en el equipo.

Estos sistemas de medición pueden registrar una amplia gama de fenómenos eléctricos y físicos en función de la aplicación. En el caso de sistemas electromecánicos suelen registrarse:

- Presión.
- Aceleración y vibración.
- Inclinación.
- Desplazamiento.
- Caudal.

[1] En el caso de la perforación mediante máquinas piloteras la clasificación que propone la norma es, cuando menos, debatible. Quizá uno de los parámetros más directamente controlables por el operador es la velocidad de penetración ya que la máquina no avanza si el operador no la acciona.

En cambio, las máquinas trabajan normalmente en modo automático, donde el sistema maximiza la velocidad de rotación en función de la resistencia del terreno (es decir, en función del par necesario). En este caso la velocidad de rotación no dependería del operador. Esto no coincide con lo que se recoge en la norma.

Para medir estas magnitudes físicas se necesita un sensor o transductor que las convierta en una señal eléctrica apta para su procesamiento. A su vez, y dependiendo del transductor específico, estas señales analógicas requerirán cierto acondicionamiento, como la amplificación o el filtrado.

Estas señales analógicas son procesadas por un convertidor analógico-digital (ADC) para transformarlas en datos digitales. Primero, el ADC realiza un muestreo de la señal a intervalos regulares. La señal muestreada se divide en niveles discretos en función de la resolución del ADC. A cada señal se le aplica un código y se transforma en datos digitales.

Estos datos digitales por el ADC pueden ser leídos y procesados por un ordenador u otro dispositivo de registro de datos.

2.3.3.1. Sensores de presiones hidráulicas

En el caso que nos ocupa, las presiones a medir serían:

- Presión de empuje p, asociada al motor del cabrestante y de la cual obtendríamos la fuerza de empuje.
- Presión de par p_{CR}, asociada a la mesa de rotación y de la cual obtendremos el par de rotación.

Además, de manera complementaria a la presión de empuje, es necesario medir la presión de retención p_X para poder medir el empuje neto aplicado sobre la herramienta de perforación.

Estas presiones se miden mediante sensores de presión colocados en el sistema hidráulico, lo más cerca posible del motor cuya presión se busca medir.

2.3.3.2. Sistemas de medición del avance

Los sistemas de medición del avance, a, pueden ser de diferentes tipos, como un codificador rotatorio montado en el eje del motor de alimentación, una cadena/cable unido entre el mástil fijo y la cabeza de perforación en movimiento, una rueda montada en la cabeza de la barrena y que corre contra el mástil u otro método adecuado.

2.3.3.3. Sistemas de medición de la velocidad de rotación

La medición de la velocidad de rotación (v_R) se realiza normalmente mediante un sensor de rotación o proximidad que cuenta el paso de un elemento metálico.

2.3.3.4. Medición del tiempo

El tiempo se mide de manera automática a partir del reloj interno del sistema de adquisición de datos.

2.3.4. Factores que influyen en los resultados del registro de parámetros de perforación

Aparte de las características del terreno atravesado existen, tres factores que influyen sobre los resultados del registro de parámetros.

2.3.4.1. Influencia del tipo de herramienta.

El tipo de herramienta determina la efectividad del método de perforación para unas condiciones del terreno dadas. Por ejemplo, el tipo de picas o palas utilizadas y el desgaste de la herramienta pueden tener gran efecto sobre el proceso de perforación.

2.3.4.2. Influencia de la máquina

Cómo es lógico, las características de la maquinaria de perforación tales como su capacidad y su diseño hidráulico influyen en los resultados del registro de parámetros. Por ejemplo, un mismo valor de presión medido en la mesa de rotación equivaldrá a diferentes valores de par en función de la máquina utilizada. Por tanto, hay que ser especialmente cuidadosos al comparar los registros obtenidos por diferentes tipos de perforadoras.

2.3.4.3. Influencia del operador

El operador debería trabajar de una manera consistente para poder comparar adecuadamente los resultados. De hecho, habría que registrar los cambios de operador. Es recomendable, por tanto, que la maquinaria funcione de la manera más automática posible, con el objetivo de reducir la influencia del operador y que los registros dependan de las propiedades del suelo perforado.

2.3.5. Obtención de los parámetros de perforación

Algunos de los parámetros de perforación se obtienen de manera directa (tiempo, velocidad de rotación) sin embargo, hay otros que es necesario calcular a partir de los valores brutos medidos.

2.3.5.1. Velocidad de penetración

La velocidad de penetración, v_A, se obtiene a partir de las medidas de avance, a, y del intervalo de tiempo d_t.

$$v_A = \frac{a_{t+dt} - a_t}{dt}$$

Ecuación 2.1: Velocidad de penetración

Normalmente, en los sistemas de registro automático, el intervalo considerado es siempre el mismo, ya sea en función de una determinada longitud de avance o de un intervalo de tiempo prestablecido.

2.3.5.2. Presión de empuje

La presión de empuje sobre la herramienta, p_O, se obtiene a partir de la presión hidráulica del motor de empuje, p. La fórmula propuesta por la *ISO 22476-15*, es la siguiente:

$$s_O \cdot p_O = \frac{p}{p_{máx.}} F_{máx.} + W_H + n \cdot W_R$$

Ecuación 2.2: Empuje sobre la herramienta

Siendo s_O el área de la perforación, $F_{máx.}$ la máxima fuerza de empuje evaluada durante la calibración con el cabezal en parada mecánica inferior, W_H el peso de la corona giratoria y W_R el peso unitario de n varillas. La componente del peso de varillas no tiene sentido en el caso de piloteras, pero en general la fórmula a aplicar es similar:

$$F_E = s_O \cdot p_O = k_E \cdot p + F_0$$

Ecuación 2.3: Empuje sobre la herramienta en perforadoras

Dónde F_0 es el peso total que ejerce la herramienta de perforación cuando no se aplica ningún empuje adicional y k_E es la relación entre la presión medida y la fuerza de empuje adicional.

2.3.5.3. Par de rotación

Según se recoge en la norma de registro de parámetros, el par de rotación, C_R, aplicado sobre la herramienta de perforación es proporcional a la presión hidráulica, p_{CR}, que recibe la mesa de rotación y el par máximo, $C_{Rmáx.}$.

$$C_R = \frac{p_{CR} - p_{CR0}}{p_{CRmáx.}} C_{Rmáx.}$$ Ecuación 2.4: Par de rotación

Sin embargo, como puntualiza la norma, el par se ve afectado cuando varían las relaciones de engranajes para cambiar la relación entre la velocidad del motor y la velocidad de rotación. Es decir, la expresión anterior sería correcta para una velocidad de rotación concreta. Para tener en cuenta los cambios de velocidad podríamos utilizar la siguiente expresión:

$$C_R = k_C(p_{CR} - p_{CR0})$$ Ecuación 2.5: Par de rotación

Dónde el coeficiente k_C varía en función de la velocidad de rotación.

2.3.6. Parámetros compuestos

Los parámetros de perforación deben estar relacionados con las características del suelo atravesado, sin embargo, la relación no es directamente evidente. Por ello, diversos autores han desarrollado parámetros compuestos que se deducen de la combinación de diversos parámetros de perforación. El objetivo de estos parámetros es facilitar la interpretación de los resultados del registro de parámetros, ayudando a la compresión y caracterización de las condiciones del terreno.

Los parámetros compuestos que aparecen en la norma *ISO 22476-15* se recogen en la Tabla 2.1.

Tabla 2.1: Parámetros compuestos

Nombre	Fórmula	Referencia
Ecuación 2.6: Resistencia a la penetración	$P_R = (t)_{z=0,2\ m}$	(Svenska Geotekniska Föreningen, 1999)
Ecuación 2.7: Resistencia suelo-roca	$R_{SR} = \frac{p_E}{v_A}$	(AEN/CTN 103 Geotecnia, 2016c)
Ecuación 2.8: Índice de Somerton	$S_d = p_E \sqrt{\frac{v_R}{v_A}}$	(Somerton, 1959)
Ecuación 2.9: Energía específica	$E_S \approx p_E + C_R \cdot \frac{v_R}{v_A}$	(Pfister, 1985), (Teale, 1965)
Ecuación 2.10: Energía normalizada	$E = \frac{\alpha \cdot p_E \cdot v_A + \beta \cdot C_R \cdot 2\pi \cdot v_R + \gamma \cdot p_M \cdot f}{v_A}$	(Nishi et al., 1998)

En geotecnia, siendo un campo expecialmente empírico, resulta conveniente comprobar cuál fue el procedimiento y variables tenidos en cuenta para las fórmulas y correlaciones que se utilizan.

La resistencia a la penetración se usa «para determinar la profundidad de la roca, la resistencia relativa de la roca y su fracturación». El caso de la resistencia suelo-roca es similar.

Por su parte el índice de Somerton se obtiene a partir de la perforación de testigos de roca en laboratorio. Igualmente, el trabajo de Teale trata la perforación de rocas y no de suelos. Por, último la fórmula de Nishi es similar a la de Teale, añadiendo los parámetros de eficiencia α, ß, γ.

Como se puede comprobar, estos parámetros compuestos han sido formulados principalmente para rocas o suelos duros, y podrían no ser lo suficientemente precisos para el estudio de suelos blandos. Además, debido a la forma de la herramienta, en la que se aplica el par a lo largo del fuste de esta, estos parámetros compuestos podrían no ser adecuados para la perforación mediante desplazamiento a rotación.

2.4. Ensayos de penetración

Los ensayos de penetración consisten en la introducción de un elemento de punta cónica en el terreno. Pueden clasificarse en dos tipos: estáticos y dinámicos.

2.4.1. Penetrómetro dinámico

El ensayo de penetración dinámica consiste en la hinca por golpeo de una puntaza cónica utilizando una maza de masa y altura de caída conocidas. La resistencia a la penetración consistiría en el número de golpes requeridos para hincar la puntaza cónica una longitud determinada. Este tipo de ensayos se tratan en la norma *UNE-EN ISO 22476-2* (CEN, 2008).

Si bien es un ensayo que no proporciona tanta información como los penetrómetros estáticos, pueden utilizarse para la determinación del perfil de suelo (junto con otros ensayos existentes), para obtener propiedades de resistencia y deformación de suelos mediante las correlaciones adecuadas, para determinar el nivel de apoyo de pilotes, etc.

2.4.1.1. Tipos de penetrómetro dinámico

Este ensayo puede clasificarse en función de la energía aplicada por golpe en los siguientes tipos:

- Ensayo de penetración dinámica ligera (*Dynamic Probing Light*, DPL).
- Ensayo de penetración dinámica mediana (*Dynamic Probing Medium*, DPM).
- Ensayo de penetración dinámica pesada (*Dynamic Probing Heavy*, DPH).
- Ensayo de penetración dinámica superpesada (*Dynamic Probing Super Heavy*, DPSH).

Las características de los diferentes tipos de ensayo pueden consultarse en la Tabla 1 de la norma (Figura 2.4).

Aparato para ensayos de penetración dinámica	Símbolo	Unidad	DPL (ligero)	DPM (mediano)	DPH (pesado)	DPSH (súper pesado)	
						DPSH-A	DPSH-B
Dispositivo de hinca masa de la maza, nueva altura de caída	 m h	 kg mm	 10 ± 0,1 500 ± 10	 30 ± 0,3 500 ± 10	 50 ± 0,5 500 ± 10	 63,5 ± 0,5 500 ± 10	 63,5 ± 0,5 750 ± 20
Cabeza de impacto diámetro masa (máx.) (barra guía incluida)	 d m	 mm kg	 $50<d<D_h$ [a] 6	 $50<d<D_h$ [a] 18	 $50<d<0,5D_h$ [a] 18	 $50<d<0,5D_h$ 18	 $50<d<0,5D_h$ [a] 30
Puntaza de 90° área nominal de la base diámetro de la base, nuevo diámetro de la base, gastada (mín.) longitud de la camisa cilíndrica longitud de la punta cónica desgaste máx. permisible de la punta	 A D L 	 cm^2 mm mm mm mm mm	 10 35,7 ± 0,3 34 35,7 ± 1 17,9 ± 0,1 3	 15 43,7 ± 0,3 42 43,7 ± 1 21,9 ± 0,1 4	 15 43,7 ± 0,3 42 43,7 ± 1 21,9 ± 0,1 4	 16 45,0 ± 0,3 43 90,0 ± 2 [b] 22,5 ± 0,1 5	 20 50,5 ± 0,5 49 51 ± 2 25,3 ± 0,4 5
Barras de hinca [c] masa (máx.) diámetro exterior (máx.) desviación de la barra [d]: en los 5 m inferiores en el resto	 m d_r 	 kg/m mm % %	 3 22 0,1 0,2	 6 32 0,1 0,2	 6 32 0,1 0,2	 6 32 0,1 0,2	 8 35 0,1 0,2
Trabajo específico por golpe	mgh/A E_n	kJ/m^2	50	100	167	194	238

[a] D_h: diámetro de la maza. En caso de forma rectangular, se asume la menor dimensión como equivalente al diámetro.
[b] Puntaza desechable (perdida) solamente.
[c] La longitud máxima de las barras no debe exceder de 2 m.
[d] Desviación de la barra respecto de la vertical.

NOTA Las tolerancias indicadas son las tolerancias de fabricación.

Figura 2.4: Tabla 1 de la *UNE-EN ISO 22476-2*

Además de los tipos presentes en la norma, pueden existir otros como el penetrómetro Borro. Este penetrómetro estaba recogido en la antigua norma *NLT-261* del CEDEX y es similar a los penetrómetros tipo DPSH.

2.4.1.2. Ejecución del penetrómetro dinámico

Durante la ejecución del ensayo se debe hincar el penetrómetro de forma continua en el terreno a una velocidad de entre 15 y 30 golpes por minuto. Las barras se deben rotar 1,5 vueltas o hasta alcanzar el par de torsión máximo, al menos, cada metro. El par de torsión necesario para el giro debe registrarse.

El número de golpes se registra cada 100 mm de penetración en el caso de los penetrómetros tipo DPL, DPM y DPH, y cada 100 o 200 mm en el caso de los tipos DPSH-A y DPSH-B.

2.4.1.3. Factores que influyen en los resultados del penetrómetro dinámico

Los resultados del penetrómetro dinámico dependen principalmente de las características del equipo utilizado y de las propiedades del suelo atravesado. Esto nos permite tener una idea de la resistencia del suelo.

Sin embargo, existen otros factores que pueden influir en la correcta interpretación de los resultados del penetrómetro. Algunos de estos factores serían los siguientes:

- En suelos finos, el rozamiento en la barra puede afectar significativamente al valor registrado del número de golpes. Este efecto es más importante con diámetro de puntaza pequeño debido a que el cono puede ser poco más ancho que la barra.
- En suelos granulares para unas mismas características, el golpeo disminuye bajo el nivel freático debido a que la presión vertical efectiva es más baja. En suelos limosos, para unas mismas características, el golpeo puede ser igual o mayor bajo el nivel freático debido al efecto de capilaridad.
- Los suelos de partículas con bordes afilados o rugosos tienen una resistencia a la penetración más alta que los suelos de partículas redondas o lisas.
- Cantos y bolos pueden aumentar significativamente la resistencia a la penetración.
- El número de golpes para atravesar un determinado terreno aumenta con una mayor longitud de barra ya que la eficiencia decrece.

En la Figura 2.5 se muestra una máquina de penetrómetros dinámicos mientras que la Figura 2.6 muestra los resultados y la gráfica obtenidos de un penetrómetro dinámico.

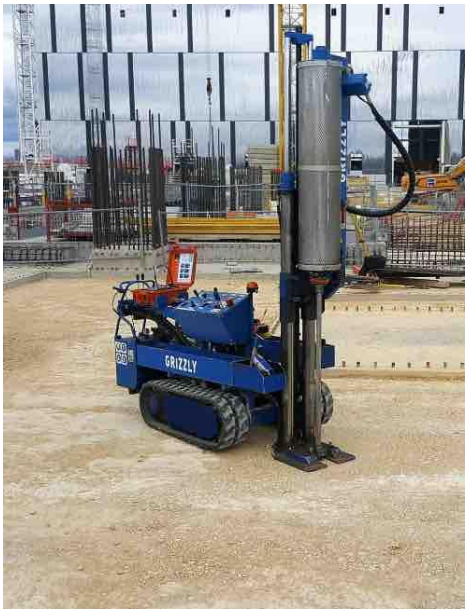


Figura 2.5: Máquina GRIZZLY capaz de ejecutar penetrómetros tipo DPSH

Prof. (m)	Golpeo	Prof. (m)	Golpeo
0-0.2	0	7.4-7.6	13
0.2-0.4	9	7.6-7.8	14
0.4-0.6	8	7.8-8	10
0.6-0.8	5	8-8.2	15
0.8-1	5	8.2-8.4	29
1-1.2	8	8.4-8.6	15
1.2-1.4	8	8.6-8.8	15
1.4-1.6	9	8.8-9	12
1.6-1.8	11	9-9.2	13
1.8-2	9	9.2-9.4	13
2-2.2	6	9.4-9.6	13
2.2-2.4	3	9.6-9.8	14
2.4-2.6	3	9.8-10	14
2.6-2.8	3	10-10.2	14
2.8-3	6	10.2-10.4	16
3-3.2	5	10.4-10.6	17
3.2-3.4	5	10.6-10.8	18
3.4-3.6	5	10.8-11	17
3.6-3.8	15	11-11.2	19
3.8-4	6	11.2-11.4	24
4-4.2	4	11.4-11.6	25
4.2-4.4	3	11.6-11.8	27
4.4-4.6	3	11.8-12	29
4.6-4.8	4	12-12.2	25
4.8-5	4	12.2-12.4	27
5-5.2	6	12.4-12.6	30
5.2-5.4	7	12.6-12.8	28
5.4-5.6	6	12.8-13	24
5.6-5.8	7	13-13.2	30
5.8-6	9	13.2-13.4	31
6-6.2	9	13.4-13.6	35
6.2-6.4	10	13.6-13.8	36
6.4-6.6	11	13.8-14	38
6.6-6.8	9	14-14.2	34
6.8-7	9	14.2-14.4	31
7-7.2	9	14.4-14.6	30
7.2-7.4	10	14.6-14.8	37

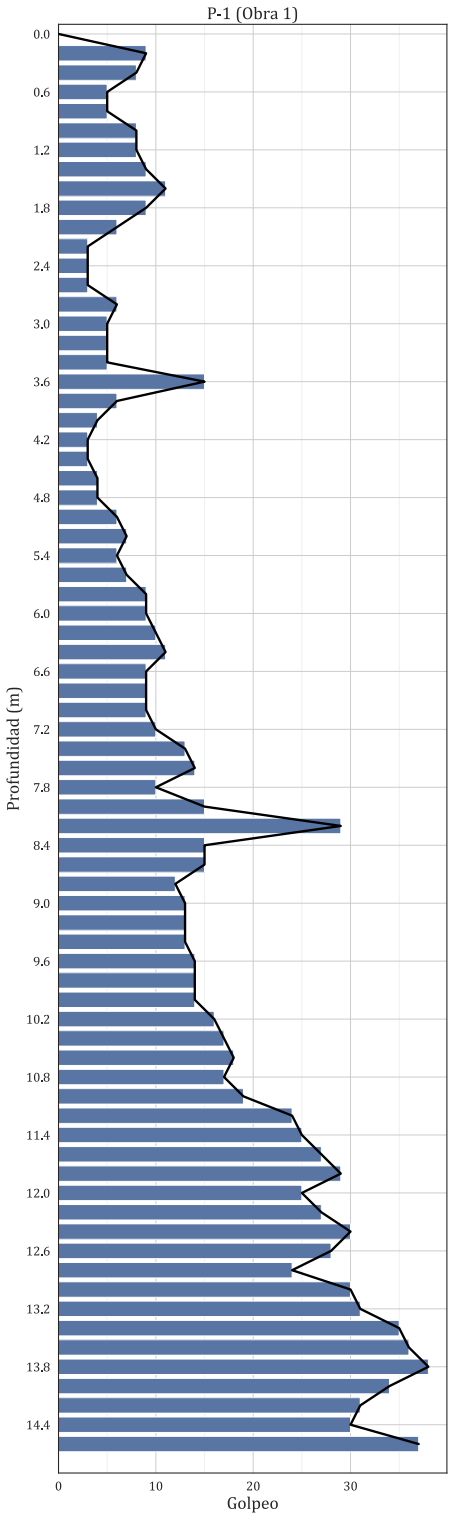


Figura 2.6: Ejemplo de resultados obtenidos de la ejecución de un penetrómetro dinámico

2.4.2. Penetrómetro estático

El ensayo de penetración con cono (*Cone Penetration Test*, CPT) consiste en introducir en el suelo un penetrómetro cónico, por medio de una serie de varillas de empuje, a un ritmo de penetración constante (aproximadamente 20 mm/s). Para este ensayo puede acudirse a dos normas distintas, la *UNE-EN ISO 22476-1* (AEN/CTN 103 Geotecnia, 2012) para el cono y piezocono eléctricos, y la *UNE-EN ISO 22476-12* (AEN/CTN 103 Geotecnia, 2010) para el cono mecánico. Esta tesis no pretende realizar un estudio exhaustivo del penetrómetro estático por lo que nos centraremos en los datos más relevantes sobre el cono y el piezocono eléctricos.

Durante la penetración pueden medirse la resistencia a la penetración del cono, la resistencia a la penetración total y/o el rozamiento por fuste. Adicionalmente, el piezocono o CPTU incluye la medición de la presión de poro alrededor del cono. El registro de los datos se realiza usualmente cada 20 mm de longitud de penetración, si bien esta frecuencia puede aumentar en función de la categoría del ensayo o del nivel de detalle requerido.

Estos tipos de ensayos proporcionan valiosa información permitiendo evaluar:

- la estratificación;
- el tipo de comportamiento del suelo;
- parámetros geotécnicos tales como:
    - la densidad del suelo;
    - los parámetros de resistencia al corte;
    - las características de deformación y consolidación;
    - la conductividad hidráulica y la presión del agua subterránea.

Además, los resultados obtenidos de un ensayo de penetración con cono pueden ser directamente utilizados en el diseño geotécnico.

2.4.2.1. Resistencia a la penetración del cono

La resistencia a la penetración del cono, q_c, se define como el cociente entre la fuerza aplicada en el cono, Q_c, y el área de la sección proyectada del cono, A_c.

$$q_c = \frac{Q_c}{A_c}$$

Ecuación 2.11: Resistencia a la penetración del cono

2.4.2.2. Rozamiento por fuste

El rozamiento por fuste medido, f_s, se calcula como la fuerza F_s medida que actúa en el manguito de fricción dividida por el área del fuste, A_s.

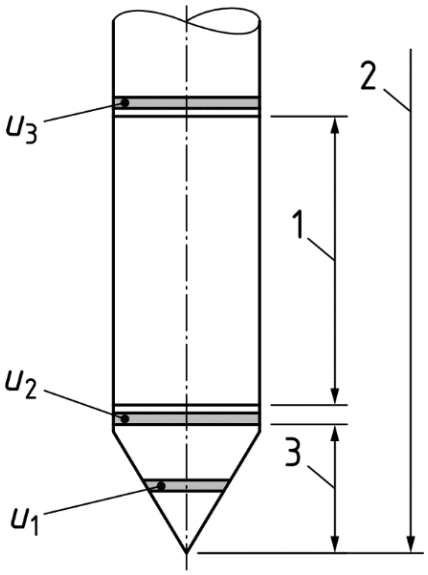
$$f_s = \frac{F_s}{A_s}$$

Ecuación 2.12: Resistencia a la penetración del cono

2.4.2.3. Presión de poro

En el caso del piezocono, la presión de poro es la presión medida en el elemento filtrante durante la penetración. Esta presión puede medirse en diferentes posiciones (Figura 2.7):

- u_1 en la cara del cono;
- u_2 en la sección del cono (entre el cono y el fuste);
- u_3 inmediatamente detrás del manguito de fricción.



Leyenda

1. Manguito de fricción
2. Penetrómetro cónico
3. Cono

Figura 2.7: Posiciones de los filtros de presión de poro *UNE-EN ISO 22476-1* (AEN/CTN 103 Geotecnia, 2012)

2.4.2.4. Resistencias corregidas

La presión de agua circundante influye en la resistencia del cono y en el rozamiento por fuste medidos. Esto se debe al efecto de la presión de agua en los huecos existentes entre el cono y el manguito de fricción, y en el hueco por encima del manguito de fricción (Figura 2.8).

Cuando se mide la presión del agua utilizando un elemento filtrante en la extensión cilíndrica del cono (u_2), se debe corregir la resistencia del cono utilizando la siguiente fórmula:

$$q_t = q_c + u_2 \cdot (1 - a)$$

Ecuación 2.13: Resistencia a la penetración del cono corregida

Donde:

q_t es la resistencia del cono corregida, en MPa;

q_c es la resistencia del cono medida, en MPa;

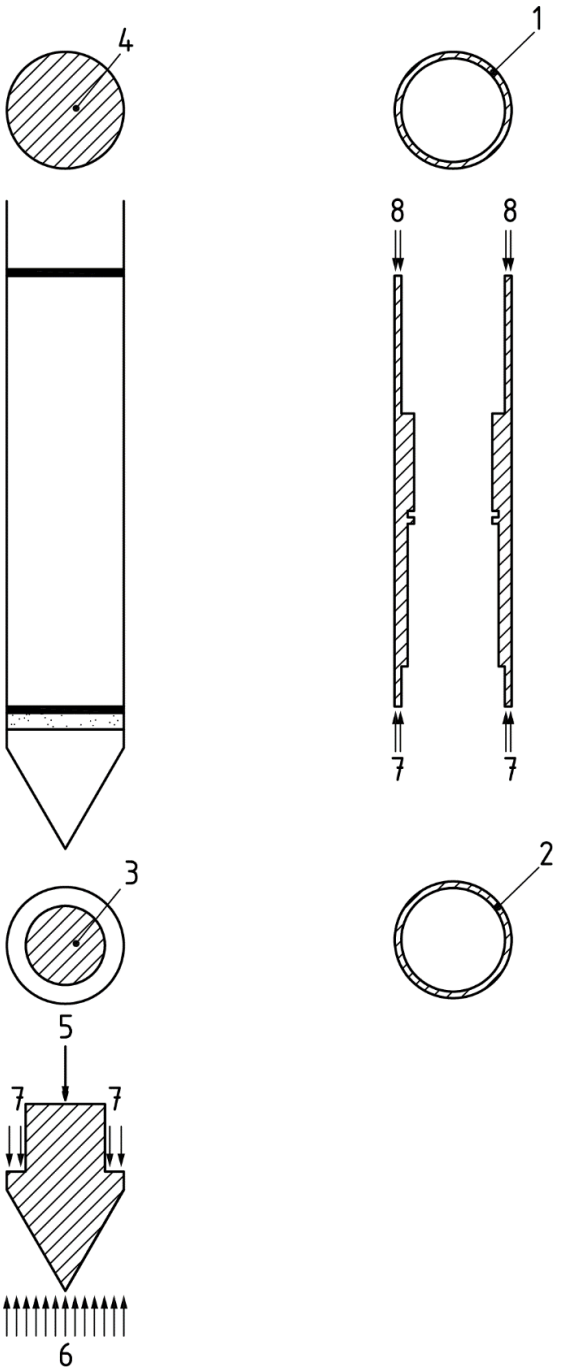
u_2 es la presión del agua de poro en la parte cilíndrica de la extensión del cono (que se asume igual a la presión de poro en el hueco entre el cono y el fuste), en MPa;

a es la relación de área neta $a = A_n/A_c$, donde:

- A_c es el área proyectada de la sección transversal del cono;
- A_n es el área de la sección transversal de la célula de carga o el vástago.

Idealmente, el penetrómetro cónico debería tener una relación de área neta lo más próxima posible a 1, estando los valores de los penetrómetros cónicos utilizado comúnmente entre a=0,5 y a=0,9 (CEN, 2023).

Las mediciones solo se deberían corregir si se mide u_2.



Leyenda

1. Área de la sección transversal (superior), A_{st}
2. Área de la sección transversal (inferior), A_{sb}
3. Área de la sección transversal del vástago, A_n
4. Área proyectada de la sección transversal, A_c
5. Fuerza medida axialmente en el cono, Q_c
6. Resistencia del cono corregida, q_t
7. Presión de poro medida en la posición 2, u_2
8. Presión de poro medida en la posición 3, u_3

Figura 2.8: Corrección de la resistencia del cono y del rozamiento por fuste debido al efecto del área final desigual *UNE-EN ISO 22476-1* (AEN/CTN 103 Geotecnia, 2012)

En el caso del rozamiento por fuste, este no se suele corregir ya que no es práctica habitual medir la presión de poro por encima del manguito de fricción (u_3). No obstante, el rozamiento por fuste corregido puede obtenerse a partir de la siguiente fórmula:

$$f_t = f_s - \frac{u_2 \cdot A_{sb} - u_3 \cdot A_{st}}{A_s}$$

Ecuación 2.14: Resistencia por fuste corregida

Donde:

f_t es el rozamiento por fuste corregido, en MPa;

f_s es el rozamiento por fuste medido, en MPa;

A_s es el área del manguito de fricción, en mm^2;

A_{sb} es el área de la sección transversal de la parte inferior del manguito de fricción, en mm^2;

A_{st} es el área de la sección transversal de la parte superior del manguito de fricción, en mm^2;

u_2 es la presión de poro medida entre el manguito de fricción y el cono, en MPa;

u_3 es la presión de poro medida por encima del manguito de fricción, en MPa.

Estas correcciones, de la resistencia del cono y de la resistencia por fuste son importantes en suelo de grano fino donde el exceso presión de poro es significativo.

En la Figura 2.9 se muestra una máquina de ejecución de CPT y en la Figura 2.10 resultados obtenidos de uno de estos ensayos.

Figura 2.9: Máquina de CPT de ConeTec

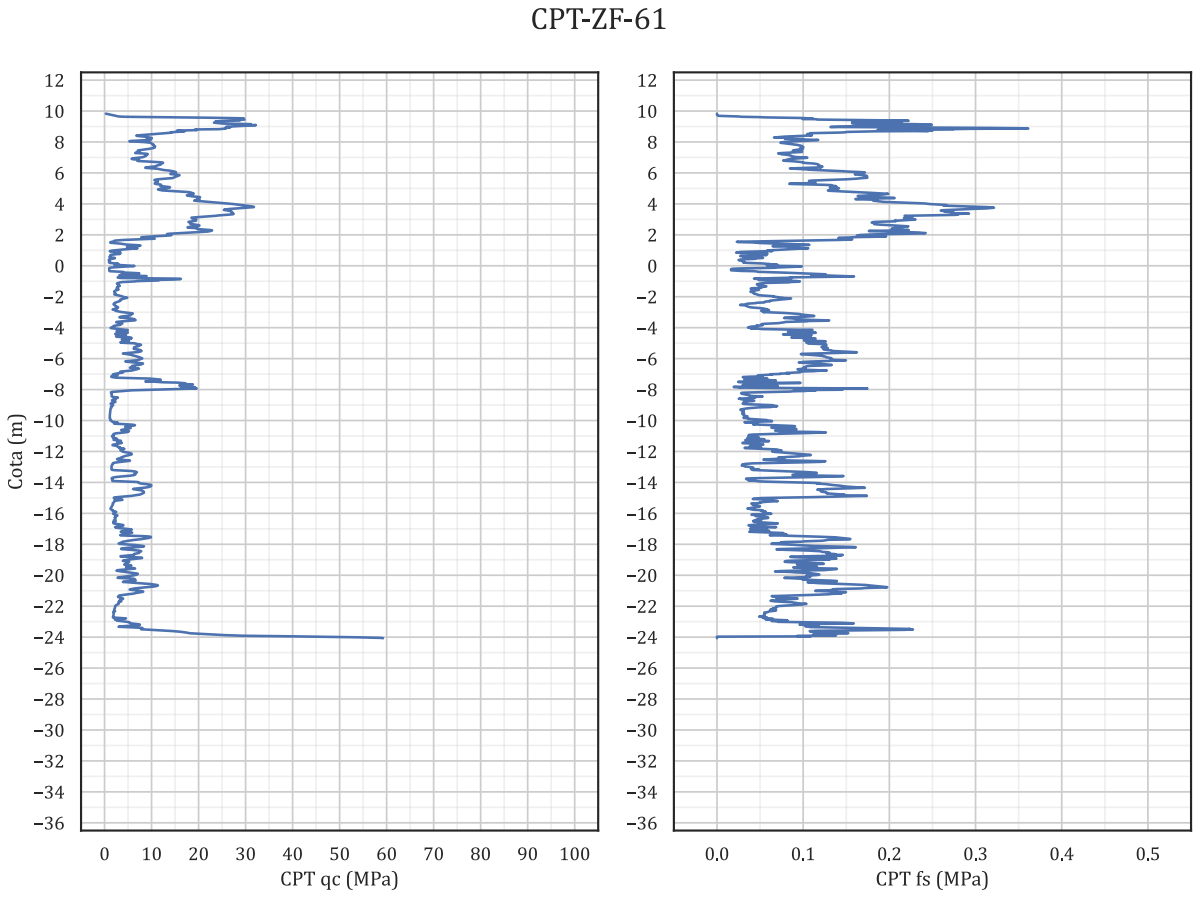


Figura 2.10: Representación de valores obtenidos durante la ejecución de un CPT

2.5. Conceptos de *Machine Learning*

«El aprendizaje automático (*Machine Learning*, ML) es un campo de la ciencia de la computación que estudia algoritmos y técnicas para automatizar soluciones a problemas complejos que son difíciles de programar utilizando métodos de programación convencionales» (Rebala et al., 2019).

De manera tradicional, un programa se diseña con una serie de reglas predefinidas para resolver un problema, pero ¿qué ocurre si se desconocen las reglas reales que rigen el problema? ¿y si dichas reglas son excesivamente complejas? En esas circunstancias resulta más eficiente crear programas que aproximen las relaciones internas del problema.

Podríamos definir un algoritmo de ML como un conjunto de instrucciones para construir un programa en función de los datos de partida. De esta manera, en función de los datos introducidos se generaría un modelo distinto, de ahí que se hable de «aprendizaje». Lo que diferenciaría un algoritmo de otro son las instrucciones utilizadas para generar el modelo.

Este trabajo no pretende tratar en profundidad el ML; existiendo tesis previas en la materia (p. ej. (Haghiri, 2019; Li, 2017; Louppe, 2015; Rolnick, 2018)). No obstante, se realiza una revisión de los conceptos básicos de ML y del funcionamiento de diferentes algoritmos que son posteriormente utilizados para el desarrollo de la tesis.

2.5.1. Aprendizaje supervisado y no supervisado

Los problemas de ML pueden clasificarse en aprendizaje supervisado y no supervisado. En el aprendizaje supervisado, se parte de un conjunto de datos de los que se conoce la respuesta esperada (datos etiquetados). El algoritmo debe ser capaz de aprender de dichos datos de tal forma que, al recibir los nuevos, sea capaz de etiquetarlos correctamente.

Por otra parte, en el aprendizaje no supervisado no existe una «respuesta correcta» sino que se buscan similitudes en los datos de cara a poder agruparlos en función de sus características.

En esta tesis buscamos generar un modelo de ML capaz de predecir el perfil penetrométrico del suelo habiéndolo entrenado previamente con datos etiquetados. Por tanto, estamos ante un problema de aprendizaje supervisado.

2.5.2. Clasificación y regresión

Los problemas de aprendizaje supervisado pueden, subdividirse en problemas de clasificación y regresión. Un problema de clasificación busca asignar una determinada categoría a cada muestra (p. ej. las categorías baja, media y alta).

Los problemas de regresión, en cambio, buscan predecir valores de una variable continua, como pueda ser el valor de una acción en bolsa o la temperatura que hará mañana.

En el caso del penetrómetro dinámico, los valores a predecir son discretos (golpeos de 5, 15 o 27, pero no de 5.2, 15.4 o 27.8) con lo que podríamos tratarlo como un problema de clasificación. Sin embargo, la naturaleza del problema es claramente continua, en el sentido de que esos valores reflejan una magnitud, la resistencia a la penetración del penetrómetro dinámico, que tiene una naturaleza continua. En ese aspecto, los valores del penetrómetro son más un redondeo a número entero que una categoría.

En esta tesis se comprueban los resultados para los enfoques tanto de clasificación como de regresión para el caso del penetrómetro dinámico. En el penetrómetro estático el problema se consideró claramente de regresión.

2.5.3. Entrenamiento y evaluación

A la hora de entrenar un algoritmo de ML para generar un modelo el objetivo es poder predecir nuevos valores. Para poder comprobar esa capacidad, no se entrena el algoritmo con todos los datos disponibles, sino que se aparta una determinada cantidad sobre la que se evaluará posteriormente el modelo.

No hay un criterio único para elegir la ratio entrenamiento/validación, aunque son de uso común 80/20, 75/25 o 60/40. En esta tesis se eligió la ratio 75/25.

2.5.4. Sesgo y varianza. Sobreajuste y subajuste

Una pareja de conceptos importantes en el análisis estadístico son el sesgo y la varianza. El sesgo se refiere al error introducido al intentar aproximar un problema real mediante un modelo mucho más simple (James et al., 2013). Se podría interpretar como la diferencia entre la forma del modelo y la forma real del problema.

Por su parte, la varianza se refiere a la magnitud en la que cambiaría nuestra predicción de haberla realizado en un conjunto de entrenamiento diferente.

Estos dos conceptos están relacionados. Mientras que un algoritmo muy flexible se adaptará bien a la forma real, será especialmente sensible respecto a los cambios en el conjunto de entrenamiento (poco sesgo, pero mucha varianza). Este fenómeno se denomina sobreajuste y se manifiesta en que el modelo no es capaz de dar resultados generales más allá del conjunto de entrenamiento (Figura 2.11, a).

Por otro lado, un algoritmo muy rígido no será capaz de adaptarse de manera precisa a los datos, pero se verá poco influenciado por variaciones de estos (poca varianza, pero mucho sesgo). Este fenómeno se denomina subajuste y puede provocar predicciones muy poco precisas del modelo (Figura 2.11, b)

Normalmente es deseable encontrar un equilibrio de tal forma que el modelo no sea ni demasiado rígido, ni demasiado flexible (Figura 2.11, c).

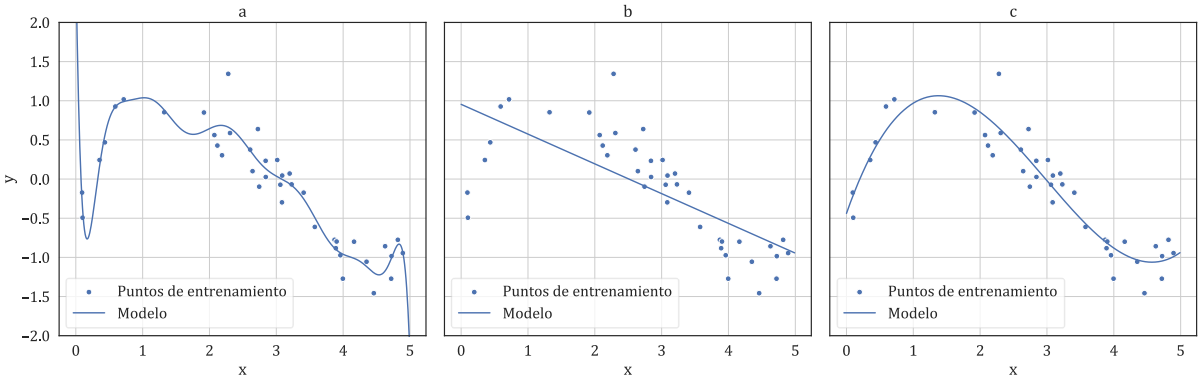


Figura 2.11: Ejemplo de sobreajuste (a), subajuste (b) y buen ajuste (c)

2.5.5. Parámetros e hiperparámetros

Los hiperparámetros limitan las opciones del modelo y son previos al entrenamiento de este. Por otro lado, los parámetros del modelo se ajustan durante el entrenamiento, alcanzándose el modelo óptimo (de aquellos posibles según los hiperparámetros previamente definidos).

Los hiperparámetros han de ajustarse para evitar tanto costes computacionales elevados como fenómenos de sobreajuste o subajuste. Este proceso de afinación de los hiperparámetros puede ser complejo, especialmente en algunos algoritmos con gran cantidad de ellos.

De manera esquemática el proceso sería el siguiente:

1. Escoger un algoritmo y analizar sus hiperparámetros
2. Definir qué hiperparámetros afinar. Puede ser necesario agruparlos ya que hay hiperparámetros que pueden estar relacionados entre sí. También puede ser relevante definir un orden de afinamiento debido a diferente importancia de los hiperparámetros y que algunos puedan ser relevantes dependiendo del valor que tomen otros.
3. Definir el rango de exploración y la escala para cada hiperparámetro.
4. Definir la métrica de evaluación para las combinaciones de hiperparámetros. Las métricas se miden en datos de validación donde no se haya entrenado el modelo.
5. Definir el método usado para explorar el espacio de hiperparámetros (*grid*, *random*, *genetic*, etc.).
6. Afinamiento.

Dos casos en los que la relación entre hiperparámetros es relevante serían:

- En algoritmos como *XGBoost,* o en aprendizaje profundo, reducir el tamaño de paso durante el proceso de optimización debe ir acompañado de un aumento en el número de epochs (el número de veces que el algoritmo utiliza por completo el conjunto de entrenamiento).
- En máquinas de vectores de soporte (*Support Vector Machines*, SVM), cada kernel tiene sus propios hiperparámetros.

El orden y la priorización de los hiperparámetros pueden influir, por ejemplo, en SVM, donde la elección de un kernel adaptado al problema estudiado tiene el mayor impacto en el funcionamiento del modelo y por tanto debería elegirse previamente al resto de hiperparámetros. En el caso de que el tamaño y el tiempo de resolución de un modelo sean limitantes, aquellos hiperparámetros que los afecten serán una prioridad.

En cuanto al rango de valores de los hiperparámetros a analizar:

- Algunos valores pueden ser irrelevantes en el problema analizado, por ejemplo, regularizaciones muy elevadas.
- Algunos hiperparámetros pueden tener un rango de valores posible muy amplio, de varios ordenes de magnitud. Utilizar una escala exponencial para estos hiperparámetros puede ser buena idea, especialmente para una primera aproximación.

- Algunos valores de los hiperparámetros pueden afectar significativamente al tiempo de entrenamiento del modelo por lo que conviene evaluar cuidadosamente si realmente son necesarios.

Una vez tenidas en cuenta todas estas consideraciones existen varias maneras de navegar por el espacio de los diferentes hiperparámetros. Esta búsqueda de los mejores hiperparámetros pueden clasificarse en dos tipos: ciegas y adaptativas. Algunos ejemplos de estos tipos se resumen a continuación.

- Búsqueda ciega de hiperparámetros.
    - *GridSearch* prueba todas las posibles combinaciones del espacio de hiperparámetros predefinido. Asegura encontrar la mejor combinación de hiperparámetros, pero el coste computacional puede ser prohibitivo.
    - *RandomSearch* prueba un determinado número de combinaciones del espacio de hiperparámetros. Proporciona una aproximación a la mejor combinación de hiperparámetros, a un menor coste, y permitir probar de manera más eficiente todo el rango del espacio de hierparámetros.
- Búsqueda adaptativa de hiperparámetros.
    - La optimización bayesiana consiste en crear gradualmente un modelo de probabilidad $p(y|x)$ mediante diversos métodos, donde y es la puntuación y x el valor de los hiperparámetros. Esto permite explorar las "zonas" donde se esperan mejores resultados y comprender como afectan los hiperparámetros al resultado final (Figura 2.12). Es posible que este método tienda a un óptimo local en lugar de a uno general.

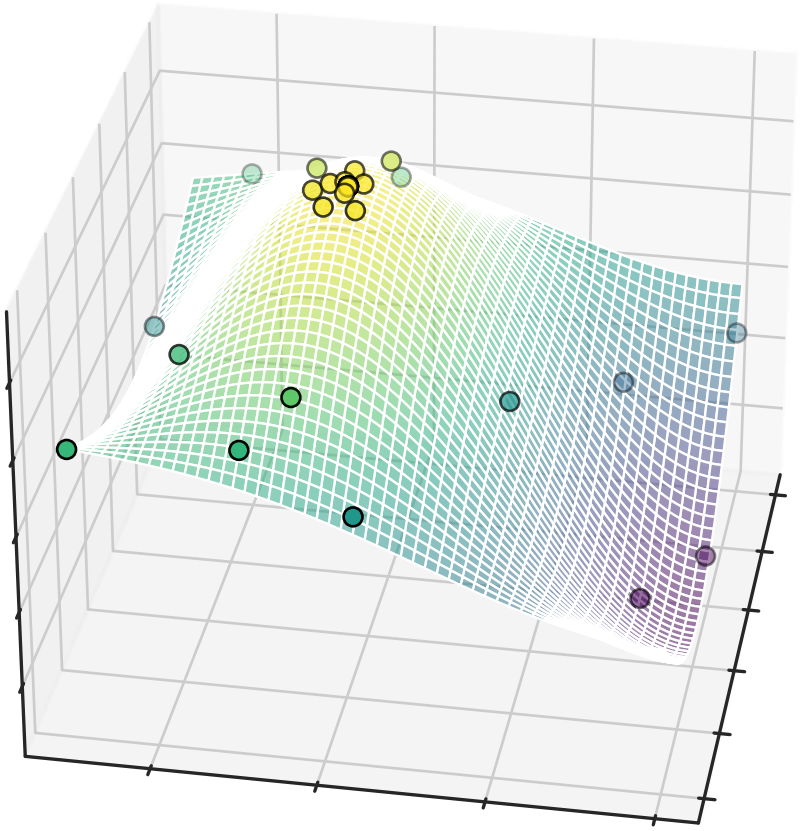


Figura 2.12: Ejemplo de optimización bayesiana con dos hiperparámetros

- La optimización "clásica" no genera un modelo de probabilidad, sino que poco a poco van buscando los mejores hiperparámetros mediante diferentes métodos:
  - El aprendizaje reforzado enfoca el problema como una serie de decisiones en un juego, y entrena el modelo para funcionar bien en dicho juego.
  - Los algoritmos genéticos crean una población de modelos que se van hibridando con un mecanismo similar a la selección natural.
  - El descenso de gradiente mueve los hiperparámetros, paso a paso, en la dirección en la que parece que se obtiene un mejor resultado.

2.6. Algoritmos de ML

Como ya se ha señalado previamente, esta tesis no pretende desgranar pormenorizadamente los algoritmos existentes. Sin embargo, se realiza a continuación una introducción a los paradigmas más importantes, lo que permite tener una idea de su funcionamiento y aplicabilidad en la estimación del perfil penetrométrico del suelo.

2.6.1. Modelos bayesianos

Los métodos bayesianos de clasificación Naïve Bayes se basan en el teorema de Bayes (Ecuación 2.15), con la hipótesis simplificadora de suponer independientes las características. De esta simplificación viene el apelativo de *naïve* (ingenuo) usado por primera vez por Kononenko (Kononenko, 1991), aunque dichos métodos venían utilizándose en aprendizaje automático desde pocos años antes (Cestnik et al., 1987).

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Ecuación 2.15: Teorema de Bayes

Donde:

$P(A|B)$ *es la probabilidad de A habiéndose dado B.*

$P(B|A)$ es la probabilidad de B habiéndose dado A.

$P(A)$ es la probabilidad de A.

$P(B)$ es la probabilidad de B.

Las diferentes variantes de los clasificadores Naive Bayes difieren principalmente en la distribución que asumen para $P(B|A)$.

Estos métodos calculan, para cada clase y característica, la distribución de probabilidad. Para clasificar una muestra se calcula la probabilidad de que se encuentre en cada una de las clases posibles suponiendo las características independientes entre sí. En la Figura 2.13 se muestra cómo se aproxima la distribución de dos clases a lo largo de los valores de una característica mediante campanas de Gauss. A partir de esa simplificación podemos obtener la probabilidad a través del teorema de Bayes.

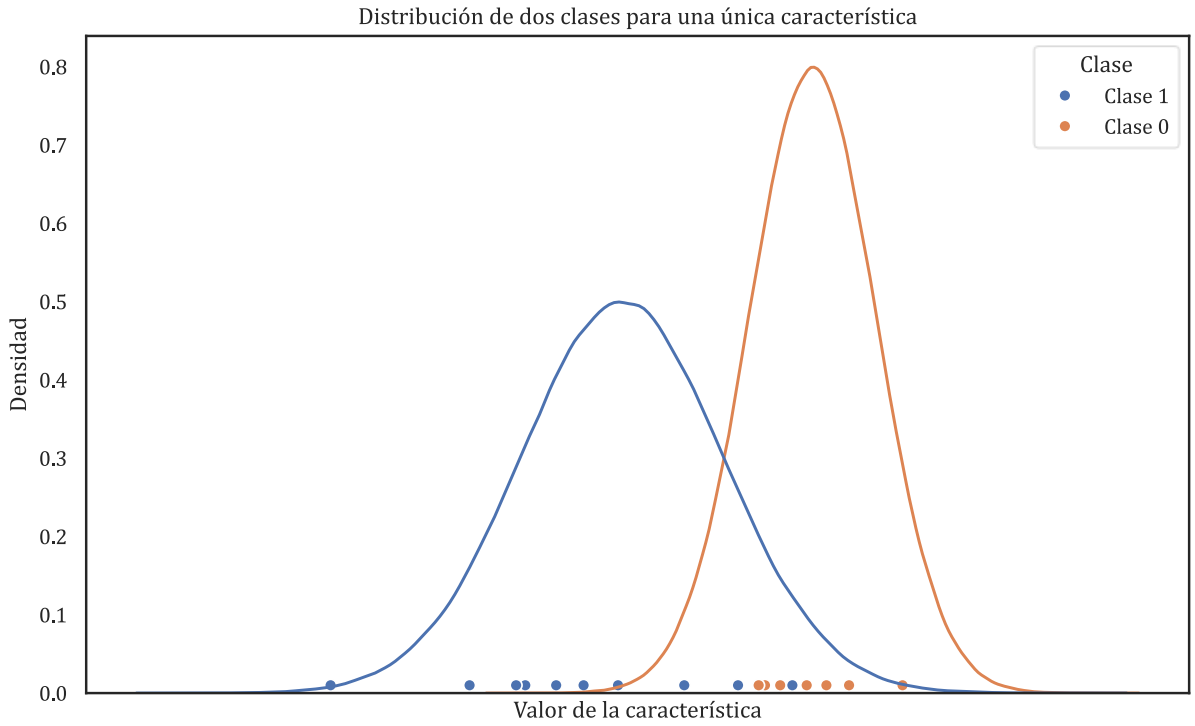


Figura 2.13: Aproximación gaussiana de la distribución de dos clases para una característica

A pesar de ser un algoritmo sencillo y con importantes hipótesis simplificadoras que no se suelen cumplir (rara vez las características son totalmente independientes entre sí), esto no suele ser un problema en la práctica (H. Zhang, 2004). A cambio tenemos un algoritmo muy eficiente que necesita pocas muestras para ser entrenado, lo que lo hace muy interesante cuando existen un gran número de datos o características.

2.6.2. Modelos lineales

Los modelos lineales asumen que existe una relación lineal entre la variable objetivo y las características. En el caso de tener una única característica, estos modelos trazarían una línea que minimice el error de la predicción (Figura 2.14). Igualmente pueden generalizarse a un mayor número de características, siendo el resultado un plano o hiperplano.

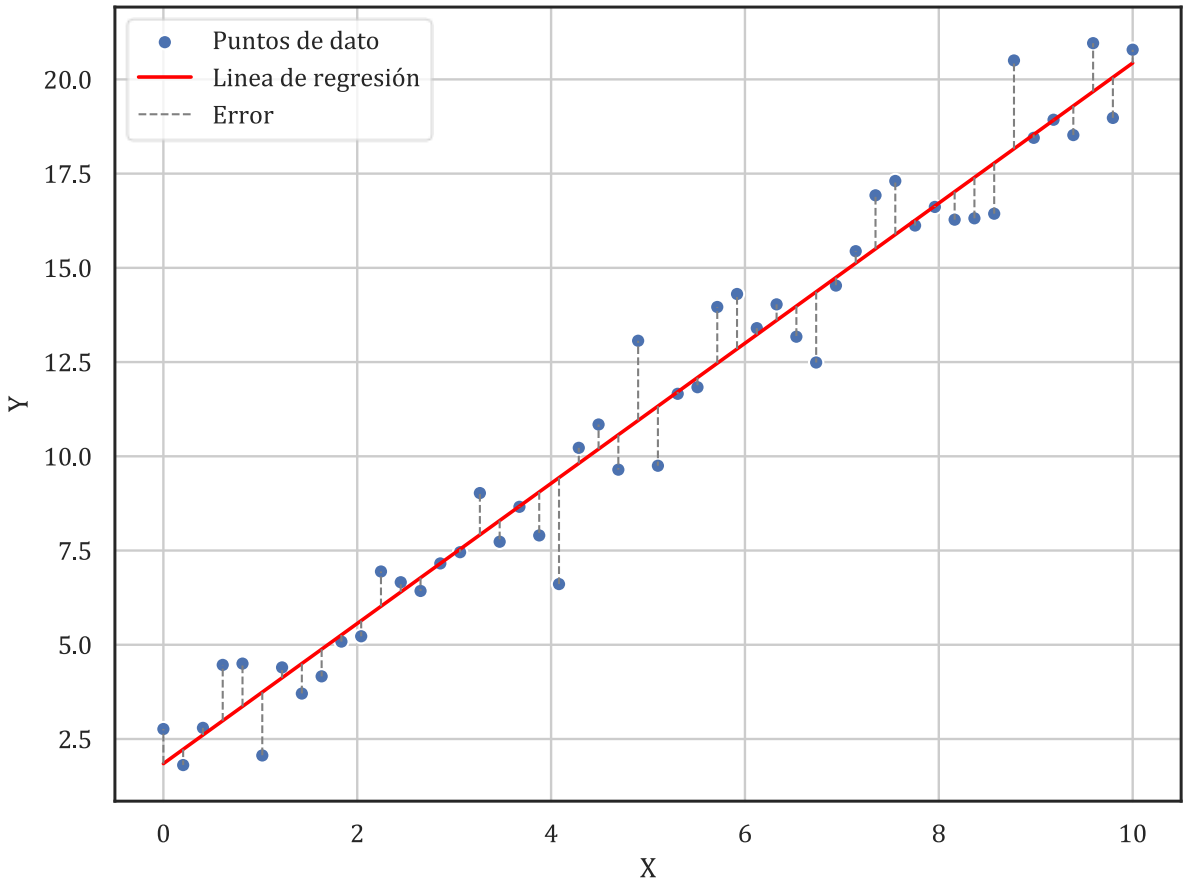


Figura 2.14: Ejemplo de regresión lineal

Esta clase de modelos son sencillos y su primera utilización nos retrotrae hasta Sir Francis Galton en el siglo XIX (Stanton, 2001). En el caso más habitual de regresión lineal se resuelve un problema de la siguiente forma:

$$\min_{\omega} ||X\omega - y||_2^2$$

Ecuación 2.16: Minimización del error cuadrático

Es decir, se busca encontrar una línea que minimiza la suma del error cuadrático.

Estos modelos son rígidos y con un alto sesgo, pero pueden funcionar muy bien cuando existen relaciones lineales entre las variables.

La colinealidad entra las variables independientes es un problema general para los modelos de aprendizaje automático y en especial para la regresión lineal. Para paliar este problema se utilizan penalizaciones directamente proporcionales a los

coeficientes ω (regresión lasso (Tibshirani, 1996)), proporcionales al cuadrado de dichos coeficientes (regresión ridge (Hoerl & Kennard, 2000)) o ambas (elastic net (Zou & Hastie, 2005)).

Otro modelo lineal, aunque en un principio pudiera no parecerlo es la regresión logística. La regresión logística es, a pesar de su nombre, un modelo lineal para clasificación en lugar de regresión. La función logística, con ese nombre, fue propuesta por Verhulst a mediados del siglo XIX, aunque tuvo que ser redescubierta en la década de los 20 del siglo pasado. El desarrollo del modelo logístico puede atribuirse al trabajo de Berkson desde 1944 en adelante (Cramer, 2003).

En este modelo, se ajusta una función logística para predecir a qué clase pertenecen las muestras. En la Figura 2.15 se muestra una función que clasifica las muestras en clase 1 y clase 0 en función del valor de la característica X, obteniendo también la probabilidad de pertenecer a la clase 1.

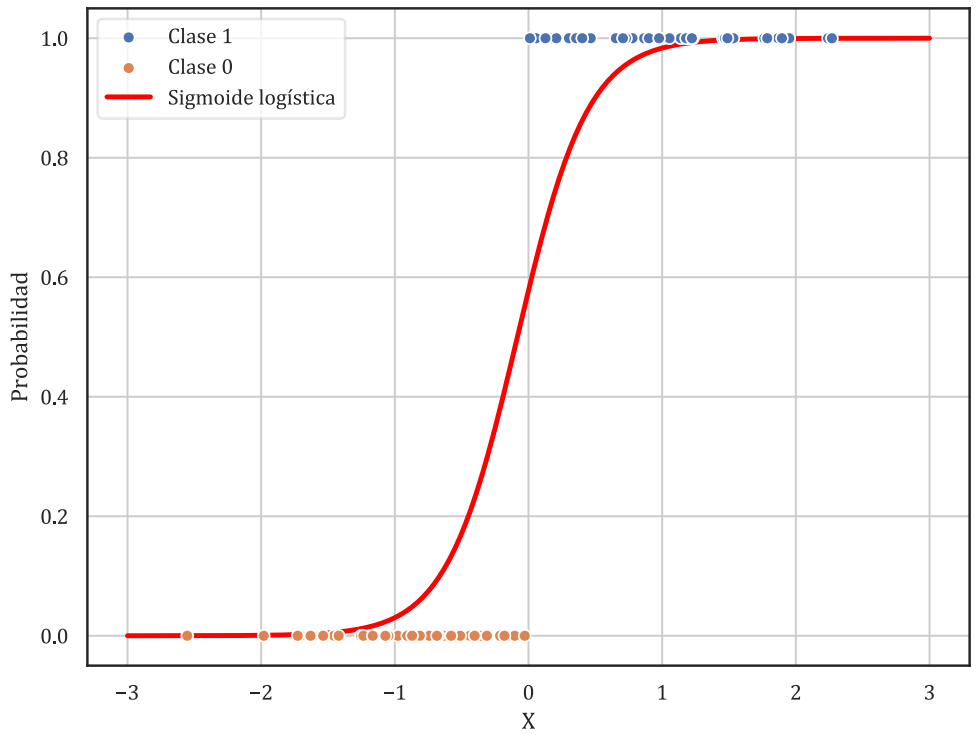


Figura 2.15: Sigmoide logística

Esta gráfica pudiera no parecer muy lineal, ya que es la función sigmoide (Ecuación 2.17), pero lo cierto es que el problema a resolver sí que lo es. Esto es así porque el exponente z es una combinación lineal de las características (Ecuación 2.18).

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Ecuación 2.17: Función sigmoide.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Ecuación 2.18: Exponente z.

Es posible también generalizar el problema para un mayor número de clases mediante diferentes estrategias (*one vs one*, *one vs rest*, método multinomial).

Podría considerarse que las máquinas de vectores de soporte también son modelos lineales, pero al funcionar de manera diferente se tratan en su propio apartado.

2.6.3. Modelos de árboles de decisión

Los árboles de decisión son un tipo de algoritmo que divide el conjunto de datos en sucesivos subconjuntos siguiendo unas reglas preestablecidas. Este proceso de división en subnodos puede ser representado en una estructura que asemeja un árbol (Figura 2.16).

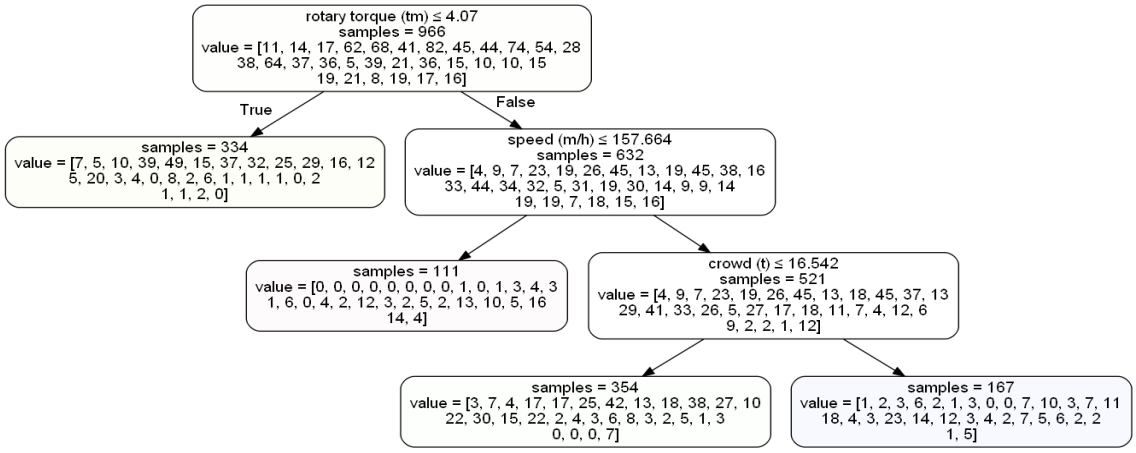


Figura 2.16: Ejemplo de árbol de decisión limitado a 4 nodos hoja

Un árbol de decisión partirá de un nodo inicial, que contendrá todas las muestras disponibles durante el entrenamiento, denominado *nodo raíz*. Este es dividido en

2 o más nodos, llamados *nodos de decisión* o *nodos internos*. Los nodos finales, que no presentan más subdivisiones, se denominan *nodos hoja* y representan la salida o respuesta del árbol para la secuencia seguida.

El origen de esta clase de algoritmos, al menos en su vertiente de aprendizaje automático, podríamos establecerla según (Loh, 2014), en el trabajo de Morgan y Sonquist (Morgan & Sonquist, 1963). Aunque quizá fueron los denominados árboles de clasificación y regresión (*Classification and Regression Trees*, CART (Breiman et al., 1984)) los que acabaron popularizándolo.

En este tiempo, los árboles de decisión han resultado ser algoritmos muy exitosos, lo que se explica por una serie de factores que los hace bastante útiles en la práctica (Louppe, 2014):

- Los árboles de decisión son no-paramétricos. No asumen la forma en la que se relacionan las variables, con lo que son capaces de modelizar interacciones complejas.
- Manejan datos heterogéneos (números, categorías o una mezcla de ambos).
- En su propio funcionamiento, los árboles de decisión implementan la elección de las variables más importantes, haciéndolos robustos (hasta cierto punto) frente a variables irrelevantes.
- Son igualmente robustos frente a valores atípicos y errores en la ordenación.
- Requieren escasa preparación previa de los datos.
- Son fáciles de interpretar, incluso para usuarios sin conocimientos previos de estadística. Pueden graficarse.
- Pueden trabajar con varias variables objetivo.

Sin embargo, también presentan una serie de factores que deben tenerse en cuenta:

- Los algoritmos tipo árbol de decisión tienden a tener un sesgo bajo y una alta varianza, es decir, tienden al sobreajuste.
- Las predicciones de los árboles de decisión no son continuas, no siendo buenos a la hora de extrapolar.
- Los algoritmos tipo árbol se entrenan a partir de óptimos locales, que no tienen por qué conducir al óptimo global.

Parte de las desventajas de los árboles de decisión se limitan realizando un adecuado proceso de poda que les permita generalizar. Sin embargo, otro posible enfoque es el uso de algoritmos ensamblados ya sea en paralelo o en serie.

2.6.4. Modelos ensamblados

Como se ha señalado en el apartado anterior, los árboles de decisión tienden al sobreajuste. Es decir, suelen adaptarse demasiado al conjunto de entrenamiento. Pero ¿Y si agregamos los resultados de varios árboles? Si cada árbol se entrena en un conjunto de entrenamiento diferente, los resultados agregados ya no tenderán tanto al sobreajuste.

Este es precisamente el enfoque de los modelos ensamblados ya que es posible demostrar que al ensamblar entre si modelos aleatorios se reduce el error de predicción disminuyendo la varianza (Louppe, 2014). Los árboles de decisión no son los únicos algoritmos que pueden ensamblarse, pero quizá han sido los más populares.

Los métodos de ensamblado pueden separarse en dos tipos, *bagging* y *boosting*.

2.6.4.1. Bagging

Bagging proviene de *Bootstrap Aggregating* y es un término acuñado por Breiman (Breiman, 1994). En estadística, *boostrapping* es un método de remuestreo utilizado para estimar la distribución de una métrica (como la media o la varianza) mediante muestreo con reemplazo (Efron, 1979). Podemos utilizar este proceso para generar diferentes conjuntos a partir de los datos de partida, entrenar predictores (como árboles de decisión) en cada conjunto, y luego agregar los resultados (por ejemplo, mediante la media).

Siguiendo esta estrategia el algoritmo más conocido es *Random Forests* (bosques aleatorios) (Breiman, 2001). Este algoritmo, además de utilizar *bagging* con árboles de decisión, no tiene en cuenta en cuenta todas las características al dividir los nodos de los árboles. Al obligar en cada división a considerar sólo un subconjunto de características, se fuerza a los árboles a ser diferentes los unos de los otros y por tanto tendentes a no estar correlacionados.

Una variante de los bosques aleatorios serían los denominados *Extra Trees* (Geurts et al., 2006). En este algoritmo, aunque se agregan los resultados de diferentes árboles, no se realiza normalmente *bootstrapping*. En cambio, a la hora de dividir un nodo, en vez de comprobar cuál es la mejor característica posible, se realiza la

división en una característica aleatoria. Esto hace que, para un mismo número de árboles, *Extra Trees* funcione peor que *Random Forests* pero con un tiempo de cálculo mucho más rápido.

2.6.4.2. Boosting

Mediante *bagging* los árboles se entrenan en paralelo y de manera independiente. Con *boosting* (Friedman, 2001) los árboles se entrenan en serie y en vez de ser independientes entre sí, cada nuevo árbol agregado al modelo se entrena para corregir los errores anteriores del conjunto. Este proceso gradual e iterativo se denomina *slow learning* ya que el modelo aprende poco a poco de sus errores.

El *slow learning* permite que el algoritmo explore patrones más complicados y relaciones más complejas entre las características, lo que dará como resultado un modelo más preciso y robusto.

Esta clase de algoritmos han tenido un tremendo éxito siendo quizá el más conocido *eXtreme Gradient Boosting*, XGBoost (T. Chen & Guestrin, 2016).

2.6.5. Modelos de vecinos

K-vecinos más cercanos (*K-Nearest Neighbors*, KNN (Cover & Hart, 1967)) es un algoritmo que puede utilizarse tanto para regresión como para clasificación. Se basa en el concepto de distancia, esto es, los puntos que presentan características cercanas pertenecerán a la misma clase o tendrán un valor objetivo similar.

En la Figura 2.17 se intenta predecir la clase de la muestra representada por un cuadrado verde observando cinco vecinos. De esos vecinos, cuatro pertenecen a la clase 1, y uno pertenece a la clase 2, por lo que lo más probable es que la muestra pertenezca a la clase 1. En caso de estar considerando un problema de regresión, el valor de la muestra sería una media ponderada de los valores de los vecinos.

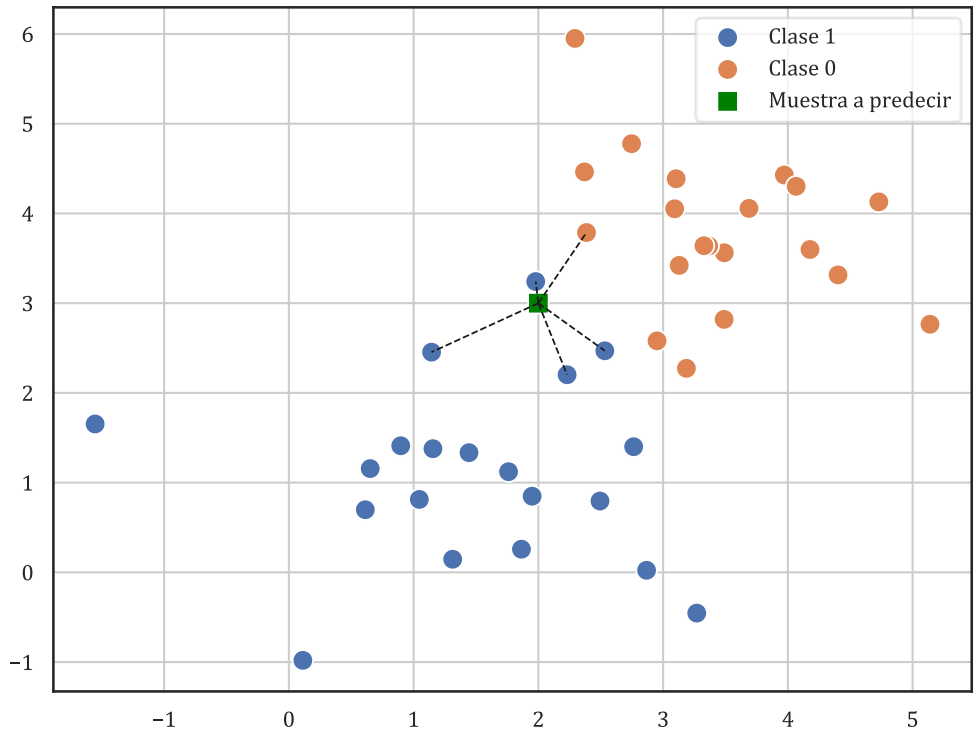


Figura 2.17: Funcionamiento del algoritmo KNN

En el caso de KNN no hay un verdadero aprendizaje, no se genera un modelo. El algoritmo se limita a comprobar los k-vecinos más cercanos cada vez que queremos predecir el valor o la clase de una muestra. Esto puede provocar, para conjuntos de datos grandes, que KNN sea lento de calcular. Sin embargo, sigue siendo un algoritmo muy sencillo de aplicar y entender, y puede funcionar muy bien cuando los datos son fácilmente clasificables por zonas.

2.6.6. Modelos de máquinas de vectores de soporte

Las máquinas de vectores de soporte (*Support Vector Machines*, SVM) son un conjunto de métodos basados en separar las muestras mediante hiperplanos. La forma actual de estos algoritmos puede rastrearse hasta la *Fifth Annual Workshop on Computational Learning Theory* (Boser et al., 1992).

Si imaginamos un conjunto de datos que son linealmente separables, será posible trazar infinitos hiperplanos que separen ambas clases, Figura 2.18.

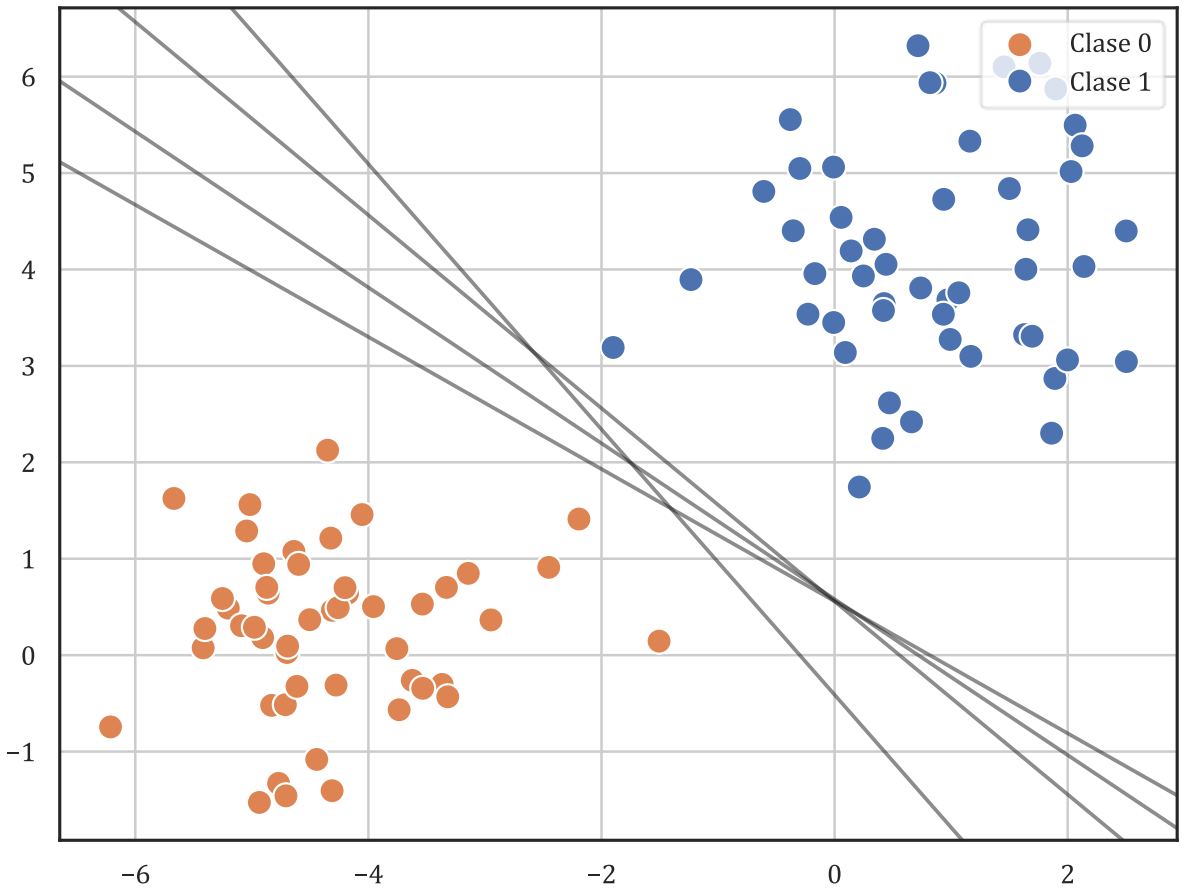


Figura 2.18: Posibles hiperplanos que separan dos clases

Es por tanto necesario determinar cuál hiperplano utilizar. Una elección natural sería el hiperplano de máximo margen, aquel que está más lejos de las muestras de entrenamiento. El hiperplano de máximo margen será aquel que presente la mayor distancia con las muestras. Se espera que aquel hiperplano con mayor margen a las muestras de entrenamiento también tenga un margen amplio respecto a las muestras de validación.

En la Figura 2.19 se observa que hay tres muestras (dos pertenecientes a la clase 1, una perteneciente a la clase 2) que son equidistantes al hiperplano de máximo margen. Estas observaciones se conocen como los vectores de soporte (*support vectors*) y «soportan» al hiperplano en el sentido de que, si estos puntos se mueven, también lo hará el hiperplano. Cualquier movimiento de las muestras que no entre dentro del margen definido por los vectores de soporte no provocará un cambio en el hiperplano.

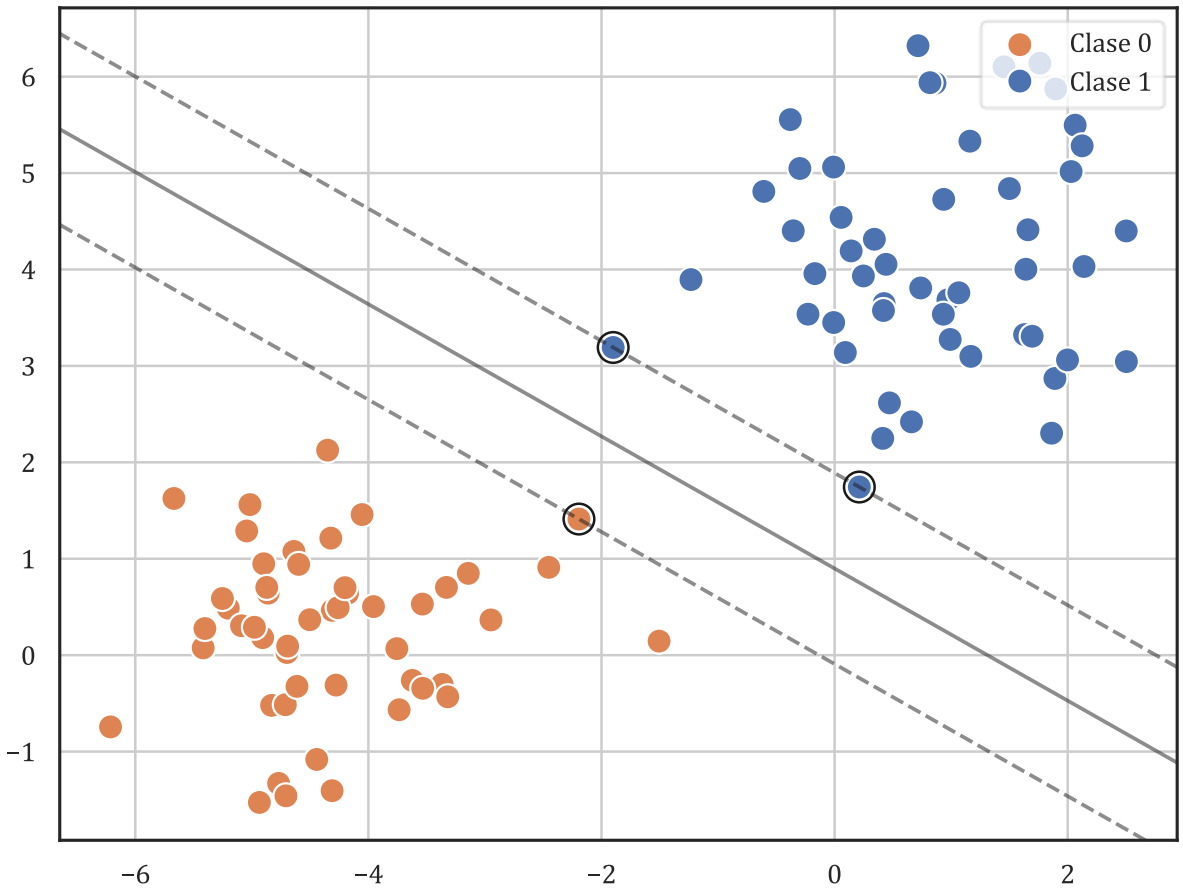


Figura 2.19: Vectores de soporte

Pero el concepto de hiperplano puede flexibilizarse permitiendo que algunas muestras queden a un lado distinto al cual les correspondería. Esta estrategia puede ser interesante por dos razones: evitar un sobreajuste excesivo a las muestras de entrenamiento y permitir crear un hiperplano con clases, en principio, no separables.

Aun permitiendo establecer hiperplanos permeables por algunas muestras, el algoritmo aún estaría limitado por la naturaleza lineal del hiperplano, por ejemplo, en el caso que una clase esté recogida "dentro" de otra.

Ante este problema podemos recurrir al uso de *kernels*; estos son funciones que ajustan los datos a un espacio de mayor dimensión, de tal manera que las clases sí sean linealmente separables. En la Figura 2.20 se muestra como dos clases concéntricas pueden ser fácilmente separables utilizando un kernel de base radial.

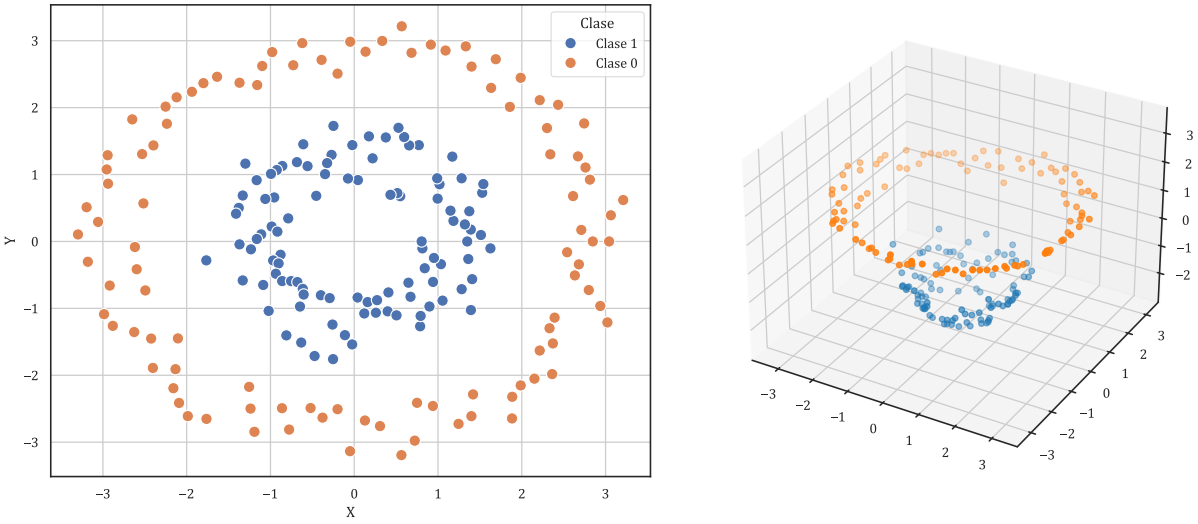


Figura 2.20: Uso de kernel radial. A la izquierda se muestran dos clases que no serían separables en el espacio bidimensional. A la derecha estas dos clases son fácilmente separables en el espacio tridimensional tras el uso de un kernel radial

2.6.7. Redes neuronales

Las redes neuronales son un tipo de algoritmo inspirado en el cerebro. La idea es que múltiples neuronas realizando tareas sencillas pueden llegar a resolver problemas realmente complejos.

El perceptrón es un modelo básico de neurona artificial desarrollado por Rosenblatt en 1957 (Rosenblatt, 1957). Dicha neurona artificial estaría formada por:

- Entradas y pesos asociados a estas.
- Una función de activación.
- Un algoritmo que ajusta los pesos para adaptarse al resultado deseado.

Actualmente conocemos la enorme potencia de las redes neuronales, sin embargo, los perceptrones recibieron poca atención durante años debido a la publicación de *Perceptrons* por Minsky y Papert (Minsky & Papert, 1969). En dicho trabajo se enfatizaban las limitaciones de lo que un único perceptrón podía conseguir. Si bien los perceptrones multicapa no formaban parte del análisis, en aquel tiempo aún se desconocía como entrenar dichos conjuntos de perceptrones. No fue hasta el trabajo de Rumelhart *et al.* sobre *backpropagation* que se inició de manera más activa el desarrollo de los perceptrones multicapa (Rumelhart et al., 1986).

Aunque pueden existir otros tipos de redes neuronales, en este apartado llamaremos así a los perceptrones multicapa. Una red neuronal está compuesta de tres partes; una capa de entrada, una o más capas ocultas, y una capa de salida. La capa de entrada está formada por una neurona para cada característica. El número de capas ocultas y el número de neuronas por capa se eligen al definir el modelo. Finalmente, la capa de salida consiste en una neurona por cada variable a predecir. Todas las neuronas están conectadas con las neuronas de las capas precedentes (Figura 2.21).

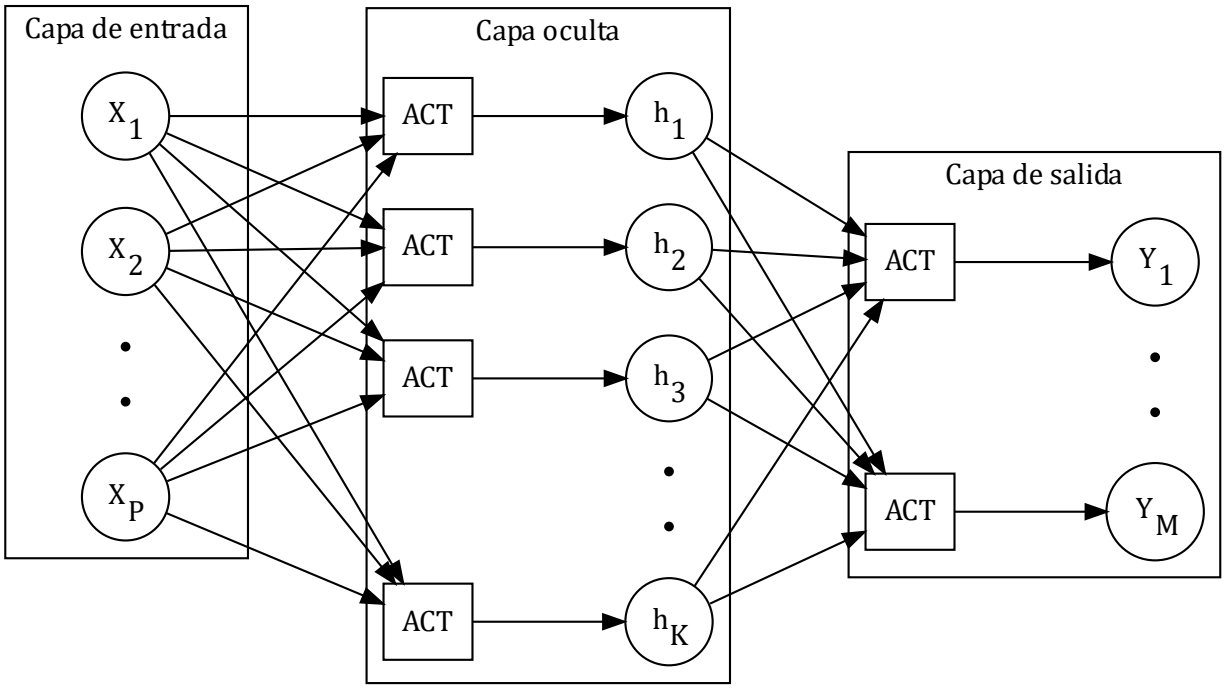


Figura 2.21: Esquema de red neuronal con capa de entrada con *P* variables, una única capa oculta con *K* neuronas y capa de salida con *M* variables a predecir

Consideremos una red neuronal de una sola capa oculta con *P* variables, *K* neuronas y *M* variables a predecir. Los datos en cada nodo se calculan como la suma de todos los nodos en las capas precedentes multiplicados por unos pesos y sumando una constante, llamada sesgo. El resultado pasa por una transformación no lineal mediante una determinada función de activación. Esta función es la que permite al modelo aprender patrones complejos y no lineales. De no existir, el modelo sería lineal, ya que la suma de transformaciones lineales es igual a una

única transformación lineal. Los rectángulos de la Figura 2.21 representan estas funciones de activación.

Podemos describir la salida m de una única capa oculta con P variables de entrada y K neuronas como:

$$h_m(x) = g(\beta_0 + \sum_{k=1}^{K} \beta_k h_k(x))$$

Ecuación 2.19: Salida de una neurona

Donde x es el valor de la variable de entrada, β_0 es el término de sesgo, $\beta_1 \dots \beta_k$ son los pesos de las capas ocultas y $g(z)$ es la función de activación. El proceso de entrenamiento del modelo consiste en encontrar valores para los pesos y el sesgo utilizando el conjunto de entrenamiento.

En el caso de que la red neuronal tenga un mayor número de capas ocultas hablaríamos de aprendizaje profundo (*Deep Learning*).

2.7. Preprocesamiento

El preprocesamiento de los datos es un paso crucial en el flujo de trabajo del aprendizaje automático. El preprocesamiento puede estar constituido por varios y diversos pasos, por ejemplo, es necesario verificar datos faltantes o con errores. Sin embargo, en este apartado nos referimos a procesos matemáticos que se aplican posteriormente a la limpieza de los datos.

Los algoritmos no dejan de ser un conjunto de fórmulas matemáticas y pueden beneficiarse de ciertos ajustes previos en los datos de entrada para optimizarlos de cara al cálculo. En algunos casos resulta incluso obligatorio realizar cierto preprocesamiento para conseguir resultados fiables. Por ejemplo, escalar las diferentes variables en los algoritmos basados en distancia.

Algunos de estos métodos de preprocesamiento se resumen a continuación.

2.7.1. Binarización

La binarización consiste en convertir una variable en booleana (es decir, en 0s y 1s) a partir de un determinado valor umbral. Esta técnica puede simplificar los

datos y mejorar la eficiencia de ciertos algoritmos. Además, permite tener en cuentas variables categóricas binarias (p.ej. noche y día) con algoritmos que sólo trabajan con números.

2.7.2. Análisis de componentes independientes

El análisis de componentes independientes (*Independent Component Analysis*, ICA) (Comon, 1994) es una técnica estadística capaz de descomponer una señal multivariable en subcomponentes aditivos independientes.

En el contexto de preprocesamiento en ML, ICA puede mejorar la calidad de los datos ya que, al separar componentes independientes, se pueden obtener características más limpias y relevantes. Esto podría mejorar el rendimiento de los modelos.

2.7.3. Análisis de componentes principales

El análisis de componentes principales (*Principal Component Analysis*, PCA) (Pearson, 1901) se utiliza para descomponer un conjunto de datos multivariable en sucesivos conjuntos ortogonales que expliquen la máxima cantidad de varianza.

PCA permite reducir el número de características manteniendo el máximo de varianza posible y, por tanto, la mayor parte de la información. Por ejemplo, en la Figura 2.22, se muestra a la izquierda un conjunto de dos características. Sin embargo, observamos que la mayor parte de la variación se produce en una cierta dirección (componente principal 1, PC1) mientras que apenas hay dispersión en la dirección perpendicular. En este caso podríamos tener en cuenta una única característica, PC1, en lugar de las dos originales.

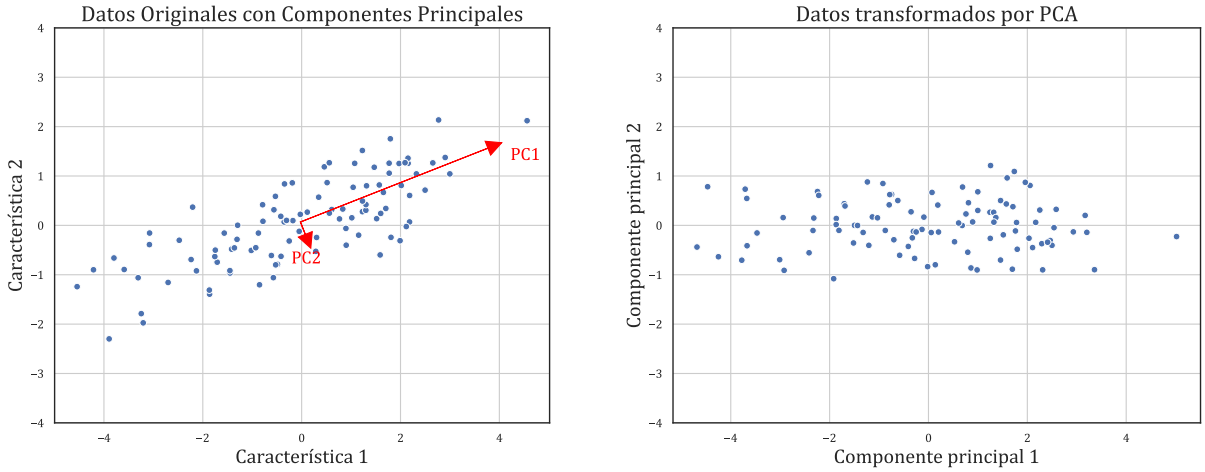


Figura 2.22: Ejemplo de funcionamiento de PCA

2.7.4. Aglomeración de características

La aglomeración de características es una técnica que consiste en agrupar variables similares entre sí, formando nuevas características agregadas. El objetivo es reducir la dimensionalidad manteniendo la mayor parte de la información.

La técnica consiste en la agrupación de características similares en función de métricas de distancia. Este proceso puede continuar hasta alcanzar un cierto número de grupos (o hasta que sólo quede un único grupo). Podemos visualizar estas agrupaciones en un dendrograma (Figura 2.23).

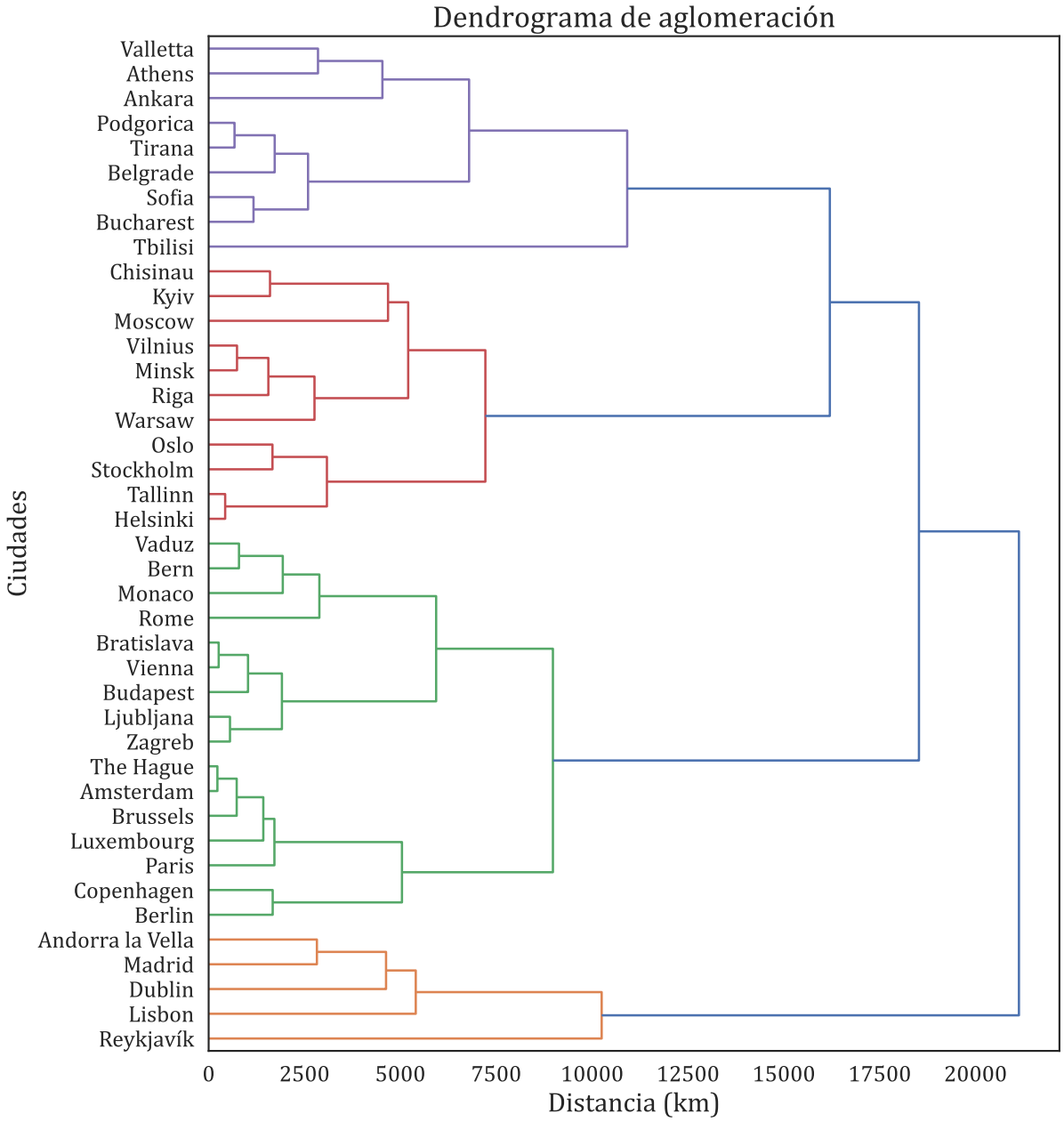


Figura 2.23: Dendrograma de aglomeración de las capitales europeas en función de la distancia entre ellas. Dependiendo de qué distancia observemos encontraremos un distinto número de agrupaciones

Este proceso consigue reducir la dimensionalidad siendo especialmente útil ante características redundantes o altamente correlacionadas.

2.7.5. Escalar las características

Como ya se ha señalado previamente, algunos algoritmos necesitan, por su funcionamiento, el que todas las características se muevan en el mismo rango de

valores. Este es el caso, por ejemplo, de los algoritmos basados en distancia (KNN, SVM) en los que, si las características tienen diferentes escalas, las de mayor rango pueden dominar la distancia, sesgando los resultados.

Existen diversos métodos para escalar las características:

- **Estandarización**: consiste en restar la media y dividir entre la desviación típica de los valores para cada característica.
- **Escalar a mínimo y máximo**: se reescalan los valores de tal manera que el máximo y el mínimo coinciden con un valor preestablecido, generalmente 1 y 0.
- **Escalador robusto**: este escalador resta la mediana y escala los datos según un rango de cuantiles, normalmente el rango intercuartil (IQR). El IQR es el rango entre el 1er cuartil (cuantil 25) y el 3er cuartil (cuantil 75).

2.7.6. Normalización

La normalización cambia la escala del vector de cada muestra para que tenga una magnitud unitaria, independientemente de la distribución de las muestras. Al igual que la estandarización, este proceso puede mejorar el rendimiento de algoritmos sensibles a la escala de las muestras.

2.7.7. Método Nyström

El método Nyström es una técnica de aproximación utilizada en el preprocesamiento de ML, especialmente en el contexto de la reducción de dimensionalidad y el escalado de algoritmos que dependen de la manipulación de grandes matrices de kernel (Williams & Seeger, 2000). El método aproxima matrices de kernel grandes y complejas mediante una versión más pequeña y computacionalmente eficiente. Esto permite a algoritmos de kernel (como SVM y PCA) escalar a conjuntos de datos más grandes.

2.7.8. Muestreo RBF

Al igual que el método anterior, el muestreo RBF (*Radial Basis Function*) es una técnica de aproximación de kernel. En el caso del muestreo RBF se proyectan los datos originales en un espacio de características de mayor dimensión utilizando funciones de base radial (Rahimi & Recht, 2007). Esta técnica, al igual que para el

método Nyström, permite funcionar de manera más eficiente a algoritmos de kernel.

2.7.9. Características polinómicas

Al tratar con relaciones no lineales en un conjunto de datos puede ser útil añadir complejidad a las características. Por ejemplo, para una muestra bidimensional [a,b], las características polinómicas de grado 2 serían [1, a, b, a^2, ab, b^2]. Estas características pueden ayudar a descubrir relaciones complejas en los datos, permitiendo así que los modelos hagan mejores predicciones. Además, permite que modelos sencillos, como la regresión lineal, puedan encontrar relaciones complejas entre las características.

2.7.10. Codificación one-hot

La codificación *one-hot* es un proceso utilizado para convertir características de datos categóricos en una serie de características binarias. Por ejemplo, imaginemos la categoría «color» que toma los valores Rojo, Amarillo y Azul. La codificación *one-hot* convertiría la categoría «color» en las categorías «Rojo», «Amarillo» y «Azul» con valor de 1, si ese era el valor de la característica, o valor de 0 en caso contrario (Tabla 2.2).

Tabla 2.2: Ejemplo de codificación *one-hot*

Característica original	**Características one-hot**		
Color	Rojo	Amarillo	Azul
Rojo	1	0	0
Amarillo	0	1	0
Azul	0	0	1
Rojo	1	0	0

2.7.11. Selectores

Los selectores son otra parte del preprocesamiento y pueden ser de dos tipos: selectores de características y selectores de muestras. Nos centraremos en los selectores de características en este apartado.

Elegir adecuadamente las características que hay que tener en cuenta puede ayudarnos a reducir la dimensionalidad y a trabajar de manera más eficiente.

Algunos de los métodos que podemos utilizar al seleccionar las características son las siguientes:

- **Eliminación de variables con baja varianza**. Una táctica sencilla es remover aquellas características que no cumplan con un mínimo de varianza. La idea subyacente es que características que apenas varíen serán poco explicativas de la variable objetivo.
- **Selección de características univariante**. Esta técnica busca detectar las características más informativas, aquellas que tienen una mayor relación con la variable objetivo. Opera considerando cada característica de forma independiente, evaluando su relación con la variable objetivo en función de ciertas pruebas estadísticas.
- **Eliminación recursiva de características**. La eliminación recursiva de características (*Recursive Feature Elimination*, RFE), funciona entrenando un modelo con el conjunto de características completo, eliminando aquellas que presenten una menor importancia. Ese procedimiento se repite recursivamente hasta que se alcanza el número deseado de características.
- **Selección desde modelo**. Un enfoque similar al anterior sería, tras entrenar un modelo, elimina todas aquellas características con una importancia menor a un cierto umbral. Al no funcionar de manera iterativa, las características eliminadas de esta manera no tienen por qué coincidir con aquellas eliminadas mediante RFE.

2.7.12. AutoML

Cómo se ha podido comprobar, en los apartados anteriores, existen diversos algoritmos entre los que elegir. Además, dentro de cada algoritmo existen muchos hiperparámetros que ajustar para conseguir un resultado óptimo. Por si fuera poco, se pueden añadir diversos preprocesos y opciones de selección con lo que las posibilidades son innumerables.

Una posible solución ante esta miríada de opciones es automatizar el proceso para encontrar el mejor algoritmo posible. Sin embargo, probar todas las opciones sería computacionalmente inabarcable. Resulta necesario limitar el número posibilidades que se analizan, pero ¿con qué criterio?

Para encontrar los flujos (*pipelines*) más prometedores podemos acudir a alguna librería de *auto machine learning* (AutoML). Estas librerías automatizan el proceso de encontrar los mejores algoritmos.

Tree-based Pipeline Optimization Tool (TPOT) (Olson et al., 2016) es una librería que optimiza pipelines de ML utilizando programación genética (*genetic programming*, GP) (Banzhaf, 1998). La programación genética en el contexto de ML se refiere a algoritmos evolutivos inspirados en la selección natural.

Para poder integrar las diferentes partes de un *pipeline* de ML, TPOT los implementa en forma de árbol, tal y como se muestra en el ejemplo de la Figura 2.24. Estos *pipelines* se generan de manera automática mediante GP según la implementación de la librería DEAP (Fortin, 2012).

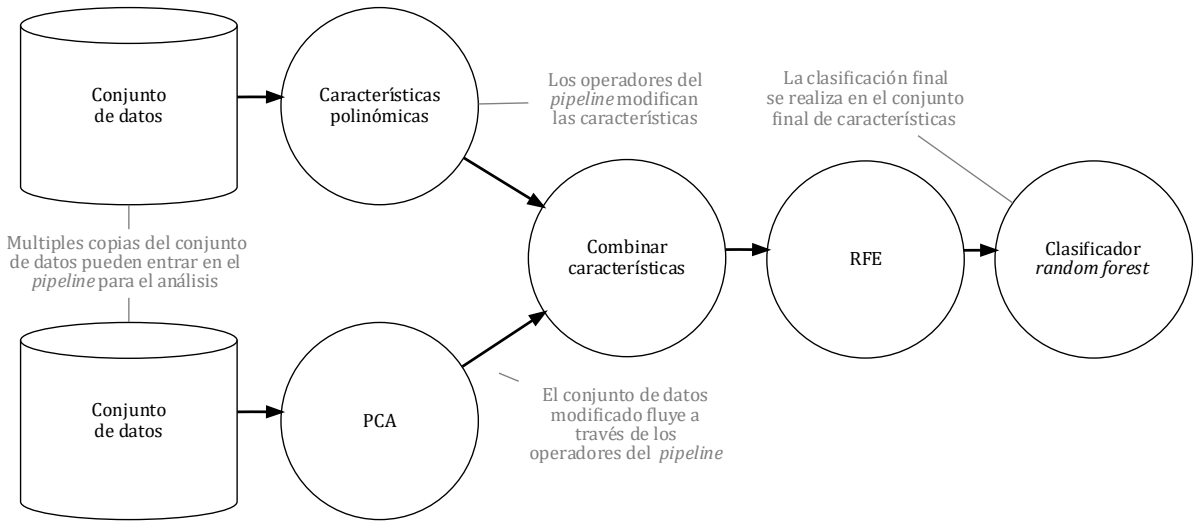


Figura 2.24: Ejemplo de *pipeline* utilizado en TPOT (Olson et al., 2016)

GP es un tipo de ML que evoluciona un programa para realizar una tarea específica inspirándose en la selección natural. El proceso sería el siguiente:

- **Población inicial**. Se parte de una población inicial que se genera aleatoriamente. En el caso de TPOT se generarían una serie de *pipelines*.
- **Evaluación de la población.** La población se evalúa mediante una métrica preestablecida para comprobar qué miembros de esta han funcionado mejor.
- **Selección.** Se seleccionan una serie de miembros de la población. Los miembros con una mejor evaluación tienen una mayor probabilidad de ser escogidos. Se produce una suerte de «selección natural».

- **Cruce y mutación.** Para engendrar una nueva generación se recurre a operaciones análogas a la reproducción biológica: cruce (donde partes de dos miembros progenitores se combinan para producir un descendiente) y mutación (donde miembros de un *pipeline* se alteran al azar).
- **Nueva generación e iteración.** La nueva generación (descendientes) reemplaza a parte o toda la generación antigua. El proceso se repite desde la etapa de evaluación durante un determinado número de generaciones o hasta que emerge una solución satisfactoria en función de su evaluación.
- **Finalización.** Una vez se cumple la condición de finalización el bucle termina y se pueden evaluar los resultados.

En la Figura 2.25 se muestra el esquema de este proceso.

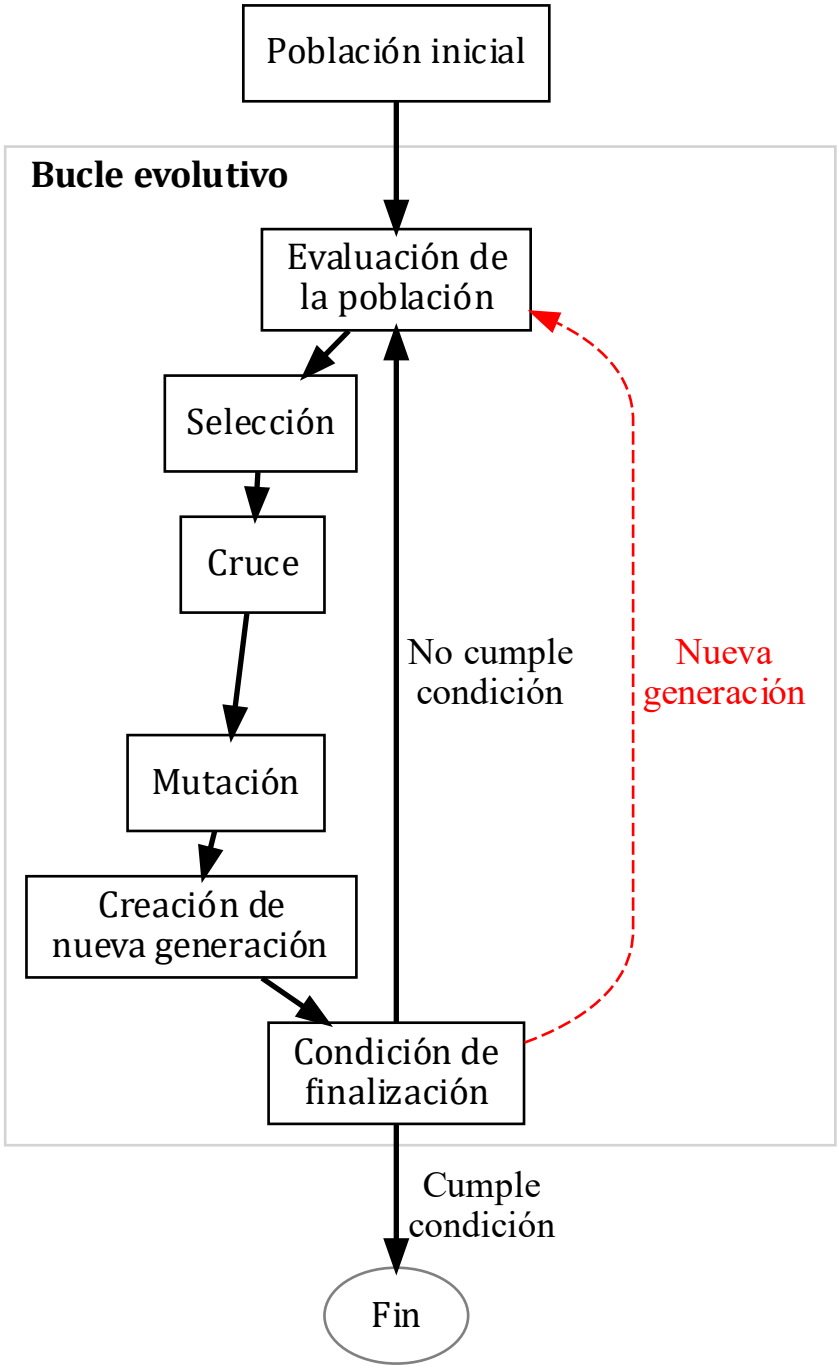


Figura 2.25: Esquema de GP

2.8. Revisión bibliográfica

En los últimos años, el uso de ML ha tenido un enorme crecimiento en multitud de ámbitos incluyendo el de la ingeniería geotécnica (Figura 2.26). En el presente apartado nos centraremos en aquellos que aplican ML a datos provenientes de

MWD. Una revisión completa sobre ML en el campo de la geotecnia está fuera del alcance de esta tesis. No obstante, recomendamos los trabajos de diversos autores.

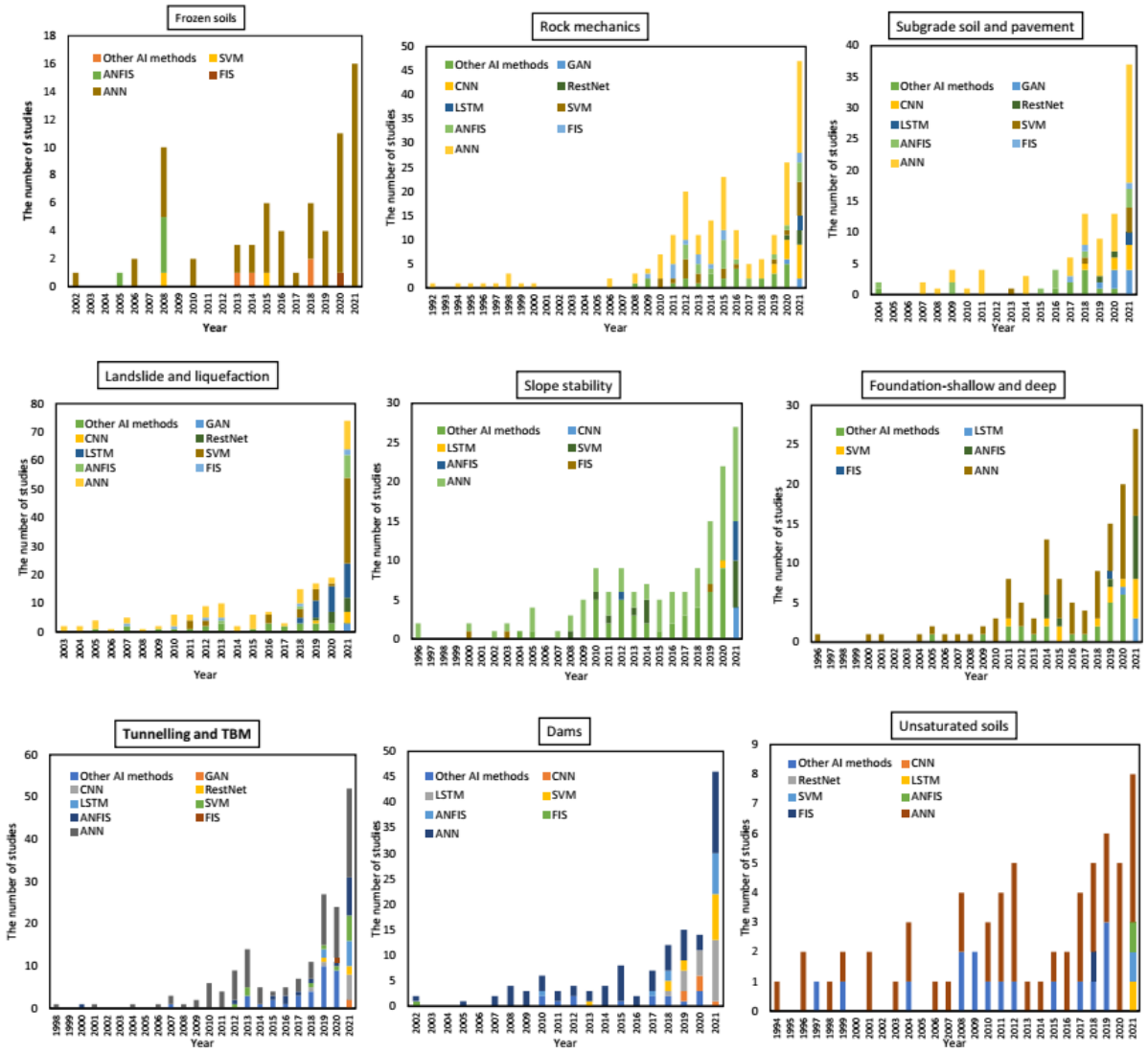


Figura 2.26: Número de artículos publicados sobre IA en diferentes ámbitos de la ingeniería geotécnica (Baghbani et al., 2022)

2.8.1. Revisiones sobre el uso de ML en el campo de la geotecnia

La *20th International Conference on Soil Mechanics and Geotechnical Engineering* (ICSMGE 2022) dedica una parte de sus actas a estados del arte y conferencias invitadas. Entre estos se encuentra el trabajo de Liu y Lacasse (Z. Liu & Lacasse, 2022) que proporciona ejemplos de aplicación de ML en caracterización, excavaciones subterráneas, diseño de túneles, mapeo de deslizamientos y

pronósticos temporales. Entre otros, se recoge el trabajo de Hansen *et al.* (Hansen et al., 2022) que utiliza datos provenientes de MWD para caracterizar el macizo rocoso durante la excavación de túneles. Concretamente se utiliza como *input* la penetración normalizada como una imagen proyectada a lo largo del perfil del túnel y los algoritmos *random forest classifier* (RFC) y *convolutional neural network* (CNN).

Por otra parte, entre los artículos generales de la ICSMGE 2022 aparece otra revisión por parte de Congress *et al.* (Congress et al., 2022) sobre la aplicación de técnicas de inteligencia artificial en las disciplinas de ingeniería geotécnica. Se recogen aplicaciones en clasificación de suelos y rocas, estabilidad de laderas, cimentaciones, licuefacción y muros de contención.

Otra interesante revisión es la que lleva a cabo Dramsch (Dramsch, 2020) ya que recoge aplicaciones, no sólo en la geotecnia, sino en toda la geociencia, remontándose las primeras a hace 70 años.

Con un enfoque similar, Ebid (Ebid, 2021) realiza una revisión de 35 años de investigaciones sobre la aplicación de IA en la ingeniería geotécnica, recopilando 626 estudios publicados entre 1984 y 2019. El autor destaca que las Redes Neuronales Artificiales (*Artificial Neural Networks*, ANN) son la técnica más utilizada, especialmente en la correlación de propiedades de suelos y rocas.

A la misma conclusión llegan otras revisiones, centrándose en las ANN. Por ejemplo, Moayedi *et al.* (Moayedi et al., 2020) realizan una revisión sistemática y un metanálisis sobre la aplicación de ANN en la ingeniería geotécnica. El estudio analiza cómo las ANN se han utilizado para afrontar problemas complejos, como la capacidad portante en pilotes, estabilidad de taludes, licuefacción o clasificación de suelos. La revisión de Chao *et al.* (Chao et al., 2018) es similar, tratando las ANN aplicadas a taludes, clasificación de suelos o asientos en cimentaciones. Juwaied (Juwaied, 2018) también revisa las aplicaciones de IA en la ingeniería geotécnica, resaltando el uso de ANN para modelar la capacidad de carga de pilotes, asentamiento de cimentaciones y licuefacción de suelos. El artículo enfatiza que las técnicas de IA manejan eficientemente múltiples variables no lineales, superando en muchos casos a los métodos tradicionales.

Otros autores han centrado sus revisiones en la capacidad portante y el asiento en pilotes. Fatehnia y Amirinia (Fatehnia & Amirinia, 2018) enfocan su revisión en el uso de ANN y la Programación Genética (GP) para estimar la capacidad portante en pilotes. Destacan cómo estas técnicas superan a los métodos tradicionales al

modelar la compleja interacción entre pilotes y suelos, ofreciendo soluciones más precisas y adaptables en la ingeniería geotécnica. Previamente Shahin (Shahin, 2016) realizaba también una revisión sobre la IA usada para estimar la capacidad portante de pilotes y calcular el asiento de estos. Las conclusiones a las que llega son similares.

Gran parte del éxito que han tenido las herramientas de IA se debe al aprendizaje profundo (*Deep learning*, DL). El DL hace referencia a redes neuronales con múltiples capas ocultas. Muchos autores han centrado sus revisiones en esta rama de la IA. Demertzis (Demertzis et al., 2023) destaca el uso de ML y DL en la clasificación de suelos, predicción de la licuefacción y modelado de daños estructurales. Zhang *et al.* (W. Zhang et al., 2021) ofrecen una revisión crítica sobre el uso de algoritmos de DL en la ingeniería geotécnica, mostrando el empleo de redes neuronales *feedforward* (FNN), recurrentes (RNN), convolucionales (CNN) y redes generativas antagónicas (GAN) en tareas como la predicción de parámetros del suelo, capacidad de carga de pilotes y mapeo de susceptibilidad a deslizamientos. Ma y Mei (Ma & Mei, 2021) revisan el uso de algoritmos de DL para el análisis de peligros geológicos, señalando el uso de datos de observación terrestre y modelos como CNN y RNN para tareas como la detección de deslizamientos, flujos de escombros, caídas de rocas y terremotos.

Khallaf y Khallaf (Khallaf & Khallaf, 2021) también realizan una revisión sobre DL pero enfocada en el proceso constructivo incluyendo la detección de grietas, seguimiento de equipos, gestión de obras y evaluación de alcantarillados.

Tanto Dikshit (Dikshit et al., 2021) como Tehrani (Tehrani et al., 2022) resaltan la aplicación de DL en el estudio de deslizamientos y otros peligros geológicos.

W. Zhang *et al.* (Tehrani et al., 2022) presentan una revisión exhaustiva sobre la aplicación de algoritmos de ML, DL y optimización en la geoingeniería y las geociencias. Analizan cómo estos métodos se han utilizado para predecir la ocurrencia de desastres geotécnicos, evaluar la seguridad de las estructuras y mejorar la precisión de las estimaciones geotécnicas. Abordan desafíos como la falta de grandes conjuntos de datos, la integración con modelos físicos y la necesidad de estándares de datos para mejorar la precisión y fiabilidad.

Las revisiones anteriores se centran, principalmente, en redes neuronales ya que siguen siendo el tipo de algoritmo más popular. Sin embargo, otros algoritmos han ido ganando popularidad en los últimos años. Por ejemplo, Jong (Jong et al., 2021) presentan una revisión sobre las técnicas de IA aplicadas a la interacción suelo-

estructura en obras subterráneas. Ponen en relieve técnicas como ANN, SVM y el método bayesiano, utilizados para predecir parámetros del suelo y comportamiento de excavaciones profundas, túneles y cimentaciones de pilotes.

Zhang *et al.* (W. Zhang et al., 2020) ofrecen una revisión del estado del arte sobre las aplicaciones de *soft computing* en excavaciones subterráneas. Subrayan técnicas como *eXtreme Gradient Boosting* (XGBoost), *Multivariate Adaptive Regression Splines* (MARS), ANN y SVM para predecir la deflexión máxima de las paredes laterales en excavaciones con soporte.

El artículo de Cabalar *et al.* (Cabalar et al., 2012) revisa las aplicaciones de los Sistemas de inferencia neuro-difuso adaptativo (*Adaptive Neuro-Fuzzy Inference System*, ANFIS) en la ingeniería geotécnica. Se señalan las aplicaciones de ANFIS en problemas como la predicción de la respuesta de cimentaciones, el análisis de estabilidad de túneles, la estimación de la permeabilidad de suelos y la predicción de la licuefacción. El estudio también presenta modelos basados en ANFIS aplicados a mezclas de arena y mica, utilizando pruebas triaxiales para modelar propiedades como el esfuerzo desviador y el módulo de corte. ANFIS muestra buenos resultados en la modelización de comportamientos complejos de suelos.

Ya hemos visto previamente como la predicción y el análisis de los riesgos geotécnicos ha suscitado gran interés para la aplicación de ANN, pero también se han utilizado otros algoritmos. Merghadi *et al.* (Merghadi et al., 2020) ofrecen una revisión comparativa de los métodos de ML para estudios de susceptibilidad a deslizamientos de tierra. Evalúan el rendimiento de varios algoritmos, como RF, SVM y Árboles de Decisión, destacando que los algoritmos basados en árboles, especialmente RF, ofrecen resultados más precisos en la cartografía de susceptibilidad. Ma *et al.* (Ma et al., 2021) revisan el uso de ML para la prevención de deslizamientos de tierra, enfocándose en la detección de deslizamientos a partir de imágenes, evaluación de la susceptibilidad y desarrollo de sistemas de alerta temprana.

Wadoux *et al.* (Wadoux et al., 2020) y P. Zhang *et al.* (P. Zhang et al., 2022) revisan la aplicación de ML a la predicción de propiedades del suelo destacándose algoritmos como RF, ANN o SVM.

Xie *et al.* (Xie et al., 2020) realizan una revisión en un ámbito muy concreto, el del mantenimiento de vías ferroviarias. Sobresalen, además del DL, los métodos de ensamblado.

El artículo de Baghbani *et al.* (Baghbani et al., 2022) es una revisión sobre la aplicación de inteligencia artificial (IA) en la ingeniería geotécnica. Aunque se resalta que las ANN son una de las técnicas más usadas, con el 52% de los estudios apoyándose en ellas, también se distinguen otros como FIS, ANFIS, SVM, y GAN, aplicados en áreas clave como la mecánica de rocas, estabilidad de taludes, túneles y licuefacción de suelos.

Muchas de las revisiones comentadas ponen de relieve las dificultades en la aplicación de ML al ámbito de la geotecnia, debido a la escasez de datos o a la ausencia de un estándar. Phoon *et al.* (Phoon et al., 2022) introducen el concepto de la geotecnia centrada en datos, mencionando iniciativas como Project DeepGeo, que utiliza datos reales e indirectos para generar mapas tridimensionales del subsuelo. Subrayan la importancia de la cuantificación de la incertidumbre y el uso de métodos bayesianos.

A pesar de que en esta tesis se trabaja con datos provenientes de la realización de inclusiones rígidas, los estudios más parecidos se encuentran en el ámbito de la minería, la exploración de hidrocarburos y los túneles. Por ejemplo, la revisión de Lawal *et al.* (Lawal et al., 2024) recoge aplicaciones de ML en la exploración de petróleo y gas, enfocándose en la exploración geológica y geofísica, así como en la caracterización de reservorios. En esta se hace mención expresa a la tesis de Bhatt (Bhatt, 2002) que utiliza ANN para predecir propiedades claves de los yacimientos petrolíferos como la porosidad, permeabilidad, saturación de fluidos y litofacies. Las variables de entrada fueron los registros durante la perforación (registros sónicos, de densidad, de resistividad, de porosidad-neutrón, etc.).

La industria minera y la ingeniería de rocas subterráneas han adoptado técnicas de ML para resolver problemas complejos y optimizar procesos. Jang y Topal (Jang & Topal, 2014) realizan una revisión sobre el uso de tecnologías de soft computing en la industria minera, señalando aplicaciones de ML, ANN y algoritmos genéticos en la selección de métodos de minería, selección de equipos y predicción del comportamiento de masas rocosas. Estas técnicas permiten manejar la incertidumbre y mejorar la toma de decisiones en entornos con información limitada.

Morgenroth *et al.* (Morgenroth et al., 2019) revisan las oportunidades del uso de ML en el diseño de ingeniería de rocas subterráneas. Destacan cómo ML puede mejorar el análisis de datos recolectados en proyectos mineros y de infraestructura subterránea, como la estabilidad de taludes y predicción de eventos geomecánicos.

Identifican la falta de estándares en la arquitectura y selección de datos y proponen un marco para integrar ML en las prácticas tradicionales.

Khalilidermi y Knez (Khalilidermani & Knez, 2022) realizan una revisión del concepto de energía específica mecánica. En su trabajo identifican el uso de la inteligencia artificial como una solución efectiva para completar los modelos existentes de energía específica. Esta revisión se centra en perforaciones para hidrocarburos, pero también trata la futura perforación en formaciones extraterrestres.

Quizá una de las revisiones más interesantes para esta tesis sea la realizada por Isheyskiy y Shanchidrián (Isheyskiy & Sanchidrián, 2020). En ella se revisa el uso de MWD para el control de las operaciones de perforación y voladura en empresas mineras. Uno de los apartados se centra en el uso conjunto de ML y MWD y destaca cuatro artículos. El primero (Klyuchnikov et al., 2019) presenta un modelo, basado en datos para la identificación en tiempo real del tipo de roca en la punta de perforación mediante algoritmos de ML. Se evaluaron métodos como la regresión logística, redes neuronales y *gradient boosting* sobre datos obtenidos en el campo de Novoportovskoe en Siberia Occidental.

En el segundo trabajo, Galende-Hernández *et al.* (Galende-Hernández et al., 2018) utilizan MWD para estimar la calidad del macizo rocoso (reflejada en el *rock mass rating*, RMR) durante la excavación de túneles. Los datos se obtienen durante el proceso de perforación de los agujeros, donde se colocan posteriormente los explosivos que permiten la excavación del túnel. El objetivo del estudio es estimar el valor del RMR del frente de excavación con los datos del anterior frente. Conocer estos valores ayuda a optimizar las labores de voladura.

El tercer trabajo (Basarir et al., 2017), también busca estimar el RMR a partir de MWD. En este caso los datos de entrada fueron el peso y la rotación aplicados sobre la herramienta y la velocidad de perforación.

Por último, Kadkhodaie-Ilkhchi *et al.* (Kadkhodaie-Ilkhchi et al., 2010) utilizan datos de MWD de perforaciones para voladuras de una mina de hierro en Australia Occidental.

2.8.2. Aprendizaje automático con datos provenientes del registro de perforación

El uso de ML a partir de datos provenientes de MWD ha sido estudiado sobre todo en la ejecución de túneles y en la industria *oil&gas.* Multitud de autores han tratado de predecir los movimientos generados por la excavación de túneles partiendo de MWD (Ahangari et al., 2015; Apoji et al., 2023; Bouayad & Emeriault, 2017; Boubou et al., 2010; R.-P. Chen et al., 2019; Ding & Ye, 2023; Gamra et al., 2023; Kim et al., 2001; Moeinossadat et al., 2018; Ocak & Seker, 2013; Pourtaghi & Lotfollahi-Yaghin, 2012; Richa, Pereira, et al., 2024; Santos & Celestino, 2008; Shi et al., 1998; Suwansawat & Einstein, 2006; Wang et al., 2013; P. Zhang et al., 2020; W. G. Zhang et al., 2021; Zhou et al., 2017). Otro caso de estudio habitual es la predicción de la velocidad de perforación (*rate of penetration*, ROP) (Mahdevari et al., 2014; Mahmoodzadeh et al., 2022).

Sin embargo, las aplicaciones más relacionadas con esta tesis son aquellas que utilizan MWD para obtener características del terreno (como el RMR). Ya hemos citado al respecto los trabajos de Galende-Hernández (Galende-Hernández et al., 2018) y Hansen (Hansen et al., 2022). También han tratado este tema Liu *et al.* (B. Liu et al., 2019), que utilizan *Support Vector Regression* (SVR) para predecir distintas propiedades del macizo rocoso (resistencia a compresión simple, índice de fragilidad, la distancia entre planos de debilidad y la orientación de discontinuidades). Las variables de entrada que utilizaron fueron el empuje, el par, las revoluciones por minuto de la cabeza de corte, la velocidad de perforación y la potencia de la cabeza de corte.

En la industria de los hidrocarburos también resulta de gran importancia poder determinar la ROP, habiendo muchos autores que aplican ML con ese objetivo (O. S. Ahmed et al., 2019; Al-AbdulJabbar et al., 2020; Amar & Ibrahim, 2012; Anemangely et al., 2018; Elkatatny et al., 2020; Eskandarian et al., 2017; Gidh et al., 2012; H. I. Bilgesu et al., 1997; Hegde et al., 2017; Kakoli et al., 2018; Mantha & Samuel, 2016; Olafadehan et al., 2023; Ong et al., 2022; Sabah et al., 2019; Yavari et al., 2018).

Además de la ROP, hay otras variables que interesa estimar con MWD en la industria de los hidrocarburos como la presión de poro (Abdelaal et al., 2021; A. Ahmed et al., 2019; Bourgoyne & Young, 1974; Khaled et al., 2022), el coeficiente de Poisson (A. Ahmed et al., 2021; Al-Anazi & Gates, 2015; Siddig et al., 2021), la resistencia a compresión simple (Asadi, 2017). Los trabajos más relacionados con

esta tesis son aquellos que buscan predecir la litología atravesada (Imamverdiyev & Sukhostat, 2019; Liang et al., 2022; Ren et al., 2022).

En la Tabla 2.3 se resumen los artículos que utilizan ML con datos de MWD.

Tabla 2.3: Resumen de los artículos que utilizan ML con datos de MWD

Referencia	Año	Tema	Métodos ML	Conclusiones
(Abdelaal et al., 2021)	2021	Predicción del gradiente de presión en formaciones rocosas usando parámetros de perforación.	SVM, FN-ANN, RF	Los modelos utilizados predijeron el gradiente de presión con alta precisión. El modelo más efectivo fue RF, alcanzando un R de 0.99 y un MSE de 0.011 psi/ft.
(Ahangari et al., 2015)	2015	Evaluación de métodos inteligentes para la predicción de asentamientos inducidos por la ejecución de túneles.	ANFIS, GEP	Los métodos inteligentes ANFIS y GEP son herramientas útiles para predecir asentamientos en túneles, siendo GEP más efectivo en escenarios de valores altos de asentamiento.
(A. Ahmed et al., 2019)	2019	Comparación de técnicas de inteligencia artificial para predecir la presión de la formación rocosa durante la perforación.	ANN, RBF-ANN, ANFIS, SVM, FN-ANN	Todas las técnicas de inteligencia artificial predijeron la presión de la formación con alta precisión, siendo SVM la más efectiva, con un R2 de 0.995 y un MAE del 0.14%.
(O. S. Ahmed et al., 2019)	2018	Comparación de técnicas de inteligencia artificial para la predicción de la ROP.	ANN, ELM-ANN, SVM, SVM-LS	SVM-LS mostró el mejor rendimiento en la predicción de la ROP, y reducir el número de variables predictoras de ocho a cinco mejoró la precisión del modelo.
(A. Ahmed et al., 2021)	2021	Predicción en tiempo real del coeficiente de Poisson estático utilizando inteligencia artificial durante la perforación.	FN-ANN, RF	Los modelos FN y RF lograron predecir el coeficiente de Poisson estático con alta precisión, siendo el modelo RF superior con un R de 0.99 y un MAE del 1.89%.
(Al-Abduljabbar et al., 2020)	2020	Predicción de la ROP en pozos en forma de S utilizando ANN.	ANN	El modelo ANN alcanzó un R promedio de 0.93 y un RMSE del 6.2%.
(Al-AbdulJabbar et al., 2020)	2020	Desarrollo de un modelo de ANN optimizado para predecir la ROP durante la perforación de reservorios carbonatados horizontales.	ANN, SaDE-ANN	El modelo optimizado ANN logró una tasa de MAE del 5.29% y un R de 0.951, superando las correlaciones anteriores en la estimación de la tasa de penetración.
(Al-Anazi & Gates, 2015)	2015	Predicción de propiedades elásticas de rocas de reservorio utilizando SVM.	SVM, ANN	SVM resultó más efectivo para predecir la relación de Poisson y el módulo de Young bajo condiciones de datos limitados.
(Amar & Ibrahim, 2012)	2012	Comparación de técnicas de machine learning para la predicción de la ROP la industria del petróleo.	ELM-ANN, RBF-ANN	El modelo ELM mostró ser el más rápido, mientras que RBF fue el más preciso en la predicción de la ROP, recomendándose RBF cuando la precisión es prioritaria.
(Anemangely et al., 2018)	2018	Predicción de la ROP utilizando datos de registros petrofísicos y mediante ANN optimizada.	PSO-ANN, COA-ANN	El modelo COA-ANN mostró una convergencia más rápida y una mejor precisión en la predicción de ROP en comparación con el modelo PSO-ANN, destacando el impacto positivo de la reducción de ruido y la selección de características.
(Apoji et al., 2023)	2023	Desarrollo de un sistema de inteligencia artificial para estimar en tiempo real los movimientos del suelo inducidos por túneles en áreas urbanas utilizando MWD.	RF	El sistema propuesto puede estimar las respuestas del suelo en tiempo real basándose únicamente en las características de la máquina de perforación y las geometrías espaciales del túnel, y puede ser utilizado como herramienta para controlar la máquina y limitar los movimientos inducidos.
(Asadi, 2017)	2017	Predicción de la resistencia a compresión uniaxial de rocas utilizando MWD.	ANN	Las ANN son efectivas para predecir la resistencia de las rocas y pueden reducir significativamente los costos de perforación al ofrecer estimaciones precisas basadas en datos de perforación.
(Basarir et al., 2017)	2017	ANFIS para predecir propiedades de la roca basándose en MWD.	ANFIS	Los resultados muestran que ANFIS puede predecir con éxito propiedades del macizo rocoso como el RQD utilizando MWD.
(H. I. Bilgesu et al., 1997)	1997	Se utiliza ANN para predecir la ROP y el desgaste de la herramienta bajo varios tipos de formación y parámetros operativos.	ANN	Las ANN fueron capaces de predecir la ROP con coeficientes de correlación entre 0.902 y 0.982.
(Bouayad & Emeriault, 2017)	2017	Modelado de asientos en superficie causados por tuneladoras mediante PCA y ANFIS.	PCA, ANFIS	Se desarrollaron modelos que predicen con alta correlación los asientos en superficie en función de parámetros operativos y geológicos, destacando la efectividad de ANFIS.
(Boubou et al., 2010)	2010	Predicción de los movimientos en superficie del suelo a partir de los parámetros de operación del túnel mediante ANN.	ANN	Las redes neuronales son adecuadas para describir los movimientos en superficie inducidos por la TBM, resaltando la importancia de ciertos parámetros operativos como la ROP y la presión hidráulica aplicada.
(R.-P. Chen et al., 2019)	2019	Evaluación de ANN al predecir el asiento máximo en superficie causado por túneles de presión de tierras (EPB).	ANN, RBF-ANN, GR-ANN	La red GR-ANN supera a las redes ANN y RBF-ANN en precisión y tiempo computacional, logrando un MAE de 1.10 y un RMSE de 1.35 en la predicción de asentamientos.

Referencia	Año	Tema	Métodos ML	Conclusiones
(Ding & Ye, 2023)	2023	Influencia de la construcción de túneles en el asentamiento de túneles existentes mediante monitoreo y predicción inteligente.	ANN	Se concluyó que el número de capas ocultas no debe exceder al número de entradas para evitar el sobreajuste y que un mayor monitoreo de datos mejora la precisión de la predicción del asentamiento.
(Elkatatny et al., 2020)	2020	ANN para predecir la ROP en operaciones de perforación.	ANN	ANN mostró R de 0.94 y un error porcentual absoluto promedio (AAPE) de 8.6%, superando a otros modelos existentes en la predicción de la ROP.
(Eskandarian et al., 2017)	2017	Estimación de la ROP en perforaciones utilizando métodos de minería de datos y aprendizaje automático.	ANN, RF	Los modelos de ML muestran una buena precisión en la estimación de la ROP.
(Galende-Hernández et al., 2018)	2017	Estimación del RMR durante la excavación de túneles utilizando MWD y ML.	PCA, FA, K-Means, X-Means, DBSCAN, S-IRL, L-IRL	Se logró una alta precisión en la estimación del RMR, alcanzando un error de predicción del 3% mediante el uso de un sistema basado en reglas difusas, lo que proporciona un soporte valioso en tiempo real para la construcción de túneles.
(Gamra et al., 2023)	2023	Estudio paramétrico sobre el impacto de la excavación de túneles en estructuras vecinas utilizando algoritmos de ML.	ANN, RF, SVM	La metodología de ML demuestra alta precisión en la predicción del daño a estructuras superficiales, superando significativamente a los métodos tradicionales.
(Gidh et al., 2012)	2012	Optimización de parámetros de perforación utilizando ANN para mejorar la ROP y gestionar el desgaste de brocas.	ANN	El sistema de optimización de parámetros de perforación basado en ANN demostró mejorar significativamente la ROP y aumentar la vida útil de las brocas, reduciendo costos operativos.
(Hansen et al., 2022)	2022	Utilización de técnicas de ML para mejorar la interpretación de datos de MWD en el proceso de decisión en túneles.	CNN, RF	Aunque se lograron algunos resultados prometedores, los modelos mostraron un desempeño insatisfactorio para la predicción de múltiples clases, destacando la necesidad de más información de características y un balance de clases en los datos.
(Hegde et al., 2017)	2017	El artículo compara modelos tradicionales y basados en datos para predecir la ROP durante la perforación.	RF	Los modelos basados en datos superaron en precisión a los modelos tradicionales en todas las formaciones evaluadas, demostrando que son más efectivos para la predicción de la ROP.
(Hegde & Gray, 2018)	2018	Optimización de la perforación utilizando modelos de aprendizaje automático ensamblados.	RF	La optimización del modelo de energía específica mecánica es la mejor estrategia de optimización de la perforación.
(Imamverdiyev & Sukhostat, 2019)	2019	Desarrollo de un modelo efectivo basado en aprendizaje profundo para la clasificación de facies geológicas en pozos.	1D-CNN, ANN, LSTM, SVM y KNN.	El modelo 1D-CNN mostró un rendimiento superior en la clasificación de facies en comparación con otros modelos de ML, alcanzando altos niveles de precisión en los conjuntos de datos de validación.
(Kadkhodaie-Ilkhchi et al., 2010)	2010	Comparación empírica de técnicas de ML para la clasificación de datos de MWD en roca.	Boosting, ANN, ANFIS	Boosting mostró el mejor rendimiento en precisión y eficiencia computacional en comparación con ANN y ANFIS para clasificar tipos de roca a partir de datos MWD.
(Kakoli et al., 2018)	2018	Predicción del rendimiento de las operaciones de perforación en un campo petrolero iraní utilizando métodos de ML.	SVM	La correlación entre la ROP real y la predicha fue de 0.989, con un MAE de solo 8.39%, mostrando que el modelo es altamente efectivo para predecir la ROP.
(Khaled et al., 2022)	2022	Desarrollo de modelos de ANN para predecir la presión de poro y de fractura durante la perforación en litologías mixtas.	ANN	Se desarrollaron modelos de ANN que lograron alta precisión en la estimación de presión de poro y de fractura, con R^2 de 0.974 y 0.998, respectivamente.
(Kim et al., 2001)	2001	Predicción de asientos en superficie causados por la excavación de túneles utilizando ANN.	ANN	ANN puede predecir eficazmente los asientos en superficie, identificando los factores más significativos que afectan dichos asientos.
(Klyuchnikov et al., 2019)	2019	Propuesta de un modelo basado en datos para la identificación del tipo de roca.	LiR, ANN, Boosting	La clasificación de tipos de roca alcanzó una precisión del 91%, reduciendo el error de clasificación del 13.5% al 9% a través de algoritmos examinados.
(Liang et al., 2022)	2022	Propuesta de un método para identificar litologías en perforación petrolera utilizando energía específica mecánica y ML.	SVM	El modelo de SVM logra una precisión de más del 91% en la identificación de litologías, superando a los métodos tradicionales.
(Mahdevari et al., 2014)	2014	Desarrollo de un modelo de regresión para predecir la ROP de TBM utilizando ML.	SVM	El modelo SVM predijo la ROP con R^2 de 0.95.

Referencia	Año	Tema	Métodos ML	Conclusiones
(Mahmoodzadeh et al., 2022)	2022	Desarrollo de un modelo híbrido LSTM optimizado para predecir la ROP de TBM.	LSTM	El modelo LSTM mostró resultados de predicción más precisos en comparación con otros modelos de ML, destacando la efectividad de la optimización en la mejora del rendimiento del LSTM.
(Mantha & Samuel, 2016)	2016	Optimización de la ROP en operaciones de perforación mediante técnicas de ML.	ANN, KNN, SVM, DT, RF, Boosting	La combinación de análisis de datos de múltiples variables y el uso de métodos de ML permite modelar y predecir la ROP con mayor precisión que los modelos tradicionales.
(Moeinossadat et al., 2016)	2016	Predicción del asiento en superficie causado por túneles EPB e identificación de parámetros más influyentes.	ANFIS	ANFIS superó a la regresión múltiple en la predicción del asiento, enfatizando la importancia de factores operacionales.
(Moeinossadat et al., 2018)	2018	Control de asientos en superficie causados por EPB utilizando ANFIS.	ANFIS	El modelo ANFIS mostró un rendimiento adecuado en la predicción de asentamientos máximos.
(Ocak & Seker, 2013)	2013	Predicción de asientos superficiales causados por EPB utilizando métodos de ML.	ANN, SVM, GNB	La metodología GNB logró la mejor precisión en la predicción de asientos superficiales superando a ANN y SVM, y resaltando la importancia de identificar variables de entrada efectivas para este tipo de predicciones.
(Olafadehan et al., 2023)	2023	Evaluación de diferentes algoritmos de ML para predecir la ROP durante la perforación de pozos petroleros.	SVM, RF, LiR, KNN, CNN	El método de stacking de modelos mostró el mejor rendimiento en la predicción de la ROP.
(Ong et al., 2022)	2022	Estudio sobre la predicción de la ROP en pozos utilizando ANN.	ANN	Los modelos ANN demostraron mayor precisión en la predicción de ROP comparado con modelos tradicionales, mostrando su aplicabilidad en operaciones de perforación reales.
(Pourtaghi & Lotfollahi-Yaghin, 2012)	2012	Comparación entre la capacidad de Wavenet y las ANN para predecir el asiento máximo en superficie causado por el tunel.	Wavenet-ANN, ANN	Wavenet mostró un rendimiento superior que las ANN tradicionales, mejorando la estimación del asiento máximo en superficie.
(Ren et al., 2022)	2022	Identificación de litología utilizando aprendizaje activo y semi-supervisado basado en una versión mejorada de NB.	NB	El algoritmo NB demostró ser efectivo en la identificación de litología, mejorando la precisión en un 15.28% y 17.36% en dos conjuntos de datos, superando otros métodos de aprendizaje activo y semi-supervisado.
(Richa et al., 2023)	2023	Predicción de subsidencias inducidas por túneles utilizando métodos de ML	XGBoost, RF	Ambos métodos mostraron un buen rendimiento en la predicción, pero es necesario incluir métricas que penalicen la subestimación para mejorar las predicciones.
(Richa, Pereira, et al., 2024)	2024	Evaluación de la precisión de cinco algoritmos de ML en la predicción de asientos inducidos por túneles con datos limitados.	SVM, DT, RF, XGBoost, ANN	Los métodos de ensamblado, especialmente XGBoost y RF, son los más precisos incluso con solo el 30% del conjunto de datos utilizado para el entrenamiento.
(Richa, Lamare, et al., 2024)	2024	Este artículo presenta dos enfoques para predecir en tiempo real los asientos inducidos por la excavación de túneles.	RF, LSTM	RF ofrece una solución más rápida y simple para la predicción del asiento máximo, mientras que LSTM proporciona un nivel de detalle adicional al modelar la evolución temporal de los asientos.
(Sabah et al., 2019)	2019	ML para predecir la ROP en perforaciones utilizando datos petrofísicos y de registro de lodo.	ANN, RBF-ANN, SVM, PSO-ANN	El modelo PSO-ANN mostró la mayor precisión en la predicción de la ROP, superando a otros métodos, incluyendo SVM, al trabajar con datos preprocesados para reducir el ruido.
(Santos & Celestino, 2008)	2008	Análisis de asientos de túneles del metro de São Paulo mediante ANN.	ANN	Las ANN son herramientas efectivas para predecir asientos en túneles, mostrando mejoras significativas en precisión tras ajustes en la topología y tratamiento de datos.
(Shi et al., 1998)	1998	ANN para predecir asentamientos en túneles mediante modelos neuronales modulares.	ANN	La implementación de un enfoque modular redujo el error promedio de predicción de asentamientos de 70 mm a 33.4 mm, demostrando una mejora significativa en la precisión respecto al modelo general de red neuronal.
(Siddig et al., 2021)	2021	Estimación en tiempo real del módulo de Poisson a partir de parámetros de perforación utilizando técnicas de ML.	ANN, ANFIS	Se logró una buena correlación entre los valores reales y los predichos del módulo de Poisson, con coeficientes de correlación entre 0.97 y 0.99, y errores promedio absolutos menores al 2.2%, siendo la ANN la más efectiva entre los métodos evaluados.
(Suwansawat & Einstein, 2006)	2006	Estudio del uso de ANN para predecir asientos superficiales máximos provocados por tuneladora EPB en Bangkok.	ANN	Las ANN pueden predecir eficazmente los asentamientos superficiales debido a las condiciones de operación de la tuneladora EPB, pero se debe considerar la influencia de las variaciones entre diferentes tipos de máquinas tuneladoras.
(Wang et al., 2013)	2013	Predicción del asiento en superficie inducido por el túnel utilizando un modelo de SVM con núcleo de wavelet.	Wavelet-SVM	El modelo propuesto proporciona predicciones precisas del asiento del suelo durante la excavación de túneles y permite una evaluación de la incertidumbre asociada a las predicciones.

Referencia	Año	Tema	Métodos ML	Conclusiones
(Yavari et al., 2018)	2018	Comparación de modelos matemáticos y de inteligencia artificial para predecir la ROP en perforación de petróleo.	ANFIS	El modelo ANFIS es más preciso que los modelos matemáticos en la predicción de la ROP cuando hay una gran cantidad de datos disponibles.
(P. Zhang et al., 2020)	2020	Comparación de algoritmos de ML para la predicción de asentamientos inducidos por tuneladora.	ANN, GR-ANN, ELM-ANN, SVM, RF	El algoritmo RF mostró el mejor rendimiento en la predicción de asentamientos, mientras que el ANN es más adecuado cuando los conjuntos de datos son pequeños.
(W. G. Zhang et al., 2021)	2021	Predicción del asiento en superficie inducido por tuneladora EPB.	XGBoost, ANN, SVM, MARS	El modelo XGBoost mostró la mayor precisión en la predicción del asiento del suelo, superando a los otros métodos de ML evaluados.
(Zhou et al., 2017)	2016	Predicción de asientos del terreno causados por la construcción de túneles utilizando RF.	RF	El método RF demostró ser efectivo para predecir los asentamientos del terreno inducidos por la construcción de túneles, identificando la cohesión y la pérdida de volumen como las variables más influyentes.

Leyenda

1D-CNN	Red neuronal convolucional unidimensional (1D Convulational Neural Network)	ANFIS	Sistema de inferencia neuro-difuso adaptativo (Adaptive Neuro-Fuzzy Inference System)
ANN	Red neuronal artificial (Artificial Neural Network)	ANN-ELM	Red neuronal de aprendizaje extremo (Artificial Neural Network - Extreme Learning Machines)
ANN-FN	Red neuronal artificial (Artificial Neural Network-Funtional Network)	ANN-GR	Red neuronal artificial de regresión general (Artificial Neural Network-General Regression)
ANN-PSO	Red neuronal con optimización de enjambre de partículas (Artificial Neural Network-Particle Swarm Optimization)	ANN-SaDE	Red neuronal artificial con Evolución Diferencial Auto-adaptativa (Artificial Neural Network-Self-adaptive Differential Evolution)
ANN-Wavenet	Red neuronal artificial redes neuronales con funciones de activación wavelet (Artificial Neural Network-Wavenet)	CNN	Red neuronal convulacional (Convulational Neural Network)
COA	Algoritmo de optimización cuco (Cuckoo Optimization Algorithm)	DT	Árboles de decisión (Decission Trees)
FA	Analisis factorial (Factorial Analysis)	GEP	Programación de expresión genética (Gene Expression Programming)
GNB	Naïve-Bayes gausiano (Gaussian Naïve-Bayes)	LiR	Regresión linear (Linear Regression)
L-IRL	Aprendizaje de reglas iterativo linguistico (Linguistic Iterative Rule Learning)	LoR	Regresión logística (Logistic Regression)
LSTM	Red neuronal de memoria larga a corto plazo (Long Short-Term Memory)	MARS	Splines de Regresión Adaptativa Multivariante (Multivariate Adaptive Regression Splines)
NB	Naïve-Bayes	PCA	Análisis de componentes principales (Principal Components Analysis)
RF	Bosques aleatorios (Random Forests)	S-IRL	Aprendizaje de reglas iterativo disperso (Scatter Iterative Rule Learning)
SVM	Máquinas de vectores de soporte (Support Vector Machines)	SVM-LS	Máquinas de vectores de soporte de mínimos cuadrados (Least Square Support Vector Machines)

Se observan multitud de enfoques a la hora de aprovechar los datos provenientes de MWD usando ML. En general las ANN y sus variantes han tenido un gran uso, pero también destacan SVM y RF. Muchos autores destacan que ML arroja mejores resultados que las formulaciones analíticas existentes.

Entre las referencias estudiadas que usan MWD, y a diferencia del trabajo desarrollado en esta tesis, apenas se encuentran trabajos en los que se perforen suelos (siendo la mayoría en roca). Por lo tanto, tampoco se han utilizado datos de penetrómetros, ya que son ensayos típicos de suelos.

3. Metodología

3.1. Obras estudiadas

En esta tesis se ha trabajado con datos del registro de perforación de diversas obras donde se ejecutaron inclusiones rígidas. Estas inclusiones se realizaron mediante desplazamiento a rotación (Figura 3.1).

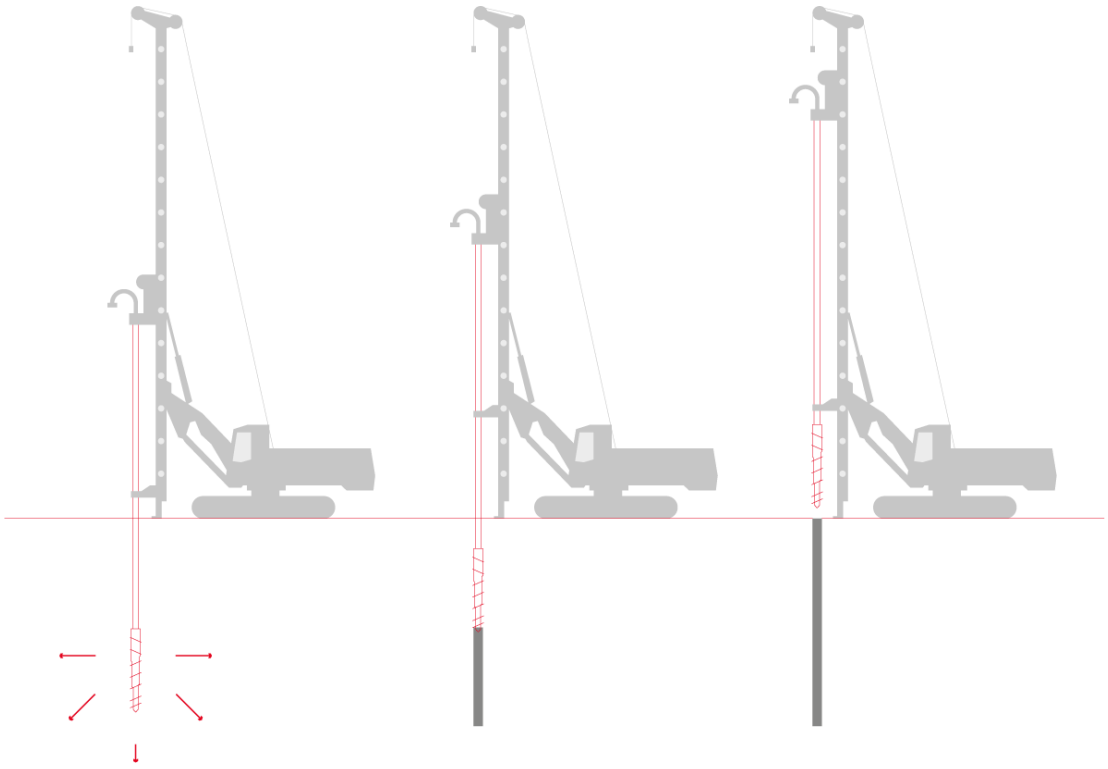


Figura 3.1: Esquema del procedimiento de ejecución de una inclusión rígida (imagen cortesía de Menard)

Para todas las obras la maquinaria de perforación fue similar a una ENTECO 6050 con las siguientes características:

Tabla 3.1: características de la maquinaria utilizada

Característica	Valor
Velocidad de perforación máxima	15 m/min
Máxima fuerza de empuje	15 t
Máximo par de rotación	18 t·m
Máxima velocidad de rotación	20 rpm

En todos los casos se realizó un registro de parámetros de perforación tomando datos, al menos, del par de rotación, la velocidad de rotación, la velocidad de perforación, el empuje aplicado y la profundidad alcanzada.

En los siguientes apartados se describen las características más relevantes de estas obras.

3.1.1. Obra 1

En esta obra se realizaron inclusiones rígidas bajo la solera y zapatas de una nave industrial. Se dispone de datos en bruto de un total de 607 inclusiones, 150 de diámetro 360 mm (bajo zapatas) y 457 de diámetro 300 mm (bajo solera). Los datos de estas inclusiones se registraron mediante un sistema EMPAREX de ABAC.

El terreno está formado superficialmente por un relleno antrópico de arenas limosas vertidas con un espesor interpretado que varía entre 1.60 y 2.20 m. Bajo estos rellenos aparece una capa de suelo vegetal de poco espesor (entre 0.4 y 1.0 m) compuesta igualmente por arenas limosas. Seguidamente aparece un nivel aluvial de arena fina con limos de compacidad muy suelta y un espesor de 1.60 a 2.20 m. La última capa investigada consiste en el granito alterado de la zona con una compacidad creciente en profundidad.

El material es, por tanto, de una naturaleza bastante homogénea y principalmente arenosa encontrándose el nivel freático cercano a la superficie.

Durante la campaña geotécnica se realizaron un total de 7 penetrómetros tipo DPSH-B entre otros trabajos. La distribución de los penetrómetros e inclusiones se muestra en la Figura 3.2, mientras que los valores de golpeo de los penetrómetros en función de la cota pueden consultarse en la Figura 3.3.

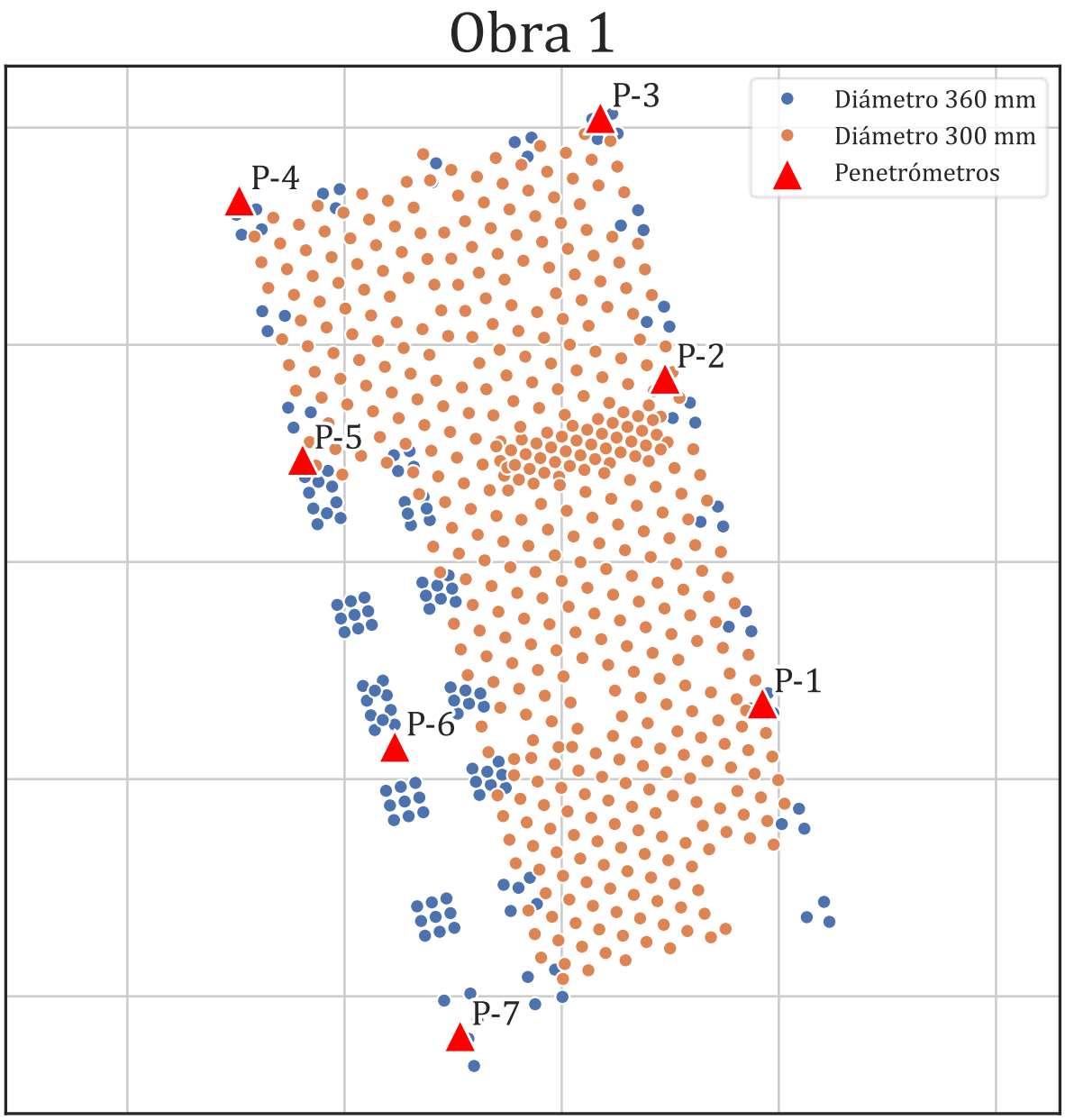
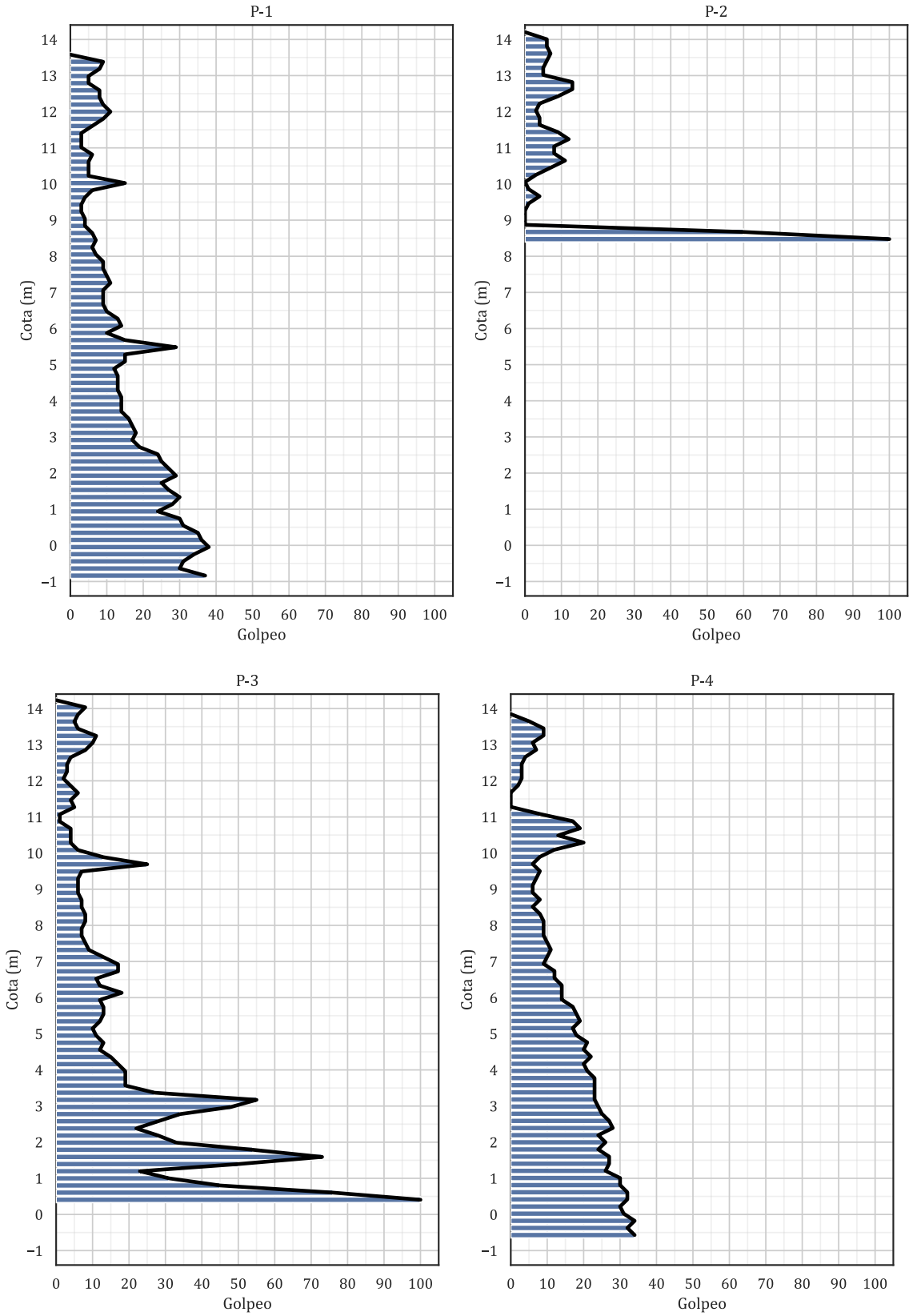


Figura 3.2: Distribución de inclusiones y penetrómetros para la Obra 1


P-1
P-2
P-3
P-4
Cota (m)
Golpeo
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
−1
0
10
20
30
40
50
60
70
80
90
100

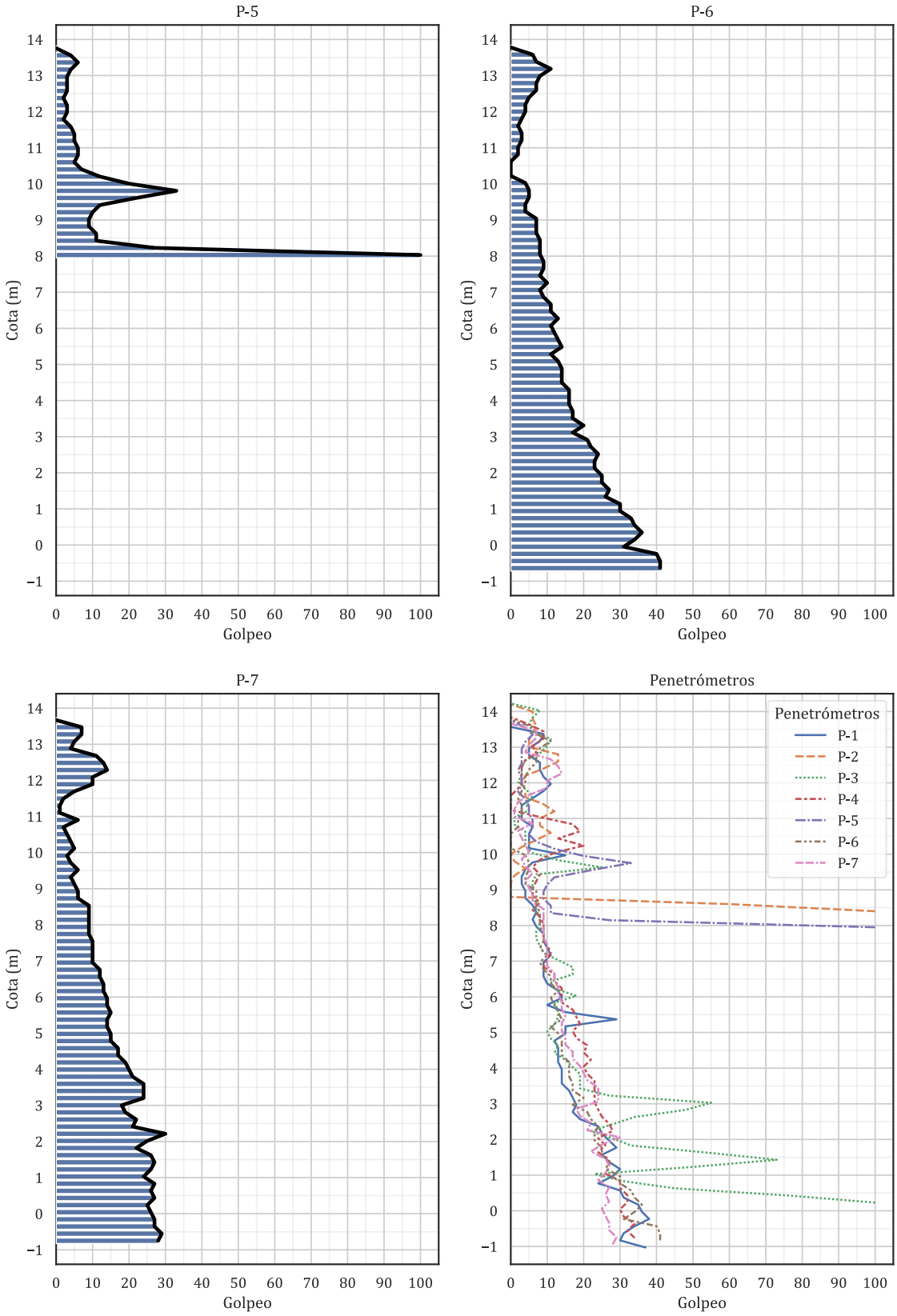


Figura 3.3: Golpeos en función de la cota para los penetrómetros de la Obra 1

3.1.2. Obra 2

La Obra 2 también consistió en el tratamiento del terreno bajo una nave industrial. En este caso se dispone de datos de 2038 inclusiones, todas ellas de 360 mm de diámetro. Los datos de estas inclusiones se tomaron mediante un sistema EMPAREX de ABAC.

Según la información geotécnica disponible, el nivel más superficial está formado por rellenos antrópicos con un espesor medio de 6 m constituidos por arenas y arenas arcillosas (SC). Bajo este nivel aparecen materiales aluviales dispuestos en capas de diversa naturaleza, formando un interestratificado desde arcillas hasta arenas, aunque con cierta continuidad lateral. Este nivel puede alcanzar una profundidad de unos 30 m. El último nivel detectado coincide con el sustrato plioceno de la zona, constituido por arenas densas de elevada compacidad. Este nivel se encuentra a una profundidad tal que no es alcanzado por el tratamiento del terreno.

El nivel freático se detectó en cotas muy distintas en función de la época de realización de las mediciones. Las medidas han variado entre los 2 y los 8 m de profundidad.

Por lo tanto, a diferencia de la Obra 1, esta obra presenta una geología mucho más heterogénea, con materiales de distinta naturaleza y una posición variable del nivel freático.

Se dispone de 7 penetrómetros tipo DPSH-B, cuya distribución respecto a las columnas se muestra en la Figura 3.4. Los valores de golpeo de los penetrómetros en función de la cota pueden consultarse en la Figura 3.5.

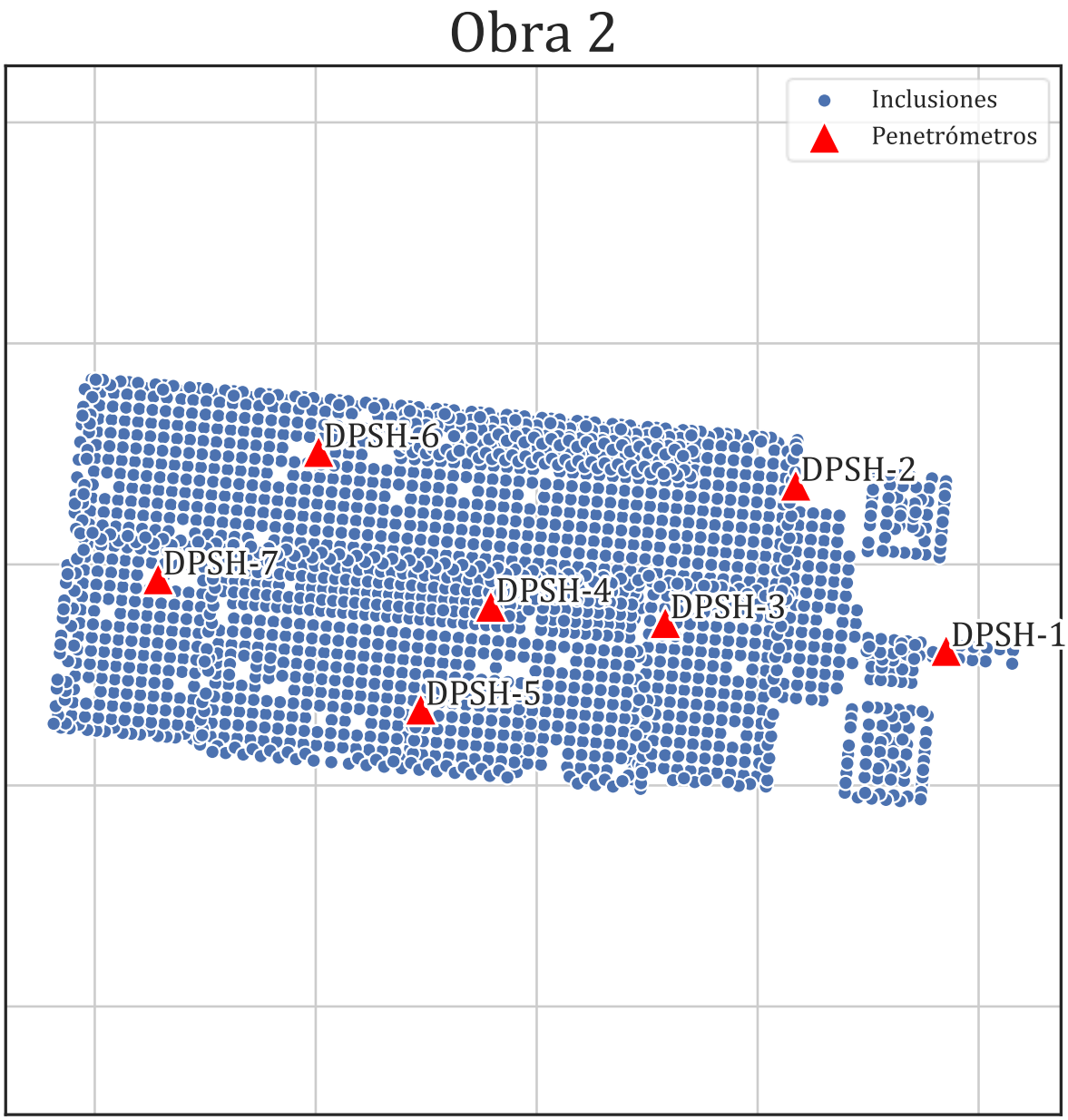
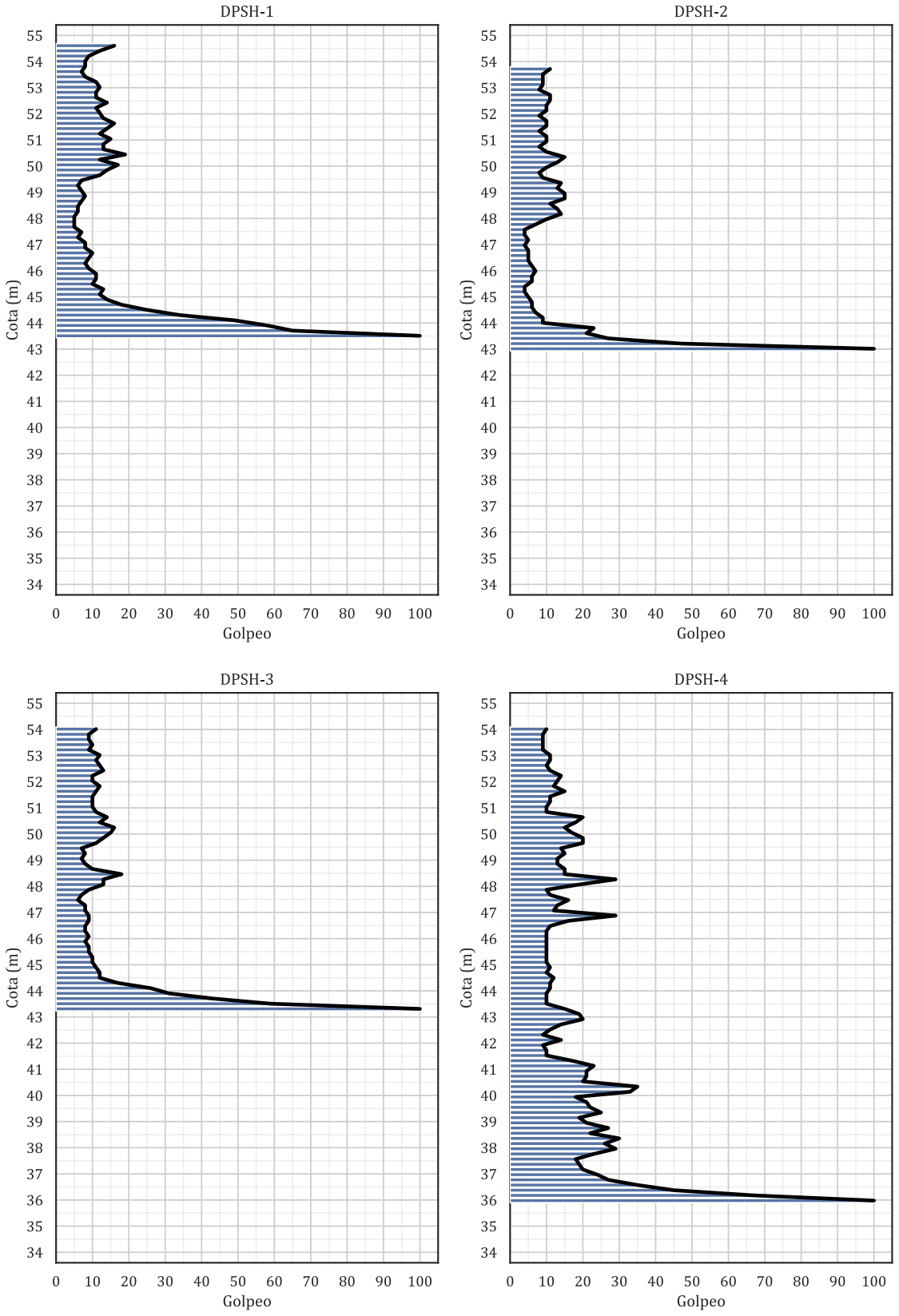


Figura 3.4: Distribución de inclusiones y penetrómetros para la Obra 2


DPSH-1
DPSH-2
DPSH-3
DPSH-4
Cota (m)
Golpeo
55
54
53
52
51
50
49
48
47
46
45
44
43
42
41
40
39
38
37
36
35
34
0
10
20
30
40
50
60
70
80
90
100

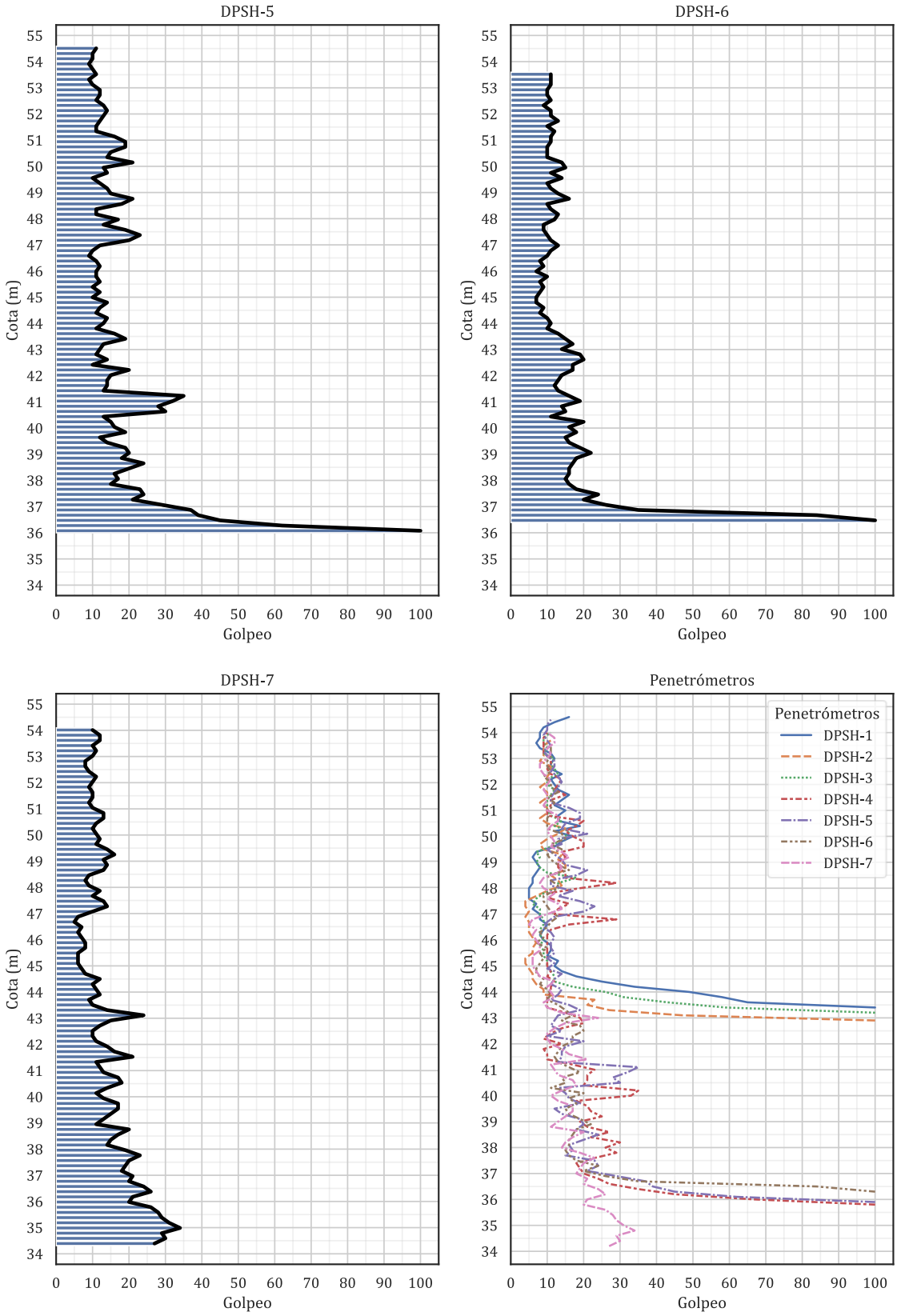


Figura 3.5: Golpeos en función de la cota para los penetrómetros de la Obra 2

3.1.3. Obra 3

En la Obra 3 las inclusiones se ejecutaron para la cimentación de unos tanques de hidrocarburos sobre un relleno portuario. En total se dispone de datos de 1488 inclusiones de diámetro 360 mm. Los datos de estas inclusiones se tomaron mediante un sistema EMPAREX de ABAC.

En esta obra se tiene información de 4 penetrómetros tipo DPSH-B. Sin embargo, estos se hicieron antes de conocer la posición final de los tanques que finalmente no coincidió con la de todos los penetrómetros. Debido a esta circunstancia sólo dos de ellos resultan relevantes (Figura 3.6).

El relleno portuario estaba formado por material procedente de dragado utilizado para rellenar varios recintos ganados al mar. La composición es variable pero principalmente granular, formada por arenas con indicios de finos y gravas. El espesor de este relleno se encuentra entre los 4 y los 8 m. Bajo dicho primer nivel se encuentra el terreno natural compuesto por arenas cuaternarias que aumentan su compacidad con la profundidad.

El nivel freático se encuentra muy superficial y dependerá principalmente del nivel del mar.

En este caso el terreno vuelve a ser principalmente arenoso y razonablemente homogéneo, como en el caso de la Obra 1.

Los valores de golpeo en los DPSH en función de la cota para la Obra 3 se muestran en la Figura 3.7.

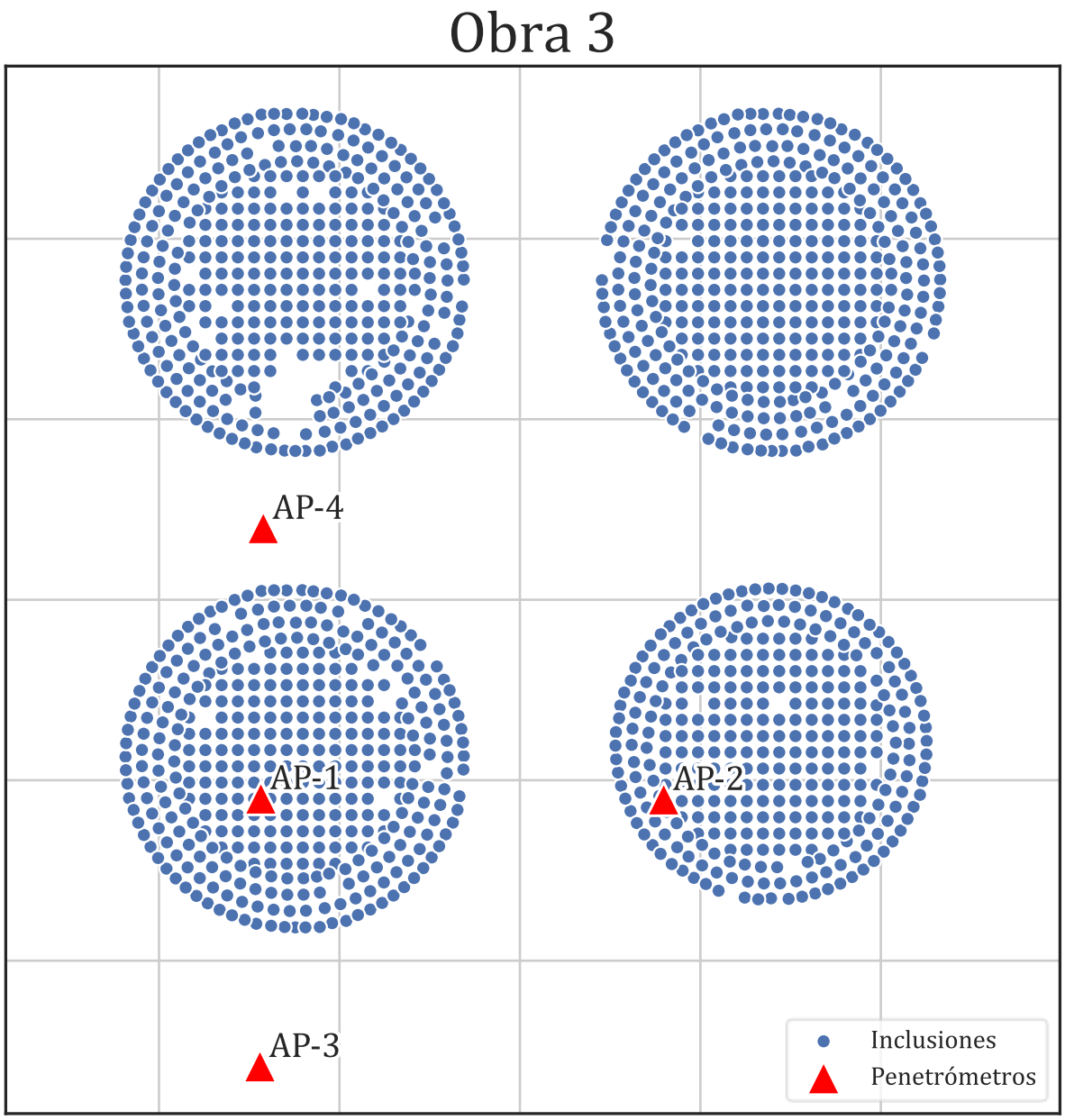
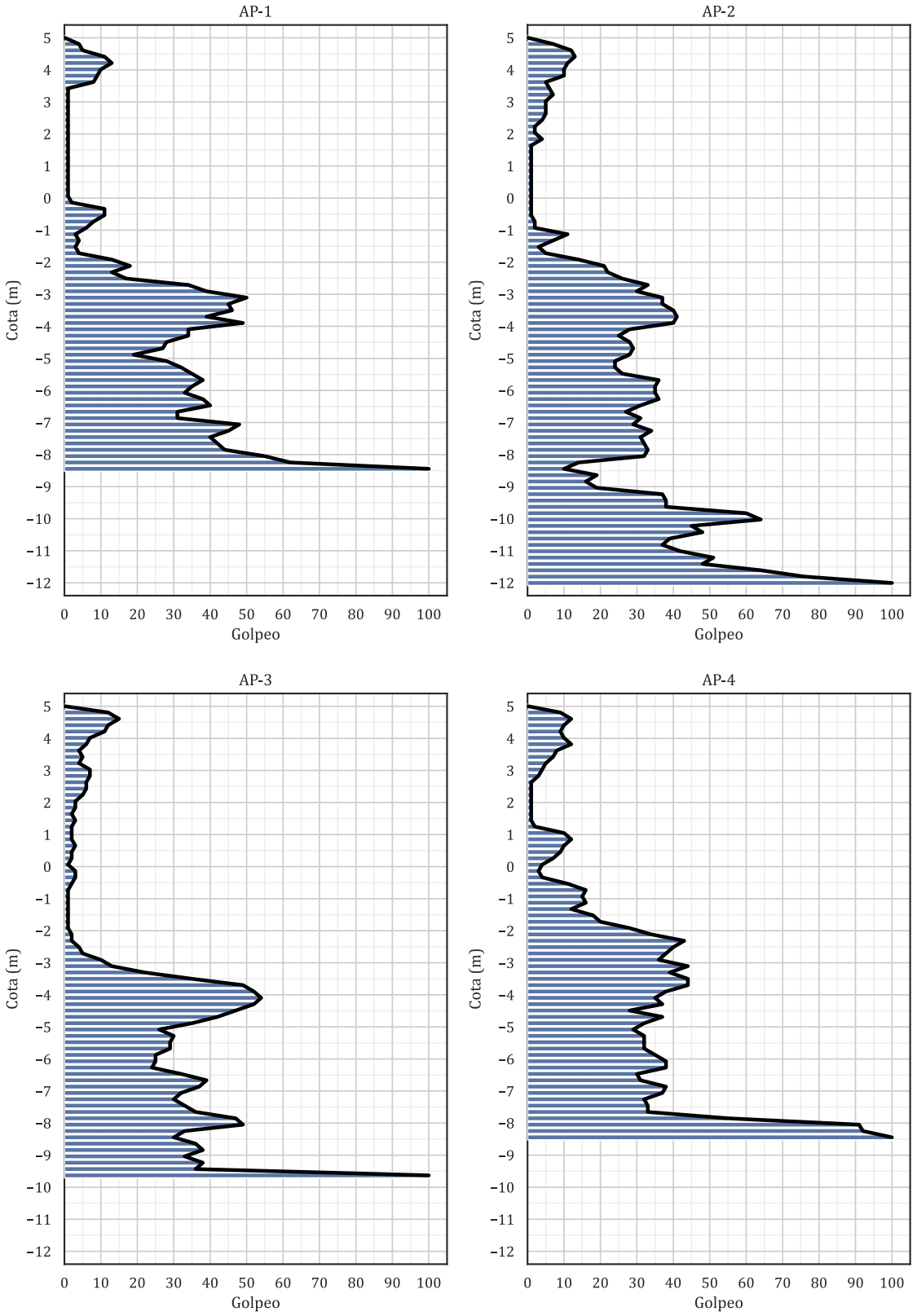


Figura 3.6: Distribución de inclusiones y penetrómetros para la Obra 3


AP-1
AP-2
AP-3
AP-4
Cota (m)
Golpeo
5
4
3
2
1
0
−1
−2
−3
−4
−5
−6
−7
−8
−9
−10
−11
−12
0
10
20
30
40
50
60
70
80
90
100

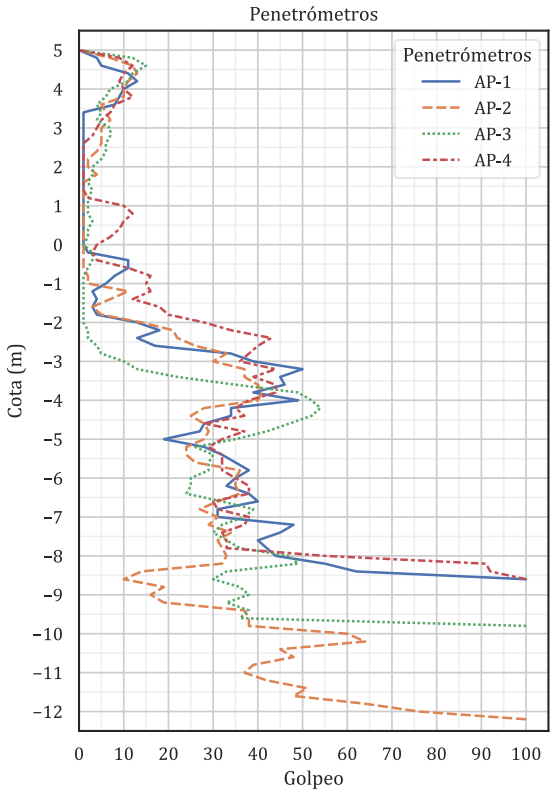


Figura 3.7: Golpeos en función de la cota para los penetrómetros de la Obra 3

3.1.4. Obra 4

En la Obra 4 se realizaron inclusiones bajo zapatas, pavimentos interiores y exteriores de una nave industrial. Se dispone de datos de 392 inclusiones de las cuales 104 tienen un diámetro de 360 mm y corresponden a las zapatas. Las 288 restantes se encuentran bajo los pavimentos y tienen un diámetro de 300 mm. Los datos de estas inclusiones se tomaron mediante un sistema EMPAREX de ABAC.

En este caso el terreno estaba formado por relleno antrópico de arenas y limos de baja plasticidad con un espesor de unos 7 m. Bajo este nivel aparecía un depósito de terraza formado por limos y arcillas arenosas y arenas limosas de colores rojizos, de grano fino y baja plasticidad. Por otro lado, el nivel freático se encontraba de media a unos 6 m de profundidad.

A diferencia de los casos anteriores, los cuatro ensayos penetrométricos realizados en esta obra fueron tipo Borros en lugar de DPSH. La distribución de los penetrómetros e inclusiones se muestra en la Figura 3.8, mientras que los valores de golpeo aparecen en la Figura 3.9.

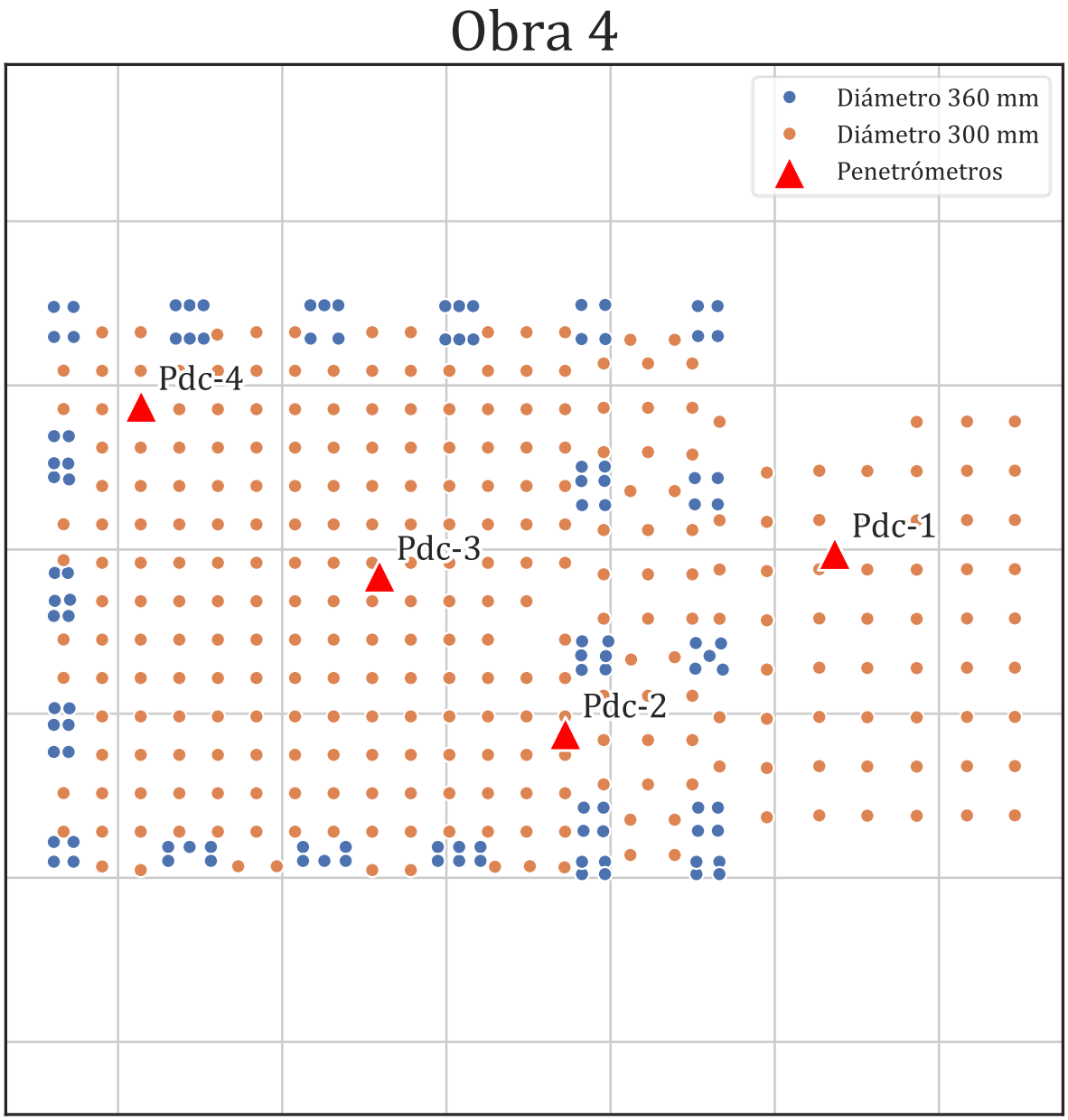
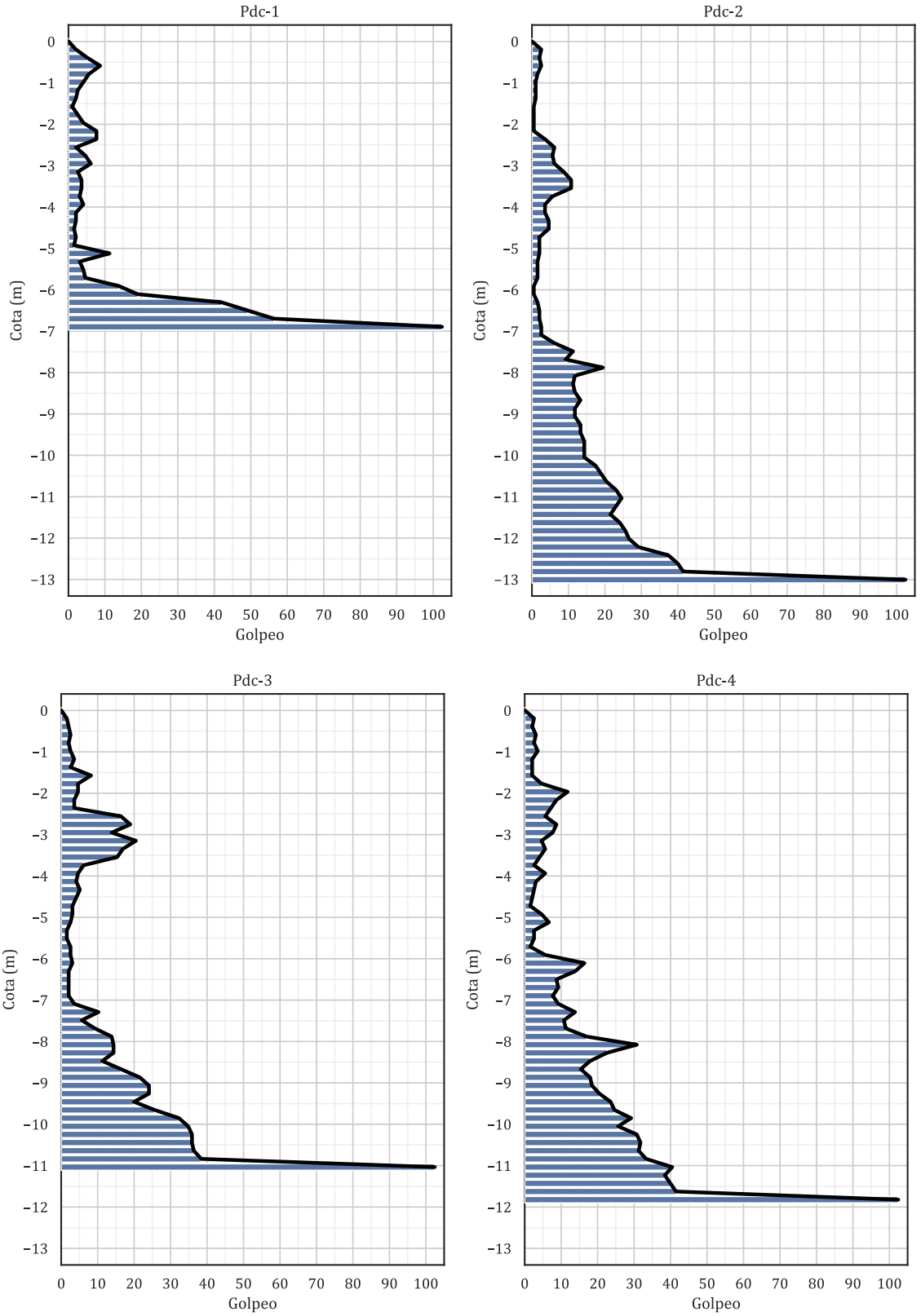


Figura 3.8: Distribución de inclusiones y penetrómetros para la Obra 4


Pdc-1
Pdc-2
Pdc-3
Pdc-4
Cota (m)
Golpeo
0
−1
−2
−3
−4
−5
−6
−7
−8
−9
−10
−11
−12
−13
0
10
20
30
40
50
60
70
80
90
100

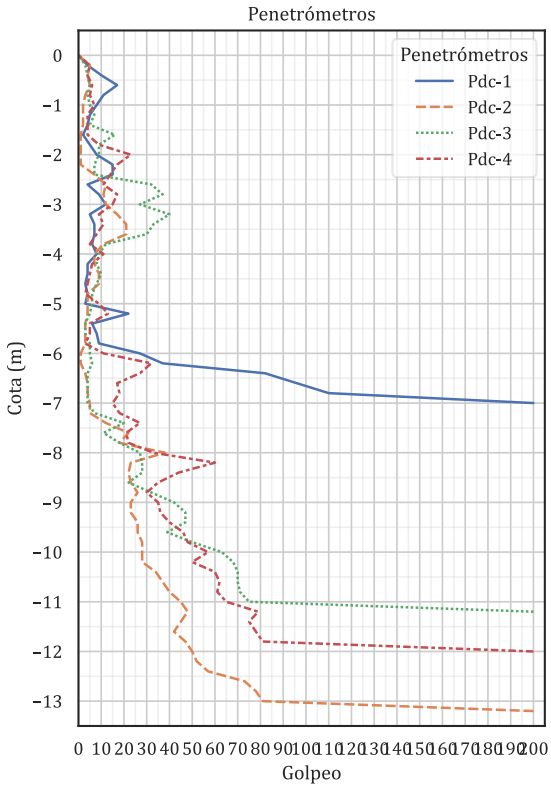


Figura 3.9: Golpeos en función de la cota para los penetrómetros de la Obra 4

3.1.5. Obra 5

En esta obra se ejecutó un tratamiento de inclusiones rígidas en una terminal portuaria que se destinaría a actividades relacionadas con la instalación de parques eólicos marinos. Es una actuación de gran envergadura y se dispone de datos de 3156 inclusiones de 300 mm de diámetro. Los datos de estas inclusiones se recogieron mediante Omnibox, un sistema de recogida de parámetros de perforación diseñado por Menard.

En este caso el terreno en superficie estaba formado por rellenos antrópicos compuestos principalmente por arenas limosas, pero con elementos de diferente naturaleza. Bajo este material aparecen depósitos de rio recientes formados mayoritariamente por arenas finas a limosas. Este material muestra presencia de materia orgánica. Por último, aparecen unas arenas grises correspondientes a depósitos de estuario y marinos compuestas por arenas finas con presencia de gravas.

De esta obra se dispone de un gran número de penetrómetros estáticos CPT; teniendo en cuenta sólo aquellos más cercanos a las inclusiones, son un total de 34. La distribución tanto de las inclusiones como de los penetrómetros se muestra en la Figura 3.10. Los valores de resistencia por punta (q_c) y fuste (f_s) del cono en función de la cota se muestran en la Figura 3.11.

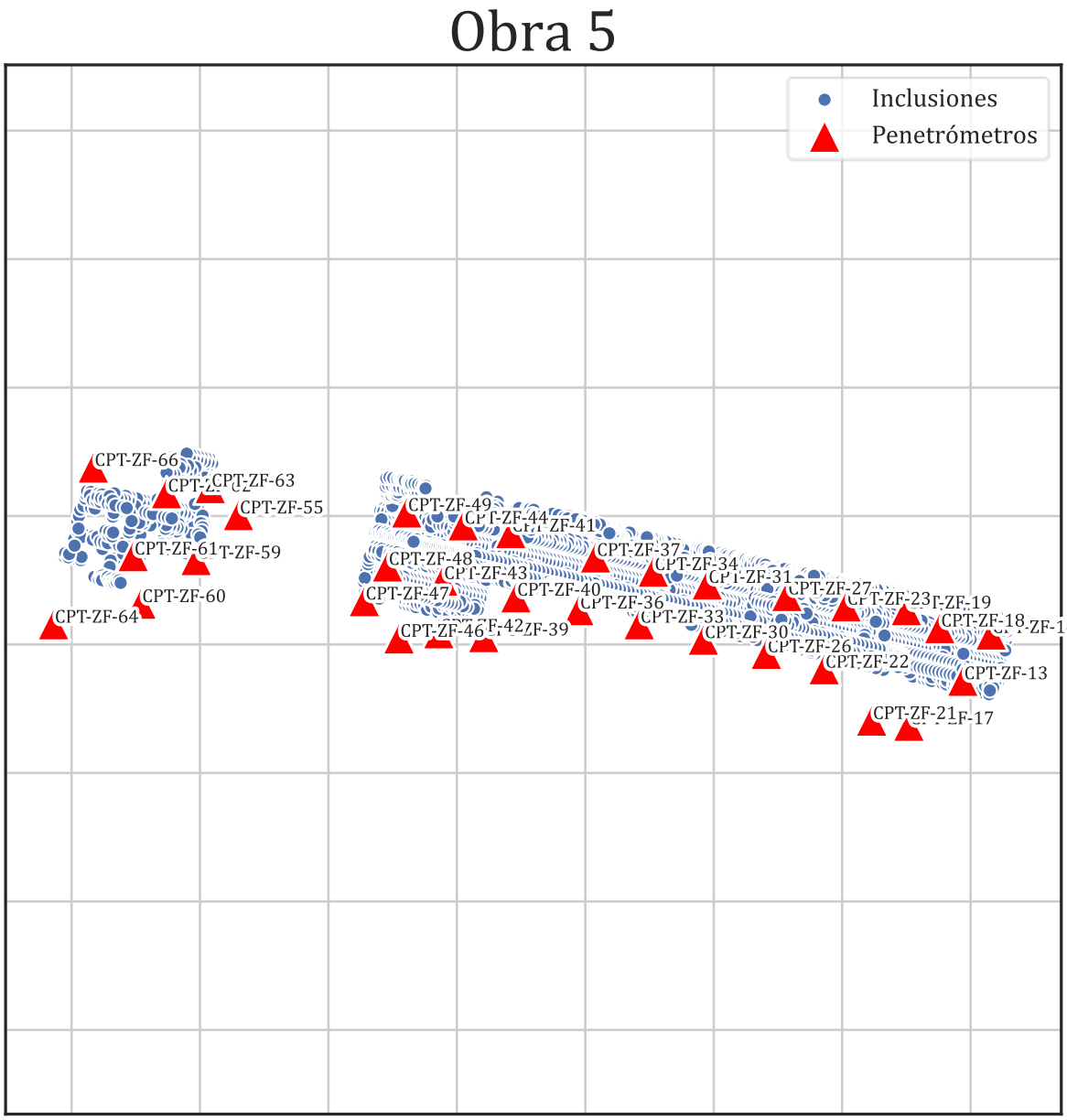
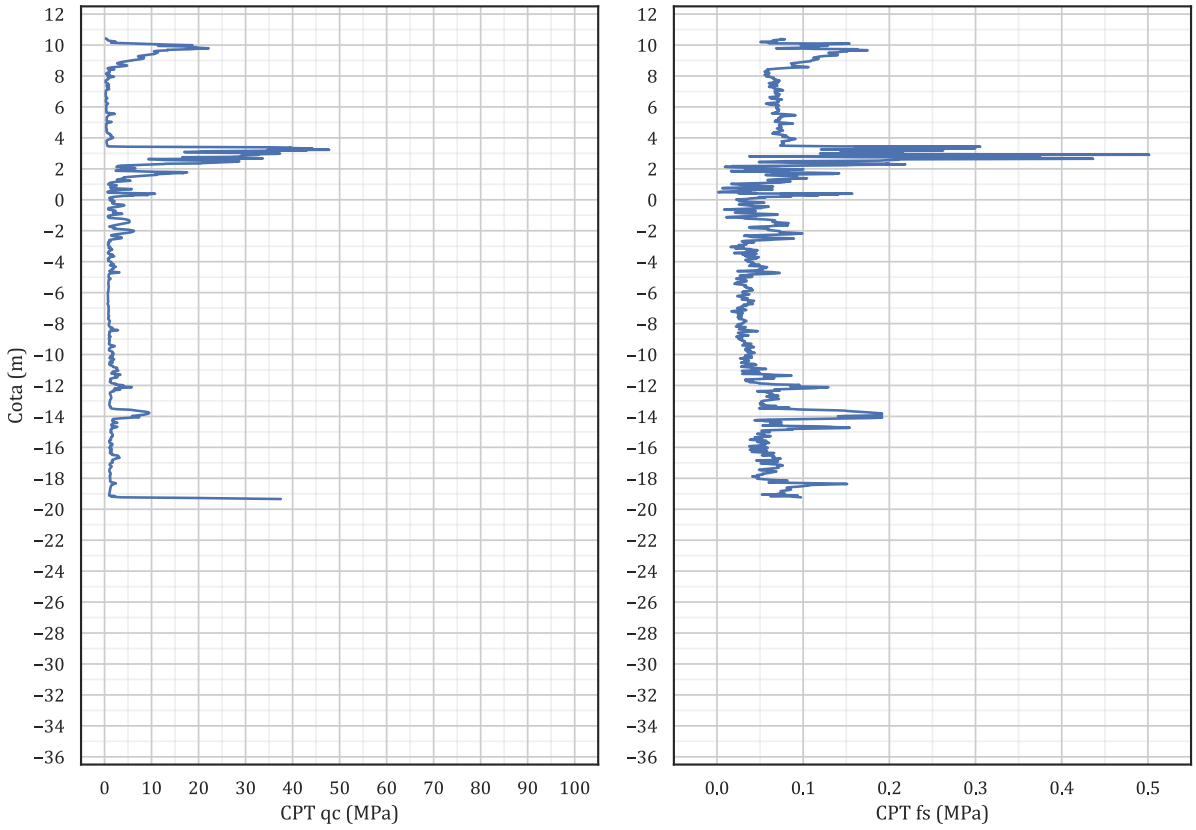
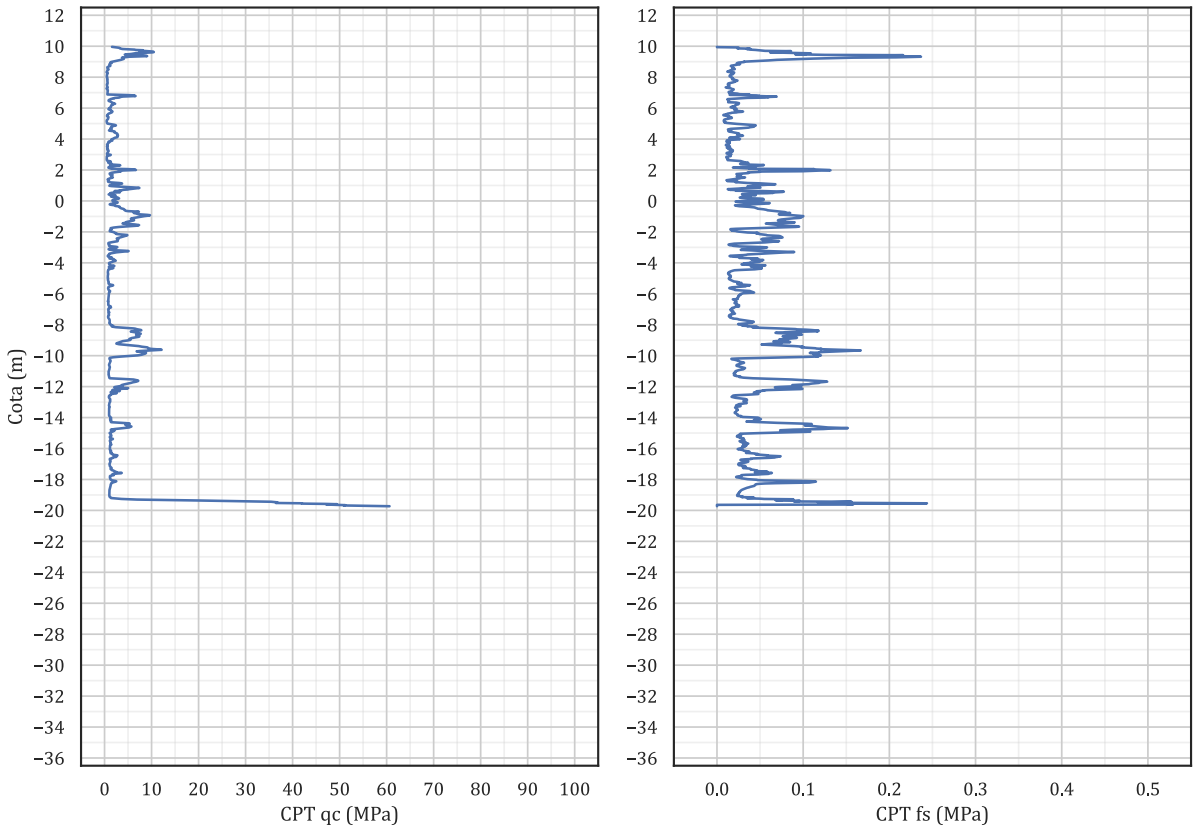


Figura 3.10: Distribución de inclusiones y penetrómetros para la Obra 5

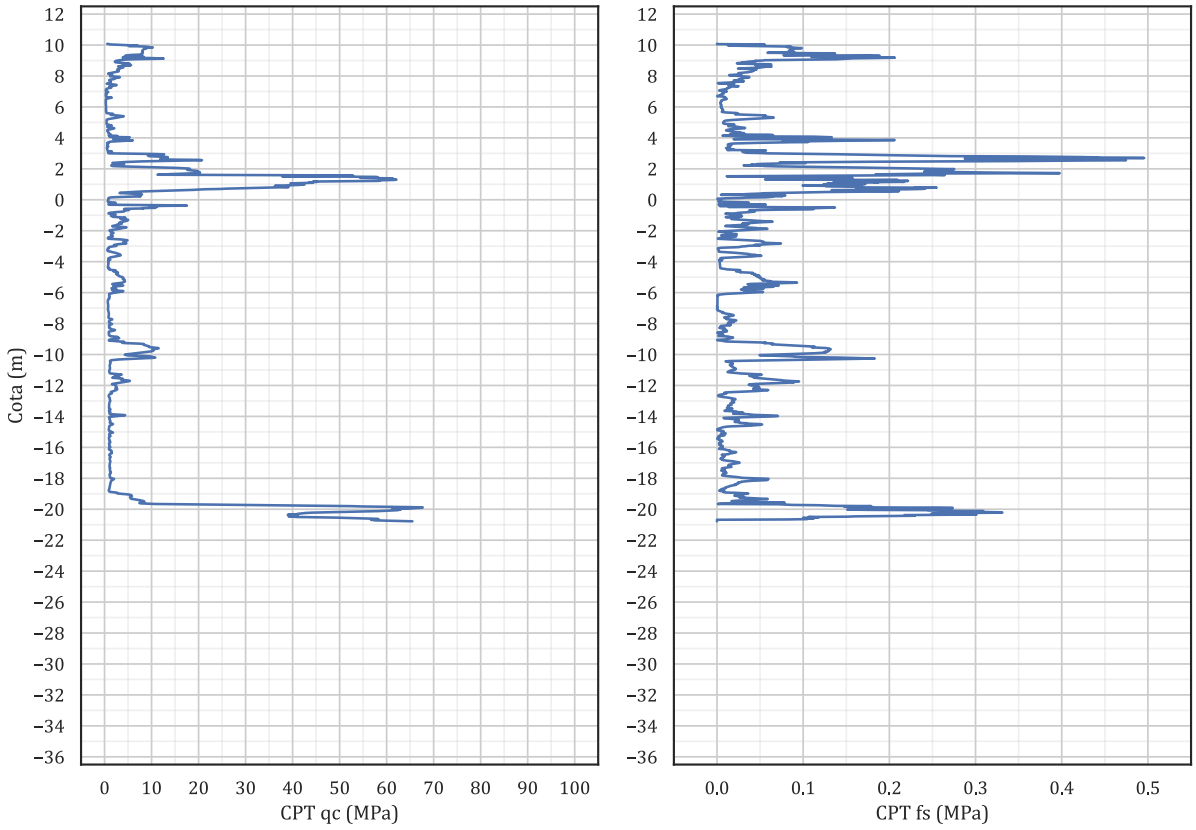
CPT-ZF-13



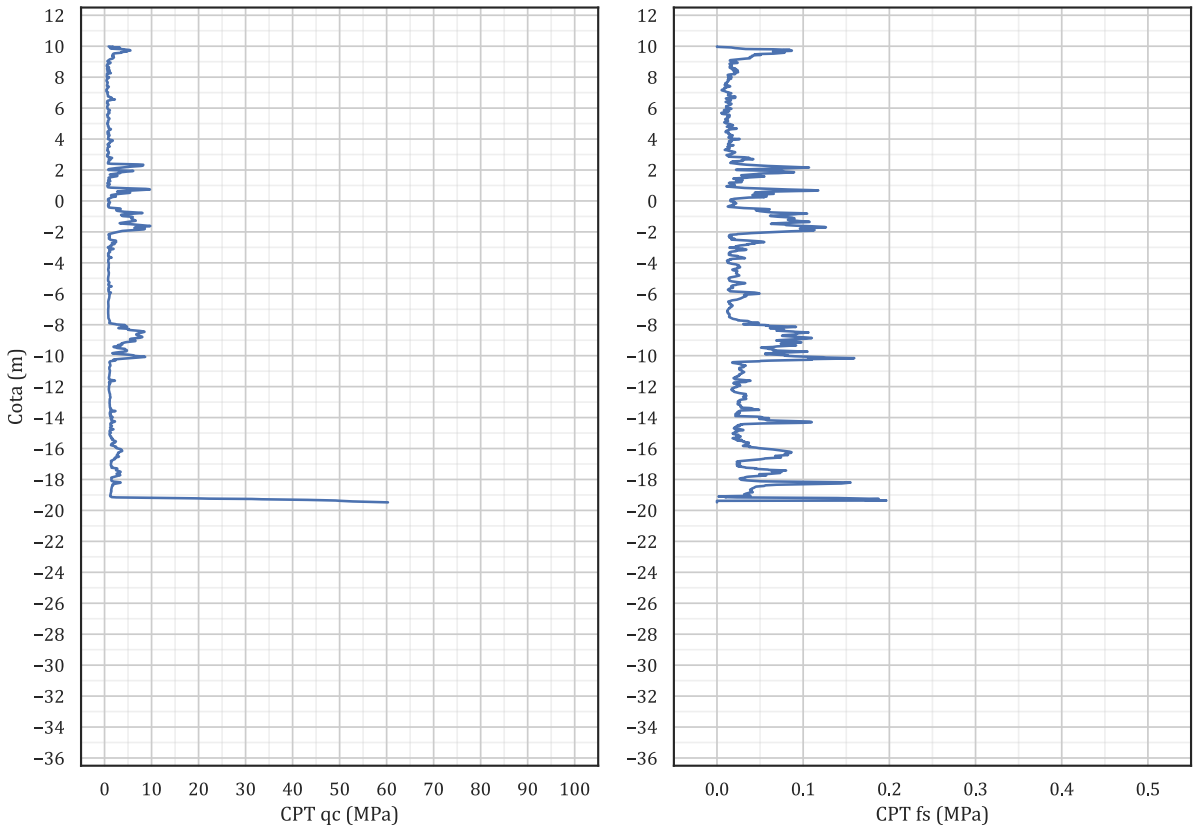
CPT-ZF-14



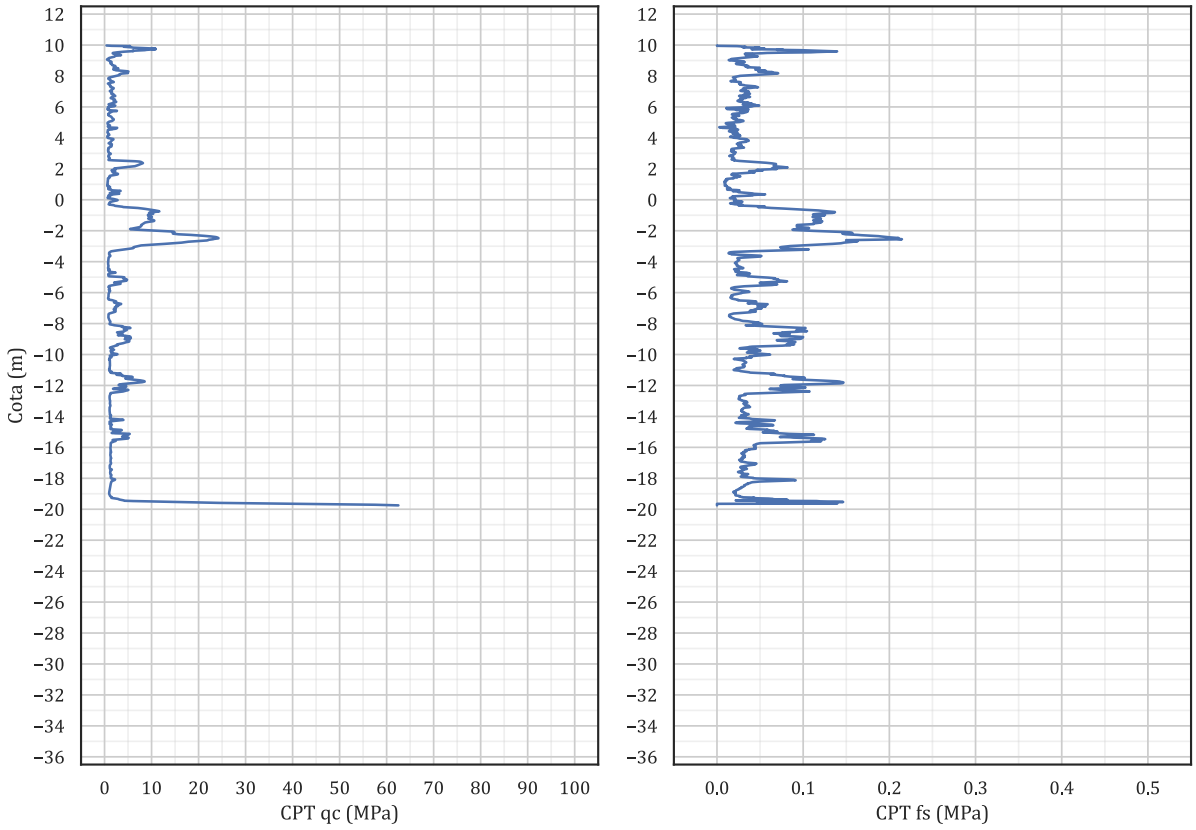
CPT-ZF-17



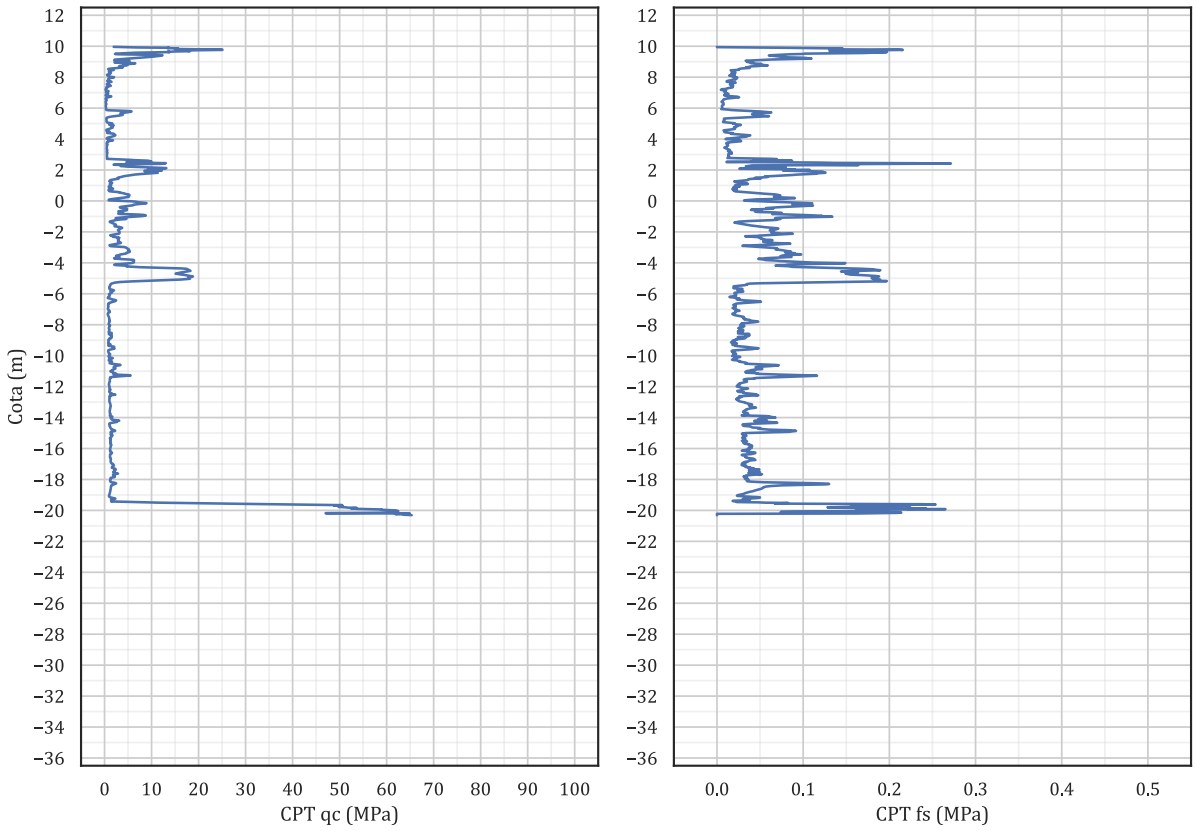
CPT-ZF-18



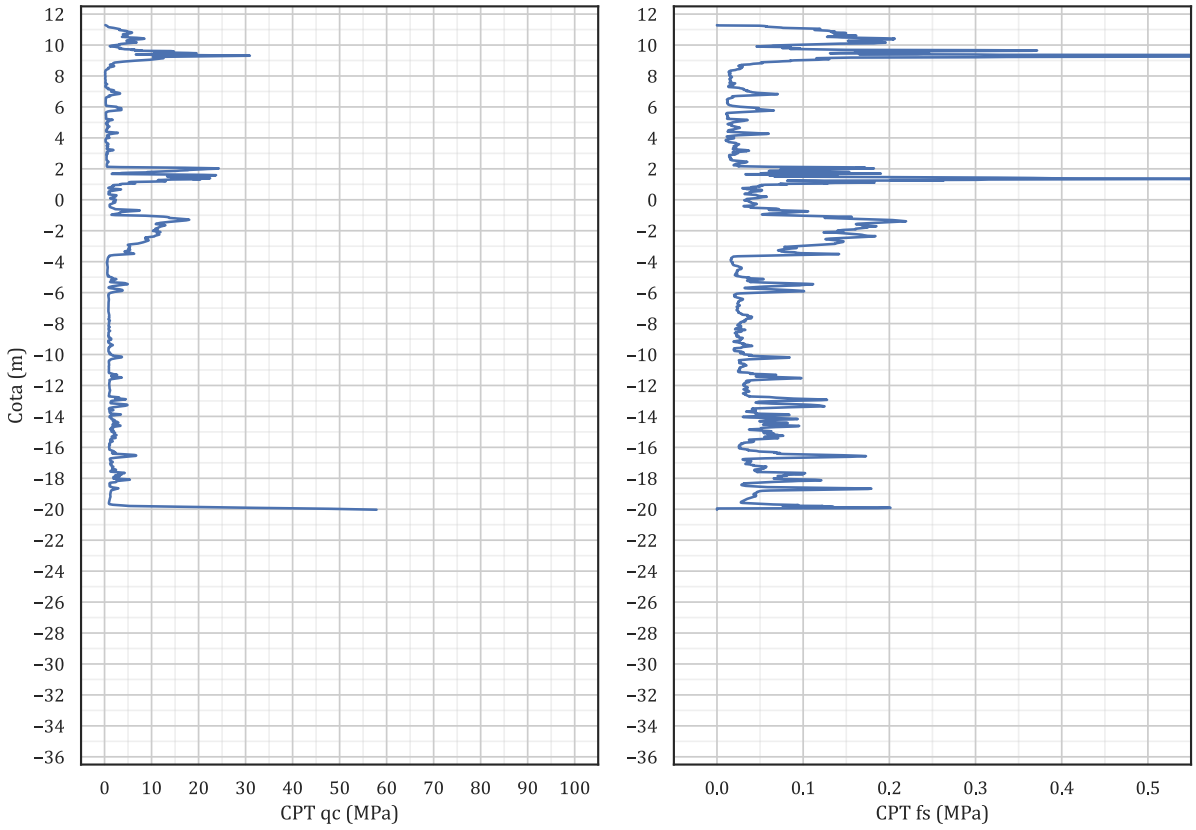
CPT-ZF-19



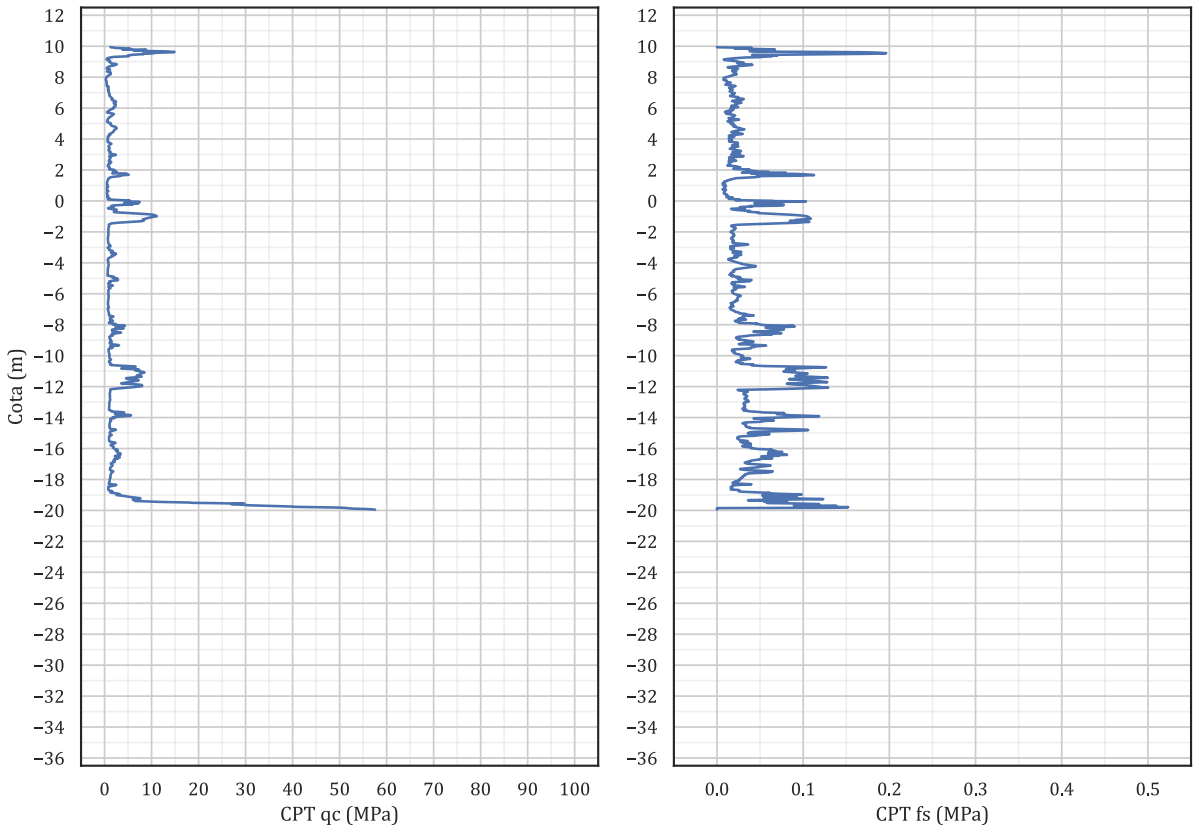
CPT-ZF-21



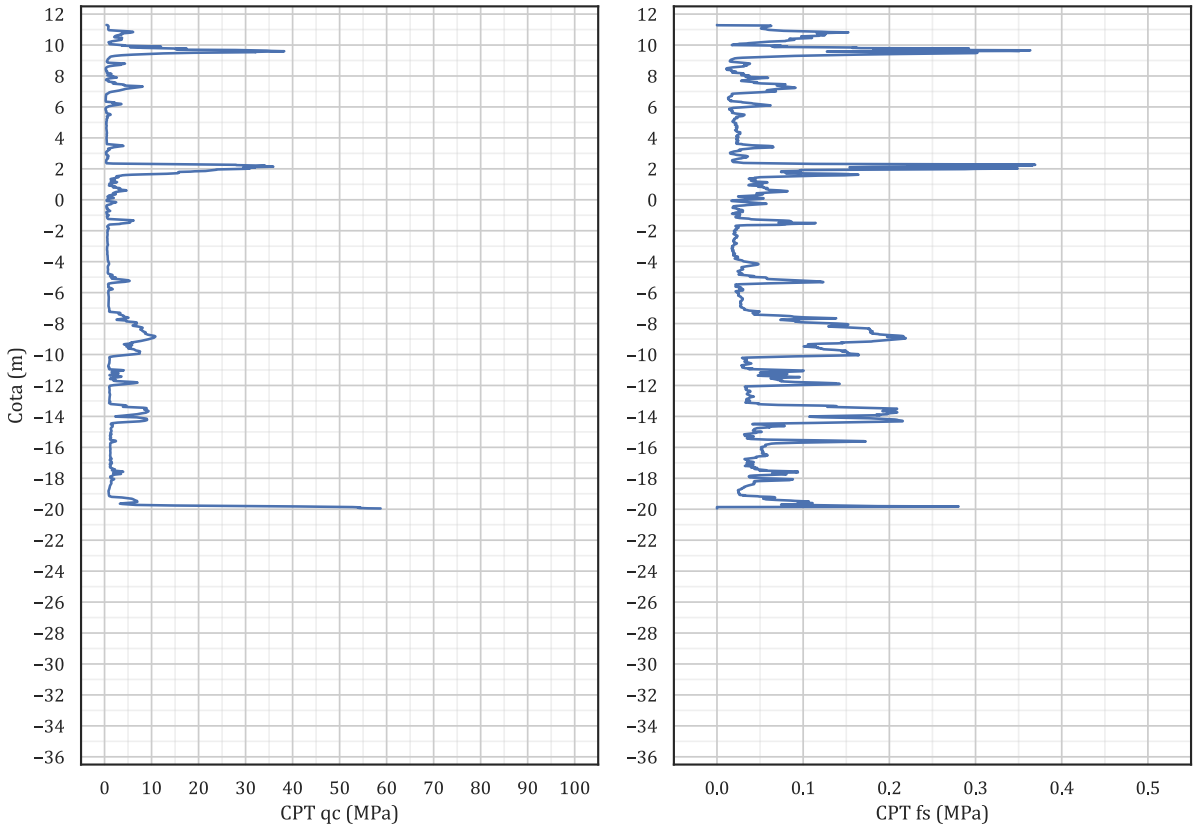
CPT-ZF-22



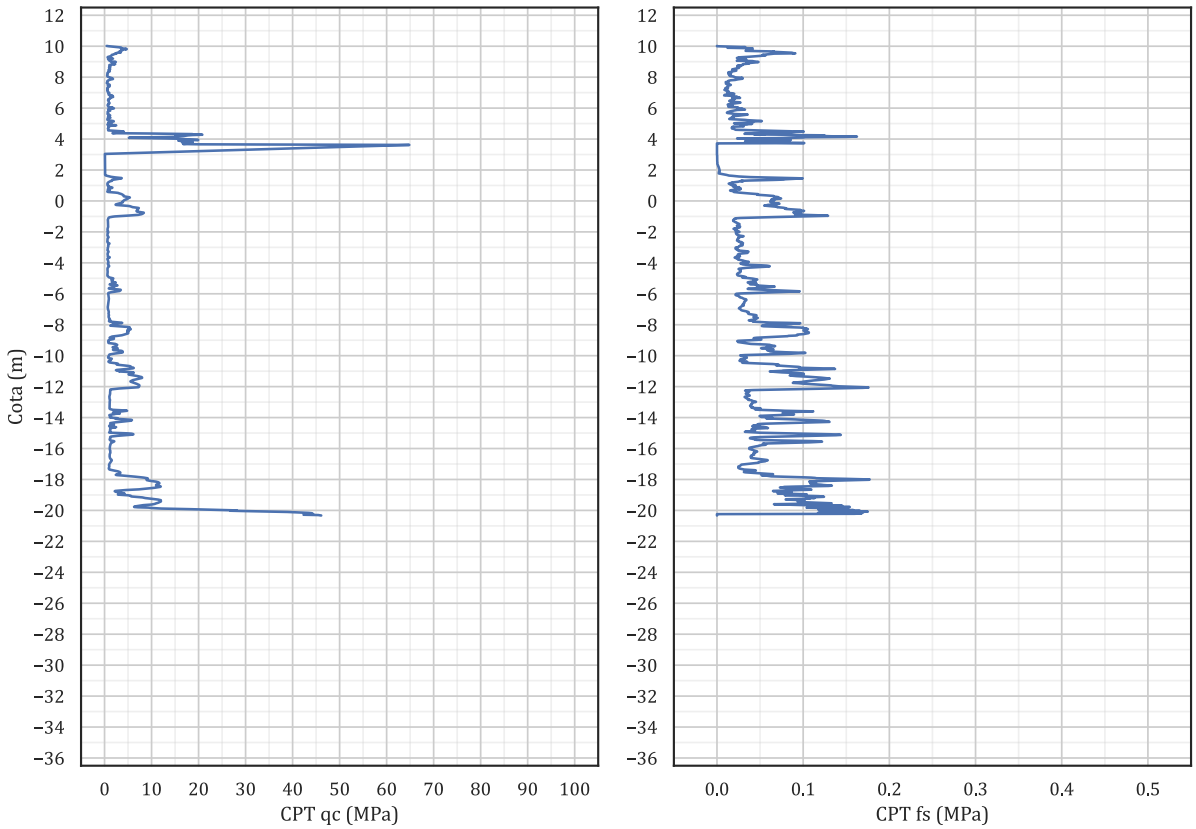
CPT-ZF-23



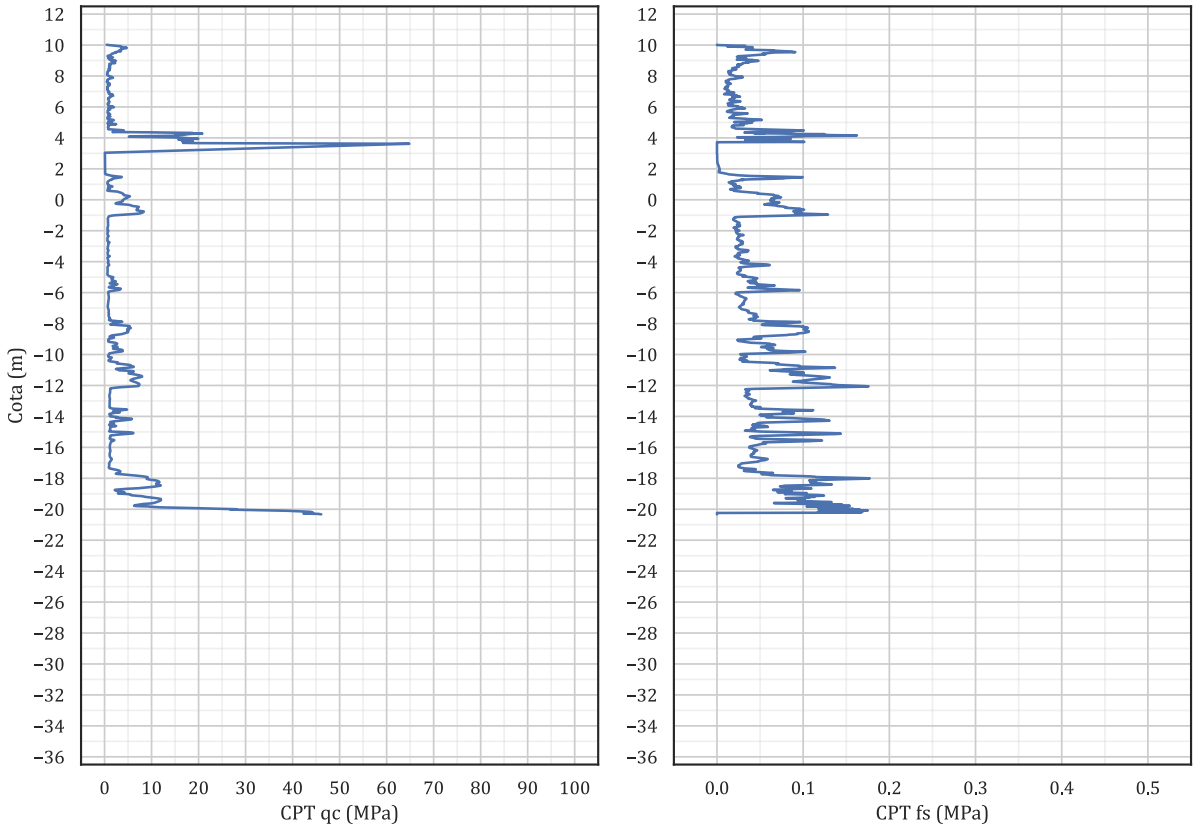
CPT-ZF-26



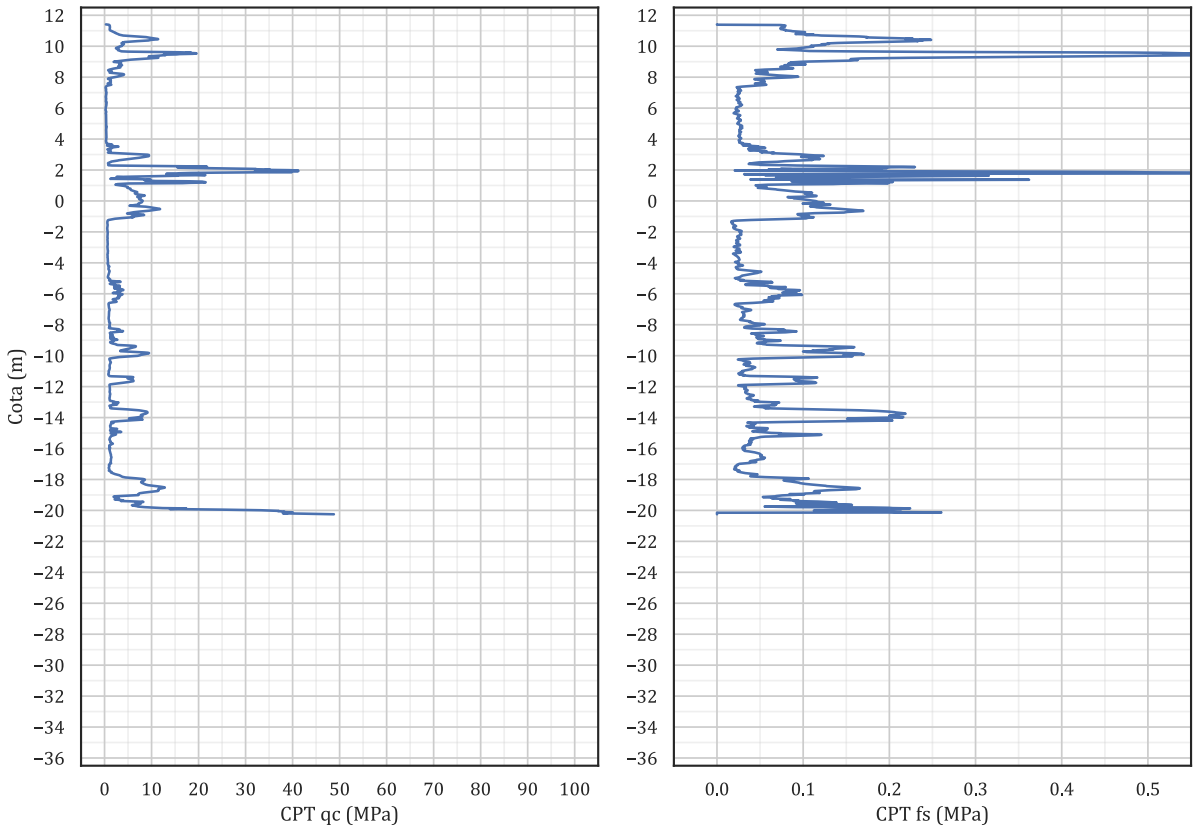
CPT-ZF-27



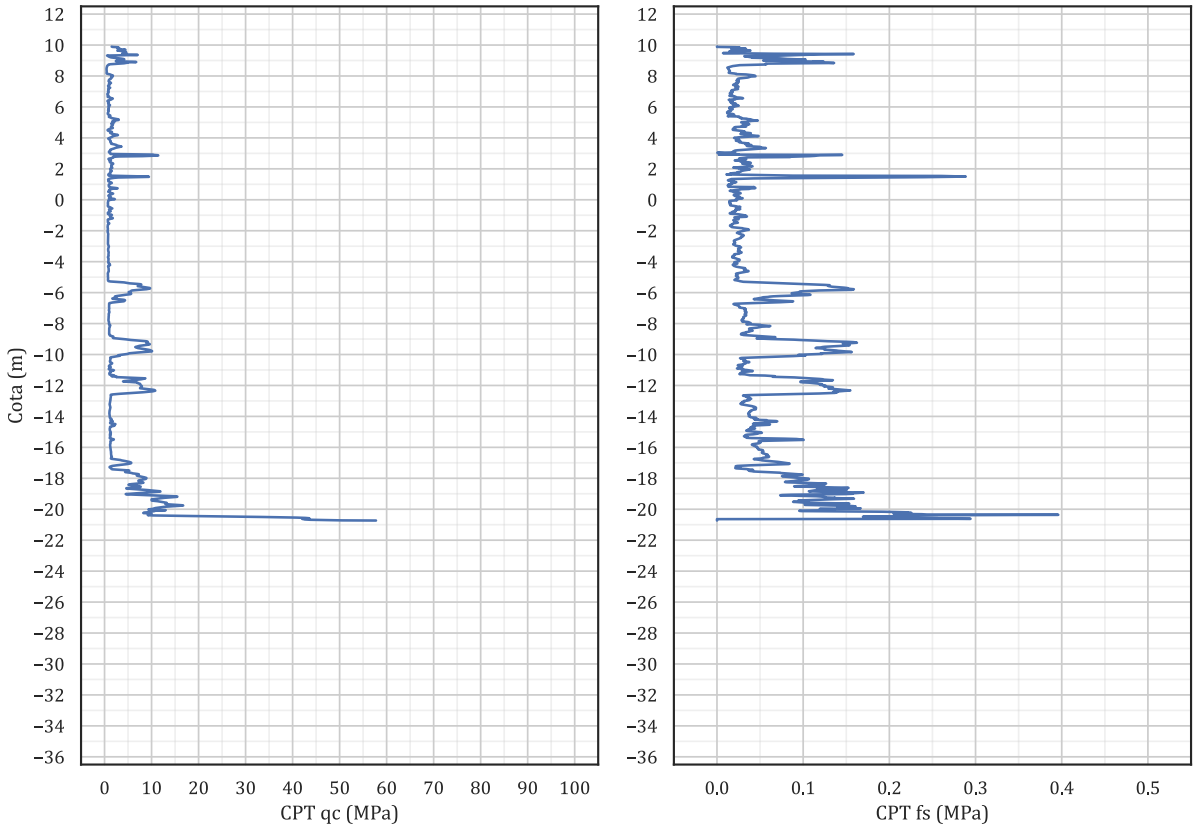
CPT-ZF-27



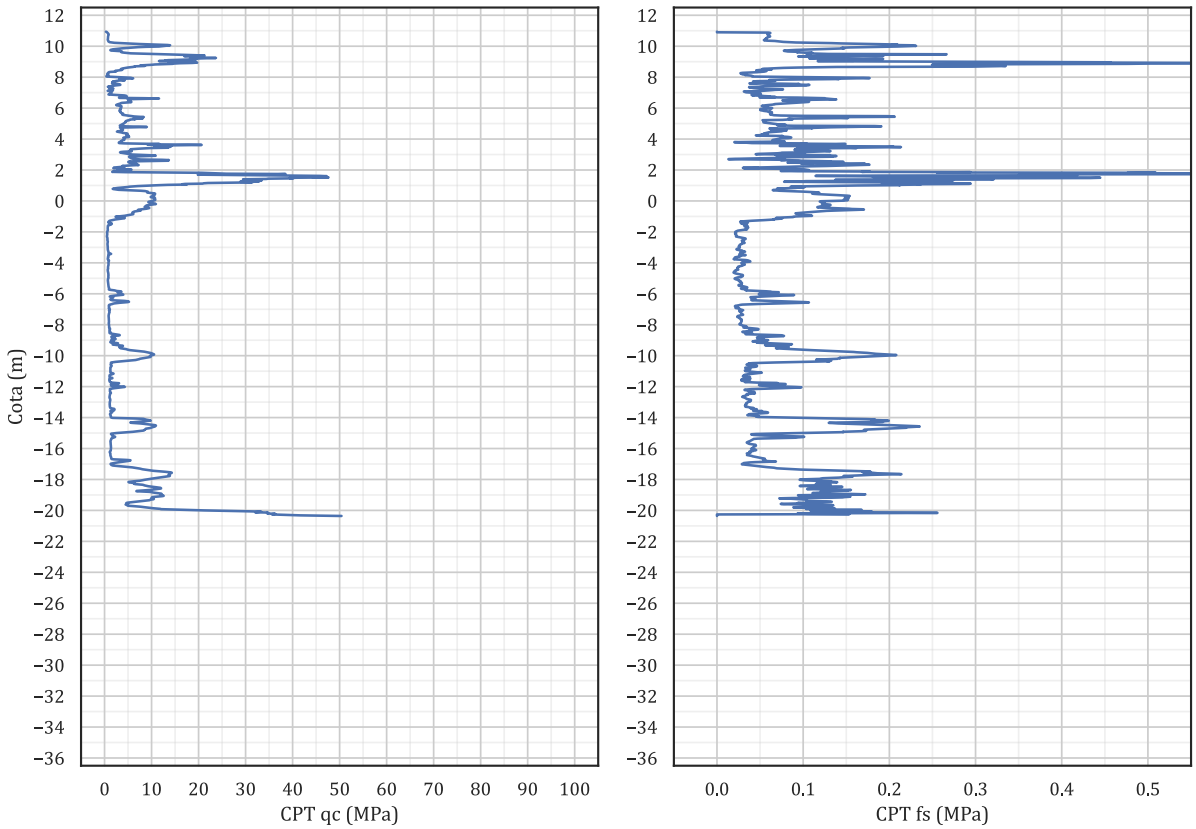
CPT-ZF-30



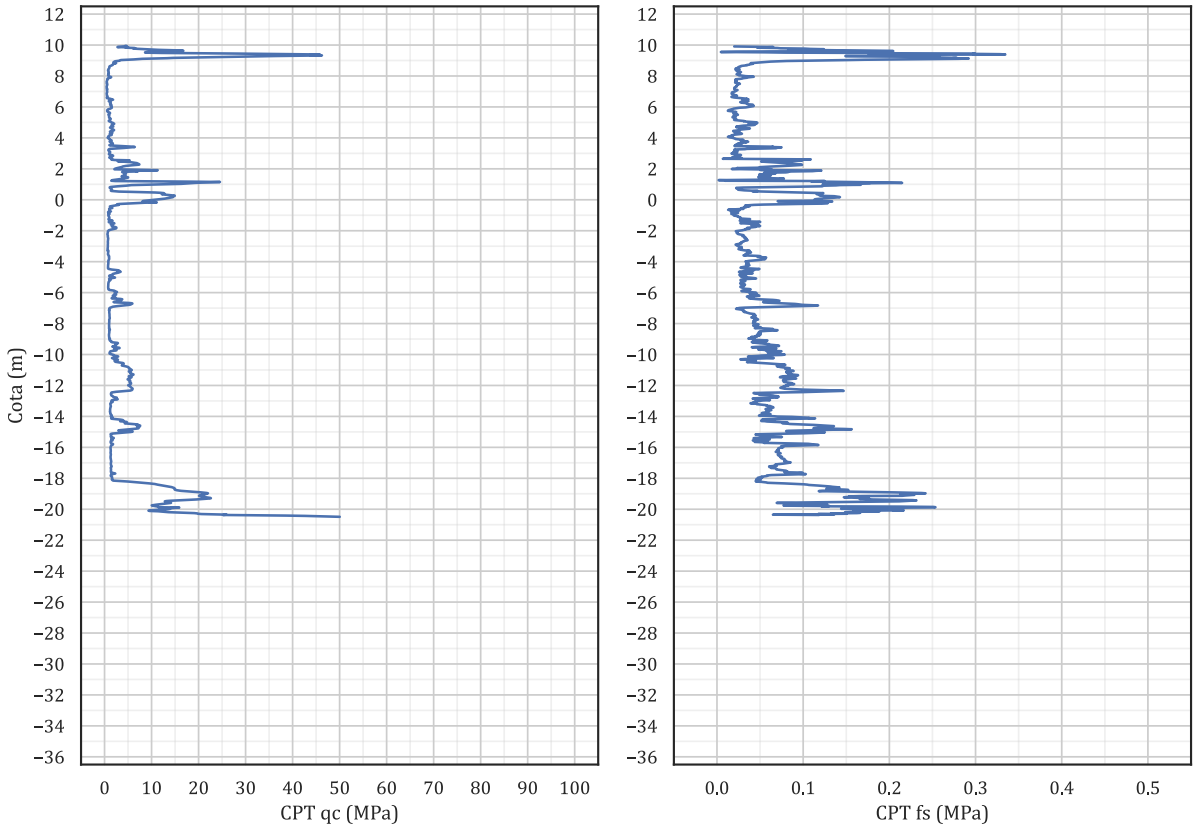
CPT-ZF-31



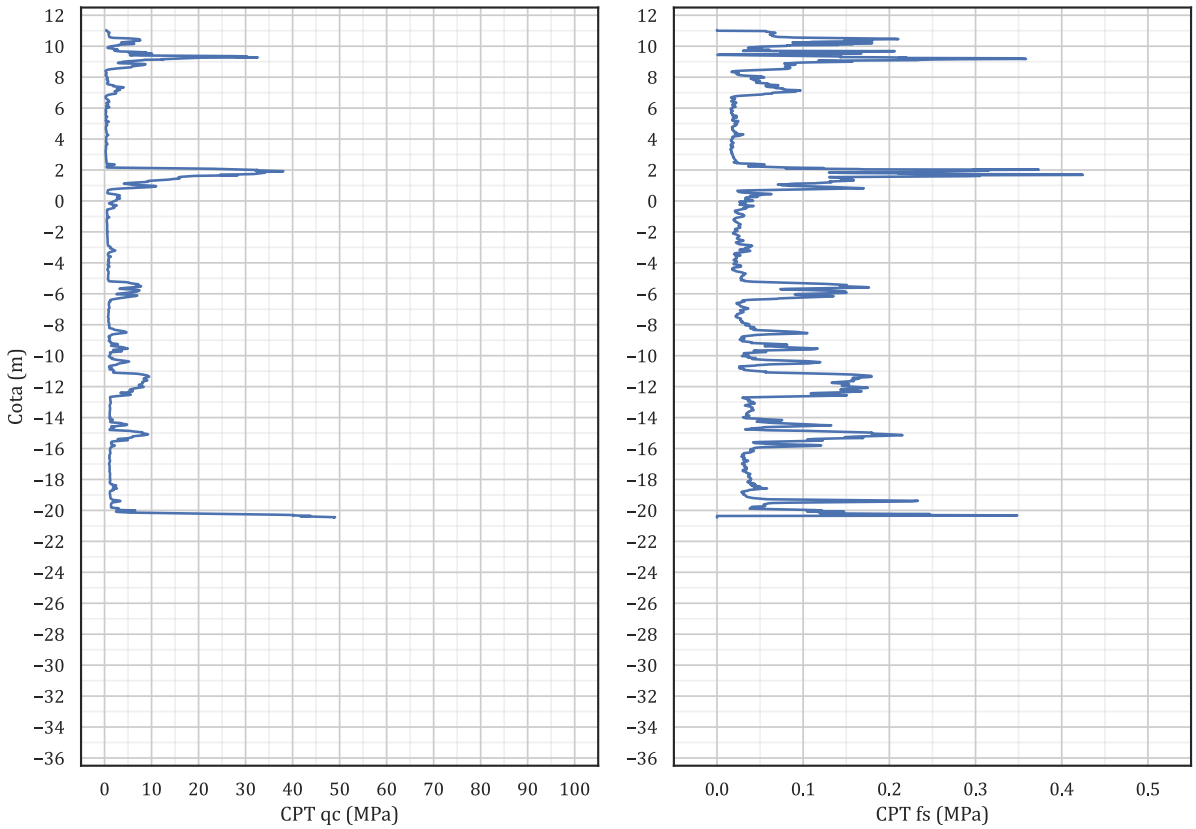
CPT-ZF-33



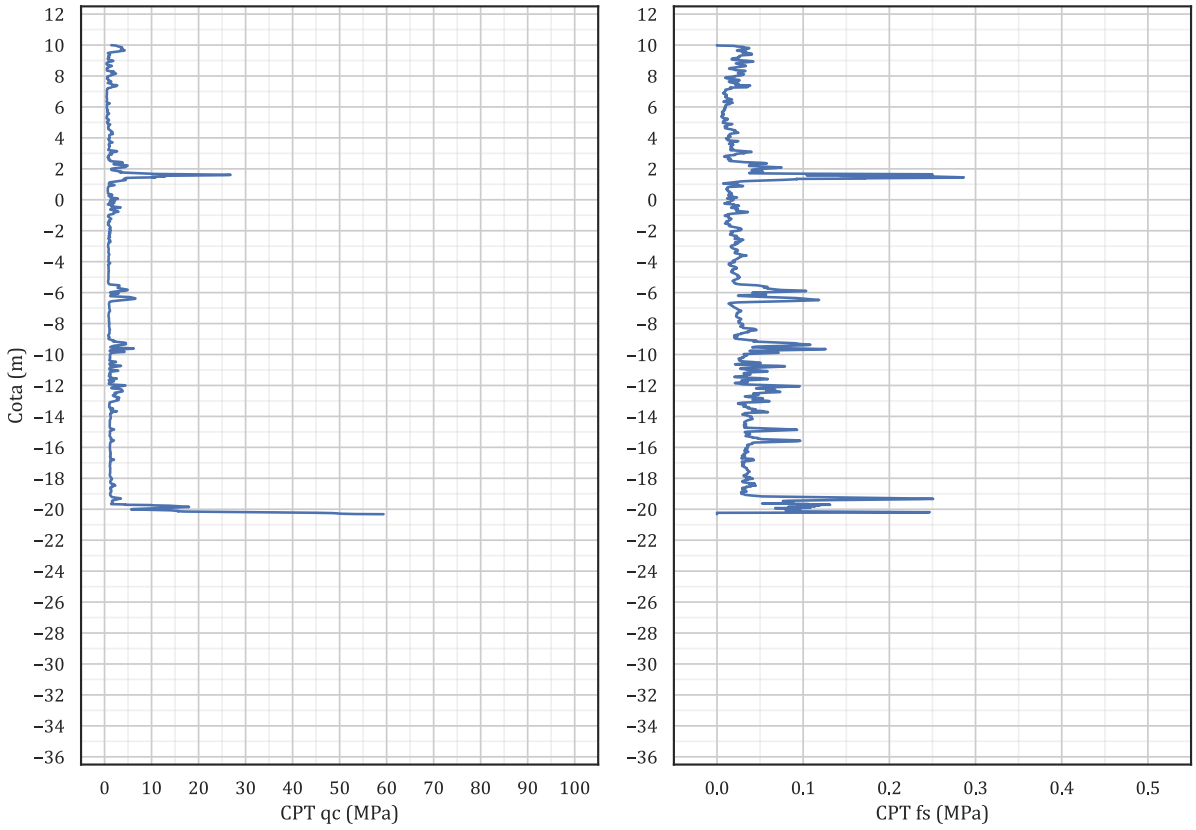
CPT-ZF-34



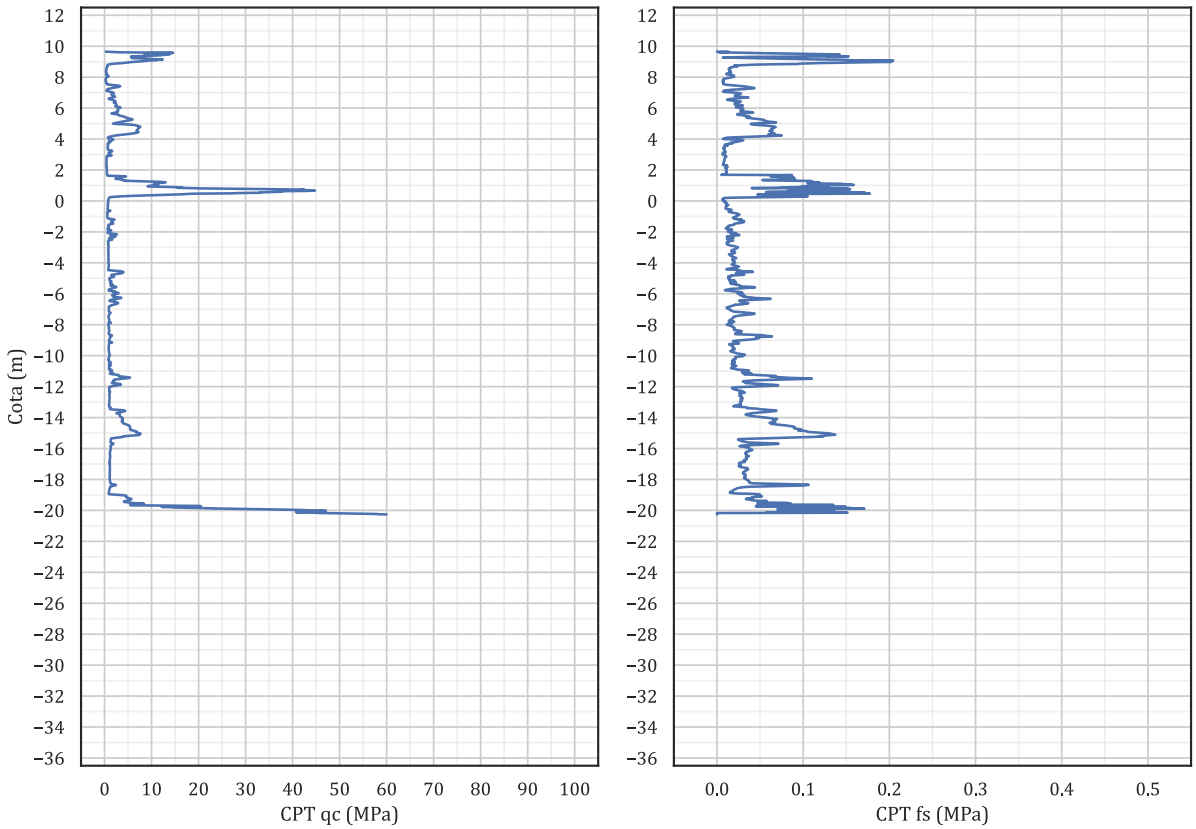
CPT-ZF-36



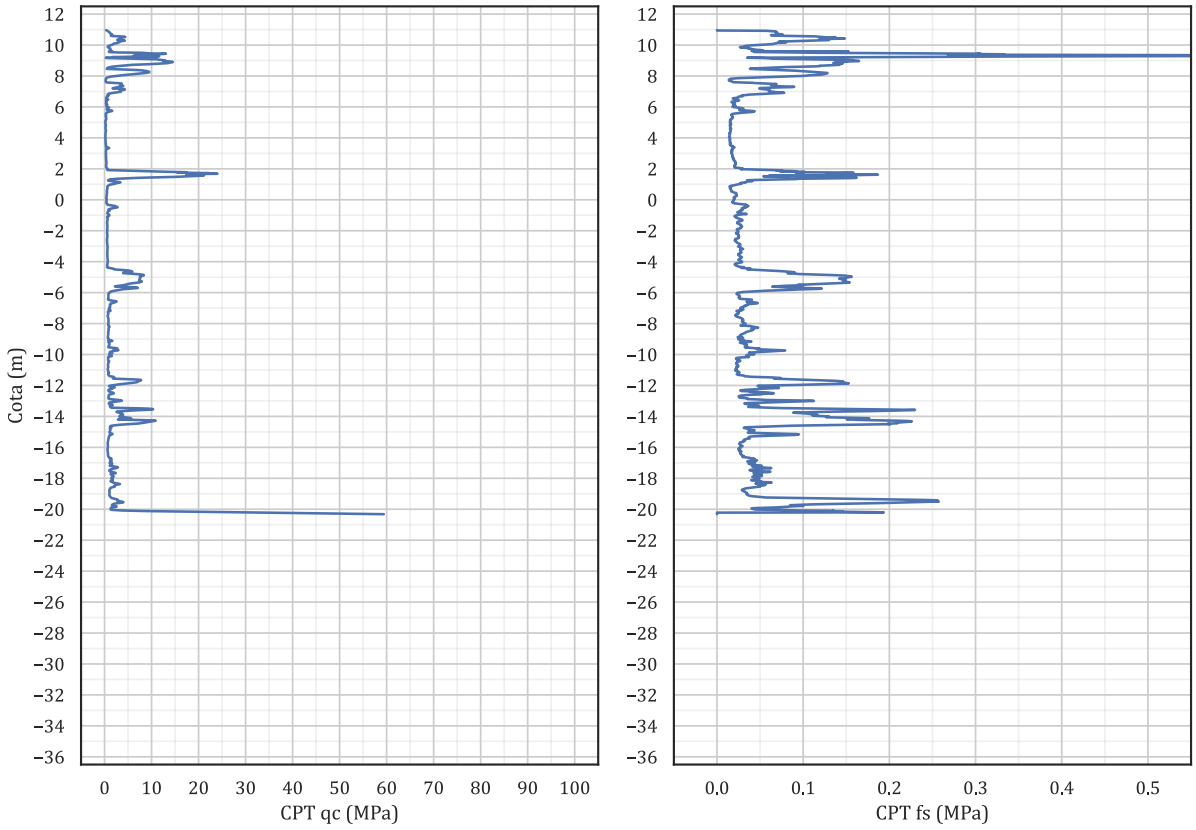
CPT-ZF-37



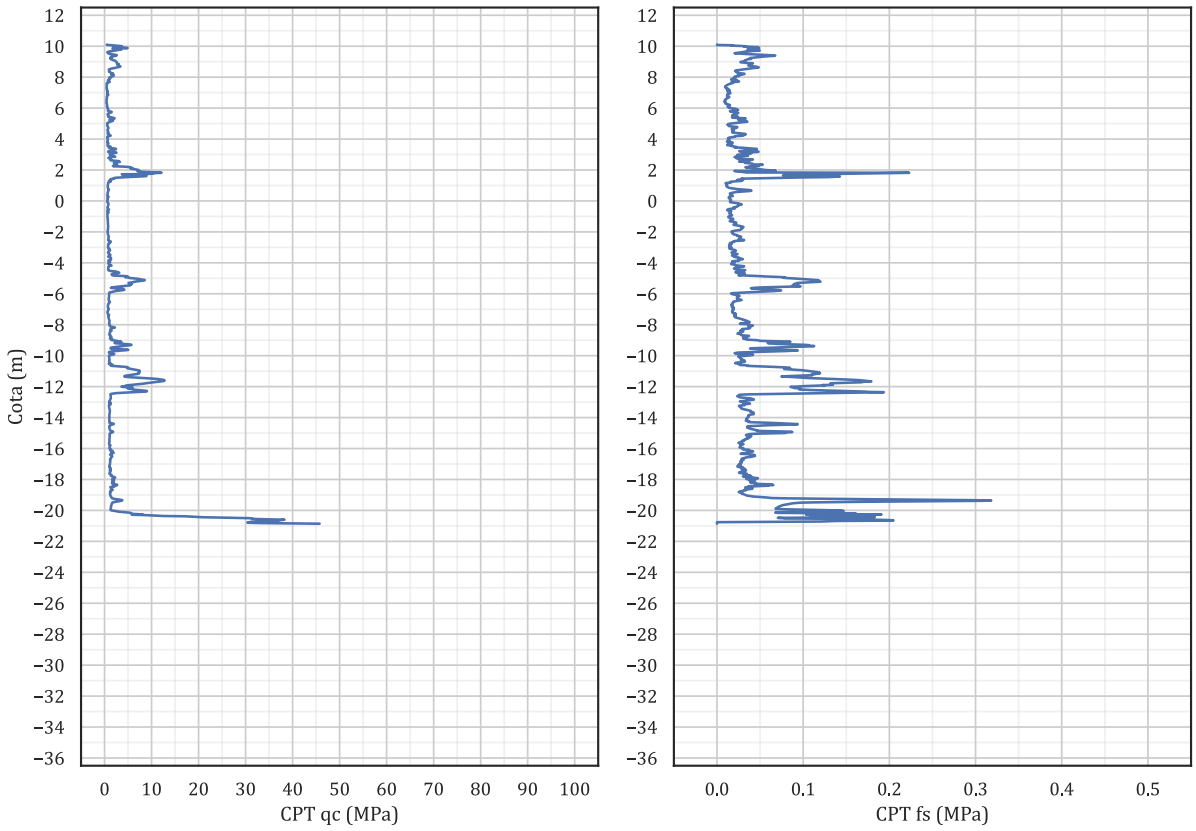
CPT-ZF-39



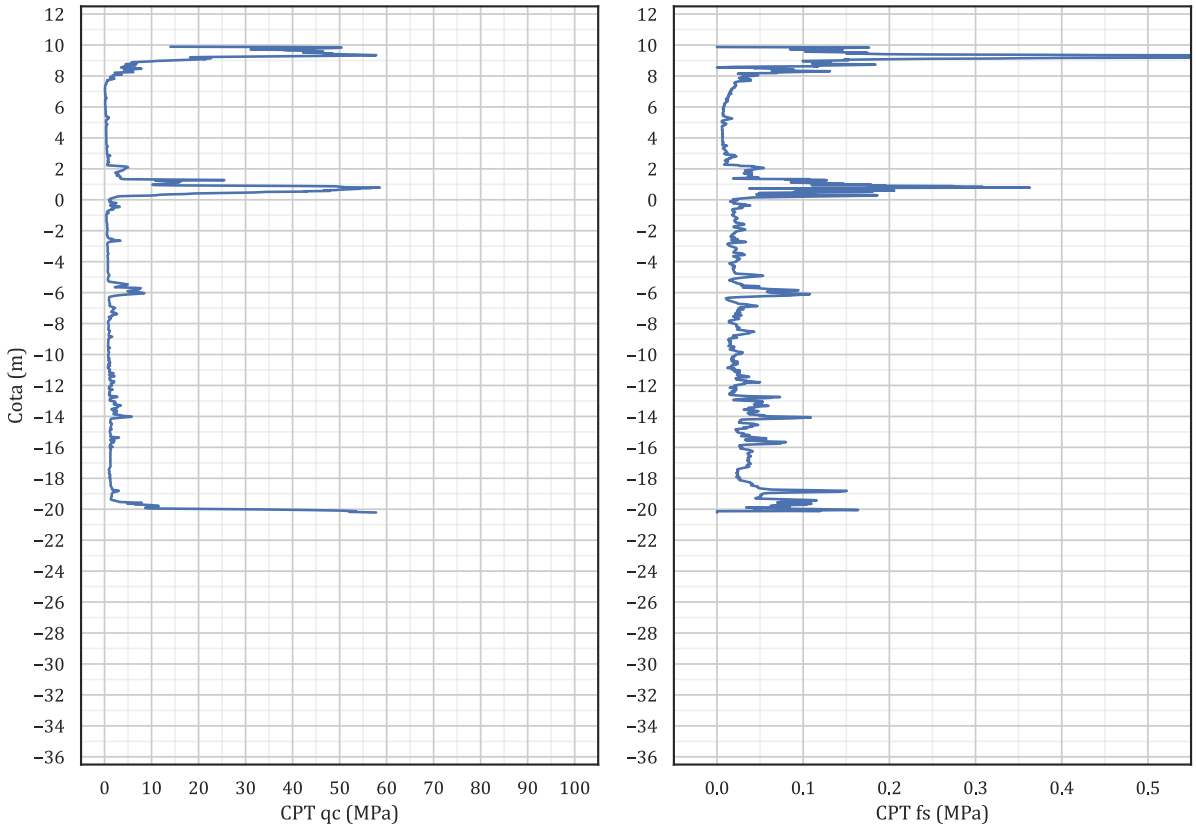
CPT-ZF-40



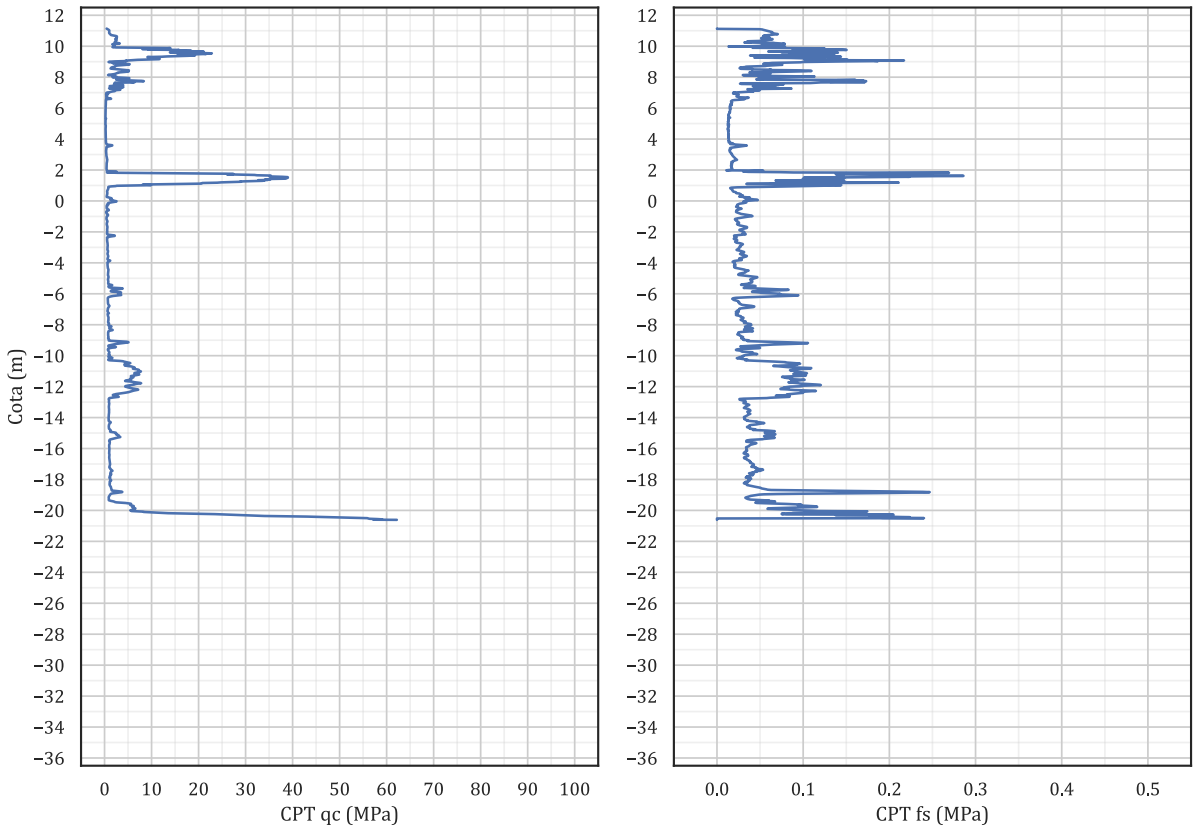
CPT-ZF-41



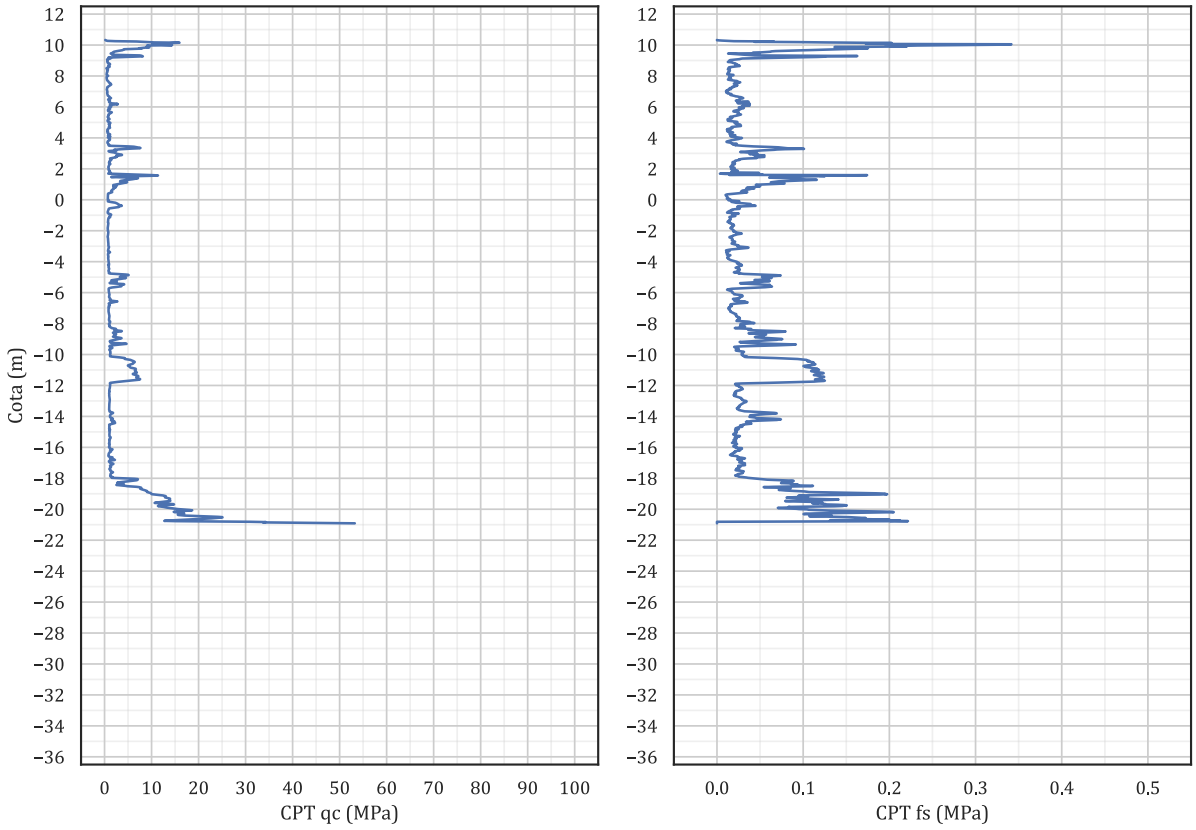
CPT-ZF-42



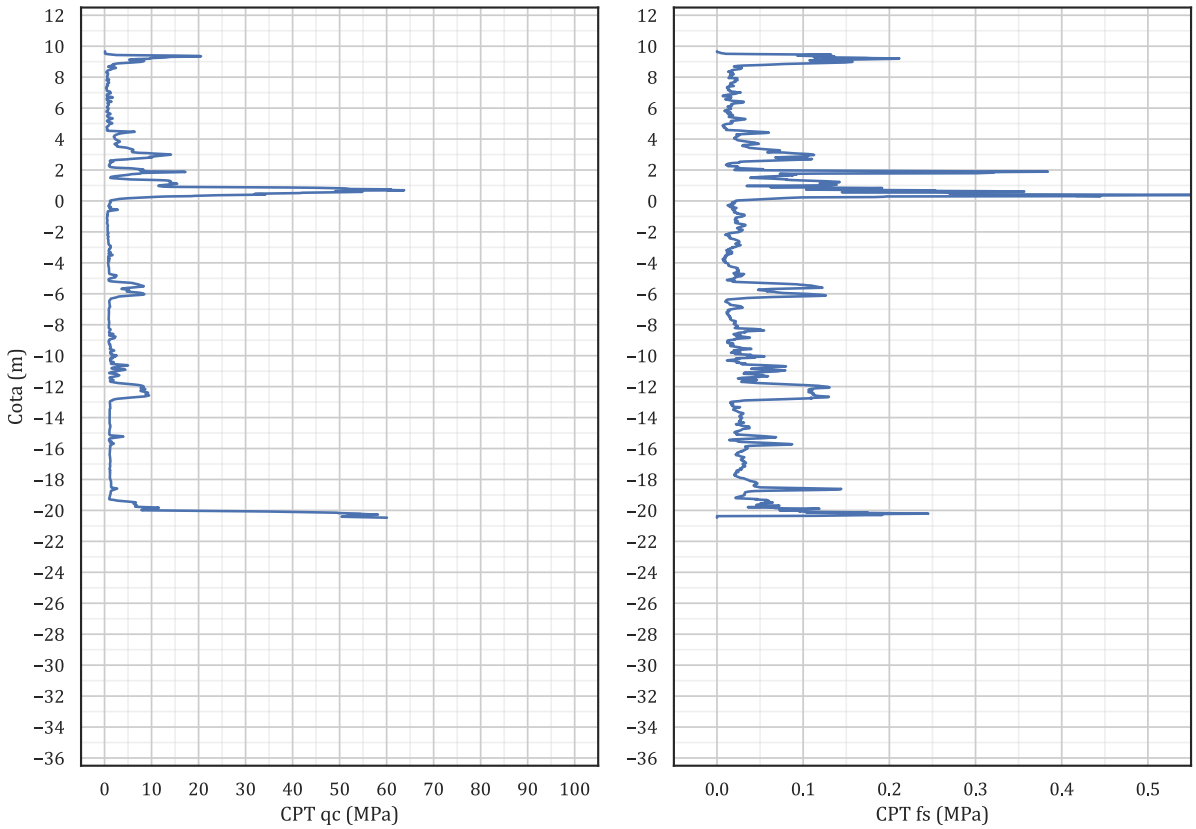
CPT-ZF-43



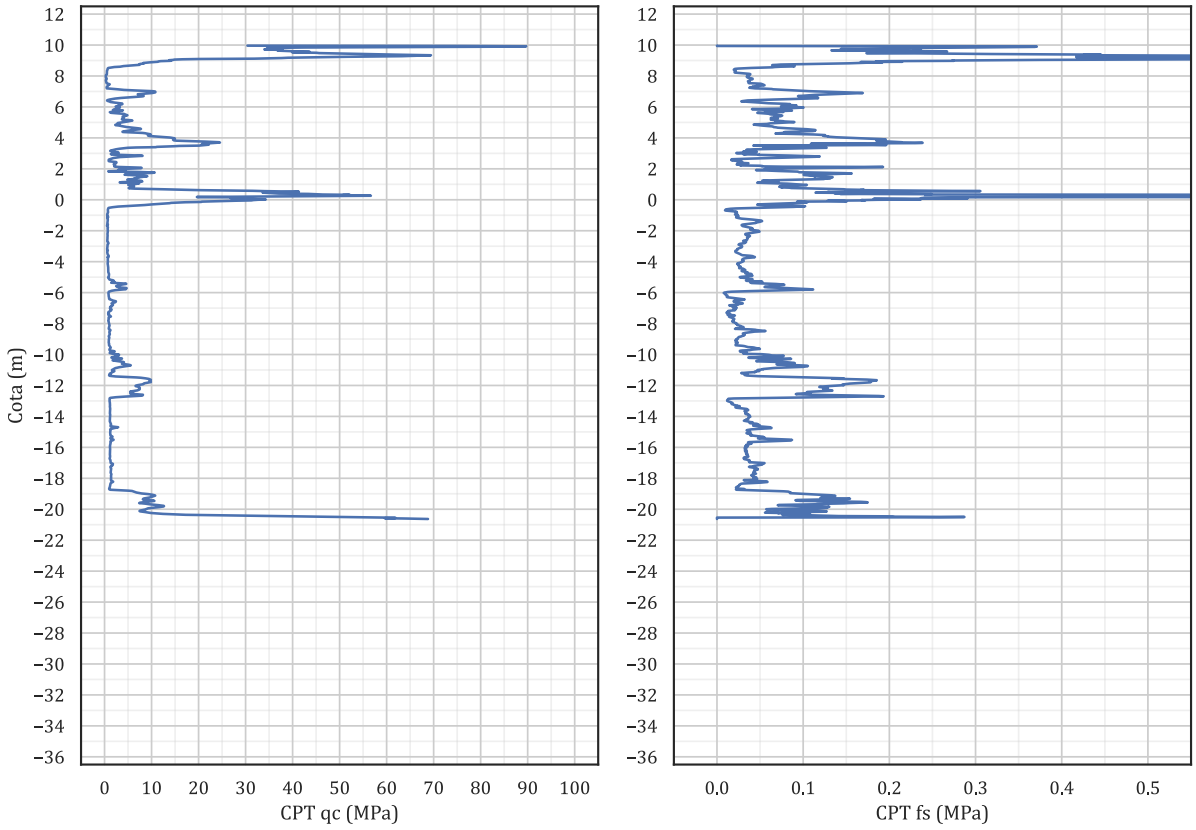
CPT-ZF-44



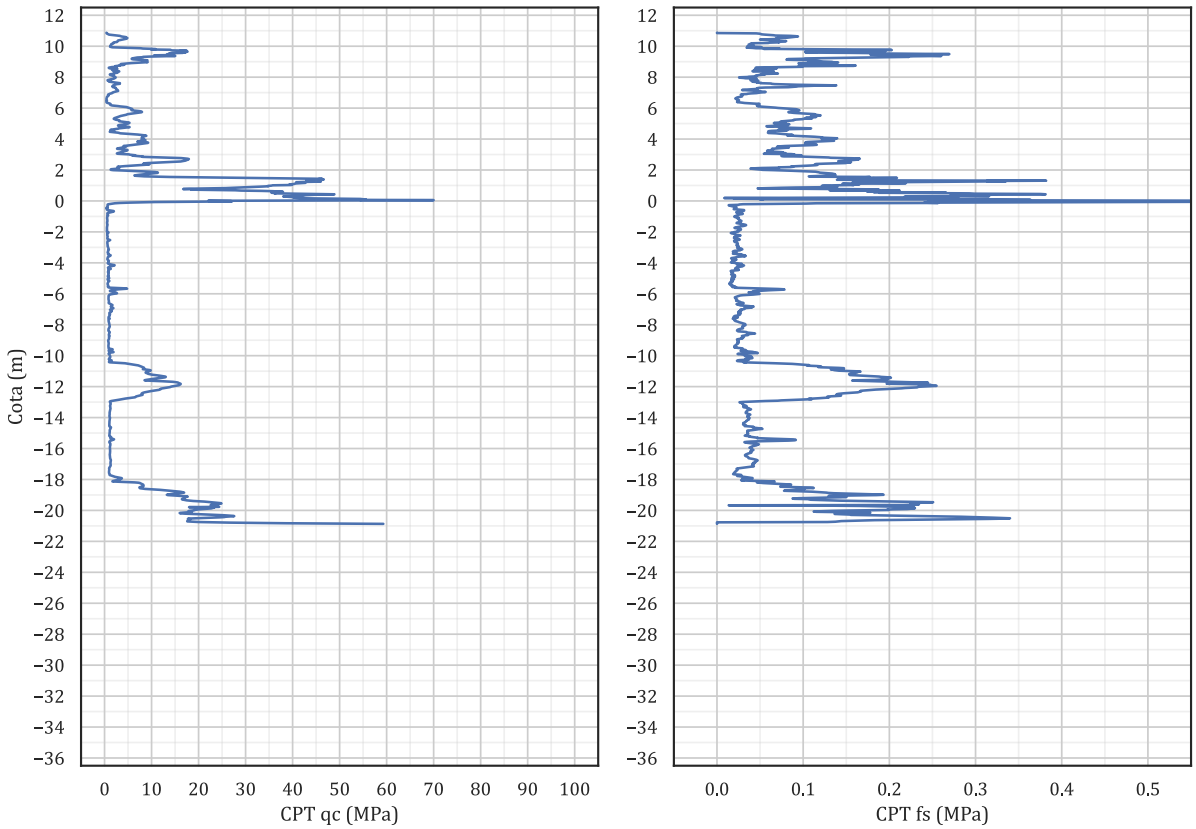
CPT-ZF-46



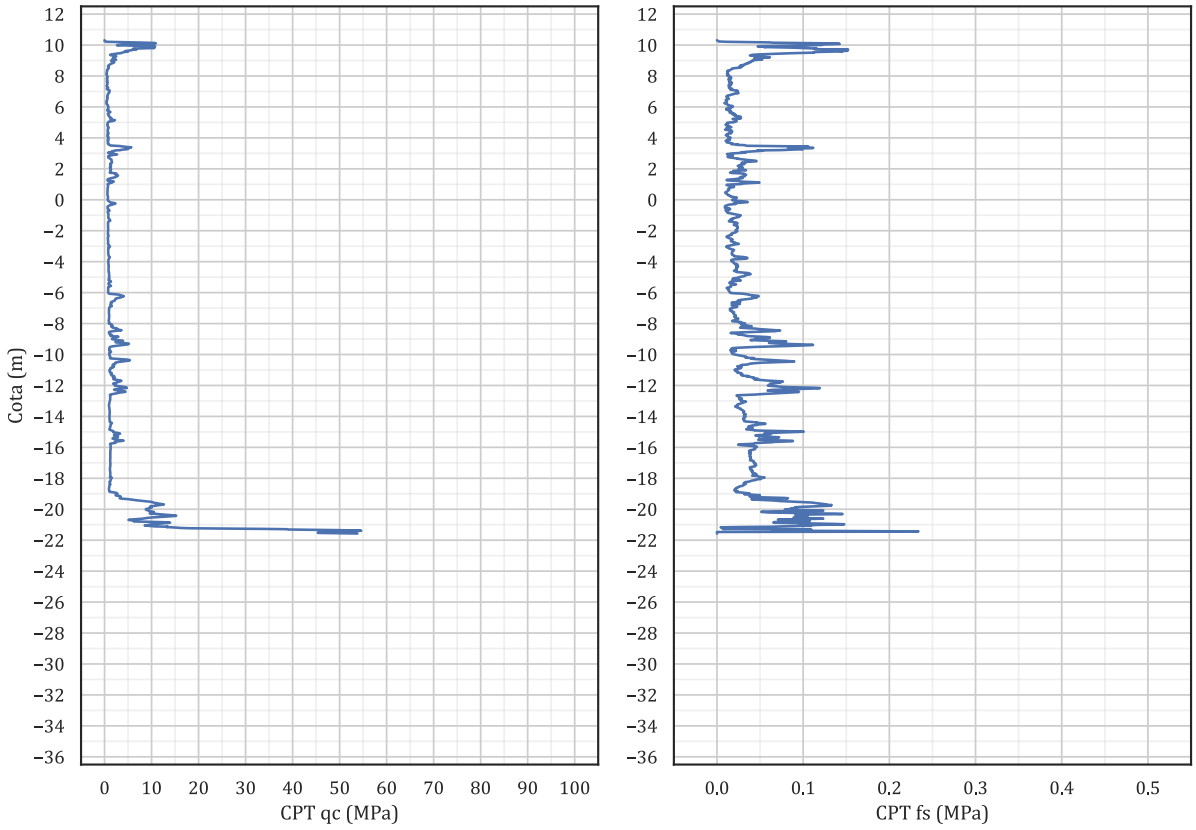
CPT-ZF-47



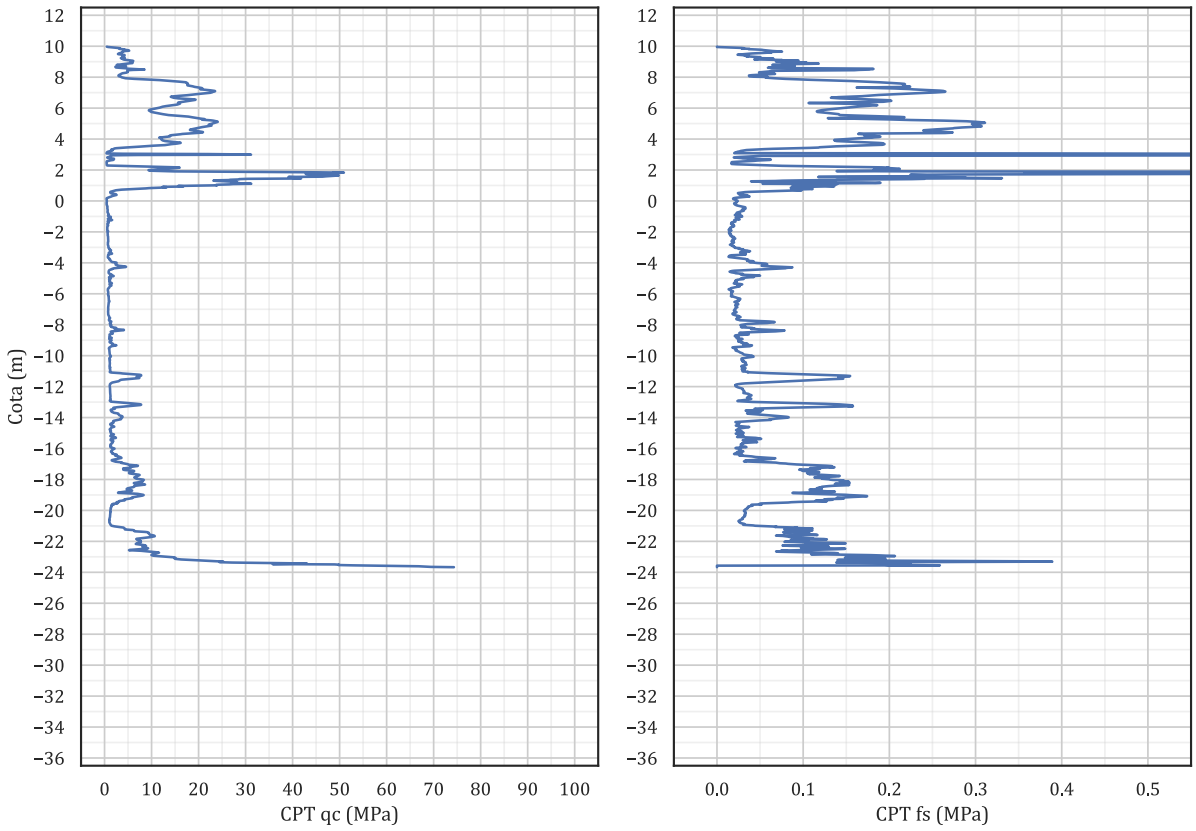
CPT-ZF-48



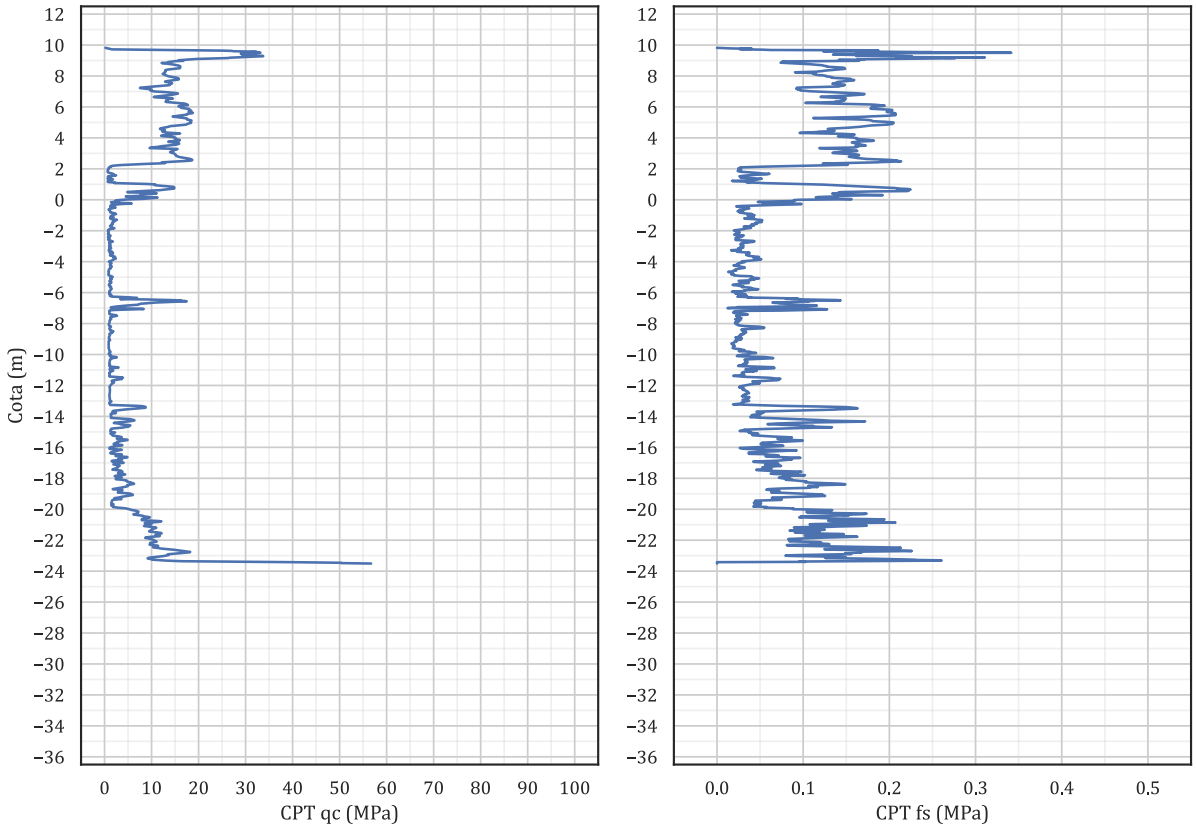
CPT-ZF-49



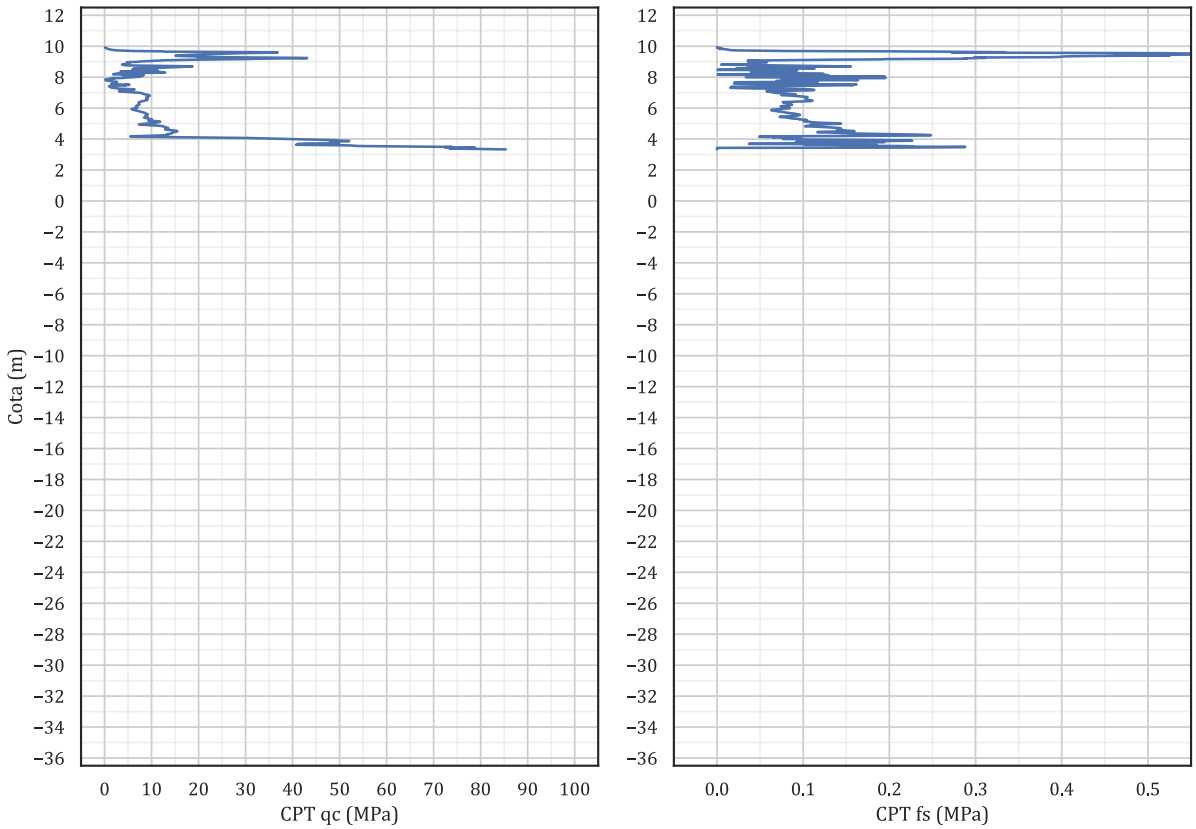
CPT-ZF-55



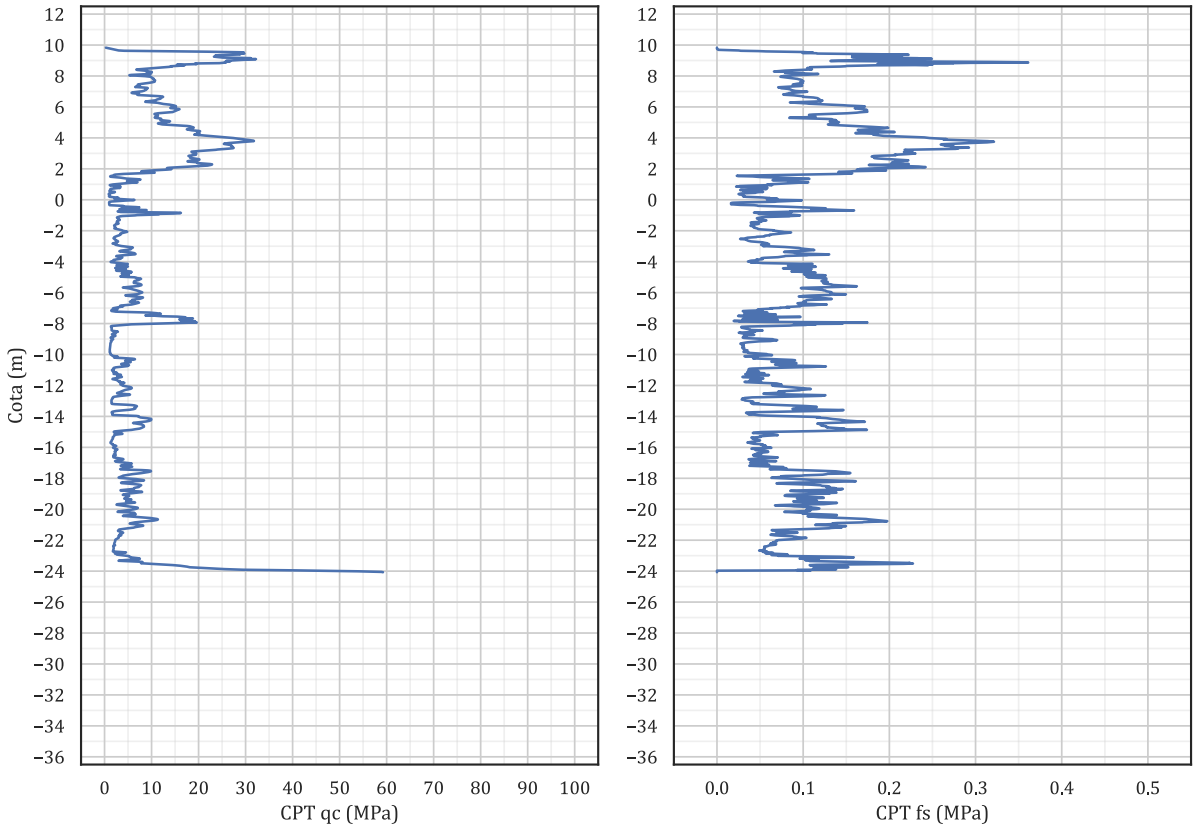
CPT-ZF-59



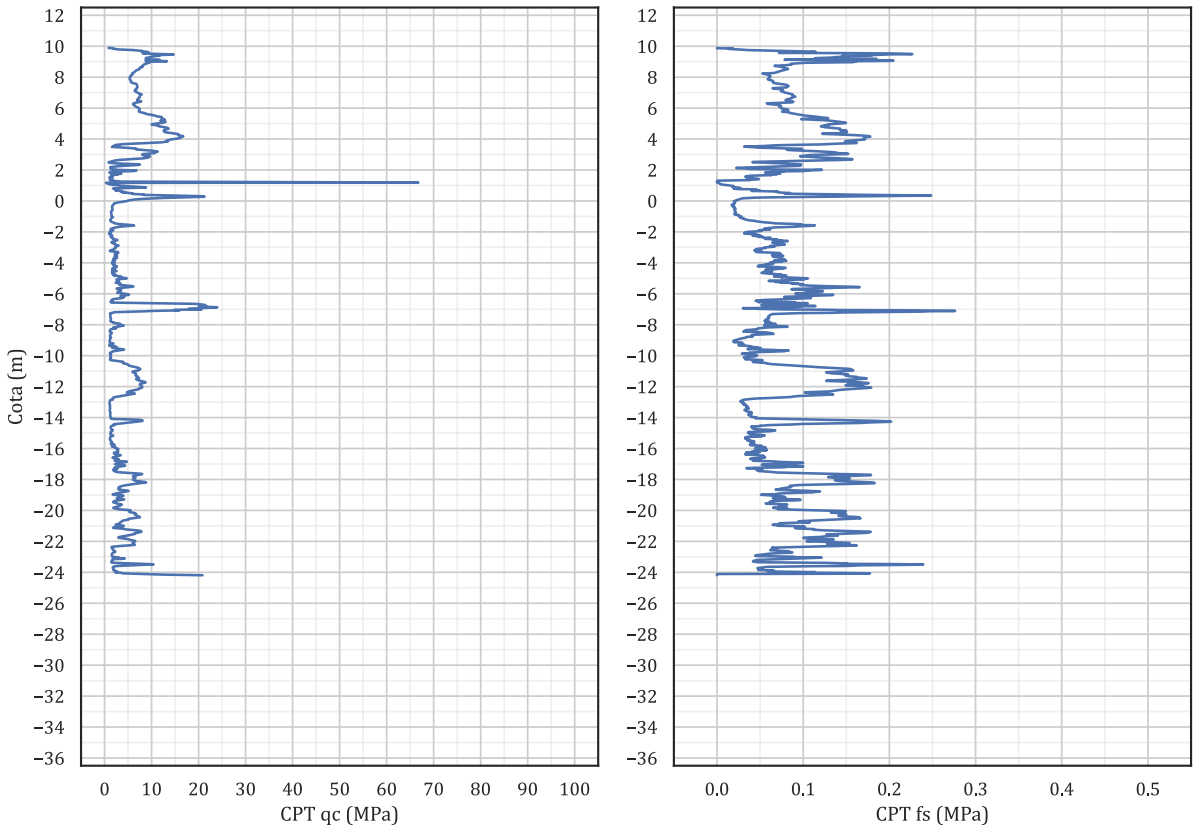
CPT-ZF-60



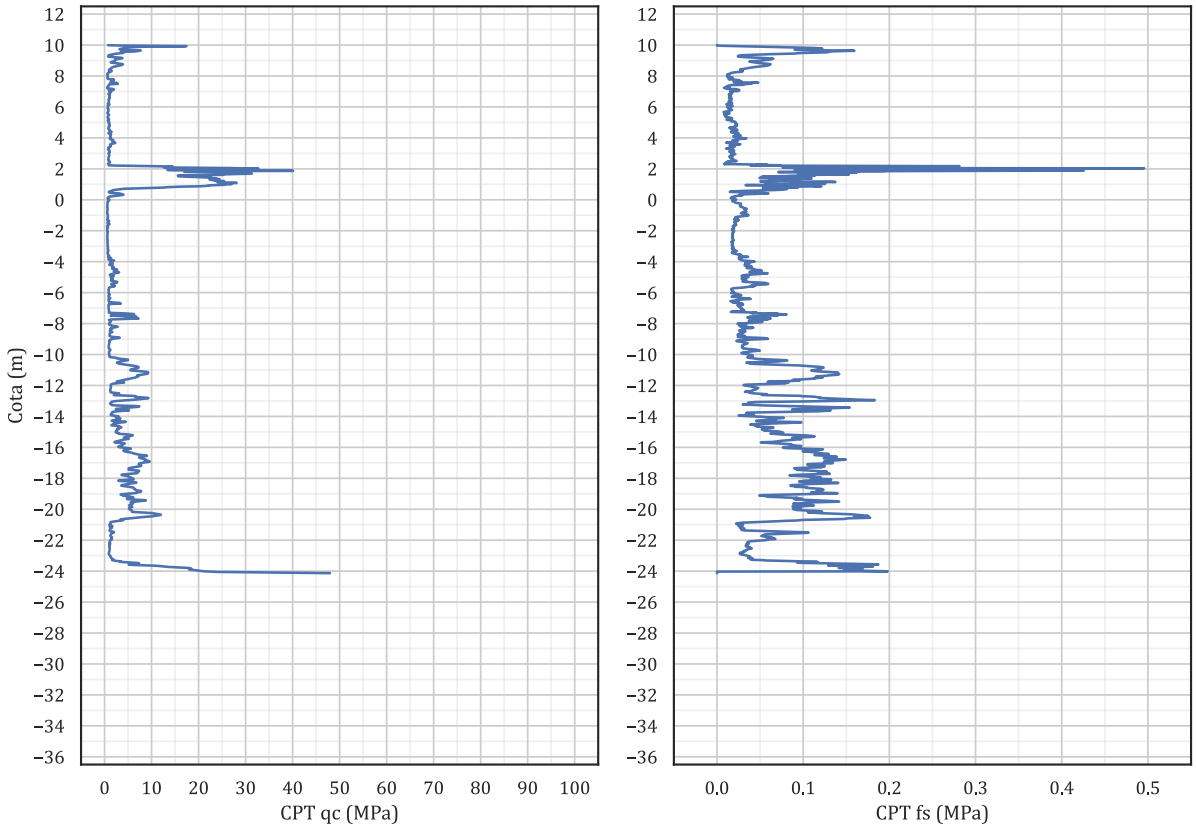
CPT-ZF-61



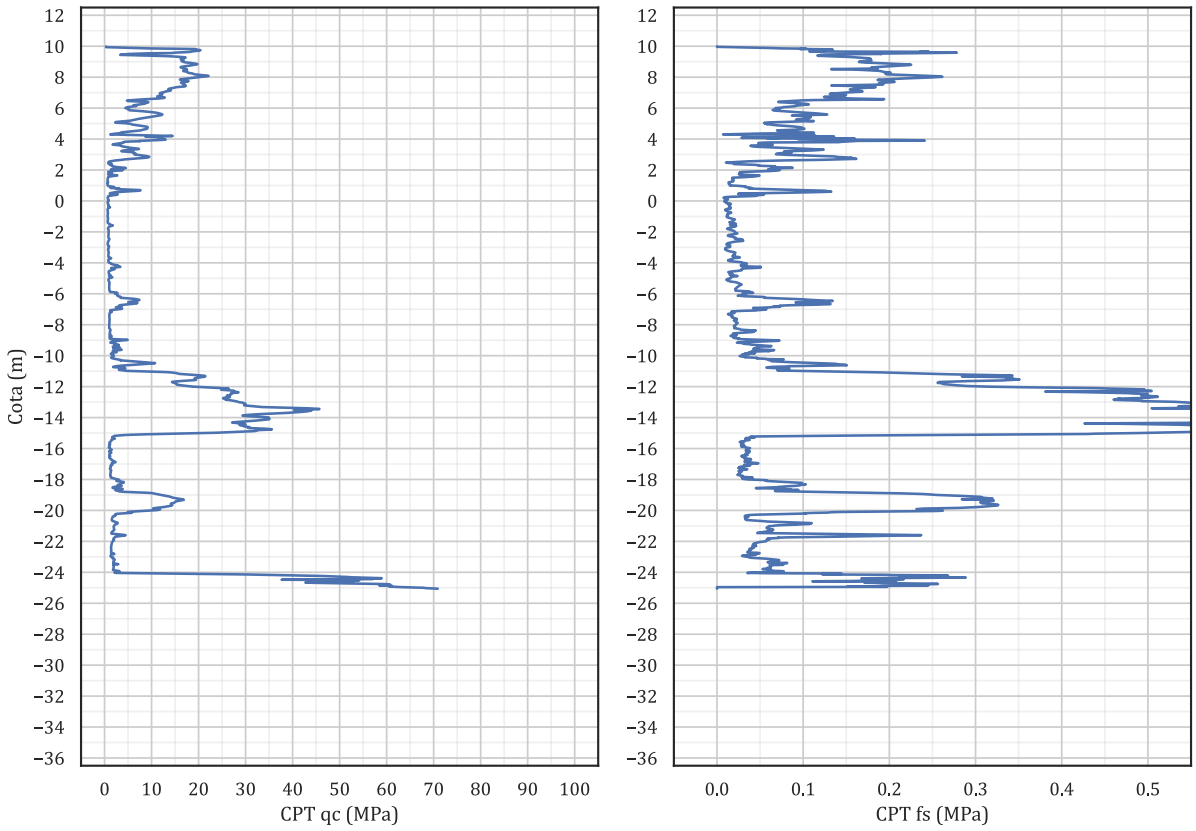
CPT-ZF-62



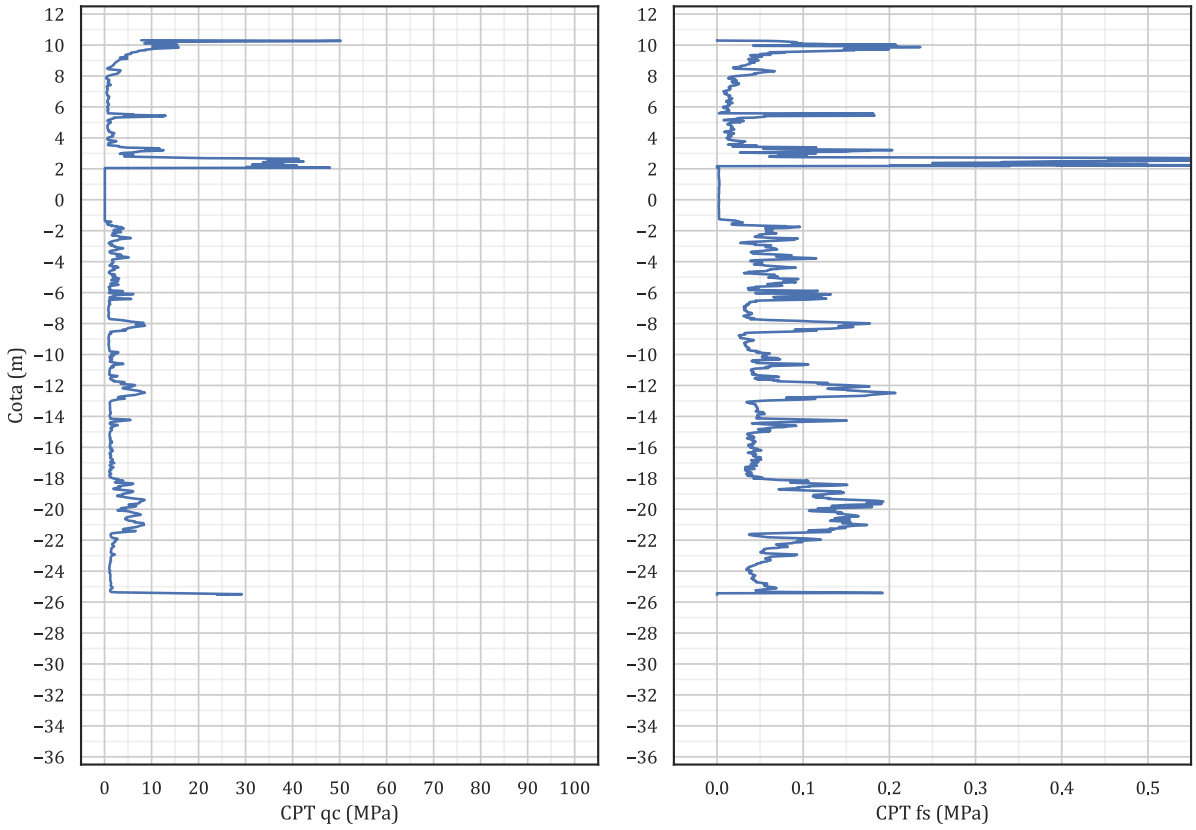
CPT-ZF-63



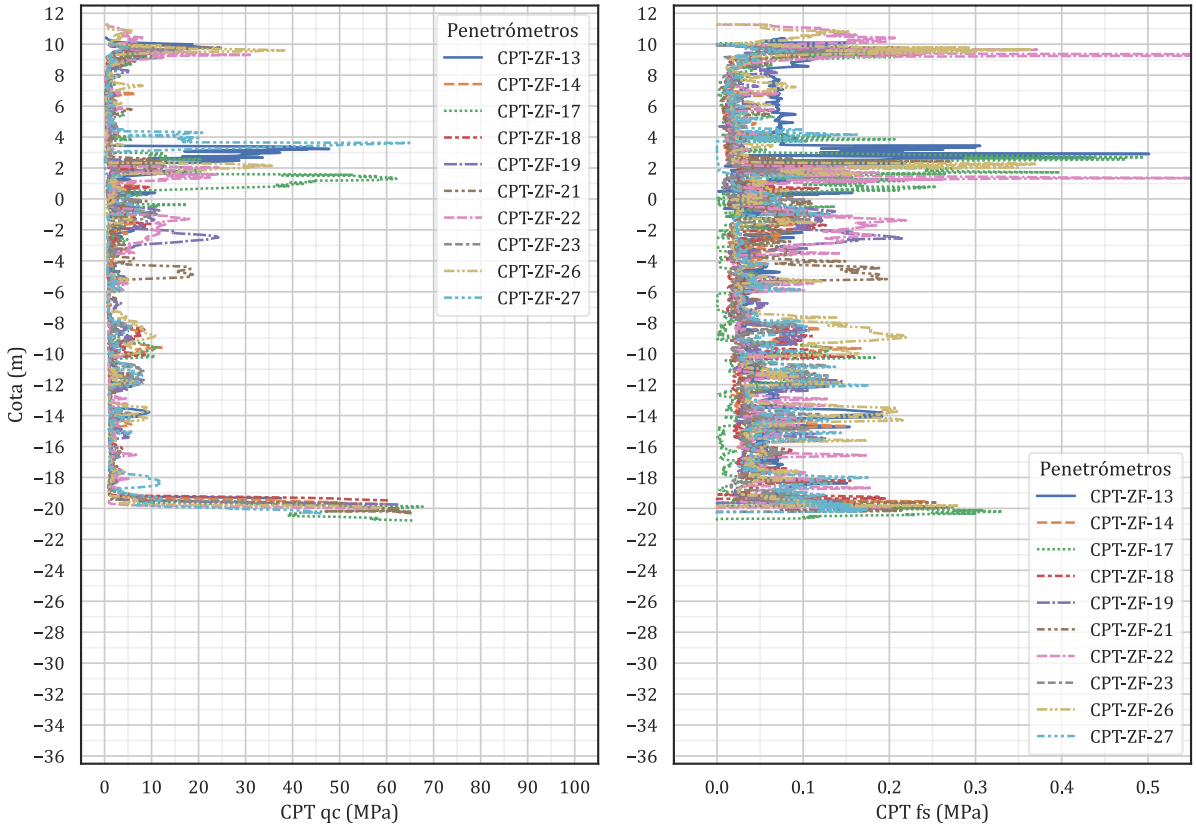
CPT-ZF-64



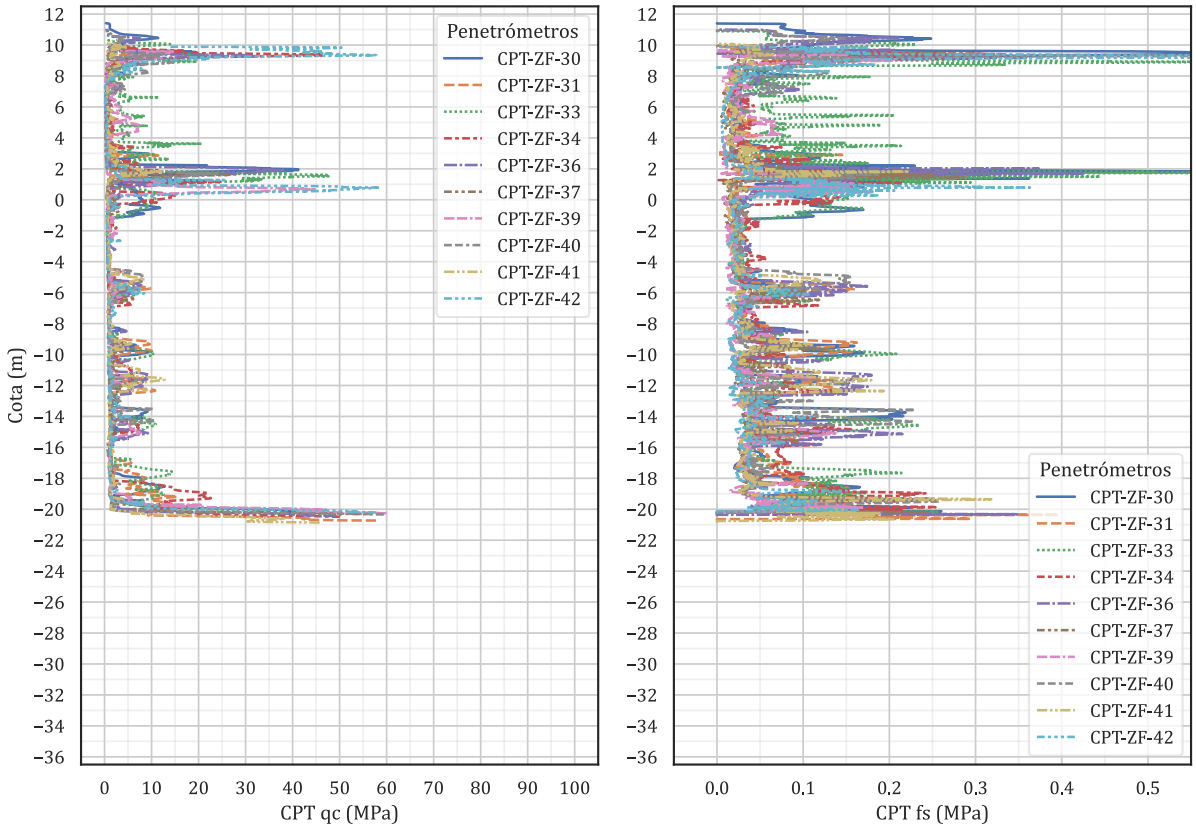
CPT-ZF-66



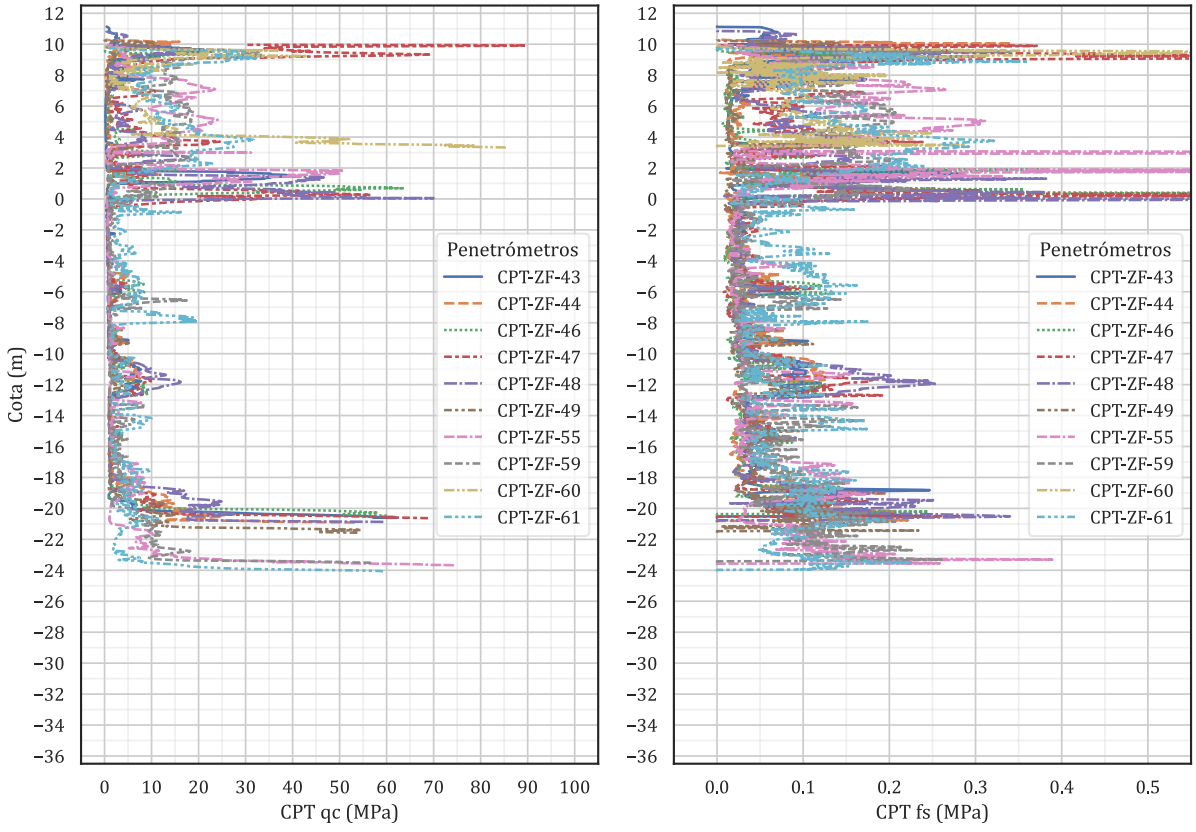
Penetrómetros



Penetrómetros



Penetrómetros



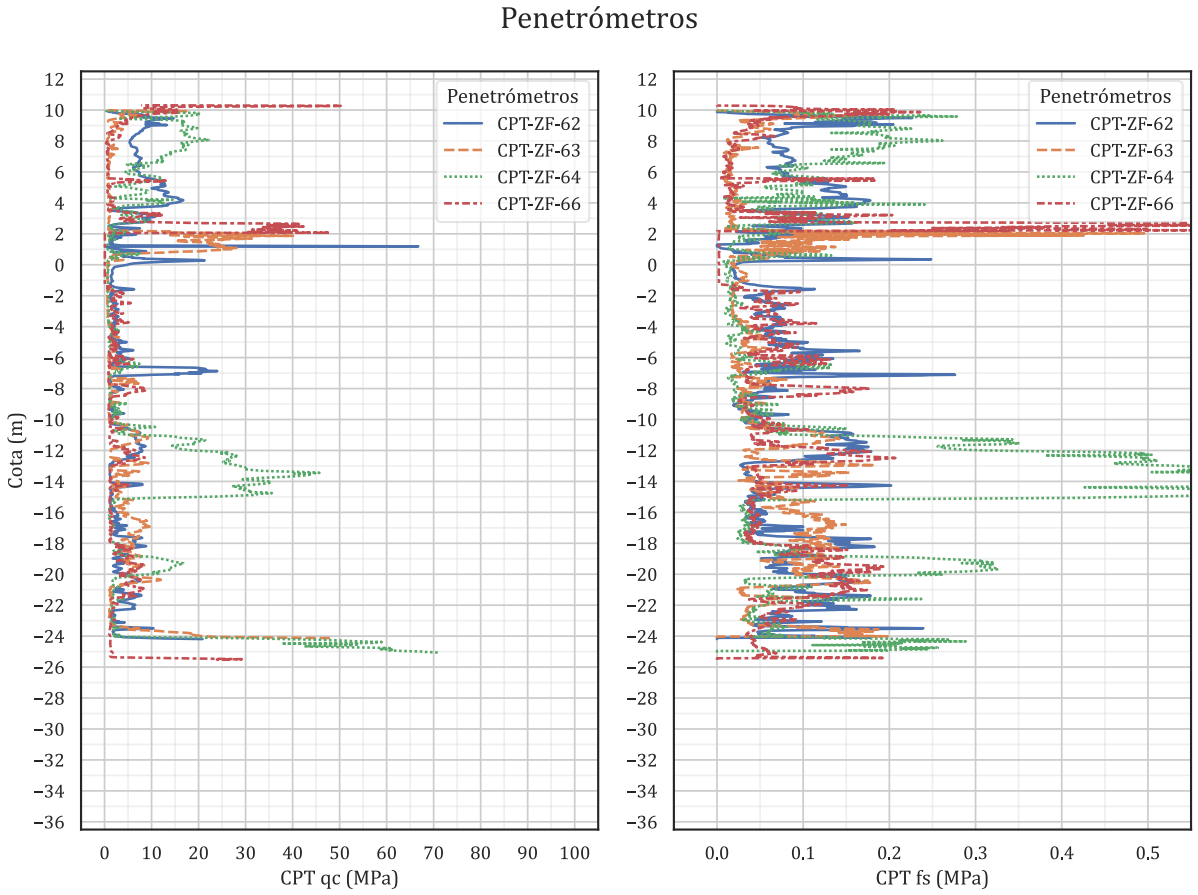


Figura 3.11: Resistencia por punta (izquierda) y por fuste (derecha) en función de la cota para los penetrómetros de la Obra 5

3.2. Herramientas utilizadas

Para conseguir este objetivo se utilizan algoritmos de ML aplicados mediante la librería scikit-learn (Pedregosa et al., 2011) de Python (Python Software Foundation, 2022). Se parte de una instalación mediante Anaconda y se utiliza como interfaz Jupyter Notebook.

Python es un lenguaje de programación muy popular en el uso de ML por su sencillez, flexibilidad y librerías disponibles. Anaconda (Anaconda Inc, 2016) es una distribución de Python y del lenguaje R pensada para computación científica (ciencia de datos, ML, análisis predictivos, etc.) y cuyo objetivo es facilitar la gestión de librerías, evitando conflictos. Scikit-learn es una librería extremadamente popular de ML que implementa algoritmos punteros, transformadores, preprocesos y demás herramientas necesarias para la ciencia de datos. Por último Jupyter Notebook (Kluyver et al., 2016) es una aplicación para

la creación de cuadernos computacionales. Los cuadernos computacionales son documentos que combinan código, descripciones en lenguaje natural, visualizaciones como gráficos y figuras, y controles interactivos. Estos cuadernos proporcionan una herramienta rápida para probar y explicar código, explorar y visualizar datos y para compartir los resultados obtenidos.

Gracias a Anaconda, la aplicación de estas herramientas es tan sencilla como instalar Anaconda Navigator, abrir un Jupyter Notebook y comenzar a programar (Figura 3.12).

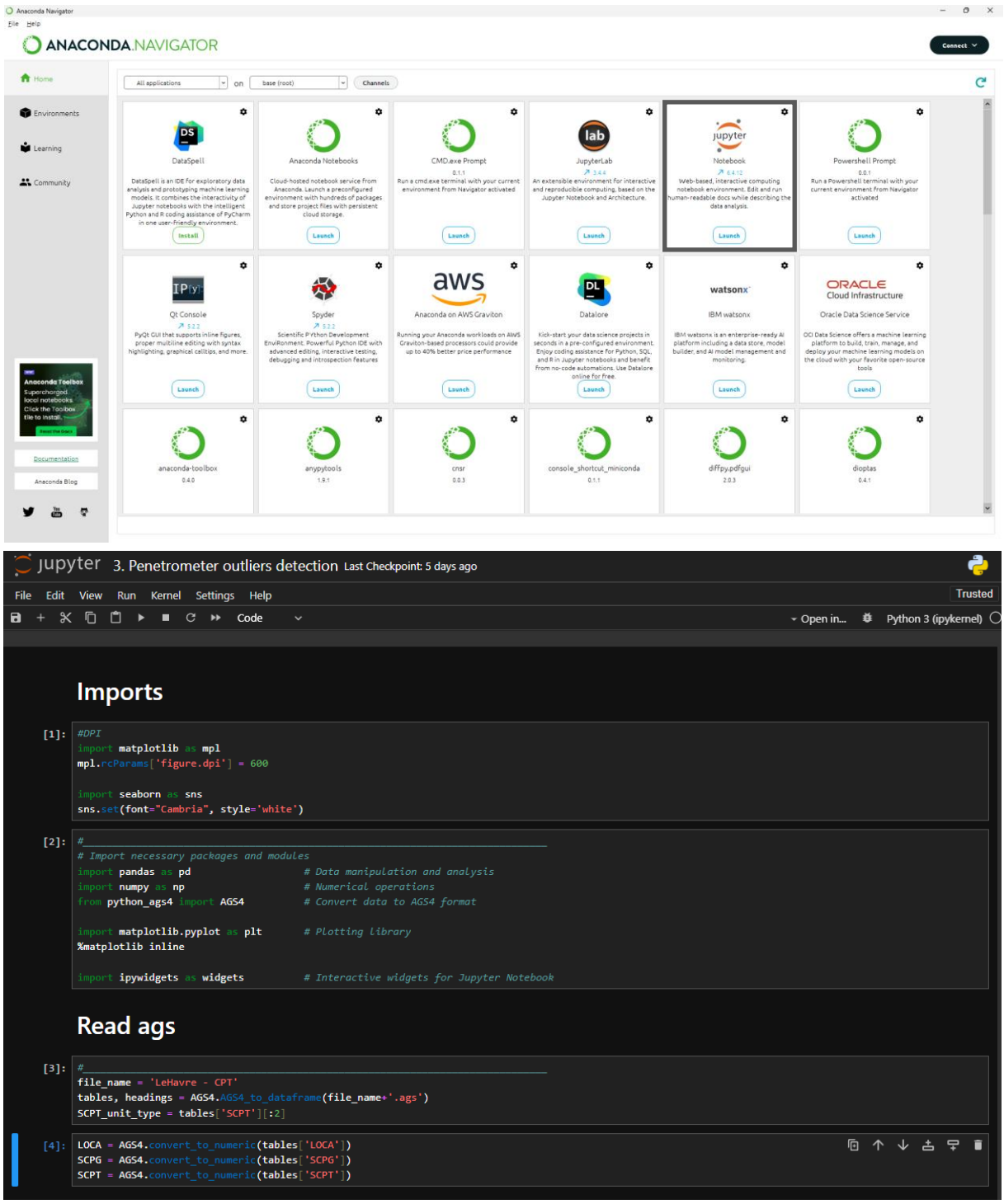


Figura 3.12: Pestaña inicial de Anaconda Navigator (superior). Interfaz de un Jupyter Notebook (inferior)

Scikit-learn tiene una API (Application Programming Interface, Interfaz de Programación de Aplicaciones) consistente. Los diferentes algoritmos (en el caso de algoritmos supervisados) se entrenan introduciendo una matriz de muestras, X,

con la forma [número de muestras, número de características] y un vector de variable objetivo, *y*, que normalmente contendrá un valor por cada muestra.

Para el caso que nos ocupa, la matriz de muestras consistiría en un número de muestras tomadas a diferentes profundidades con el par de rotación, velocidad de rotación, velocidad de perforación, etc., como características. El vector de variable objetivo contendría los valores asignados del penetrómetro para cada profundidad (Tabla 3.2).

Tabla 3.2: Ejemplo de dos filas de la matriz de muestras con los valores correspondientes de la variable objetivo

Características				**Variable objetivo**
Par (Nm)	**Velocidad de rotación (rpm)**	**Velocidad de penetración (m/hr)**	**Empuje (kN)**	**Golpeos**
18548.6	14.0625	143.5871	151.2445	7.0
20035.6	14.0625	216.7883	149.6975	7.0

Una vez entrenado el modelo se puede introducir en el mismo otra matriz de muestras para obtener una predicción.

Este proceso se muestra de manera simplificada en el siguiente código extraído de la página de scikit-learn.

```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(random_state=0)
X = [[ 1, 2, 3], # 2 muestras, 3 características
   [11, 12, 13]]
y = [0, 1] # clases para cada muestra
clf.fit(X, y) # entrenamiento del modelo

clf.predict(X) # predicción de las clases de los datos de entrenamiento
clf.predict([[4, 5, 6], [14, 15, 16]]) # predicción de las clases de nuevos datos
```

3.3. Análisis inicial de los datos

El sistema de registro EMPAREX de las obras 1 a 4 proporciona los datos en formato XML (eXtensible Markup Language, Lenguaje de Marcado Extensible). Este lenguaje trabaja con datos estructurados, que permiten separar la información fácilmente por partes. Para nuestro caso, es fácil separar la información primero para cada inclusión y luego para cada profundidad. Los datos generales (obra, maquinaria utilizada, diámetro, etc.) se indican al principio del archivo. Estos archivos son a su vez fácilmente convertibles a una hoja de cálculo.

Por otro lado, los datos de la Obra 5 obtenidos mediante Omnibox (una plataforma de toma de datos propia de Menard) pueden extraerse directamente como una hoja de cálculo.

Estas hojas de cálculo pueden leerse como un DataFrame de pandas. Pandas (McKinney, 2010) es una librería de Python optimizada para el análisis de datos tipo tabla. Un DataFrame sería equivalente a una tabla en una pestaña de una hoja de cálculo.

Una vez con los datos en DataFrames podemos analizarlos en busca de errores y prepararlos para el análisis posterior. A continuación, se recogen los ajustes realizados a los datos:

- No se tienen en cuenta aquellas columnas o filas en las que falten datos.
- Se eliminan los datos de inclusiones de prueba o que no aparezcan en los planos.
- Existen inclusiones repetidas. Se desconoce si esto se debe a un error de numeración o a otro tipo de problema, por lo que también se eliminan esas columnas.
- Para la Obra 1 sólo se tienen en cuenta las columnas de 360 mm de diámetro por ser las que se encuentran más cercanas a los penetrómetros.
- Para la Obra 4 sólo se tienen en cuenta las columnas de 300 mm de diámetro por ser las más cercanas a los penetrómetros.

Los datos provenientes de EMPAREX se proporcionan, cada 8 cm de perforación aproximadamente, de una manera muy consistente. Para Omnibox los datos se proporcionan a una frecuencia muy alta, según los leen los sensores, sin embargo, ello provoca que haya un cierto retraso entre unos datos y otros. Esto se ejemplifica en la Figura 3.13 donde las columnas representan los diferentes datos registrados

y las filas los valores en función del tiempo. En amarillo se marcan los valores faltantes.

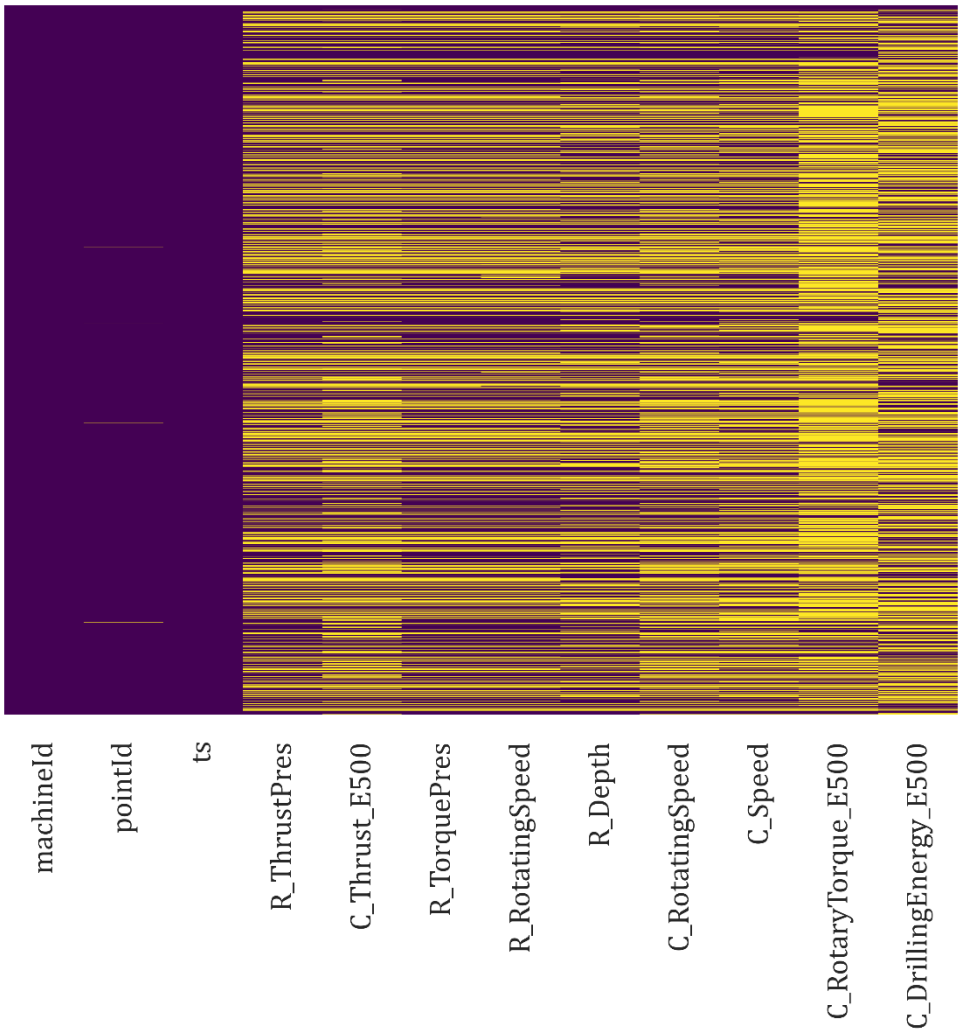


Figura 3.13: Representación de los datos para la Obra 5

Como se observa hay muchos huecos, pero no significa que falten datos, sino que se han tomado en momentos ligeramente distintos. Esto es un problema que se soluciona agrupando las filas en intervalos de 200 ms. De esta manera, la diferencia de profundidad entre muestras dependerá de la velocidad de penetración.

Además, mientras que en las Obras 1 a 4 se utilizó un único equipo de perforación, en la Obra 5 se utilizaron tres. Uno de ellos era un modelo distinto de máquina,

por lo que sus mediciones no se tienen en cuenta. Los datos que se utilizan provienen del mismo modelo de perforadora.

3.4. Formato de archivo

Si bien es posible realizar los análisis que se presentan en esta tesis a partir de hojas de cálculo, esto puede derivar en inconsistencias como distintos nombres para las columnas, o considerar diferentes unidades. Por ello, resulta conveniente convertir los datos disponibles a un formato consistente que permita tratar la información proveniente de diversas fuentes.

Desgraciadamente no existe un formato que se haya impuesto claramente a la hora de transferir datos geotécnicos. De hecho, en la mayoría de las ocasiones, los datos se recogen dentro de informes en PDF.

No obstante, la Asociación de Especialistas Geotécnicos y Geoambientales de Reino Unido (*Association of Geotechnical & Geoenvironmental Specialists*, AGS) mantiene el formato de datos AGS (extensión .ags). El formato de datos AGS (Association of Geotechnical & Geoenvironmental Specialists, 2023a) es un formato de archivo de texto usado para transferir datos de manera fiable dentro de la industria de la investigación geotécnica.

Este formato organiza la información en grupos. Cada uno de estos grupos recoge una parte concreta de los datos geotécnicos, como la localización o los valores de golpeo. Los siguientes cinco grupos son obligatorios ya que contienen información necesaria para identificar los datos:

- Grupo PROJ. Contiene información general sobre el proyecto incluyendo como mínimo un identificador, pero también datos como la localización, el proyectista o el cliente.
- Grupo TRAN. Contiene información respecto al propio archivo como quien lo ha generado, la fecha, el tipo de dato transferido, etc.
- Grupo ABBR. Contiene información sobre las abreviaturas usadas y su descripción.
- Grupo TYPE. Contiene información sobre el tipo de dato utilizado para almacenar la información.

- Grupo UNIT. Contiene información sobre las unidades en las que se proporcionan los datos.

Adicionalmente el grupo LOCA contiene detalles de posición como las coordenadas o la cota de ejecución de una prospección.

Existen otros muchos grupos para todo tipo de información geotécnica. Los relevantes para esta tesis son los siguientes:

- Grupo DOBS. Contiene información sobre parámetros observados durante la perforación. En este grupo se incluirían todos los datos del registro de perforación (par de rotación, velocidad de rotación, velocidad de penetración, etc.).
- Grupo DPRG. Contiene información general de ensayos de penetración dinámica como la normativa aplicada, el peso de la masa, la fecha de ejecución, etc.
- Grupo DPRB. Contiene los datos propios del ensayo de penetración dinámica como el número de golpeos, la profundidad o el incremento para el que se cuentan estos golpeos.
- Grupo SCPT. Contiene todos los datos referentes a los CPT como la posición, la resistencia por fuste o la resistencia por punta, entre otros.

Este formato dispone, además, de una librería (Association of Geotechnical & Geoenvironmental Specialists, 2023b) que permite la conversión desde DataFrame a .ags. También proporciona un programa que permite verificar que el fichero generado cumple con las características del formato.

Una vez con la información en un formato robusto, se puede empezar a analizar esta en profundidad.

3.5. Anomalías en los penetrómetros

Una parte importante a la hora de enfocar un problema de ML es la detección de anomalías en los datos. Estas anomalías pueden deberse a múltiples razones, como errores en los aparatos de medida o fallos de numeración, pero también a cambios bruscos en la naturaleza del terreno o a presencia de bolos. Todas aquellas anomalías que no estén relacionadas con circunstancias reales y que no se reflejen en la resistencia del terreno, deberían eliminarse.

Una posible aproximación al problema podría ser considerar como anomalías aquellas muestras que se separan mucho de un valor medio. Sin embargo, esto no es posible para los penetrómetros. En la Figura 3.14 se muestra el resultado de un penetrómetro de la Obra 1 en el que puede observarse que un valor de golpeo de 15, a unos 4 m de profundidad, es probablemente una anomalía, ya que a esa profundidad el valor esperable es de 5 aproximadamente. En cambio, ese valor de 15 a una profundidad de 10 m es perfectamente normal. En este caso parece, además, que haya otra anomalía a unos 8,5 m de profundidad.

Los penetrómetros son secuencias y los valores están ordenados. Un valor es o no una anomalía en función de su posición en la secuencia. Esto impide utilizar técnicas simples, como definir las anomalías a partir de un cierto número de desviaciones típicas.

Este es un problema propio de series temporales, un tipo de datos que se ordenan en función del tiempo que son comunes en ML, si bien los penetrómetros no tienen el carácter cíclico de las series temporales (i.e. repetirse cada día o cada año). En este caso podríamos utilizar modelos complejos, como aproximar una serie temporal y marcar como anomalías aquellos puntos lejanos a la predicción, sin embargo, se ha optado por una opción sencilla. En vez de utilizar los valores del penetrómetro en bruto, se utiliza una media móvil centrada con una ventana de 5 mediciones, en el caso de los penetrómetros dinámicos, y de 50 en el caso de los penetrómetros estáticos. El resultado es un perfil suavizado que debería recoger de manera más adecuada la resistencia del terreno (Figura 3.15).

Este método tiene la ventaja de conservar todos los valores, adaptándolos, pero podría no funcionar bien con anomalías extremas, por ejemplo, un valor de golpeo de 80 en una zona con golpeos de 5. En ese caso suavizar el perfil no sería posible. Otra opción sería utilizar la mediana en vez de la media, pero ello eliminaría la información que proporciona el valor de golpeo elevado.

En una circunstancia así podría acudirse a una diferenciación, una técnica común en el análisis de series temporales. La diferenciación consiste en restar cada valor por el anterior en la secuencia, si el valor resultado es muy elevado es probable que ese punto sea una anomalía y se actuaría en consecuencia. Sin embargo, en este trabajo no se han detectado esas grandes anomalías y se ha utilizado simplemente la media móvil.

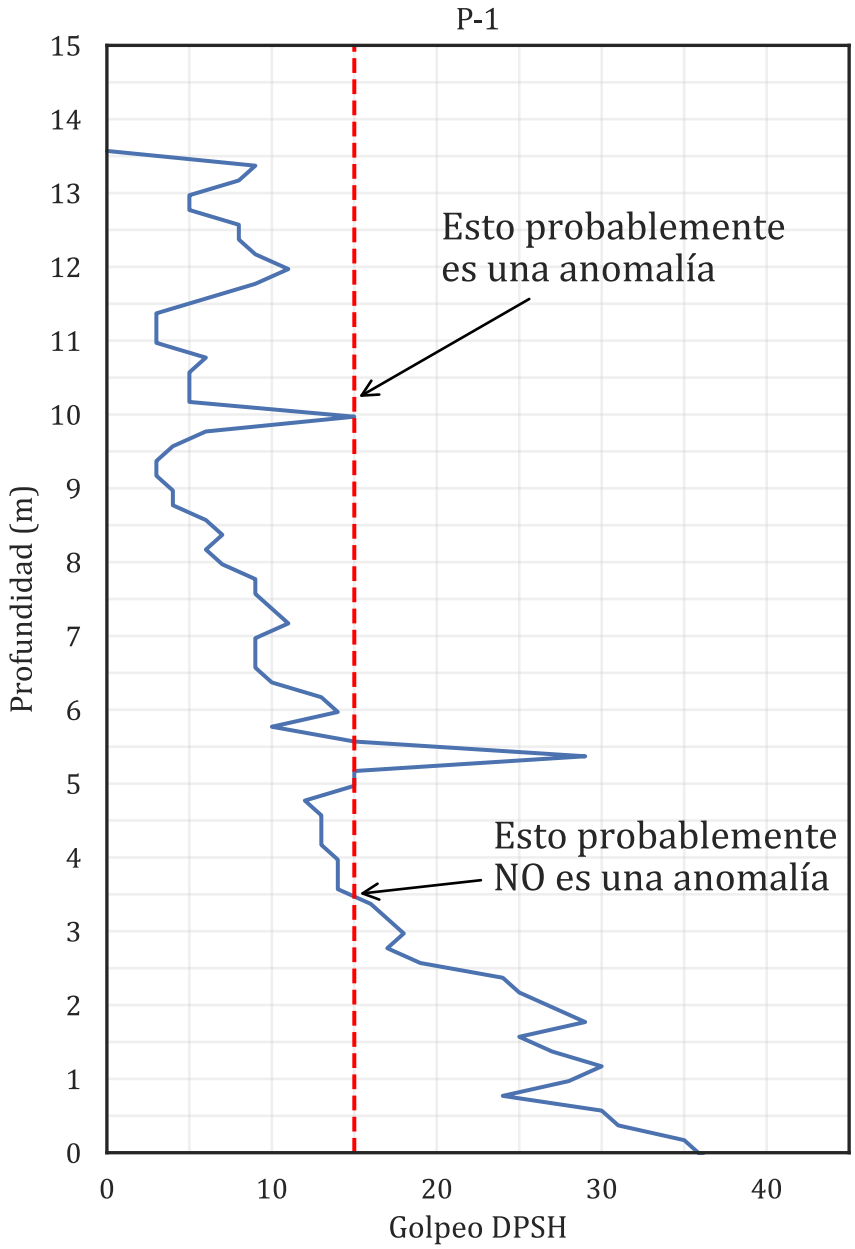


Figura 3.14: Presencia de anomalías en un penetrómetro DPSH perteneciente a la Obra 1

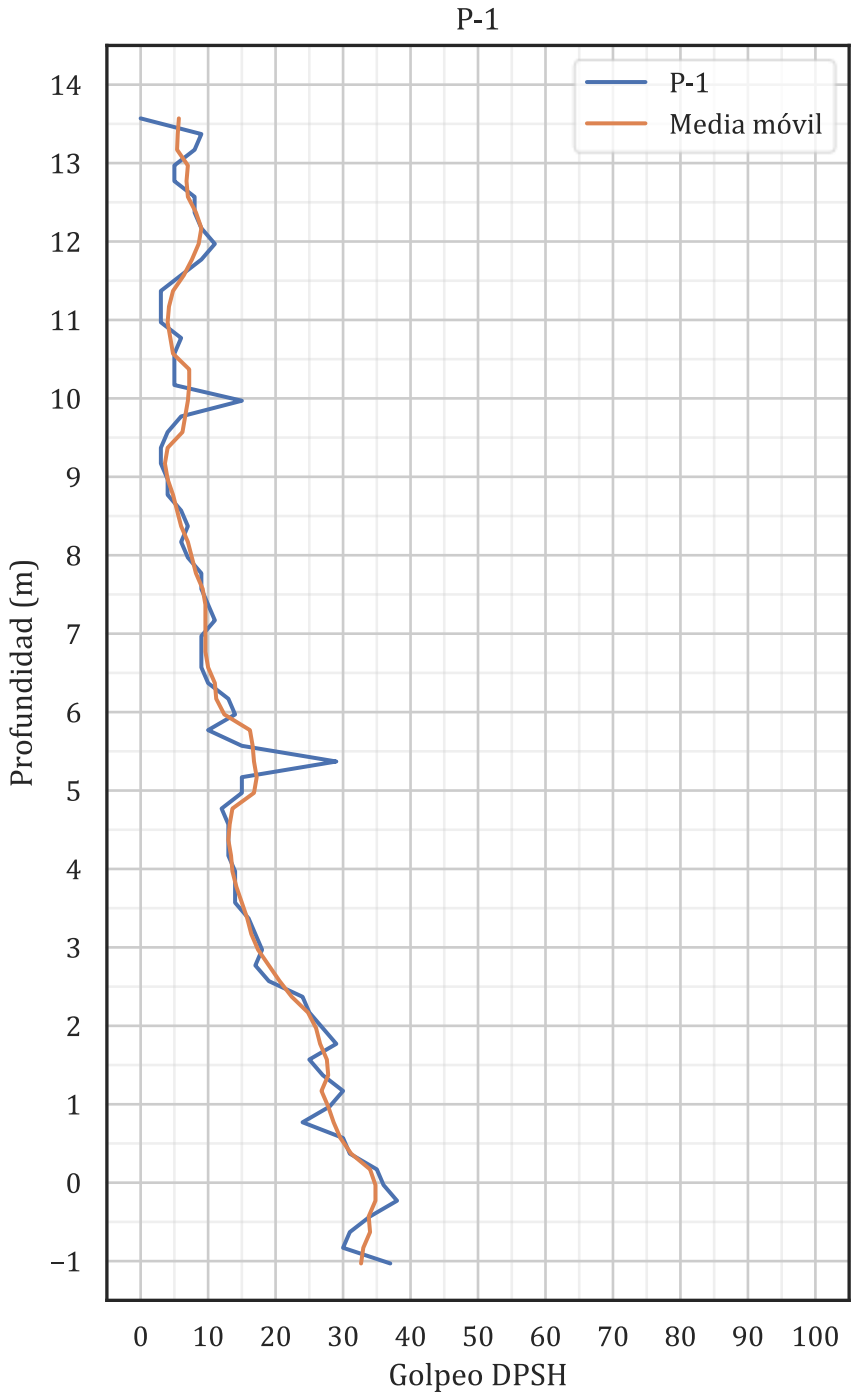


Figura 3.15: Penetrómetro original y perfil suavizado mediante media móvil

3.6. Anomalías en los parámetros de perforación

La detección de posibles anomalías en los datos del registro de parámetros es mucho más compleja que en el caso de los penetrómetros. Existe el mismo problema

con la naturaleza secuencial de los datos: un mismo valor puede ser o no una anomalía en función de su posición en la secuencia. Pero además son datos secuenciales multivariable por lo que un valor puede ser una anomalía en función de su relación con el resto de los datos. Por ejemplo, un valor de par de rotación muy elevado no cuadra con una velocidad de penetración alta.

Para enfrentar este problema utilizaremos el algoritmo *isolation forest* (F. T. Liu et al., 2008). *Isolation forest* es un algoritmo para detectar anomalías, el cual parte de la idea de que una anomalía tiene características diferentes al resto de los datos y que debe ser más fácil de aislar. Este algoritmo crea un árbol, similar a los árboles de decisión, que divide de manera recursiva el conjunto de datos a partir de un valor aleatorio de alguna de las características, hasta que todas las muestras quedan aisladas. Usando varios de estos árboles es posible detectar las anomalías como aquellas muestras que se aíslan más fácilmente.

En la Figura 3.16 las líneas discontinuas representan las divisiones aleatorias realizadas. Como se observa, con tres divisiones ha quedado aislado un punto en la esquina superior derecha del gráfico. Sin embargo, ese mismo punto se podría haber aislado en una única división. Aun así, se ve claramente cómo los puntos con características distintas serán aislados en un número menor de divisiones, de media. En cambio, los puntos acumulados en el centro, que presentan valores similares entre sí, serán difíciles de aislar y requerirán, de media, un mayor número de pasos.

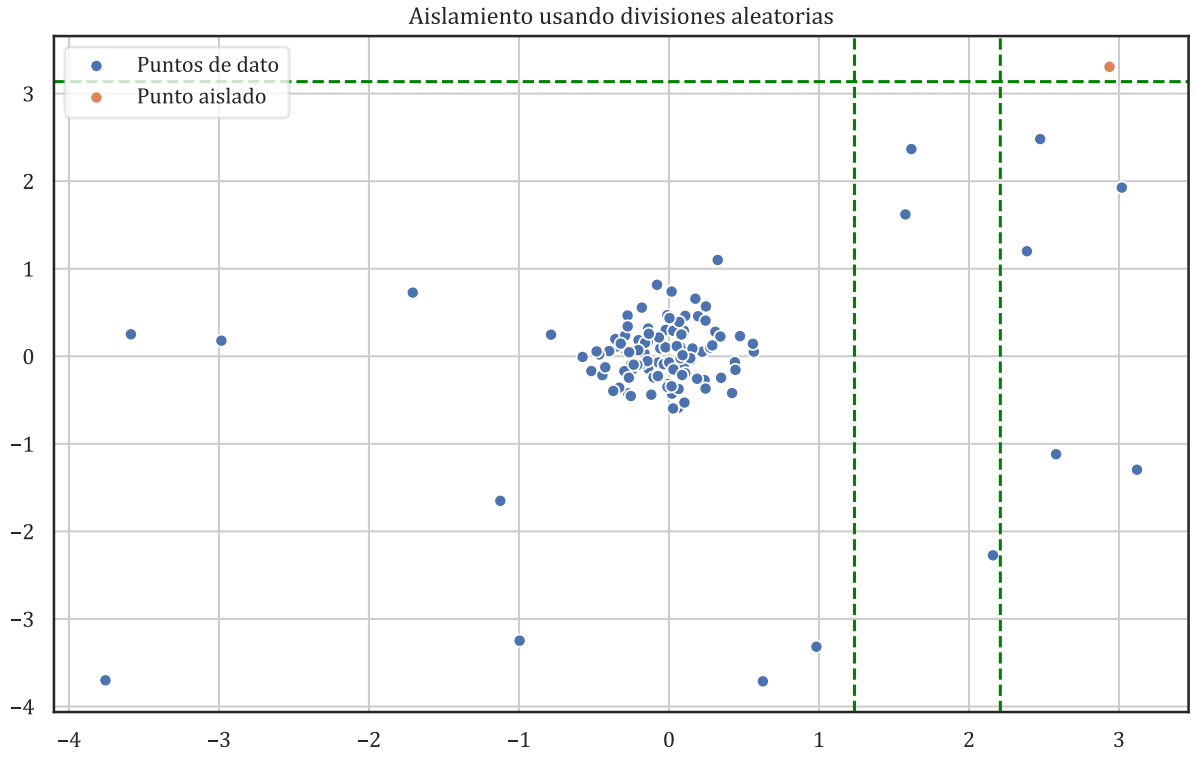


Figura 3.16: Aislamiento mediante divisiones aleatorias

En la Figura 3.17 se muestra un gráfico por pares que muestra la relación entre diferentes características de la Obra 1 (profundidad, par, velocidad de rotación, velocidad de penetración, empuje y energía específica) así como la distribución de cada una en la diagonal. En este gráfico se pueden observar claramente algunos puntos aislados que, con toda probabilidad, serán anomalías. Sin embargo, hay otras muestras que seguramente también sean anomalías pero que no resultan tan evidentes de detectar. Por ejemplo, no debería haber valores de par superiores a 180000 Nm, por ser el par máximo del equipo.

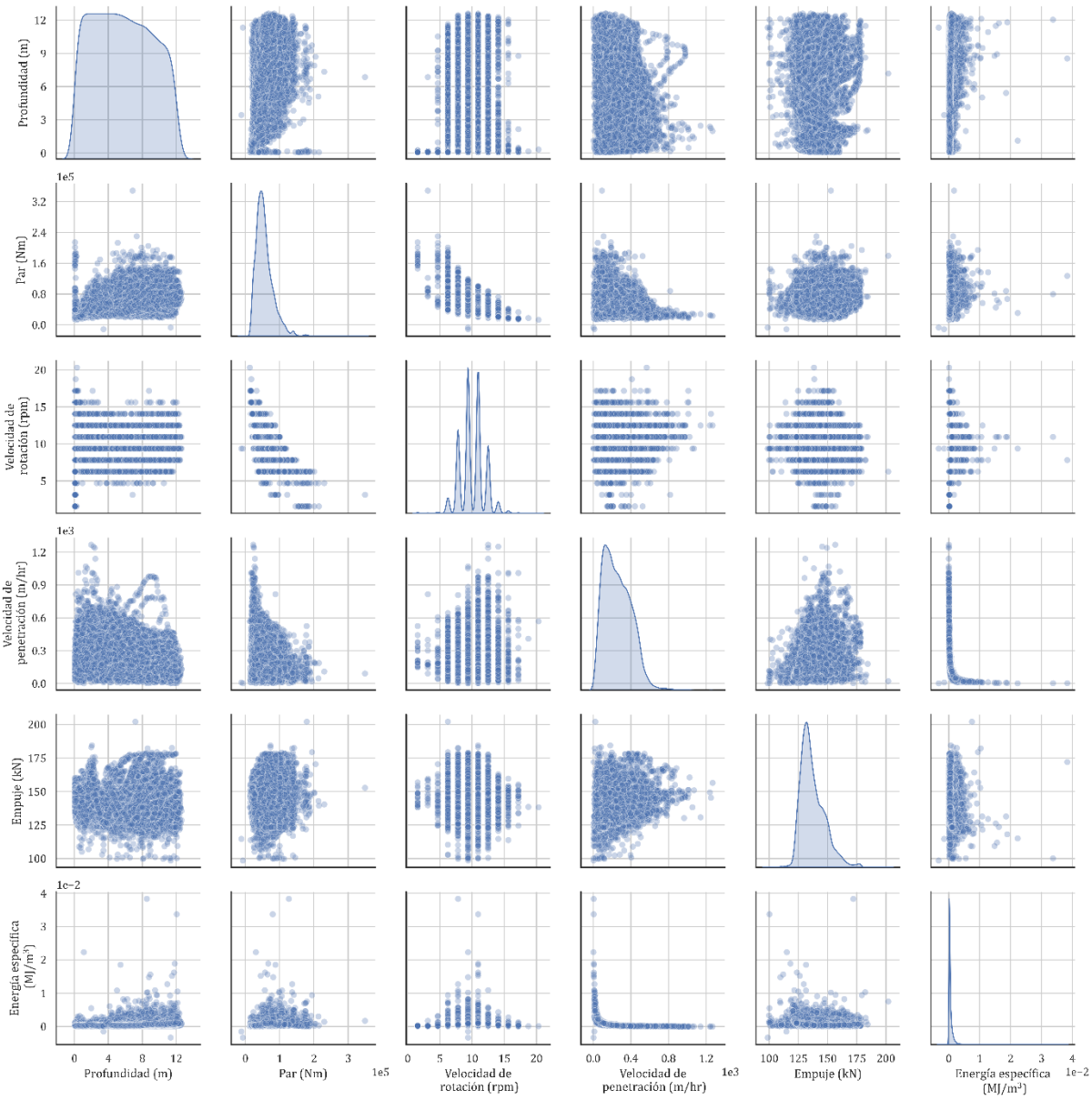


Figura 3.17: Gráfico por pares con características de la Obra 1

Por otro lado, en la Figura 3.18 se muestran los valores considerados normales y las anomalías tras aplicar *Isolation Forest*, con un porcentaje de anomalías del 5%. El algoritmo es capaz de detectar como anomalías valores de par o de velocidad de perforación muy elevados. Además, también detecta valores anómalos de par que, si bien no son tan elevados, no cuadra que se produzcan al inicio de la perforación. Esto puede deberse a picos de presión en el sistema hidráulico asociados al arranque de la perforación, y son anomalías que serían difíciles de detectar de manera automática con métodos más sencillos.

Figura 3.18: Detección de anomalías en la Obra 1

3.7. Etiquetado de las muestras

El etiquetado de las muestras se refiere, en este caso, a asignar a cada muestra (a cada fila de datos de perforación) un valor apropiado del ensayo penetrométrico. En principio es una tarea sencilla, tan sólo tenemos que escoger el penetrómetro más cercano y asignar a la fila el valor de su cota. Las cotas no tienen por qué coincidir, por lo que se interpolarían linealmente los datos para obtener un valor al nivel correspondiente. En la Figura 3.19 se muestra una interpolación lineal sencilla para obtener valores a cotas diferentes a las originales.

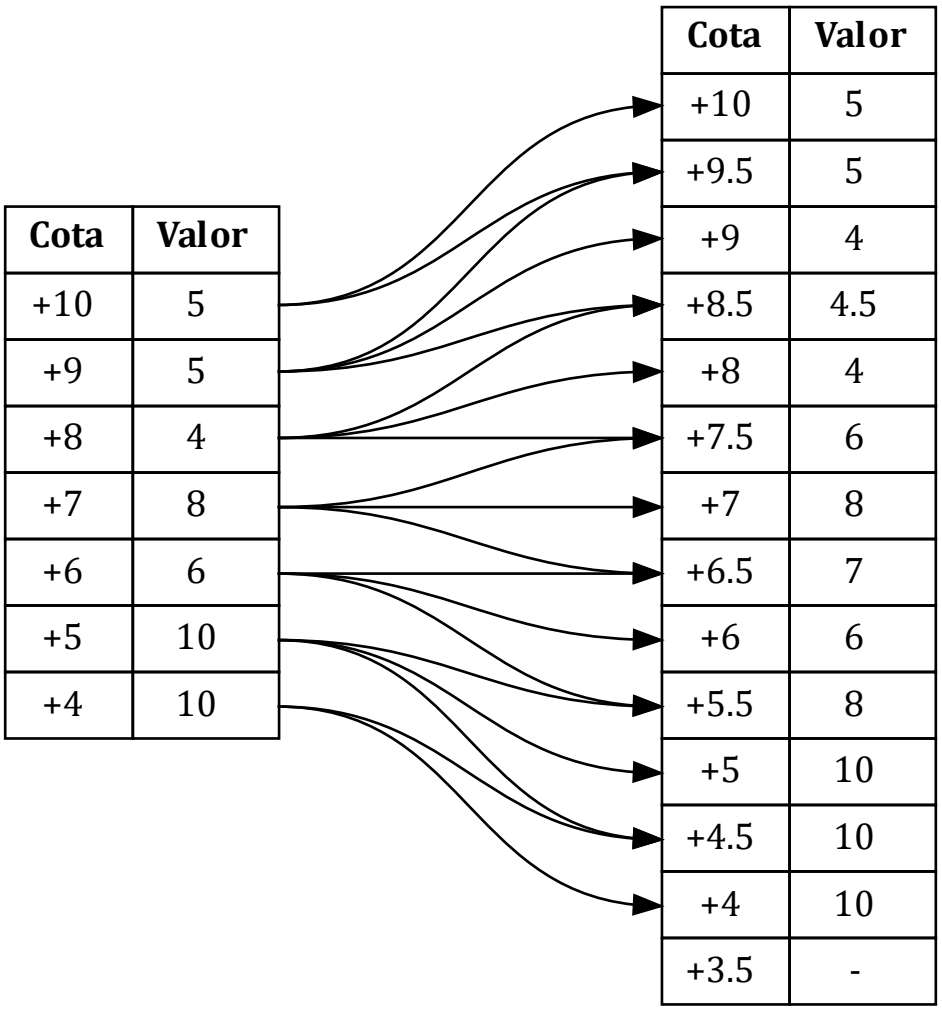


Figura 3.19: Interpolación lineal para obtener valores con cotas diferentes a las originales

Cabe señalar que, en el ejemplo de la Figura 3.19, no es posible obtener un valor para la cota +3.5 ya que no se dispone de información a esa profundidad. Cuando no hay datos a una determinada cota, no tenemos en cuenta esa fila en los cálculos.

3.8. Constitución del conjunto de datos

Una vez detectadas las posibles anomalías y con las muestras etiquetadas podríamos aplicar directamente los algoritmos. Sin embargo, hemos de tener en cuenta una fuente de incertidumbre adicional: la distancia entre los penetrómetros y las inclusiones realizadas. Como se observa en la Figura 3.20, el terreno puede cambiar sus características según nos alejamos de los penetrómetros. Si

consideramos inclusiones muy alejadas a su penetrómetro más cercano, estaremos introduciendo errores debido a los cambios en el suelo.

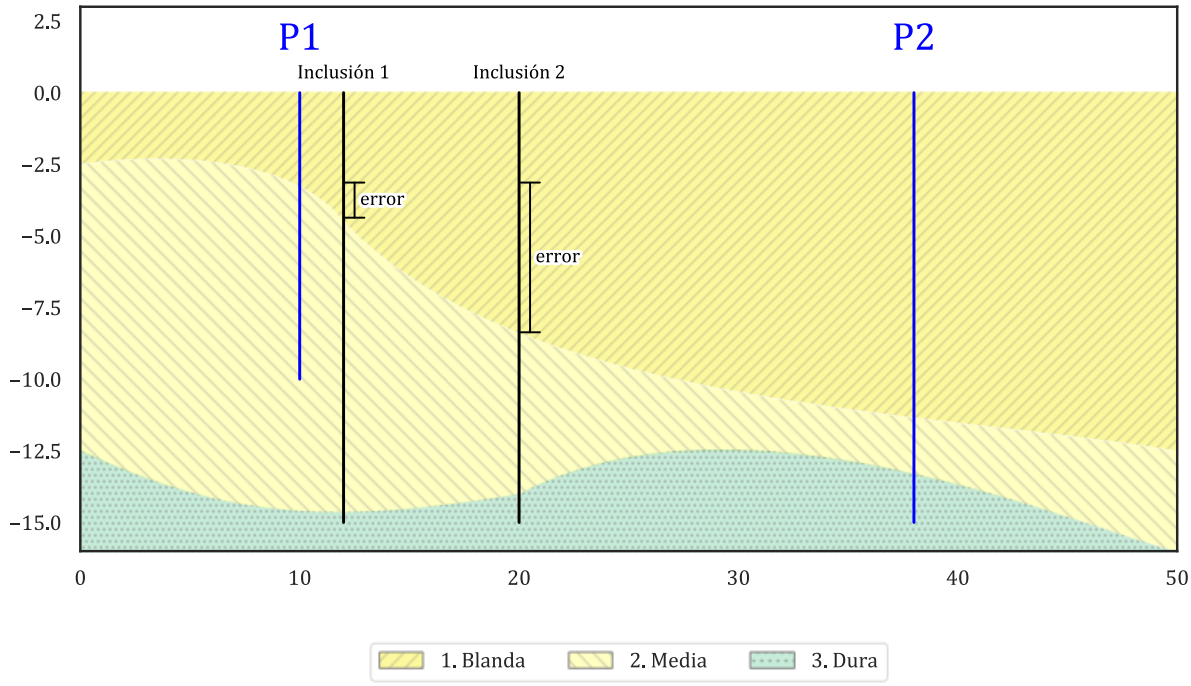


Figura 3.20: Perfil geotécnico. Se observa cómo cambia el terreno según aumenta la distancia a los penetrómetros

Una posible opción ante este problema sería considerar únicamente aquellas inclusiones realizadas a muy poca distancia de los penetrómetros. Sin embargo, esto puede reducir mucho el número de datos disponible. Esa situación no es deseable ya que un conjunto de datos amplio tenderá a reflejar mejor la realidad (Ley de los Grandes Números), y da la capacidad al modelo de aprender patrones más robustos y generales. A mayor número de muestras es más probable tener ejemplos suficientes de las posibles situaciones. Además, cuantas más características se tengan en cuenta, más separados quedan los datos, siendo más difícil generalizar (maldición de la dimensionalidad, (Bellman, 1957)). Por estas razones, suponiendo que se mantiene la calidad de los datos y que tenemos una capacidad de cálculo suficiente, es preferible disponer de un conjunto de datos lo más amplio posible.

Por tanto, debemos encontrar la manera de considerar el mayor número de muestras posibles controlando el error introducido. Para ello, Bunieski (Bunieski, 2022), propone dos métodos, el método de variabilidad y el método continuo, como

alternativa a simplemente contar con el penetrómetro más cercano, independientemente de su distancia a la inclusión.

3.8.1. Método de variabilidad

El método de variabilidad se basa en definir una distancia máxima entre el penetrómetro y la inclusión en la que, con un nivel alto de seguridad, el terreno no varíe significativamente. De esta manera se define la variabilidad V_p de un parámetro p, como el percentil 90 de la diferencia absoluta del valor del parámetro para una determinada cota entre cada pareja de prospecciones, dividida entre la distancia entre prospecciones (Ecuación 3.1).

$$V_p = P90\left(\frac{|p_i - p_j|_k}{d_{ij}}\right)_{i,j,k,i\neq j}$$

Ecuación 3.1: Variabilidad

Dónde k es una profundidad determinada; i, j son dos prospecciones distintas; p_i es el valor del parámetro p en la prospección i para la cota k; $d_{i,j}$ es la distancia entre las prospecciones i y j; y P90 es el percentil 90. Esta formulación puede interpretarse como el percentil 90 de las pendientes del parámetro p entre las diferentes prospecciones. A partir del valor de V_p podemos calcular una distancia máxima para una determinada variación de p según la Ecuación 3.2:

$$D_{max}(m) = \frac{Variación}{V_p}$$

Ecuación 3.2: Distancia máxima a partir de V_p

En la Figura 3.21 se muestra la distancia entre las inclusiones y su prospección más cercana para la Obra 1.

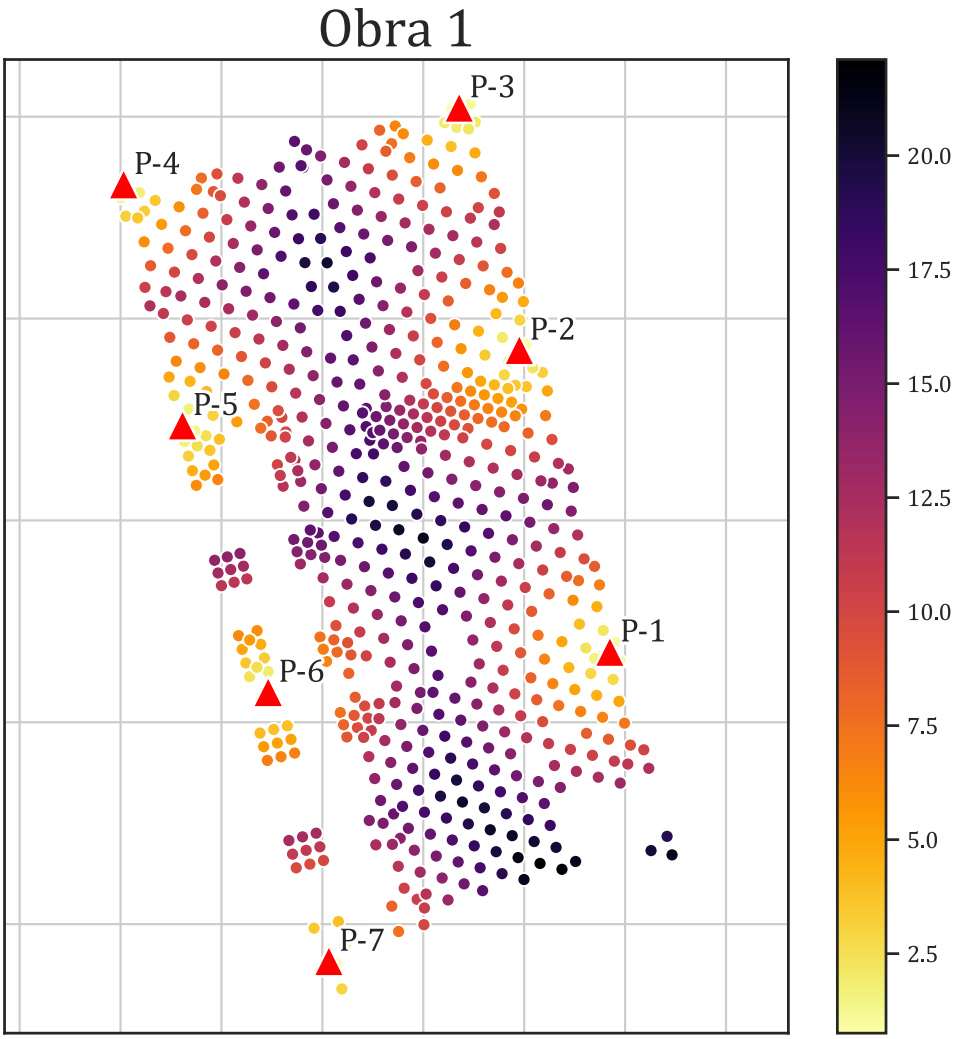


Figura 3.21: Distancia entre las inclusiones y su prospección más cercana para la Obra 1

Si calculamos V_p para el golpeo del penetrómetro dinámico en la Obra 1 obtenemos un valor de $V_p = 0.2746$. Suponiendo una variación de 2 en el valor de golpeo obtendríamos una distancia $D_{max} = 7.284$. Podemos ver cómo afecta esta distancia en la Figura 3.22.

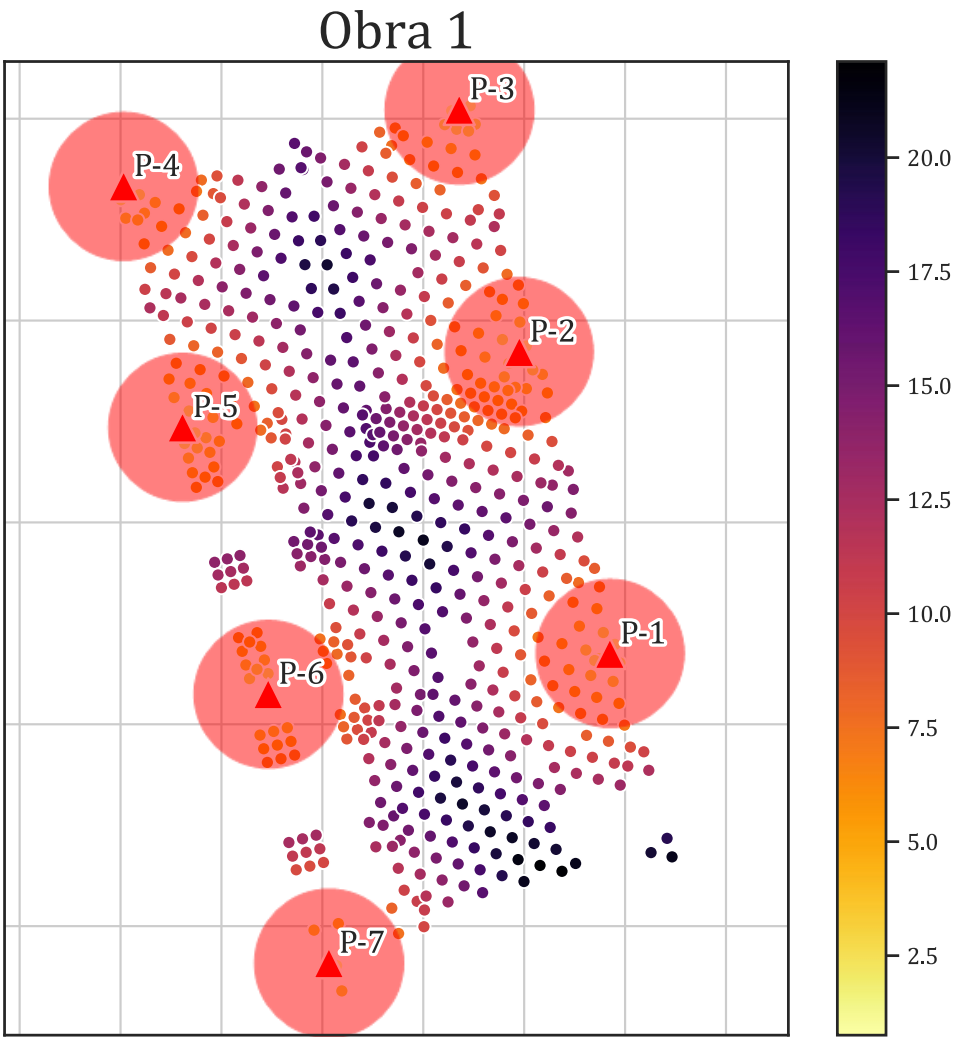


Figura 3.22: Representación de la distancia máxima según el método de variabilidad para la Obra 1

3.8.2. Método continuo

El método de variabilidad, aunque intuitivo, presenta algunos inconvenientes:

- Puede implicar una reducción significativa en la cantidad de datos disponibles.
- Obtiene una única distancia máxima independiente de la dirección. Puede darse el caso que el terreno presente mayores cambios en una dirección que en otras (por ejemplo, en la dirección perpendicular a una vaguada, frente a la dirección paralela a esta).

- Las prospecciones se consideran de manera independiente, no hay continuidad.

Frente a estos problemas Bunieski propone el método continuo. Este método interpola los valores para todos los puntos, pero propone una función de información L. Esta función L (Ecuación 3.4) define la cantidad de "información" disponible en una localización determinada como la suma de las informaciones individuales l_i (Ecuación 3.3) proporcionadas por cada prospección i donde d_i es la distancia entre la localización objetivo y dicha prospección. α es un parámetro que controla el área de impacto de la función.

$$l_i = 100\left(\frac{1}{\cosh(d_i)}\right)^{\alpha}$$

Ecuación 3.3: Información individual

$$L = \sum_i l_i$$

Ecuación 3.4: Información

La información individual se multiplica por 100 para obtener un valor tipo porcentual. Justo en la prospección tendríamos un valor de 100 que iría decreciendo según nos alejamos. La velocidad de dicho decrecimiento depende de α. A mayor α, más rápido decrece la información (Figura 3.23).

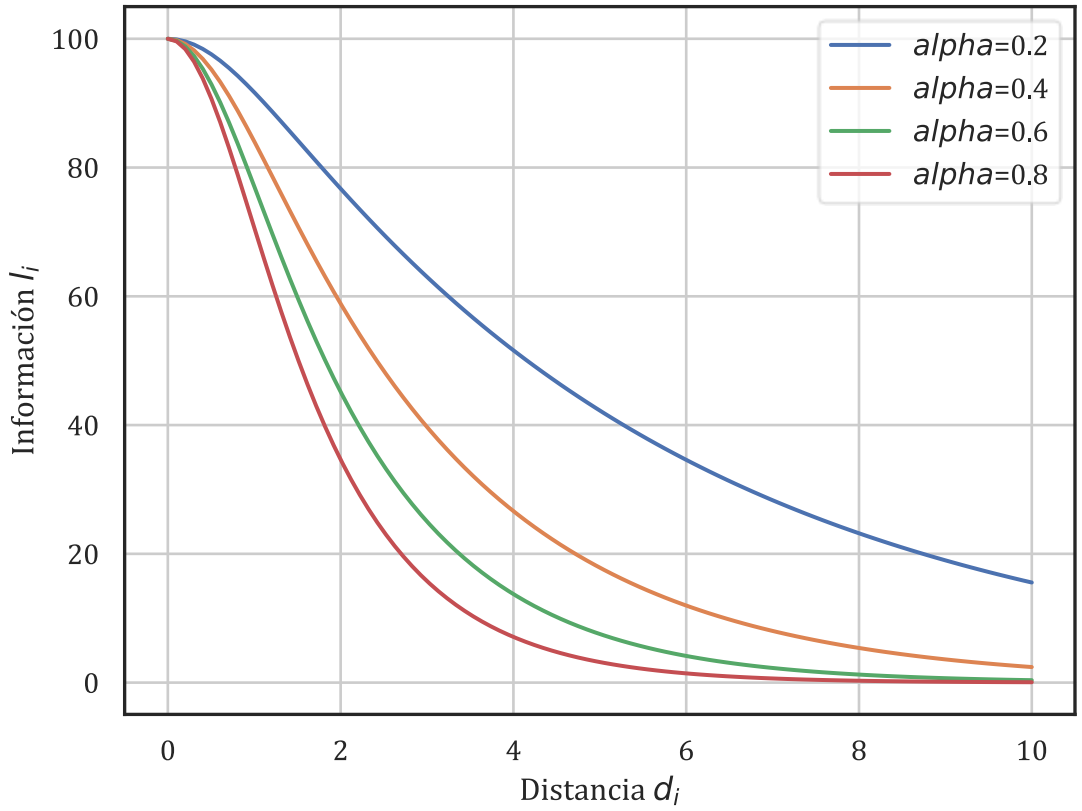


Figura 3.23: Decrecimiento de la información individual l_i en función de la distancia d_i a la prospección y de distintos valores de α

Con este método podríamos tener en cuenta todas las inclusiones de la obra, pero podemos definir un valor mínimo de L para mantener contenido el error. En la Figura 3.24 se muestra el resultado del cálculo de L a lo largo de la Obra 1 para $\alpha = 0.2$ y $L_{min} = 10$.

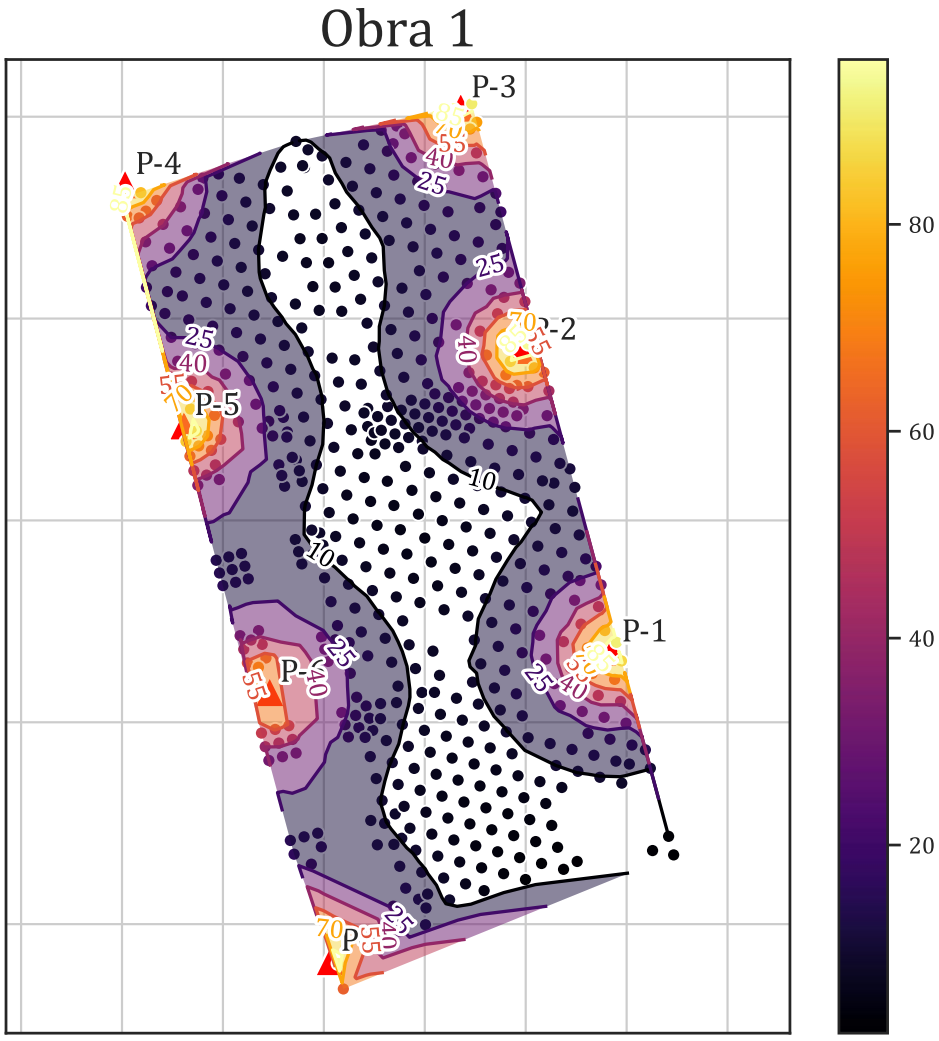


Figura 3.24: Información L a lo largo de la Obra 1 para $\alpha = 0.2$ y $L_{min} = 10$

Cabe señalar que la Figura 3.24 representa el valor L a cota superior de las inclusiones. Puede darse el caso de que las prospecciones no tengan la misma longitud y que, a partir de cierta profundidad, la información disminuya al no alcanzar los penetrómetros dicha cota.

En esta tesis se estudian ambos métodos, variabilidad y continuo, además de estudiarse el efecto de la distancia entre la inclusión y su prospección asociada.

3.9. Ingeniería de características

La elección de las características a utilizar es un parte muy relevante en el proceso de aplicación de ML. En este caso, simplemente observando los parámetros

compuestos existentes en la literatura, ya destacan cuatro parámetros: la velocidad de rotación, la velocidad de penetración, el empuje aplicado y el par. Estas características medidas durante la perforación parecen claramente relacionadas con las propiedades del terreno atravesado.

Otra característica que podríamos tener en cuenta es la profundidad. La resistencia del suelo suele estar relacionada con la profundidad y es un parámetro que podría ayudar a los algoritmos. Sin embargo, si no se dispone de suficiente variación en los datos de los penetrómetros, estos estarán muy correlacionados con la cota. Si esto ocurre el algoritmo podría dar demasiada importancia a la profundidad, perdiendo la capacidad de detectar diferencias en la naturaleza del terreno en función del registro de perforación.

Por otro lado, debido al método de ejecución mediante rotación por desplazamiento, el par aplicado se distribuye a lo largo del fuste de la herramienta de perforación. Esta herramienta, por su forma, desplaza el terreno hacia arriba hasta aproximadamente 1.5 m por encima de la punta, donde el terreno pasa a ser desplazado hacia los laterales. Para tener en cuenta este hecho podrían agregarse todos los valores del registro de parámetros a lo largo de esos 1.5 m, sin embargo, esto añadiría una gran cantidad de características pudiendo caer en la «maldición de la dimensionalidad».

Para añadir esta información se utiliza la media de los valores de las cuatro características mencionadas (velocidad de rotación, velocidad de perforación, empuje aplicado y par) en dicha longitud.

Sin embargo, la media podría no reflejar adecuadamente la naturaleza del terreno, por lo que se utiliza también un vector de correlación. El vector de correlación fue propuesto por Sapronova y Marcher (Sapronova & Marcher, 2023) como una manera de extraer patrones de datos agregados. Este vector consistiría en los valores de coeficiente de correlación *vis a vis* de las diferentes características. En la Tabla 3.3 se muestra cual sería el vector de correlación resultante de la matriz de correlación de la Figura 3.25.

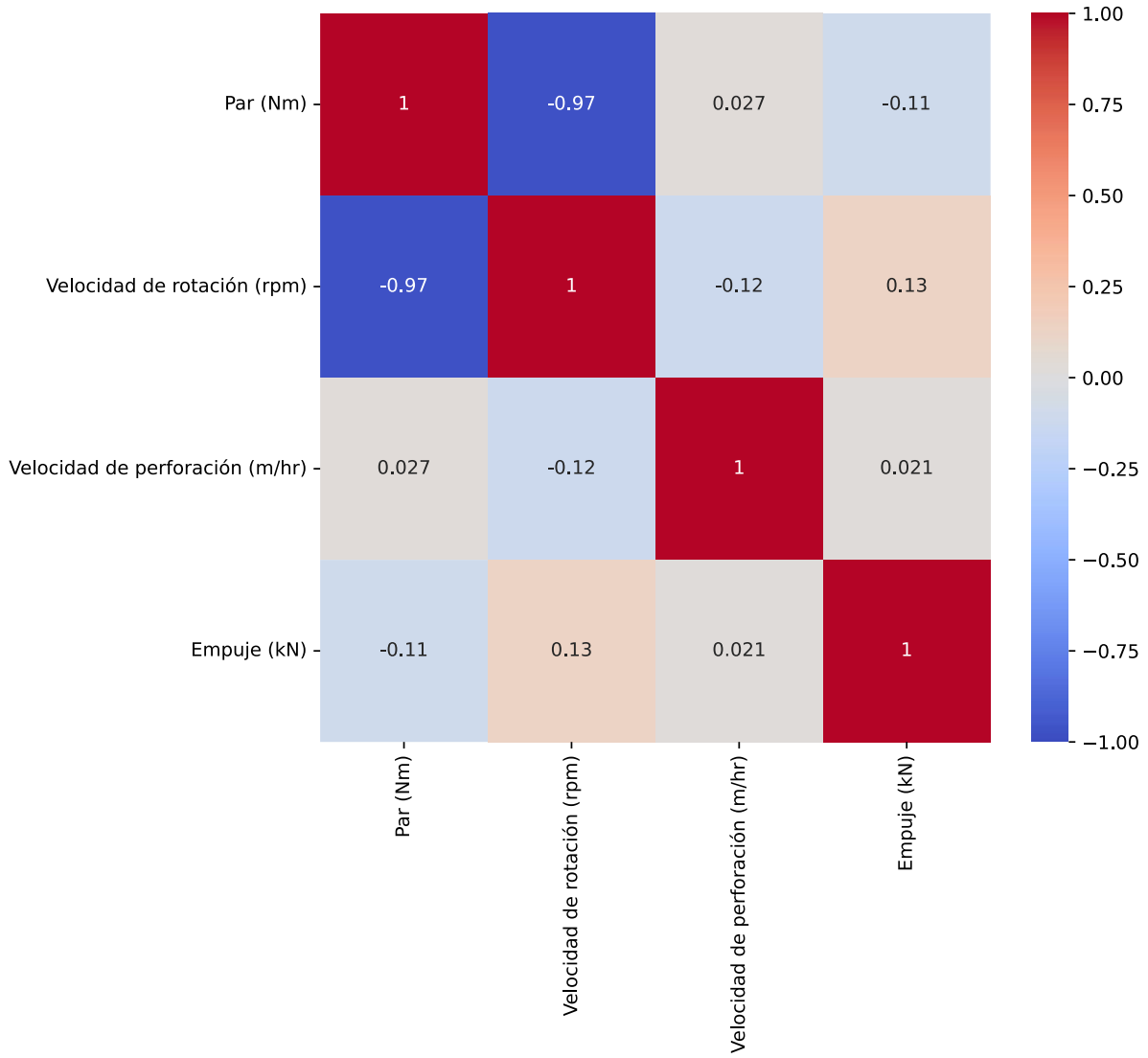


Figura 3.25: Ejemplo de matriz de correlación

Tabla 3.3: Vector de correlación resultante.

Par/ velocidad de rotación	**Par/ Velocidad de perforación**	**Par/ Empuje**	**Velocidad de rotación/ Velocidad de perforación**	**Velocidad de rotación/ Empuje**	**Velocidad de perforación/ Empuje**
-0.97	0.027	-0.11	-0.12	-0.13	0.021

En este trabajo se tienen en cuenta diferentes conjuntos de características según lo descrito en el presente apartado, como se refleja en la Tabla 3.4.

Tabla 3.4: Diferentes conjuntos de características utilizados.

Par	Velocidad de rotación	Velocidad de perforación	Empuje	Profundidad	Media 1.5 m	Vector de correlación
X	X	X	X			
X	X	X	X	X		
X	X	X	X		X	
X	X	X	X			X
X	X	X	X	X	X	
X	X	X	X	X		X
X	X	X	X		X	X
X	X	X	X	X	X	X

3.10. Elección de *pipelines*

Tal y como se ha expuesto previamente, existen multitud de algoritmos de ML entre los que elegir (ANN, SVM, KNN, RF, etc.), cada uno de ellos con su propio conjunto de hiperparámetros para afinar. Además, nos encontramos con muchos preprocesamientos que pueden ser de aplicación. Existen por tanto una infinidad de posibles opciones.

Para acercarnos a los mejores *pipelines* posibles, utilizamos la librería de AutoML TPOT, ya presentada en el apartado correspondiente. Concretamente se ha utilizado la versión 0.12.1 de TPOT, la cual tiene una serie de configuraciones de las que se han utilizado la opción por defecto al enfocar el problema como regresión y la opción «TPOT NN» con el enfoque de clasificación. En las tablas 3. a 3. se recogen los distintos elementos que pueden formar parte de un *pipeline* de TPOT según las configuraciones utilizadas.

De cara al cálculo, se mantiene el resto de la configuración de TPOT por defecto, por lo que el proceso comenzará con una población de 100 *pipelines* que hibridará, mutará y evaluará a lo largo de 100 generaciones. Este proceso lleva tiempo y es normal que tome horas e incluso días en finalizar. Por otro lado, también se mantienen las métricas por defecto para la evaluación de los modelos: la precisión para los algoritmos de clasificación y el negativo del error cuadrático medio para regresión.

Otro parámetro de configuración importante de TPOT que mantenemos por defecto es el tiempo de evaluación de un único *pipeline*. Este valor es de 5 min lo que impide

que se evalúen *pipelines* demasiado complejos, pero también limita los resultados que pueden obtenerse.

Tabla 3.5: Preprocesos considerados con TPOT

Preproceso	Módulo	Hiperparámetros
Binarizador	sklearn.preprocessing.Binarizer	'threshold': np.arange(0.0, 1.01, 0.05)
ICA	sklearn.decomposition.FastICA	'tol': np.arange(0.0, 1.01, 0.05)
Aglomeración de características	sklearn.cluster.FeatureAgglomeration	'linkage': ['ward', 'complete', 'average'], 'affinity': ['euclidean', 'l1', 'l2', 'manhattan', 'cosine']
Escalador MaxAbs	sklearn.preprocessing.MaxAbsScaler	
Escalador MinMax	sklearn.preprocessing.MinMaxScaler	
Normalizador	sklearn.preprocessing.Normalizer	'norm': ['l1', 'l2', 'max']
Nystroem	sklearn.kernel_approximation.Nystroem	'kernel': ['rbf', 'cosine', 'chi2', 'laplacian', 'polynomial', 'poly', 'linear', 'additive_chi2', 'sigmoid'], 'gamma': np.arange(0.0, 1.01, 0.05), 'n_components': range(1, 11)
PCA	sklearn.decomposition.PCA	'svd_solver': ['randomized'], 'iterated_power': range(1, 11)
Características polinómicas	sklearn.preprocessing.PolynomialFeatures	'degree': [2], 'include_bias': [False], 'interaction_only': [False]
Muestreo RBF	sklearn.kernel_approximation.RBFSampler	'gamma': np.arange(0.0, 1.01, 0.05)
Escalador robusto	sklearn.preprocessing.RobustScaler	
Escalador estandar	sklearn.preprocessing.StandardScaler	
Cuenta de ceros	tpot.builtins.ZeroCount	
Codificador OneHot	tpot.builtins.OneHotEncoder	'minimum_fraction': [0.05, 0.1, 0.15, 0.2, 0.25], 'sparse': [False], 'threshold': [10]

Tabla 3.6: Algoritmos de regresión considerados en TPOT

Algoritmo	Módulo scikit-learn	Hiperparámetros
Elastic-Net	sklearn.linear_model.ElasticNetCV	'l1_ratio': np.arange(0.0, 1.01, 0.05), 'tol': [1e-5, 1e-4, 1e-3, 1e-2, 1e-1]
ET	sklearn.ensemble.ExtraTreesRegressor	'n_estimators': [100], 'max_features': np.arange(0.05, 1.01, 0.05), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21), 'bootstrap': [True, False]

GB	sklearn.ensemble.GradientBoostingRegressor	'n_estimators': [100], 'loss': ["ls", "lad", "huber", "quantile"], 'learning_rate': [1e-3, 1e-2, 1e-1, 0.5, 1.], 'max_depth': range(1, 11), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21), 'subsample': np.arange(0.05, 1.01, 0.05), 'max_features': np.arange(0.05, 1.01, 0.05), 'alpha': [0.75, 0.8, 0.85, 0.9, 0.95, 0.99]
Adaboost	sklearn.ensemble.AdaBoostRegressor	'n_estimators': [100], 'learning_rate': [1e-3, 1e-2, 1e-1, 0.5, 1.], 'loss': ["linear", "square", "exponential"]
DT	sklearn.tree.DecisionTreeRegressor	'max_depth': range(1, 11), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21)
KNN	sklearn.neighbors.KNeighborsRegressor	'n_neighbors': range(1, 101), 'weights': ["uniform", "distance"], 'p': [1, 2]
Lasso	sklearn.linear_model.LassoLarsCV	'normalize': [True, False]
SVM	sklearn.svm.LinearSVR	'loss': ["epsilon_insensitive", "squared_epsilon_insensitive"], 'dual': [True, False], 'tol': [1e-5, 1e-4, 1e-3, 1e-2, 1e-1], 'C': [1e-4, 1e-3, 1e-2, 1e-1, 0.5, 1., 5., 10., 15., 20., 25.], 'epsilon': [1e-4, 1e-3, 1e-2, 1e-1, 1.]
RF	sklearn.ensemble.RandomForestRegressor	'n_estimators': [100], 'max_features': np.arange(0.05, 1.01, 0.05), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21), 'bootstrap': [True, False]
Ridge	sklearn.linear_model.RidgeCV	
XGB	xgboost.XGBRegressor	'n_estimators': [100], 'max_depth': range(1, 11), 'learning_rate': [1e-3, 1e-2, 1e-1, 0.5, 1.], 'subsample': np.arange(0.05, 1.01, 0.05), 'min_child_weight': range(1, 21), 'n_jobs': [1], 'verbosity': [0], 'objective': ['reg:squarederror']
SGD	sklearn.linear_model.SGDRegressor	'loss': ['squared_loss', 'huber', 'epsilon_insensitive'], 'penalty': ['elasticnet'], 'alpha': [0.0, 0.01, 0.001] , 'learning_rate': ['invscaling', 'constant'] , 'fit_intercept': [True, False], 'l1_ratio': [0.25, 0.0, 1.0, 0.75, 0.5], 'eta0': [0.1, 1.0, 0.01], 'power_t': [0.5, 0.0, 1.0, 0.1, 100.0, 10.0, 50.0]

Tabla 3.7: Selectores considerados con TPOT para regresión

Selector	Módulo	Hiperparámetros
FWE	sklearn.feature_selection.SelectFwe	'alpha': np.arange(0, 0.05, 0.001), 'score_func': {'sklearn.feature_selection.f_regression': None}

Percentil	sklearn.feature_selection.SelectPercentile	'percentile': range(1, 100), 'score_func': {'sklearn.feature_selection.f_regression': None}
Límite de varianza	sklearn.feature_selection.VarianceThreshold	'threshold': [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2]
Selección desde modelo	sklearn.feature_selection.SelectFromModel	'threshold': np.arange(0, 1.01, 0.05), 'estimator': {'sklearn.ensemble.ExtraTreesRegressor': { 'n_estimators': [100], 'max_features': np.arange(0.05, 1.01, 0.05)}

Tabla 3.8: Algoritmos de clasificación considerados en TPOT

Algoritmo	**Módulo**	**Hiperparámetros**
LR	tpot.builtins.PytorchLRClassifier	'learning_rate': [1e-3, 1e-2, 1e-1, 0.5, 1.], 'batch_size': [4, 8, 16, 32], 'num_epochs': [5, 10, 15], 'weight_decay': [0, 1e-4, 1e-3, 1e-2]
MLP	tpot.builtins.PytorchMLPClassifier	'learning_rate': [1e-3, 1e-2, 1e-1, 0.5, 1.], 'batch_size': [4, 8, 16, 32], 'num_epochs': [5, 10, 15], 'weight_decay': [0, 1e-4, 1e-3, 1e-2]
NB Gaussiana	sklearn.naive_bayes.GaussianNB	
NB Bernoulli	sklearn.naive_bayes.BernoulliNB	'alpha': [1e-3, 1e-2, 1e-1, 1., 10., 100.], 'fit_prior': [True, False]
NB multinomial	sklearn.naive_bayes.MultinomialNB	'alpha': [1e-3, 1e-2, 1e-1, 1., 10., 100.], 'fit_prior': [True, False]
DT	sklearn.tree.DecisionTreeClassifier	'criterion': ["gini", "entropy"], 'max_depth': range(1, 11), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21)
ET	sklearn.ensemble.ExtraTreesClassifier	'n_estimators': [100], 'criterion': ["gini", "entropy"], 'max_features': np.arange(0.05, 1.01, 0.05), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21), 'bootstrap': [True, False]
RF	sklearn.ensemble.RandomForestClassifier	'n_estimators': [100], 'criterion': ["gini", "entropy"], 'max_features': np.arange(0.05, 1.01, 0.05), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21), 'bootstrap': [True, False]
GB	GradientBoostingClassifier	'n_estimators': [100], 'learning_rate': [1e-3, 1e-2, 1e-1, 0.5, 1.], 'max_depth': range(1, 11), 'min_samples_split': range(2, 21), 'min_samples_leaf': range(1, 21), 'subsample': np.arange(0.05, 1.01, 0.05), 'max_features': np.arange(0.05, 1.01, 0.05)
KNN	sklearn.neighbors.KNeighborsClassifier	'n_neighbors': range(1, 101), 'weights': ["uniform", "distance"], 'p': [1, 2]
SVM	sklearn.svm.LinearSVC	'penalty': ["l1", "l2"],

		'loss': ["hinge", "squared_hinge"], 'dual': [True, False], 'tol': [1e-5, 1e-4, 1e-3, 1e-2, 1e-1], 'C': [1e-4, 1e-3, 1e-2, 1e-1, 0.5, 1., 5., 10., 15., 20., 25.]
LR	sklearn.linear_model.LogisticRegression	'penalty': ["l1", "l2"], 'C': [1e-4, 1e-3, 1e-2, 1e-1, 0.5, 1., 5., 10., 15., 20., 25.], 'dual': [True, False]
XGB	xgboost.XGBClassifier	'n_estimators': [100], 'max_depth': range(1, 11), 'learning_rate': [1e-3, 1e-2, 1e-1, 0.5, 1.], 'subsample': np.arange(0.05, 1.01, 0.05), 'min_child_weight': range(1, 21), 'n_jobs': [1], 'verbosity': [0]
Linear SGD	sklearn.linear_model.SGDClassifier	'loss': ['log', 'hinge', 'modified_huber', 'squared_hinge', 'perceptron'], 'penalty': ['elasticnet'], 'alpha': [0.0, 0.01, 0.001], 'learning_rate': ['invscaling', 'constant'], 'fit_intercept': [True, False], 'l1_ratio': [0.25, 0.0, 1.0, 0.75, 0.5], 'eta0': [0.1, 1.0, 0.01], 'power_t': [0.5, 0.0, 1.0, 0.1, 100.0, 10.0, 50.0]
MLP	sklearn.neural_network.MLPClassifier	'alpha': [1e-4, 1e-3, 1e-2, 1e-1], 'learning_rate_init': [1e-3, 1e-2, 1e-1, 0.5, 1.]

Tabla 3.9: Selectores considerados con TPOT para clasificación

Selector	Módulo	Hiperparámetros
FWE	sklearn.feature_selection.SelectFwe	'alpha': np.arange(0, 0.05, 0.001), 'score_func': {'sklearn.feature_selection.f_regression': None}
Percentil	sklearn.feature_selection.SelectPercentile	'percentile': range(1, 100), 'score_func': {'sklearn.feature_selection.f_regression': None}
Límite de varianza	sklearn.feature_selection.VarianceThreshold	'threshold': [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2]
Eliminación de características	sklearn.feature_selection.RFE	'step': np.arange(0.05, 1.01, 0.05), 'estimator': {'sklearn.ensemble.ExtraTreesClassifier': { 'n_estimators': [100], 'criterion': ['gini', 'entropy'], 'max_features': np.arange(0.05, 1.01, 0.05)}}

Selección desde modelo	sklearn.feature_selection.SelectFromModel	'sklearn.feature_selection.SelectFromModel':
		{
		'threshold': np.arange(0, 1.01, 0.05),
		'estimator':
		{'sklearn.ensemble.ExtraTreesClassifier': {
		'n_estimators': [100],
		'criterion': ['gini', 'entropy'],
		'max_features': np.arange(0.05, 1.01, 0.05)}}}

3.11. Métricas de evaluación

El objetivo buscado en este trabajo es poder correlacionar el registro de parámetros con el perfil penetrométrico del suelo atravesado. Para comparar el perfil real con los resultados obtenidos podemos utilizar diferentes métricas descriptivas como serían la media, la mediana, los cuartiles 25 y 75 o la desviación típica. Todas estas estadísticas deberían presentar valores muy similares entre los valores reales y los calculados.

Además, al ser una predicción, podemos utilizar el coeficiente de correlación de Pearson, que nos da una idea de la relación lineal entre los valores reales y los valores predichos. Los valores pueden variar entre -1 y 1, mostrando el signo el sentido de la relación y siendo el valor absoluto de 1 una correlación perfecta. Quizá la métrica más común en predicciones sea el coeficiente de determinación (R^2) que mide la variación de los resultados que puede explicarse por el modelo.

No obstante, tal y como muestra la docena del datasaurio (Figura 3.26, Tabla 3.10) relaciones muy diferentes entre variables pueden dar lugar a métricas similares (Matejka & Fitzmaurice, 2017).

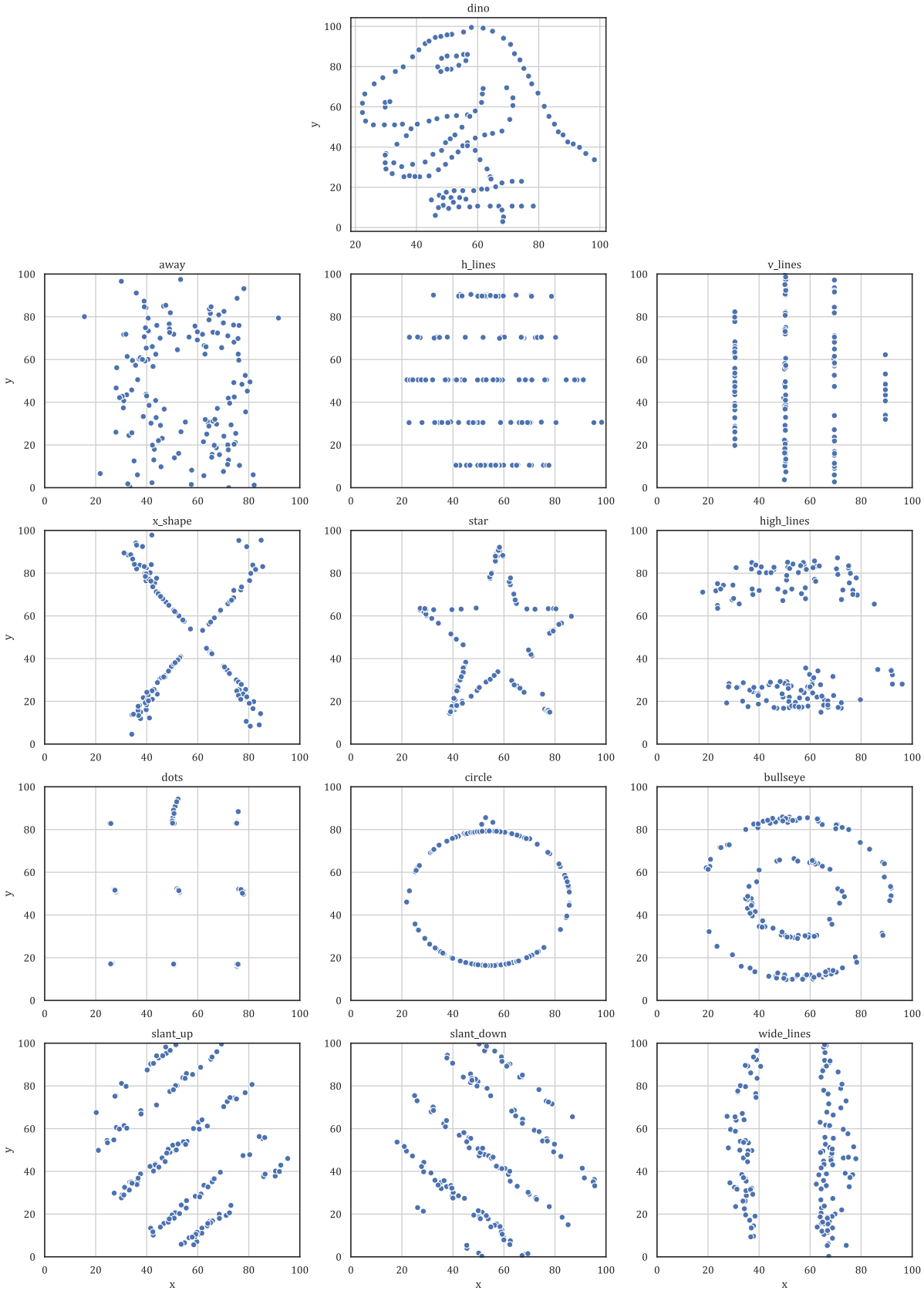


Figura 3.26: Distribución de los diferentes conjuntos de la docena del datasaurio

Tabla 3.10: Métricas de los diferentes conjuntos de la docena del datasaurio

	x		y		
Conjunto	**Muestras**	**Media**	**Desviación típica**	**Media**	**Desviación típica**	**Coef. Correlación**
away	142	54.2661	16.7698	47.8347	26.9397	-0.0641
bullseye	142	54.2687	16.7692	47.8308	26.9357	-0.0686
circle	142	54.2673	16.7600	47.8377	26.9300	-0.0683
dino	142	54.2633	16.7651	47.8323	26.9354	-0.0645
dots	142	54.2603	16.7677	47.8398	26.9302	-0.0603
h_lines	142	54.2614	16.7659	47.8303	26.9399	-0.0617
high_lines	142	54.2688	16.7667	47.8355	26.9400	-0.0685
slant_down	142	54.2678	16.7668	47.8359	26.9361	-0.0690
slant_up	142	54.2659	16.7689	47.8315	26.9386	-0.0686
star	142	54.2673	16.7690	47.8395	26.9303	-0.0630
v_lines	142	54.2699	16.7700	47.8370	26.9377	-0.0694
wide_lines	142	54.2669	16.7700	47.8316	26.9379	-0.0666
x_shape	142	54.2602	16.7700	47.8397	26.9300	-0.0656

Resulta por tanto deseable graficar los datos en la medida de lo posible. Por ejemplo, podemos utilizar diagramas de caja a lo largo de la profundidad para comparar la distribución real con la predicha (Figura 3.27).

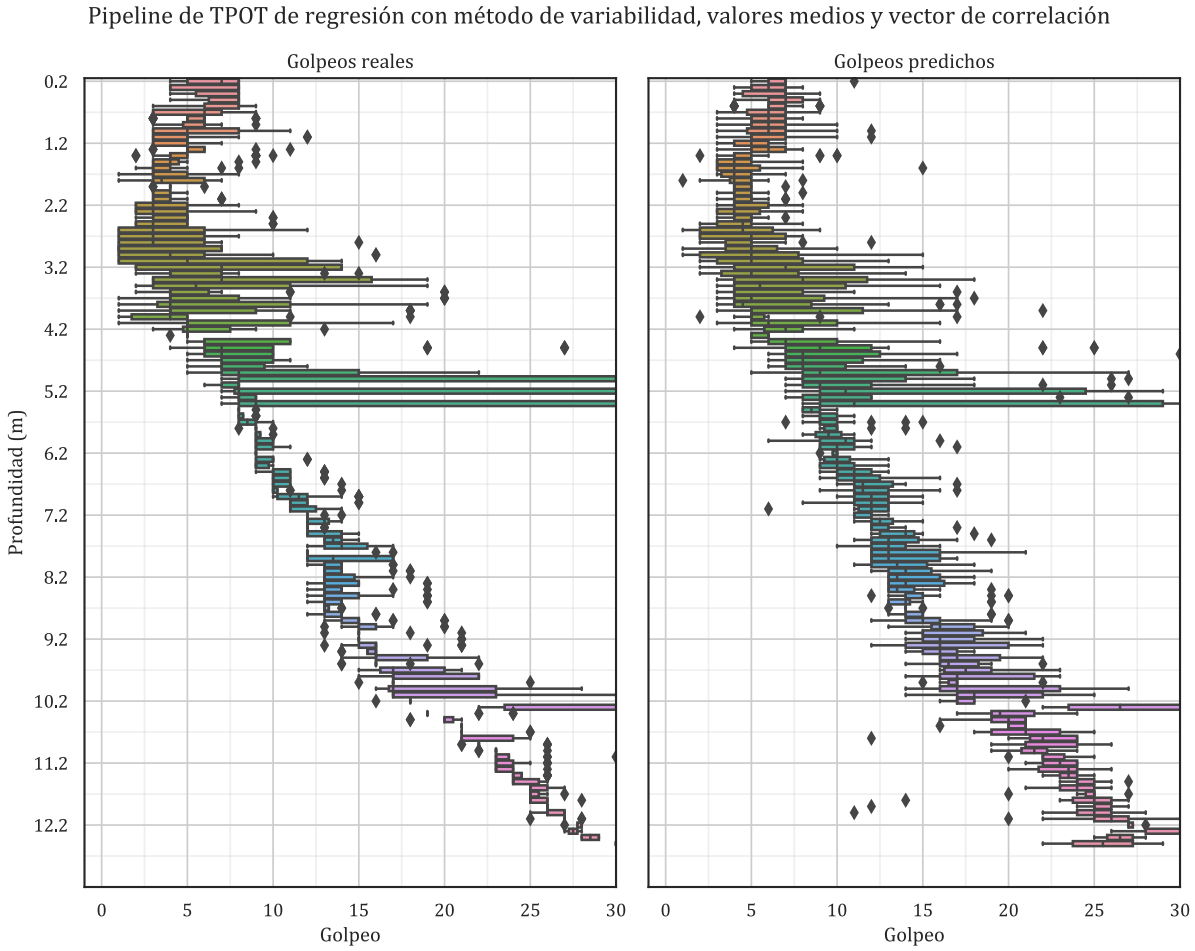


Figura 3.27: Representación mediante diagrama de caja de la distribución real de los golpeos de los penetrómetros de la Obra 1 junto con la distribución de los golpeos predichos para uno de los algoritmos analizados

Además, otras métricas necesarias para comparar los valores reales con los predichos serían el error medio (*mean average error*, MAE) y la raíz del error cuadrático medio (*root of mean squared error*, RMSE). El coeficiente de correlación, la raíz del error cuadrático medio de patrón centrado y la desviación típica pueden representarse en un único gráfico denominado diagrama de Taylor (Taylor, 2001) (Figura 3.28). En estos gráficos el radio representa la desviación típica, el ángulo representa el coeficiente de correlación y los círculos concéntricos que parten desde la referencia representan la RMSE de patrón centrado. Estos datos permiten comparar rápidamente distintos modelos entre sí.

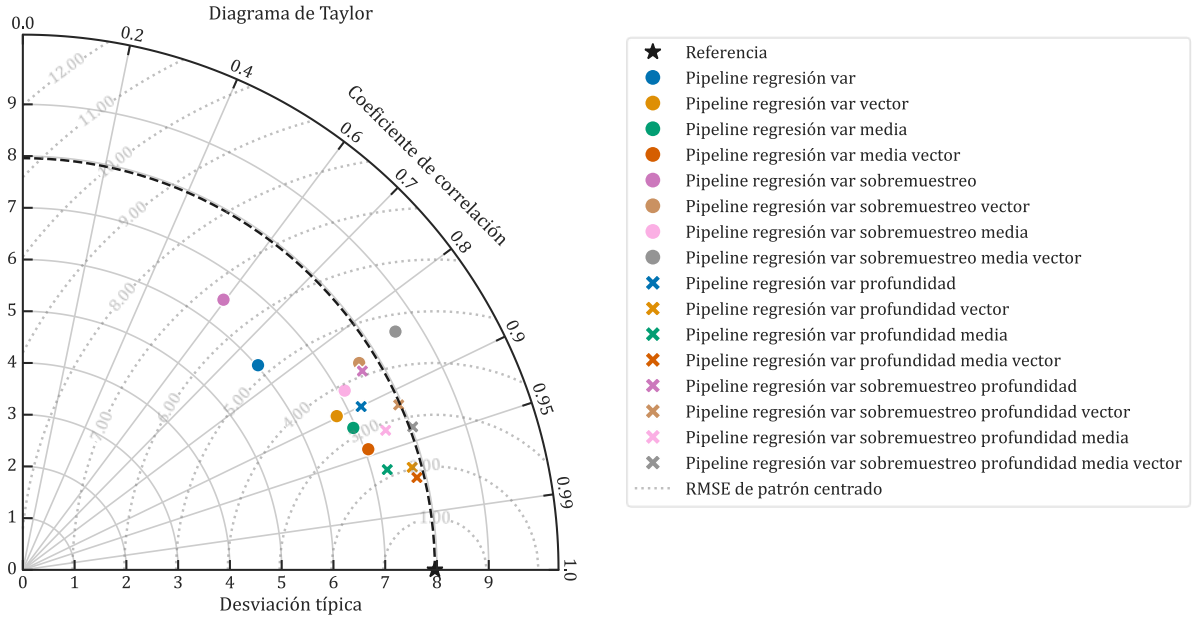


Figura 3.28: Diagrama de Taylor para los pipelines de TPOT de regresión con método de variabilidad para la Obra 1

Es necesario hacer hincapié en que la RMSE que se representa no es la RMSE normalmente utilizada sino que se resta previamente la media.

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

Ecuación 3.5: RMSE

$$RMSE\ de\ patrón\ centrado = \sqrt{\sum_{i=1}^{n} \frac{\left((\hat{y}_i - \bar{\hat{y}}) - (y_i - \bar{y})\right)^2}{n}}$$

Ecuación 3.6: RMSE de patrón centrado

Donde $\hat{y}_i$ es el valor predicho e y_i el valor real.

Esta RMSE de patrón centrado puede también interpretarse como la desviación típica de los errores (Elvidge et al., 2014). Los valores obtenidos serán, en general, similares a los de la RMSE normal, pero no exactamente iguales.

Por otro lado, también resulta conveniente graficar los errores residuales, ya sea de manera general (Figura 3.29) o en función de la profundidad (Figura 3.30).

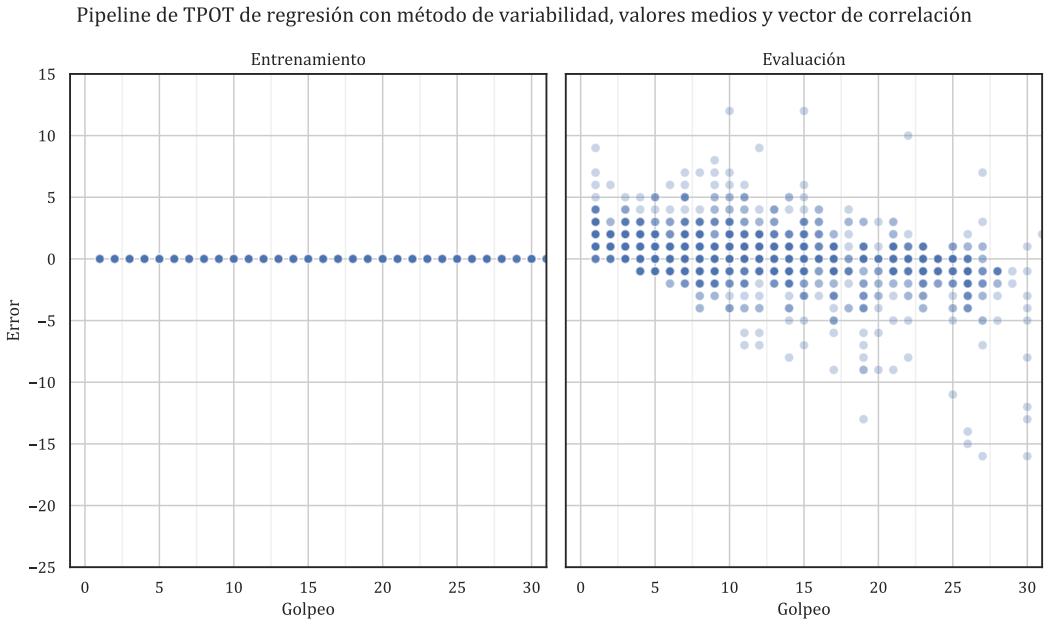


Figura 3.29: Error residual en la Obra 1 para uno de los algoritmos analizados

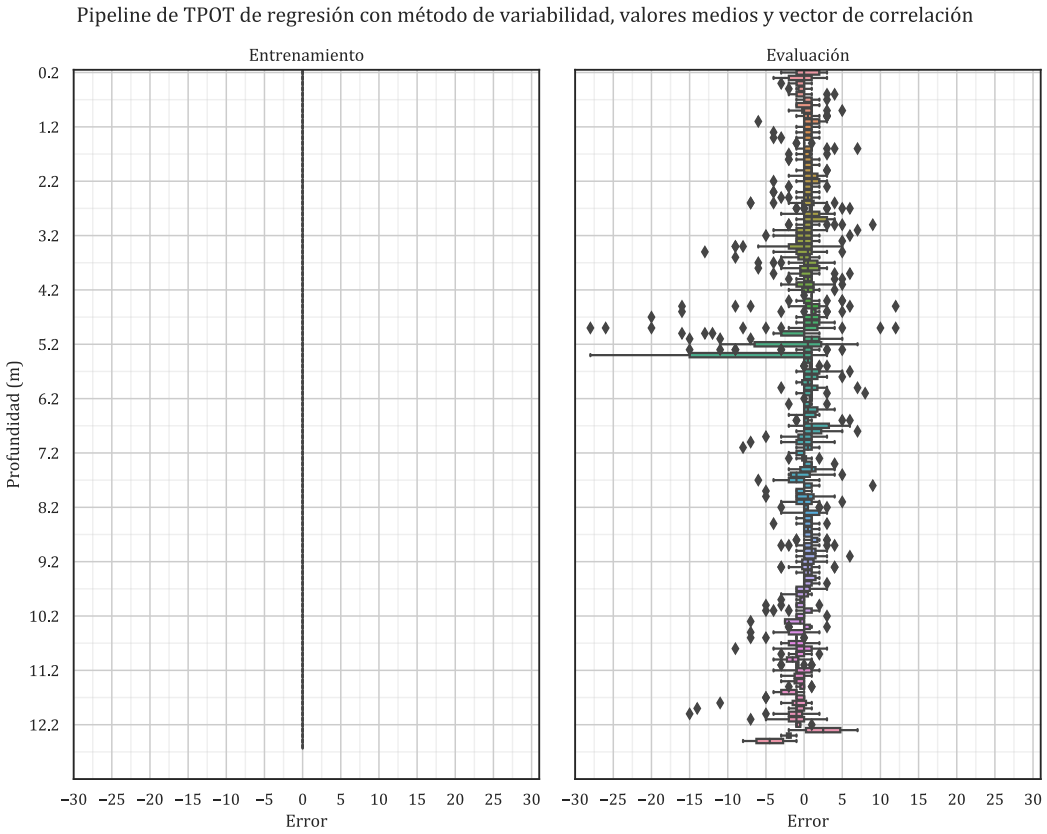


Figura 3.30: Error residual en función de la profundidad para la Obra 1 según uno de los algoritmos analizados

Por último, otra opción interesante es la propuesta de Charles et al (Charles et al., 2023) de utilizar *dynamic time warping* (DTW, (Morse & Patel, 2007)) para comparar penetrómetros entre sí. DTW es un algoritmo pensado para la comparación de series temporales. Este algoritmo prueba todas las posibles combinaciones para desplazar una serie temporal hasta igualarla con otra. La distancia mínima de desplazamiento nos da una idea de la similitud entre ambas series. Las combinaciones pueden mostrarse como una matriz de costos de alineación (Figura 3.31) y también es posible mostrar los movimientos para convertir la serie temporal (Figura 3.32), sin embargo, al comprobar muchas inclusiones no resulta razonable mostrar estos gráficos para todas ellas.

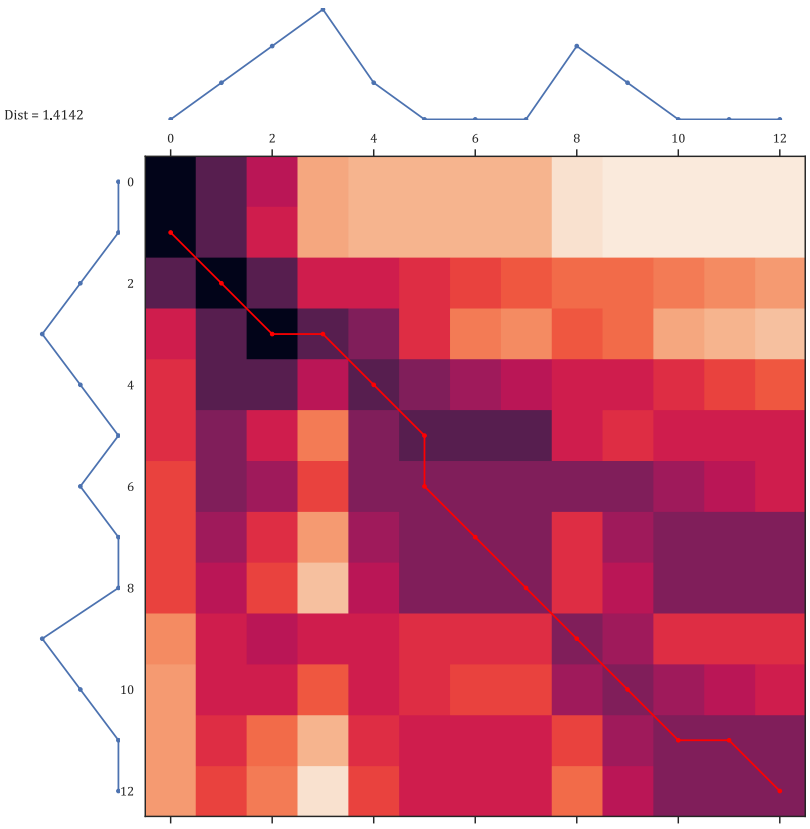


Figura 3.31: Matriz de costos de alineación

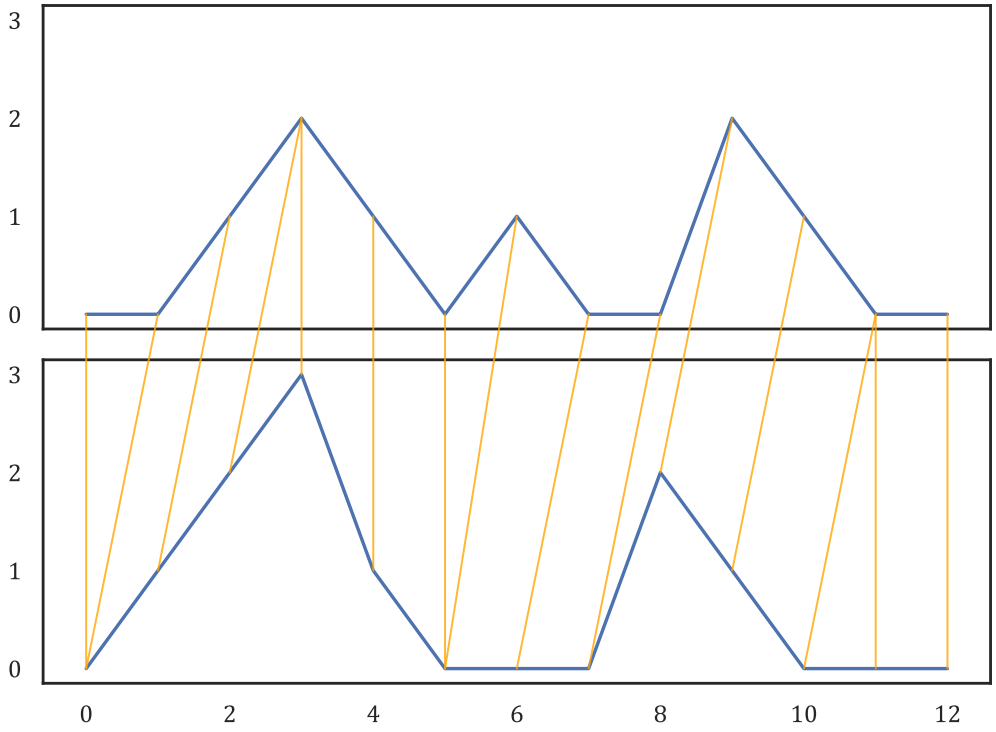


Figura 3.32: Conversión utilizando DTW

La idea detrás de utilizar DTW es calcular la distancia media entre los valores del penetrómetro real y los predichos como una métrica de la correlación entre ambas series. Hay que tener en cuenta que las series han de estar ordenadas para calcular la distancia. No es un cálculo que se pueda hacer a partir de muestras, si no teniendo en cuenta los valores a lo largo de cada inclusión. Por tanto, estos cálculos tienen una complejidad mayor que calcular simplemente el coeficiente de correlación o los errores, pero tienen la gran ventaja de tener en cuenta la forma de los resultados.

Por ejemplo, en la Figura 3.33 se muestra el perfil original frente una predicción que sobreestima en 10 puntos los golpeos, y frente a una predicción con cierto desfase. Si calculamos el coeficiente de correlación entre el perfil original y el sobrestimado el resultado es 1, mientras que la correlación con el perfil desfasado es de 0.96. Sin embargo, el perfil desfasado es claramente preferible al perfil sobrestimado. Además, que haya un cierto desfase tiene sentido geotécnico cuando aumenta la distancia entre la inclusión y el penetrómetro por la inclinación de las capas de suelo.

En cambio, utilizando DTW, la distancia para el perfil sobrestimado es de 74.6 mientras que para el perfil desfasado es de 7.4, diez veces menor.

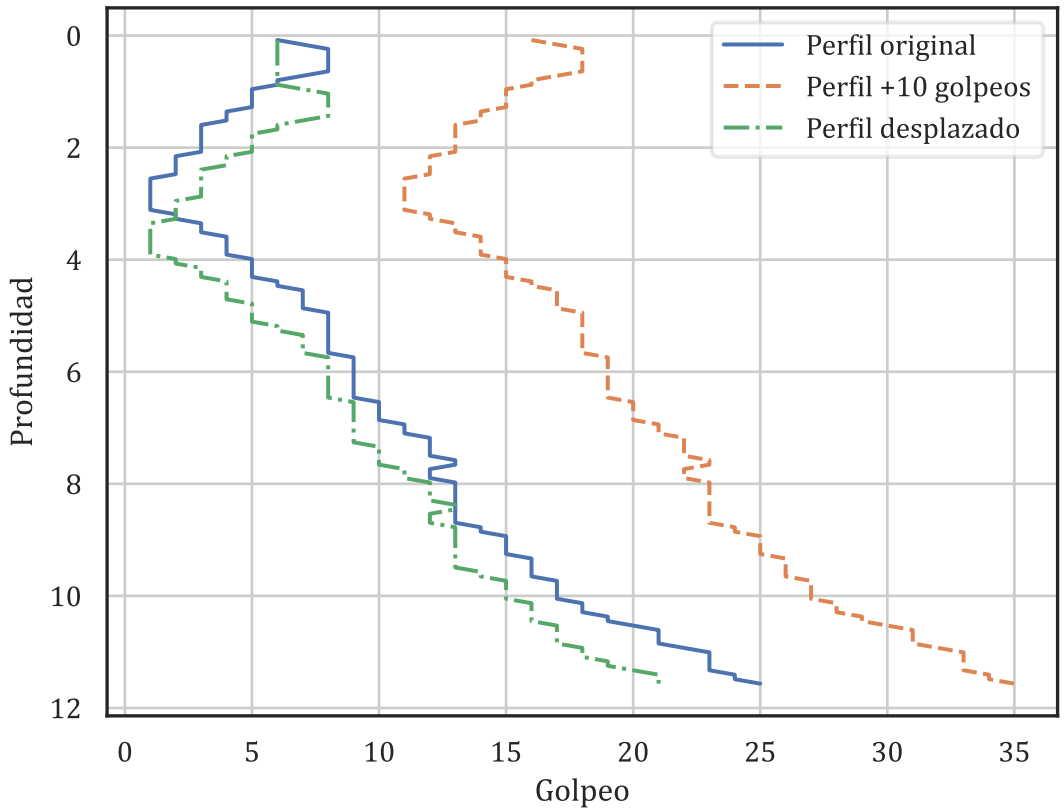


Figura 3.33: Perfil original, perfil sobrestimado en 10 puntos y perfil desplazado

De esta manera, las métricas utilizadas para comprobar la eficacia de los algoritmos son las siguientes:

- Coeficiente de correlación de Pearson, cuya formulación sería la siguiente:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} \cdot \sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

Ecuación 3.7: Coeficiente de correlación de Pearson

- MAE

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

Ecuación 3.8: MAE

- RMSE, cuya formulación ya se ha mostrado previamente.
- Coeficiente de determinación (R^2)

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

Ecuación 3.9: Coeficiente de determinación

- Distancia media obtenida mediante DTW
- Valor F

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

Ecuación 3.10: Valor F

Donde:

TP: verdaderos positivos (*true positives*)

FP: falsos positivos (*false positives)*

FN: falsos negativos (*false negatives*)

El valor F, a diferencia de las otras métricas planteadas, mide la eficiencia de la predicción teniendo en cuenta un problema de clasificación, y no de regresión.

3.12. Explicabilidad

La explicabilidad, dentro del marco de ML, se refiere al grado en que las personas podemos comprender la causa de una decisión. En algoritmos simples, los resultados son fácilmente entendibles, por ejemplo, en una regresión lineal con pocas características o en un único árbol de decisión todo el proceso del algoritmo puede representarse gráficamente de manera relativamente sencilla.

Sin embargo, con algoritmos más complejos o en aquellos que forman parte de un proceso mayor, el explicar de dónde provienen los resultados se vuelve más complicado. En esos casos, aunque no analicemos pormenorizadamente el modelo generado, si podemos observar cuál es el efecto de las diferentes características sobre los resultados obtenidos. Una manera sería utilizar gráficos de expectativa individual condicionada (*individual conditional expectation*, ICE). La idea tras estos gráficos es que, para cada muestra, se prueban diferentes valores de una característica, manteniendo el resto fijas. Se ha de tener en cuenta, no obstante, que si dos características están relacionadas pueden darse combinaciones imposibles, ya que en la realización del gráfico sólo se varía una cada vez. El resultado es un conjunto de líneas que muestran la variación del resultado en función del cambio de cada característica (Figura 3.34).

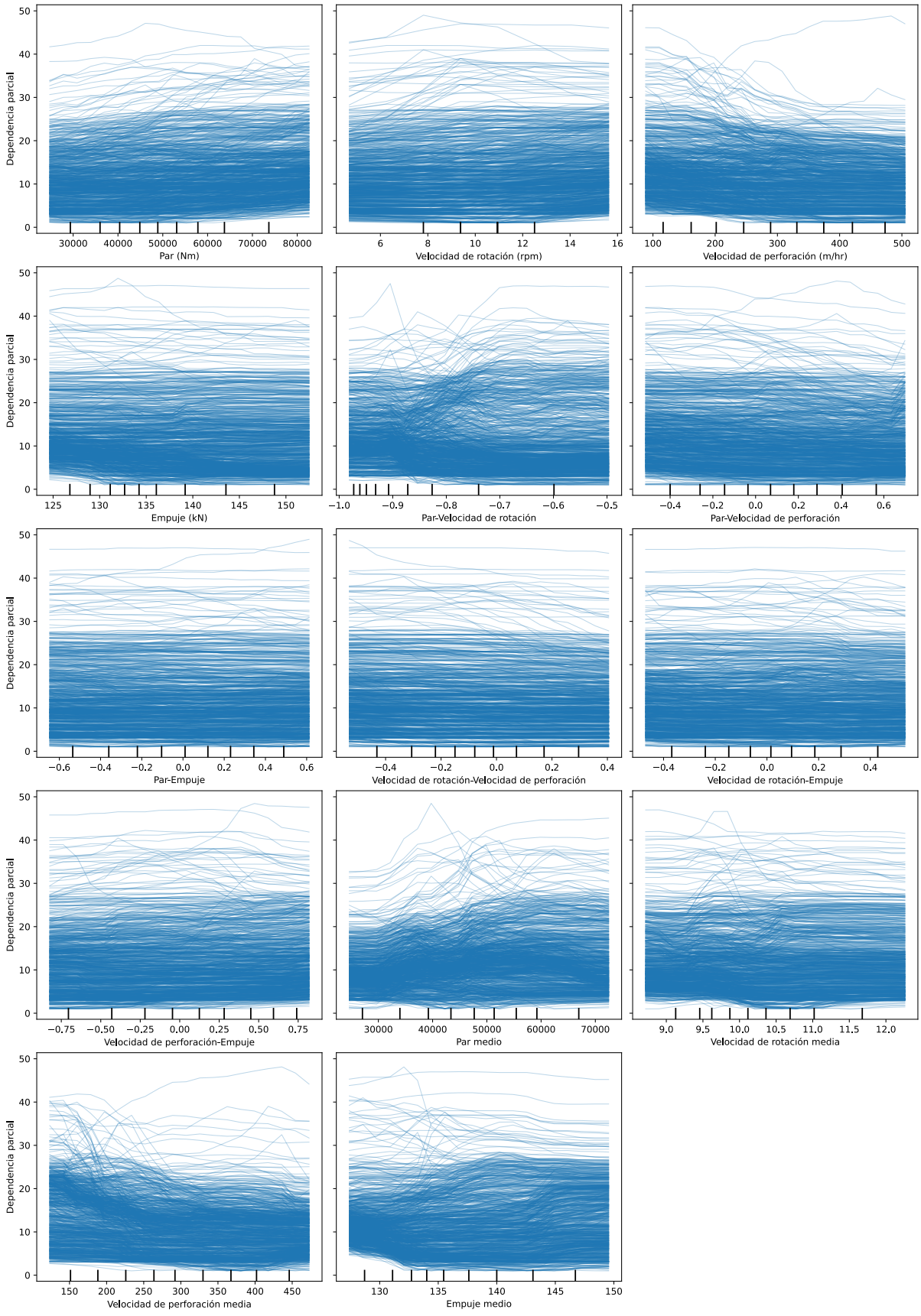


Figura 3.34: Gráfico de expectativa individual condicionada

Si se realiza la media de los resultados obtenidos en un gráfico ICE obtenemos un gráfico de dependencia parcial (*partial dependency plot*, PDP, Figura 3.35).

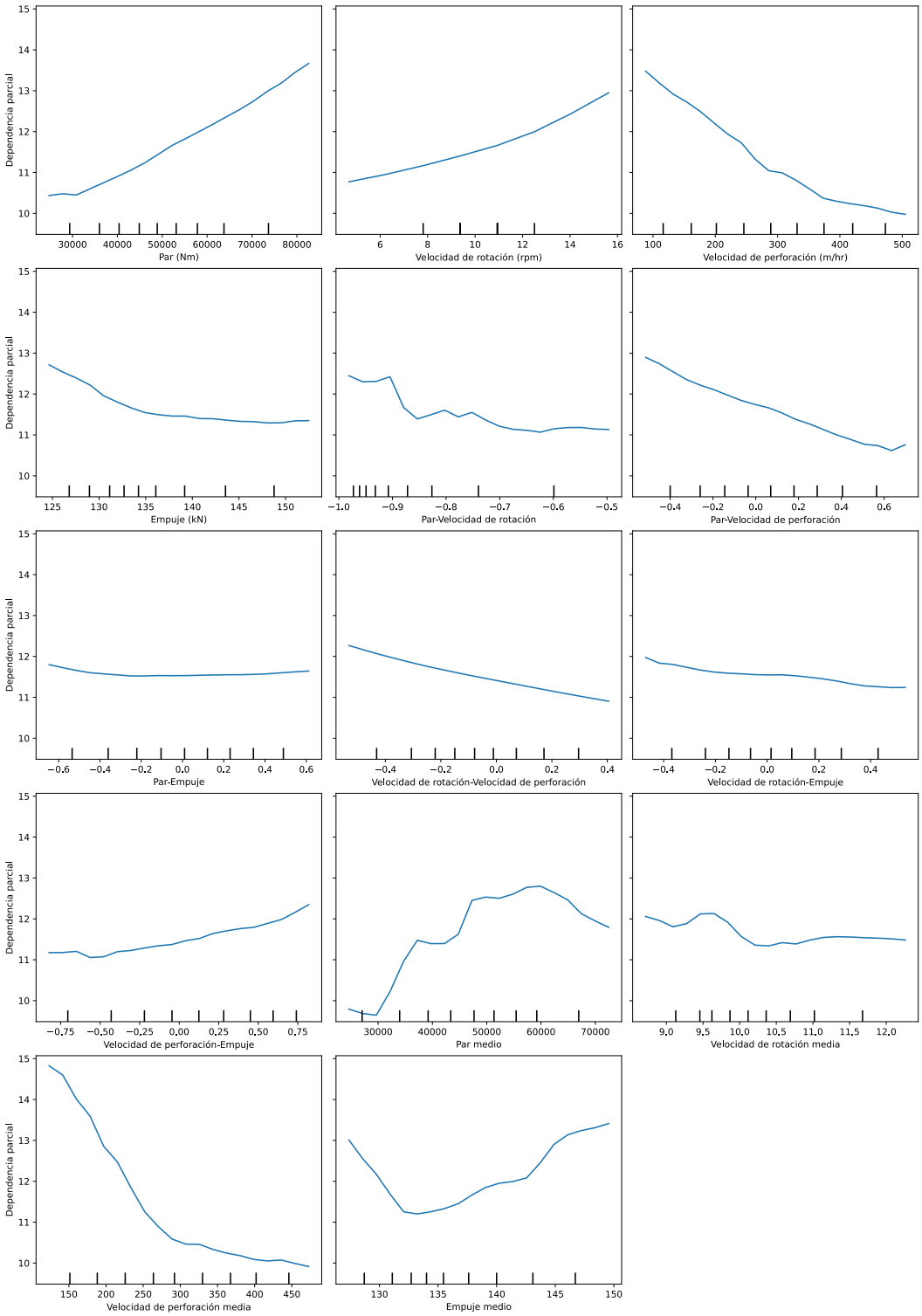


Figura 3.35. Gráfico de dependencia parcial

Otra manera de conocer el efecto de las diferentes características en el resultado final del modelo, es recurrir a la teoría de juegos donde los valores Shapley (Shapley, 1953) hacen referencia a la contribución de cada característica al resultado final del modelo. Aplicado a ML podemos utilizar SHAP (*SHapley Additive exPlanations*, (Lundberg & Lee, 2017)). El cálculo de la contribución de cada variable se realiza, para una única muestra, calculando el modelo para todas las posibles combinaciones de características, y ponderando el efecto de añadir una característica en concreto a un modelo que antes no la incluía. Si bien sería posible calcular de manera exacta estos valores, el número de modelos a realizar aumenta exponencialmente con el número de características, por lo que SHAP se limita a aproximar los valores de contribución. Una vez calculados los valores de contribución de las diferentes características para cada muestra, los resultados pueden presentarse en un diagrama de enjambre donde se muestra el efecto que los valores altos y bajos de cada característica tienen sobre el resultado del modelo (Figura 3.36).

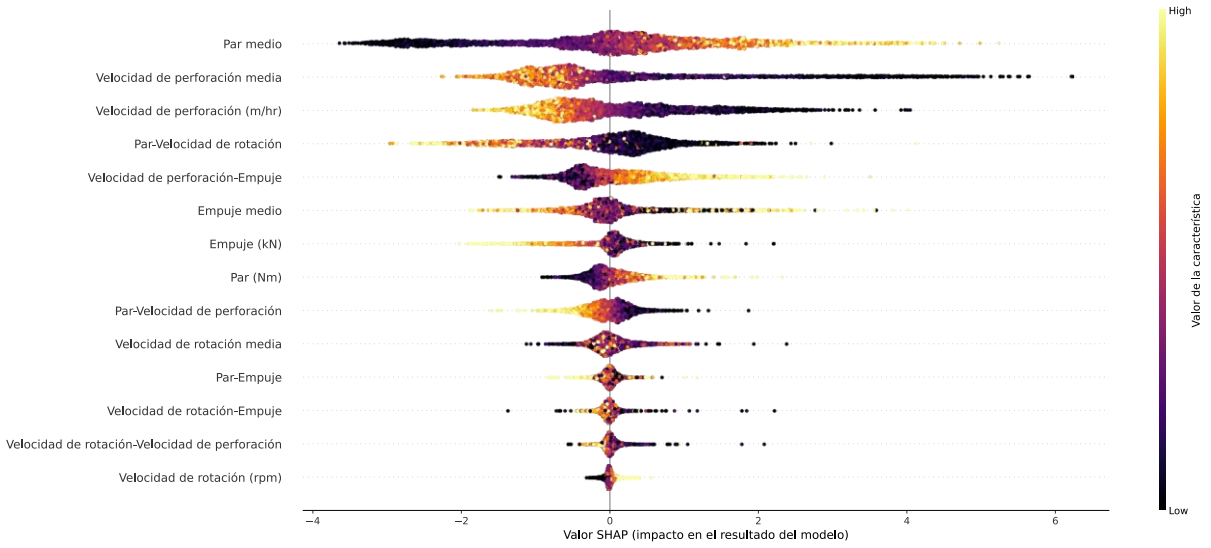


Figura 3.36: Diagrama de enjambre

3.13. Procedimiento aplicado

La metodología seguida en este trabajo ha sido la siguiente:

- Mediante Python y Jupyter Notebook se observan inicialmente los datos disponibles del registro de parámetros y de los penetrómetros.
- Se transforman los datos al formato AGS.
- Se eliminan las anomalías en los penetrómetros mediante el uso de media móvil.
- Se eliminan las anomalías en el registro de parámetros usando *isolation forest*.
- Se realiza el etiquetado de las muestras.
- Se constituyen los conjuntos de datos a utilizar usando el método de variabilidad y el método continuo.
- Se generan las nuevas características (media y vector de correlación).
- Se calcula el mejor proceso de ML utilizando TPOT para las diferentes combinaciones de las siguientes características de los algoritmos (Tabla 3.11):
    - Tipo de modelo: clasificación o regresión.
    - Conjunto de datos: método de variabilidad o método continuo.
    - Conjunto de características: características iniciales (velocidad de rotación, velocidad de penetración, par de rotación y empuje aplicado); profundidad de la perforación; valores medios de las características a lo largo de 1,5 m por encima de la punta; vector de correlación a lo largo de 1,5 m por encima de la punta.
    - Uso o no de sobremuestreo.
- Se calculan las métricas de evaluación para los diferentes procesos obtenidos en el paso previo. Se calculan como cambian estas métricas para diferentes divisiones entrenamiento/evaluación, para diferentes tamaños del conjunto de datos y para diferentes distancias entre la inclusión y el penetrómetro.
- Se realizan los mismos cálculos, pero para un algoritmo único (sin *pipeline* completo).
- Se comparan la capacidad de aproximar el perfil penetrómetrico de los algoritmos con la de los parámetros compuestos.

Tabla 3.11: Diferentes combinaciones analizadas

Clasificación/regresión		Conjunto de datos		Sobremuestreo	Características			
Clasificación	Regresión	Variabilidad	Continuo	Sobremuestreo	Iniciales	Prof.	Media 1,5 m	Vector de correlación
■		■			■			
■		■			■	■		
■		■			■		■	
■		■			■			■
■		■			■	■	■	
■		■			■	■		■
■		■			■		■	■
■		■			■	■	■	■
■		■		■	■			
■		■		■	■	■		
■		■		■	■		■	
■		■		■	■			■
■		■		■	■	■	■	
■		■		■	■	■		■
■		■		■	■		■	■
■		■		■	■	■	■	■
■			■		■			
■			■		■	■		
■			■		■		■	
■			■		■			■
■			■		■	■	■	
■			■		■	■		■
■			■		■		■	■
■			■		■	■	■	■
■			■	■	■			
■			■	■	■	■		
■			■	■	■		■	
■			■	■	■			■
■			■	■	■	■		■
■			■	■	■	■	■	■
	■	■			■			
	■	■			■	■		
	■	■			■		■	
	■	■			■			■
	■	■			■	■	■	
	■	■			■	■		■
	■	■			■		■	■
	■	■			■	■	■	■
	■	■		■	■			
	■	■		■	■	■		
	■	■		■	■		■	
	■	■		■	■			■
	■	■		■	■	■	■	
	■	■		■	■	■		■
	■	■		■	■		■	■
	■	■		■	■	■	■	■
	■		■		■			
	■		■		■	■		
	■		■		■		■	
	■		■		■			■
	■		■		■	■	■	
	■		■		■	■		■
	■		■		■		■	■
	■		■		■	■	■	■
	■		■	■	■			
	■		■	■	■	■		
	■		■	■	■		■	
	■		■	■	■			■
	■		■	■	■	■	■	
	■		■	■	■	■		■
	■		■	■	■		■	■
	■		■	■	■	■	■	■

En la Figura 3.37 se muestra el esquema del procedimiento aplicado.

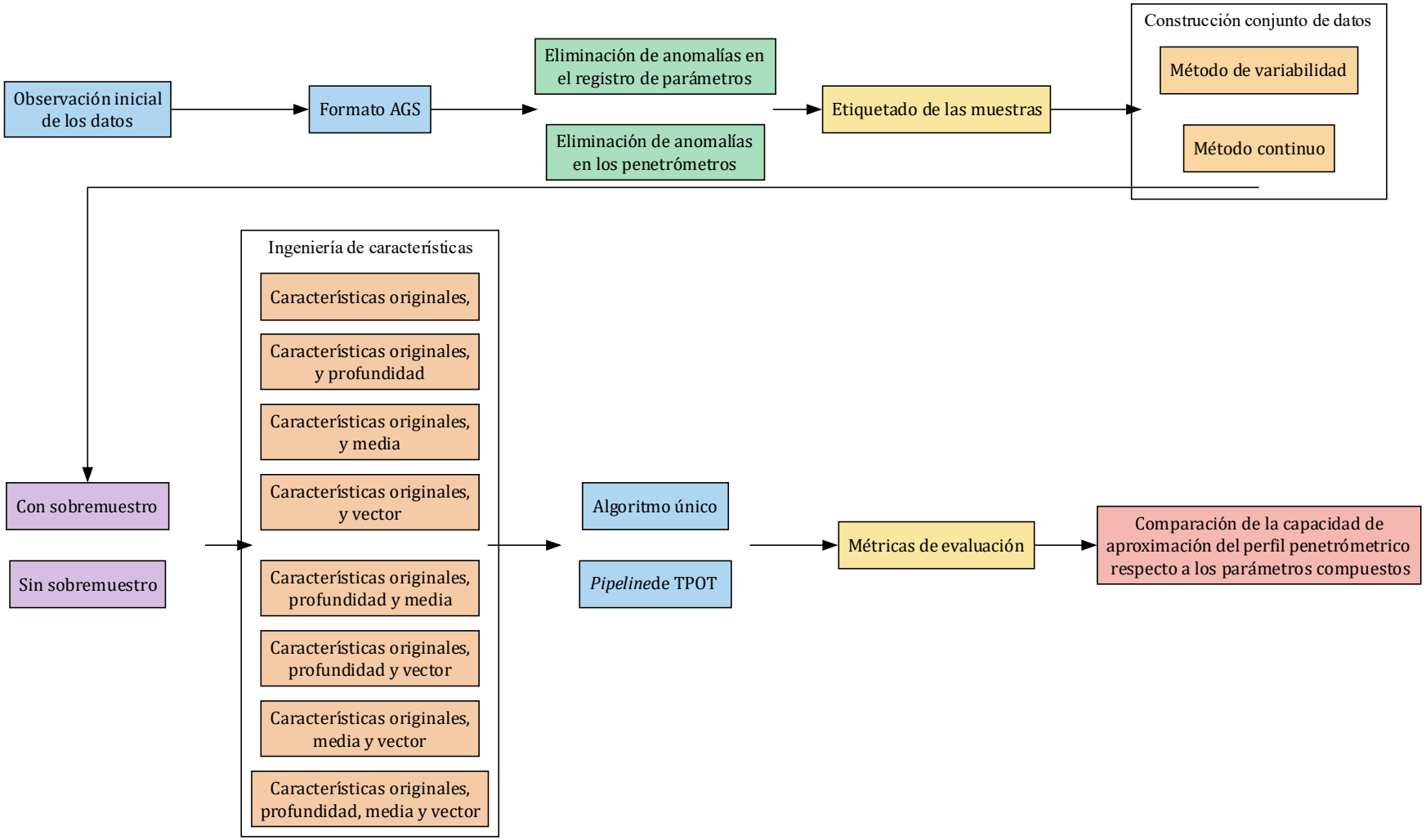


Figura 3.37: Esquema del procedimiento aplicado

Este procedimiento se repite en todas las obras descritas. Adicionalmente se comprueban los resultados que se obtienen cuando se aplica un modelo a una obra en la que no ha sido entrenado previamente.

4. Resultados y discusión

4.1. Introducción

En el presente capítulo, *Resultados*, se muestran comentados y resumidos los principales valores obtenidos tras los cálculos realizados a lo largo de la presente tesis. De esta manera se analiza el efecto de las acciones llevadas a cabo y de las diferentes herramientas utilizadas.

Los principales resultados presentados son los siguientes:

- Detección de anomalías,
- Construcción del conjunto de datos en función de la información geotécnica,
- Resultados obtenidos con TPOT,
- Capacidad predictiva de los algoritmos,
- Efecto de la distancia entre el penetrómetro y la inclusión,
- Efecto del tamaño del conjunto de datos,
- Comparación con los parámetros compuestos.

No obstante, y antes de discutir con mayor profundidad los resultados, se muestran una serie de figuras (4.1 a 4.5) en las que aparece, para una serie de columnas y para las diferentes obras, el perfil penetrométrico real y el perfil calculado como muestra de la capacidad de los modelos para aproximar los valores del penetrómetro. Hay que señalar que los resultados de los modelos no se suavizan mediante media móvil, por lo que presentan un perfil con picos algo más acusados que los del perfil penetrométrico. Se muestran así para comparar el valor objetivo (el perfil penetrométrico suavizado) con los resultados en bruto de los modelos.

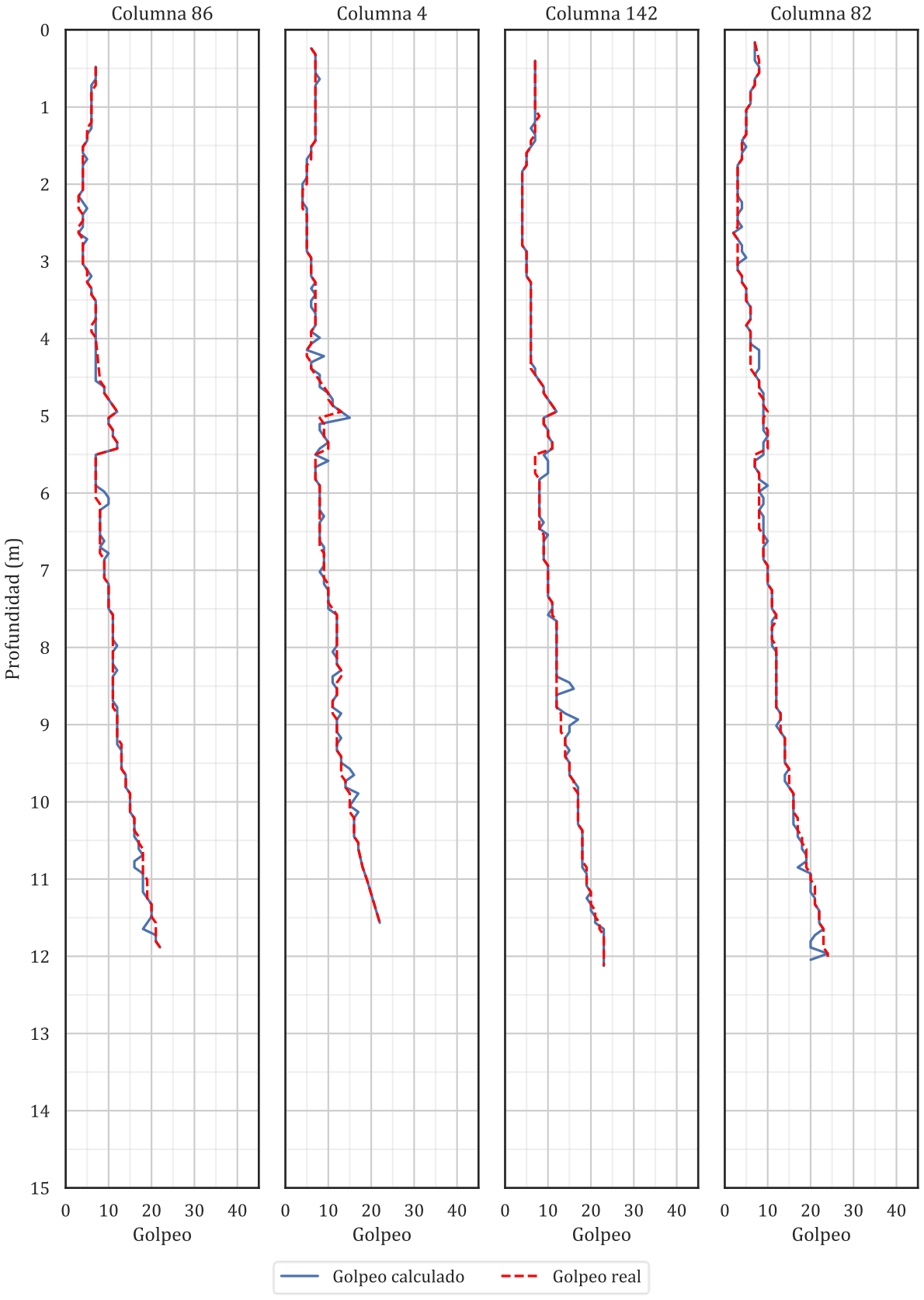


Figura 4.1: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 1 para ETR con características originales, medias y vector de correlación.

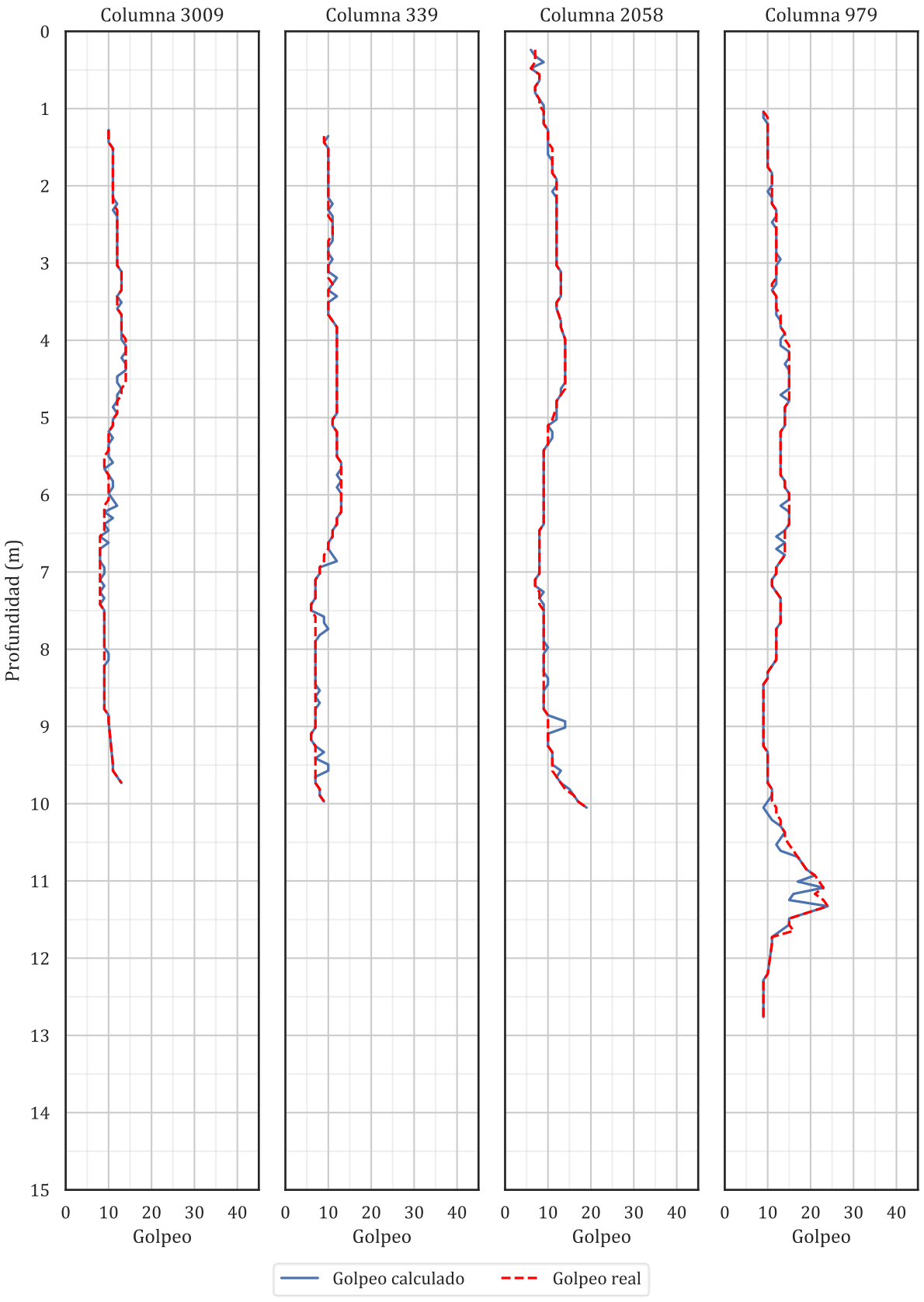


Figura 4.2: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 2 para ETR con características originales, medias y vector de correlación.

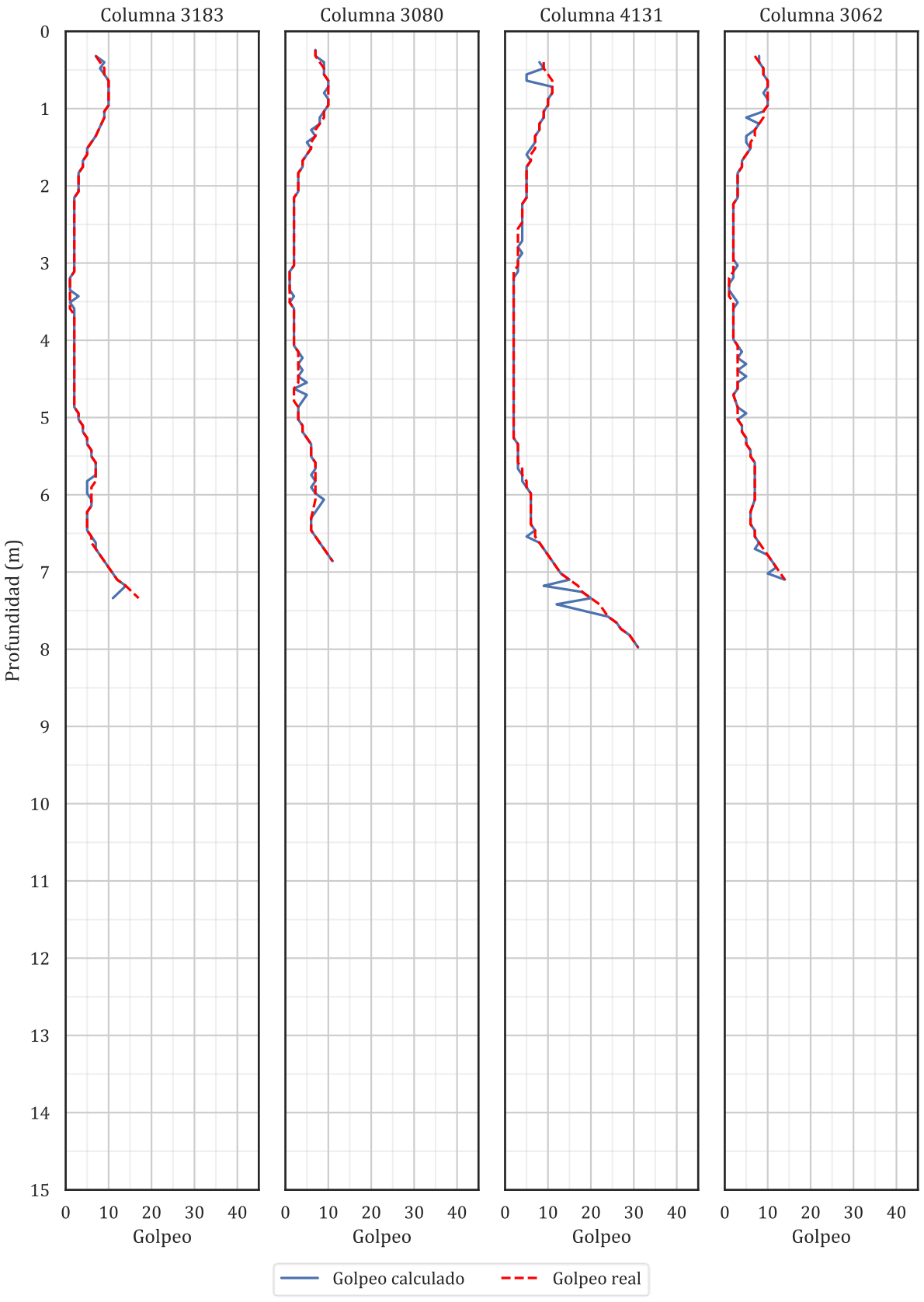


Figura 4.3: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 3 para ETR con características originales, medias y vector de correlación.

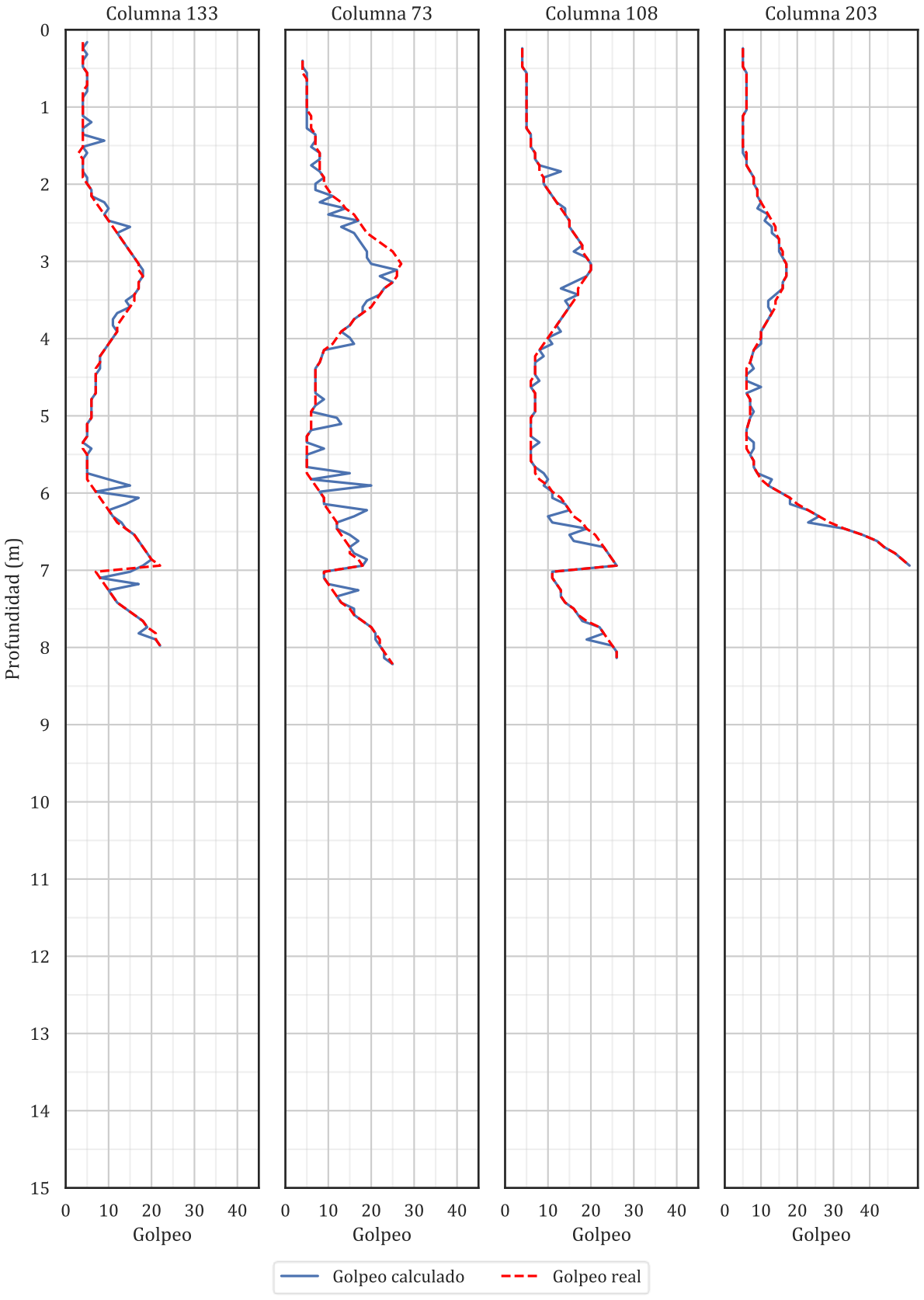


Figura 4.4: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 4 para ETR con características originales, medias y vector de correlación.

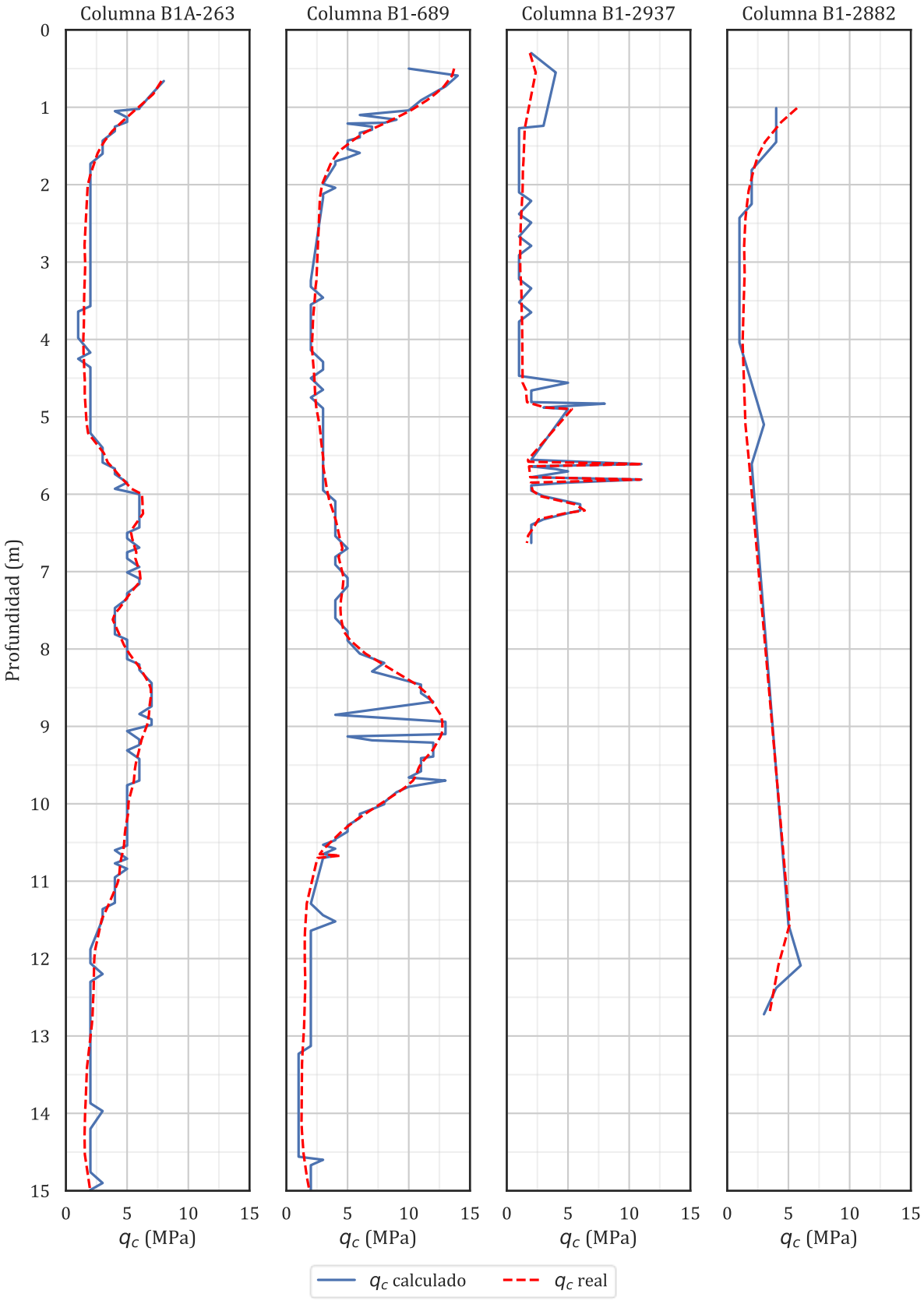


Figura 4.5: Ejemplos de perfiles penetrométricos reales y calculados para la Obra 5 para ETR con características originales, medias y vector de correlación.

4.2. Detección de anomalías

Las anomalías en los penetrómetros se han enfrentado utilizando una media móvil, tal y como se ha reflejado en el apartado de metodología. Por otro lado, las anomalías en los datos de MWD se han detectado utilizando *isolation forest* para, posteriormente, eliminar esas muestras.

En la Figura 4.6 se muestra la distribución de las diferentes características y la relación entre éstas (gráfico por pares) para la Obra 1, a modo de ejemplo.

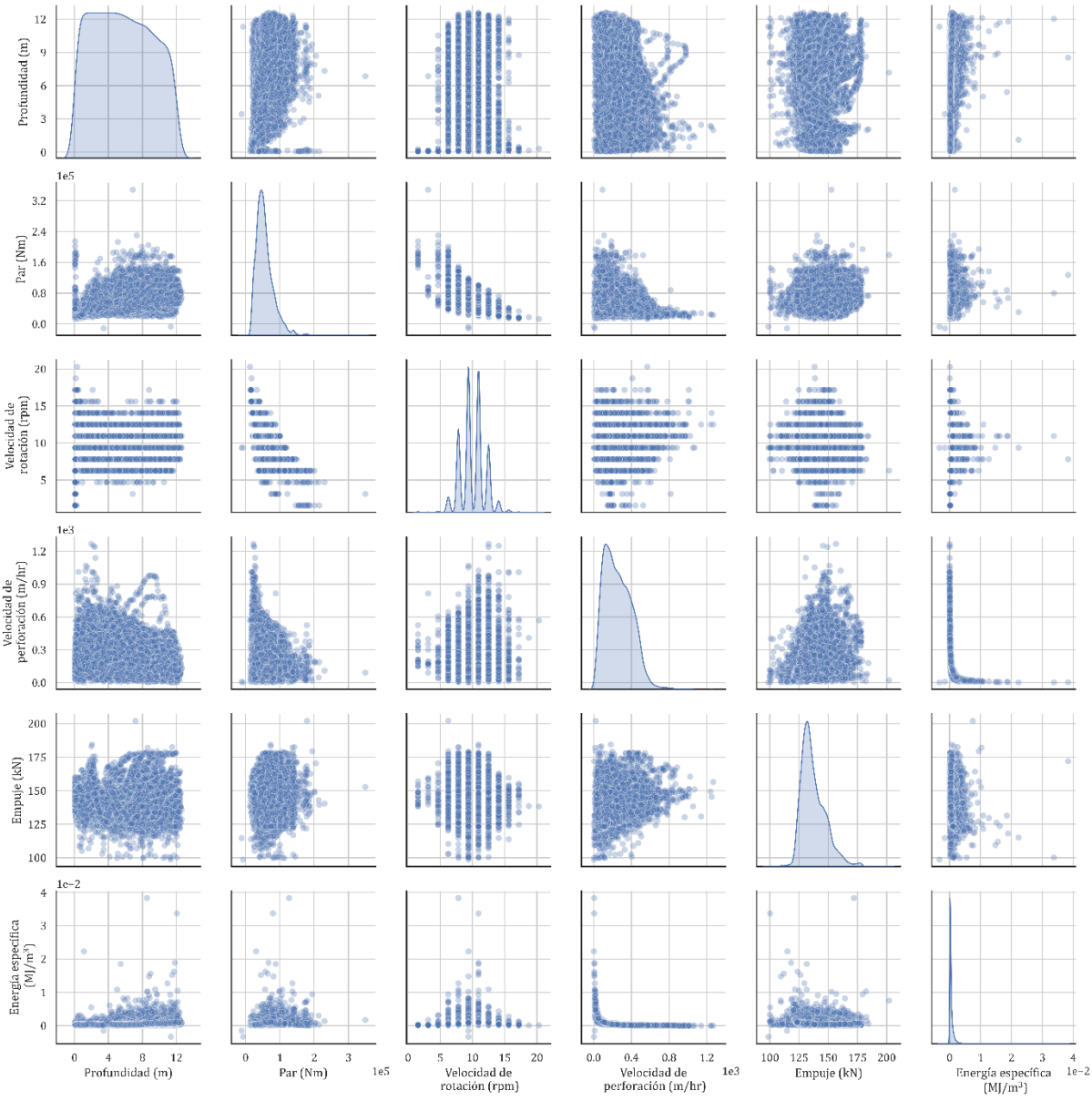


Figura 4.6: Gráfico por pares con características de la Obra 1

No se observan relaciones fuertes entre las diferentes características, aunque sí aparecen tendencias lógicas como menor velocidad de perforación y rotación a mayor par aplicado. También pueden apreciarse valores anómalos como par mayores de 180000 N·m (superiores a lo que se espera del equipo) o valores de par elevados a profundidades someras. Estas distribuciones de las características se mantienen de forma similar para las diferentes obras.

En la Figura 4.7 se muestra el gráfico por pares de la Obra 1 mostrando, en rojo, las muestras que *isolation forest* considera anomalías y, en verde, el resto. En este caso se marcan como anomalías el 5% de las muestras que el algoritmo considera más «extrañas».

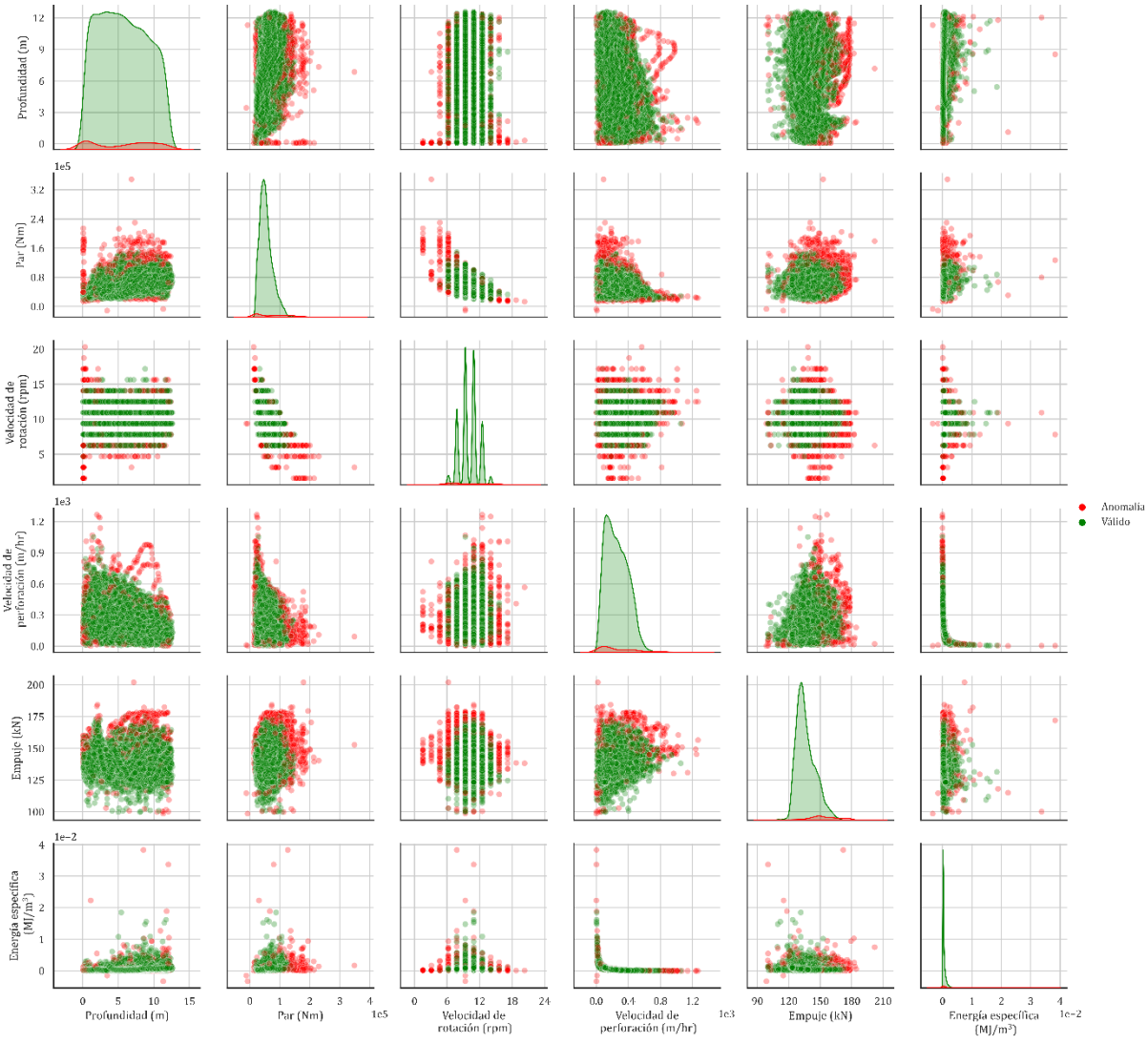


Figura 4.7: Detección de anomalías para la Obra 1

Puede comprobarse cómo el algoritmo detecta con éxito los valores de par elevados, así como aquellos valores de par altos a escasa profundidad. Estas últimas anomalías no serían detectadas simplemente estableciendo un valor máximo de par.

Sin embargo, no en todas las obras se tiene tanto éxito. En la Figura 4.8 se muestran los valores considerados válidos por *isolation forest* para la Obra 4. Como vemos, aún se mantienen, por ejemplo, valores de par negativos.

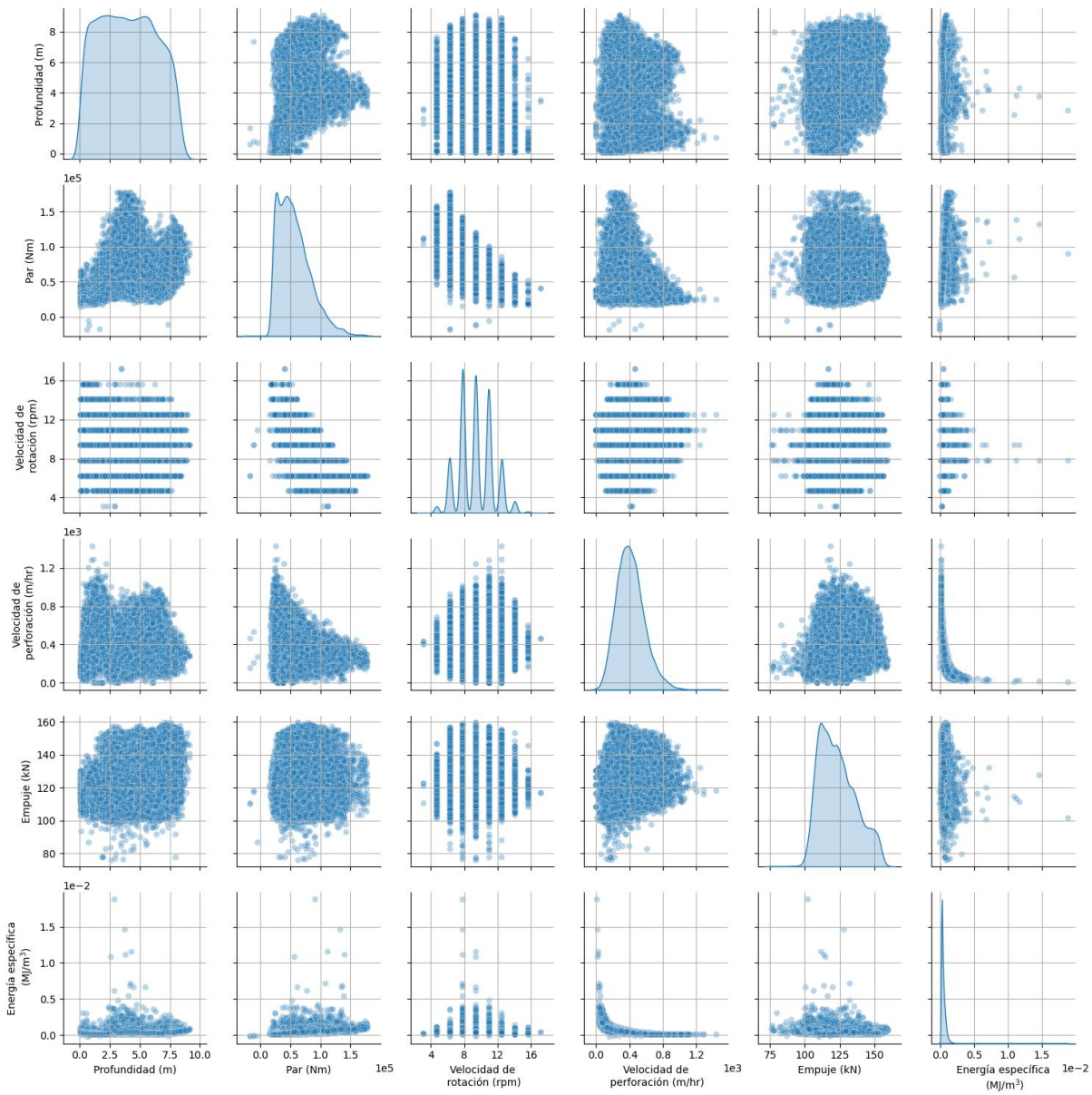


Figura 4.8: Gráfico por pares de la Obra 4 tras la eliminación de anomalías

Aún sin ser perfecto, *isolation forest* parece ser capaz de eliminar casi la totalidad de los valores que parecen anomalías.

4.3. Construcción del conjunto de datos en función de la distancia

Se han utilizado los métodos de variabilidad y continuo. Para el método de variabilidad se considera una variación de golpeo de 2 para el penetrómetro dinámico y una variación de resistencia por punta de 0.5 MPa para el penetrómetro estático (Obra 5). Para el método continuo se utiliza $\alpha = 0.2$ y $L_{min} = 10$. El efecto de ambos métodos difiere en función de la obra, como se aprecia en la Tabla 4.1.

Tabla 4.1: Número de muestras según los métodos de variabilidad y continuo para las diferentes obras

	Número de muestras		
Obra	Original	Variabilidad	Continuo
Obra 1	17535	6472	14018
Obra 2	252189	54129	43566
Obra 3	105488	4394	28341
Obra 4	26242	411	21471
Obra 5	358925	42125	64965

La proporción de muestras sobre el total varía mucho entre obras y métodos. Aquellas obras donde los penetrómetros varían mucho entre sí tendrán menos muestras por el método de variabilidad. Aquellas otras que tengan pocos penetrómetros o estos estén muy separados entre sí, tendrán menos muestras por método continuo (manteniendo α constante).

Hay que destacar el caso de la Obra 4. En la Figura 4.9 se muestra el resultado de aplicar el método de variabilidad. La distancia al penetrómetro que se obtiene es de, tan solo, 1.29 m reduciéndose mucho el número de muestras. Esto ocurre en

este caso porque los penetrómetros, aunque similares entre sí, presentan importantes diferencias en la altura a la que aparecen las diferentes capas.

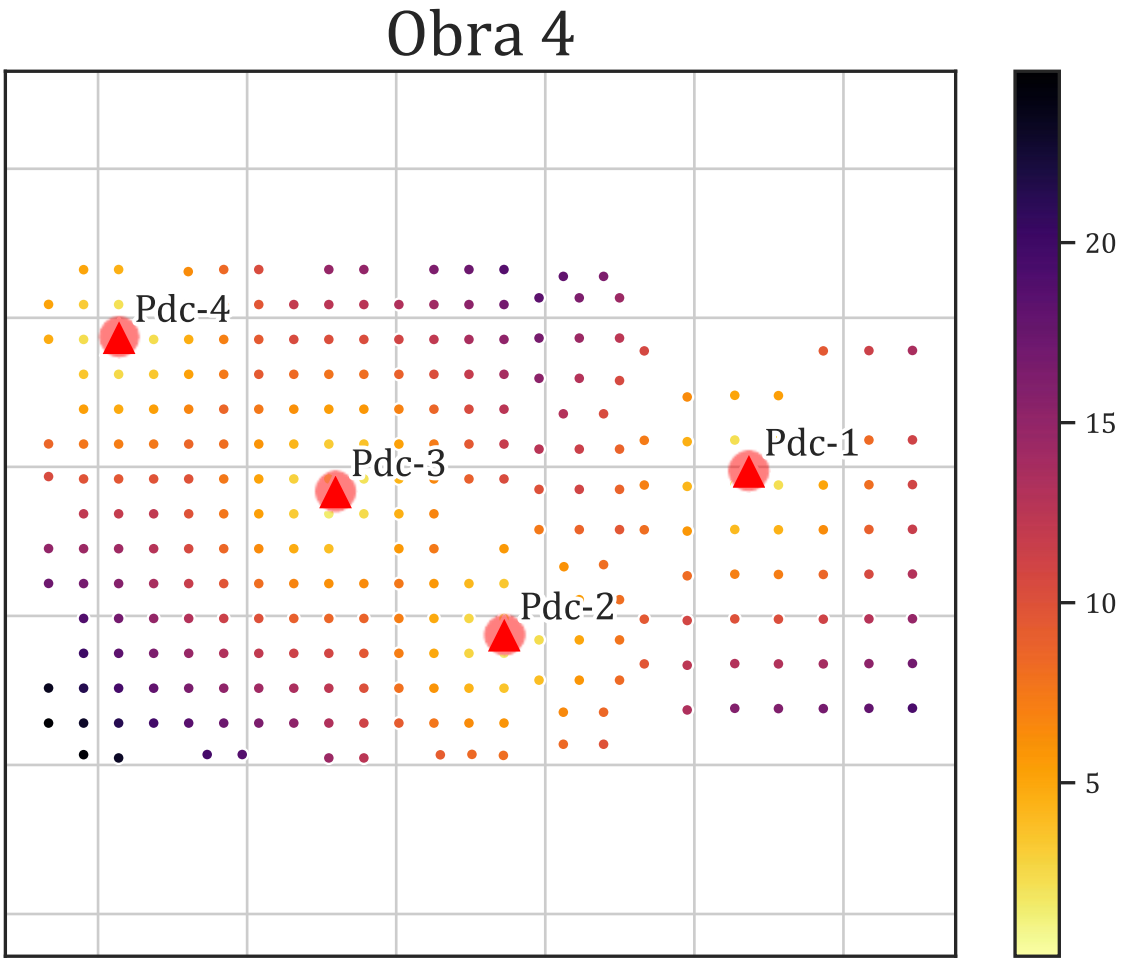


Figura 4.9: Representación de la distancia máxima según el método de variabilidad para la Obra 4

Sin embargo, tal y como se observa en la Figura 4.10, el método continuo tiene en cuenta un área mucho mayor al haber una densidad alta de penetrómetros en un área pequeña. Resulta por tanto importante adaptar el parámetro α a las condiciones del lugar. Puede utilizarse el valor de v_p del método de variabilidad como referencia, tal y como ya sugiere Bunieski al proponer estos métodos. Aún así, la mejor opción siempre será recopilar un número suficiente de muestras como para poder considerar una distancia entre penetrómetro y columna pequeña.

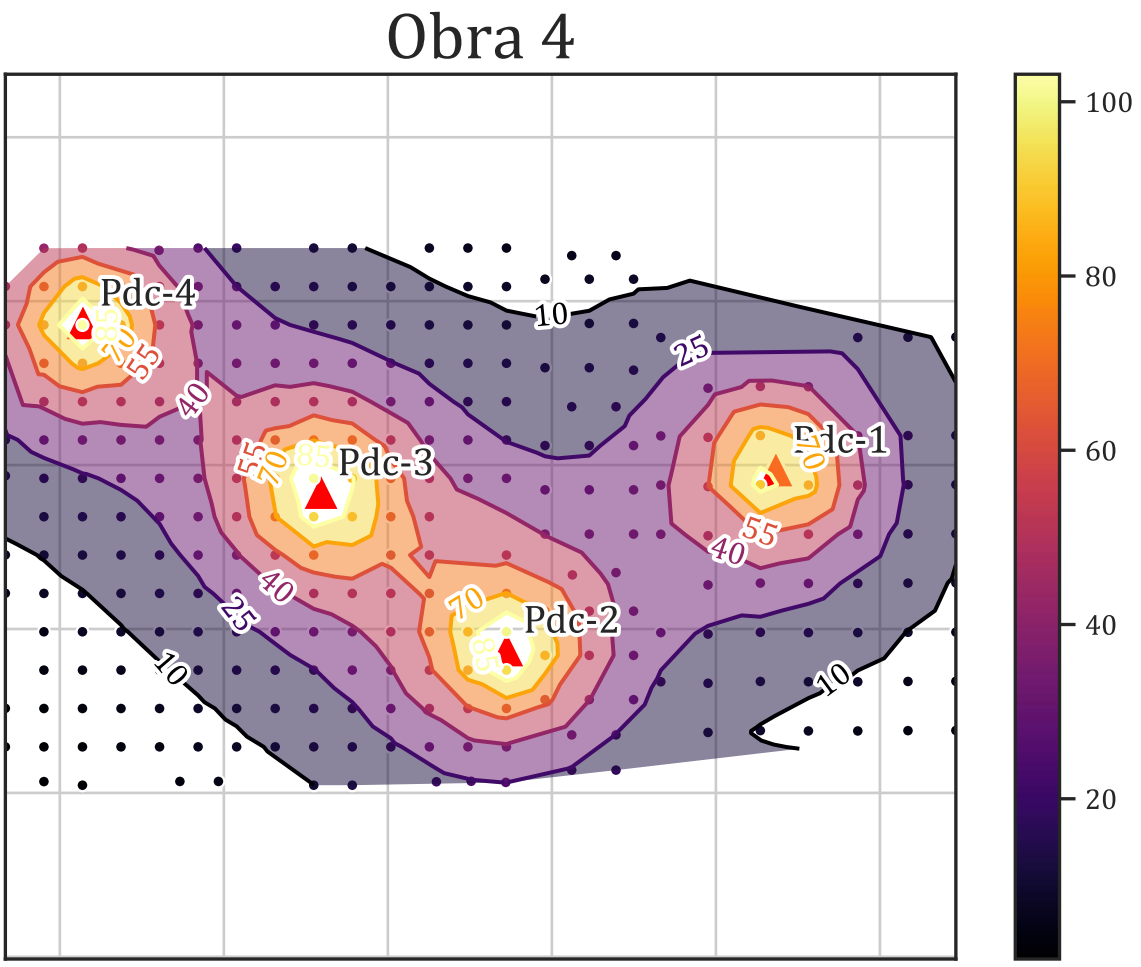


Figura 4.10: Información L a lo largo de la Obra 4 para $\alpha = 0.2$ y $L_{min} = 10$

4.4. Resultados obtenidos con TPOT

TPOT es una librería de AutoML que, mediante algoritmos que simulan la selección natural, intenta aproximar cuál es el mejor *pipeline* posible. Esta librería está pensada para proponer los mejores *pipelines* pero es posible obtener los datos de todos los analizados.

En el proceso de selección, aquellos algoritmos que estén funcionando mejor tenderán a ser escogidos con mayor frecuencia. Podemos comprobar así cuáles son los que suelen dar mejores resultados. En las figuras 4.11 y 4.12 (escala horizontal logarítmica) podemos comprobar cómo los algoritmos de árboles ensamblados *Random Forest* y *ExtraTrees* son los que más se utilizan. Esto ocurre tanto para el enfoque de clasificación como de regresión y se repite para todas las obras analizadas.

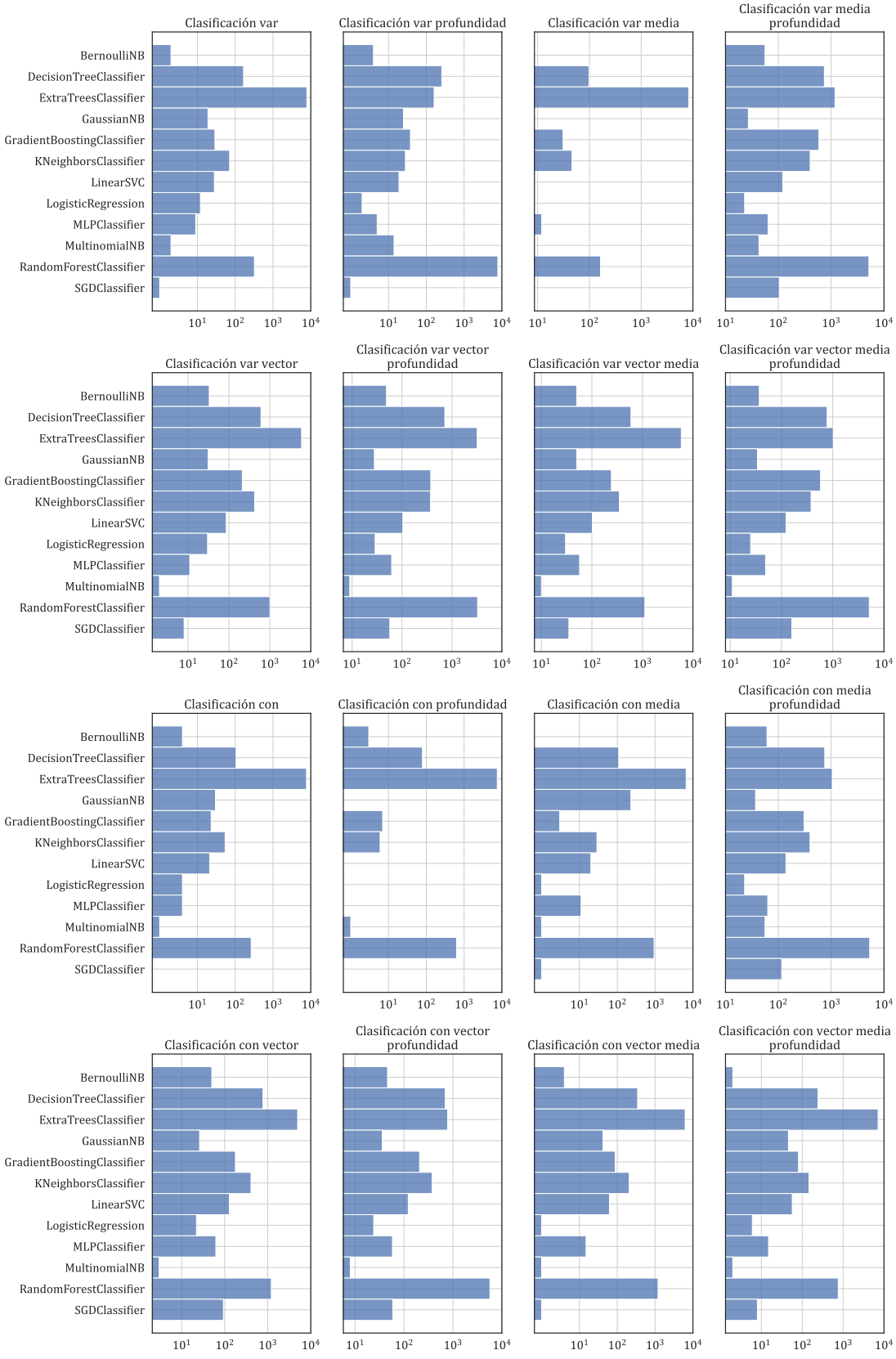


Figura 4.11: Número de veces que se escoge un determinado algoritmo en TPOT, para enfoque de clasificación en la Obra 1

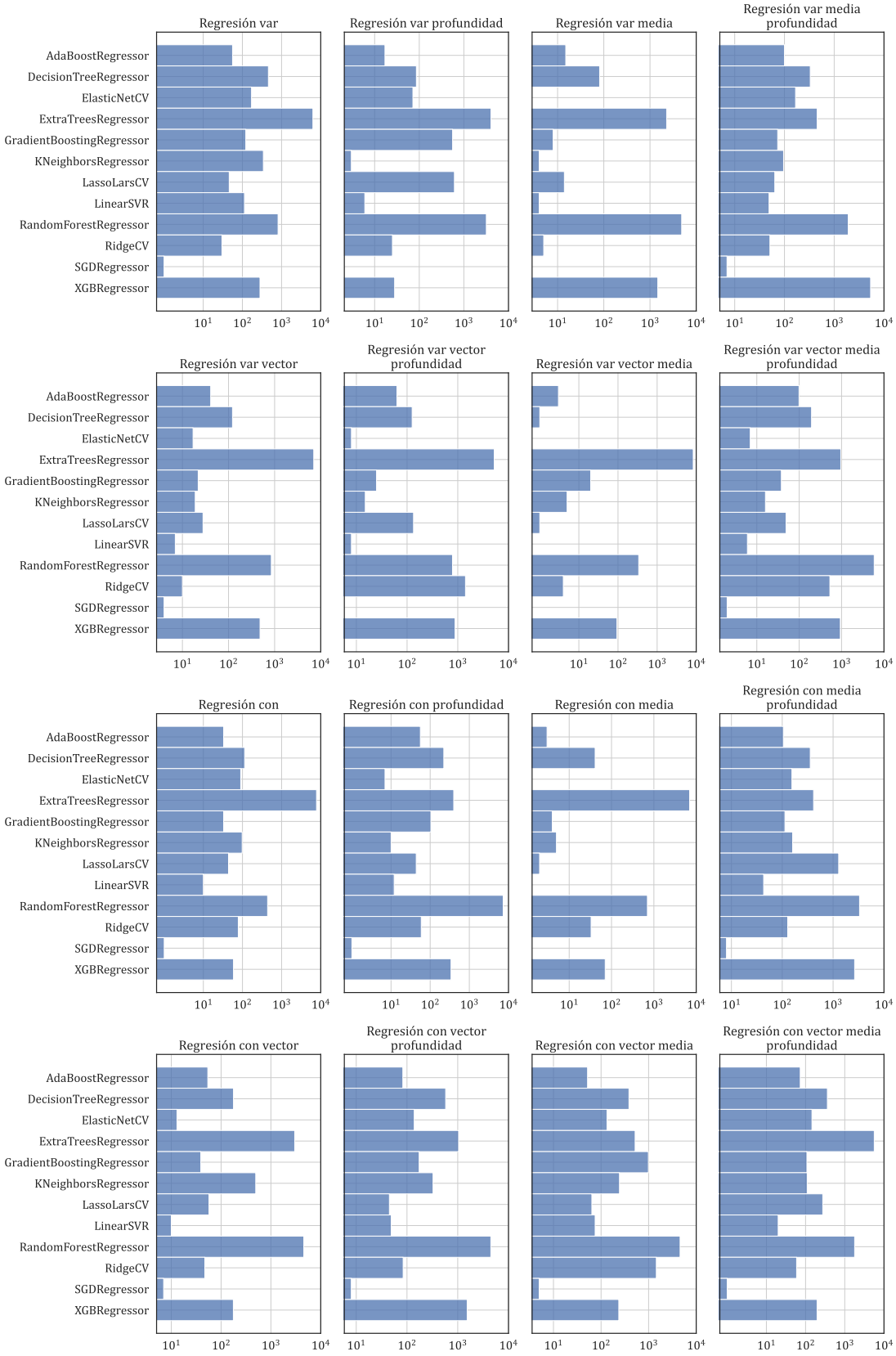


Figura 4.12: Número de veces que se escoge un determinado algoritmo en TPOT, para enfoque de regresión en la Obra 1

A partir del mejor *pipeline* de TPOT en cada caso, se han obtenido una serie de tablas con diversas métricas. Para la obtención de éstas se realiza la media de los resultados de 100 combinaciones distintas de los conjuntos de entrenamiento y evaluación. En la Tabla 4.2: Métricas para los pipelines obtenidos con TPOT de regresión y con el método continuo para la Obra 1 se muestran las métricas para los mejores *pipelines* obtenidos por TPOT con el enfoque de regresión, método continuo y sin sobremuestreo para la Obra 1.

Tabla 4.2: Métricas para los pipelines obtenidos con TPOT de regresión y con el método continuo para la Obra 1

Conjunto	Valor F	Valor F - Entrenamiento	Coef. Correlación	Coef. Correlación - Entrenamiento	MAE	MAE - Entrenamiento	RMSE	RMSE - Entrenamiento	R2	R2 - Entrenamiento	R2 Ajustado	R2 Ajustado - Entrenamiento	Distancia
Regresión con	0.07 ±0.0	0.08 ±0.0	0.76 ±0.01	0.85 ±0.0	2.56 ±0.03	2.07 ±0.01	3.63 ±0.06	2.96 ±0.02	0.57 ±0.01	0.71 ±0.0	0.57 ±0.01	0.71 ±0.0	3.46 ±0.01
Regresión con vector	0.18 ±0.02	1.0 ±0.0	0.9 ±0.01	1.0 ±0.0	1.48 ±0.03	0.0 ±0.0	2.47 ±0.07	0.0 ±0.0	0.8 ±0.01	1.0 ±0.0	0.8 ±0.01	1.0 ±0.0	2.37 ±0.01
Regresión con media	0.18 ±0.01	1.0 ±0.0	0.92 ±0.0	1.0 ±0.0	1.33 ±0.03	0.0 ±0.0	2.17 ±0.07	0.0 ±0.0	0.85 ±0.01	1.0 ±0.0	0.85 ±0.01	1.0 ±0.0	2.35 ±0.01
Regresión con media vector	0.23 ±0.02	1.0 ±0.0	0.95 ±0.0	1.0 ±0.0	1.07 ±0.02	0.0 ±0.0	1.77 ±0.06	0.0 ±0.0	0.9 ±0.01	1.0 ±0.0	0.9 ±0.01	1.0 ±0.0	2.3 ±0.01
Regresión con profundidad	0.19 ±0.01	0.34 ±0.01	0.94 ±0.0	0.99 ±0.0	1.15 ±0.02	0.47 ±0.01	1.95 ±0.06	0.95 ±0.01	0.88 ±0.01	0.97 ±0.0	0.88 ±0.01	0.97 ±0.0	2.49 ±0.01
Regresión con profundidad vector	0.27 ±0.02	0.95 ±0.03	0.96 ±0.0	1.0 ±0.0	0.85 ±0.02	0.0 ±0.0	1.47 ±0.05	0.02 ±0.0	0.93 ±0.0	1.0 ±0.0	0.93 ±0.0	1.0 ±0.0	2.22 ±0.01
Regresión con profundidad media	0.29 ±0.02	0.78 ±0.04	0.97 ±0.0	1.0 ±0.0	0.75 ±0.02	0.09 ±0.01	1.35 ±0.05	0.3 ±0.02	0.94 ±0.0	1.0 ±0.0	0.94 ±0.0	1.0 ±0.0	2.27 ±0.01
Regresión con profundidad media vector	0.36 ±0.03	1.0 ±0.0	0.98 ±0.0	1.0 ±0.0	0.61 ±0.02	0.0 ±0.0	1.15 ±0.06	0.0 ±0.0	0.96 ±0.0	1.0 ±0.0	0.96 ±0.0	1.0 ±0.0	2.18 ±0.01

Las conclusiones de estas tablas en las diferentes obras son similares, *ceteris paribus*:

- Realizar sobremuestreo es, en el conjunto de evaluación, peor que no realizarlo. Se obtienen peores correlaciones y mayores errores.
- No hay grandes diferencias entre utilizar el método continuo y utilizar la variabilidad.
- El enfoque de regresión funciona algo mejor.

En general, elegir entre estas opciones no genera resultados muy diferentes. Sin embargo, añadir nuevas características (vector de correlación, media y profundidad) sí conlleva aumentos importantes del desempeño de los modelos. En la Figura 4.13 se muestra en un diagrama de Taylor el progreso que supone el añadir las diferentes características para el enfoque de regresión, método continuo y sin realizar sobremuestreo.

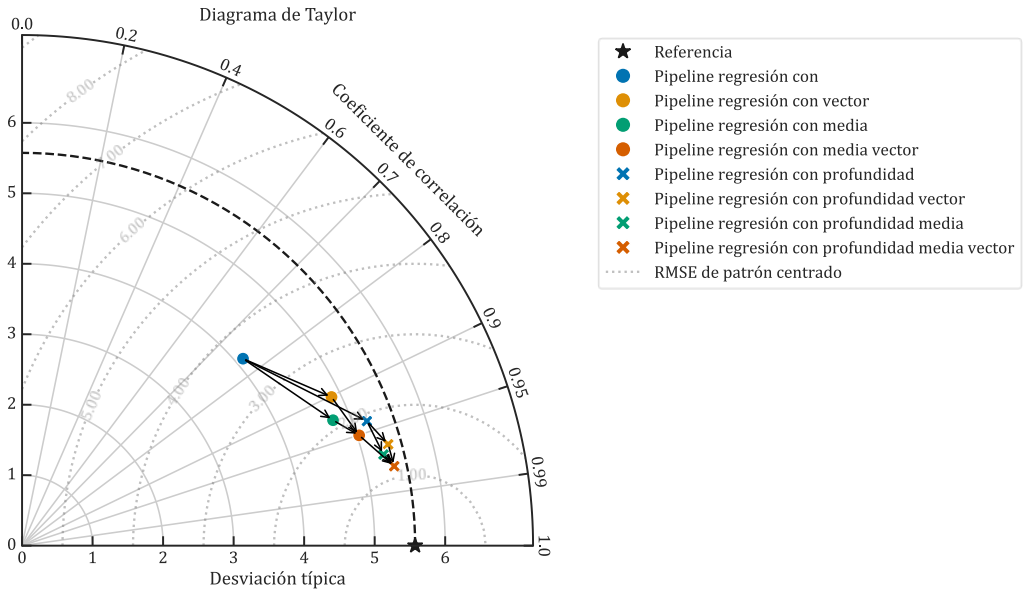


Figura 4.13: Evolución de las métricas del *pipeline* de TPOT con regresión, método continuo y sobremuestreo en función de las características consideradas para la Obra 1

Se observa cómo se produce una mejora importante con coeficientes de correlación que van desde 0.76 hasta 0.97.

Lo que estos valores implican puede apreciarse muy bien si graficamos la distribución de los golpeos con la profundidad. En la Figura 4.14 se muestra la distribución a lo largo de la profundidad con los golpeos reales (izquierda) y predichos (derecha); los golpeos se predicen únicamente teniendo en cuenta las características originales. Se observa cómo la predicción es bastante rígida, no consiguiendo reproducir del todo la distribución real de los golpeos. La Figura 4.15 es similar a la anterior pero la predicción tiene en cuenta, además de las características originales, el vector de correlación y la media; en este caso observamos cómo la predicción es mucho más ajustada a la distribución real de los golpeos. Este hecho se reproduce también en todas las obras analizadas.

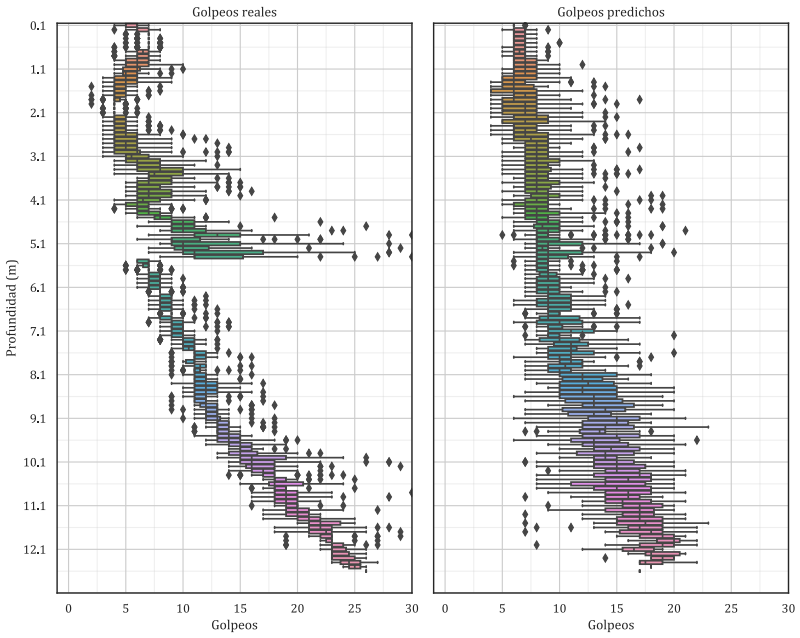


Figura 4.14: Representación mediante diagrama de caja de la distribución real de los golpeos de los penetrómetros de la Obra 1 junto con la distribución de los golpeos predichos para regresión, método continuo y sin sobremuestreo

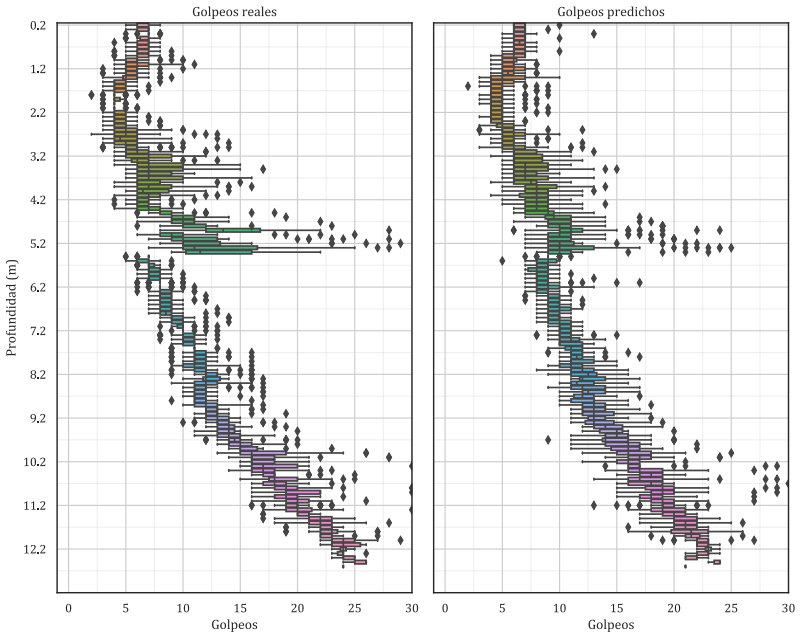


Figura 4.15: Representación mediante diagrama de caja de la distribución real de los golpeos de los penetrómetros de la Obra 1 junto con la distribución de los golpeos predichos para regresión, método continuo y sin sobremuestreo. Se añaden como características el vector y la media

En cuanto a la distribución del error, este no se produce de forma homogénea para los diferentes valores de golpeo, que sería lo deseable. Los algoritmos tienden a sobrestimar los valores bajos y a subestimar los valores altos (figuras 4.16 y 4.17).

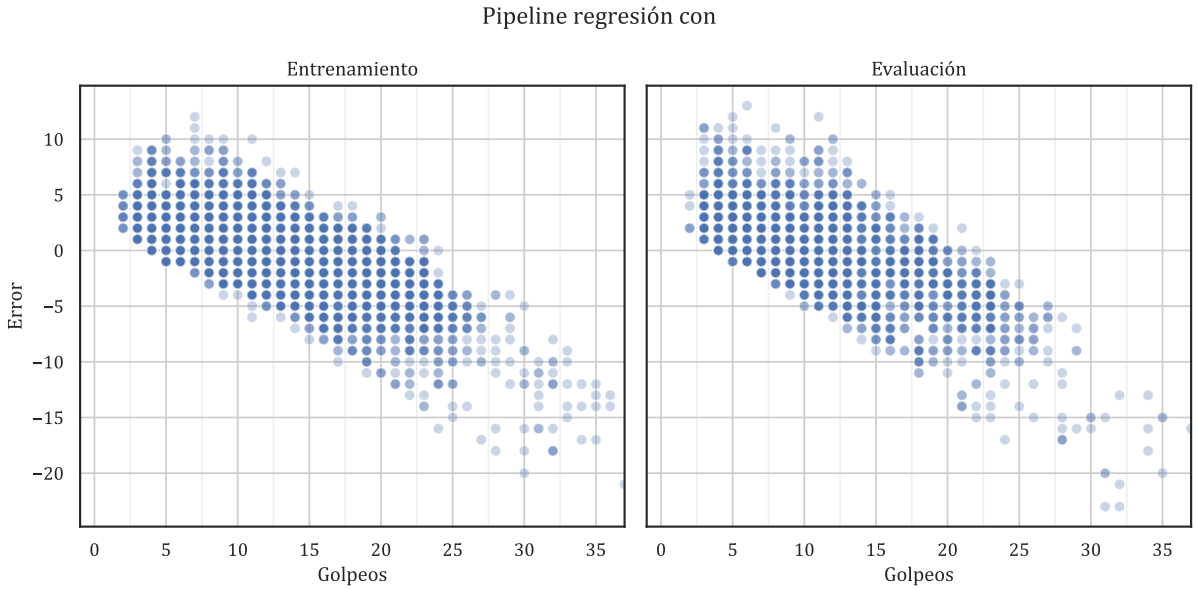


Figura 4.16: Error residual para regresión y método continuo con las características originales y para la Obra 1

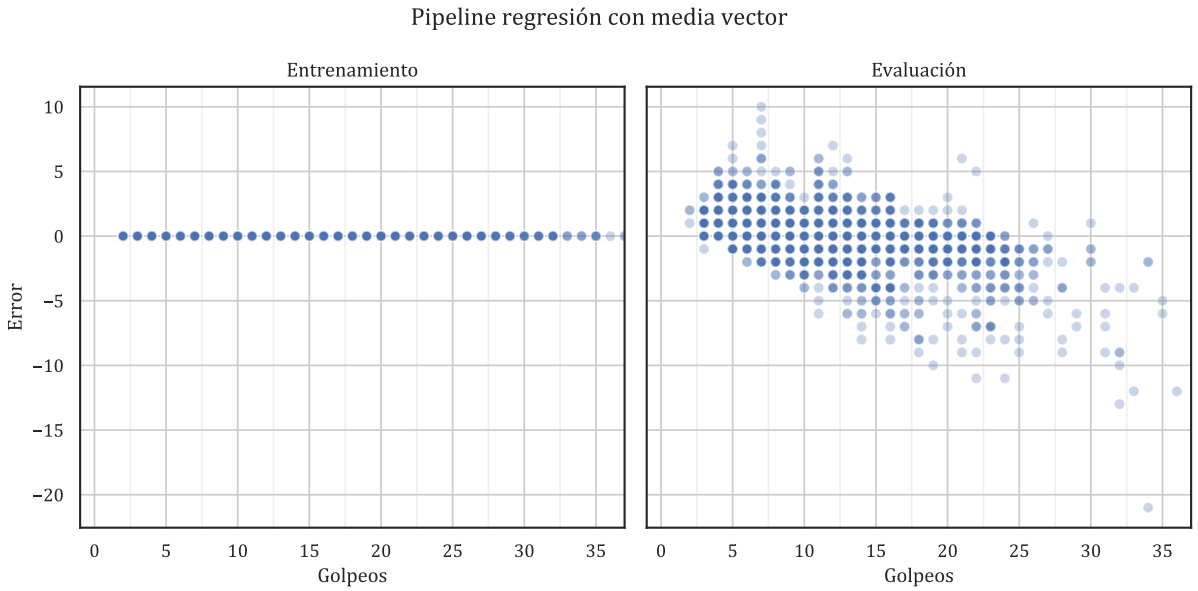


Figura 4.17: Error residual para regresión y método continuo con media y vector de correlación para la Obra 1

Este efecto se atenúa, en parte, al aplicar sobremuestreo. Sin embargo, los valores altos siguen siendo subestimados (Figura 4.18).

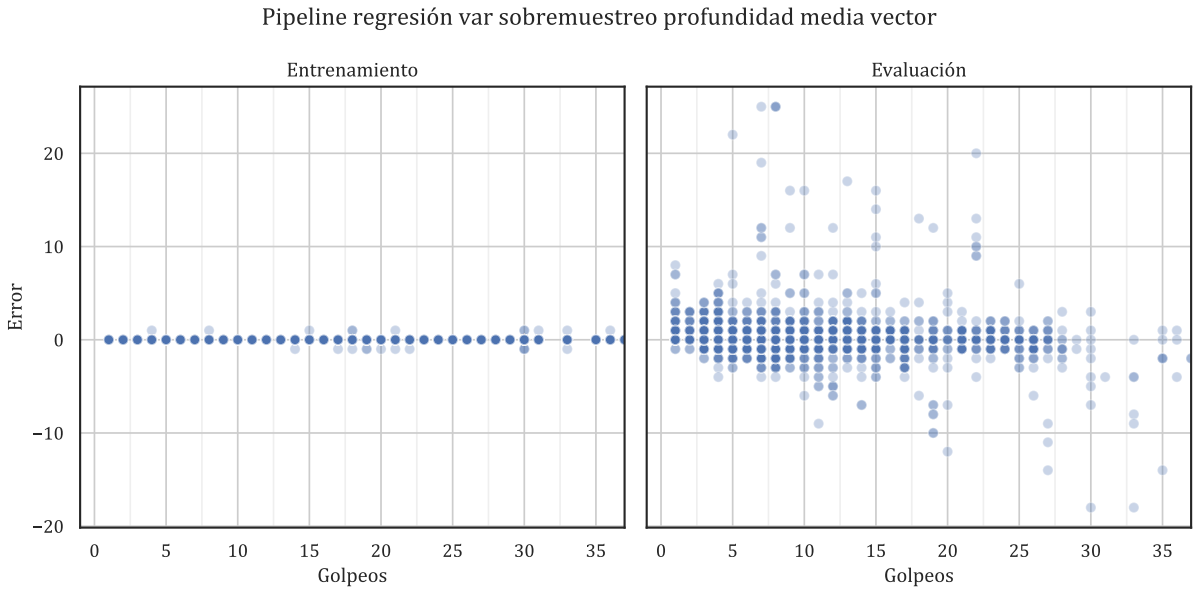


Figura 4.18: Error residual para regresión y método continuo con profundidad, media, vector de correlación y sobremuestreo

No obstante, ya hemos comprobado que el sobremuestreo, en términos generales, ha producido peores resultados.

La subestimación de los valores altos de golpeo puede tener una explicación física. En este caso, dado el procedimiento de ejecución por desplazamiento del terreno y la potencia de las perforadoras utilizadas, no es posible atravesar paquetes de espesores relevantes con golpeos superiores a 20. Es posible que los valores superiores en los datos no sean realmente representativos de la resistencia, o que estén incluidos debido a la distancia entre el penetrómetro y la columna.

También observamos en las Figura 4.17 y Figura 4.18 que se produce un sobreajuste a los datos de entrenamiento (estos no presentan apenas error).

Si representamos el error a lo largo de la profundidad, observamos también la rigidez del modelo que sólo considera las características originales (Figura 4.19), frente a los modelos que consideran las características adicionales (Figura 4.20). En estos últimos la magnitud del error está bastante contenida.

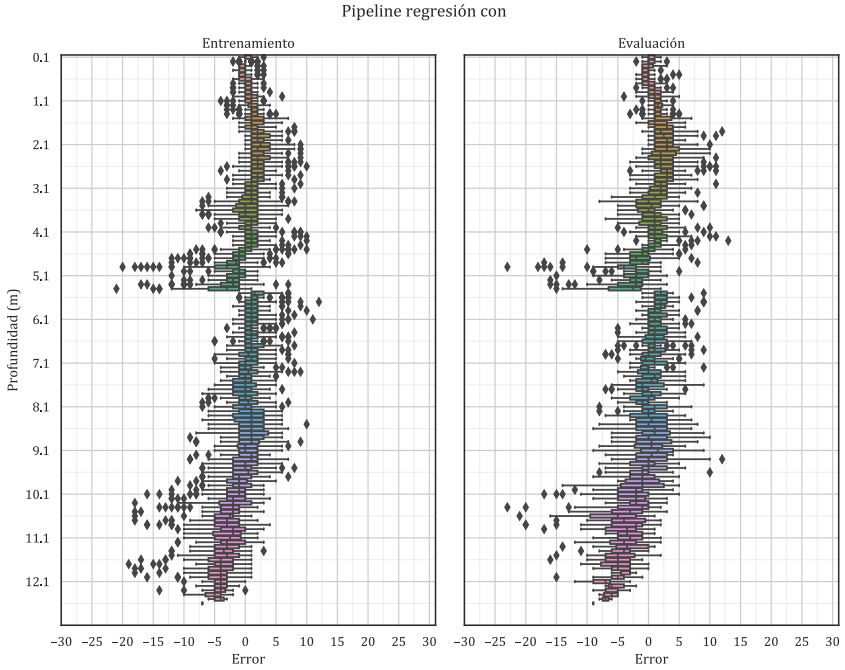


Figura 4.19: Error residual en función de la profundidad para regresión y método continuo para la Obra 1

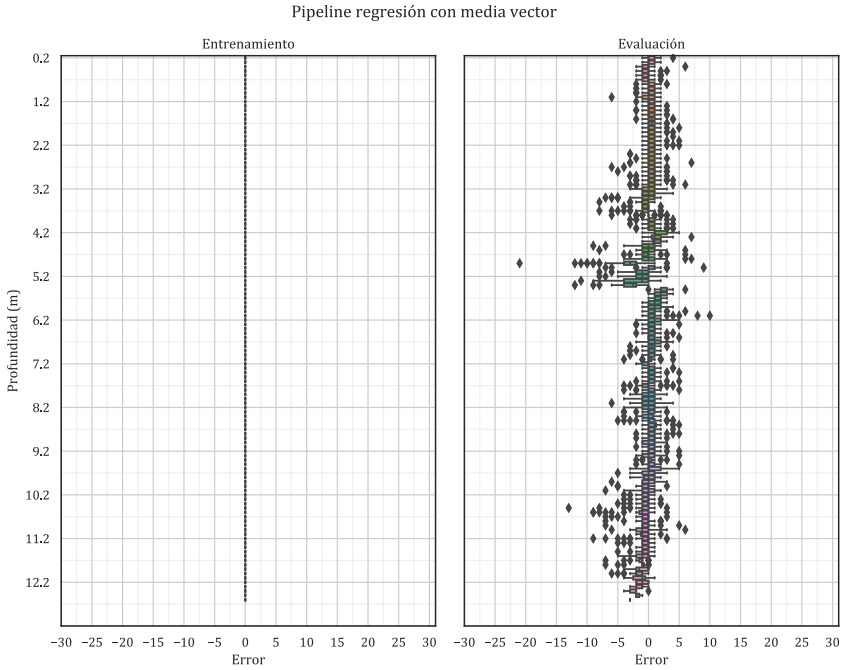


Figura 4.20: Error residual en función de la profundidad para regresión y método continuo con media y vector de correlación para la Obra 1

4.4.1. Explicabilidad de los resultados

Para analizar cómo afectan las diferentes características al resultado final se utilizan los gráficos ICE y valores Shapley. Lo que más destaca de estos gráficos es la excesiva importancia que dan los algoritmos a la profundidad cuando esta se incluye como característica (figuras 4.21 y 4.22).

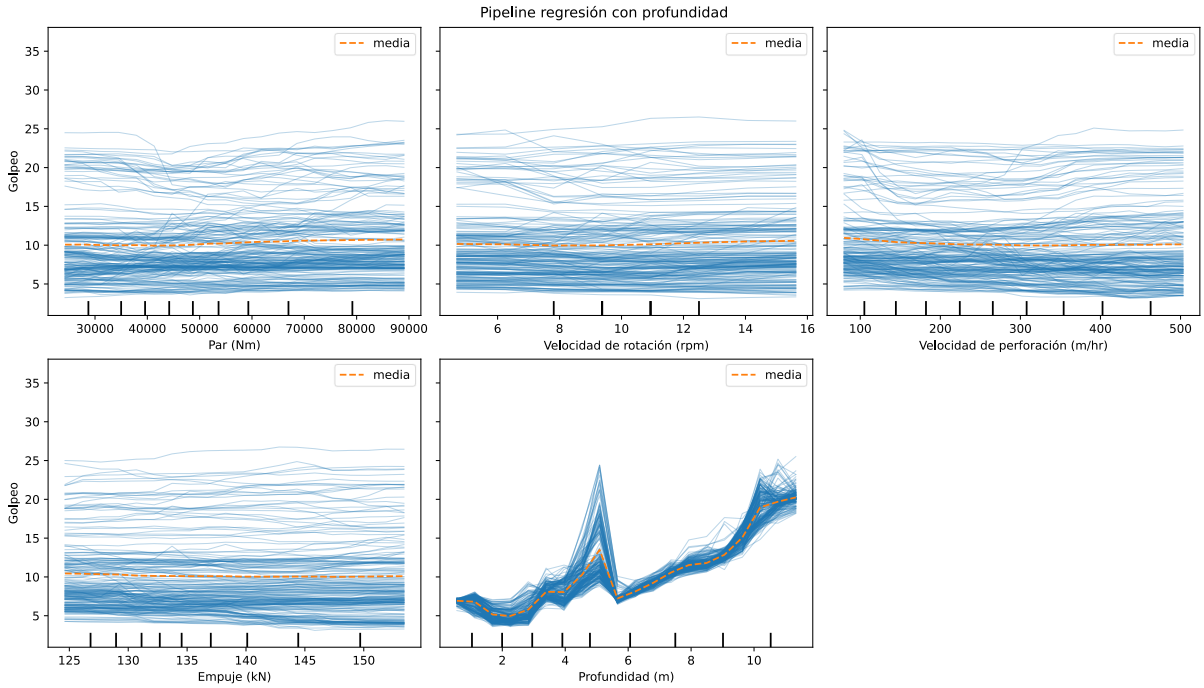


Figura 4.21: gráfico ICE para regresión y método continuo con profundidad para la Obra 1

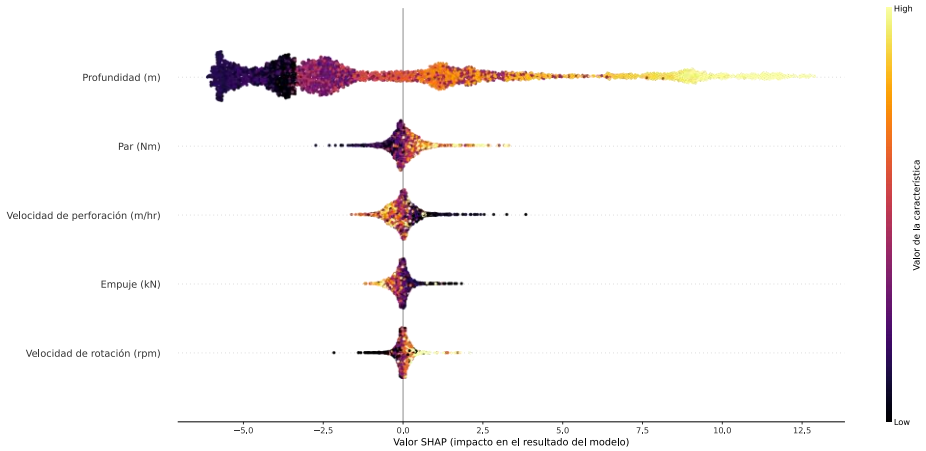


Figura 4.22: diagrama de enjambre para regresión y método continuo con profundidad para la Obra 1

En ambos casos se observa como el modelo se adapta mucho a la profundidad y el resto de las características apenas afectan. Este hecho puede dar lugar a unas métricas muy «buenas», pero podrían impedir la detección de anomalías al adaptar la predicción a la profundidad y no a los valores necesarios de par o empuje. Esto se produce porque los penetrómetros en las diferentes obras son muy parecidos entre sí y, por tanto, los golpeos están muy correlacionados con la profundidad.

En cuanto a los modelos que no tienen en cuenta la profundidad, en los gráficos ICE se observa en general que los golpeos aumentan con el par y la velocidad de rotación, disminuyen con la velocidad de perforación y se ven poco afectados por los cambios en el empuje. Cuando se añaden otras características resulta más complicado evaluar el efecto individual de cada una, lo que quiere decir que la variación de una sola de estas no produce grandes cambios en la predicción; es el conjunto de variaciones de diferentes características lo que hace variar la predicción. En todo caso, puede apreciarse efecto del par medio (Figura 4.23).

En cuanto a la importancia de las características que obtenemos a partir de los valores Shapley, no es la misma para las diversas obras, si bien las características más importantes suelen ser el par, el par medio, la velocidad de penetración y la velocidad de penetración media (figuras 4.24 y 4.25).

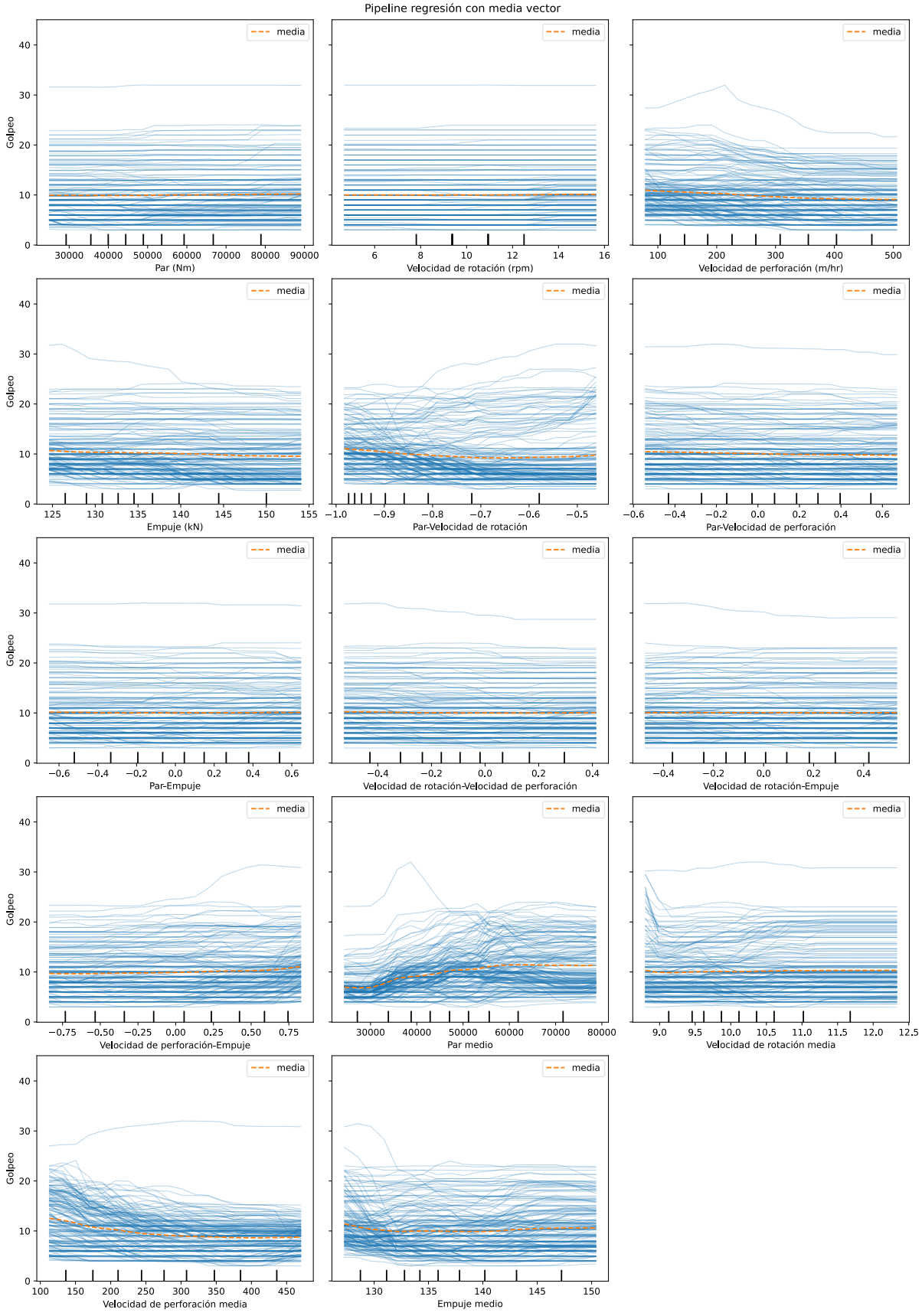


Figura 4.23: gráfico ICE para regresión y método continuo con media y vector de correlación para la Obra 1

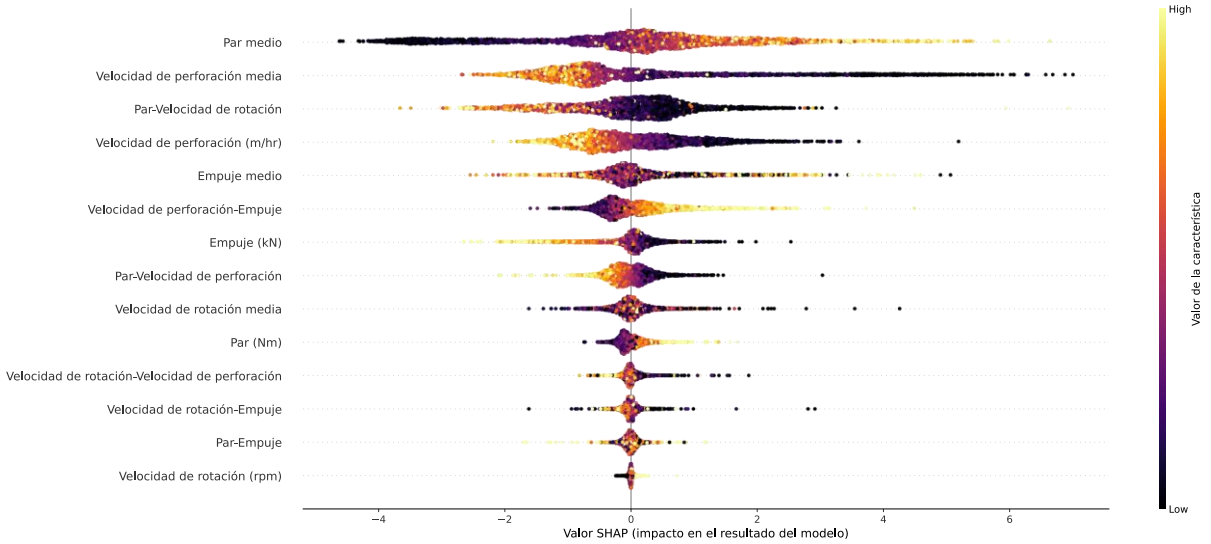


Figura 4.24: diagrama de enjambre para regresión y método continuo con media y vector de correlación para la Obra 1

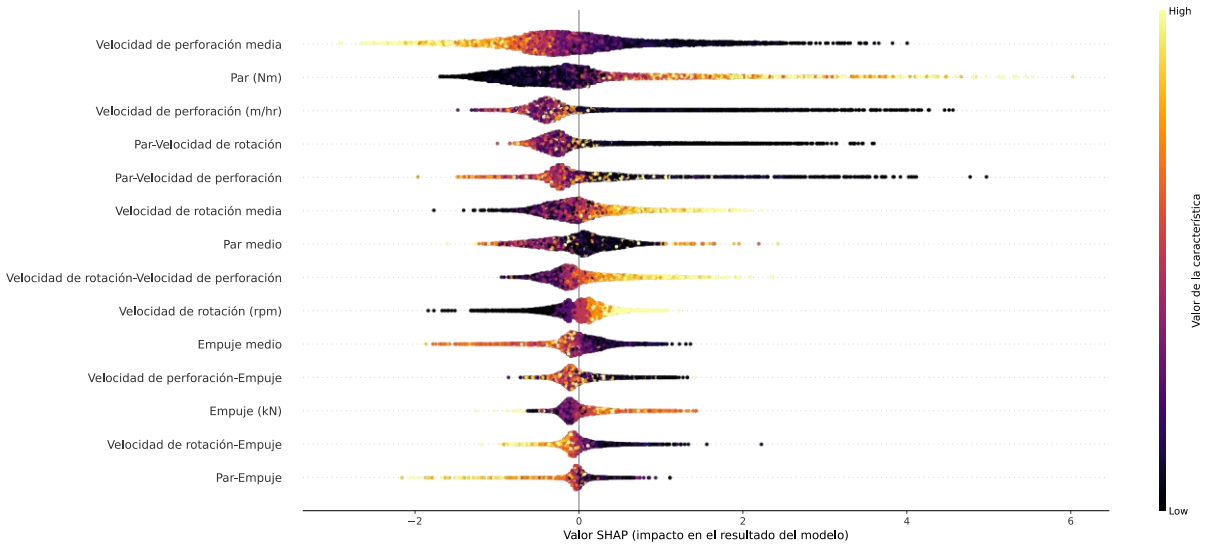


Figura 4.25: diagrama de enjambre para regresión y método continuo con media y vector de correlación para la Obra 3

4.4.2. Efecto de la distancia

Como se ha indicado previamente, a mayor distancia entre la inclusión y el penetrómetro, mayores errores podemos estar introduciendo en el conjunto. Ante esto se han utilizado los métodos de variabilidad y continuo. No obstante, resulta

interesante evaluar cuál es el efecto que supone el aumento del error con la distancia para las diferentes obras y comprobar si existe una distancia óptima.

Para ello se calculan las métricas obtenidas con conjuntos de datos formados teniendo en cuenta una determinada distancia entre el penetrómetro y la columna. Para comprobar si existe algún cambio de tendencia nos valemos de un gráfico de suma acumulada (gráfico CUSUM). En este tipo de gráficos se muestra la suma acumulada de la diferencia con respecto a algún valor de referencia, en este caso, la media.

En la Figura 4.26 se muestra, desde la gráfica superior hacia abajo:

1. Evolución del coeficiente de correlación, en el conjunto de evaluación, con la distancia.
2. Evolución del valor de CUSUM con la distancia.
3. Evolución del número de muestras con la distancia.

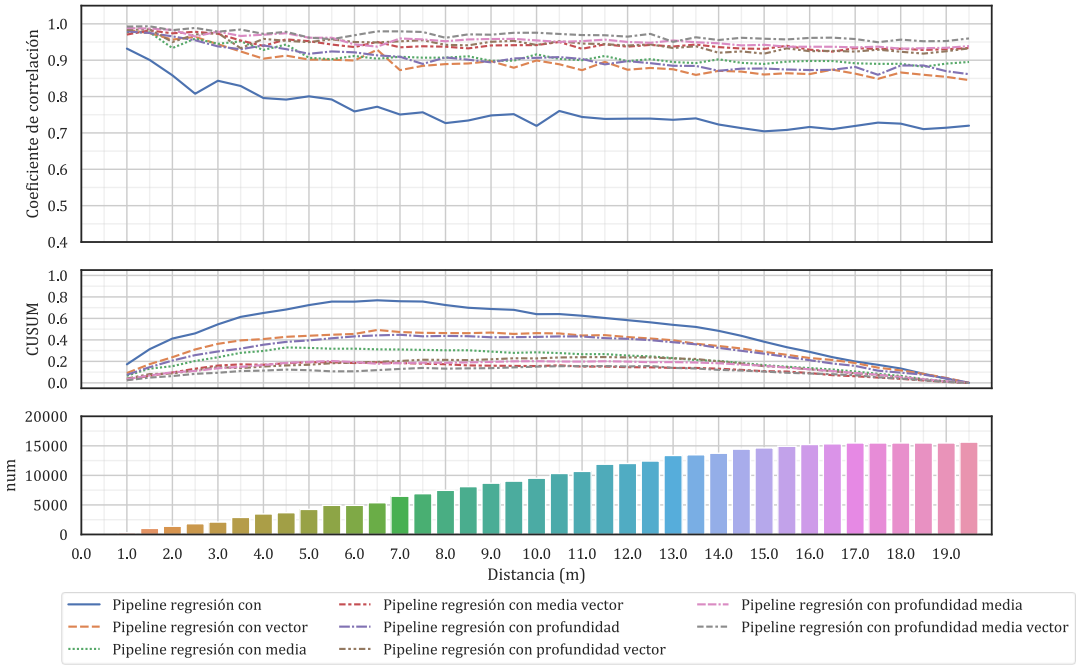


Figura 4.26. Evolución del coeficiente de correlación, CUSUM y número de muestras con la distancia para los algoritmos de regresión, método continuo y sin sobremuestreo para la Obra 1

Lo que se observa es que, prácticamente en el momento en que nos alejamos del penetrómetro, el error empieza a aumentar. En algunos de los conjuntos analizados

se observa que el mayor valor del coeficiente de correlación no se produce a la distancia mínima, sino a una distancia algo mayor (Figura 4.27). Esto puede deberse a que, en distancias muy bajas, el número de muestras es muy pequeño e insuficiente para entrenar adecuadamente los modelos.

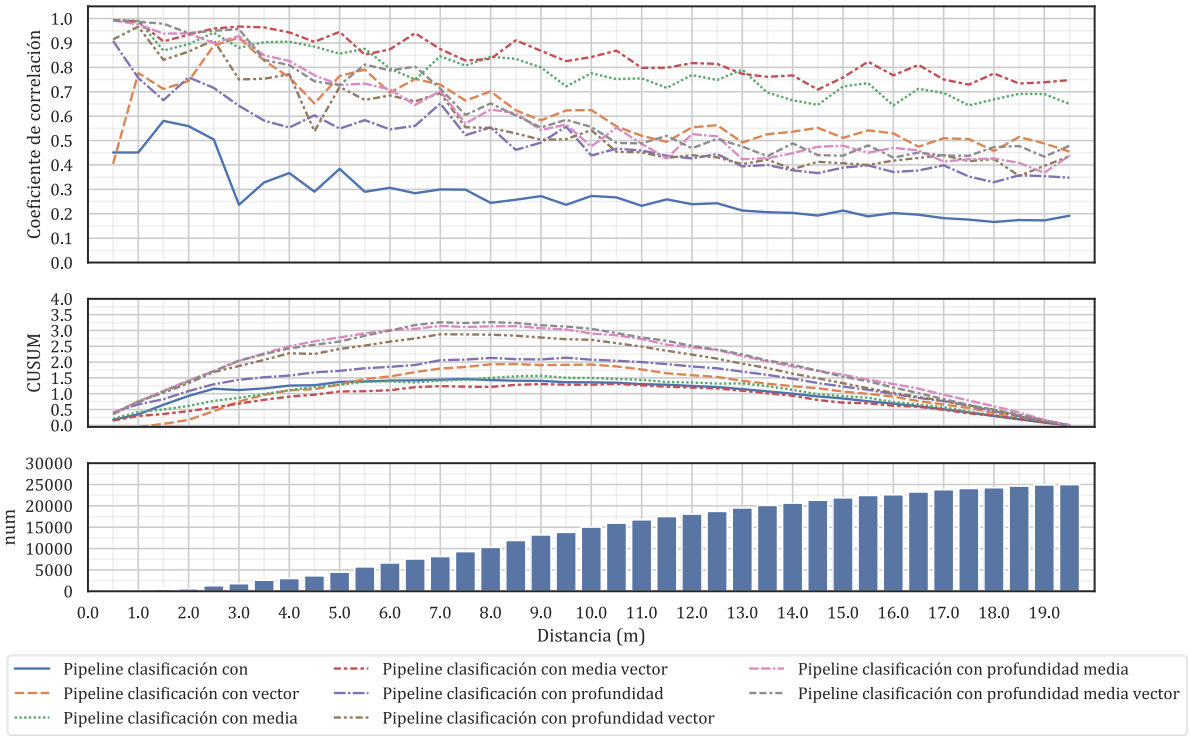


Figura 4.27. Evolución del coeficiente de correlación, CUSUM y número de muestras con la distancia para los algoritmos de clasificación, método continuo y sin sobremuestreo para la Obra 4

4.4.3. Efecto del tamaño del conjunto de datos

Para evaluar el efecto del conjunto de datos se crean conjuntos de diferentes tamaños construidos de modo aleatorio a partir de los conjuntos continuo o de variabilidad. A diferencia de con la distancia, en el caso del tamaño del conjunto sí que se aprecia un punto de inflexión. Al aumentar el tamaño del conjunto, el coeficiente de correlación aumenta rápidamente para después seguir creciendo de manera paulatina pero ininterrumpida (Figura 4.28). En principio, cuanto más grande sea el conjunto de datos para entrenar los modelos, mejor. Sin embargo, un conjunto de entrenamiento muy grande también aumenta el tiempo de cálculo, por lo que puede ser razonable limitar su tamaño si sabemos que van a obtenerse

buenos resultados. Resulta, no obstante, complicado establecer un valor mínimo en el número de muestras ya que, para las obras analizadas, conseguir un coeficiente de correlación de 0.8 (en los algoritmos sin profundidad) ha variado entre unos pocos y varias decenas de miles de muestras.

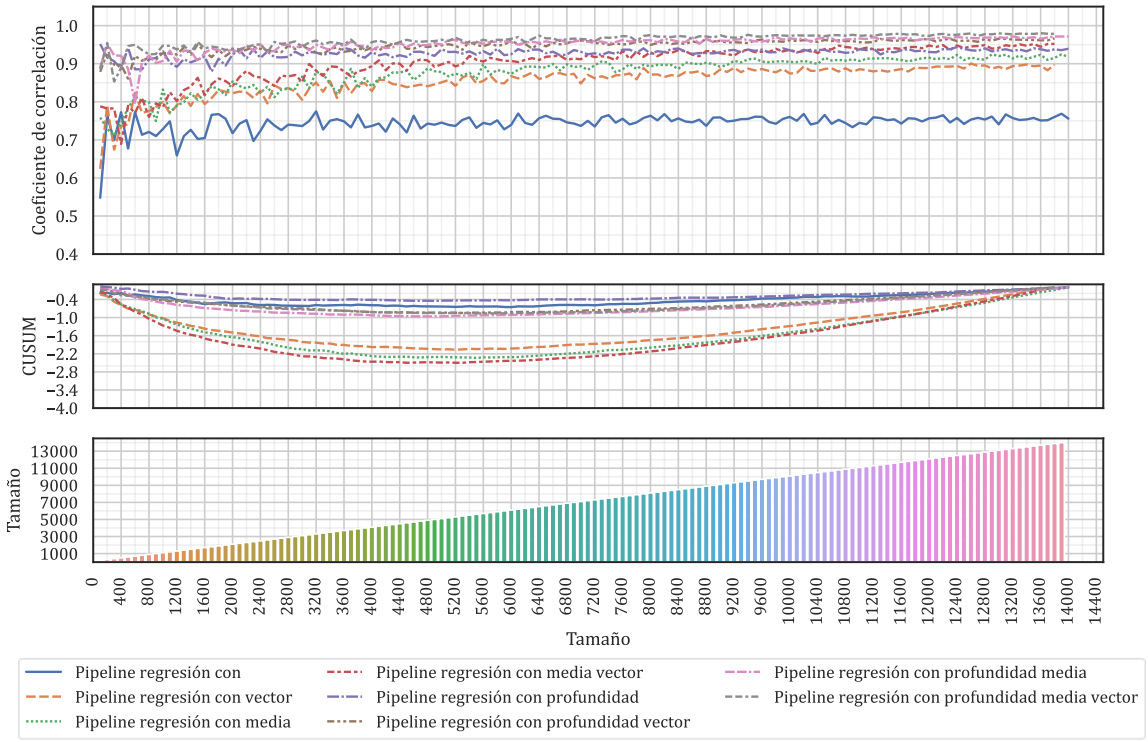


Figura 4.28: Evolución del coeficiente de correlación y CUSUM con el tamaño de la muestra para los algoritmos de regresión, método continuo y sin sobremuestreo

4.5. Resultados con *pipeline* simplificada

Es posible que los *pipelines* de TPOT, en su configuración por defecto, sean muy complejos. Por ejemplo, el *pipeline* sugerido por TPOT para la Obra 1, con el enfoque de regresión, método continuo y teniendo en cuenta la media y el vector de correlación es el siguiente:

1. Duplicación del conjunto de datos original.
2. Escalado con MaxAbsScaler.
3. Combinación del conjunto de datos original y el escalado.
4. Escalado adicional con MinMaxScaler.

5. Combinación con el conjunto de datos original.
6. Aplicación de ExtraTreesRegressor.
7. Escalado de la salida del ExtraTreesRegressor.
8. Combinación del conjunto de datos original con la salida del modelo escalada.
9. Estandarización de todo el conjunto de datos combinado.
10. Aplicación final del modelo RandomForestRegressor.

En este punto puede surgir la pregunta ¿Realmente necesitamos *pipelines* tan largos? ¿Qué mejora suponen? En principio son preferibles *pipelines* lo más cortos posibles, ya que son más fáciles de entender y rápidos de entrenar.

Por ello se ha acudido a uno de los algoritmos más utilizados en los *pipelines* de TPOT, *ExtraTreesRegressor* (ETR), sin usar preprocesos previos. Concretamente se ha empleado ETR con los parámetros por defecto de *scikit-learn*, que son los siguientes:

- n_estimators=100,
- criterion='squared_error',
- max_depth=None,
- min_samples_split=2,
- min_samples_leaf=1,
- min_weight_fraction_leaf=0.0,
- max_features=1.0,
- max_leaf_nodes=None,
- min_impurity_decrease=0.0,
- bootstrap=False, oob_score=False,
- n_jobs=None,
- random_state=None,
- verbose=0,
- warm_start=False,
- ccp_alpha=0.0,

- max_samples=None,
- monotonic_cst=None

Los resultados que se obtienen con este único algoritmo son muy similares a los del *pipeline* completo, si bien algo inferiores (Figura 4.29). Esto se repite en todas las obras.

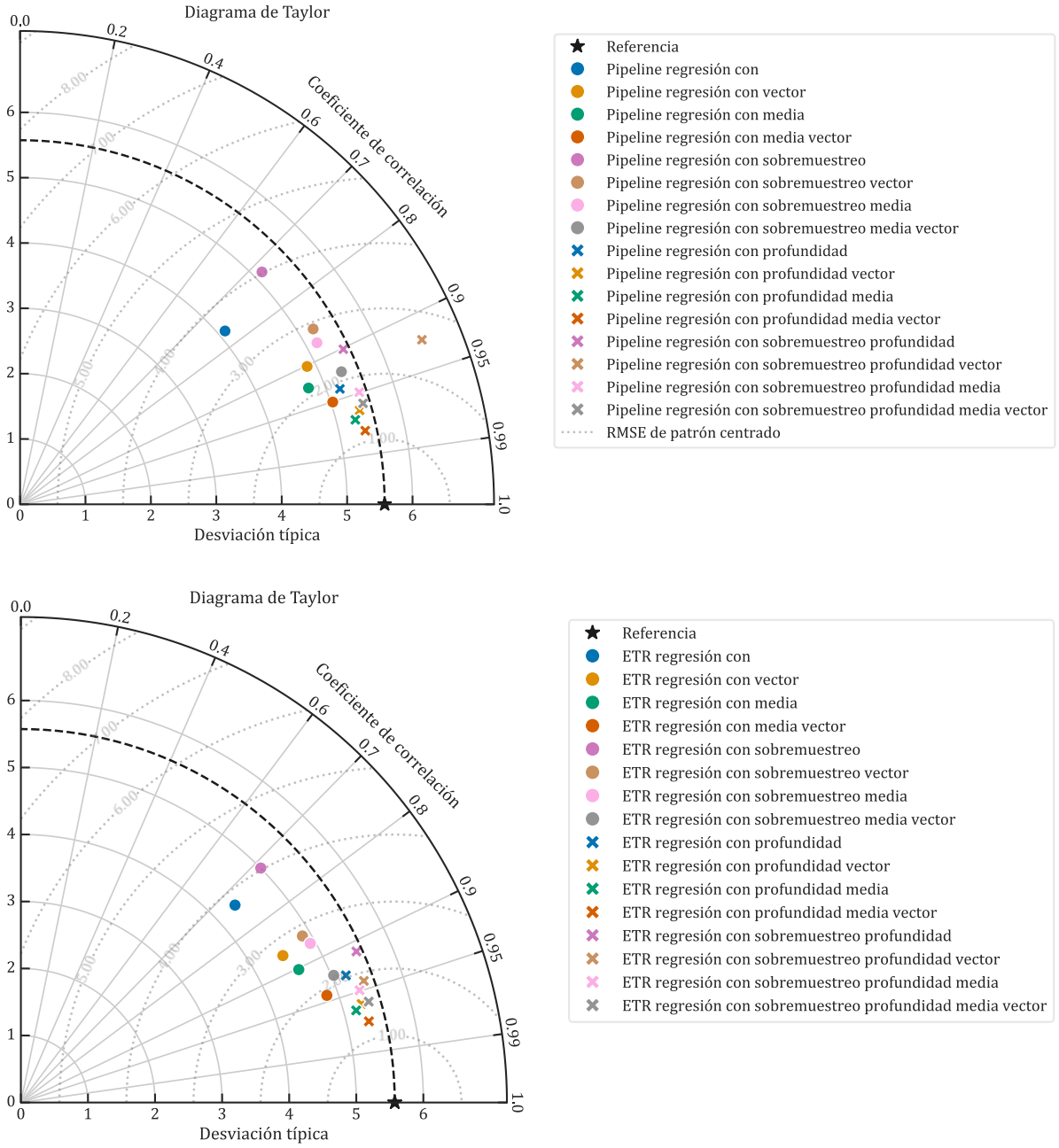


Figura 4.29: Diagramas de Taylor para los conjuntos con método continuo y considerando regresión para la Obra 1. Arriba, *pipelines* de TPOT. Abajo ETR

4.6. Comparación con los parámetros compuestos

Es evidente que, si se va a utilizar aprendizaje automático, este ha de proporcionar una información mucho mejor que los parámetros compuestos, ya que su aplicación es más compleja.

Para comparar la capacidad de aproximación del perfil penetrométrico del suelo de las fórmulas analíticas y de los algoritmos de aprendizaje automático, se utiliza el siguiente proceso:

1. Se predicen los valores de golpeo mediante los algoritmos.
2. Se calculan los valores resultantes de los parámetros compuestos.
3. Se escalan los valores obtenidos, así como el golpeo esperado, entre 0 y 1. Esto es necesario porque los parámetros compuestos no están en la misma escala que los golpeos.
4. Se obtienen las siguientes métricas: R, MAE, RMSE y d.
5. Se grafican los perfiles medios.

Se ha tenido en cuenta el conjunto de datos obtenido mediante el método continuo. Aun teniendo en cuenta un conjunto de datos que, en principio, ya no tiene anomalías, se detectan unos pocos puntos con valores especialmente altos de los parámetros compuestos. En este caso basta con eliminar el 1% más alto de los parámetros para tener una distribución lógica.

En la Tabla 4.3 pueden comprobarse los valores obtenidos de coeficiente de correlación, MAE, RMSE y distancia para los diferentes parámetros compuestos, y para ETR con método continuo media y vector de correlación.

Tabla 4.3: Resultados obtenidos en la comparativa entre los parámetros compuestos, y ETR con método continuo, media y vector de correlación

Obra	Métrica	Resistencia a la penetración	Índice de Somerton	Energía específica de perforación	Energía específica	ETR
Obra 1	R	0.45	0.37	0.44	0.59	0.94
	MAE	0.13	0.13	0.12	0.14	0.05
	RMSE	0.18	0.17	0.15	0.18	0.06
	d	0.17	0.17	0.18	0.16	0.02
Obra 2	R	0.03	-0.03	0.11	0.10	0.86
	MAE	0.10	0.08	0.19	0.10	0.02
	RMSE	0.16	0.13	0.23	0.16	0.03
	d	0.22	0.23	0.22	0.20	0.03
Obra 3	R	0.40	0.38	0.37	0.44	0.90
	MAE	0.13	0.15	0.21	0.11	0.08
	RMSE	0.18	0.19	0.24	0.16	0.10
	d	0.17	0.17	0.18	0.17	0.02
Obra 4	R	0.24	0.24	0.33	0.30	0.84
	MAE	0.16	0.21	0.35	0.11	0.08
	RMSE	0.21	0.38	0.38	0.12	0.09
	d	0.18	0.19	0.21	0.18	0.03
Obra 5	R	0.35	0.38	0.21	0.34	0.85
	MAE	0.09	0.13	0.28	0.09	0.04
	RMSE	0.15	0.16	0.31	0.13	0.07
	d	0.19	0.19	0.23	0.23	0.06

La diferencia entre los resultados es muy importante. Los valores del coeficiente de correlación para ETR se encuentra entre 0.84 y 0.94 cuando los parámetros compuestos no alcanzan, en el mejor de los casos, 0.6. En la Obra 2, ni siquiera existe una relación lineal entre los valores de golpeo y los parámetros compuestos.

4.7. Predicción de datos ajenos al entrenamiento

Idealmente un modelo entrenado debería ser capaz de predecir valores de manera precisa en todo tipo de obras ¿es posible entrenar un modelo en una obra para obtener valores en otra? La respuesta rápida es que no.

Para comprobar esto se han tenido en cuenta dos situaciones:

- Se entrena el algoritmo ExtraTreesRegressor con datos de la Obra 1 y se intenta predecir los valores de las obras 2 y 3.

- Se entrena el algoritmo ExtraTreesRegressor con datos de las obras 1 y 2 y se intentan predecir los valores de la Obra 3.

En las figuras 4.30 y 4.31 se muestran los gráficos de Taylor con los resultados obtenidos para el método continuo con el modelo entrenado en la Obra 1 para predecir los valores de las obras 2 y 3 respectivamente.

Se obtienen valores francamente bajos y el modelo está muy lejos de aproximar la distribución real de los golpeos. La Obra 3 se predice algo mejor que la Obra 2, probablemente debido a que el comportamiento del terreno es más parecido entre las obras 1 y 3.

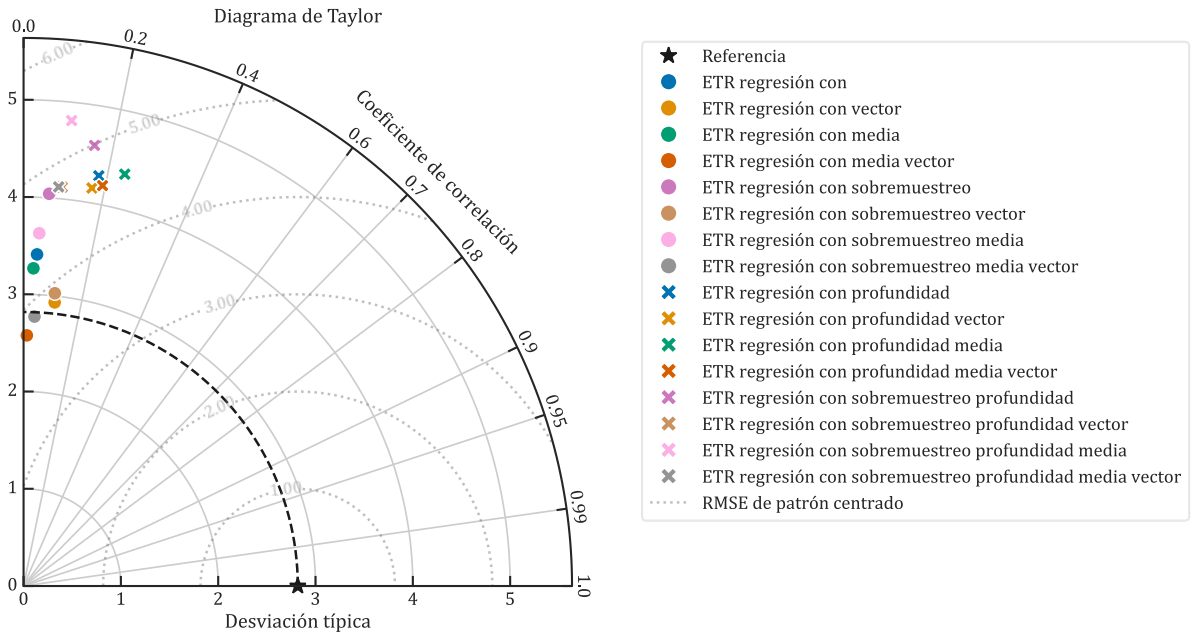


Figura 4.30: Diagrama de Taylor de la Obra 2 con algoritmo entrenado en la Obra 1 para los conjuntos de regresión y con método continuo

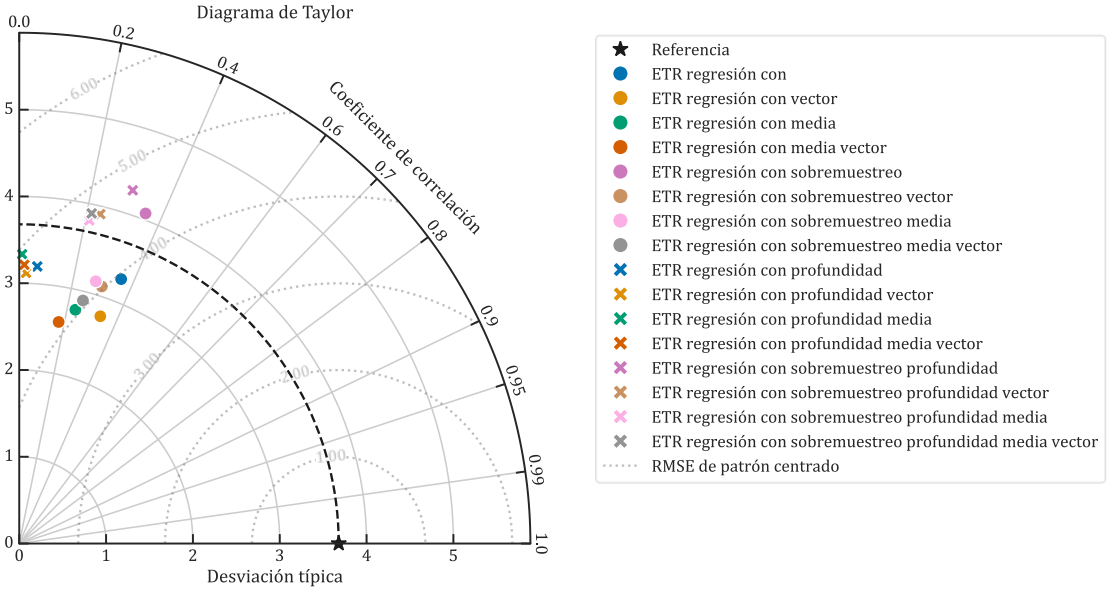


Figura 4.31: Diagrama de Taylor de la Obra 3 con algoritmo entrenado en la Obra 1 para los conjuntos de regresión y con método continuo

Por otro lado, en la Figura 4.32 se muestra la gráfica de Taylor con los resultados obtenidos para el método continuo con el modelo entrenado en las obras 1 y 2 para predecir los valores de la Obra 3.

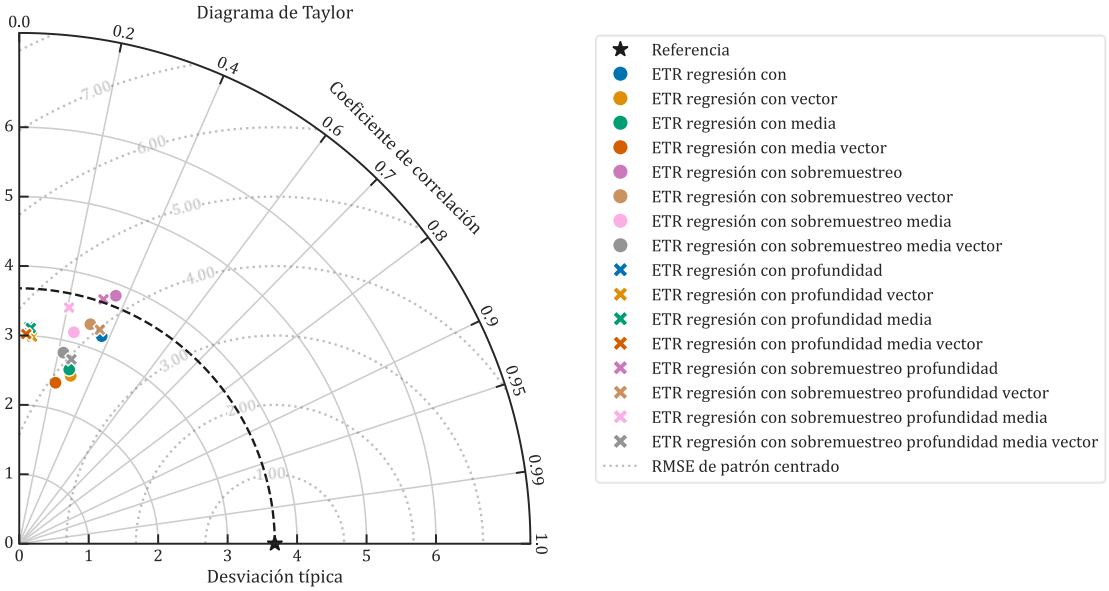


Figura 4.32: Diagrama de Taylor de la Obra 3 con algoritmo entrenado en las obras 1 y 2 para los conjuntos de regresión y con método continuo

Los resultados que se obtienen son similares a aquellos obtenidos entrenando el modelo únicamente en la Obra 1. Dado que la naturaleza del terreno entre las obras 1 y 3 es similar, parece que añadir al entrenamiento la Obra 2 no aporta mucha información adicional.

4.8. Predicción de un conjunto de obras

Hemos visto que un modelo obtenido en una sola obra no es extrapolable a otras, pero un modelo entrenado en un conjunto de obras lo suficientemente grande sí podría tener capacidad de generalización.

Se ha entrenado el modelo en conjuntos formados por el mismo número de muestras de cada obra. Se consideran las obras 1, 2 y 3.

En la Figura 4.33 se muestra el diagrama de Taylor del conjunto de las obras para el enfoque de regresión y método continuo. Se observa que obtenemos igualmente una predicción muy buena.

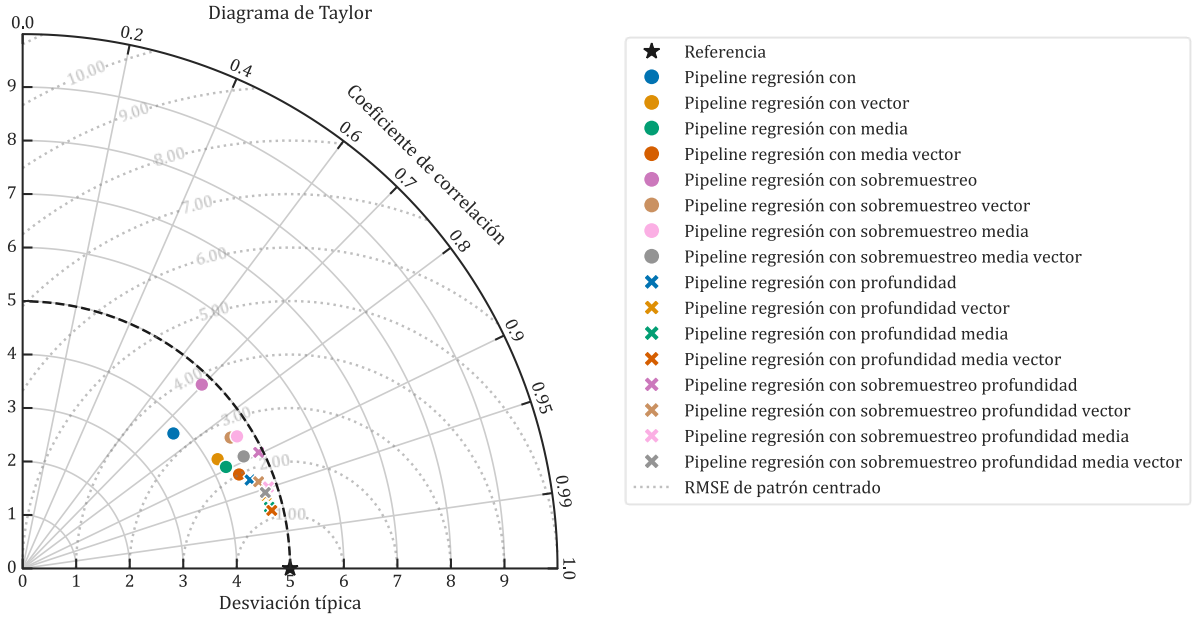


Figura 4.33: Diagrama de Taylor del conjunto de las obras 1, 2 y 3 para los conjuntos de regresión y con método continuo

Los resultados que se obtienen para un modelo entrenado en las tres obras son similares a la media de los modelos por separado. Sin embargo, tenemos la ventaja de no sobreajustar el modelo a los datos de una única obra.

Cabría pensar que, al considerar un conjunto de obras, la profundidad no sería tan influyente como característica. Sin embargo, esto no es así y en los conjuntos donde se usa la profundidad ésta sí cobra una importancia capital (Figura 4.34).

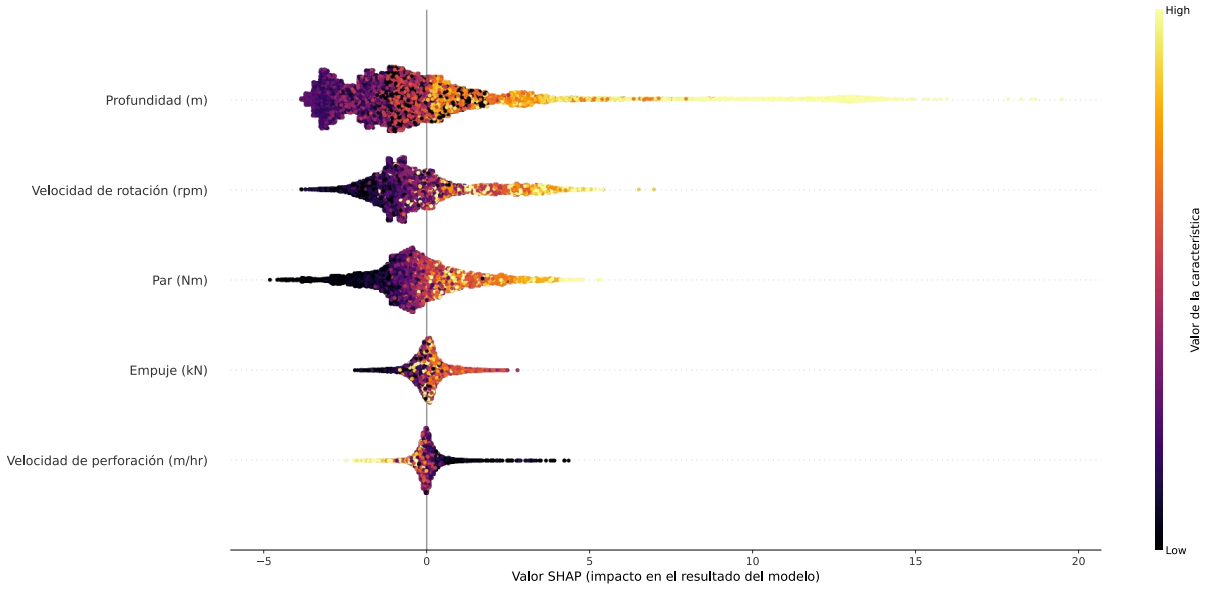


Figura 4.34: diagrama de enjambre para regresión y método continuo con profundidad para el conjunto de las obras 1, 2 y 3

Si bien es cierto que el perfil penetrométrico en función de la profundidad puede ser similar en todas las obras de tratamientos del terreno (golpeos bajos hasta alcanzar una punta algo más competente), las diferencias no son despreciables. Lo que parece ocurrir atendiendo a los gráficos ICE (Figura 4.35), es que el modelo es capaz, a partir del resto de características, de distinguir en cuál obra se encuentra y utilizar después la profundidad. Es de esperar que en un conjunto que contenga un número lo suficientemente alto de obras, este efecto desaparezca.

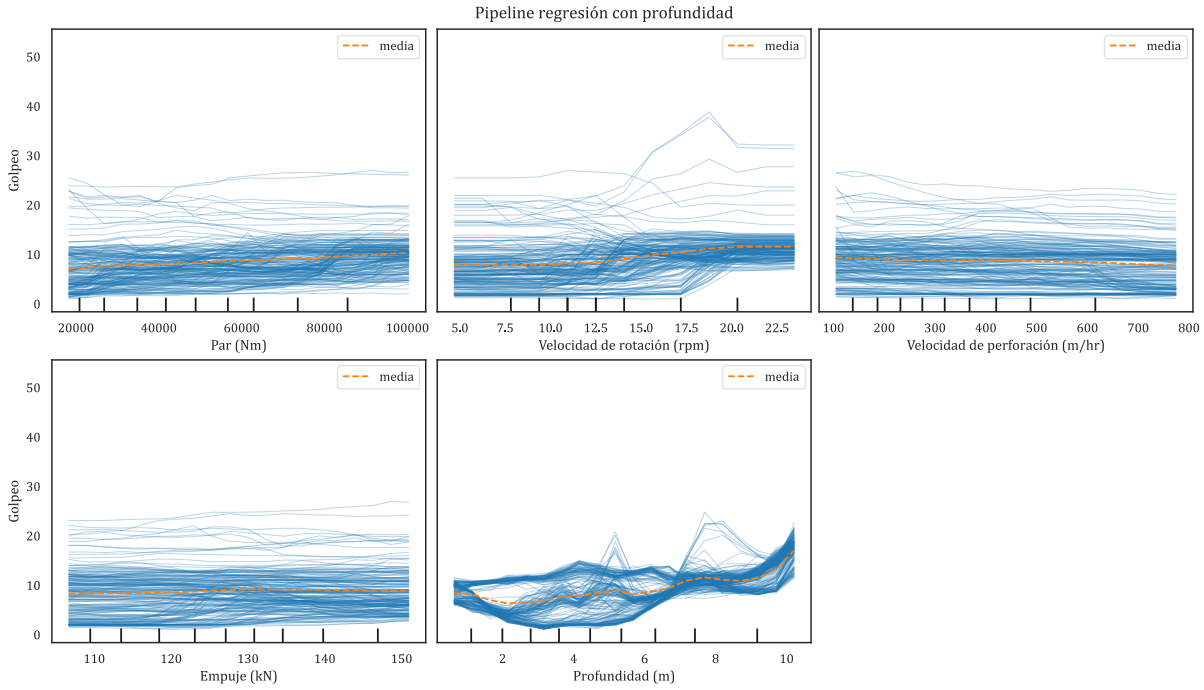


Figura 4.35: gráfico ICE para regresión y método continuo con profundidad. Conjunto de las Obras 1, 2 y 3

5. Hitos alcanzados y conclusiones

El análisis realizado en este trabajo ha evaluado la utilización de algoritmos de aprendizaje automático para obtener el perfil penetrométrico del suelo a partir del registro de parámetros de perforación. Concretamente se han utilizado datos provenientes de perforación de inclusiones rígidas ejecutadas mediante desplazamiento en cinco obras distintas. En una de estas obras la variable objetivo ha sido la resistencia a la penetración del penetrómetro estático. En el resto de las obras el objetivo ha sido el número de golpeos del penetrómetro dinámico.

A lo largo del trabajo se ha analizado cómo aplicar el aprendizaje automático a esta problemática. Puede comprobarse que la aplicación de ML a datos geotécnicos tiene sus propias peculiaridades.

Los pasos e hitos más relevantes han sido los siguientes:

- Conversión de los datos al formato AGS. Si bien los datos de perforación se toman de un sistema digital de registro, los parámetros del terreno raramente están en un formato estandarizado.
- Detección de anomalías. Es común que los penetrómetros presenten valores puntuales que no reflejen la resistencia real del terreno. Por otro lado, el sistema de registro también puede mostrar valores anómalos asociados, por ejemplo, al inicio de los movimientos.
- Constitución de los conjuntos de datos. La distancia entre las inclusiones y los penetrómetros es un factor limitante a la hora de constituir el conjunto de datos.
- Ingeniería de características. Además de las características más evidentes, se han comprobado también el vector de correlación y la media.
- Elección de *pipelines*. Se ha utilizado programación genética para comprobar cuáles son los posibles mejores *pipelines*.
- Se han comprobado los resultados para las distintas obras por separado, para el conjunto de las obras 1, 2 y 3, y para la predicción fuera del conjunto de la obra.

A continuación, se muestran las principales conclusiones que pueden obtenerse, asociadas a los diferentes objetivos propuestos:

Predicción del perfil penetrométrico del suelo:

- Los modelos obtenidos son capaces de reproducir fielmente el perfil penetrométrico del suelo cuando nos movemos en el ámbito de una única obra.

Evaluación y comparación de los algoritmos de ML:

- Los algoritmos que han obtenido mejores resultados en los *pipelines* son principalmente aquellos que utilizan algún tipo de árboles ensamblados.
- Los *pipelines* más complejos no presentan una mejora significativa frente a modelos tipo ExtraTrees con los hiperparámetros por defecto.
- La aproximación con técnicas de ML es muy superior en precisión a los parámetros compuestos.

Capacidad de generalización de los modelos:

- Al intentar predecir datos de obras ajenas al conjunto de entrenamiento, los resultados son notablemente menos precisos. Sin embargo, los resultados en un conjunto de varias obras sí son elevados. Esto hace pensar que, con un volumen de datos suficiente proveniente de numerosas y diversas obras, se podría alcanzar un modelo general.

Estrategias en el tratamiento de los datos:

- En este caso realizar sobremuestreo mediante SMOTE no ha dado lugar a mejores resultados.
- En general, el método continuo ha dado mejores resultados que el método de variabilidad.
- Enfocar el problema como regresión proporciona resultados algo mejores que como clasificación.

Ingeniería de características:

- Cuando se usa la profundidad como característica, los algoritmos tienden a darle demasiada importancia. Esto podría evitar la detección de terrenos diferentes a lo esperado para una determinada cota y, por lo tanto, no se recomienda su uso.

- Las características adicionales propuestas (vector de correlación y la media de los valores que abarca el 1,5 m desde la punta de la barrena hacia arriba) han proporcionado una mejora importante en los resultados.

El hecho de que los métodos de árboles ensamblados hayan dado los mejores resultados es coherente con la literatura existente, ya que estos métodos son de los más adecuados para datos tabulares (Grinsztajn et al., 2022).

Que los algoritmos funcionen bien en una obra, pero que no podamos extrapolar ese modelo a otras, muestra el diferente comportamiento que pueden presentar los suelos. Esto era algo esperable, sin embargo, un modelo entrenado en varias obras sí funciona bien en dicho conjunto, por lo que es posible que, con un número de obras suficiente, se pueda entrenar un modelo general aplicable a todas las futuras situaciones.

En cuanto al sobremuestreo mediante SMOTE, esta técnica genera muestras adicionales para las clases subrepresentadas, pero, aunque con ciertos criterios lógicos detrás, no dejan de ser muestras «inventadas». En este caso, no ha resultado útil y, con el objetivo de predecir mejor los valores altos de golpeo, quizá sería posible recurrir al submuestreo. Igualmente, la subestimación de las resistencias elevadas no sería un gran problema ya que estaríamos del lado de la seguridad.

Sobre los métodos continuos y de variabilidad resultan herramientas muy interesantes para conocer la cantidad de información de la que se dispone en una obra respecto a la heterogeneidad del terreno. Sin embargo, como herramientas para construir conjuntos de datos no parecen necesarios por dos razones:

- El número de muestras necesarias para entrenar un modelo en una única obra no es muy elevado, por lo que es posible limitar bastante la distancia entre el penetrómetro y su columna asociada.
- Cuando se entrena un conjunto de obras también aumenta el número de muestras disponibles que se encuentran cerca de algún penetrómetro.

En cuanto a los enfoques de clasificación y regresión, este último funciona tanto para los penetrómetros estáticos (que dan valores continuos) como para los dinámicos (que dan valores discretos). Además, proporciona valores ligeramente mejores en los coeficientes de correlación y en los errores. Sin embargo, en el valor F o la distancia, en muchos casos el enfoque de clasificación ha dado mejores resultados, por lo que no resulta evidente cuál de los dos es mejor. La

recomendación sería utilizar el método de regresión, ya que permite analizar ambos tipos de penetrómetro.

El uso de la profundidad como característica mejora todas las métricas, pero, si se observan los gráficos de explicabilidad, se comprueba que los algoritmos se adaptan demasiado a esta. Ello ocurre porque los penetrómetros de una única obra se parecen demasiado entre sí; la profundidad, como único predictor, ya obtiene valores elevados, sin embargo, esto limita la capacidad de detectar suelos distintos a lo esperado, que es precisamente el objetivo buscado.

Cabría esperar que, al analizar un conjunto de obras, como los penetrómetros entre ellas no se parecen tanto en función de la profundidad, esta característica cobrara menos importancia. Y, sin embargo, sigue siendo extraordinariamente relevante, ya que el modelo parece capaz de distinguir de cuál obra es la muestra, para luego guiarse por la profundidad a la hora de realizar la predicción. Esto es una capacidad interesante de los modelos, ya que pueden detectar el terreno en que se encuentran a partir de las características introducidas.

Con un número suficiente de obras y situaciones, es posible que los algoritmos no den esa importancia excesiva a la profundidad como característica y, a cambio, se obtendría una información importante, ya que en función de la profundidad la dificultad de desplazar el terreno cambia, aunque tenga valores parecidos de resistencia a la penetración en los penetrómetros. Además, con la profundidad se reduce la eficiencia de los penetrómetros que avanzan con mayor dificultad. Utilizar la profundidad como característica daría información al algoritmo en esta dirección.

Las características adicionales propuestas tienen un sentido físico al estar relacionadas con la longitud de la herramienta en la que se produce el desplazamiento del terreno. Añadirlas supone una mejora muy importante en la capacidad de los algoritmos.

Por último, y al menos con el número de características usadas, no parece necesario un gran preprocesamiento o afinamiento de los algoritmos, habiéndose obtenido resultados muy similares utilizando ExtraTrees por defecto que con los *pipelines*, mucho más complejos, obtenidos con TPOT.

En esta tesis se ha analizado la aplicación de algoritmos de aprendizaje automático para obtener el perfil penetrométrico del suelo a partir del registro de parámetros de perforación. Concretamente los datos se han obtenido de la ejecución de inclusiones rígidas con desplazamiento del terreno.

Los resultados obtenidos son prometedores y los algoritmos son capaces de obtener el perfil penetrométrico del suelo con precisión, especialmente al tener en cuenta las características adicionales propuestas. Si bien no es posible afirmar con rotundidad que se pueda entrenar un modelo general, tampoco es descartable que se alcance con una cantidad suficiente de datos.

Igualmente, tener la capacidad de predecir el perfil penetrométrico en una única obra ya supone un importante incremento en la información disponible, pasando a tener valores del penetrómetro en cientos de puntos, frente a unos pocos de la información inicial. Estos datos adicionales pueden mitigar o llegar a eliminar la necesidad de ensayos complementarios y permiten una mejor adaptación a las condiciones reales, por ejemplo, adaptando la longitud de las perforaciones.

No hay duda de que el uso del aprendizaje automático puede reducir la incertidumbre y aumentar la seguridad de los trabajos geotécnicos. Esto permitirá revisar los diseños *in situ*, optimizar las mediciones y rendimientos, reducir los tiempos de trabajo y las emisiones contaminantes.

Esta tesis resulta un primer paso indispensable para conseguir todos esos objetivos.

6. Publicaciones derivadas de esta tesis

El trabajo que se ha realizado a lo largo de esta tesis se ha presentado en diversos congresos, tanto nacionales como internacionales, y en dos artículos científicos.

- Congresos:
    - Estimación de valores de golpeo de penetrómetro dinámico mediante árboles de decisión y bosques aleatorios a partir del registro de parámetros durante la perforación. CITE 2021 (Martínez García, 2021)
    - Caracterización del terreno atravesado durante la ejecución de Columnas de Módulo Controlado mediante algoritmos de machine learning. XI Simposio Nacional de Ingeniería Geotécnica (Martínez García et al., 2022)
    - Measurement-While-Drilling Based Estimation of Dynamic Penetrometer Values using Machine Learning. ISMLG 2023 (García et al., 2023)
    - Application of Machine Learning algorithms to drilling data for deep foundations and soil treatments. ECSMGE 2024 (Martínez García et al., 2024)
- Artículos:
    - Measurement-While-Drilling Based Estimation of Dynamic Penetrometer Values Using Decision Trees and Random Forests. Applied Sciences (García et al., 2022)
    - Generation of Penetrometric Profile of the Soil Applying Machine Learning to Measure While Drilling Data from Deep Foundation Machinery. Applied Sciences (Martínez García et al., 2025)

7. Futuras líneas de investigación

Esta tesis resulta un comienzo muy prometedor del uso de ML para el tipo de perforaciones analizadas. Pero también pone de relevancia varias líneas de investigación para el futuro.

Quizá la más evidente es que no se ha analizado el efecto del diámetro de la perforación, que tiene una importancia capital en la potencia necesaria para atravesar el terreno.

En las obras analizadas no se disponía de datos de la forma exacta de la barrena. Tal y como muestra Krasiński (Krasinski, 2024), la forma de la herramienta tiene un efecto relevante en el par necesario.

También es posible cambiar el tipo de pica que se coloca en la herramienta, que se suele adaptar al tipo de terreno a perforar.

Todos estos parámetros (diámetro, forma de la herramienta y tipo de pica) deberían formar parte de los futuros avances. El análisis con ellos podría permitir la optimización de la herramienta de perforación.

Además, conociendo para cada perforación el terreno que se atraviesa, podrían detectarse posibles anomalías en el funcionamiento de la perforadora, comparando los parámetros esperables con los reales.

Lógicamente, el tipo de ejecución, mediante extracción o desplazamiento, cambia totalmente el funcionamiento del proceso de perforación. El análisis para perforaciones con extracción puede llevarse de manera complementaria.

En cuanto a las inclusiones con desplazamiento, estas tienen un efecto de mejora del terreno en el contorno de la perforación. En las obras analizadas, el espaciado entre columnas es lo suficientemente amplio como para que este efecto no tenga relevancia en la ejecución de columnas cercanas. Pero el poder obtener el perfil penetrométrico del suelo puede facilitar el análisis de dicha mejora perimetral.

En el caso del penetrómetro estático resulta necesario comprobar la diferencia que supone el uso de la resistencia a la penetración corregida q_t en lugar de q_c.

Por último, analizar un mayor número de obras resulta indispensable para cumplir un rango mayor de situaciones. Analizar más obras que, además, tengan menor continuidad lateral (perfiles penetrométricos dispares) permitirá entrenar modelos generales y poder realizar un análisis del uso de la profundidad como característica.

Bibliografía

Abdelaal, A., Elkatatny, S., & Abdulraheem, A. (2021). Data-Driven Modeling Approach for Pore Pressure Gradient Prediction while Drilling from Drilling Parameters. *ACS Omega*, *6*(21), 13807–13816. https://doi.org/10.1021/acsomega.1c01340

AEN/CTN 103 Geotecnia. (2010). *UNE-EN ISO 22476-12. Investigación y ensayos geotécnicos. Ensayos de campo. Parte 12: Ensayo de penetración con el cono mecánico (CPTM)*. AENOR.

AEN/CTN 103 Geotecnia. (2012). *UNE-EN ISO 22476-1. Investigación y ensayos geotécnicos. Ensayos de campo. Parte 1: Ensayos de penetración con el cono eléctrico y el piezocono*.

AEN/CTN 103 Geotecnia. (2016a). *UNE-EN 1536:2011+A1. Ejecución de trabajos geotécnicos especiales. Pilotes perforados*. AENOR.

AEN/CTN 103 Geotecnia. (2016b). *UNE-EN 12699. Ejecución de trabajos geotécnicos especiales. Pilotes de desplazamiento*.

AEN/CTN 103 Geotecnia. (2016c). *UNE-EN ISO 22476-15. Investigación y ensayos geotécnicos. Ensayos de campo. Parte 15: Registro de parámetros de perforación*. AENOR.

AEN/CTN 140 Eurocódigos estructurales. (2016). *UNE-EN 1997-1. Eurocódigo 7: Proyecto geotécnico. Parte 1: Reglas generales*. AENOR.

Ahangari, K., Moeinossadat, S. R., & Behnia, D. (2015). Estimation of tunnelling-induced settlement by modern intelligent methods. *Soils and Foundations*, *55*(4), 737–748. https://doi.org/10.1016/j.sandf.2015.06.006

Ahmed, A., Elkatatny, S., Ali, A., & Abdulraheem, A. (2019). Comparative analysis of artificial intelligence techniques for formation pressure prediction while drilling. *Arabian Journal of Geosciences*, *12*(18), 592. https://doi.org/10.1007/s12517-019-4800-7

Ahmed, A., Elkatatny, S., & Alsaihati, A. (2021). Applications of Artificial Intelligence for Static Poisson's Ratio Prediction While Drilling. *Computational Intelligence and Neuroscience*, *2021*, 1–10. https://doi.org/10.1155/2021/9956128

Ahmed, O. S., Adeniran, A. A., & Samsuri, A. (2019). Computational intelligence based prediction of drilling rate of penetration: A comparative study. *Journal of Petroleum Science and Engineering*, *172*, 1–12. https://doi.org/10.1016/j.petrol.2018.09.027

Al-AbdulJabbar, A., Elkatatny, S., Abdulhamid Mahmoud, A., Moussa, T., Al-Shehri, D., Abughaban, M., & Al-Yami, A. (2020). Prediction of the Rate of Penetration while Drilling Horizontal Carbonate Reservoirs Using the Self-

Adaptive Artificial Neural Networks Technique. *Sustainability*, *12*(4), 1376. https://doi.org/10.3390/su12041376

Al-Abduljabbar, A., Gamal, H., & Elkatatny, S. (2020). Application of artificial neural network to predict the rate of penetration for S-shape well profile. *Arabian Journal of Geosciences*, *13*(16), 784. https://doi.org/10.1007/s12517-020-05821-w

Al-Anazi, A. F., & Gates, I. D. (2015). On Support Vector Regression to Predict Poisson's Ratio and Young's Modulus of Reservoir Rock. In C. Cranganu, H. Luchian, & M. E. Breaban (Eds.), *Artificial Intelligent Approaches in Petroleum Geosciences* (pp. 167–189). Springer International Publishing. https://doi.org/10.1007/978-3-319-16531-8_5

Amar, K., & Ibrahim, A. (2012). Rate of Penetration Prediction and Optimization using Advances in Artificial Neural Networks, a Comparative Study: *Proceedings of the 4th International Joint Conference on Computational Intelligence*, 647–652. https://doi.org/10.5220/0004172506470652

Anaconda Inc. (2016). *Anaconda Software Distribution* (Version 2-2.4.0) [Computer software]. https://anaconda.com

Anemangely, M., Ramezanzadeh, A., Tokhmechi, B., Molaghab, A., & Mohammadian, A. (2018). Drilling rate prediction from petrophysical logs and mud logging data using an optimized multilayer perceptron neural network. *Journal of Geophysics and Engineering*, *15*(4), 1146–1159. https://doi.org/10.1088/1742-2140/aaac5d

Apoji, D., Ning, Z., & Soga, K. (2023). *Real-Time Ground Movement Estimation for Urban Tunneling*. 4th International Symposium on Machine Learning & Big Data in Geoscience.

Asadi, A. (2017). Application of Artificial Neural Networks in Prediction of Uniaxial Compressive Strength of Rocks using Well Logs and Drilling Data. *Procedia Engineering*, *191*, 279–286. https://doi.org/10.1016/j.proeng.2017.05.182

Association of Geotechnical & Geoenvironmental Specialists. (2023a). *AGS Data Format*. https://www.ags.org.uk/data-format/

Association of Geotechnical & Geoenvironmental Specialists. (2023b). *AGS Python Library*. https://gitlab.com/ags-data-format-wg/ags-python-library

Baghbani, A., Choudhury, T., Costa, S., & Reiner, J. (2022). Application of artificial intelligence in geotechnical engineering: A state-of-the-art review. *Earth-Science Reviews*, *228*, 103991. https://doi.org/10.1016/j.earscirev.2022.103991

Banzhaf, W. (Ed.). (1998). *Genetic programming: An introduction on the automatic evolution of computer programs and its applications*. Morgan Kaufmann Publishers ; Dpunkt-verlag.

Basarir, H., Wesseloo, J., Karrech, A., Pasternak, E., & Dyskin, A. (2017). *The use of soft computing methods for the prediction of rock properties based on*

measurement while drilling data. 537–551. https://doi.org/10.36487/ACG_rep/1704_36_Basarir

Bellman, R. (1957). *Dynamic programming*. Princeton Univ. Pr.

Bhatt, A. (2002). *Reservoir Properties from Well Logs using neural Networks*. Norwegian University of Science and Technology.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

Bouayad, D., & Emeriault, F. (2017). Modeling the relationship between ground surface settlements induced by shield tunneling and the operational and geological parameters based on the hybrid PCA/ANFIS method. *Tunnelling and Underground Space Technology*, *68*, 142–152. https://doi.org/10.1016/j.tust.2017.03.011

Boubou, R., Emeriault, F., & Kastner, R. (2010). Artificial neural network application for the prediction of ground surface movements induced by shield tunnelling. *Canadian Geotechnical Journal*, *47*(11), 1214–1233. https://doi.org/10.1139/T10-023

Bourgoyne, A. T., & Young, F. S. (1974). A Multiple Regression Approach to Optimal Drilling and Abnormal Pressure Detection. *Society of Petroleum Engineers Journal*, *14*(04), 371–384. https://doi.org/10.2118/4238-PA

Breiman, L. (1994). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). *Random forests*. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification And Regression Trees. In *Classification and Regression Trees*. Routledge. https://doi.org/10.1201/9781315139470

Bunieski, S. (2022). *Practical methodologies to apply machine learning algorithms to ground improvement data*. 20th International Conference on Soil Mechanics and Geotechnical Engineering.

Cabalar, A. F., Cevik, A., & Gokceoglu, C. (2012). Some applications of Adaptive Neuro-Fuzzy Inference System (ANFIS) in geotechnical engineering. *Computers and Geotechnics*, *40*, 14–33. https://doi.org/10.1016/j.compgeo.2011.09.008

CEN. (2008). *UNE-EN ISO 22476-2. Investigación y ensayos geotécnicos. Ensayos de campo. Parte 2: Ensayo de penetración dinámica*. AENOR.

CEN. (2023). *UNE-EN ISO 22476-1. Investigación y ensayos geotécnicos. Ensayos de campo. Parte 1: Ensayo de penetración con el cono y piezocono eléctrico*. AENOR.

Cestnik, B., Kononenko, I., & Bratko, I. (1987). ASSISTANT-86: A knowledge elicitation tool for sophisticated users. *Progress in Machine Learning,* 31–45.

Chao, Z., Ma, G., Zhang, Y., Zhu, Y., & Hu, H. (2018). The application of artificial neural network in geotechnical engineering. *IOP Conference Series: Earth and Environmental Science*, *189*, 022054. https://doi.org/10.1088/1755-1315/189/2/022054

Charles, D. J., Axtell, M. D., & Gourvenec, S. (2023). *Assessing the quality of synthetic CPT training data using time series similarity*. 4th International Symposium on Machine Learning & Big Data in Geoscience.

Chen, R.-P., Zhang, P., Kang, X., Zhong, Z.-Q., Liu, Y., & Wu, H.-N. (2019). Prediction of maximum surface settlement caused by earth pressure balance (EPB) shield tunneling with ANN methods. *Soils and Foundations*, *59*(2), 284–295. https://doi.org/10.1016/j.sandf.2018.11.005

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Comon, P. (1994). Independent component analysis, A new concept? *Signal Processing*, *36*(3), 287–314. https://doi.org/10.1016/0165-1684(94)90029-9

Congress, S. S. C., Gajurel, A., Chimauriya, H., & Puppala, A. J. (2022). *A review of the applications of artificial intelligence techniques in geotechnical engineering disciplines*. 20th International Conference on Soil Mechanics and Geotechnical Engineering.

Cover, T. M., & Hart, P. E. (1967). *Nearest Neighbor Pattern Classification*.

Cramer, J. S. (2003). The Origins of Logistic Regression. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.360300

Demertzis, K., Demertzis, S., & Iliadis, L. (2023). A Selective Survey Review of Computational Intelligence Applications in the Primary Subdomains of Civil Engineering Specializations. *Applied Sciences*, *13*(6), 3380. https://doi.org/10.3390/app13063380

Dikshit, A., Pradhan, B., & Alamri, A. M. (2021). Pathways and challenges of the application of artificial intelligence to geohazards modelling. *Gondwana Research*, *100*, 290–301. https://doi.org/10.1016/j.gr.2020.08.007

Ding, Y., & Ye, X.-W. (2023). *Influence of adjacent shield tunnelling construction on existing tunnel settlement: Field monitoring and intelligent prediction*. 4th International Symposium on Machine Learning & Big Data in Geoscience.

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. In *Advances in Geophysics* (Vol. 61, pp. 1–55). Elsevier. https://doi.org/10.1016/bs.agph.2020.08.002

Ebid, A. M. (2021). 35 Years of (AI) in Geotechnical Engineering: State of the Art. *Geotechnical and Geological Engineering*, *39*(2), 637–690. https://doi.org/10.1007/s10706-020-01536-7

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, *7*(1). https://doi.org/10.1214/aos/1176344552

Elkatatny, S., Al-AbdulJabbar, A., & Abdelgawad, K. (2020). A New Model for Predicting Rate of Penetration Using an Artificial Neural Network. *Sensors*, *20*(7), 2058. https://doi.org/10.3390/s20072058

Elvidge, S., Angling, M. J., & Nava, B. (2014). On the use of modified Taylor diagrams to compare ionospheric assimilation models. *Radio Science*, *49*(9), 737–745. https://doi.org/10.1002/2014RS005435

Eskandarian, S., Bahrami, P., & Kazemi, P. (2017). A comprehensive data mining approach to estimate the rate of penetration: Application of neural network, rule based models and feature ranking. *Journal of Petroleum Science and Engineering*, *156*, 605–615. https://doi.org/10.1016/j.petrol.2017.06.039

Fatehnia, M., & Amirinia, G. (2018). A review of Genetic Programming and Artificial Neural Network applications in pile foundations. *International Journal of Geo-Engineering*, *9*(1), 2. https://doi.org/10.1186/s40703-017-0067-6

Fortin, F.-A. (2012). *DEAP: Evolutionary Algorithms Made Easy*.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Galende-Hernández, M., Menéndez, M., Fuente, M. J., & Sainz-Palmero, G. I. (2018). Monitor-While-Drilling-based estimation of rock mass rating with computational intelligence: The case of tunnel excavation front. *Automation in Construction*, *93*, 325–338. https://doi.org/10.1016/j.autcon.2018.05.019

Gamra, A., Ninic, J., & Ghiassi, B. (2023). *A comprehensive parametric study on the impact of underground tunnelling on neighbouring structures using machine learning*. 4th International Symposium on Machine Learning & Big Data in Geoscience.

García, E. M., Alberti, M. G., & Arcos, A. A. (2023, August). Measurement-While-Drilling Based Estimation of Dynamic Penetrometer Values using Machine Learning. *Abstract of the 4th International Symposium on Machine Learning and Big Data in Geoscience*. 4th International Symposium on Machine Learning and Big Data in Geoscience, Cork (Ireland).

García, E. M., Alberti, M. G., & Arcos Álvarez, A. A. (2022). Measurement-While-Drilling Based Estimation of Dynamic Penetrometer Values Using Decision Trees and Random Forests. *Applied Sciences*, *12*(9), 4565. https://doi.org/10.3390/app12094565

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gidh, Y., Purwanto, A., & Bits, S. (2012). Artificial Neural Network Drilling Parameter Optimization System Improves ROP by Predicting/Managing Bit Wear. *All Days*, SPE-149801-MS. https://doi.org/10.2118/149801-MS

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). *Why do tree-based models still outperform deep learning on tabular data?*

Gwizdała, K., & Więcławski, P. (2020). Displacement piles—Classification and new methods for the calculation of bearing capacity. *Scientific Review Engineering and Environmental Studies (SREES)*, *29*(3), 308–318. https://doi.org/10.22630/PNIKS.2020.29.3.26

H. I. Bilgesu, L. T. Tetrick, U. Altmis, S. Mohaghegh, & S. Ameri. (1997). A New Approach for the Prediction of Rate of Penetration (ROP) Values. *All Days*, SPE-39231-MS. https://doi.org/10.2118/39231-MS

Haghiri, S. (2019). *Comparison-based methods in machine learning*.

Hansen, T. F., Erharter, G. H., Marcher, T., Liu, Z., & Tørresen, J. (2022). Improving face decisions in tunnelling by machine learning - based MWD analysis. *Geomechanics and Tunnelling*, *15*(2), 222–231. https://doi.org/10.1002/geot.202100070

Hegde, C., Daigle, H., Millwater, H., & Gray, K. (2017). Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models. *Journal of Petroleum Science and Engineering*, *159*, 295–306. https://doi.org/10.1016/j.petrol.2017.09.020

Hegde, C., & Gray, K. (2018). Evaluation of coupled machine learning models for drilling optimization. *Journal of Natural Gas Science and Engineering*, *56*, 397–407. https://doi.org/10.1016/j.jngse.2018.06.006

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *42*(1), 80–86. https://doi.org/10.1080/00401706.2000.10485983

Imamverdiyev, Y., & Sukhostat, L. (2019). Lithological facies classification using deep convolutional neural network. *Journal of Petroleum Science and Engineering*, *174*, 216–228. https://doi.org/10.1016/j.petrol.2018.11.023

Isheyskiy, V., & Sanchidrián, J. A. (2020). Prospects of Applying MWD Technology for Quality Management of Drilling and Blasting Operations at Mining Enterprises. *Minerals*, *10*(10), 925. https://doi.org/10.3390/min10100925

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Jang, H., & Topal, E. (2014). A review of soft computing technology applications in several mining problems. *Applied Soft Computing*, *22*, 638–651. https://doi.org/10.1016/j.asoc.2014.05.019

Jong, S. C., Ong, D. E. L., & Oh, E. (2021). State-of-the-art review of geotechnical-driven artificial intelligence techniques in underground soil-structure

interaction. *Tunnelling and Underground Space Technology*, *113*, 103946. https://doi.org/10.1016/j.tust.2021.103946

Juwaied, N. S. (2018). APPLICATIONS OF ARTIFICIAL INTELLIGENCE IN GEOTECHNICAL ENGINEERING. *ARPN Journal of Engineering and Applied Sciences*.

Kadkhodaie-Ilkhchi, A., Monteiro, S. T., Ramos, F., & Hatherly, P. (2010). Rock Recognition From MWD Data: A Comparative Study of Boosting, Neural Networks, and Fuzzy Logic. *IEEE Geoscience and Remote Sensing Letters*, *7*(4), 680–684. https://doi.org/10.1109/LGRS.2010.2046312

Kakoli, M., Ebrahimabadi, A., Mirshekari, B., & Sheikhzakariaee, J. (2018). A Model to Predict the Performance of Drilling Operation Using SVM Approach in One of the Iranian Oilfields. *International Journal of Petroleum and Petrochemical Engineering*. https://doi.org/10.20431/2454-7980.0402002

Khaled, S., Soliman, A. A., Mohamed, A., Gomaa, S., & Attia, A. M. (2022). New Models for Predicting Pore Pressure and Fracture Pressure while Drilling in Mixed Lithologies Using Artificial Neural Networks. *ACS Omega*, *7*(36), 31691–31699. https://doi.org/10.1021/acsomega.2c01602

Khalilidermani, M., & Knez, D. (2022). A Survey of Application of Mechanical Specific Energy in Petroleum and Space Drilling. *Energies*, *15*(9), 3162. https://doi.org/10.3390/en15093162

Khallaf, R., & Khallaf, M. (2021). Classification and analysis of deep learning applications in construction: A systematic literature review. *Automation in Construction*, *129*, 103760. https://doi.org/10.1016/j.autcon.2021.103760

Kim, C. Y., Bae, G. J., Hong, S. W., Park, C. H., Moon, H. K., & Shin, H. S. (2001). Neural network based prediction of ground surface settlements due to tunnelling. *Computers and Geotechnics*, *28*(6–7), 517–547. https://doi.org/10.1016/S0266-352X(01)00011-8

Kluyver, T., Ragan-Kelley, B., Pérez, F., Bussonnier, M., Frederic, J., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Abdalla, S., & Willing, C. (2016). *Jupyter Notebooks—A publishing format for reproducible computational workflows*.

Klyuchnikov, N., Zaytsev, A., Gruzdev, A., Ovchinnikov, G., Antipova, K., Ismailova, L., Muravleva, E., Burnaev, E., Semenikhin, A., Cherepanov, A., Koryabkin, V., Simon, I., Tsurgan, A., Krasnov, F., & Koroteev, D. (2019). Data-driven model for the identification of the rock type at a drilling bit. *Journal of Petroleum Science and Engineering*, *178*, 506–516. https://doi.org/10.1016/j.petrol.2019.03.041

Kononenko, I. (1991). Semi-naive bayesian classifier. In Y. Kodratoff (Ed.), *Machine Learning—EWSL-91* (Vol. 482, pp. 206–219). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0017015

Krasinski, A. (2015). The Analysis of Soil Resistance During Screw Displacement Pile Installation. *Studia Geotechnica et Mechanica*, *36*(3), 49–56. https://doi.org/10.2478/sgem-2014-0026

Krasinski, A. (2024). *New concept of drilling auger for displacement pile installation*.

Lawal, A., Yang, Y., He, H., & Baisa, N. L. (2024). Machine Learning in Oil and Gas Exploration: A Review. *IEEE Access*, *12*, 19035–19058. https://doi.org/10.1109/ACCESS.2023.3349216

Li, X. (2017). *Classification with Large Sparse Datasets: Convergence Analysis and Scalable Algorithms*.

Liang, H., Chen, H., Guo, J., Bai, J., & Jiang, Y. (2022). Research on lithology identification method based on mechanical specific energy principle and machine learning theory. *Expert Systems with Applications*, *189*, 116142. https://doi.org/10.1016/j.eswa.2021.116142

Liu, B., Wang, R., Guan, Z., Li, J., Xu, Z., Guo, X., & Wang, Y. (2019). Improved support vector regression models for predicting rock mass parameters using tunnel boring machine driving data. *Tunnelling and Underground Space Technology*, *91*, 102958. https://doi.org/10.1016/j.tust.2019.04.014

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Liu, Z., & Lacasse, S. (2022). Machine learning in geotechnical engineering: Opportunities and applications. In M. Rahman & Jaksa Mark (Eds.), *Proceedings of the 20th ICSMGE-State of the Art and Invited Lectures* (pp. 543–558). Australian Geomechanics Society.

Loh, W.-Y. (2014). Fifty Years of Classification and Regression Trees: Fifty Years of Classification and Regression Trees. *International Statistical Review*, *82*(3), 329–348. https://doi.org/10.1111/insr.12016

Louppe, G. (2014). *Understanding Random Forests: From Theory to Practice*. *July*.

Louppe, G. (2015). *Understanding Random Forests: From Theory to Practice* [arXiv]. http://arxiv.org/abs/1407.7502

Lundberg, S., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions* (No. arXiv:1705.07874). arXiv. http://arxiv.org/abs/1705.07874

Ma, Z., & Mei, G. (2021). Deep learning for geological hazards analysis: Data, models, applications, and opportunities. *Earth-Science Reviews*, *223*, 103858. https://doi.org/10.1016/j.earscirev.2021.103858

Ma, Z., Mei, G., & Piccialli, F. (2021). Machine learning for landslides prevention: A survey. *Neural Computing and Applications*, *33*(17), 10881–10907. https://doi.org/10.1007/s00521-020-05529-8

Mahdevari, S., Shahriar, K., Yagiz, S., & Akbarpour Shirazi, M. (2014). A support vector regression model for predicting tunnel boring machine penetration rates. *International Journal of Rock Mechanics and Mining Sciences*, *72*, 214–229. https://doi.org/10.1016/j.ijrmms.2014.09.012

Mahmoodzadeh, A., Nejati, H. R., Mohammadi, M., Hashim Ibrahim, H., Rashidi, S., & Ahmed Rashid, T. (2022). Forecasting tunnel boring machine penetration rate using LSTM deep neural network optimized by grey wolf optimization algorithm. *Expert Systems with Applications*, *209*, 118303. https://doi.org/10.1016/j.eswa.2022.118303

Mantha, B. ., & Samuel, R. . (2016). ROP Optimization Using Artificial Intelligence Techniques with Statistical Regression Coupling. *Day 3 Wed, September 28, 2016*, D031S041R007. https://doi.org/10.2118/181382-MS

Martínez García, E. (2021, March). *Estimación de valores de golpeo de penetrómetro dinámico mediante árboles de decisión y bosques aleatorios a partir del registro de parámetros durante la perforación*. VI International Conference on Technological Innovation in Building, Madrid.

Martínez García, E., Alberti, M. G., & Arcos Álvarez, A. A. (2025). Generation of Penetrometric Profile of the Soil Applying Machine Learning to Measure While Drilling Data from Deep Foundation Machinery. *Applied Sciences*, *15*(3), 1331. https://doi.org/10.3390/app15031331

Martínez García, E., García Alberti, M., & Arcos Álvares, A. A. (2022). *Caracterización del terreno atravesado durante la ejecución de Columnas de Módulo Controlado mediante algoritmos de machine learning*. XI Simposio Nacional de Ingeniería Geotécnica.

Martínez García, E., García Alberti, M., & Arcos Álvarez, A. A. (2024). *Application of Machine Learning algorithms to drilling data for deep foundations and soil treatments*.

Matejka, J., & Fitzmaurice, G. (2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 1290–1294. https://doi.org/10.1145/3025453.3025912

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Merghadi, A., Yunus, A. P., Dou, J., Whiteley, J., ThaiPham, B., Bui, D. T., Avtar, R., & Abderrahmane, B. (2020). Machine learning methods for landslide susceptibility studies: A comparative overview of algorithm performance. *Earth-Science Reviews*, *207*, 103225. https://doi.org/10.1016/j.earscirev.2020.103225

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press. https://doi.org/10.7551/mitpress/11301.001.0001

Moayedi, H., Mosallanezhad, M., Rashid, A. S. A., Jusoh, W. A. W., & Muazu, M. A. (2020). A systematic review and meta-analysis of artificial neural network application in geotechnical engineering: Theory and applications. *Neural Computing and Applications*, *32*(2), 495–518. https://doi.org/10.1007/s00521-019-04109-9

Moeinossadat, S. R., Ahangari, K., & Shahriar, K. (2016). Calculation of maximum surface settlement induced by EPB shield tunnelling and introducing most effective parameter. *Journal of Central South University*, *23*(12), 3273–3283. https://doi.org/10.1007/s11771-016-3393-5

Moeinossadat, S. R., Ahangari, K., & Shahriar, K. (2018). Control of Ground Settlements Caused by EPBS Tunneling Using an Intelligent Predictive Model. *Indian Geotechnical Journal*, *48*(3), 420–429. https://doi.org/10.1007/s40098-017-0253-7

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the Analysis of Survey Data, and a Proposal. *Journal of the American Statistical Association*, *58*(302), 415–434. https://doi.org/10.1080/01621459.1963.10500855

Morgenroth, J., Khan, U. T., & Perras, M. A. (2019). An Overview of Opportunities for Machine Learning Methods in Underground Rock Engineering Design. *Geosciences*, *9*(12), 504. https://doi.org/10.3390/geosciences9120504

Morse, M. D., & Patel, J. M. (2007). An efficient and accurate method for evaluating time series similarity. *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, 569–580. https://doi.org/10.1145/1247480.1247544

Nishi, K., Suzuki, Y., & Sasao, H. (1998). Estimation of soil resistance using rotary percussion drill. *Proceedings First International Conference on Site Characterization*.

Ocak, I., & Seker, S. E. (2013). Calculation of surface settlements caused by EPBM tunneling using artificial neural network, SVM, and Gaussian processes. *Environmental Earth Sciences*, *70*(3), 1263–1276. https://doi.org/10.1007/s12665-012-2214-x

Olafadehan, O. A., Ahaotu, I. D., Oa, O., & Id, A. (2023). *Comparative Analysis of Machine Learning Algorithms in Predicting Rate of Penetration during Drilling*. *1*.

Olson, R. S., Bartley, N., Urbanowicz, R. J., & Moore, J. H. (2016). Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science. *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, 485–492. https://doi.org/10.1145/2908812.2908918

Ong, V. K. P., Do, Q. K., Nguyen, T., Vo, H. L., Nguyen, N. A. T., & Ly, N. Y. L. (2022). *Prediction of Rate of Penetration for wells at Nam Con Son basin using Artificial Neural Networks models*. 116–109. https://doi.org/10.29007/4sdt

Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space*. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,

A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.

Pfister, P. (1985). Recording Drilling Parameters in Ground Engineering. *Ground Engineering*, *18*(3), 16–21.

Phoon, K.-K., Ching, J., & Cao, Z. (2022). Unpacking data-centric geotechnics. *Underground Space*, *7*(6), 967–989. https://doi.org/10.1016/j.undsp.2022.04.001

Pourtaghi, A., & Lotfollahi-Yaghin, M. A. (2012). Wavenet ability assessment in comparison to ANN for predicting the maximum surface settlement caused by tunneling. *Tunnelling and Underground Space Technology*, *28*, 257–271. https://doi.org/10.1016/j.tust.2011.11.008

Python Software Foundation. (2022). *Python programming language*. https://www.python.org/

Rahimi, A., & Recht, B. (2007). *Random Features for Large-Scale Kernel Machines*.

Rebala, G., Ravi, A., & Churiwala, S. (2019). *An Introduction to Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-15729-6

Ren, Q., Zhang, H., Zhang, D., Zhao, X., Yan, L., Rui, J., Zeng, F., & Zhu, X. (2022). A framework of active learning and semi-supervised learning for lithology identification based on improved naive Bayes. *Expert Systems with Applications*, *202*, 117278. https://doi.org/10.1016/j.eswa.2022.117278

Richa, T., Lamare, T., Pereira, J.-M., Guayacan, L.-M., Chapron, G., & Lanquette, F. (2024). *PREVISION EN TEMPS REEL DES TASSEMENTS INDUITS PAR LE CREUSEMENT AU TUNNELIER*.

Richa, T., Lebdaoui, M. S., Pereira, J.-M., Guayacán Carrillo, L. M., Chapron, M. G., & Lanquette, M. F. (2023). *Ensemble methods for the prediction of tunnel induced settlements*. 4th International Symposium on Machine Learning & Big Data in Geoscience.

Richa, T., Pereira, J. M., Chapron, G., & Lanquette, F. (2024). *Accuracy of machine learning techniques in forecasting tunnelling-induced soil settlements with limited data*.

Rolnick, D. (2018). *Towards an Integrated Understanding of Neural Networks*.

Rosenblatt, F. (1957, January). *The perceptron—A perceiving and recognizing automaton*. Cornell Aeronautical Laboratory.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Sabah, M., Talebkeikhah, M., Wood, D. A., Khosravanian, R., Anemangely, M., & Younesi, A. (2019). A machine learning approach to predict drilling rate

using petrophysical and mud logging data. *Earth Science Informatics*, *12*(3), 319–339. https://doi.org/10.1007/s12145-019-00381-4

Santos, O. J., & Celestino, T. B. (2008). Artificial neural networks analysis of São Paulo subway tunnel settlement data. *Tunnelling and Underground Space Technology*, *23*(5), 481–491. https://doi.org/10.1016/j.tust.2007.07.002

Sapronova, A., & Marcher, T. (2023). *Correlational analysis for extracting patterns in geotechnical data*. 4th International Symposium on Machine Learning & Big Data in Geoscience.

Shahin, M. A. (2016). State-of-the-art review of some artificial intelligence applications in pile foundations. *Geoscience Frontiers*, *7*(1), 33–44. https://doi.org/10.1016/j.gsf.2014.10.002

Shapley, L. S. (1953). 17. A Value for n-Person Games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the Theory of Games (AM-28), Volume II* (pp. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Shi, J., Ortigao, J. A. R., & Bai, J. (1998). Modular Neural Networks for Predicting Settlements during Tunneling. *Journal of Geotechnical and Geoenvironmental Engineering*, *124*(5), 389–395. https://doi.org/10.1061/(ASCE)1090-0241(1998)124:5(389)

Siddig, O., Gamal, H., Elkatatny, S., & Abdulraheem, A. (2021). Real-time prediction of Poisson's ratio from drilling parameters using machine learning tools. *Scientific Reports*, *11*(1), 12611. https://doi.org/10.1038/s41598-021-92082-6

Somerton, W. H. (1959). A Laboratory Study of Rock Breakage by Rotary Drilling. *Transactions of the AIME*, *216*(01), 92–97. https://doi.org/10.2118/1163-G

Stanton, J. M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, *9*(3), 3. https://doi.org/10.1080/10691898.2001.11910537

Suwansawat, S., & Einstein, H. H. (2006). Artificial neural networks for predicting the maximum surface settlement caused by EPB shield tunneling. *Tunnelling and Underground Space Technology*, *21*(2), 133–150. https://doi.org/10.1016/j.tust.2005.06.007

Svenska Geotekniska Föreningen. (1999). *Metodbeskrivning för Jord-bergsondering*.

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Teale, R. (1965). *THE CONCEPT OF SPECIFIC ENERGY IN ROCK DRILLING*.

Tehrani, F. S., Calvello, M., Liu, Z., Zhang, L., & Lacasse, S. (2022). Machine learning and landslide studies: Recent advances and applications. *Natural Hazards*, *114*(2), 1197–1245. https://doi.org/10.1007/s11069-022-05423-7

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Wadoux, A. M. J.-C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Science Reviews*, *210*, 103359. https://doi.org/10.1016/j.earscirev.2020.103359

Wang, F., Gou, B., & Qin, Y. (2013). Modeling tunneling-induced ground surface settlement development using a wavelet smooth relevance vector machine. *Computers and Geotechnics*, *54*, 125–132. https://doi.org/10.1016/j.compgeo.2013.07.004

Williams, C. K. I., & Seeger, M. (2000). Using the Nyström method to speed up kernel machines. *Proceedings of the 13th International Conference on Neural Information Processing Systems*, 661–667.

Xie, J., Huang, J., Zeng, C., Jiang, S.-H., & Podlich, N. (2020). Systematic Literature Review on Data-Driven Models for Predictive Maintenance of Railway Track: Implications in Geotechnical Engineering. *Geosciences*, *10*(11), 425. https://doi.org/10.3390/geosciences10110425

Yavari, H., Sabah, M., Khosravanian, R., & Wood, D. A. (2018). *Application of an Adaptive Neuro-fuzzy Inference System and Mathematical Rate of Penetration Models to Predicting Drilling Rate*. *7*(3).

Zhang, H. (2004). *The Optimality of Naive Bayes*.

Zhang, P., Wu, H.-N., Chen, R.-P., & Chan, T. H. T. (2020). Hybrid meta-heuristic and machine learning algorithms for tunneling-induced settlement prediction: A comparative study. *Tunnelling and Underground Space Technology*, *99*, 103383. https://doi.org/10.1016/j.tust.2020.103383

Zhang, P., Yin, Z.-Y., & Jin, Y.-F. (2022). Machine Learning-Based Modelling of Soil Properties for Geotechnical Design: Review, Tool Development and Comparison. *Archives of Computational Methods in Engineering*, *29*(2), 1229–1245. https://doi.org/10.1007/s11831-021-09615-5

Zhang, W. G., Li, H. R., Wu, C. Z., Li, Y. Q., Liu, Z. Q., & Liu, H. L. (2021). Soft computing approach for prediction of surface settlement induced by earth pressure balance shield tunneling. *Underground Space*, *6*(4), 353–363. https://doi.org/10.1016/j.undsp.2019.12.003

Zhang, W., Li, H., Li, Y., Liu, H., Chen, Y., & Ding, X. (2021). Application of deep learning algorithms in geotechnical engineering: A short critical review. *Artificial Intelligence Review*, *54*(8), 5633–5673. https://doi.org/10.1007/s10462-021-09967-1

Zhang, W., Zhang, R., Wu, C., Goh, A. T. C., Lacasse, S., Liu, Z., & Liu, H. (2020). State-of-the-art review of soft computing applications in underground excavations. *Geoscience Frontiers*, *11*(4), 1095–1106. https://doi.org/10.1016/j.gsf.2019.12.003

Zhou, J., Shi, X., Du, K., Qiu, X., Li, X., & Mitri, H. S. (2017). Feasibility of Random-Forest Approach for Prediction of Ground Settlements Induced by the Construction of a Shield-Driven Tunnel. *International Journal of Geomechanics*, *17*(6), 04016129. https://doi.org/10.1061/(ASCE)GM.1943-5622.0000817

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x