UNIVERSIDAD NACIONAL MICAELA BASTIDAS DE APURÍMAC

FACULTAD DE INGENIERÍA

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA INFORMÁTICA Y SISTEMAS

Tesis en formato de artículo científico

Determinación del mejor modelo de Machine Learning para la predicción del California Bearing Ratio de suelos en Abancay, 2024

Presentado por:

Flor de Cantuta Tello Sarmiento

Para optar el título de Ingeniero Informático y Sistemas

Abancay, Perú

2025

UNIVERSIDAD NACIONAL MICAELA BASTIDAS DE APURÍMAC

FACULTAD DE INGENIERÍA

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA INFORMÁTICA Y SISTEMAS

TESIS

Determinación del mejor modelo de Machine Learning para la predicción del California Bearing Ratio de suelos en Abancay, 2024.

Presentado por **Flor de Cantuta Tello Sarmiento**, para optar el título profesional de Ingeniero Informático y Sistemas

Sustentado y aprobado el 28 de agosto del 2025 ante el jurado evaluador:

Presidente: *Mag. Mario Aquino Cruz*

Primer miembro: *Mtro. Rafael Ricardo Quispe Merma*

Segundo miembro: *Mtro. Virgilio Martinez Duran*

Asesor: *Dr. Manuel Jesús Ibarra Cabrera*

"Año de la recuperación y consolidación de la economía peruana"

CONSTANCIA DE ORIGINALIDAD
Nº 192-2025

La Universidad Nacional Micaela Bastidas de Apurímac, a través de la Unidad de Investigación de la Facultad de Ingeniería declara que, la Tesis en formato de articulo científico que lleva por título: **Determinación del mejor modelo de Machine Learning para la predicción del California Bearing Ratio de suelos en Abancay, 2024",** presentado por la Bach**: Flor de Cantuta Tello Sarmiento,** Para optar el Título de **Ingeniero Informático y Sistemas ;** ha sido sometido a un mecanismo de evaluación y verificación de similitud, a través del Software Turnitin, siendo el índice de similitud ACEPTABLE de **(13%)** por lo que, cumple con los criterios de originalidad establecidos por la Universidad.

Abancay, 20 de agosto del 2025

Atentamente,

UNIVERSIDAD NACIONAL MICAELA BASTIDAS
FACULTAD DE INGENIERÍA
Dra. Hesmeraldo Tejas Enriquez
DIRECTORA DE LA UNIDAD DE INVESTIGACIÓN
FACULTAD DE INGENIERÍA

C. c.
Archivo
REG. Nº 648

Agradecimiento

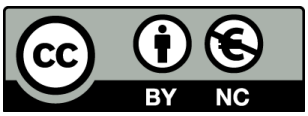
Agradezco a Dios por la fortaleza, salud, sabiduría y perseverancia que me permitió alcanzar mis objetivos; a mis padres, por su amor incondicional y apoyo constante, pilares fundamentales de mi vida; a los docentes de la facultad, cuya dedicación y compromiso fueron esenciales para mi formación; y de manera especial, a mi asesor, Dr. Manuel Jesús Ibarra Cabrera, por su paciencia, guía y apoyo continuo, claves para el desarrollo y culminación de este proyecto.

Dedicatoria

Dedico este logro a mis padres por su amor incondicional, su apoyo constante y por creer siempre en mi, Este logró es reflejo de su dedicación y esfuerzo. También a todos aquellos que, de una manera u otra, han sido parte de este proceso, este trabajo está dedicado a usteds con profunda gratitud.

Determinación del mejor modelo de Machine Learning para la predicción del California Bearing Ratio de suelos en Abancay, 2024

Línea de investigación: Ingeniería de software.



ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

C&T Riqchary
Revista de Investigación en Ciencia y Tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x

Vol. 6 Núm. 1 (2024)
Publicado: 25/07/2024
https://doi.org/10.57166/riqchary.v6.n1.2024.121

Determinación del mejor modelo de Machine Learning para la predicción del California Bearing Ratio de suelos en Abancay, 2024

Determination of the best Machine Learning model for the prediction of the California Bearing Ratio of soils in Abancay, 2024

Flor de Cantuta Tello-Sarmiento [A], Manuel J. Ibarra-Cabrera [B]
ORCID: 0009-0006-9567-5220 [A], 0000-0001-6711-4916 [B]
(Recepción: 16/07/2024 y aceptación 25/07/2024)

Resumen— El California Bearing Ratio (CBR) es un índice fundamental en la ingeniería geotécnica para evaluar la capacidad de soporte de los suelos, especialmente en el diseño y construcción de pavimentos y otras estructuras sobre terreno natural. Pero la determinación de este índice es una tarea costosa y laboriosa, por dicha razón en este estudio se propone la predicción del CBR mediante modelos de machine learning. Se desarrollaron 3 modelos de aprendizaje automático, redes neuronales profundas (DNN), árboles de decisión y máquinas de vectores de soporte. El trabajo consistió en recolectar 310 registros con características del suelo, de los cuales 217 registros fueron considerados para el entrenamiento, 62 para la validación y 31 para las pruebas; los datos fueron recolectados en 3 laboratorios de mecánica de suelos de la ciudad de Abancay, provincia de Abancay en la región Apurímac de Perú, donde se obtuvieron las siguientes características físicas del suelo: el porcentaje de grava, porcentaje de finos, el óptimo contenido de humedad (OCH), límite líquido, límite plástico, índice de plasticidad, máxima densidad seca (MDS) y para la característica a predecir el valor del CBR al 100%. Los modelos fueron evaluados con el coeficiente de determinación (R^2), el error absoluto medio (MAE), el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE). Los resultados muestran que el algoritmo o modelo de árboles de decisión es el más eficiente para predecir el CBR al 100% porque tiene el mejor coeficiente de determinación $R^2 = 0.9307$ y también los valores más bajos para el MSE = 9.199, MAE = 1.216 y RMSE = 3.033; estos valores son los mejores en relación con los encontrados para los modelos de redes neuronales profundas y el de máquina de vectores de soporte para regresión.

Palabras clave: CBR, machine learning, árboles de decisión, red neuronal, regresión lineal múltiple

Abstract— The California Bearing Ratio (CBR) is a fundamental index in geotechnical engineering to evaluate the bearing capacity of soils, especially in the design and construction of pavements and other structures on natural ground. However, the determination of this index is a costly and laborious task, for that reason in this study, the prediction of CBR using machine learning models is proposed. Three machine learning models were developed, deep neural networks (DNN), decision trees, and support vector machines. The work consisted of collecting 310 records with soil characteristics, of which 217 records were considered for training, 62 for validation and 31 for testing; the data were collected in 3 soil laboratories in the city of Abancay, province of Abancay in the Apurimac region of Peru, where the following physical soil characteristics were obtained: gravel percentage, percentage of fines, optimum moisture content (OCH), liquid limit, plastic limit, plasticity index and maximum dry density (MDS) and for the characteristic to be predicted the CBR value at 100%. The models were evaluated with the coefficient of determination (R^2), the mean absolute error (MAE), the mean square error (MSE), and the root mean square error (RMSE). The results show that the decision tree algorithm or model is the most efficient for predicting the CBR at 100% because it has the best coefficient of determination $R^2 = 0.9307$ and also the lowest values for the MSE = 9.199, MAE = 1.216 and RMSE = 3.033; these values are the best in relation to those found for the deep neural network and support vector regression machine models.

Keywords: CBR, machine learning, decision tree, neural network, multiple linear regression.

A. Flor de Cantuta Tello-Sarmiento, Escuela Profesional de Ingenieria Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac-Perú, 111181@unamba.edu.pe.
B. Manuel J. Ibarra-Cabrera, Departamento Académico de Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac-Perú, mibarra@unamba.edu.pe.

1 Introducción

El empleo de técnicas de aprendizaje automático han sido fundamentales para resolver problemas en la ingeniería, especialmente cuando las variables implicadas tienen relaciones no lineales y resulta difícil representar el problema mediante funciones matemáticas convencionales [1].

California Bearing Ratio (CBR) es una medida de la resistencia del suelo a la penetración, el CBR se mide en laboratorio aplicando una carga a una muestra de suelo compactado y midiendo la resistencia a la penetración de un pistón estándar, el CBR viene siendo considerado como la única forma para evaluar la capacidad de soporte de los suelos, especialmente en el diseño y construcción de pavimentos y otras estructuras sobre terreno natural. Un CBR alto indica que el suelo tiene una buena capacidad de soporte y puede reducir

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 1 (2024) - publicado:25/07/2024
DOI https://doi.org/10.57166/riqchary.v6.n1.2024.121

los costos de construcción al requerir menos materiales de refuerzo [2].

Sin embargo, debido a que el ensayo de CBR es un proceso laborioso y que consume mucho tiempo aproximadamente 4 días y en ocasiones, los resultados pueden carecer de precisión debido a errores humanos y a la posible falta de habilidades técnicas suficientes del personal de laboratorio [2]. La aplicación de algoritmos de machine learning para calcular el CBR basado en ciertas características del suelo es una excelente opción, principalmente en lo que respecta a una mayor precisión y reducción de tiempo.

Una de las alternativas para mejorar el proceso de diseño y construcción de pavimentos es la aplicación de modelos de machine learning para predecir el CBR de los suelos, ya que teniendo valores aproximados se pueden estimar presupuestos preliminares en obras viales.

Asimismo, la ingeniería ha sido transformada por la revolución digital, donde el crecimiento masivo de datos y el avance de la inteligencia artificial, en particular el aprendizaje automático (ML), han sido cruciales. Machine learning se ha convertido en una herramienta esencial en diversas áreas de ingeniería, desde mejorar procesos industriales hasta diseñar sistemas autónomos y prever comportamientos complejos. Adicionalmente, el aprendizaje automatizado, es un área de estudio de la inteligencia artificial y de las ciencias de la computación que aporta bastante a la ingeniería geotécnica [3], el adecuado uso de técnicas de machine learning trae consigo beneficios económicos, técnicos y contribuye a la optimización de procesos, así como, por ejemplo, se usa para la estimación de la densidad del suelo [4], también para la evaluación de algunas propiedades del suelo para la caracterización de macizos rocosos [5].

En el Perú, existe una variedad de suelos y la construcción de carreteras y pavimentos es esencial para el desarrollo socioeconómico del país. Estas infraestructuras mejoran la conectividad, impulsan el comercio, promueven la integración regional, facilitan el acceso a servicios básicos y aumentan la seguridad vial. Estos proyectos viales son cruciales para fomentar el progreso y elevar la calidad de vida de la población peruana. En Abancay, como en todo el Perú, los pavimentos suelen ser descuidados tras su construcción, lo que resulta en su deterioro visible en el centro de la ciudad. Las autoridades a menudo no toman medidas de mantenimiento adecuadas, lo que lleva a reparaciones mal ejecutadas o incompletas, acelerando el deterioro de las infraestructuras. Asimismo, la falta de un plan de mantenimiento para las infraestructuras viales se refleja en los daños no reparados, empeorando las condiciones viales. La congestión de tráfico se agrava por la gran cantidad de vehículos pesados, afectando tanto a vehículos particulares como al transporte público. Además, el sistema de drenaje deficiente causa obstrucciones durante las lluvias, afectando el sistema de agua y desagüe y provocando daños en las pistas debido a filtraciones y colapsos de tuberías subterráneas [6].

En cuanto al espesor de la capa de asfalto, es esencial realizar estudios adecuados para la construcción de carreteras, ya que estas vías unen diferentes regiones y facilitan el desarrollo económico, ignorar estos aspectos puede conllevar a altos costos logísticos para la economía, por lo que es vital considerar estos factores durante la planificación y construcción de infraestructuras viales [7]. Al diseñar un pavimento para una carretera, es esencial determinar la capacidad de soporte del suelo subyacente, ya que esto influye en la selección de materiales y el espesor del pavimento necesario para garantizar una estructura segura, duradera y eficiente. El índice California Bearing Ratio (CBR) se usa para evaluar la capacidad de soporte del suelo, siendo crucial para decisiones informadas sobre el diseño de pavimentos en vías y campos de aterrizaje [8]. Sin embargo, la escasez de datos y la ausencia de análisis de suelos son problemas significativos en la red vial de carreteras en el Perú [9].

Este estudio busca comparar distintas técnicas de machine learning para determinar cuál es la más eficiente, para calcular el indicador de la calidad de suelo basado en su resistencia (California Bearing Ratio), considerando varias características del suelo como entrada. Después de un análisis exploratorio de datos (EDA) [10] donde tenemos múltiples variables predictoras para predecir la variable de respuesta (en este caso el CBR al 100%) se opta por hacer la comparación de 3 modelos de regresión múltiple para determinar el mejor modelo de machine learning, utilizando varias métricas de evaluación como el coeficiente de determinación(R^2), error cuadrático medio (MSE), error absoluto medio (MAE) y raíz del error cuadrático medio (RMSE) [11].

2 TRABAJOS RELACIONADOS

En este contexto, existen algunos autores que han realizado estudios para predecir características del suelo, por ejemplo, Johan y José realizaron una investigación cuyo objetivo fue construir bibliotecas espectrales con técnicas de machine learning para los suelos tropicales de Costa Rica y determinar las bandas hiper-espectrales óptimas en el rango espectral visible infrarrojo cercano e infrarrojo de onda corta para caracterizar propiedades de suelo. El método propuesto logró una estimación precisa del contenido de diferentes componentes (Ca, Mg, Fe, C, N y CICE) con un R^2 superior a 0,8 y un error cuadrático medio (RMSE) inferior al 10% [12].

Por otro lado, Frank y Heber realizaron una investigación cuyo objetivo fue desarrollar un modelo predictivo de las propiedades del suelo usando RNA del tipo perceptrón multicapa, para predecir individualmente la máxima densidad seca (MDD), el óptimo contenido de humedad (OMC), el valor de la relación de soporte de California (CBR) al 95% y el CBR al 100% usando como variables de entrada otras propiedades del suelo. La investigación contó con un dataset de 285 ejemplos. Teniendo como resultados la predicción de la MDD con R^2=0,90, OMC con R^2=0,87, CBR al 95% con R^2=0,92 y CBR al 100% con R^2=0,89, respectivamente, demostrando que los modelos son eficientes para predecir las propiedades del suelo [13].

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 1 (2024) - publicado:25/07/2024
DOI https://doi.org/10.57166/riqchary.v6.n1.2024.121

También Saúl y Alaín realizaron un proyecto de investigación donde el objetivo principal fue determinar el modelo de redes bayesianas para predecir los parámetros de resistencia al corte del suelo (ángulo de fricción y cohesión) en función de sus propiedades físicas (granulometría, límites de Atterberg, contenido de humedad y peso específico). La investigación concluyó con un modelo de redes bayesianas con un coeficiente de determinación R^2 de 0.89790 para el ángulo de fricción y de 0.92819 para la cohesión y reafirmó el potencial de las técnicas avanzadas de aprendizaje automático, como las redes bayesianas, como herramientas innovadoras y valiosas para la predicción y la resolución de problemas complejos en el campo de la ingeniería [14].

Asimismo, Marisabel y Rodrigo en su investigación donde su objetivo fue evaluar el desempeño de un modelo de inteligencia artificial basado en redes neuronales artificiales (RNA) para predecir los parámetros de resistencia al corte de suelos, específicamente el ángulo de fricción interna y la cohesión, a partir de sus propiedades físicas como los límites de Atterberg, granulometría, humedad y peso específico. El modelo de RNA fue entrenado utilizando la metodología de propagación inversa (Feed-Forward Backprop) en Matlab, con un 72% de los datos para el entrenamiento y el 28% restante para la validación. Los resultados, evaluados mediante el análisis estadístico de error medio cuadrático (MSE), mostraron un coeficiente de determinación R-cuadrado de 0.93927 en el entrenamiento, 0.99746 en la validación, y 0.96465 en la prueba, obteniendo un modelo final con R^2=0.95507. Esto demuestra que el modelo de RNA es eficaz con un error menor al 5%, proponiéndose, así como una alternativa viable para estudios geotécnicos en la planificación, diseño y ejecución de proyectos de construcción [15].

Como se observa, existen varios estudios donde se aplicaron técnicas de machine learning en el área de la ingeniería geotécnica como una alternativa en la predicción de características del suelo.

El objetivo de este proyecto de investigación fue determinar el mejor algoritmo de machine learning para determinar el valor del CBR al 100% en base a ciertas características del suelo ya que la predicción del CBR es esencial para determinar el espesor y los materiales adecuados para la construcción de pavimentos y carreteras, para lo cual se recolectó registros de ensayos mecánicos llevados a cabo en los diferentes laboratorios especializados en suelos y concretos y algunos expedientes técnicos de obras de carreteras en el departamento de Apurímac.

El entrenamiento se realizó con tres modelos o algoritmos de machine learning: árboles de decisión, máquinas de vectores de soporte y redes neuronales profundas.

3 Método

3.1 Ámbito de estudio

Este estudio se realizó en la Universidad Nacional Micaela Bastidas de Apurímac (UNAMBA) en la ciudad de Abancay. Los datos fueron recolectados de informes de estudios de mecánica de suelos de los laboratorios de suelos de la ciudad de Abancay y expedientes técnicos de obras. Los datos proporcionados pertenecen a diferentes proyectos de desarrollo y mejoramiento de pavimentos en el departamento de Apurímac entre los años 2022 y 2024.

3.2 Tipo y nivel de investigación.

Este trabajo es una investigación de tipo aplicada y de nivel descriptivo.

3.3 Procedimiento

El procedimiento fue el siguiente:

a) Recopilación y adquisición de datos con las características de los suelos y el indicador de calidad de suelo basado en su resistencia (CBR).
b) Preparación de datos, tener los datos limpios y preparados para en el entrenamiento de los modelos, esto incluye la normalización de datos, manejo de valores faltantes y la separación de datos de entrenamiento, validación y pruebas.
c) Entrenamiento de los modelos, en esta etapa se entrena cada modelo con el 70% de datos.
d) Validación de los modelos, en esta etapa de utiliza el 20% de datos para evaluar el rendimiento de cada modelo con las métricas relevantes.
e) Ajuste de los modelos, en esta etapa, si corresponde se ajusta la arquitectura del modelo o los parámetros de entrenamiento y se vuelve a entrenar los modelos para mejorar su rendimiento.
f) Proceso pruebas de los modelos, en esta etapa se utilizan los modelos entrenados para realizar predicciones con los datos de pruebas donde se usa el 10% de los datos.
g) Comparación y análisis de los resultados obtenidos con cada uno de los modelos de machine learning y se selecciona el más eficiente y con menor margen de error.

Para la partición de los datos de entrenamiento y validación se utilizó el método train_test_split de Scikit-learn el cual realiza de manera aleatoria por defecto.

Para la implementación de los modelos de machine learning se utilizó el lenguaje de programación python v3.10.12 junto a las librerías sklearn v1.2.2, tensorflow v2.15.0 y keras v2.15.0. Asimismo, para el análisis y visualización de datos pandas v2.0.3 numpy v1.25.2, seaborn v0.13.1 y matplotlib v3.7.1.

3.4 Población y muestra

Los datos han sido recolectados en 3 laboratorios de suelos: Obregon S.C.R.L, Geolef E.I.R.L y Ecx ingenieros, ubicados en la ciudad de Abancay y también de expedientes técnicos de obras de pavimentos y carreteras del departamento de Apurímac. Se obtuvieron registros históricos de estudios de mecánica de suelos, los valores de las 8 características del suelo fueron registradas manualmente a partir de los informes de laboratorio de cada muestra de suelo, alcanzando a recopilar un total de 310 registros esto debido a la disponibilidad limitada de datos. Donde 7 columnas son características del suelo que intervienen en la capacidad de carga del

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 1 (2024) - publicado:25/07/2024
DOI https://doi.org/10.57166/riqchary.v6.n1.2024.121

suelo asimismo como variable dependiente el valor del soporte de california (CBR).

Las variables independientes se obtuvieron de los ensayos granulométricos (porcentaje de grava, porcentaje de finos), de los límites de Atterberg (límite líquido, límite plástico e índice de plasticidad) y del ensayo del Proctor Modificado (óptimo contenido de humedad y máxima densidad seca) y la columna objetivo con el valor del CBR al 100%. Del total de datos, el 70% se usa para el entrenamiento y el 20% se usa para evaluar la eficiencia de los modelos y 10% para las pruebas. Para el proceso de entrenamiento se obtuvo una muestra de 217 registros, 62 registros para la validación y 31 para las pruebas. Este muestreo fue de forma aleatoria.

Fig. 1. Imágenes tomadas de los ensayos de laboratorio

La tabla 1 muestra las variables con los 10 primeros registros de la base de datos considerando las primeras 7 características del suelo como los datos de entrada y el CBR al 100% como dato de salida. Siendo la Grava el porcentaje de partículas gruesas del suelo, Finos el porcentaje de partículas del suelo con diámetro menor como arcilla y limos, Óptimo Contenido de Humedad(OCH) el porcentaje de agua en el suelo que permite alcanzar la máxima densidad durante el proceso de compactación, Límite Líquido(LL) el porcentaje de agua contenido en el suelo donde pasa de comportamiento plástico a líquido, Límite Plástico(LP) el porcentaje de agua donde el suelo deja de ser plástico y comienza a comportarse como un material quebradizo, **IP (%)**: Índice Plástico que indica la plasticidad del suelo, Densidad Máxima Seca(DMS) es la mayor densidad que un suelo puede alcanzar cuando se compacta y el California Bearing Ratio(CBR al 100%) que expresa la medida de la capacidad de carga del suelo al ser sometido a presión, expresada al 100% de compactación. Donde se observan que todos los valores de las variables independientes son datos de tipo cuantitativos continuos, todos se miden en porcentajes a excepción de la densidad máxima seca (MDS) que se mide en gramos por centímetro cúbico. Mientras que la variable dependiente (CBR) se mide en porcentaje y también los valores son de tipo cuantitativos continuos, mientras más alto el valor del CBR indica mayor capacidad de soporte.

TABLA 1
Registro de las 10 primeras observaciones del dataset

Grava (%)	Finos (%)	OCH (%)	LL (%)	LP (%)	IP (%)	MDS (gr/cm³)	CBR al 100%
57.5	42.5	7.3	30.5	22.7	7.8	2.117	38.1
22.6	77.4	10.2	29.8	17.7	12.1	1.928	17.4
52.6	47.4	7.1	25.5	19.1	6.4	2.165	34.3
34.1	65.9	12.2	24.1	16.7	7.4	1.975	14.1
40.9	59.1	7.7	23.1	18.2	4.9	2.113	38.4
8	92	5.7	31	22.4	8.6	1.681	8.5
67.7	32.3	5.2	32.8	24.5	8.3	2.234	45.1
36.9	63.1	7	34	26.5	7.5	2.095	39.7
0	100	16.3	39.1	31.7	7.4	1.696	6.8
43.2	56.8	5.4	34.2	24.4	9.7	2.143	32.4

3.5 Materiales e Instrumentos

Dentro de los materiales utilizados tenemos:

a) Hojas de calculo Excel.
b) Google Colab.

Dentro de los instrumentos utilizados están:

a) Registro de los resultados de informes de ensayos de laboratorios de estudios de mecánica de suelos en formato .xlsx y .csv.
b) Cuaderno de notebooks con extensión .ipynb.

4 Resultados y Discusión

4.1 Analisis exploratorio de datos(EDA)

La Fig. 2 muestra el mapa de calor, en el cual se puede observar la correlación entre variables, donde podemos observar que la característica de Grava tiene una alta correlación positiva con el CBR al 100%, por lo que se puede inferir que a medida que el porcentaje de Grava incrementa también se incrementa el valor del CBR al 100%; asimismo, con la densidad máxima seca (MDS) que también tiene una alta correlación positiva con el CBR al 100%. Por otro lado, el porcentaje de Finos, el óptimo contenido de humedad (OCH) y el Límite Líquido (LL) tienen una correlación negativa con el CBR al 100% con lo que se puede inferir que a medida que estos valores bajan también baja el CBR al 100%. En cuanto al límite plástico (LP) y el índice de plasticidad (IP) tienen una correlación baja, por lo que podemos inferir que tienen menor influencia en el CBR al 100%.

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 1 (2024) - publicado:25/07/2024
DOI https://doi.org/10.57166/riqchary.v6.n1.2024.121

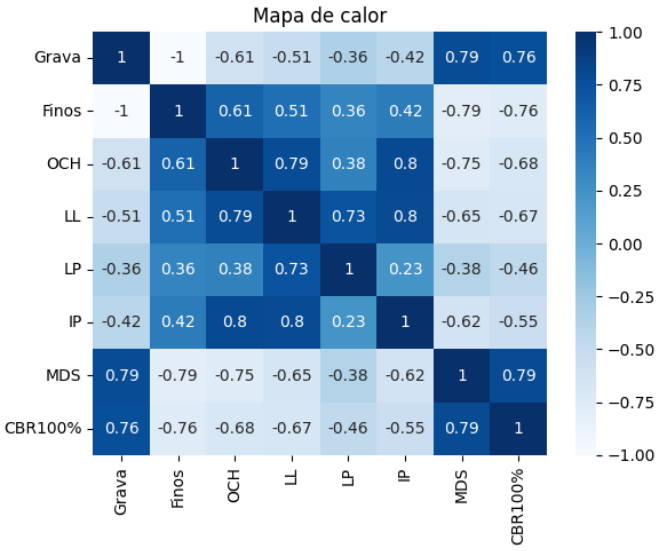


Fig. 2. Mapa de Calor (Heatmap).

La Fig. 3 muestra la distribución de la variable dependiente(CBR al 100%) en el cual se puede observar que tenemos valores atípicos o outliers a partir del valor 50, en total se presentaron 4 registros del total de las observaciones, y revisando los informes de los ensayos de laboratorio concluimos que estos datos corresponden a muestras con algún adherido o añadidos para el mejoramiento del CBR por lo que se tomó la decisión de remover estos registros ya que son irrelevantes para este estudio porque estamos tomando muestras de suelos naturales.

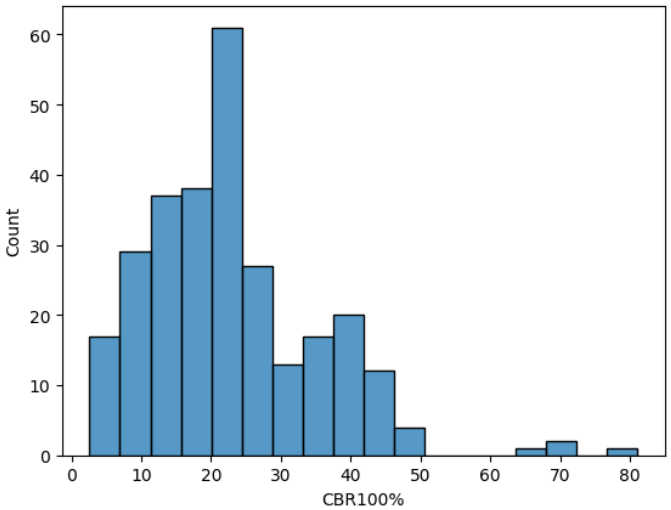


Fig. 3. Distribución del CBR al 100%

Como la variable dependiente(CBR) es un valor continuo el problema es de regresión por lo que no se aplico técnicas de balanceo de datos solo eliminación de valores atipicos. Para el caso de las variables independientes no se contaban con valores atipicos relevantes pero si se aplicó normalización optando el método estándar scaler para uniformizar la escala de datos.

4.2 Pruebas y resultados para el modelo de máquinas de vectores de soporte (SVM)

El modelo SVM implementado emplea un kernel lineal para predecir valores del CBR100% utilizando características geotécnicas estandarizadas mediante StandardScaler, excluyendo el índice de plasticidad (IP) del análisis. La arquitectura incorpora una división de datos (entrenamiento/prueba) con inicialización de semilla aleatoria.

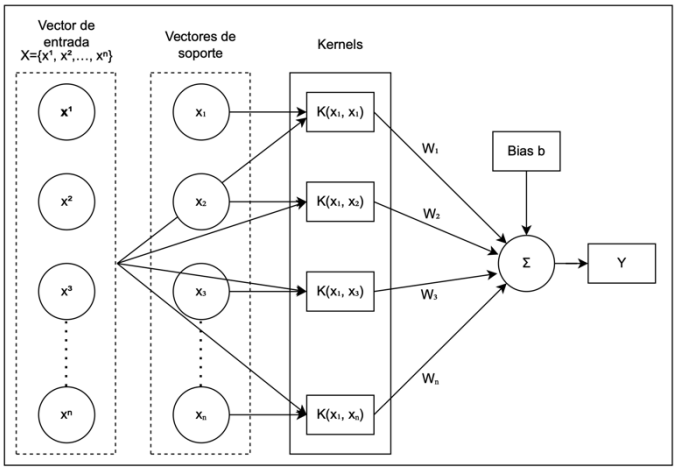


Fig. 4. Diagrama de Arquitectura del modelo de máquina de vectores de soporte

La tabla 2 muestra las métricas para el modelo de Máquinas de Vectores de Soporte, en la cual se puede observar que el modelo SVM tiene un buen rendimiento en la predicción del CBR al 100%, con un R^2 superior a 0.7, lo que sugiere que el modelo es capaz de capturar una parte significativa de la variabilidad en los datos.

El R^2 de 0.701 sugiere que el modelo explica aproximadamente el 70.15% de la variabilidad en los valores del CBR basados en las características del suelo. Este es un indicador positivo del rendimiento del modelo, aunque también muestra que hay un 29.85% de la variabilidad que no está siendo capturada. El MSE de 36.83 indica el promedio de los cuadrados de los errores. El RMSE de 6.07 indica que en promedio las predicciones del modelo tienen un error de aproximadamente 6.07 unidades del CBR al 100%. El MAE de 4.41 indica que, en promedio, las predicciones del modelo están desviadas en 4.41 unidades del CBR al 100%.

TABLA 2
Métricas para el modelo SVM

Modelo de machine learning	R^2	MSE	RMSE	MAE
Máquinas de vectores de soporte (SVM)	0.701	36.834	6.069	4.409

La Fig. 5 muestra el gráfico de dispersión de modelo para la

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 1 (2024) - publicado:25/07/2024
DOI https://doi.org/10.57166/riqchary.v6.n1.2024.121

data de entrenamiento de las predicciones versus los valores reales del modelo.

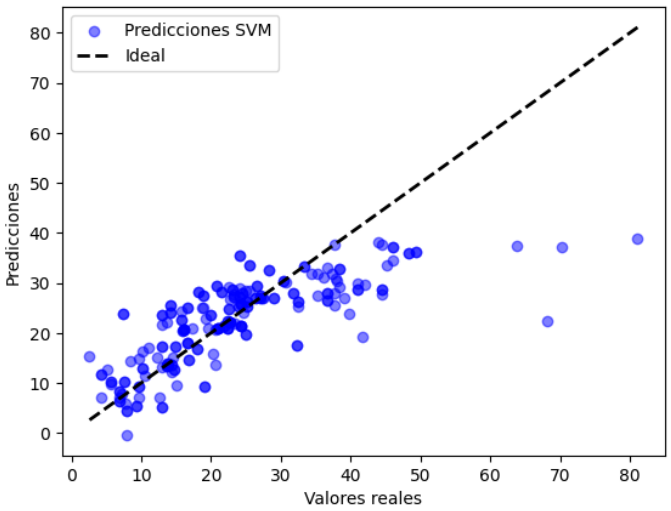


Fig. 5. Gráficos de dispersión del modelo de SVM para la data de entrenamiento

La Fig. 6 muestra el gráfico de dispersión de modelo para la data de validación de las predicciones versus los valores reales del modelo.

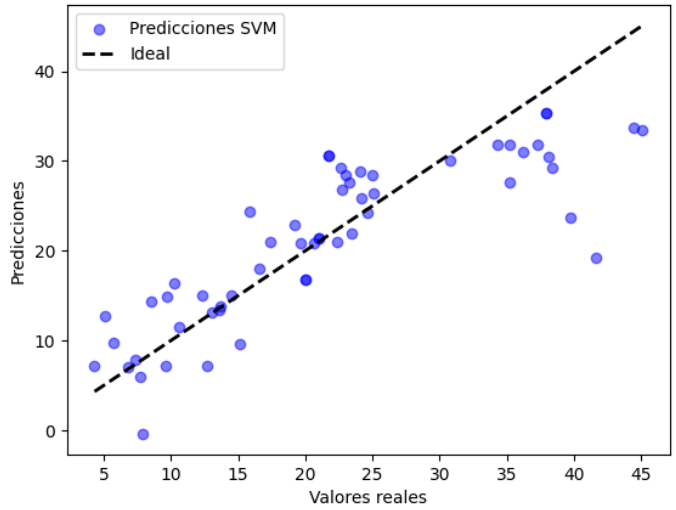


Fig. 6. Gráficos de dispersión del modelo de SVM para la data de validación.

4.3 Pruebas y resultados para el modelo con redes neuronales profundas(DNN)

El modelo de Red Neuronal Profunda (DNN) implementado para predecir CBR100% elimina valores atípicos de la variable objetivo y del Índice de Plasticidad (IP) antes del procesamiento, utilizando una arquitectura secuencial con tres capas: dos capas ocultas de 64 neuronas cada una con activación ReLU, y una capa de salida lineal para regresión

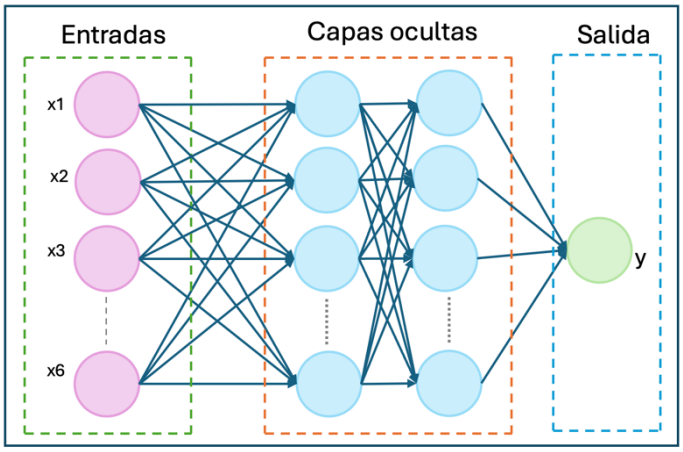


Fig. 7. Diagrama de Arquitectura del modelo de redes neuronales profundas

La tabla 3 muestra las métricas para el modelo de redes neuronales profundas, en la cual se puede observar que el modelo tiene un buen rendimiento en la predicción del CBR al 100%, con un R^2 superior a 0.784, lo que sugiere que el modelo es capaz de capturar una parte significativa de la variabilidad en los datos. Estas métricas sugieren que el modelo DNN tiene un rendimiento bastante bueno para predecir el CBR basado en las características del suelo. El valor de R^2 de 0.7875 indica que el modelo explica bien la variabilidad en los datos. Tanto el RMSE de 5.12 como el MAE de 3.76 son razonablemente bajos, lo que sugiere que el modelo tiene un error de predicción aceptable. Adicionalmente, en las pruebas se pudo observar que para el entrenamiento del modelo no afecta significativamente tener los datos normalizados, asimismo, el excluir la característica del índice de plasticidad (IP) tuvo un impacto positivo en el rendimiento del modelo.

TABLA 3
Métricas para el modelo DNN

Modelo de machine learning	R^2	MSE	RMSE	MAE
Redes neuronales profundas (DNN)	0.784	26.606	5.158	3.743

La Fig. 8 muestra el gráfico de dispersión de modelo para la data de entrenamiento de las predicciones versus los valores reales del modelo.

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 1 (2024) - publicado:25/07/2024
DOI https://doi.org/10.57166/riqchary.v6.n1.2024.121

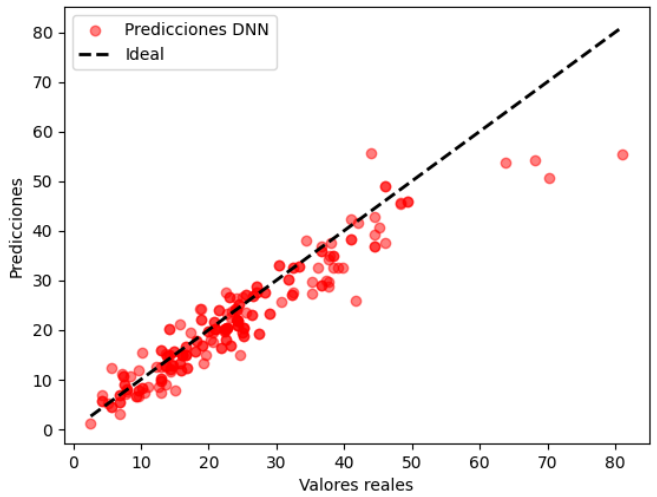


Fig. 8. Gráficos de dispersión del modelo de DNN para la data de entrenamiento

La Fig. 9 muestra el gráfico de dispersión de modelo para la data de validación de las predicciones versus los valores reales del modelo.

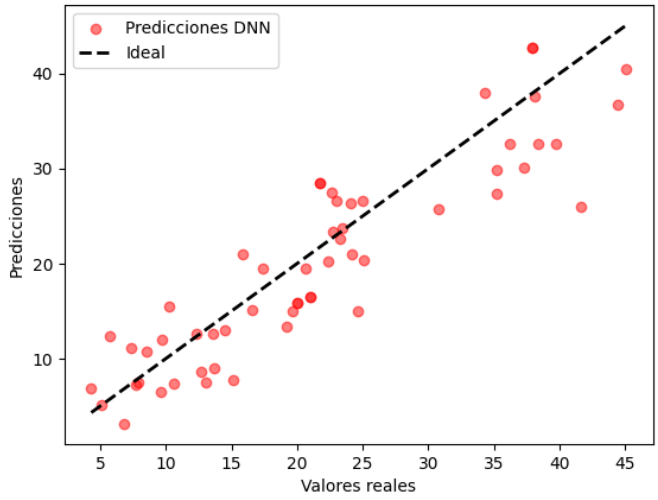


Fig. 9. Gráficos de dispersión del modelo de DNN para la data de validación

4.4 Pruebas y resultados para el modelo con Árboles de decisión(Decision tree)

El modelo de Árbol de Decisión para regresión (DecisionTreeRegressor) incorpora un preprocesamiento que elimina valores atípicos de la variable objetivo (CBR100%), Límite Líquido (LL) e Índice de Plasticidad (IP). Utiliza como predictores todas las características disponibles excepto el IP, con semilla aleatoria fija para reproducibilidad. El árbol emplea el criterio 'best' el cual selecciona la mejor división posible en cada nodo.

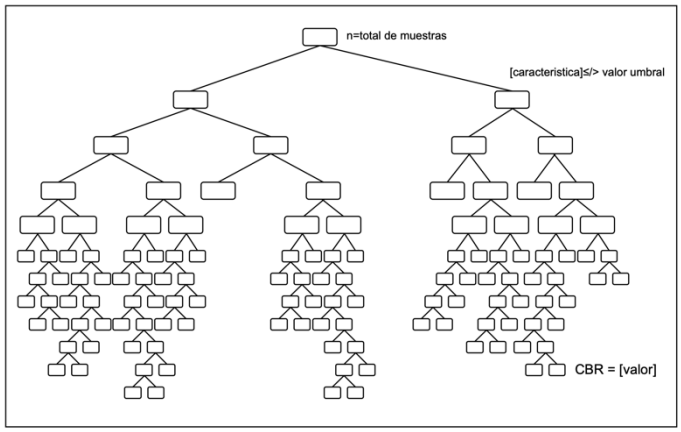


Fig. 10. Diagrama de Arquitectura del modelo de árboles de decisión

La tabla 4 muestra las métricas para el modelo de árboles de decisión, en la cual se puede observar que el modelo tiene un buen rendimiento en la predicción del CBR al 100%. Este valor indica que el modelo explica aproximadamente el 93.07% de la variabilidad en los valores de CBR al 100%. El valor de R^2 cercano a 1 sugiere que el modelo tiene un muy buen ajuste, capturando la mayor parte de la variabilidad en los datos. En cuanto a la importancia de las características, el orden es el siguiente: la densidad máxima seca (MDS) contribuye aproximadamente en 53.5.% a las decisiones del modelo, el porcentaje de finos es la segunda más importante, con una contribución del 23.1%, el límite plástico (LP) tiene una importancia del 10.6%, El óptimo contenido de humedad (OCH) tiene una importancia del 9.0%, aunque menos influyente que las tres primeras características, OCH sigue siendo relevante en la predicción del CBR al 100% y la característica Grava tiene la menor importancia, con una contribución del 3.8%. Asimismo, se excluyó el límite líquido (LL) y el índice de plasticidad (IP) porque influyen en la mejora del rendimiento del modelo. El número de nodos del árbol es 393, que nos indica la complejidad del modelo que está capturando detalles en los datos. Las métricas de error (MAE, RMSE y MSE) son relativamente bajas, sugiriendo que el modelo tiene una buena precisión.

TABLA 4
Métricas para el modelo árboles de decisión

Modelo de machine learning	R^2	MSE	RMSE	MAE
árboles de decisión	0.9307	9.199	3.033	1.216

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 1 (2024) - publicado:25/07/2024
DOI https://doi.org/10.57166/riqchary.v6.n1.2024.121

La Fig. 11 muestra el gráfico de dispersión de modelo para la data de entrenamiento de las predicciones versus los valores reales del modelo.

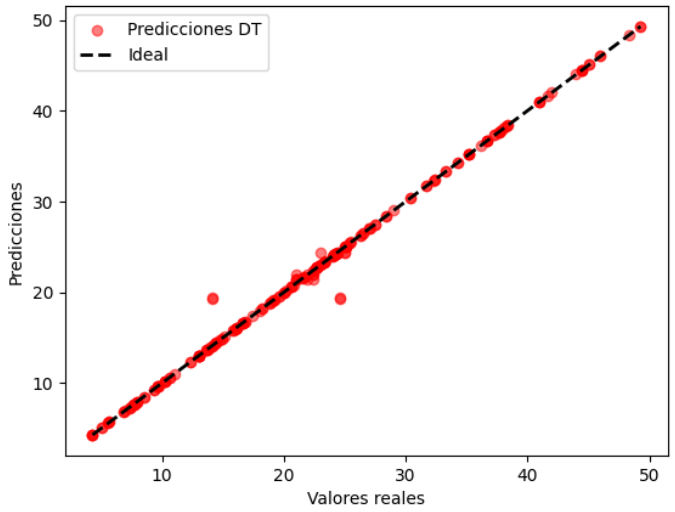


Fig. 11. Gráficos de dispersión del modelo árboles de decisión para la data de entrenamiento

La Fig. 12 muestra el gráfico de dispersión de modelo para la data de validación de las predicciones versus los valores reales del modelo.

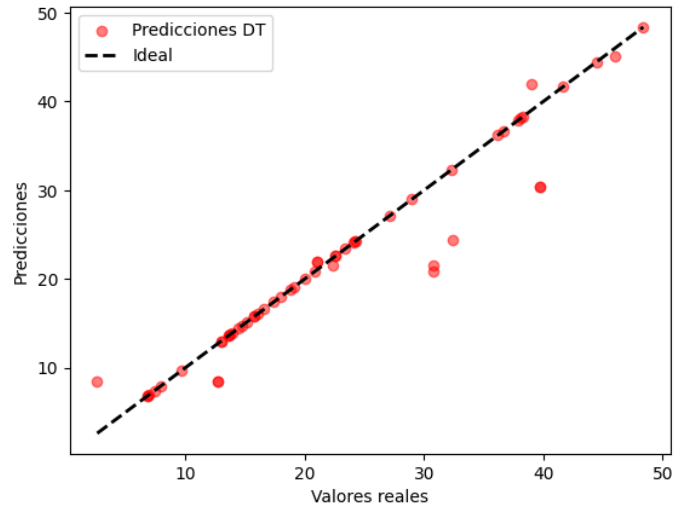


Fig. 12. Gráficos de dispersión del modelo de árboles de decisión para la data de validación

4.5 Discusión

En este estudio, los árboles de decisión (DT) demostraron ser el modelo más efectivo para predecir el California Bearing Ratio (CBR) basado en las características del suelo, alcanzando un R^2 de 0.9307 y superando a las redes neuronales profundas (DNN) y las máquinas de vectores de soporte (SVM). Estos resultados son consistentes con investigaciones previas, como la de Frank y Heber [13], quienes también obtuvieron un R^2 de 0.89 al predecir el CBR al 100% utilizando redes neuronales. Sin embargo, los árboles de decisión destacan por su menor complejidad y facilidad de interpretación, lo que los hace ideales para aplicaciones geotécnicas.

Además, estudios como los de Marisabel y Rodrigo [15], que aplicaron redes neuronales para predecir propiedades de resistencia al corte, y el trabajo de Johan y José [12] sobre suelos tropicales, validan la eficacia de las técnicas de machine learning en la predicción de propiedades del suelo. Este análisis reafirma que los árboles de decisión no solo ofrecen un alto rendimiento, sino que también son herramientas prácticas y accesibles para abordar problemas complejos en la ingeniería geotécnica.

La tabla 5 muestra la comparación de métricas de los modelos El modelo de árboles de decisión no solo ofrece el mejor coeficiente de determinación (R^2), sino que también tiene los valores más bajos de MSE, RMSE y MAE, lo que indica una menor variabilidad del error y una mayor precisión en las predicciones.

TABLA 5
Comparación de los modelos

Modelo de machine learning	R^2	MSE	RMSE	MAE
Máquinas de Vectores de Soporte (SVM)	0.701	36.834	6.069	4.409
Redes neuronales profundas (DNN)	0.784	26.606	5.158	3.743
árboles de decisión (DT)	0.9307	9.199	3.033	1.216

5 Conclusiones

En esta investigación se corrobora que las propiedades físicas del suelo pueden ser utilizadas para predecir otras características como el CBR, puesto que, utilizando variables como el porcentaje de grava, el porcentaje de finos, el óptimo contenido de humedad, la densidad máxima seca, el límite líquido, el límite plástico, el índice de plasticidad del suelo y aplicando modelos de machine learning se puede predecir el índice de CBR al 100%. Asimismo, a pesar del tamaño limitado de muestras, los hallazgos son significativos para proporcionar una base sólida para futuras investigaciones.

Adicionalmente, este trabajo de investigación muestra que, el algoritmo más eficiente para la predicción del CBR al 100% basádonos en el coeficiente de determinación y las métricas de error es el modelo de árboles de decisión, seguido de redes neuronales profundas (DNN) y finalmente el modelo de máquinas de vectores de soporte de regresión (SVM).

6 Trabajos futuros

A partir de los resultados de esta investigación se puede ver el potencial que tiene el desarrollo de modelos de machine

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 1 (2024) - publicado:25/07/2024
DOI https://doi.org/10.57166/riqchary.v6.n1.2024.121

learning para la predicción de propiedades del suelo, por lo cual se espera que en trabajos futuros se puedan mejorar los modelos propuestos, incorporando características adicionales del suelo, como propiedades de compactación y características de deformación. Asimismo, se podrían aplicar y experimentar diferentes algoritmos como random forest. También se puede optar por el desarrollo de aplicaciones web que permitan a los profesionales predecir el CBR en tiempo real utilizando modelos de machine learning entrenados. Adicionalmente, se puede incorporar la automatización con el uso de sensores y dispositivos IoT para la recolección automática de datos del suelo. También sería genial fomentar la colaboración entre ingenieros civiles, geotécnicos, científicos de datos para desarrollar modelos más integrales y precisos, asimismo desarrollar programas de educación y capacitación para ingenieros sobre el uso de técnicas avanzadas de machine learning aplicadas a la geotecnia.

AGRADECIMIENTOS

Se agradece la colaboración de los ingenieros del grupo corporativo Obregon S.C.R.L por sus conocimientos compartidos durante el desarrollo de la investigación. Agradecer también a los laboratorios de suelos y concreto Geolef E.I.R.L y Ecx ingenieros por el apoyo en la recolección de datos.

REFERENCIAS

[1] J. R. Lópe, "Herramientas digitales y uso de inteligencia artificial en geotecnia: Un enfoque en la evaluación de taludes con Matlab," no. 29 de enero de 2024, 2024. https://doi.org/10.56712/latam.v5i1.1640

[2] D. K. Talukdar, "A Study of Correlation Between California Bearing Ratio (CBR) Value With Other Properties of Soil," vol. 4, India, 2014.

[3] J. V.-L. J. E. A.-B. K. F.-F. C. L. C.-E. K. M. M.-A. E. N. L.-M. M. T. &. D. D. PERRET, Desarrollo de métodos de análisis de espectroscopia y algoritmos de aprendizaje automático para la evaluación de algunas propiedades del suelo en Costa Rica. Agronomía Costarricense, Universidad de Costa Rica. Colegio de Ingenieros y Agrónomos. Ministerio de Agricultura y Ganadería, 2022.

[4] D. L. COBA, M. HERRERA SUAREZ, M. M. GARCIA LORENZO and R. y BELTRAN, Modelo computacional para la estimación de la densidad del suelo a través del sensoramiento continuo. Revista Ciencia y Técnica Agrícola, vol. 27, 2018.

[5] V. C. LOME, "Análisis fotogramétrico de nube de puntos y aprendizaje automático como herramientas útiles en la caracterización de macizos rocosos," Puebla, Benemérita Universidad Autónoma de Puebla, 2023.

[6] M. J. CCASANI and Y. I. FERRO, "Evaluación y análisis de pavimentos en la ciudad de Abancay, para proponer una mejor alternativa estructural en el diseño de pavimentos," Abancay, Universidad Tecnológica de los Andes, 2017.

[7] M.T.C, "MANUAL DE CARRETERAS: DISEÑO GEOMÉTRICO DG – 2018," Lima, Ministerio de Transportes y Comunicaciones, 2018.

[8] M.T.C, "MANUAL DE ENSAYO DE MATERIALES," Lima, Ministerio de Transportes y Comunicaciones, 2016.

[9] H. O. Becerra, "ANÁLISIS Y ESTUDIOS DE SUELOS Y SU APLICACIÓN PARA EL MEJORAMIENTO DEL TRAMO 19 DE UNA CARRETERA EN LA PROVINCIA DE CORONEL PORTILLO UCAYALI 2018," tesis de grado, Lima, 2020.

[10] J. W. Tukey, Exploratory Data Analysis, Massachusetts: Addison-Wesley, 1977.

[11] G. James, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, Springer, 2013. https://doi.org/10.1007/978-1-4614-7138-7

[12] J. E. V.-L. K. A.-B. Johan Perret, Desarrollo de métodos de análisis de espectroscopia y algoritmos de aprendizaje automático para la evaluación de algunas propiedades del suelo en Costa Rica, 2019.

[13] F. J. Valderrama, H. I. Mejía, S. P. Muñoz Pérez and V. Tuesta, "Desarrollo de un modelo predictivo de las propiedades mecánicas del suelo usando redes neuronales artificiales," 2021.

[14] A. R. Paucar and S. Esteban, "Parámetros de resistencia al corte del suelo en función a sus propiedades físicas, empleando redes bayesianas y ensayo triaxial-Callqui Grande," 2023.

[15] M. M. Boza Capani y R. O. Merino Ortiz, "PARÁMETROS DE RESISTENCIA AL CORTE DE SUELOS A PARTIR DE SUS PROPIEDADES FÍSICAS, UTILIZANDO REDES NEURONALES ARTIFICIALES Y EQUIPO TRIAXIAL," tesis de grado, Universidad Nacional de Huancavelica, 2018.

[16] K. TERZAGHI, R. B. PECK and G. MESRI, Soil mechanics in engineering practice., John wiley & sons, 1996.

BIOGRAFÍA

Flor de Cantuta Tello Sarmiento, bachiller en Ingeniería Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac, actualmente es software engineer en la empresa Rappi S.A.C.

Manuel Jesús Ibarra Cabrera, doctor en ciencias de la computación e investigador en las áreas de ingeniería de software, serious game, informática educativa, computación móvil, IoT e industria y sociedad. Docente universitario de pre y pos grado en la Universidad Nacional Micaela Bastidas de Apurimac, Universidad Tecnológica de los Andes, Universidad Nacional del Altiplano y la Universidad Nacional Mayor de San Marcos. Actualmente, es miembro activo de conferencias importantes en latinoamerica: LACLO, CLEI, CONTIE, HCI, SABTIC, CISTI, Decisioning y otros.