UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS

FACULTAD DE INGENIERÍA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Presentado por

BACH. FRANKLIN DANIEL CARBAJAL OSIS

EFICIENCIA DE UN MODELO PREDICTIVO DE LAS PROPIEDADES MECÁNICAS DE LOS SUELOS BASADO EN TÉCNICAS DE MACHINE LEARNING

Asesor:
M.SC. RICHARD CARRIÓN ABOLLANEDA

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

ANDAHUAYLAS – APURIMAC – PERÚ

2024

APROBACIÓN DEL ASESOR

Quién suscribe:

MSc. Ing. Richard Carrión Abollaneda, por la presente:

CERTIFICA,

Que, el Bachiller en Ingeniería de Sistemas, **FRANKLIN DANIEL CARBAJAL OSIS** haculminado satisfactoriamente el Informe Final intitulado: *"***EFICIENCIA DE UN MODELO PREDICTIVO DE LAS PROPIEDADES MECÁNICAS DE LOS SUELOS BASADO EN TÉCNICAS DE MACHINE LEARNING***"* para optar el Título Profesional de Ingeniero de Sistemas.

Andahuaylas, 02 de mayo de 2024.

MSc. Richard Carrión Abollaneda
Asesor

Bach. Franklin Daniel Carbajal Osis
Tesista

UNIVERSIDAD NACIONAL
JOSÉ MARÍA ARGUEDAS
Licenciada por la Sunedu

ACTA DE SUSTENTACIÓN DE TESIS

En la Av. José María Arguedas del Local Académico SL01 (Ccoyahuacho) en el auditorio de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional José María Arguedas ubicado en el distrito de San Jerónimo de la Provincia de Andahuaylas, siendo las 11:00 horas del día viernes 28 de junio del año 2024, se reunieron los docentes: Dr. Enrique Edgardo Condor Tinoco, M.Sc. Neptalí Menejes Palomino, M.Sc. Yovana Flores Ccorisapra, en condición de integrantes del Jurado Evaluador del Informe Final de Tesis intitulado: "EFICIENCIA DE UN MODELO PREDICTIVO DE LAS PROPIEDADES MECÁNICAS DE LOS SUELOS BASADO EN TÉCNICAS DE MACHINE LEARNING", cuyo autor es el Bachiller en Ingeniería de Sistemas **FRANKLIN DANIEL CARBAJAL OSIS** y el asesor M.Sc. Richard Carrión Abollaneda, con el propósito de proceder a la sustentación y defensa de dicha tesis.

Luego de la sustentación y defensa de la tesis, el Jurado Evaluador **ACORDÓ**: APROBAR por UNANIMIDAD al Bachiller en Ingeniería de Sistemas **FRANKLIN DANIEL CARBAJAL OSIS,** obteniendo la siguiente calificación y mención:

Nota escala vigesimal		Mención
Números	Letras	
16	DIECISÉIS	MUY BUENO

En señal de conformidad, se procedió a la firma de la presente acta en 03 ejemplares.

.....
Dr. Enrique Edgardo Condor Tinoco
Presidente del Jurado Evaluador

.....
M.Sc. Neptalí Menejes Palomino
Primer Miembro del Jurado Evaluador

.....
M.Sc. Yovana Flores Ccorisapra
Segundo Miembro del Jurado Evaluador

ISO 9001:2015
ISO 21001:2018

ANEXO 31

APROBACIÓN DEL JURADO DICTAMINADOR

LA TESIS: **"EFICIENCIA DE UN MODELO PREDICTIVO DE LAS PROPIEDADES MECÁNICAS DE LOS SUELOS BASADO EN TÉCNICAS DE MACHINE LEARNING"**; para optar el Título Profesional de Ingeniero de Sistemas, ha sido evaluada por el Jurado Dictaminador conformado por:

PRESIDENTE: Dr. Enrique Edgardo Condor Tinoco
PRIMER MIEMBRO: M.Sc. Neptalí Menejes Palomino
SEGUNDO MIEMBRO: M.Sc. Yovana Flores Ccorisapra

Habiendo sido aprobado por UNANIMIDAD/MAYORIA, en la ciudad de Andahuaylas el día 28 del mes de junio de 2024

Andahuaylas, 01 de julio de 2024.

Dr. Enrique Edgardo Condor Tinoco
PRESIDENTE DEL JURADO DICTAMINADOR

M.Sc. Neptalí Menejes Palomino
PRIMER MIEMBRO DEL JURADO DICTAMINADOR

M.Sc. Yovana Flores Ccorisapra
SEGUNDO MIEMBRO DEL JURADO DICTAMINADOR

Universidad Nacional José María Arguedas

Identidad y Excelencia para el Trabajo Productivo y el Desarrollo

Unidad de Investigación de la Facultad Ingeniería
C018-2024

Andahuaylas, 08 de julio de 2024

La Unidad de Investigación de la Facultad de Ingeniería, expide la:

Constancia

De porcentaje de similitud (13%) según el software Turnitin, al informe final de investigación: Eficiencia de un modelo predictivo de las propiedades mecánicas de los suelos basado en técnicas de Machine Learning. Presentado por el **Bach. Franklin Daniel Carbajal Osis** cuyo Asesor es el M.Sc. Richard Carrión Abollaneda.

Dra. María del Carmen Delgado Laime
Presidente de la Unidad de Investigación de la
Facultad de Ingeniería

MSc. Fidelia Tapia Tadeo
Miembro de la
Unidad de Investigación de la Facultad de Ingeniería

M.Sc. Richard Carrión Abollaneda
Miembro de la
Unidad de Investigación de la Facultad de Ingeniería

C.c
Archivo.

NOMBRE DEL TRABAJO

EFICIENCIA DE UN MODELO PREDICTIVO DE LAS PROPIEDADES MECÁNICAS DE LOS SUELOS BASADO EN TÉCNICAS DE

AUTOR

FRANKLIN DANIEL CARBAJAL OSIS

RECUENTO DE PALABRAS

46322 Words

RECUENTO DE CARACTERES

266659 Characters

RECUENTO DE PÁGINAS

219 Pages

TAMAÑO DEL ARCHIVO

10.2MB

FECHA DE ENTREGA

Jul 2, 2024 6:14 AM GMT-5

FECHA DEL INFORME

Jul 2, 2024 6:17 AM GMT-5

13% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base de datos.

- 12% Base de datos de Internet
- 2% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Crossref

Excluir del Reporte de Similitud

- Material bibliográfico
- Material citado
- Material citado
- Coincidencia baja (menos de 10 palabras)

UNIVERSIDAD NACIONAL JOSE MARIA ARGUEDAS

Ing. Richard Carrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Osis
DNI: 72680375

13% de similitud general

Principales fuentes encontradas en las siguientes bases de datos:

- 12% Base de datos de Internet
- 2% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Crossref

FUENTES PRINCIPALES

Las fuentes con el mayor número de coincidencias dentro de la entrega. Las fuentes superpuestas no se mostrarán.

1	hdl.handle.net Internet	1%
2	repositorio.unajma.edu.pe Internet	1%
3	repositorio.uss.edu.pe Internet	1%
4	inecon.org Internet	1%
5	core.ac.uk Internet	<1%
6	gredos.usal.es Internet	<1%
7	repositorio.urp.edu.pe Internet	<1%
8	secgen.unajma.edu.pe Internet	<1%

UNIVERSIDAD NACIONAL JOSE MARIA ARGUEDAS
Ing. Richard Cabrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Osis
DNI: 72680375

9	pirhua.udep.edu.pe Internet	<1%
10	intellectum.unisabana.edu.co Internet	<1%
11	upc.aws.openrepository.com Internet	<1%
12	repositorio.ucv.edu.pe Internet	<1%
13	diariocorreo.pe Internet	<1%
14	repositorio.ual.edu.pe Internet	<1%
15	repositorio.une.edu.pe Internet	<1%
16	coursehero.com Internet	<1%
17	gob.pe Internet	<1%
18	repositorio.unsaac.edu.pe Internet	<1%
19	repositorio.upla.edu.pe Internet	<1%
20	scribd.com Internet	<1%

UNIVERSIDAD NACIONAL JOSE MARIA ARGUEDAS

Ing. Richard Carrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Osis
DNI : 72680375

21	slideshare.net Internet	<1%
22	github.com Internet	<1%
23	jalayo.blogspot.com Internet	<1%
24	rua.ua.es Internet	<1%
25	gmsc-uce.blogspot.com Internet	<1%
26	repository.usta.edu.co Internet	<1%
27	vdocuments.mx Internet	<1%
28	fastercapital.com Internet	<1%
29	repositorio.unc.edu.pe Internet	<1%
30	repositorio.unprg.edu.pe Internet	<1%
31	repositorio.unphu.edu.do Internet	<1%
32	andina.pe Internet	<1%

JOSE MARIA ARGUEDAS

Ing. Richard Carrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Osis
DNI : 72600375

33	preview.marsh.com Internet	<1%
34	repositorio.upt.edu.pe Internet	<1%
35	repositorio.uladech.edu.pe Internet	<1%
36	list.waikato.ac.nz Internet	<1%
37	es.scribd.com Internet	<1%
38	larepublica.pe Internet	<1%
39	repositorio.uniremington.edu.co Internet	<1%
40	repositorioacademico.upc.edu.pe Internet	<1%
41	biblioteca.usac.edu.gt Internet	<1%
42	dspace.unl.edu.ec Internet	<1%
43	repositorio.uap.edu.pe Internet	<1%
44	repositorio.udec.cl Internet	<1%

UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS
Ing. Richard Carrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Oris
DNI: 72680375

45	repositorio.ug.edu.ec Internet	<1%
46	rstudio-pubs-static.s3.amazonaws.com Internet	<1%
47	documentop.com Internet	<1%
48	guiassuelos.blogspot.com Internet	<1%
49	repositorio.unab.edu.pe Internet	<1%
50	hindawi.com Internet	<1%
51	detodopython.com Internet	<1%
52	César Aristóteles Yajure Ramírez. "Pronóstico de consumo de energía ... Crossref	<1%
53	"Informatyka i zarządzanie na przełomie wieków. Metody, narzędzia, s... Crossref	<1%
54	es.sciencewal.com Internet	<1%
55	repositorio.uisrael.edu.ec Internet	<1%
56	repositorio.unitec.edu Internet	<1%

UNIVERSIDAD NACIONAL JOSE MARIA ARGUEDAS
Ing. Richard Currión Abollaneda
DOCENTE

Franklin Daniel Cortéjad Osis
DNI: 72680375

57	doku.pub Internet	<1%
58	fr.slideshare.net Internet	<1%
59	oposinet.com Internet	<1%
60	siftdesk.org Internet	<1%
61	virtualpro.co Internet	<1%
62	openwebinars.net Internet	<1%
63	qdoc.tips Internet	<1%
64	repository.pedagogica.edu.co Internet	<1%
65	tesla.puertomaderoeditorial.com.ar Internet	<1%
66	José Francisco Pedrero Sánchez. "Desarrollo de procedimientos de val... Crossref posted content	<1%
67	Luis Felipe Lalinde Castrillón. "Estudio de compuestos de GRC y adicio... Crossref posted content	<1%
68	Manh Duc Nguyen, Binh Thai Pham, Tran Thi Tuyen, Hoang Phan Hai Y... Crossref	<1%

UNIVERSIDAD ADVENTISTA JOSÉ MARÍA SANDOVAL

Ing. Richard Garrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Dias
DNI: 72680375

69	cjascience.com Internet	<1%
70	pdfcoffee.com Internet	<1%
71	repositorio.espe.edu.ec Internet	<1%
72	upo.es Internet	<1%
73	goconqr.com Internet	<1%
74	Danil D. Shesternev. "Changes in physical and mechanical properties o... Crossref	<1%
75	alicia.concytec.gob.pe Internet	<1%
76	catalonica.bnc.cat Internet	<1%
77	cloud.google.com Internet	<1%
78	docplayer.es Internet	<1%
79	e-archivo.uc3m.es Internet	<1%
80	repositorio.uandina.edu.pe Internet	<1%

UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS

Ing. Richard Carrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Arias
DNI: 72680375

81	repositorio.unas.edu.pe Internet	<1%
82	repositorio.upse.edu.ec Internet	<1%
83	repositorio.utc.edu.ec Internet	<1%
84	repository.unipiloto.edu.co Internet	<1%
85	ricardollarves.com.ve Internet	<1%
86	runebook.dev Internet	<1%
87	sedici.unlp.edu.ar Internet	<1%
88	tesis.pucp.edu.pe Internet	<1%
89	vdoc.pub Internet	<1%
90	esan.edu.pe Internet	<1%
91	euskadi.net Internet	<1%
92	nature.com Internet	<1%

UNIVERSIDAD NACIONAL JOSE MARIA ARGUEDAS

Ing. Richard Carrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Osis
DNI: 72680375

93	novicell.es Internet	<1%
94	repositorio.unach.edu.pe Internet	<1%
95	researchgate.net Internet	<1%
96	Jiménez-Cortés, Rocío. "The influence of informal learning processes ... Crossref	<1%
97	Jorge Luis Andrade, Valencia. "MODELING LAPSE RATES USING MAC... Crossref	<1%
98	José Garrigues Baixauli. "Implementación de algoritmos de redes neur... Internet	<1%
99	RALPH MAC NALLY. "A Successful Predictive Model of Species Richne... Crossref	<1%
100	SILVIA SOLEDAD MORENO GUTIERREZ, MARGARITA GARCIA LOPEZ. "... Crossref	<1%
101	de.slideshare.net Internet	<1%
102	deepnote.com Internet	<1%
103	doaj.org Internet	<1%
104	es.coursera.org Internet	<1%

UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS

Ing. Richard Carrión Abollaneda
DOCENTE

Franklin Daniel Cardojal Dias
DNI: 72680375

105	negociointernacional.bancsabadell.com Internet	<1%
106	repositorio.untumbes.edu.pe Internet	<1%
107	repositorio.usmp.edu.pe Internet	<1%
108	repository.ucatolicaluisamigo.edu.co Internet	<1%
109	rraae.cedia.edu.ec Internet	<1%
110	tesis.ucsm.edu.pe Internet	<1%
111	tesis.usat.edu.pe Internet	<1%
112	dspace.espol.edu.ec Internet	<1%
113	swissinfo.org Internet	<1%
114	theinsightpartners.com Internet	<1%

UNIVERSIDAD NACIONAL JOSE MARIA ARGUEDAS

Ing. Richard Garrión Abollaneda
DOCENTE

Franklin Daniel Carbajal Osis
DNI: 72680375

UNIVERSIDAD NACIONAL
JOSÉ MARÍA ARGUEDAS
Licenciada por la Sunedu

FORMATO DE AUTORIZACIÓN DE PUBLICACIÓN EN EL REPOSITORIO INSTITUCIONAL DE LA UNAJMA

Formulario de autorización para la publicación electrónica en la página web del Repositorio Institucional de la Universidad Nacional de José María Arguedas, de conformidad con la Ley Nº 30035 Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto y Resolución de Comisión Organizadora N° 0203-2021 que aprueba el Reglamento que regula el Repositorio Institucional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto de la Universidad Nacional de José María Arguedas.

1. DATOS DEL AUTOR:

Nombres y Apellidos: Bach. Franklin Daniel Carbajal Osis

DNI / Carnet de Extranjería: 72680375

Pasaporte Nº:

Correo Institucional: 1006120182@unajma.edu.pe

Teléfono/Celular: 982076766

Correo Electrónico: danielcco.12@gmail.com

2. GRADO, TITULO O ESPECIALIZACIÓN (Marque con una X)

Bachiller	Título Profesional	Magíster	Doctor	Segunda Especialidad
X				

3. TIPO DE PUBLICACIÓN (Marque con una X)

Tesis	Trabajo Académico	Trabajo de investigación	Trabajo de Suficiencia profesional
X			

UNIVERSIDAD NACIONAL
JOSÉ MARÍA ARGUEDAS
Licenciada por la Sunedu

4. **DATOS DE LA PUBLICACION:**

Título: EFICIENCIA DE UN MODELO PREDICTIVO DE LAS PROPIEDADES MECÁNICAS DE LOS SUELOS BASADO EN TÉCNICAS DE MACHINE LEARNING

Fecha de Publicación:

Resolución de Acta de Sustentación: Nº 0365-2024-D-FI-UNAJMA

5. **DATOS DEL ASESOR:**

Nombres y Apellidos: M.Sc. Richard Carrión Abollaneda

DNI/Pasaporte/Cédula: 41877537

ORCID: 0000-0002-1875-4180

.....

Firma

DOCUMENTO DE IDENTIDAD: 72680375

NOMBRES Y APELLIDOS: Bach. Franklin Daniel Carbajal Osis

UNIVERSIDAD NACIONAL
JOSÉ MARÍA ARGUEDAS
Licenciada por la Sunedu

LICENCIAS

Licencia Estándar

Bajo los siguientes términos, autorizo el depósito de mi trabajo de investigación en el Repositorio Institucional Digital de la UNAJMA.

Con la autorización de depósito de mi trabajo de investigación otorgo a la Universidad Nacional de José María Arguedas, una Licencia No Exclusiva para reproducir. distribuir comunicar al público, transformar (únicamente mediante su traducción a otros idiomas) y poner a disposición del público mi trabajo de investigación (incluido el resumen), en formato físico o digital en cualquier medio, conocido o por conocerse, a través de los diversos servicios provistos por la Universidad, creados o por crearse, tales como el Repositorio Institucional Digital de la UNAJMA, Colección de Tesis, entre otros, en el Perú y en el extranjero, por el tiempo y veces que considere necesarias y libre de remuneraciones.

En virtud de dicha Licencia, la Universidad Nacional de José María Arguedas, podrá reproducir mi trabajo de investigación en cualquier tipo de soporte y en más de un ejemplar, sin modificar su contenido, solo con los propósitos de seguridad, respaldo y preservación.

Declaro que el trabajo de investigación es una creación de mi autoría y exclusiva titularidad, o coautoría con titularidad compartida, y me encuentro facultado a conceder la presente licencia y, asimismo, garantizo que dicho trabajo de investigación no infringe derechos de autor de terceras personas.

La Universidad Nacional de José María Arguedas consignara el nombre del/los autores/es del trabajo de investigación, y no le hará ninguna modificación más de la permitida en la presente licencia.

Autorizo el depósito (marque con una X).

(X) SI autorizo que se deposite inmediatamente.
( ) No autorizo.

Fecha: 16/07/2024

.....
Firma

Documento de identidad: 72680375

Nombres y apellidos: Bach. Franklin Daniel Carbajal Osis

DEDICATORIA

El presente trabajo de tesis lo dedico a todas las personas que me acompañaron en mi proceso de formación profesional, en especial, a mis padres, por sus consejos y amor que me han permitido cumplir un objetivo importante en mi vida.

A mis primos y familiares, por su continua enseñanza, formaron una persona de bien en mí.

A mis abuelos, tíos y familiares por sus apoyos incondicionales en toda la etapa de mi formación profesional.

A mis docentes por su enseñanza y dedicación que inculcaron las bases para mi formación profesional.

Carbajal Osis Franklin Daniel

AGRADECIMIENTO

A Dios por permitirme estudiar está bonita carrera y haberme brindado vida y salud, cuidar de mis seres queridos y darme la oportunidad de cumplir uno de mis objetivos profesionales.

A mis padres y seres queridos por todo el esfuerzo que gracias a ellos pude cumplir uno de mis objetivos profesionales.

A mi asesor de tesis, por su aporte, tiempo y orientación durante el desarrollo de la presente investigación.

A la Universidad Nacional José María Arguedas y docentes, por las enseñanzas durante este periodo de formación profesional que me ayudará a ser un profesional exitoso.

Gracias.

ÍNDICE

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

El presente trabajo de investigación se realizó con el propósito de aumentar la eficiencia de un modelo predictivo de las propiedades mecánicas de los suelos. Para ello, se desarrollaron diferentes modelos predictivos empleando 3 técnicas de Machine Learning denominados: Extra Trees Regressor, Random Forest Regressor y Redes Neuronales Artificiales (ANN), con la finalidad de determinar la técnica que aumenta la eficiencia de un modelo para predecir las propiedades de Capacidad de Soporte (CBR), Máxima Densidad Seca (MDD) y Optimo Contenido de Humedad (OMC). El Set de Datos conformado por 3294 datos de todos los departamentos del Perú fueron recopilados de 311 expedientes técnicos de estudios y proyectos definitivos de construcción del Sistema Electrónico de Contrataciones del Estado (SEACE 3.0) y de investigaciones anteriores. En donde, se consideraron las variables de entrada del modelo: Cantidad de Grava, Cantidad de Arena, Cantidad de Finos, Límite Líquido, Límite Plástico, Índice de Plasticidad, Departamento y SUCS; los cuales fueron utilizados para predecir los valores de: CBR, MDD y OMC. Los resultados destacan la eficiencia de la técnica de Random Forest Regressor, particularmente cuando se emplean herramientas avanzadas de Auto-ML que hacen uso de técnicas innovadoras como el morfismo de red, la optimización bayesiana y entre otras técnicas que ya vienen implementadas en la herramienta. Estos resultados superan a otras técnicas en la predicción de propiedades mecánicas del suelo, alcanzando valores de R-Squared de 0.9342 para OMC, 0.9696 para MDD y 0.9848 para CBR. Comparado con investigaciones anteriores, los resultados muestran mejoras significativas, incluso superando a investigaciones más recientes. Los resultados de la evaluación de los modelos indican que, aunque el rendimiento de los tres modelos es bueno, el modelo de Random Forest Regressor resulta ser significativamente mejor en la predicción de las propiedades mecánicas de los suelos. Finalmente, la aplicación de técnicas de Machine Learning y la metodología KKD de Ciencia de Datos y Minería de Datos ayudaron a aumentar significativamente la eficiencia de los modelos predictivos. Los resultados de este estudio podrían ayudar a seleccionar los modelos de Machine Learning adecuados, los parámetros e hiperparámetros de entrada de forma más rápida y eficiente para predecir las propiedades mecánicas de los suelos.

Palabras clave: Machine Learning, Ciencia de Datos, Minería de Datos, Propiedades del Suelo, Mecánica de Suelos, Geotecnia.

"EFFICIENCY OF A PREDICTIVE MODEL OF THE MECHANICAL PROPERTIES OF SOILS BASED ON MACHINE LEARNING TECHNIQUES"

ABSTRACT

The present research work was carried out with the purpose of increasing the efficiency of a predictive model of the mechanical properties of soils. To this end, different predictive models were developed using 3 Machine Learning techniques called: Extra Trees Regressor, Random Forest Regressor and Artificial Neural Networks (ANN), with the purpose of determining the technique that increases the efficiency of a model to predict the properties of Support Capacity (CBR), Maximum Dry Density (MDD) and Optimum Moisture Content (OMC). The Data Set made up of 3,294 data from all the departments of Peru were compiled from 311 technical files of studies and definitive construction projects of the State Electronic Contracting System (SEACE 3.0) and from previous investigations. Where, the input variables of the model were considered: Gravel, Sand, Fines, Liquid Limit, Plastic Limit, Plasticity Index, Department and SUCS; which were used to predict the values of: CBR, MDD and OMC. The results highlight the efficiency of the Random Forest Regressor technique, particularly when advanced Auto-ML tools are used that make use of innovative techniques such as network morphism, Bayesian optimization and among other techniques that are already implemented in the tool. These results outperform other techniques in predicting soil mechanical properties, reaching R-Squared values of 0.9342 for OMC, 0.9696 for MDD and 0.9848 for CBR. Compared to previous research, the results show significant improvements, even surpassing more recent research. The results of the evaluation of the models indicate that, although the performance of the three models is good, the Random Forest Regressor model turns out to be significantly better in predicting the mechanical properties of soils. Finally, the application of Machine Learning techniques and the KKD Data Science and data mining methodology helped to significantly increase the efficiency of the predictive models. The results of this study could help to select the appropriate Machine Learning models, input parameters and hyperparameters more quickly and efficiently to predict the mechanical properties of soils.

Keywords: Machine Learning, Data Science, Data Mining, Soil Properties, Soil Mechanics, Geotechnics.

CHUMASQA

Kay kunan investigacion llamkayqa ruwakurqa allpakunapa propiedades mecánicas nisqamanta modelo predictivo nisqapa allin ruwayninta yapanapaqmi. Chaypaqmi ruwakurqa hukniray modelos predictivos nisqakuna 3 técnicas de Aprendizaje de Maquina nisqawan sutichasqa: Regresor de Sacha Extra, Regresor de Bosques Aleatorios y Redes Neuronales Artificiales (ANN), chaywanmi determinanapaq chay técnica nisqa, chaymi yapan huk modelopa eficiencianta, chaywanmi willakunman propiedades nisqakunata Yanapakuy atiy (CBR), Densidad Seca Máxima (MDD) chaymanta Contenido de Humedad Óptima (OMC). Chay Conjunto de Datos nisqa 3.294 datos nisqamanta ruwasqa llapan departamentokunamanta Perú suyupi, 311 archivo técnico de estudios nisqamanta, proyectos de construcción definitivo nisqamanta Sistema Electrónico de Contratación Estatal (SEACE 3.0) nisqamanta, hinallataq ñawpaq investigacionkunamanta. Maypichus, chay modelopa variables de entrada nisqakuna qawarisqa karqa: Grava nisqapa hayka kaynin, Arena nisqapa hayka kaynin, Multas nisqapa hayka kaynin, Límite Líquido nisqapas, Límite Plástico nisqapas, Índice de Plasticidad nisqapas, Departamento nisqapas, SUCS nisqapas; chaykunam llamkachirqaku: CBR, MDD hinaspa OMC nisqakunapa chaninkunata willanapaq. Chay ruwasqakunam qawarichin chay técnica Regresor de Bosques Aleatorios nisqapa allin ruwayninta, aswantaqa Auto-ML nisqa ñawpaqman puriq yanapakuykunata servichikuptinku, chaykunam musuq ruwaykunata servichikunku, chaykunam kanku: morfismo de red, optimización bayesiana hinaspa huk ruwaykunapas chay herramientapi ruwasqaña. Chay ruwasqakunam aswan allinta ruwan huk técnicas nisqamanta allpapa propiedades mecánicas nisqamanta willakuypi, chaymi chayarun valores R-Cadrado nisqaman 0,9342 OMC nisqapaq, 0,9696 MDD nisqapaq, 0,9848 CBR nisqapaq. Ñawpaq investigacionkunawan tupachisqaqa, chay ruwasqakunaqa hatun allinchakuykunatam qawachin, chayllaraq investigacionkunataraqmi atipan. Chay modelos nisqakuna chaninchasqapa ruwakuyninqa qawarichinmi, kimsa modelokunapa ruwaynin allin kaptinpas, chay modelo Regresor de Bosques Aleatorios nisqahmi aswan allin rikurirun allpakunapa propiedades mecánicas nisqamanta willakuypi. Tukuyninpiqa, Machine Learning nisqa técnicas nisqawan, KKD Data Science nisqawan, datos minería nisqawan ima, yanapakurqa, chay modelos predictivos nisqap allin ruwayninta anchata yapananpaq. Kay yachaypa ruwakuyninqa yanapanmanmi chay allin modelos de Aprendizaje Máquina nisqakunata, parámetros de entrada

nisqakunata, hiperparámetros nisqakunatapas aswan utqaylla hinaspa allinta akllayta, chaynapi allpakunapa propiedades mecánicas nisqamanta willakunanpaq.

Sapaq simikuna: Maquina Yachay, Willayta Yachay, Willayta Minería, Allpa Kawsay, Allpa Mecánica, Geotecnia.

INTRODUCCIÓN

La eficiencia de un modelo predictivo en el análisis de las propiedades mecánicas de los suelos es crucial para la toma de decisiones en diferentes ámbitos, tales como la construcción de infraestructuras, la planificación urbana, geotecnia y la gestión ambiental. En este contexto, el uso de técnicas de Machine Learning ha demostrado ser una herramienta valiosa para la modelación y predicción de dichas propiedades. En esta tesis, se aborda el estudio de la eficiencia de un modelo predictivo basado en técnicas de machine learning, con el objetivo de evaluar su capacidad para predecir las propiedades mecánicas de los suelos de manera precisa y confiable. Por lo consiguiente, el presente informe de tesis está constituido de la siguiente manera:

PROBLEMA DE INVESTIGACIÓN: En este capítulo se presenta la descripción del problema, la formulación del problema, los objetivos, la justificación y el alcance de la investigación.

ANTECEDENTES: En este capítulo se encuentran los Antecedentes internacionales, nacionales y locales de la presente investigación.

MARCO TEÓRICO: En este capítulo se presentan las bases teóricas científicas de las variables y las definiciones conceptuales.

METODOLOGÍA DE INVESTIGACIÓN: En este capítulo se presentan la operacionalización de las variables, el tipo de investigación, el nivel de investigación, el diseño de investigación, la población y muestra, el método de investigación, las técnicas e instrumentos de recolección de datos y las técnicas estadísticas de análisis de datos.

RESULTADOS: En el capítulo de los resultados se presentan los resultados estadísticos aplicados, resultados de la metodología aplicada.

DISCUSIÓN: En el capítulo de la discusión se presentan el análisis, discusiones y comparaciones de resultados con investigaciones similares citadas en los antecedentes.

Y por último se presentan las conclusiones, recomendaciones, referencias bibliográficas y anexos del presente informe de tesis.

CAPÍTULO I

PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción del Problema

Actualmente a nivel mundial se realizan una gran cantidad de proyectos de construcción e infraestructura civil de diferentes tipos, como proyectos viales, puentes, edificaciones y todo tipo de infraestructuras, las cuales demandan una gran cantidad de estudios de mecánica de suelos de diferentes tipos. Determinar las propiedades mecánicas de los suelos viene a ser una tarea costosa, complicada y poco precisa al momento de realizar las pruebas de laboratorio de mecánica de suelos en los proyectos de construcción. El crecimiento de la población a nivel mundial demanda la construcción de infraestructuras de diferentes tipos para las diferentes actividades que los seres humanos realizan. En un informe de un estudio hecho por la transnacional Marsh menciona un pronóstico global para la construcción hasta el 2030, donde destaca que la industria de la construcción liderará la recuperación económica mundial de la pandemia en el mediano plazo con un crecimiento más rápido que los otros sectores, se pronostica que el mercado mundial de la construcción crezca en 4,5 billones de dólares en los diez últimos años llegando a 15,2 billones de dólares. Así mismo, se espera que los 10 principales mercados de la construcción mundial representen casi el 70% del crecimiento en el mismo periodo de 10 años (Marsh McLennan, 2021). Dicha demanda de la construcción a nivel mundial al mismo tiempo demanda el estudio de las propiedades mecánicas de los suelos para cada tipo de proyecto de construcción, ya que el estudio de la mecánica de los suelos viene ser el primer paso para la toma de decisiones acerca del diseño y planificación de un proyecto de construcción. Según Day (2010), en su libro de Ingeniería de Cimentaciones, menciona que las pruebas de laboratorio deben ser lo más simples posible. Las pruebas que utilizan equipos elaborados requieren mucho tiempo y, por lo tanto, son costosas, y están sujetas a errores graves a menos que sean realizadas de manera cuidadosa y concienzuda por técnicos altamente experimentados. Dichos métodos pueden estar bastante injustificados si las muestras son pocas o si el costo es alto en relación con el costo del proyecto. Las pruebas elaboradas y costosas se justifican solo si la mayor precisión de los datos generará ahorros valiosos en el diseño o eliminará el riesgo de una falla costosa (Day, 2010). Lo que significa que a nivel mundial crece la demanda de estudios de los suelos para predecir las propiedades mecánicas de

los suelos de un proyecto de construcción, pero muchas veces no son realizados correctamente o son obviadas.

A nivel nacional, el sector de la construcción en el Perú es muy demandada al ser un país en vías de desarrollo, con necesidades de infraestructura y obras civiles en las diferentes regiones del País. Según un informe de la Revista Digital de la Cámara del Comercio de Lima menciona que el aporte de la construcción al PBI en el Perú durante la pandemia fue el 6,7% del PBI nacional, lo que se traduce en 16.500 millones de dólares americanos, destacando su importancia en la economía peruana. Al mismo tiempo recalca que en el plano laboral, la construcción genera 211.455 puestos de trabajo formales a nivel nacional (La Cámara, 2022). Según un reporte del INEI 2021 menciona que la actividad Construcción creció en 231,9% debido a la mayor ejecución de obras del sector privado y público, de acuerdo con el plan de reactivación de las actividades económicas en el contexto del Estado de Emergencia Sanitaria Nacional (INEI, 2021). Así mismo, el sector de construcción permite el dinamismo de la economía lo que conlleva el crecimiento del país y otros sectores. El INEI 2021 dio a conocer que el incremento del consumo final del gobierno se explica por el aumento en remuneraciones (7,5%) y bienes y servicios (39,4%). Asimismo, la inversión bruta fija se incrementó notablemente como resultado de la mayor ejecución de obras de construcción (231,7%) y adquisiciones en maquinaria y equipo (86,6%); las exportaciones aumentaron en 43,9% y las importaciones de bienes y servicios subieron en 54,9% (INEI, 2021). Por otro lado, en un informe del diario El Peruano 2022 menciona que la historia sísmica del Perú ha evidenciado que, ante la ocurrencia de un sismo, el mayor nivel de sacudimiento del suelo se percibe sobre suelos blandos y/o inestables, a diferencia de los suelos rígidos, que atenúan este sacudimiento. Entonces, estudiar los suelos de las áreas urbanas del Perú permite conocer las propiedades físicas de la estructura del subsuelo y su comportamiento dinámico a la demanda sísmica. Las ondas sísmicas recorren diferentes tipos de suelos que, dependiendo de su consistencia, composición y estructura, pueden modificar su contenido frecuencial, amplificación y duración del sacudimiento del suelo (El Peruano, 2022). Ante la demanda y el crecimiento de la construcción es importante realizar estudios de mecánica de suelos en el Perú ya que muchas veces estos estudios no son realizados o ignorados y en consecuencia se ven los resultados de esas malas decisiones en el transcurso del proyecto, al final del proyecto o lamentablemente ante un sismo.

A nivel local, el Instituto Geofísico del Perú en el 2018 realizó estudios de suelos tras rajaduras en paredes, viviendas e instituciones educativas de las comunidades de Antuyo y Curanco en el distrito de Tambobamba de la provincia de Cotabambas de la región de Apurímac, con la finalidad de determinar el origen de los agrietamientos que afectan las viviendas de los pobladores. Donde se menciona que los agrietamientos de las viviendas se originaron por la mala calidad de los suelos, aunado a los efectos de eventos geodinámicos en la zona. Los eventos geodinámicos tuvieron como factores detonantes: las precipitaciones pluviales y fuerzas externas (sismos o vibraciones), que podrían estar generando movimientos en masa, que afectan la seguridad física de las viviendas de los pobladores. Cabe precisar que, los estudios realizados son: el cartografiado geológico (sirve para determinar la litología de la zona), geomorfológico (determina las formas de relieve de la zona), geodinámico (para inventariar y delimitar algunos peligros de origen geológico que puedan estar afectando la seguridad física de los pobladores) y estudio geotécnico (tipo de estudio de la presente investigación), que consistió en elaborar calicatas y obtener muestras de suelos para conocer las propiedades físicas y la capacidad portante de estos (Instituto Geofísico del Perú, 2018). Actualmente, en la ciudad de Andahuaylas de la región de Apurímac se ha podido observar que existen edificios de instituciones, edificios comerciales, casas domésticas, pavimentos y construcciones en general en donde no se realizan un adecuado estudio del suelo para predecir las propiedades mecánicas del suelo y en consecuencia las obras se paralizan, no se concluyen, surgen agrietamientos y fallas generando pérdidas económicas o incluso poniendo en peligro la vida de las personas de dicha ciudad. El problema principal en la ciudad de Andahuaylas es que no se realizan correctamente, o simplemente no se realizan los estudios geotécnicos, específicamente estudios de mecánica de suelos que permitan predecir las propiedades mecánicas del suelo. Propiedades mecánicas básicas como: CBR (California Bearing Ratio), MDD (Máxima densidad seca) y OMC (Óptimo contenido de humedad) que son imprescindibles en un proyecto de construcción y que no son realizados ya que requieren ensayos de laboratorios complicados, costosos y demandan tiempo para su elaboración.

Por un lado, el CBR (California Bearing Ratio) permite saber la resistencia, capacidad portante o compactación del suelo. Este ensayo puede ser evaluado tanto en laboratorio como en campo, siendo en ambos casos de un alto costo y tiempo. Para caracterizar el material de base se usa el CBR de laboratorio, con muestras recopiladas en campo, remoldeadas y sometidas a

condiciones saturadas por 4 días, lo que supone aproximadamente una semana de duración para la obtención del resultado. Para obtener el valor del CBR en subrasantes y poder caracterizar un suelo se necesitan como mínimo 6 valores de CBR y éstos se deben tomar cada 2 Km de espaciamiento como máximo. Sin embargo, la cantidad de muestras se ve influenciada por la variabilidad del terreno, longitud, importancia del proyecto, recursos técnicos y económicos disponibles. Además, el valor de un CBR equivale al promedio de tres especímenes ensayados (Araujo Navarro, 2014). Por último, el costo de un ensayo de CBR se encuentra alrededor de los S/530.00 y realizar un ensayo de CBR en cada una de las muestras adquiridas de las calicatas resultaría muy costoso si se quiere conocer el comportamiento exacto del suelo a lo largo de una sección de estudio (Valderrama Purizaca, 2020).

Por otro lado, la MDD (Máxima densidad seca) que permite saber la máxima densidad para un esfuerzo de compactación requiere pruebas de laboratorios laboriosas, lentas y costosas. Ensayos que requieren equipos, materiales, procedimientos complejos y conocimientos técnicos para calcular la MDD, algunos ensayos para calcular la MDD son: Ensayo de Proctor, Este ensayo se realiza en una máquina llamada compactador de Proctor, que se utiliza para aplicar una carga progresiva a una muestra de material granular mientras se mantiene constante la humedad de la muestra. El ensayo se realiza variando el contenido de humedad de la muestra y midiendo la densidad máxima seca y la humedad óptima para cada nivel de humedad. Ensayo Proctor Modificado: Este ensayo es similar al ensayo de Proctor, pero se realiza a una velocidad de compactación más baja para permitir que el suelo se compacte de manera más uniforme. Ensayo de compresión Uniaxial: Este ensayo se realiza en una máquina llamada compresómetro uniaxial, que se utiliza para aplicar una carga progresiva a una muestra de material granular mientras se mantiene constante la humedad de la muestra. El ensayo se realiza variando el contenido de humedad de la muestra y midiendo la densidad máxima seca para cada nivel de humedad (Manrique Dávila, 2019). En consecuencia, para proceder a determinar la MDD viene ser una tarea poco eficiente ya que se requiere de muchas muestras para su correcto cálculo, demanda ensayos costosos y conocimientos técnicos complicados.

Por último, el OMC (Óptimo contenido de humedad) que es el contenido de humedad con el que se obtiene la máxima densidad seca por medio de pruebas de laboratorios laboriosas, lentas

y costosas. Ensayos que requieren equipos, materiales, procedimientos complejos y conocimientos técnicos para calcular el OMC, algunos ensayos para calcular la OMC son: Ensayo de Casagrande: Este ensayo se realiza en un dispositivo llamado penetrómetro de Casagrande, que se utiliza para medir la resistencia del suelo a la penetración. Se realiza variando el contenido de humedad de la muestra y midiendo la resistencia a la penetración para cada nivel de humedad. Ensayo de Atterberg: Este ensayo se utiliza para determinar las propiedades del suelo en estado líquido y plástico. Se realiza variando el contenido de humedad de la muestra y midiendo la humedad óptima para cada nivel de humedad. Ensayo de Proctor: Este ensayo se realiza en una máquina llamada compactador de Proctor, que se utiliza para aplicar una carga progresiva a una muestra de material granular mientras se mantiene constante la humedad de la muestra. El ensayo se realiza variando el contenido de humedad de la muestra y midiendo la densidad máxima seca y la humedad óptima para cada nivel de humedad. Ensayo de Proctor modificado: Este ensayo es similar al ensayo de Proctor, pero se realiza a una velocidad de compactación más baja para permitir que el suelo se compacte de manera más uniforme (Manrique Dávila, 2019). En consecuencia, para proceder a determinar el OMC viene ser una tarea poco eficiente ya que se requiere de muchas muestras para su correcto cálculo, demanda ensayos costosos y conocimientos técnicos complicados.

Con los aspectos mencionados anteriormente es importante investigar y conocer los diferentes modelos predictivos de las propiedades mecánicas de los suelos que permiten obtener los resultados de dichos ensayos de forma rápida y exacta. Para lo cual se propuso desarrollar y aumentar la eficiencia de un modelo predictivo de las propiedades mecánicas de los suelos mediante el uso de técnicas de Machine Learning.

1.2. Formulación del Problema

1.2.1. Problema General

- ¿De qué manera se puede aumentar la eficiencia de un modelo predictivo de las propiedades mecánicas de los suelos basado en técnicas de Machine Learning?

1.2.2. Problemas Específicos

- ¿Cuál de las técnicas de Regresión de Machine Learning aumenta la eficiencia de un modelo para predecir el resultado del ensayo de CBR (California Bearing Ratio)?
- ¿Cuál de las técnicas de Regresión de Machine Learning aumenta la eficiencia de un modelo para predecir el resultado del ensayo de MDD (Máxima densidad seca)?
- ¿Cuál de las técnicas de Regresión de Machine Learning aumenta la eficiencia de un modelo para predecir el resultado del ensayo de OMC (Contenido óptimo de humedad)?

1.3. Objetivos

1.3.1. Objetivo General

- Aumentar la eficiencia de un modelo predictivo de las propiedades mecánicas de los suelos basado en técnicas de Machine Learning.

1.3.2. Objetivos Específicos

- Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para para predecir el resultado del ensayo de CBR (California Bearing Ratio).
- Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para para predecir el resultado del ensayo de MDD (Máxima densidad seca).
- Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para predecir el resultado del ensayo de OMC (Contenido óptimo de humedad).

1.4. Justificación

La eficiencia de un modelo predictivo en el análisis de las propiedades mecánicas de los suelos es crucial para la toma de decisiones en diferentes ámbitos, tales como la construcción de infraestructuras, la planificación urbana, geotecnia y la gestión ambiental. La predicción de las propiedades mecánicas de los suelos en los proyectos de construcción es de mucha ayuda ya que permite conocer de forma rápida y exacta las propiedades mecánicas del suelo para tomar decisiones acerca del diseño y la planificación del proyecto de construcción. Especialmente en

proyectos en el que se necesita realizar muchos ensayos y por el tiempo de demora de cada ensayo no se obtienen de manera correcta o no son realizados.

La importancia del uso de Machine Learning en la predicción de las propiedades mecánicas de los suelos permite una gran cantidad de soluciones innovadoras que se podrían aplicar para optimizar tiempos y resultados en los diferentes ensayos de laboratorio. Así como también, permite tener un modelo capaz de predecir las propiedades mecánicas del suelo de cualquier zona de una región determinada de forma eficiente.

Asimismo, la investigación beneficia en la predicción de las propiedades mecánicas de los suelos en las dimensiones de ensayo de CBR (California Bearing Ratio), ensayo de MDD (Máxima densidad seca), ensayo de OMC (Contenido óptimo de humedad) y como guía para plantear nuevos modelos predictivos de las propiedades mecánicas de los suelos y geotécnicos. La investigación mejora la eficiencia de los modelos de predicción de los ensayos anteriormente mencionados y la exactitud de los resultados, ya que los modelos existentes se basan en modelos matemáticos estadísticos y los que se basan en Machine Learning no utilizan las técnicas más eficientes de esta rama.

El sector de la construcción pública y privada son beneficiados con los objetivos de la investigación porque permitirá tener resultados de ensayos de laboratorio en segundos, ya que dichos ensayos mencionados anteriormente demandan altos costos y tiempos.

La investigación tiene impacto en factores económicos ya que permite ahorrar dinero de forma directa e indirecta al momento de automatizar los ensayos de la predicción de las propiedades mecánicas de los suelos. De forma directa se ahorran los costos que conllevan realizar dichos ensayos, así como también los costos indirectos como los costos de los materiales y maquinas que se necesitan para cada ensayo, el costo del personal especializado, el tiempo de demora de cada ensayo. Por último, en un proyecto de construcción civil pública o privada permite ahorrar costos y gastos al saber con exactitud las propiedades mecánicas de los suelos, ya que un mal estudio de suelos repercute en fallas estructurales y en consecuencia en gastos por reparaciones.

La investigación tiene impactos sociales porque ayuda a incentivar a realizar los ensayos de las propiedades mecánicas de los suelos de forma fácil y rápida para cualquier proyecto de construcción o investigación con una exactitud considerable. Así mismo, se incentiva a realizar estudios de suelos para la construcción de infraestructuras en el Perú, ya que en la mayoría de las construcciones del Perú no se realizan estudios de suelos al comenzar un proyecto de construcción.

La investigación tiene impactos ambientales porque ayuda a reducir la huella de carbono y el cambio climático al automatizar los ensayos de laboratorio de las propiedades mecánicas de los suelos. Así mismo de forma indirecta, se ayuda al medio ambiente al no incurrir en gastos innecesarios de materiales de construcción, reparación e incluso de demolición.

Finalmente, la presente investigación beneficia a la ciudad de Andahuaylas y a las ciudades de todo el Perú ya que se cuenta con un modelo predictivo de las propiedades mecánicas del suelo que permite predecir las propiedades mecánicas de los suelos con las propiedades físicas de los suelos del Perú.

1.5. Limitaciones

El presente trabajo de tesis presentó limitaciones con respecto a los siguientes puntos:

- La recolección de datos, ya que existen pocos estudios de mecánica de suelos en la ciudad de Andahuaylas y en el Perú.
- La calidad y disponibilidad de los datos, ya que los datos fueron recolectados de numerosos expedientes técnicos del portal SEACE 3.0 donde la mayoría de los expedientes técnicos no contenían los datos que se requerían para la presente investigación o presentaban datos incompletos, así mismo, se encontraban en formatos PDF, imagen, incluso copias borrosas en blanco y negro dificultando la recolección de los datos.
- Laboratorios y especialistas de geotecnia y mecánica de suelos, ya que la universidad no cuenta con dichos laboratorios.
- Recolección y almacenamiento de base de datos o data sets con respecto a las pruebas de laboratorio de mecánica de suelos, ya que las universidades no guardan una base de datos de sus ensayos de laboratorio.

- El financiamiento del presente proyecto, ya que el investigador financió la totalidad del presupuesto.

CAPÍTULO II

ANTECEDENTES

2.1. Antecedentes de investigación

En los siguientes párrafos se mencionan siete antecedentes internacionales, dos antecedentes nacionales, donde se presentarán distintas investigaciones realizadas con respecto a trabajos cuyo tema se asemeja al tema tratado en el presente proyecto de investigación.

2.1.1 Antecedentes Internacionales

Los autores Li et al., China, 2024 en su artículo científico "Prediction of Soil Compaction Parameters Using Machine Learning Models" de la revista Applied Sciences, en su investigación plantean como objetivo general evaluar la viabilidad del aprendizaje automático en la predicción de parámetros de compactación del suelo y comparar las ventajas y desventajas de diferentes modelos. Para ello, los autores recolectaron 168 conjuntos de datos de muestras de suelo y utilizaron cuatro técnicas de aprendizaje automático: Máquina de Vectores de Soporte (SVM), Red Neuronal Artificial (ANN), Bosques Aleatorios (RF) y Árboles de Impulso de Gradiente Extremo (XGBoost) para desarrollar modelos predictivos de MDD y OMC. Los resultados señalan que XGBoost, como método de aprendizaje de conjunto (Tree Ensembles), ofrece las predicciones más precisas tanto para MDD como para OMC, logrando un valor de R-Squared de 0.9234 para OMC y 0.9098 para MDD. Además, el análisis de la importancia de las características revela que el límite plástico (PL) y el límite líquido (LL) son los factores que más inciden en OMC y MDD; llegando a la conclusión, que los Ensemble Algorithms llegan a ser efectivos en cuestiones de rendimiento y generalización. También concluye que la ANN muestra menos éxito en la predicción de OMC y MDD, lo que indica la necesidad de grandes conjuntos de datos y de alta calidad, así como algoritmos optimizados para mejorar el rendimiento del modelo, sugiere la importancia de la optimización de la arquitectura de redes neuronales (Li et al., 2024).

Los autores Gowda et al., India, 2024 en su artículo científico "Prediction of the Subgrade Soil California Bearing Ratio Using Machine Learning and Neuro-Fuzzy Inference System Techniques: A Sustainable Approach in Urban Infrastructure Development" de la revista Urban

Science, en su investigación plantean como objetivo general desarrollar modelos predictivos de CBR utilizando técnicas estadísticas y de Soft Computing. Para ello los autores utilizan técnicas de Machine Learning como la regresión lineal multivariable (MLR), las redes neuronales artificiales (ANN) y el sistema de inferencia neuro-difuso adaptativo (ANFIS), para predecir indirectamente el CBR basándose en el tipo de suelo, el índice de plasticidad (PI) y el MDD. Donde se analizó 2191 muestras de suelo para parámetros de PI, MDD, tamaño de partículas y CBR, utilizando tanto cálculos teóricos como análisis de grandes volúmenes de datos. El ANFIS mostró un rendimiento sobresaliente en la predicción de CBR, con un valor de R-Squared de 0.81, superando a MLR y ANN. Se demostró mediante el análisis de sensibilidad que el PI es el parámetro más influyente en el CBR, con una importancia relativa del 46%; llegando a la conclusión, que los hallazgos subrayan la eficacia de las técnicas de Soft Computing en la predicción precisa del CBR, proporcionando valiosa información para la gestión sostenible de recursos urbanos y el diseño eficiente de infraestructuras (Gowda et al., 2024).

Los autores Ofrikhter et al., Rusia, 2022 en su artículo científico "Estimation of soil properties by an artificial neural network" de la universidad Peter the Great St. Petersburg Polytechnic University, en su investigación plantean como objetivo general usar redes neuronales artificiales (ANN) para estimar los parámetros mecánicos de suelos basados en características físicas conocidas. Para ello, se recopiló 420 datos de pruebas de laboratorio y como resultado se obtuvo una red neuronal artificial que permite predecir la cohesión específica y el ángulo de fricción de un suelo arcilloso con un valor de MAPE de 15.33% y 9.1% respectivamente. Con una arquitectura de red neuronal de 3 capas ocultas con 150 neuronas para el ángulo de fricción y para la cohesión especifica 4 capas ocultas con 200 neuronas. De acuerdo con la comparación de los resultados la ANN permite aumentar la precisión de estimación de ambos parámetros; llegando a la conclusión, que las ANN es un método prometedor de análisis en geotecnia que permite predecir las características mecánicas de los suelos a partir de sus parámetros físicos. La precisión es superior a la de los métodos estadísticos conocidos. Así mismo, recomienda utilizar datos de diferentes regiones, ya que puede generar un mayor error en otras regiones (Ofrikhter et al., 2022).

Los autores Taffese y Abegaz, Suiza, 2021 en su artículo cientifico "Artificial Intelligence for Prediction of Physical and Mechanical Properties of Stabilized Soil for Affordable Housing"

de la revista Multidisciplinary Digital Publishing Institute, como objetivo general proponen desarrollar cuatro modelos basados en inteligencia artificial para predecir el contenido de humedad óptimo (OMC), la densidad seca máxima (MDD) y la resistencia a la compresión no confinada (UCS). Se utilizaron datos conformados por 408 observaciones y 7 características que describen la proporción y los tipos de suelos estabilizados, los límites de Atterberg y los grupos de clasificación de suelos, adicionalmente, se utilizaron 2 variables (OMC y MDD) para predecir UCS. Teniendo como mejor resultado los modelos de Ensemble of Regression Trees para cada variable objetivo de OMC, MDD, UCS+ y UCS-. Con un valor de MSE igual a 13.23, 21.201, 422.149 y 694.752 respectivamente. Con un valor de R-Squared igual a 0.76, 0.59, 0.69 y 0.49 respectivamente. Llegando a la conclusión de que los modelos de Ensemble of Regression Trees superan a los modelos ANN. El rendimiento de todos los modelos podría mejorarse aún más con más datos y se puede aplicar para determinar la proporción óptima de estabilizadores de suelos que podrían cumplir con el OMC, MDD y UCS deseados (Taffese & Abegaz, 2021).

Los autores Diaz et al., España, 2021 en su artículo científico "Machine learning techniques for relating liquid limit obtained by Casagrande cup and fall cone test in low-medium plasticity fine grained soils" de la Universidad de Alicante, como objetivo general de aplicar Técnicas de Machine Learning para correlacionar el LLcup (Límite líquido por el ensayo de copa) y LLcone (Límite líquido por el ensayo de cono). Para ello se ha llevado a cabo una recopilación de 113 datos de muestras de suelos finos. Luego, se realizó un estudio comparativo de diferentes algoritmos teniendo en consideración otras características como el límite plástico y la granulometría. El ajuste de hiperparámetros del modelo Extra-Trees se ha realizado mediante GridSearchCV, se ha utilizado la validación cruzada de k-folds igual a 5 para obtener la combinación con mayor precisión obteniendo un valor de coeficiente de determinación R-Squared de 0,99. Concluyendo que, el algoritmo Extra-Trees adoptado tiene un desempeño claramente superior a la regresión lineal aplicada tradicionalmente para relacionar LLcup y LLcone. En consecuencia, se recomienda el uso de técnicas sofisticadas como ML. Así mismo recomienda, apoyar al modelo con una gran cantidad de datos variados para tener una mejor predicción en otras regiones diferentes al de España (Diaz et al., 2021).

Los autores Nguyen et al., Vietnam, 2020 en su artículo científico "Backpropagation Neural Network-Based Machine Learning Model for Prediction of Soil Friction Angle" de la revista Mathematical Problems in Engineering, en su investigación plantean como objetivo general construir el algoritmo de retro propagación de regularización bayesiana (BR) para predecir el ángulo de fricción interna del suelo. Para ello se propone un modelo ANN con algoritmo BR para predecir el ángulo de fricción interna del suelo. Se recopilan un total de 145 resultados experimentales del proyecto de carretera Danang-Quang Ngai, Vietnam, para la construcción del modelo ANN-BR. Los datos de entrada para el proceso de entrenamiento de la red son: El contenido de arcilla, el contenido de humedad natural, el límite líquido, el límite plástico, la gravedad específica y la relación de vacíos. Los resultados obtenidos fueron un valor de R = 0,8885, RMSE = 0,0442 (rad) y MAE = 0,0328 (rad) para el conjunto de datos de prueba. Concluyendo que, en los problemas de aprendizaje automático, los datos son el factor clave para crear una herramienta predictiva confiable. Por lo tanto, la recopilación de datos adicionales para mejorar el algoritmo es el principal objetivo del estudio, lo que ayuda a evitar costosos experimentos en el campo (Nguyen et al., 2020).

Los autores Pham et al., Vietnam, 2019 en su artículo científico "Development of artificial intelligence models for the prediction of Compression Coefficient of soil: An application of Monte Carlo sensitivity analysis" de la revista Science of the Total Environment, en su investigación plantean como objetivo general desarrollar diferentes modelos de IA: Red Neural Artificial (ANN), Sistema de Inferencia Difusa basado en Red Adaptativa (ANFIS) y Máquina de Vector de Soporte (SVM) para la predicción del Coeficiente de Compresión del suelo (Cc). Para la construcción y validación de modelos de IA, se utilizaron datos de 189 muestras de suelo arcilloso blando. En total, se utilizaron 13 parámetros del suelo: profundidad de la muestra, densidad aparente, índice de plasticidad, contenido de humedad, contenido de arcilla, gravedad específica, relación de vacíos, límite líquido, densidad seca, porosidad, límite plástico, grado de saturación e índice de liquidez. Como resultado el R-Squared calculado para el Cc son 0,939, 0,957 y 0,989 para los modelos ANN, ANFIS y SVM, respectivamente. Concluyendo que, El modelo SVM es superior en la predicción del Cc. Los resultados del análisis de sensibilidad muestran que, de los 13 parámetros de entrada, cuatro parámetros, denominados, arcilla, grado de saturación, gravedad específica y profundidad de la muestra, son los más representativos para la predicción de Cc. Así

mismo, sugieren, aplicar la técnica de preprocesamiento del análisis de multicolinealidad en los datos para estudiar la posibilidad de una mayor reducción en el espacio de entrada (Pham et al., 2019).

2.1.2. Antecedentes Nacionales

El autor Valderrama Purizaca, Perú, 2020 en su Tesis "Desarrollo de un modelo de Redes Neuronales Artificiales para predecir las Propiedades Mecánicas del Suelo" de la Universidad Señor de Sipán, en su informe de tesis plantea como objetivo general desarrollar un modelo de redes neuronales artificiales para predecir las propiedades mecánicas del suelo en la región Lambayeque. Para ello se recopilaron un total de 285 resultados de pruebas de proyectos de carreteras construidas en la región Lambayeque. Donde las variables de entrada utilizadas fueron los valores de contenido de humedad, contenido de sales, análisis granulométrico y límites de consistencia. Los modelos alcanzaron un MSE para el MDD (0.003 g/cm3), el OMC (2.42%), el CBR95 (6.35 %) y el CBR100 (21.80 %). En cuanto al MAE, para el MDD (0.035 g/cm3), el OMC (1.13 %), el CBR95 (1.65 %) y le CBR100 (2.79 %). Llegando a la conclusión de que se demostró que cada ANN aprendió correctamente comparando los valores reales con los valores predichos por el modelo. Lo que ayudó a identificar los hiperparámetros óptimos fue el MSE porque muestra que tan bien un modelo redujo el error y cómo se comportó con respecto a otro. Menciona que lo difícil en la identificación de los hiperparámetros fue realizar el código en Spyder porque hubo diversas técnicas aplicables (Valderrama Purizaca, 2020).

La autora Manrique Dávila, Perú, 2019 en su tesis "Predicción de propiedades mecánicas de suelos utilizando redes neuronales artificiales" de la Universidad Continental, en su informe de tesis plantea como objetivo general encontrar los resultados de los ensayos de las propiedades mecánicas de los suelos usando un software basado en algoritmos de redes neuronales artificiales. Para ello se recopiló 222 datos de las propiedades mecánicas y físicas del suelo de 5 expedientes técnicos de estudios y proyectos definitivos realizados por Provias Nacional. Se entreno el modelo mediante las variables de entrada de porcentaje de grava, arena, finos, limite líquido, limite plástico e índice de plasticidad. Variables de salida de CBR al 100%, CBR al 95%, OMC y MDD. Con asistencia del software NeuralTools se obtuvo 4 mejores modelos neuronales, en la que se utilizó el 80% de datos para el entrenamiento de los modelos neuronales y 20% para el proceso de prueba

del modelo. Donde se obtuvieron un coeficiente de determinación en la etapa de validación de R-Squared = 0.7317 para CBR al 100%, un R-Squared = 0.8229 para CBR al 95%, R-Squared = 0.8445 para el OMC y un R-Squared = 0.9265 para la MDD. Llegando a la conclusión que la base de datos de los ensayos de suelos realizados y supervisados por Provías Nacional es confiable, ya que se usó la base de datos de Provías Nacional y ensayos de su localidad de Hauncayo para lograr dichos resultados (Manrique Dávila, 2019).

2.1.3. Antecedentes Locales

No se encontraron antecedentes locales de temas relacionados a Machine Learning (ML) y/o Inteligencia artificial (IA) para predecir las propiedades mecánicas de los suelos.

CAPÍTULO III

MARCO TEÓRICO

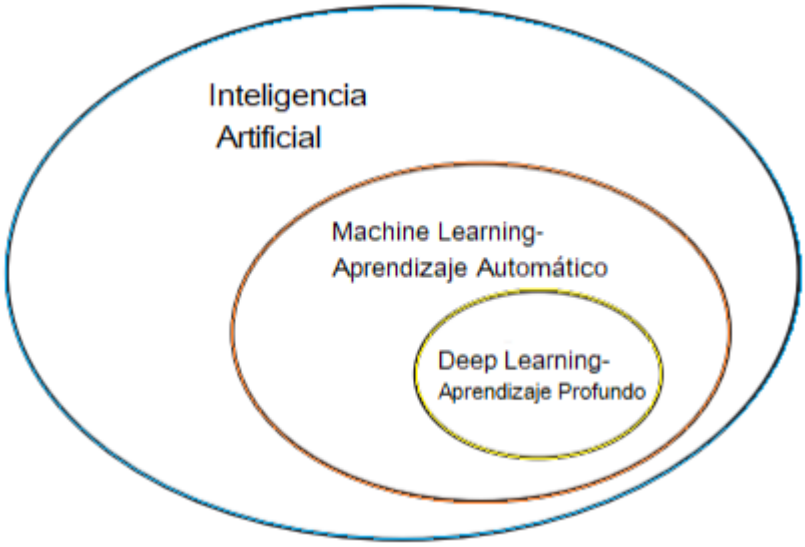
3.1. Machine Learning

El aprendizaje automático o Machine Learning, también conocido como análisis aumentado, se considera un subconjunto de la inteligencia artificial (IA). Está estrechamente relacionado con la estadística computacional. Los algoritmos de aprendizaje automático son programas informáticos, que se dice que aprenden de la experiencia en una variedad de actividades. Realiza estas tareas con un nivel de precisión que mejora con la experiencia. Los algoritmos de aprendizaje automático deben entrenarse con datos de entrenamiento para lograr el nivel de precisión deseado. Una vez entrenados, estos algoritmos generan información valiosa que puede usarse para predecir resultados futuros (Konasani & Kadre, 2021).

Así mismo Dark (2019), Define el aprendizaje automático como un concepto relacionado con las máquinas de inteligencia artificial, que a menudo son responsables de la detección, la predicción y el diagnóstico. Estos dispositivos a menudo funcionan con datos. Los datos, a menudo denominados datos de entrenamiento, pueden ser datos de muestra o datos históricos que ayudan a entrenar el sistema. Estas máquinas aprenden a analizar patrones en los datos y los usan para crear su propio análisis (Dark, 2019).

Se puede entender que el aprendizaje automático se encuentra dentro del campo más amplio de la inteligencia artificial, mientras que el aprendizaje profundo o Deep Learning se ubica dentro del aprendizaje automático. Por lo tanto, todos los programas de aprendizaje automático se consideran programas de IA, pero no todos los programas de IA son de aprendizaje automático. En cuanto a si los enfoques de IA y aprendizaje automático son similares, se puede decir que sí, ya que cada problema de aprendizaje automático se considera un problema de IA, y el aprendizaje profundo mejora los algoritmos de aprendizaje automático para hacerlos más precisos y eficientes en ciertas tareas, como la extracción de funciones. En resumen, se puede decir que el aprendizaje profundo es simplemente una rama del aprendizaje automático que se enfoca en mejorar los métodos de este campo, y que la principal diferencia entre los dos es que el aprendizaje profundo es un subconjunto del aprendizaje automático (El-Amir & Hamdy, 2020).

Figura 01 Inteligencia artificial, Machine Learning y Deep Learning



Fuente: (Centeno Franco, 2019, pág. 2)

3.1.1. Tipos de sistemas de Machine Learning

Según Russel (2018) dependiendo del tipo y alcance del monitoreo, existen diferentes tipos de sistemas de aprendizaje automático. Se pueden dividir en cuatro categorías principales en función de si han sido entrenados con humanos o no:

- Machine Learning supervisado
- Machine Learning no supervisado
- Machine Learning semi-supervisado
- Machine Learning de refuerzo

a) Machine Learning Supervisado

En machine learning supervisado una etiqueta son los datos proporcionados al algoritmo por la solución deseada, es decir, se le enseña a través de los datos. Un caso representativo, es el ejemplo del correo que se le asigna una etiqueta para su pronóstico, es posible ya que entrena múltiples correos electrónicos a la vez a través de la clase (Russell, 2018).

b) Machine Learning No Supervisado

En este tipo de sistema de aprendizaje automático, los datos pueden considerarse anónimos. Esto significa que no necesita intervención humana para clasificar las etiquetas, pero puede crear grupos o agruparlos por sí mismo (Russell, 2018).

3.1.2. Métodos de Machine Learning

Según (Gonzáles, 2019) menciona a dos tipos populares de métodos de Machine learning basados en el formato de sus salidas:

a) Algoritmos de Clasificación

Se utiliza cuando el resultado deseado es una etiqueta discreta, es decir, el resultado es un conjunto finito de respuestas. De manera similar, cuando se espera que el modelo entrenado prediga dos clases de características, se denomina clasificación binaria. Finalmente, si necesitamos predecir más de dos clases objetivas, lo llamamos clasificación multiclase, multiclase o multidimensional (Gonzáles, 2019).

b) Algoritmos de Regresión

Su uso es para predecir resultados futuros continuos. Además, para responder a una pregunta, se ingresa una cantidad que se puede determinar fácilmente en función de las entradas del modelo. Finalmente, el valor predictivo se puede usar para identificar relaciones lineales entre atributos (Gonzáles, 2019).

3.1.3. Técnicas de Machine Learning

Las herramientas o técnicas de aprendizaje automático también se denominan específicamente algoritmos de aprendizaje automático, en muchos libros de aprendizaje automático, muchos autores se refieren a ellas por los nombres mencionados en sus libros. Algunas técnicas o algoritmos supervisados de regresión más importantes según la literatura son:

- Regresión lineal y múltiple

- Regresión Logística
- Máquinas de soporte de vectores (SVM)
- Arboles de decisiones y bosques aleatorios
- Regression Tree Ensembles (Extra Trees)
- Redes Neuronales Artificiales
- BPNN

a) Regresión Lineal

Los autores El-Amir y Hamdy (2020) recomiendan comenzar con la regresión lineal como algoritmo básico de aprendizaje automático porque es simple y ayuda al lector a comprender conceptos fundamentales como el aprendizaje supervisado, la función de costo y el descenso de gradiente. Después de aprender la regresión lineal, se vuelve más fácil comprender la regresión logística, que se puede categorizar como una pequeña red neuronal. La regresión lineal es un método estadístico común que se utiliza para aprender una función o relación a partir de un conjunto de datos continuos, donde la hipótesis predice el mejor modelo para ajustarse a los datos de entrada, generalmente representados como una línea recta.

b) Regresión Lineal Múltiple

En escenarios de la vida real, generalmente existen múltiples variables que pueden afectar la variable objetivo, como el costo de marketing, las calificaciones de los clientes y las condiciones climáticas. Para mejorar la precisión de nuestras predicciones, deberíamos considerar el uso de todas estas variables en el modelo. Por lo tanto, es necesario construir un modelo de regresión con múltiples variables predictoras, lo que puede aumentar la precisión las predicciones. La ecuación para una línea de regresión simple es $y = \beta 0 + \beta 1x$, mientras que la ecuación para variables predictoras múltiples es $y = \beta 0 + \beta 1x1 + \beta 2x2 + ... + \beta kxk$. Mediante el uso de múltiples variables predictoras, se puede crear un modelo más completo que tenga en cuenta todos los factores que pueden afectar la variable de destino (Konasani & Kadre, 2021).

c) Regresión logística

La regresión lineal y la regresión logística son los dos algoritmos más básicos y esenciales en el aprendizaje automático y se clasifican como simples. Como se sabe, los algoritmos de aprendizaje automático pueden resolver muchos problemas complejos, pero al mismo tiempo, los algoritmos subyacentes también pueden ser complejos. Los algoritmos de regresión lineal y logística son relativamente simples de construir y fáciles de interpretar incluso para personas de negocios sin conocimientos técnicos. Son en su mayoría intuitivos y tardan un tiempo relativamente corto en ejecutarse. Empresas de todo el mundo han estado utilizando estos dos algoritmos de aprendizaje automático durante más de 25 años. La mayoría de los bancos e instituciones financieras utilizan la regresión logística en sus modelos de riesgo crediticio (Konasani & Kadre, 2021).

d) Máquinas de soporte de vectores (SVM)

Otro algoritmo de aprendizaje potente y ampliamente utilizado es la máquina de vectores (SVM), que puede considerarse una extensión de la detección. Al usar el algoritmo Perceptron, se limita la clasificación errónea. Sin embargo, en SVM, el objetivo de optimización es maximizar el umbral. La clasificación de margen se define como la distancia entre el hiperplano (límite de la solución) y las muestras de entrenamiento más cercanas al hiperplano, que son los denominados vectores de apoyo (Raschka & Mirjalili, 2017).

Figura 02 Función de pérdida ε-insensitive

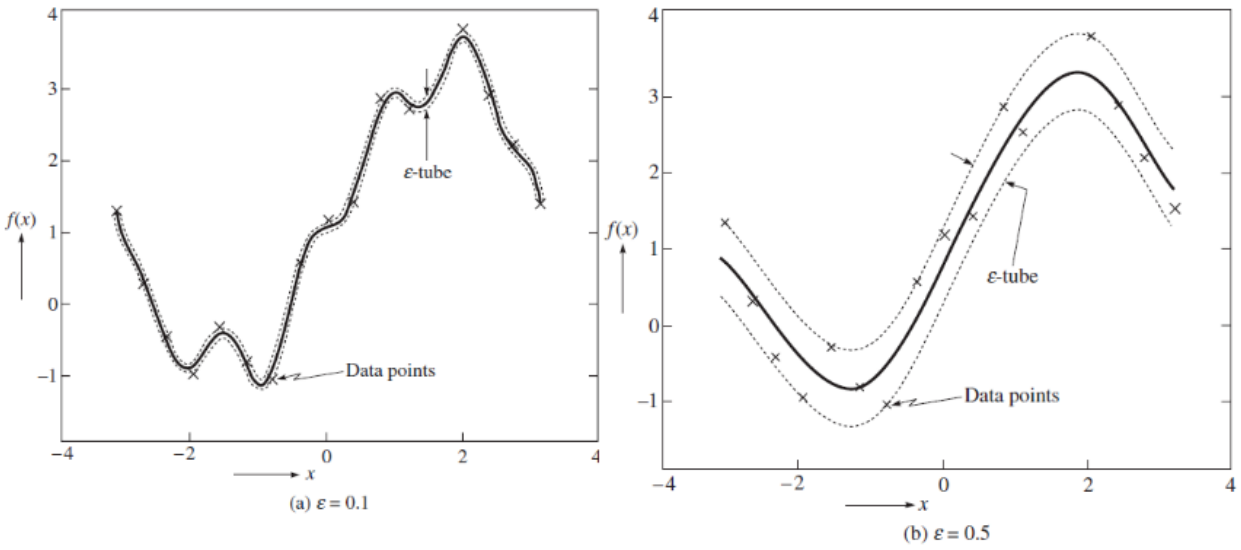
$$|y - f(\mathbf{x})|_{\varepsilon} \Delta \begin{cases} 0 & \text{if } |y - f(\mathbf{x})| \leq \varepsilon \\ |y - f(\mathbf{x})| - \varepsilon & \text{otherwise} \end{cases}$$

Fuente: (Gopal, 2019)

Los métodos de SVM desarrollados inicialmente para problemas de clasificación también se pueden aplicar con éxito en problemas de regresión (predicción numérica). En la regresión, se utiliza una medida de error de aproximación en lugar del margen entre un hiperplano de separación óptimo y los vectores de soporte utilizados en el diseño de los clasificadores de SVM. El objetivo en el aprendizaje de regresión es encontrar una función de regresión no lineal que tenga una desviación máxima de ε para todas las muestras de entrenamiento. La función de pérdida de error

ε-insensible (Figura 02) se utiliza para tener en cuenta el error de regresión en la formulación de SVM. Esta función de error es más tolerante al ruido y, por lo tanto, es más robusta que la función de pérdida de error de suma de errores al cuadrado utilizada en la regresión convencional. El parámetro ε define los requisitos de precisión de la aproximación, y un aumento en ε significa una reducción en los requisitos de precisión, lo que da como resultado efectos suavizantes en la modelización de datos altamente ruidosos. Por otro lado, una disminución en ε puede resultar en un modelo complejo que se sobreajuste a los datos (Gopal, 2019).

Figura 03 Regresión SVM unidimensional



Fuente: (Gopal, 2019)

e) Arboles de decisiones

Los árboles de decisión son uno de los algoritmos más utilizados. Siguen un enfoque completamente diferente a la clasificación. Mientras que la regresión logística, se trabaja en un enfoque de ajuste de curvas, donde se trata de ajustar una función logística a un conjunto de datos. Se observa que la función logística era una función en forma de "S" que predecía la clase 0 o la clase 1. Cuando se trabaja con árboles de decisión, se adopta un enfoque de clasificación basado en árboles. Donde se intenta dividir los datos en pequeños para que cada subconjunto tenga un dominante, ya sea clase 0 o clase 1 (Konasani & Kadre, 2021).

f) Bosques aleatorios

Bosques aleatorios o Random Forest es un conjunto de árboles de decisión que generalmente se entrena mediante bagging or pasting. El parámetro max_samples normalmente se establece en el tamaño del conjunto de entrenamiento. En lugar de crear un BaggingClassifier y proporcionar un DecisionTreeClassifier, se puede usar la clase RandomForestClassifier, que está optimizada para Decision Trees. Lo mismo se aplica a las tareas de regresión, donde se usa la clase RandomForestRegressor equivalente a RandomForestClassifier que comparte casi todos los hiperparámetros de DecisionTreeClassifier para controlar cómo crecen los árboles, además de todos los hiperparámetros de BaggingClassifier para controlar el conjunto en sí, el algoritmo introduce aleatoriedad adicional al buscar la mejor característica entre un subconjunto aleatorio de características, lo que da como resultado una mayor diversidad de árboles. Esto da como resultado un mayor sesgo y una menor varianza, lo que generalmente conduce a un mejor modelo (Géron, 2019).

g) Extra Trees (Extremely Randomized Trees)

Se puede hacer que los árboles en un Random Forest sean aún más aleatorios al considerar solo un subconjunto aleatorio de características y umbrales aleatorios para cada característica en lugar de buscar los mejores umbrales posibles. Esto se conoce como el conjunto de árboles extremadamente aleatorios o Extra-Trees. Esta técnica intercambia más sesgo por una menor varianza y es mucho más rápida de entrenar que los Random Forest. Algunas librerías de Machine Learning como Scikit-Learn proporciona las clases ExtraTreesClassifier y ExtraTreesRegressor que tienen la misma API que las clases RandomForestClassifier y RandomForestRegressor (Géron, 2019).

Extra-Trees es un tipo de modelo de aprendizaje de ensemble que utiliza un enfoque de bagging. Tiene técnicas de entrenamiento similares al modelo Random Forest, utilizando muestreo Bootstrap y selección de variables aleatorias. Sin embargo, Extra-Trees aleatoriza más el punto de corte óptimo durante el entrenamiento del árbol de decisión. Esta técnica ofrece beneficios como una disminución en la varianza general y tiempos de entrenamiento más rápidos, lo que la hace

tres veces más rápida que Random Forests. Además, Extra-Trees tiene un rendimiento comparable o incluso mejor que Random Forests (Diaz et al., 2021).

Los métodos de Ensemble combinan múltiples modelos para crear un poderoso modelo predictivo. Dos modelos de conjunto exitosos que utilizan árboles de regresión como modelos base son los árboles de regresión de Bagging y Boosting. Se ha demostrado que estos modelos resuelven de manera efectiva problemas de regresión complejos y proporcionan interpretabilidad e importancia predictiva (Taffese & Abegaz, 2021).

- **Bagging Regression Tree**

  En el árbol de regresión de Bagging, se forman múltiples modelos base extrayendo aleatoriamente muestras de arranque del conjunto de datos original, que se realizan varias veces. Cada conjunto de datos de entrenamiento de arranque formado contiene en promedio 0,63 N instancias, donde N es el número total de muestras en el conjunto de datos original. El resultado final del modelo es el promedio del resultado previsto de los modelos base individuales, lo que reduce la varianza y mejora la estabilidad. La estimación empaquetada es la predicción media de los árboles y se obtiene promediando la salida prevista de los modelos base individuales (Taffese & Abegaz, 2021).

- **Boosting Regression Tree**

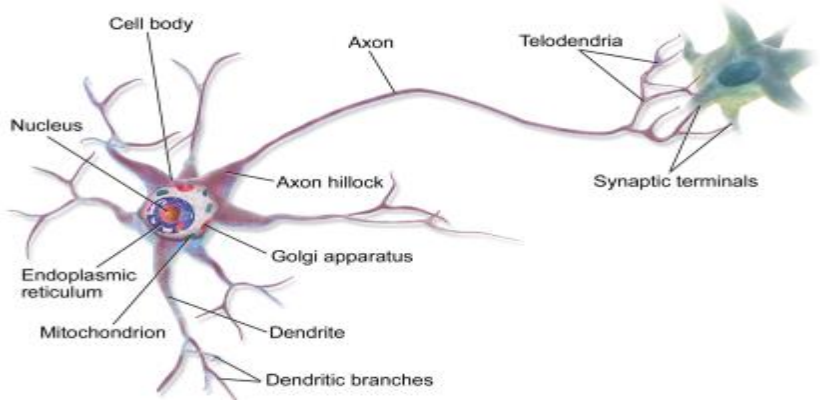
  Boosting es una mejora de Bagging que se enfoca en casos desafiantes en el rendimiento del modelo. Construye en serie modelos de árboles simples con mejoras de un modelo de árbol a otro, y cada árbol se cultiva a partir de un conjunto de datos de entrenamiento utilizando el conocimiento de árboles cultivados previamente. El vector de ponderación se ajusta en cada modelo base en función de los errores percibidos, y el resultado final de los modelos es una suma ponderada de los resultados de los modelos individuales. El algoritmo LSBoost es un enfoque de impulso popular que ajusta los árboles pequeños a los residuos, lo que aumenta lentamente el modelo en áreas de bajo rendimiento. La salida del modelo potenciado es una suma ponderada de las salidas del modelo individual, siendo los pesos mayores para los mejores modelos (Taffese & Abegaz, 2021).

h) Redes neuronales artificiales

La regresión lineal, la regresión logística y los árboles son algoritmos de aprendizaje automático son relativamente menos complicados. Sin embargo, se debe desarrollar una buena comprensión de estos para comprender los algoritmos avanzados de aprendizaje automático. Los algoritmos de regresión y árbol son simples; los resultados de estos algoritmos son fáciles de interpretar para explicar. Estos modelos son buenos para identificar patrones simples en los datos y toman muy poco tiempo de ejecución. Sin embargo, estos algoritmos básicos fallan debido a patrones no lineales en los datos. Tenemos algoritmos más sofisticados que pueden encontrar los efectos de interacción e identificar automáticamente patrones ocultos no lineales. Uno de los algoritmos de aprendizaje automático más sofisticados son las redes neuronales, también conocidas como aproximadores universales (Konasani & Kadre, 2021).

La investigación sobre redes neuronales se ha inspirado en el deseo de producir sistemas artificiales capaces de un procesamiento "inteligente" sofisticado similar al cerebro humano. El perceptrón es el paradigma más antiguo de la red neuronal artificial. En 1982, John Hopfield introdujo una red neuronal de tipo recurrente basada en la interacción de las neuronas a través de un mecanismo de retroalimentación. La investigación en la década de 1980 provocó un auge en la comunidad científica. Se han propuesto nuevos y mejores modelos. Muchos de los problemas tecnológicos actuales se encuentran en áreas donde la tecnología de redes neuronales ha mostrado potencial (Gopal, 2019).

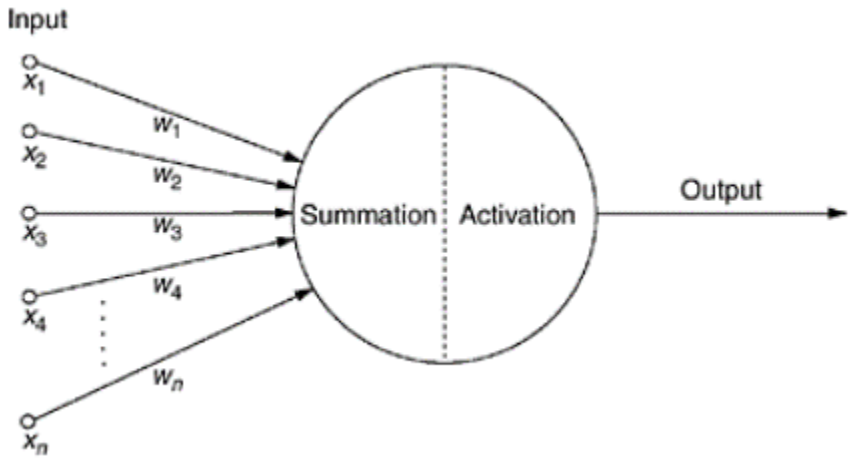
Figura 04 Interacción de redes neuronales bilógicas



Fuente: (Géron, 2019)

Las neuronas artificiales son unidades de procesamiento que reciben entradas de nodos conectados (x) y generan una salida (Output) a través de una función de transferencia, combinando lineal o no linealmente las señales de entrada. Cada unidad de procesamiento funciona como una máquina de reconocimiento de patrones simple en la que se evalúan los datos de entrada en comparación con la fuerza sináptica. La neurona se puede implementar como una suma ponderada de todas las señales de entrada, con un valor numérico o sesgo asociado a cada conexión sináptica. Los pesos (w) indican la fuerza de la conexión sináptica, y modulan la salida de la unidad de procesamiento. La salida de la neurona se convierte en una señal de activación para las unidades conectadas a ella, y así sucesivamente a través de la red neuronal. En las redes de varias capas, esta señal de activación se convierte en entrada para la siguiente capa oculta (Gao, 2009).

Figura 05 Procesamiento de datos en una neurona artificial típica



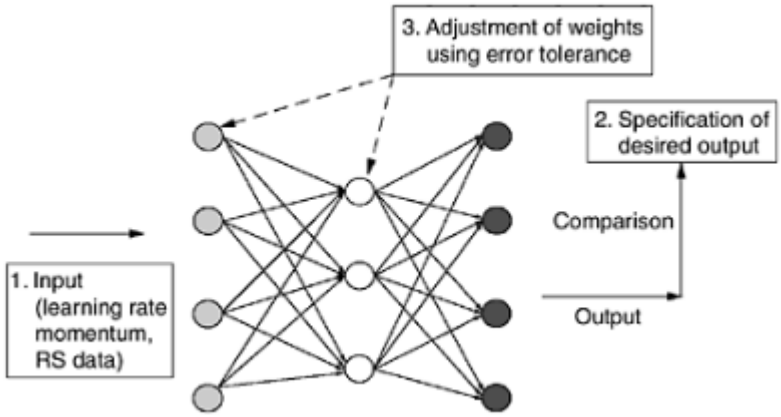
Fuente: (Gao, 2009)

i) Redes Neuronales Back Propagation (BPNN)

Una red neuronal de retropropagación (BPNN) utiliza una topología de avance (feed-forward) de aprendizaje supervisado y el algoritmo de aprendizaje de retropropagación (backpropagation). Funciona pasando datos desde la capa de entrada a la capa de salida a través de las capas ocultas. En el aprendizaje supervisado, se conoce el resultado deseado para cualquier

entrada dada, después de que se presenta una entrada, se produce una salida basada en pesos asignados aleatoriamente y luego se compara con la salida deseada. La discrepancia entre los dos es la señal de error que se propaga hacia atrás desde los nodos de salida a los nodos de entrada a través de la red utilizando el algoritmo de retropropagación de errores. Este proceso se repite iterativamente y los pesos sinápticos entre nodos se ajustan para minimizar el error. Una red de retropropagación simple comprende una sola capa oculta de nodos, una capa de entrada y una capa de salida, y puede modelar funciones continuas con precisión si hay suficientes nodos de procesamiento en la capa oculta con funciones de activación no lineal. Agregar más capas ocultas convierte la red lineal en una no lineal, que puede realizar una regresión logística multivariante, a diferencia de los clasificadores estadísticos tradicionales. El uso de una BPNN para la regresión logística también permite modelar múltiples salidas simultáneamente, lo que puede capturar efectos confusos de múltiples parámetros de entrada (Gao, 2009).

Figura 06 Configuración de una BPNN

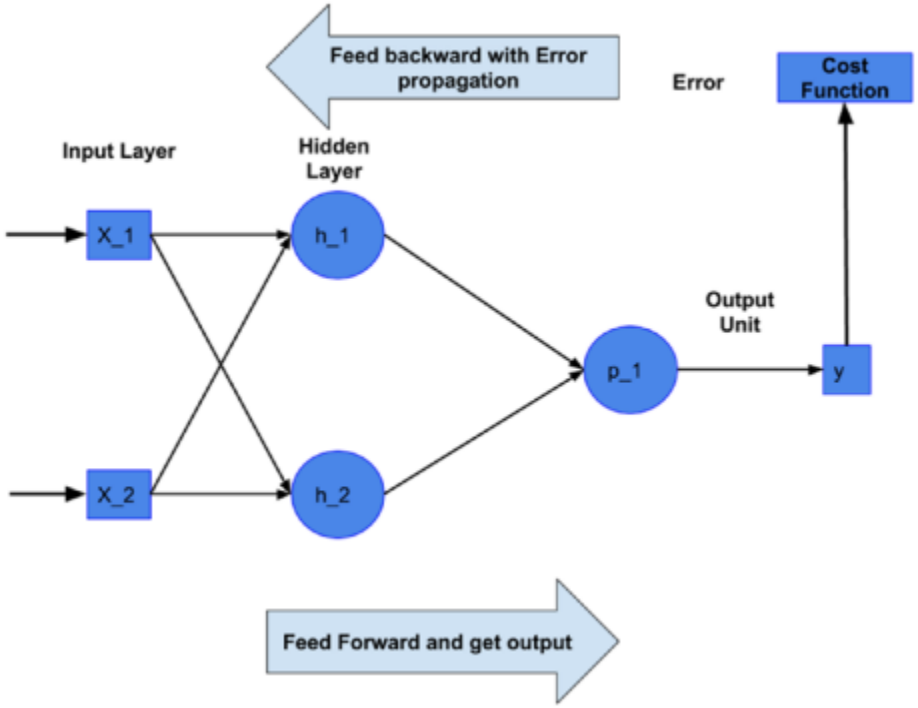


Fuente: (Gao, 2009)

El algoritmo de retropropagación es un método rápido y eficiente para encontrar la derivada parcial de la pérdida con respecto a cada peso en una red neuronal. Si bien las bibliotecas modernas pueden realizar este procedimiento automáticamente, comprender la propagación hacia atrás puede proporcionar información sobre los problemas relacionados con el aprendizaje profundo. El descenso de gradiente es un algoritmo poderoso, pero se vuelve lento y complejo cuando se trata de grandes redes neuronales. La retropropagación es un caballo de batalla del procedimiento de entrenamiento para algoritmos de aprendizaje profundo y es la forma más eficiente de calcular el

gradiente exacto, con un costo computacional siempre de la misma complejidad que calcular la pérdida en sí (El-Amir & Hamdy, 2020).

Figura 07 Arquitectura de una BPNN



Fuente: (El-Amir & Hamdy, 2020)

La retropropagación se basa en dos suposiciones sobre la función de error: que el error total se puede escribir como la suma de los errores individuales de las muestras/minilotes de entrenamiento, y que el error se puede escribir como una función de las salidas de la red. El algoritmo consta de un paso hacia adelante para inicializar los pesos y una red de avance, y un paso hacia atrás para actualizar los pesos en función de los valores almacenados. La retropropagación requiere derivadas parciales, reglas de cadena y álgebra lineal para implementarse (El-Amir & Hamdy, 2020).

Backpropagation es un algoritmo clave para entrenar redes neuronales artificiales. Su idea principal es calcular el gradiente de la función de pérdida con respecto a los pesos de la red usando la regla de la cadena del cálculo y luego usar este gradiente para actualizar los pesos en la dirección que minimiza la función de pérdida. Las ecuaciones clave utilizadas en la retropropagación son:

- Cálculo del error de la capa de salida:
- Cálculo del error de la capa anterior:
- Cálculo del gradiente de la función de costo con respecto al sesgo de una neurona:
- Calculando el gradiente de la función de costo con respecto al peso que conecta dos neuronas:

Figura 08 Ecuaciones utilizadas en una BPNN

$$\delta^L = \nabla_a C \odot \sigma'(z^L)$$

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

$$\frac{\partial C}{\partial b^l_j} = \delta^l_j$$

$$\frac{\partial C}{\partial w^l_{jk}} = a^{l-1}_k \delta^l_j$$

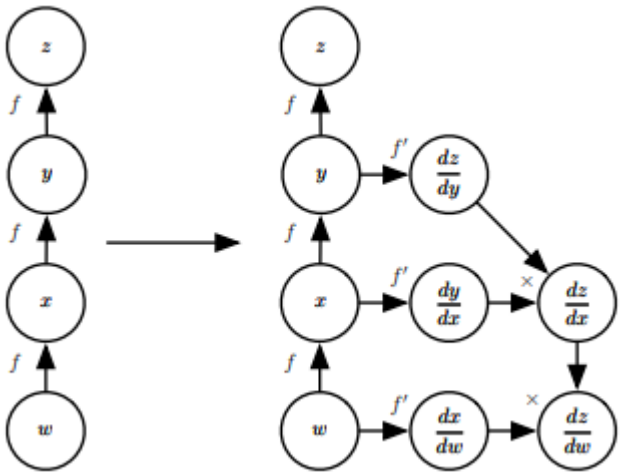
Fuente: (El-Amir & Hamdy, 2020)

Estas ecuaciones se utilizan de forma recursiva para calcular los gradientes de todas las capas de la red. Una vez que se han calculado los gradientes, se utilizan para actualizar los pesos y sesgos de la red mediante un algoritmo de optimización como el descenso de gradiente estocástico.

Para describir el algoritmo de propagación hacia atrás de una manera más precisa y formal, es necesario un lenguaje grafo computacional (GCL). Este lenguaje utiliza nodos para representar variables, que pueden ser escalares, vectores, matrices, tensores u otro tipo de variables. El lenguaje también incluye operaciones, que son funciones simples de una o más variables. El grafo se construye conectando nodos con aristas, lo que representa el flujo de variables a través de operaciones. La salida de una operación está representada por un borde dirigido que apunta a la variable resultante. Si bien las implementaciones de software pueden admitir operaciones con múltiples salidas, esto se evita en la descripción del lenguaje para simplificar la comprensión. El

lenguaje permite una descripción más precisa del algoritmo de propagación hacia atrás (Goodfellow et al., 2016).

Figura 09 Lenguaje grafo computacional de una BPNN



Fuente: (Goodfellow et al., 2016)

En el algoritmo de propagación hacia atrás, inicialmente, todos los pesos de los bordes se asignan aleatoriamente. La ANN se activa para cada entrada en el conjunto de datos de entrenamiento y su salida se compara con la salida deseada. Luego, el error se propaga de nuevo a la capa anterior y los pesos se ajustan en consecuencia. Este proceso continúa hasta que el error de salida está por debajo de un umbral predeterminado, lo que da como resultado una red neuronal "aprendida". La mayoría de las bibliotecas admiten la diferenciación automática, por lo que no es necesario implementar el algoritmo manualmente. En términos simples, la propagación hacia atrás actualiza los pesos secuencialmente haciendo un pase hacia adelante en la red, actualizando los pesos de la última capa usando la etiqueta y las salidas de la última capa, y recursivamente usando esta información en la capa anterior (El-Amir & Hamdy, 2020).

3.1.4. Métricas de Evaluación de modelos de regresión

En los modelos de regresión, la variable de salida es una cantidad continua que puede tomar cualquier valor dentro de su rango. El valor de R-cuadrado o coeficiente de determinación es una medida útil para evaluar la precisión del modelo. Esta medida indica tanto la precisión general del

modelo como la cantidad de varianza en los datos que es explicada por el modelo. Además de R-cuadrado, existen otras medidas que se utilizan para evaluar qué tan cerca o lejos están los valores predichos del modelo de los valores reales (Konasani & Kadre, 2021).

Figura 10 Cuadro de resumen de algunas métricas de evaluación para modelos de regresión

Error at one point	$y_i - \hat{y}_i$
Sum of squares of errors (SSE)	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
Mean absolute deviation (MAD)	$\sum_{i=1}^{n} \frac{\lvert y_i - \hat{y}_i \rvert}{n}$
Mean absolute percent error (MAPE)	$\frac{100}{n} \sum_{i=1}^{n} \frac{\lvert y_i - \hat{y}_i \rvert}{y_i}$
Mean square error (MSE)	$\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}$
Root mean square error (RMSE)	$\sqrt{MSE}$

Fuente: (Konasani & Kadre, 2021)

A) Coeficiente de determinación R-cuadrado

El coeficiente de determinación (R-cuadrado) también indica la precisión con la que el modelo de conversión se ajusta a los datos. En otras palabras, un valor alto de "R" indica un análisis más preciso y, por lo tanto, un modelo excelente. Esta precisión se determina mediante el coeficiente de determinación (R2), que se calcula como el cociente entre dos cantidades (Cordova Valentin & Mori Montalvo, 2021). Se expresa de la siguiente manera:

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

Dónde:

$$SCT = SCE + SCR$$

SCR: Suma de los cuadrados debido a la regresión.

SCT: Sumas de los cuadrados total.

SCE: Suma de los cuadrados de los errores.

B) Error cuadrático medio (MSE)

El MSE (Error Cuadrático Medio) es una medida de la calidad de un modelo de aprendizaje automático, específicamente, un modelo de regresión. Representa el promedio de los cuadrados de las diferencias entre los valores reales y los valores previstos por el modelo. En otras palabras, el MSE mide la diferencia promedio entre lo que se espera y lo que el modelo ha predicho. Cuanto más pequeño sea el MSE, mejor será el ajuste del modelo a los datos de entrenamiento (Valderrama Purizaca, 2020).

$$MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}$$

Donde:

Y: Valores predichos del modelo.

X: Valores Reales del modelo.

N: Cantidad de datos del modelo.

C) Error absoluto medio (MAE)

El MAE (Error Absoluto Medio) es una medida de la calidad de un modelo de aprendizaje automático, específicamente, un modelo de regresión. Representa el promedio de las diferencias absolutas entre los valores reales y los valores previstos por el modelo. En otras palabras, el MAE mide la diferencia promedio entre lo que se espera y lo que el modelo ha predicho, en términos de

valor absoluto. A diferencia del MSE, el MAE no penaliza tanto las predicciones erróneas que están muy lejos de la verdad, por lo que es más robusto ante outliers (valores atípicos) (Valderrama Purizaca, 2020).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Donde:

Y: Valores predichos del modelo.

X: Valores Reales del modelo.

N: Cantidad de datos del modelo.

D) Desviación absoluta media (MAD)

Sirve para encontrar la desviación del valor predicho del valor real. Tanto las desviaciones negativas como las positivas son errores, por lo que se opta por la desviación absoluta. El promedio de todas las desviaciones absolutas se conoce como desviación absoluta media (MAD). Calculamos MAD en ambos tipos de datos: datos de entrenamiento y datos de prueba. Un buen modelo tendrá una MAD cercana a cero en ambos tipos de datos. MAD da la desviación promedio de las predicciones de los valores reales, donde se debe tener en cuenta la escala de los datos, es decir, un MAD de 100 para 1000 mil datos es aceptable pero no para 100 datos (Konasani & Kadre, 2021).

$$MAD = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

E) Porcentaje del error absoluto medio (MAPE)

Se realiza una modificación de la fórmula MAD, transformando cada desviación en un porcentaje del valor real. Si estamos interesados en conocer el porcentaje de desviación en lugar de la desviación real, entonces podemos utilizar MAPE. Con MAPE, no es necesario preocuparse

por la escala de la variable. Un valor MAPE del 2 por ciento siempre será más bajo que un valor MAPE del 10 por ciento, independientemente de la escala (Konasani & Kadre, 2021).

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - x_i|}{y_i}$$

F) Raíz del error cuadrático medio (RMSE)

Es una medida de evaluación de la precisión de un modelo de regresión. Es la raíz cuadrada de la media de los errores cuadráticos, es decir, la diferencia entre los valores predichos y los valores reales al cuadrado. Una RMSE más baja indica una menor dispersión de los errores y, por lo tanto, un mejor ajuste del modelo a los datos. En el contexto de la comparación de dos modelos, se prefiere aquel que tenga un valor más bajo de RMSE. Estas medidas se basan en diferentes fórmulas para medir los errores. Dependiendo de la organización o del profesional a cargo, se pueden utilizar diferentes medidas en distintos proyectos. Por lo tanto, es importante tener en cuenta las prácticas comunes al elegir una medida adecuada para evaluar la precisión de un modelo en particular (Konasani & Kadre, 2021).

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

3.1.5. Técnicas de validación del rendimiento y selección de modelos

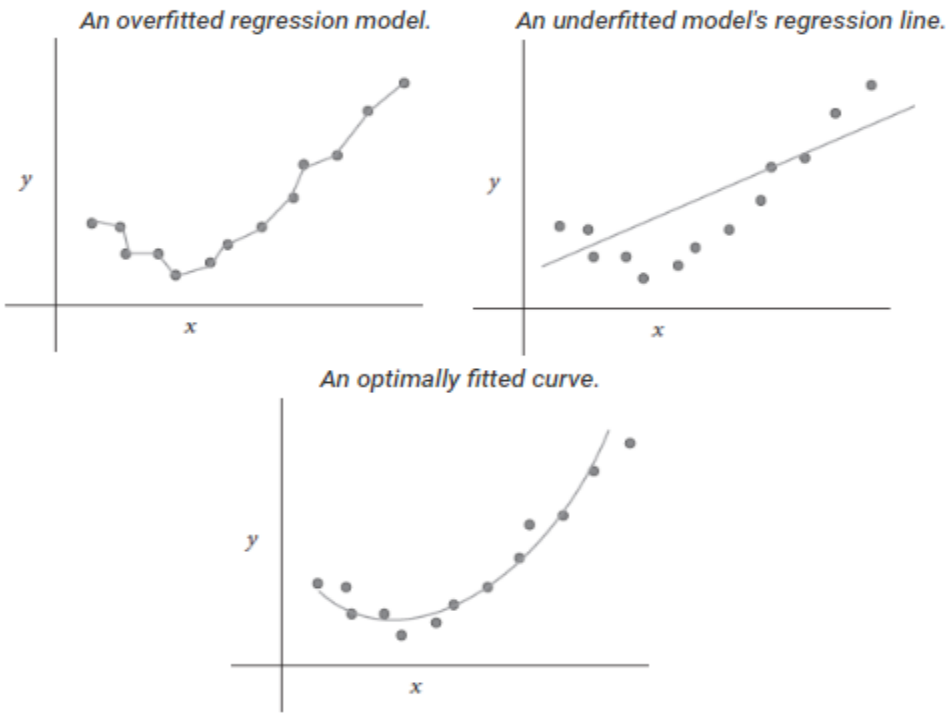
Antes de construir un modelo a partir de datos, es importante validar y explorar los datos para detectar problemas como valores atípicos o erróneos, y limpiar los datos antes de construir cualquier modelo. Es crucial validar exhaustivamente el modelo para asegurarse de que cumple con todas las condiciones de validación y demuestre una precisión constante con los datos entrenamiento, los datos de prueba y los conjuntos de datos de producción. Se utilizan métricas estándar de validación de modelos como R-cuadrado en regresión y precisión en modelos de clasificación, pero en algunos casos específicos, se pueden adoptar medidas alternativas de

validación. En algunas empresas la validación de modelos es una fase separada en el ciclo de vida de construcción de modelos y se realiza por un equipo independiente que supervisa y valida los modelos para garantizar que sean robustos y estén listos para su implementación (Konasani & Kadre, 2021).

A) Compensación de sesgo-varianza (Bias-Variance Trade-Off)

El modelo no debe estar sobre ajustado (overfitting) ni desajustado (underfitting); en otras palabras, el modelo no debe tener varianza (variance) ni sesgo (bias). En un modelo sobreajustado, la regresión pasa por todos los puntos. En un modelo insuficientemente ajustado, la línea de regresión pierde la mayoría de los puntos. En un modelo óptimo, la regresión pasa por la mayoría de los puntos y captura el patrón general en los datos (Konasani & Kadre, 2021).

Figura 11 Modelo de regresión sobre ajustado, desajustado y de ajuste optimo

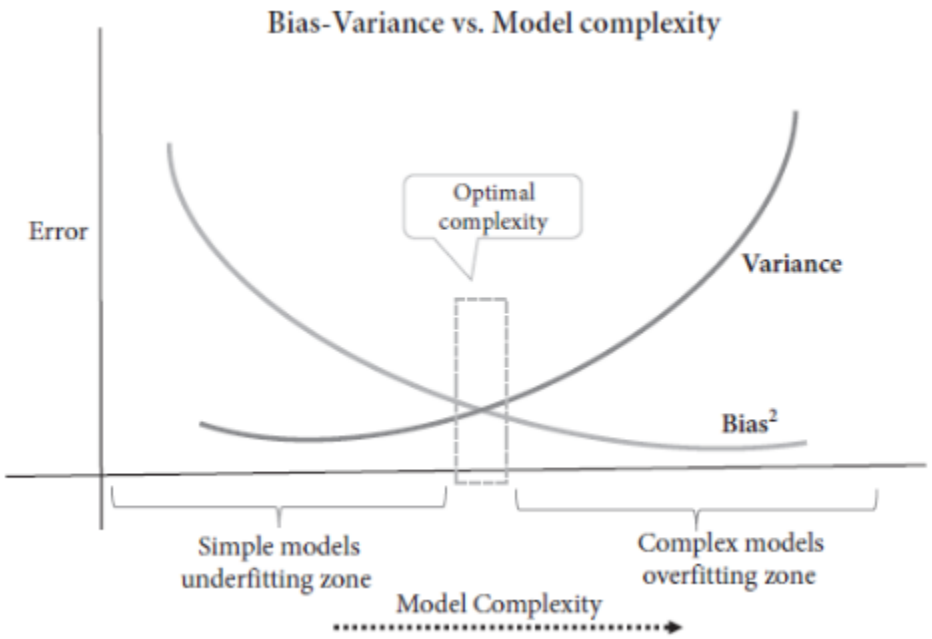


Fuente: (Konasani & Kadre, 2021)

En la siguiente figura se puede observar que a medida que se incrementa la complejidad del modelo, disminuye el sesgo o bias, pero aumenta su varianza. La varianza permanece baja

hasta un cierto punto, después del cual aumenta de manera significativa. Por ello, para conseguir un buen rendimiento de un modelo se debe considerar el punto de equilibrio entre las curvas de la varianza y el sesgo, es decir, se debe considerar los modelos con una complejidad optima.

Figura 12 Bias, varianza y complejidad optima del modelo



Fuente: (Konasani & Kadre, 2021)

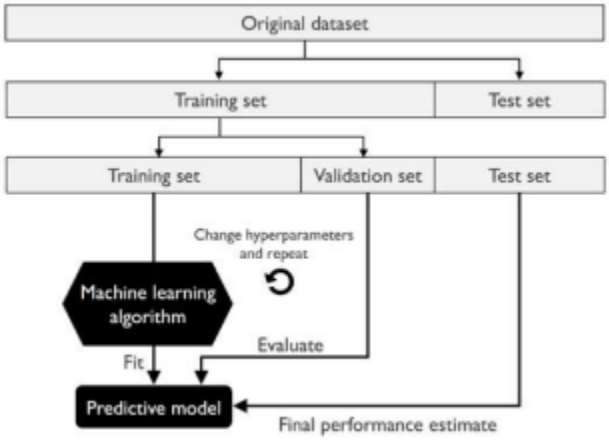
B) Validación cruzada (Cross-Validation)

Al construir un modelo de aprendizaje automático, es esencial evaluar su desempeño en datos que no ha visto antes. Si ajustamos nuestro modelo con un conjunto de datos de entrenamiento y lo utilizamos para evaluar su rendimiento en nuevos datos, es posible que sufra de desajuste (alto sesgo) si el modelo es demasiado simple o de sobreajuste (alta varianza) si el modelo es demasiado complejo para los datos de entrenamiento subyacentes. Para encontrar un equilibrio aceptable entre sesgo y varianza, es fundamental evaluar el modelo de manera cuidadosa mediante las técnicas comunes de validación cruzada, como la validación cruzada de retención (Holdout) y la validación cruzada k-fold, que ayudan a obtener estimaciones confiables del rendimiento de generalización del modelo, es decir, cómo se desempeña el modelo en datos que no ha visto antes (Raschka & Mirjalili, 2017).

- **Validación cruzada de retención (Holdout)**

La validación cruzada de retención es un método popular para estimar el rendimiento de generalización de los modelos de aprendizaje automático. Sin embargo, el uso del mismo conjunto de datos de prueba para la selección del modelo puede provocar un sobreajuste. Un mejor enfoque es usar un conjunto de datos de validación separado para la selección del modelo, además de los conjuntos de datos de entrenamiento y prueba. Esto permite una estimación menos sesgada de la capacidad del modelo para generalizar a nuevos datos. El conjunto de datos de validación se utiliza para evaluar el rendimiento del modelo después del entrenamiento con diferentes valores de hiperparámetros. Una vez que se seleccionan los hiperparámetros óptimos, el rendimiento de generalización del modelo se estima utilizando el conjunto de datos de prueba (Raschka & Mirjalili, 2017).

Figura 13 Validación cruzada de retención



Fuente: (Raschka & Mirjalili, 2017)

- **Validación cruzada de K-fold**

La validación cruzada de K-fold implica dividir aleatoriamente el conjunto de datos de entrenamiento en k pliegues, donde los pliegues k-1 se usan para el entrenamiento del modelo y un pliegue se usa para la evaluación del rendimiento. Este proceso se repite k veces para obtener k modelos y estimaciones de rendimiento. A continuación, se calcula el rendimiento medio para obtener una estimación menos sesgada. La validación cruzada de

K-fold se usa normalmente para el ajuste del modelo, encontrando valores de hiperparámetro óptimos que producen un rendimiento de generalización satisfactorio. Un buen valor estándar para k es 10, que ofrece la mejor compensación entre el sesgo y la varianza, pero el número de pliegues se puede aumentar o disminuir según el tamaño del conjunto de datos (Raschka & Mirjalili, 2017).

Figura 14 Validación cruzada de K-fold

Fuente: (Raschka & Mirjalili, 2017)

C) Grid Search

Ajuste fino de modelos de aprendizaje automático a través de la búsqueda de cuadrícula (Grid Search). En aprendizaje automático existen dos tipos de parámetros: aquellos que se aprenden a partir de los datos de entrenamiento y los parámetros de un algoritmo de aprendizaje que se optimizan por separado. Estos últimos son los parámetros de ajuste (o hiperparámetros) de un modelo. Para encontrar la combinación óptima de valores de hiperparámetros y mejorar el rendimiento del modelo, se utiliza una técnica llamada búsqueda en cuadrícula (grid search), que consiste en evaluar el modelo con diferentes combinaciones de valores de hiperparámetros y seleccionar aquellos que den el mejor rendimiento (Raschka & Mirjalili, 2017).

Ajustar los hiperparámetros de un modelo puede implicar modificar configuraciones manualmente para lograr resultados óptimos. Este proceso es tedioso y es posible que no haya tiempo para probar muchas variaciones. Afortunadamente, existe una herramienta llamada GridSearchCV de Scikit-Learn que automatiza esta búsqueda. Simplemente le dices qué configuraciones quieres explorar y qué valores probar, y revisará eficientemente todas las combinaciones utilizando una técnica de validación como la validación cruzada para evaluar su efectividad (Géron, 2019).

D) Randomized Search

Cuando se buscan los mejores valores para los hiperparámetros de un modelo de aprendizaje automático, GridSearchCV es una herramienta útil para explorar un número limitado de combinaciones. Sin embargo, si el espacio de búsqueda es grande, RandomizedSearchCV ofrece una mejor alternativa. Esta herramienta evalúa un conjunto de combinaciones aleatorias, permitiendo una mayor exploración y un control más preciso del tiempo dedicado a la búsqueda (Géron, 2019).

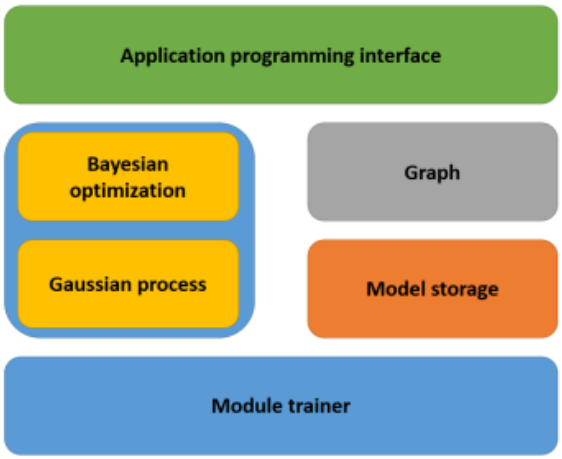
E) Auto Machine Learning

Construir modelos de aprendizaje automático puede ser complicado, costoso y poco eficiente. El Auto-ML (Auto Machine Learning) enfrenta esto automatizando varios pasos como la selección del algoritmo, el ajuste de hiperparámetros y la optimización de la arquitectura del modelo. Esto ahorra tiempo y reduce recursos como la necesidad de mano de obra de expertos en aprendizaje automático. Las herramientas Auto-ML utilizan técnicas de búsqueda y optimización para encontrar el mejor modelo para una tarea específica. Incluso pueden aprovechar el esemble learning y transfer learning para mejorar el rendimiento. Si bien Auto-ML es poderoso, los datos de entrenamiento de alta calidad siguen siendo esenciales para obtener resultados óptimos. Actualmente, se continúa investigando en la generación de datos automática, ya que estas herramientas no pueden generar nuevos datos para tareas sin conjuntos de entrenamiento existentes, lo que limita su capacidad para automatizar completamente la adquisición de datos y así lograr un rendimiento del modelo aún mejor (Hottong et al., 2024).

- **Auto-Keras**

El Auto-Keras es una herramienta de Auto-ML que aborda el desafío de Neural Architecture Search (Conocido como NAS) mediante el uso de técnicas innovadoras como el morfismo de red y la optimización bayesiana. Al aprovechar el morfismo de red, Auto-Keras permite modificar la arquitectura de una red neuronal manteniendo su funcionalidad, lo que reduce significativamente el tiempo medio de entrenamiento durante la búsqueda de arquitecturas. Además, la integración de la optimización bayesiana proporciona una estrategia eficiente para explorar el espacio de búsqueda y seleccionar operaciones de morfismo de red de manera más inteligente, lo que mejora la eficacia del proceso de AutoML (Jin et al., 2019).

Figura 15 Arquitectura de la Herramienta de Auto-ML: AutoKeras



Fuente: (Alaiad et al., 2023)

La arquitectura de AutoKeras, aprovecha eficientemente los recursos de CPU y GPU al almacenar únicamente los datos actualmente en uso en la RAM, mientras que el resto se guarda en dispositivos de almacenamiento. La API se utiliza para invocar módulos de funcionalidad específica, mientras que el algoritmo de búsqueda de AutoKeras, basado en optimización bayesiana, controla las operaciones de morfismo de red en la CPU, utilizando un proceso gaussiano y un optimizador bayesiano para guiar la búsqueda. Además, la unidad de entrenamiento de modelos opera en la GPU para lograr el paralelismo en el proceso de entrenamiento de redes neuronales. Los modelos entrenados se almacenan

en dispositivos de almacenamiento debido a su gran tamaño y la dificultad de almacenarlos completamente en la memoria (Alaiad et al., 2023).

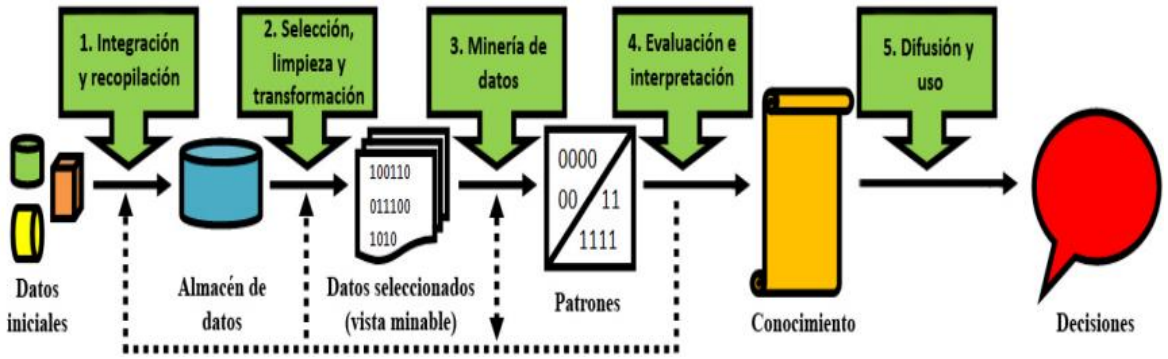
- **Auto-WEKA**

  Auto-WEKA es una extensión del popular software de aprendizaje automático WEKA, facilita la selección del algoritmo y la configuración de hiperparámetros adecuados de forma automatizada para un conjunto de datos. Auto-WEKA emplea la optimización bayesiana para navegar por el espacio combinado de los algoritmos de aprendizaje de WEKA y sus hiperparámetros, con el objetivo de minimizar la pérdida de validación cruzada. Esta optimización conjunta, denominada Combined Algorithm Selection and Hyperparameter Optimization (CASH por sus siglas en ingles), es similar a resolver un problema de optimización de caja negra, donde el objetivo es identificar la configuración óptima de algoritmo e hiperparámetros para un conjunto de datos dado. La optimización bayesiana, impulsada por un modelo probabilístico, explora iterativamente este espacio, siendo los modelos basados en árboles efectivos para el espacio altamente condicional de los algoritmos y hiperparámetros de WEKA. Auto-WEKA ofrece ejecuciones paralelas, interfaces de línea de comandos y gráficas para facilitar su uso (Kotthoff et al., 2017).

3.1.6. Metodología KDD

La metodología KDD (Knowledge Discovery from Databases) es un proceso que se repite en diferentes etapas para obtener información útil y práctica a partir de grandes cantidades de datos.

Figura 16 Proceso de la metodología KDD



Fuente: (Galán Cortina, 2015)

- Recopilación e integración de datos: Se seleccionan las fuentes de información relevantes, los datos se transforman a un formato común (almacén de datos).
- Selección, limpieza y transformación: Se corrigen o eliminan datos erróneos o incompletos, se seleccionan los datos relevantes para el análisis.
- Minería de datos: Se elige la tarea a realizar (predecir, agrupar, clasificar, etc.), se selecciona y aplica el método de minería de datos adecuado.
- Evaluación e interpretación: Se evalúan e interpretan los resultados de la minería de datos, se busca comprender el conocimiento obtenido.
- Difusión y uso: Se comparte el conocimiento adquirido con los usuarios, se implementa el conocimiento en aplicaciones prácticas (Galán Cortina, 2015).

3.2. Mecánica de suelos

La mecánica es la rama de la ciencia física que estudia la acción de las fuerzas sobre los cuerpos. De manera similar, la mecánica de suelos es la parte de la mecánica que se ocupa de la acción de las fuerzas sobre la masa de los suelos. Últimamente, los significados de la geotecnia y la ingeniería geotécnica se han incorporado a la terminología del suelo, donde a menudo se aplican los principios y aplicaciones de la mecánica del suelo y la geología y la mecánica de rocas (Crespo Villalaz, 2004).

3.2.1. Clasificación de los suelos

a) Clasificación de los suelos mediante el sistema SUCS

La mecánica de suelos ha desarrollado métodos para clasificar los suelos, ya que la naturaleza presenta variedades de suelo, dependiendo del tamaño de las partículas y la plasticidad que presenta.

Figura 17 Tipos de suelos según el sistema SUCS

Símbolo	G	S	M	C	O	Pt	H	L	W	P
Descripción	Grava	Arena	Limo	Arcilla	Limos orgánicos y arcilla	Turba y suelos altamente orgánicos	Alta plasticidad	Baja plasticidad	Bien graduado	Mal graduado

Fuente: (Das, 2012)

El Sistema Unificado de Clasificación de Suelos fue propuesto originalmente por A. Casagrande en 1942, luego revisado y adoptado por la Oficina de Reclamación de EE. UU. y el Cuerpo de Ingenieros del Ejército de EE. UU. Hoy en día el sistema se utiliza en prácticamente todos los trabajos geotécnicos (Das, 2012).

Figura 18 Clasificación de los suelos según SUCS

	Criterios para asignar símbolos y nombres de grupo utilizando pruebas de laboratorio[a]			Clasificación del suelo	
				Símbolo de grupo	Nombre de grupo[b]
Suelos de grano grueso Más de 50% retenido en la malla núm. 200	Gravas Más de 50% de la fracción gruesa retenida en la malla núm. 4	Gravas limpias Menos de 5% finos[c]	$C_u \geq 4$ y $1 \leq C_c \leq 3$[e]	GW	Grava bien graduada[f]
			$C_u < 4$ y/o $1 > C_c > 3$[e]	GP	Grava mal graduada[f]
		Gravas con finos Más de 12% finos[c]	Los finos se clasifican como ML o MH	GM	Grava limosa[f,g,h]
			Los finos se clasifican como CL o CH	GC	Grava arcillosa[f,g,h]
	Arenas 50% o más de la fracción gruesa pasa la malla núm. 4	Arenas limpias Menos de 5% finos[d]	$C_u \geq 6$ y $1 \leq C_c \leq 3$[e]	SW	Arena bien graduada[i]
			$C_u < 6$ y/o $1 > C_c > 3$[e]	SP	Arena mal graduada[i]
		Arena con finos Más de 12% finos[d]	Los finos se clasifican como ML o MH	SM	Arena limosa[g,h,i]
			Los finos se clasifican como CL o CH	SC	Arena arcillosa[g,h,i]
Suelos de grano fino 50% o más pasa la malla núm. 200	Limos y arcillas Límite líquido menor que 50	Inorgánicos	IP > 7 y se encuentra en o arriba de la línea "A"[j]	CL	Arcilla de baja compresibilidad[k,l,m]
			IP < 4 o se encuentra debajo de la línea "A"[j]	ML	Limo de baja compresibilidad[k,l,m]
		Orgánicos	$\frac{\text{Límite líquido—secado en horno}}{\text{Límite líquido—no secado}} < 0.75$	OL	Arcilla orgánica[k,l,m,n] / Limo orgánico[k,l,m,o]
	Limos y arcillas Límite líquido 50 o mayor	Inorgánicos	IP se encuentra en o arriba de la línea "A"	CH	Arcilla de alta compresibilidad[g,l,m,n]
			IP se encuentra debajo de la línea "A"	MH	Limo de alta compresibilidad[k,l,m,n]
		Orgánicos	$\frac{\text{Límite líquido—secado en horno}}{\text{Límite líquido—no secado}} < 0.75$	OH	Arcilla orgánica[k,l,m,p] / Limo orgánico[k,l,m,q]
Suelos altamente orgánicos	Principalmente materia orgánica, de color oscuro y olor orgánico			PT	Turba

Fuente: (Das, 2012)

- **Suelos gruesos**

Los suelos gruesos son gravas y arenas, en donde si más de la mitad de la fracción gruesa es retenida por la malla no. 4, pertenece al grupo de grava (G), de lo contrario

pertenecerá al grupo de arena (S). Estos dos grupos se dividen en otros cuatro grupos que son respectivamente GM, GP, GW, GC y SW, SP, SM, SC para gravas y arenas, debido a que estas pueden tener porcentajes de finos. Los símbolos GW y GP denotan gravas de buena calidad y gravas de mala o mala calidad respectivamente, lo mismo para aquellas con finos, que son GM, que denota gravas francas y GC, denota gravas arcillosas. Lo mismo se aplica a las arenas SW, SP, SM y SC, que designan respectivamente arenas de grano fino, arenas de grano pobre, arenas francas y arenas arcillosas (Crespo Villalaz, 2004, pág. 92).

- **Suelos finos**

  Los suelos finos se clasifican en limos y arcillas, se agrupan en tres grupos con un límite líquido por debajo del 50%, aquellos con un límite líquido por encima del 50% también se agrupan en tres grupos, y un grupo para suelos finos muy orgánicos. Si el suelo tiene un límite de liquidez inferior al 50%, es decir, si el suelo tiene una compresibilidad baja o media, se añadirá el símbolo L a los prefijos M, C y O, obteniendo los grupos ML, CL y OL que se denominarán limo respectivamente, arcillas inorgánicas de baja compresibilidad, arcillas inorgánicas de baja compresibilidad y limos orgánicos de baja compresibilidad. Para las turberas que son suelos altamente orgánicos, se designan con el símbolo Pt (Crespo Villalaz, 2004, pág. 93).

b) Clasificación de los suelos mediante el sistema AASHTO

El Sistema de clasificación de suelos AASHTO se propuso originalmente al Comité de clasificación de materiales de subbase y tipos de carreteras del Consejo de investigación de carreteras (1945). Según la forma actual de este sistema, los suelos se pueden clasificar en ocho grupos principales, del A-1 al A-8, según su distribución granulométrica, su límite líquido y su índice de plasticidad. Los suelos enumerados en los grupos A-1, A-2 y A-3 son materiales de grano grueso, mientras que los de los grupos A-4, A-5, A-6 y A-7 son materiales de grano fino. La turba, el lodo y otros suelos altamente orgánicos se clasifican en el grupo A-8 y se identifican mediante inspección visual (Das, 2012).

Figura 19 Clasificación de los suelos según AASHTO A1-A2

Clasificación general	Materiales granulares (35% o menos de la muestra total pasa la malla núm. 200)						
	A-1			A-2			
Clasificación de grupo	A-1-a	A-1-b	A-3	A-2-4	A-2-5	A-2-6	A-2-7
Análisis por mallas (% que pasa)							
Malla núm. 10	50 máx						
Malla núm. 40	30 máx	50 máx	51 mín				
Malla núm. 200	15 máx	25 máx	10 máx	35 máx	35 máx	35 máx	35 máx
Para la fracción que pasa Malla núm. 40							
Límite líquido (LL)				40 máx	41 mín	40 máx	41 mín
Índice de plasticidad (IP)	6 máx		No plástico	10 máx	10 máx	11 mín	11 mín
Tipo usual de material	Fragmentos de roca, grava y arena		Arena fina	Grava y arena limosa o arcillosa			
Clasificación de la capa	Excelente a buena						

Fuente: (Das, 2012)

Figura 20 Clasificación de los suelos según AASHTO A4-A7

Clasificación general	Materiales de limo y arcilla (más de 35% de la muestra total pasa la malla núm. 200)			
Clasificación de grupo	A-4	A-5	A-6	A-7
				A-7-5[a] A-7-6[b]
Análisis por mallas (% que pasa)				
Malla núm. 10				
Malla núm. 40				
Malla núm. 200	36 mín	36 mín	36 mín	36 mín
Para la fracción que pasa Malla núm. 40				
Límite líquido (LL)	40 máx	41 mín	40 máx	41 mín
Índice de plasticidad (IP)	10 máx	10 máx	11 mín	11 mín
Tipo usual de material	Principalmente suelos limosos		Principalmente suelos arcillosos	
Calificación subrasante	Regular a malo			

[a]Si IP ≤ LL − 30, la clasificación es A-7-5.
[b]Si IP > LL − 30, la clasificación es A-7-6.

Fuente: (Das, 2012)

3.2.2. Propiedades físicas de los suelos

De acuerdo con la norma de Suelos y Cimentaciones (E.050, 2015), se establece que se deben realizar ensayos de campo y laboratorio de acuerdo con las normas técnicas de la NTP peruana para determinar las propiedades del suelo. En mecánica de suelos, conocer las propiedades físicas claves de un suelo es muy importante, ya que la interpretación correcta de los resultados puede predecir el comportamiento futuro de un suelo determinado bajo cargas cuando dicho suelo presente diferentes contenidos de humedad (Crespo Villalaz, 2004).

a) Humedad natural del suelo

De acuerdo con la Norma Técnica Peruana (NTP 339.127, 2006), define humedad natural o contenido de humedad del suelo como la cantidad de agua presente en el suelo. Se expresa como una relación porcentual entre el peso del agua en una masa de suelo dada y el peso de las partículas sólidas.

b) Análisis granulométrico del suelo

El tamaño de las partículas varía considerablemente dentro de una determinada masa de suelo. Para clasificar adecuadamente el suelo, necesitamos conocer la distribución del tamaño de sus partículas. La distribución del tamaño de grano de los suelos gruesos generalmente se determina mediante el análisis del tamaño de malla. Para suelos de grano fino, la distribución del tamaño de las partículas se puede obtener mediante análisis hidrométrico (Das, 2012).

Según la Norma Técnica Peruana (NTP 339.128, 2006), Muestra la distribución de partículas de suelo por tamiz estandarizado. Se debe tener en cuenta que esta distribución de partículas es útil para clasificar suelos utilizando los sistemas SUCS o AASHTO. El análisis del tamaño de las partículas del suelo las clasifica según su tamaño de partícula. Consiste en remover gradualmente la muestra a través de un tamiz estandarizado con pequeñas aberturas (MTC, 2016).

Figura 21 Tamaño de partículas

TIPO DE MATERIAL		TAMAÑO DE PARTICULAS (mm.)
Grava		75 – 4.75
Arena		Arena gruesa: 4.75 – 2.00
		Arena media: 2.00 – 0.425
		Arena fina: 0.425 – 0.075
Material Fino	Limo	0.075 – 0.005
	Arcilla	Menor a 0.005

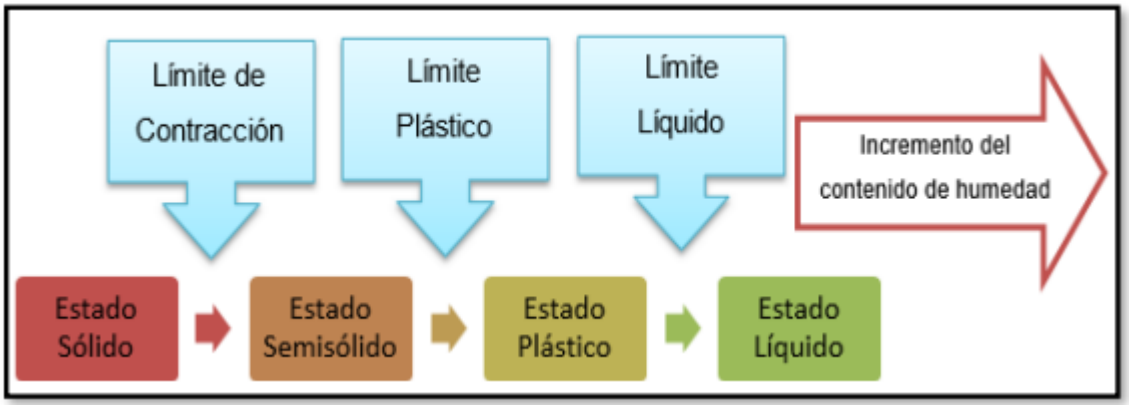
Fuente: (Manrique Dávila, 2019)

c) Consistencia del suelo

La consistencia de los suelos establece que, con un contenido de agua muy bajo, el suelo se comporta como un sólido friable, pero con un contenido de agua muy alto, el suelo y el agua pueden hundirse como un líquido. Por tanto, el comportamiento natural de los suelos se puede dividir en cuatro estados básicos: sólido, semisólido, plástico y líquido (Das, 2012).

Los límites de líquido, plasticidad y contracción a menudo se denominan límites de Atterberg, y sus respectivos contenidos de agua distinguen los límites de diferentes estados de consistencia en suelos cohesivos (MTC, 2016).

Figura 22 Límites de Atterberg



Fuente: (Manrique Dávila, 2019)

Según el MTC 2016 la resistencia a la compresión no confinada o uniaxial también utiliza para clasificar la consistencia del suelo en muy blando, blando, medio, compacto, muy compacto y duro, según el valor obtenido de la siguiente forma:

Figura 23 Consistencia del suelo y resistencia a la compresión confinada

Consistencia del suelo	Resistencia a la Compresión no confinada	
	kg/cm²	(kPa)
Muy blanda	< 0,25	(< 25)
Blanda	0,25-0,50	(25- 50)
Mediana	0,50-1,00	(50-100)
Firme	1,00-2,00	(100-200)
Muy firme	2,00-4,00	(200-400)
Dura	> 4,00	(> 400)

Fuente: (MTC, 2016)

- **Límite líquido (LL)**

  Según la Norma Técnica Peruana (NTP 339.129, 2006), define el límite líquido como la humedad presente en el suelo. Se encuentra entre los límites de los estados líquido y plástico. Esta humedad se mide colocando un terrón de tierra en la Copa de Casagrande separado por ranuras, dejando caer el vaso desde una altura de 1 cm, cerrando las ranuras 1,3 cm y golpeándolo 25 veces con la cuchara. base. El cálculo del límite líquido se realiza probando varios contenidos de humedad. Utiliza un método gráfico para determinar la humedad en 25 golpes en un gráfico semilogarítmico (humedad según el número de golpes).

  Según Day (2010) define el límite líquido como el contenido de agua correspondiente al cambio de comportamiento entre el estado líquido y plástico del limo o arcilla. El límite líquido se coloca en una copa de latón ligeramente sucia, se parte por la mitad con una herramienta ranurada y luego se coloca en un límite líquido estándar. En términos específicos, el límite líquido se define como el contenido de humedad del suelo cortado por una herramienta de ranurado que se desliza una distancia de 0,5 pulgadas. (12,7 mm) 25 choques en un limitador de líquido estándar. Para obtener más información sobre las pruebas de laboratorio, consulte la norma ASTM D 4318-00

(2004), "Métodos de prueba estándar para el límite líquido, el límite plástico y el índice de plasticidad de los suelos" o la Norma Técnica Peruana (NTP 339.129, 2006).

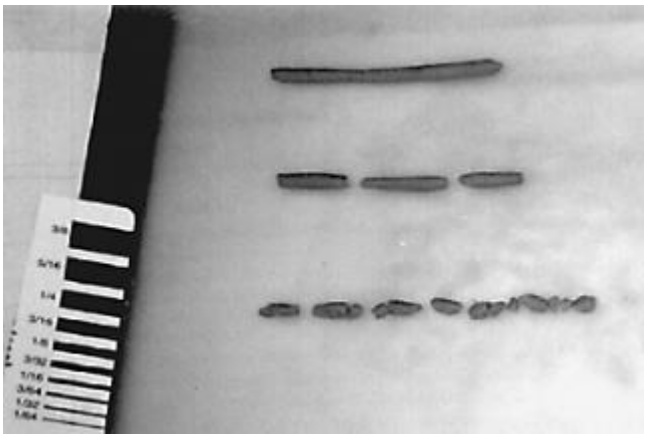
Figura 24 Prueba del Limite Líquido

Fuente: (Day, 2010)

- **Límite plástico (LP)**

Según Day (2010) define el límite plástico como el contenido de agua correspondiente al cambio de comportamiento entre el estado plástico y semisólido de un limo o arcilla. El límite plástico se determina presionando y haciendo rodar una pequeña porción del suelo plástico para que su contenido de agua se reduzca lentamente con el resultado final de que el hilo del suelo se desmorone. En términos específicos, el límite plástico se define como el contenido de agua en el que un limo o arcilla comenzará a desmoronarse cuando se enrolle en un hilo de aproximadamente 1/8 pulg. (3,2 mm) de diámetro. Para obtener detalles sobre las pruebas de laboratorio, consulte ASTM D 4318-00 (2004), “Métodos de prueba estándar para el límite líquido, el límite plástico y el índice de plasticidad de los suelos” o la Norma Técnica Peruana (NTP 339.129, 2006).

Figura 25 Prueba del Limite Plástico



Fuente: (Day, 2010)

El límite de plasticidad se define como el contenido de humedad del suelo, y si se producen grietas o roturas al formar un cilindro de suelo con un diámetro de 3 mm, la muestra de suelo utiliza malla pasante número 40, y el ensayo del límite líquido es el mismo para el límite plástico (NTP 339.129, 2006).

- **Índice de plasticidad (IP)**

Según Day (2010) el índice de plasticidad es la diferencia entre el límite líquido y el límite plástico se denomina índice de plasticidad del suelo, denotado por la siguiente formula:

$$IP = LL - LP$$

Donde:

- PI = índice de plasticidad de un suelo cohesivo
- LL = límite líquido determinado a partir de la prueba de límite líquido
- PL = límite plástico determinado a partir del ensayo de límite plástico

Asimismo, el índice de plasticidad describe el comportamiento del suelo en función de su contenido de agua o humedad, para que pueda funcionar resistiendo la deformación sin hundimientos y fallas.

Figura 26 Característica del suelo según índice plasticidad

ÍNDICE DE PLASTICIDAD	CARACTERÍSTICAS
IP >20	Suelos muy arcillosos
20>IP >10	Suelos arcillosos
10>IP >4	Suelos poco arcillosos
IP=0	Suelos exentos de arcillas

Fuente: (Manrique Dávila, 2019)

- **Límite de contracción**

Según la (NTP 339.129, 2006), el límite de contracción se define como la humedad máxima del suelo en la que una disminución de la humedad no provoca un cambio en el volumen del suelo. Así mismo, Day (2010) menciona que el límite de contracción es el contenido de agua correspondiente al cambio de comportamiento entre el estado semisólido a sólido de un limo o arcilla. El límite de contracción también se define como el contenido de agua en el cual cualquier reducción adicional en el contenido de agua no resultará en una disminución en el volumen de la masa del suelo. El límite de contracción rara vez se obtiene en la práctica debido a las dificultades de las pruebas y métodos de laboratorio. Para obtener detalles sobre las pruebas, consulte ASTM D 427-98 (2004), "Método de prueba estándar para factores de contracción de suelos por el método de mercurio", o D 4943-02, 2004, "Método de prueba estándar para factores de contracción de suelos por el método de la cera" (Day, 2010).

3.2.3. Propiedades mecánicas de los suelos

Las propiedades mecánicas de los suelos se definen como aquellas obtenidas a partir de ensayos relacionados con la resistencia, deformabilidad y estabilidad, tales como ensayos de compactación o compresión, ensayos de corte y entre otros (Manrique Dávila, 2019).

a) Compactación de suelos

La compactación de los suelos es la mejora artificial de las propiedades mecánicas del suelo por medios mecánicos. Se diferencia de la consolidación del suelo en que el peso específico del material aumenta paulatinamente bajo la acción natural de las sobrecargas impuestas que provocan la expulsión del agua a través de un proceso de difusión, ambos procesos implican una disminución de volumen, por lo que al final son equivalentes (Juárez Badillo & Rico Rodríguez, 2012).

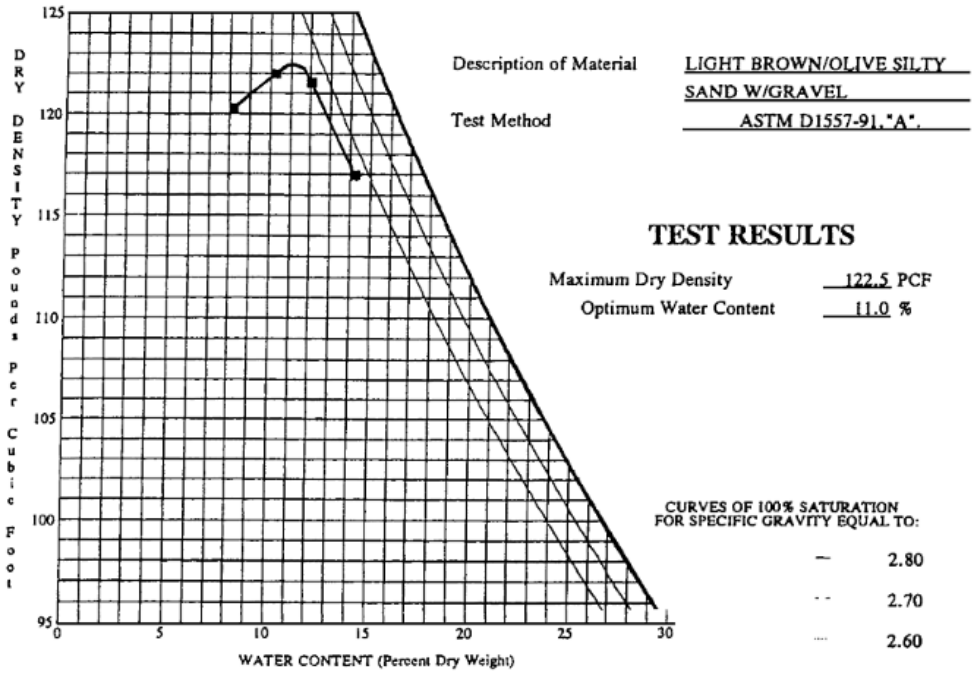
La prueba de compactación de laboratorio consiste en compactar un suelo con un contenido de agua conocido en un molde de dimensiones específicas utilizando una cierta energía de compactación. El procedimiento se repite para varios contenidos de agua para establecer la curva de compactación. Los procedimientos de prueba más comunes (energía de compactación, número de capas de suelo en el molde y similares) son el Proctor modificado (ASTM D 1557-02, 2004) y el Proctor estándar (ASTM D 698-00, 2004). El término "Proctor" es en honor a R. R. Proctor, quien en 1933 demostró que la densidad seca de un suelo para un esfuerzo de compactación dado depende de la cantidad de agua que contiene el suelo durante la compactación (Day, 2010).

- **Prueba de Proctor Modificado**

  Para el Proctor modificado (ASTM D 1557-02, 2004, procedimiento A), el suelo se compacta en una capa de 10,2 cm de diámetro que tiene un volumen de 944 cm3, donde se compactan cinco capas de suelo en el molde y cada capa recibe 25 golpes de un apisonador de 44,5 N que tiene 0,46 m de caída. El Proctor modificado tiene una energía de compactación de 2700 kN-m/m3. El procedimiento de prueba consiste en preparar el suelo con un cierto contenido de agua, compactar el suelo en el molde y luego, al registrar la masa de suelo dentro del molde, se obtiene la Densidad Húmeda del Suelo Compactado. Al conocer el contenido de agua del suelo compactado, se puede calcular la densidad seca, dividiendo la densidad húmeda entre el contenido de agua en forma decimal aumentando en uno. Este procedimiento de compactación se repite para el suelo con diferentes contenidos de agua y luego la densidad seca frente al contenido de agua se traza en un gráfico para obtener la curva de compactación. El punto máximo de la curva de compactación se conoce como la Máxima Densidad Seca (MDD) de

laboratorio (Mg/m3). El contenido de agua correspondiente al punto máximo de la curva de compactación de laboratorio se conoce como el Óptimo Contenido de Humedad (OMC) (Day, 2010).

Figura 27 Curva de compactación para una arena limosa basada en las especificaciones de la prueba Proctor modificada



Fuente: (Day, 2010)

Las tres líneas a la derecha de la curva de compactación se conocen como “curva de cero vacíos de aire”. Estas curvas representan la relación entre el contenido de agua y la densidad seca para una condición de saturación (S = 100 por ciento) para una gravedad específica especificada. Por lo general, el lado derecho de la curva de compactación será aproximadamente paralelo a la curva de cero vacíos de aire y se puede usar como control de los resultados de las pruebas de laboratorio. La siguiente ecuación se puede utilizar para trazar una "curva de vacíos de aire cero" (Day, 2010).

$$\rho_z = \frac{G_s \rho_w}{1 + G_s w}$$

Donde:

- ρz = valor del eje vertical de la densidad seca del suelo para el trazado de la curva de vacíos de aire cero (Mg/m3).
- Gs = gravedad específica de los sólidos del suelo (adimensional).
- ρw = densidad del agua, que es igual a 62,4 1,0 Mg/m3.
- w = contenido de agua del suelo, expresado como decimal.

- **Prueba de Proctor Estándar**

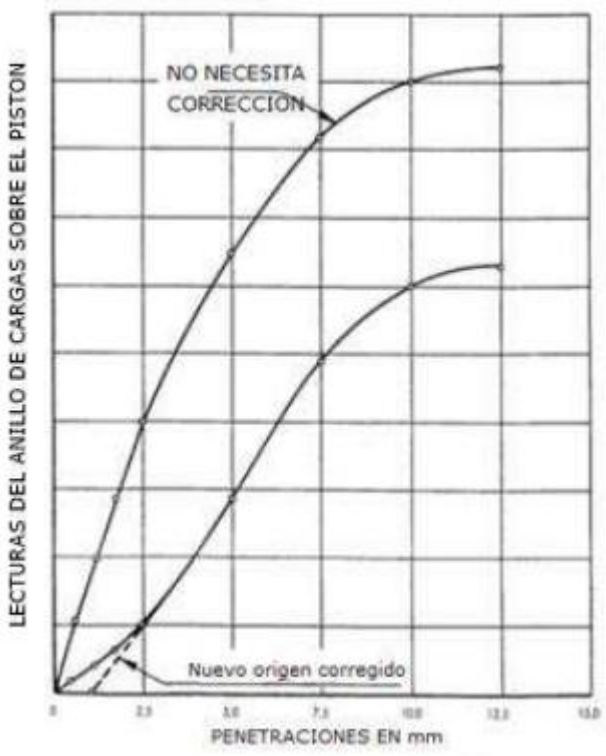
Para el Standard Proctor (ASTM D 698-00, 2004, Procedimiento A), el suelo se compacta en un molde de 10,2 cm de diámetro que tiene un volumen de 944 cm3, donde tiene tres capas de suelo y se compactan en el molde y cada capa recibe 25 golpes de un apisonador de 24,4 N que tiene una caída de 0,31 m. Por lo tanto, una diferencia entre las pruebas de compactación de laboratorio es que el Proctor modificado tiene cinco capas de suelo compactado, mientras que el Proctor estándar tiene solo tres capas de suelo compactado. Para ambas pruebas de laboratorio, cada capa se somete a 25 golpes, pero el apisonador Proctor modificado imparte mucha más energía en comparación con la prueba Proctor estándar que utiliza un apisonador diferente (Day, 2010).

b) Capacidad de soporte CBR (California Bearing Ratio)

Este método de prueba se utiliza para evaluar la resistencia potencial o capacidad de soporte de suelos subrasante, cimentación, material de base, incluidos los materiales reciclados para pavimentar caminos, campos aterrizaje. El valor de CBR (California Bearing Ratio) obtenido en esta prueba es una parte integral de varios métodos de diseño de pavimento flexible (MTC, 2016). De acuerdo con la norma técnica peruana (NTP 339.145, 2006), la cual indica que el valor de la relación de soporte conocida como CBR es un índice de resistencia del suelo, que se refiere a que el suelo está sometido a la penetración de un eje a una velocidad constante, el resultado del índice CBR nos dice que esta es la capacidad de carga del suelo en comparación con la de una grava estándar o de un suelo patrón. Se determinan dos valores de CBR para cada muestra de suelo, uno a 0,1 pulgada y otro a 0,2 pulgada de penetración. ASTM D-1883 recomienda usar CBR con una

penetración de 0,1 pulgadas. siempre que este índice sea inferior a 0,2 pulgadas. Si el valor de CBR para 0,1 pulgadas es superior a 0,2 pulgadas, se debe repetir la prueba para esa muestra.

Figura 28 Curva para el cálculo del índice del CBR



Fuente: (MTC, 2016)

En base a la gráfica de la curva para el cálculo del índice del CBR se puede calcular el CBR con la siguiente formula:

$$CBR = \frac{Presión\ en\ muestra\ problema}{Presión\ en\ muestra\ patron} \times 100$$

Fuente: (Manrique Dávila, 2019)

Los datos y resultados de la prueba que deberán suministrarse son los siguientes:

- Método usado para la preparación y compactación de los especímenes.

- Descripción e identificación de la muestra ensayada.
- Humedad al fabricar el espécimen.
- Peso unitario.
- Sobrecarga de saturación y penetración.
- Expansión del espécimen.
- Humedad después de la saturación.
- Humedad óptima y densidad máxima determinados mediante la norma MTC E 115.
- Curva presión-penetración.
- Valor de relación de soporte CBR. (MTC, 2016).

3.2.4. Ensayos de laboratorio de la mecánica de suelos

Las pruebas de laboratorio generalmente comienzan una vez que se completa la exploración del subsuelo. El primer paso en las pruebas de laboratorio es registrar todos los materiales (suelo, roca o agua subterránea) recuperados de la exploración del subsuelo. Luego, el ingeniero geotécnico y el ingeniero geólogo preparan un programa de pruebas de laboratorio, que básicamente consiste en asignar pruebas de laboratorio específicas para las muestras de suelo. Técnicos experimentados, que están bajo la supervisión del ingeniero geotécnico, a menudo realizan las pruebas de laboratorio reales de las muestras de suelo. Debido a que las muestras de suelo pueden secarse o puede haber cambios en la estructura del suelo con el tiempo, es importante realizar las pruebas de laboratorio lo antes posible (Day, 2010).

Las pruebas de laboratorio deben realizarse de acuerdo con los procedimientos estándar, como los recomendados por la Sociedad Estadounidense para Pruebas y Materiales (ASTM) o los procedimientos enumerados en los libros de texto estándar, manuales de especificaciones o normas técnicas.

Se presenta una lista de pruebas de laboratorio de suelo comunes utilizadas en ingeniería geotécnica y estudios de mecánica de suelos según la norma E.050 (2015).

Tabla 01 Lista de principales ensayos de laboratorio de mecánica de suelos

TIPO DE CONDICIÓN	PROPIEDADES DE LOS SUELOS	ESPECIFICACIÓN
Pruebas de índice	Prueba de contenido de agua (contenido de humedad)	ASTM D 2216–98 y D 4643–00
	Prueba de gravedad específica	ASTM D 854–02 y D 5550–00
	Densidad relativa	ASTM D 4253–00 y D 4254–00
Tamaño de partículas y límites de Atterberg	Análisis granulométrico	ASTM D 422–02
	Prueba de hidrómetro	ASTM D 422–02
	Prueba de límites de Atterberg	ASTM D 4318–00
	Clasificación de suelos (USCS)	ASTM D 2487–00
Asentamiento	Prueba de colapso	ASTM D 5333–03
	Prueba de consolidación	ASTM D 2435–03
Suelo expansivo	Prueba de índice de expansión	ASTM D 4829–03
	Prueba de hinchamiento de HUD	HUD especificaciones (1971)
	Prueba de hinchamiento intacto	ASTM D 4546–03
	Prueba de edómetro (método C)	ASTM D 4546–03
Ensayos de resistencia al corte	Ensayo de corte directo	ASTM D 3080–03
	Resistencia a la compresión no confinada	ASTM D 2166–00
	Prueba de paletas en miniatura	ASTM D 4648–00
	Triaxial no drenado no consolidado	ASTM D 2850–03
	Ensayo triaxial no drenado consolidado	ASTM D 4767–02

	Prueba de corte de anillo torsional	ASTM D 6467–99
Compactación	Prueba de Proctor estándar	ASTM D 698–00
	Prueba de Proctor modificado	ASTM D 1557–02
	Prueba de cono de arena	ASTM D 1556–00
	Prueba del cilindro de accionamiento	ASTM D 2937–00
Permeabilidad	Prueba de cabeza constante	ASTM D 2434–00
	Prueba de caída de cabeza	ASTM D 5084–00
Erosión	Arcilla dispersiva	ASTM D 4647-93
	Potencial de erosión	Day 1990b
Pavimentos y deterioraciones	Pavimentos: CBR	ASTM D 1883-94
	Pavimentos: valor R	ASTM D 2844-94
	Sulfatos, Cloruros y sales solubles	BS1377 – Parte 3

Fuente: (E.050, 2015, pág. 13), (Day, 2010)

3.3. Modelos para predecir las propiedades mecánicas de los suelos

Como se ha mencionado en el apartado de los antecedentes del capítulo 2, según la literatura revisada, existen autores que realizaron varios modelos para predecir las propiedades mecánicas de los suelos basados en modelos de inteligencia artificial y/o Machine Learning. En este apartado se mencionarán modelos donde se utilizan técnicas estadísticas prometedoras para la predicción de las propiedades mecánicas de los suelos aparte de la inteligencia artificial.

3.3.1. Modelos de correlación del CBR existentes

Estos autores en su investigación recopilan información de diferentes estudios pasados que han realizado varios intentos para desarrollar una relación entre el valor del CBR en función del

óptimo contenido de humedad (OMC) y las pruebas más simples como son: límite plástico, límite líquido, compactación, máxima densidad seca (MDD) (Cordova Valentin & Mori Montalvo, 2021). Así como también se presenta una tabla de principales autores y sus modelos propuestos con respeto a la predicción del CBR mediante las diferentes propiedades de los suelos:

Tabla 02 Modelos de Correlación del CBR Existentes

Autor	Consideraciones	Modelo	Parámetro Estadístico
AASHTO (2002)	Suelos granules con finos	$CBR = 75 / (1 + 0.728 * (P200 * IP)$	-
Patel & Desai (2010)	Suelos Finos	CBR (lab)=43.907-0.093(IP)-18.78(MDD)-0.3081(OMC)	-
Saklecha et al.(2011)	CBR=1-70%	CBR=0.62 OMC+58.9MDD+0.11LL+0.53LP-126.18	**R=0.80**
Yildirim & Gunaydin (2011)	Suelos granulares y finos	CBR = 4.739MDD + 0.122 OMC +0.22G + 0.045S	**R^2=0.88**
Ramasubbarao (2013)	Suelos FinosCL, CH CBR=0.8-6%	CBR lab = 0.064F - 1.810MDD - 0.061OMC + 0.082S + 0.033G - 0.069LL + 0.157LP	**R^2=0.92**
Kumar (2014)	Suelos finosML CBR=5-7%	CBR lab = 0.127LL-0.1598IP + 1.405MDD - 0.259OMC + 4.618	-
Araujo (2014)	**Todos los tipos de suelo**	**CBR=0.681G-2.917 OMC+0.032F-17.991MDD+0.510LL-0.82IP+0LP+0.205S+64.89**	**R=0.89**

Adaptado de (Cordova Valentin & Mori Montalvo, 2021)

3.4. Marco conceptual

- Python: Python es un lenguaje de programación de alto nivel, interpretado y de propósito general.
- Librerías Python: Python tiene una gran cantidad de librerías disponibles que permiten a los programadores ahorrar tiempo y esfuerzo al utilizar funciones predefinidas para realizar diferentes tareas.
- Pandas: Es una librería para la manipulación y análisis de datos. Permite trabajar con estructuras de datos como DataFrames y Series, y se utiliza en análisis de datos, ciencia de datos y finanzas.
- Numpy: Es una librería que permite trabajar con arreglos multidimensionales y matrices. Se utiliza comúnmente en ciencia de datos, aprendizaje automático y cálculo científico.
- Matplotlib: Es una librería de visualización de datos en 2D que permite crear gráficos, histogramas, diagramas de dispersión y más.
- Keras: Es una librería de aprendizaje profundo de alto nivel que permite crear y entrenar redes neuronales de manera sencilla.
- TensorFlow: Es una librería de aprendizaje automático de código abierto desarrollada por Google. Permite construir modelos de aprendizaje profundo y se utiliza en reconocimiento de voz, procesamiento de lenguaje natural, visión por computadora, entre otros.
- Scikit-learn: Es una librería para aprendizaje automático y minería de datos. Incluye una amplia variedad de algoritmos para la clasificación, regresión, agrupación, entre otros.
- API: Application Programming Interface es un conjunto de reglas, protocolos y herramientas que permiten a los programadores acceder y utilizar datos y servicios de una aplicación o plataforma sin necesidad de acceder directamente a su código fuente.
- Google Colaboratory: También conocido como Google Colab, es una plataforma en línea de Google que permite a los usuarios ejecutar y colaborar en código de Python. Es una herramienta útil para el desarrollo de proyectos de aprendizaje automático, ciencia de datos, y otros proyectos de programación.
- Propiedades físicas del suelo: Las propiedades físicas del suelo son las que se definen como las propiedades obtenidas con los ensayos físicos mas no ensayos de resistencia,

como contenido de humedad, límites de consistencia, granulometría, etc (Manrique Dávila, 2019).

- Propiedades mecánicas del suelo: Las propiedades mecánicas del suelo son las que se definen como las propiedades obtenidas con los ensayos que se relacionan con la resistencia, deformabilidad y estabilidad, como ensayos de compactación, ensayos de corte, etc (Manrique Dávila, 2019).
- Calicata: Excavaciones a cielo abierto, hasta la profundidad deseada, tomando las precauciones necesarias para evitar el desprendimiento de material de las paredes que pueda afectar la seguridad del trabajador o contaminar la muestra que se espera obtener (MTC, 2016).
- Compresión: Acción de comprimir un material aplicando una carga que puede ser axial, existiendo variantes en ensayos como: no confinada, triaxial y entre estos el ensayo consolidado no drenado; el ensayo drenado, el ensayo no consolidado no drenado y que sirven para medir el ángulo de fricción interna (ϕ) y la cohesión (C), cuyos valores se emplean en análisis de estabilidad en estructuras (fundaciones), cortes, taludes, muros de contención, etc (MTC, 2016).
- Capacidad de carga: La capacidad de carga por corte del suelo o presión última o de falla, se determinará utilizando las fórmulas aceptadas por la mecánica de suelos (E.050, 2015).
- Consistencia: Relativa facilidad con que el suelo puede fluir y deformarse (MTC, 2016).
- Módulo de finura: Número empírico que se obtiene sumando los porcentajes retenidos en cada una de las mallas que se indican más abajo y luego dividiendo el resultado entre 100 (MTC, 2016).
- Picnómetro: Recipiente de vidro, forma cilíndrica o cónica (MTC, 2016).
- Eficiencia: Sinónimo de productividad o la capacidad de disponer de algo para conseguir un efecto determinado utilizando el mínimo de recursos y esfuerzo.
- Eficacia: Capacidad de lograr el efecto que se espera o se desea.

3.5. Definición de variables

- Técnicas de Machine Learning: Son herramientas o técnicas de aprendizaje automático también se denominan específicamente algoritmos de aprendizaje automático. Las técnicas

de machine learning abarcan una variedad de métodos que permiten a los sistemas aprender y hacer predicciones a partir de datos (Konasani & Kadre, 2021).

- Propiedades mecánicas del suelo: Las propiedades mecánicas del suelo son las que se definen como las propiedades obtenidas con los ensayos que se relacionan con la resistencia, deformabilidad y estabilidad, como ensayos de compactación, ensayos de corte, etc (Manrique Dávila, 2019).

CAPITULO IV

METODOLOGIA DE INVESTIGACIÓN

4.1. Operacionalización de variables

Tabla 03 Operacionalización de Variables

VARIABLE	DIMENSION	INDICADORES	INDICE/ESCALA
Independiente: **Técnicas de Machine Learning**	Modelos predictivos	-Extra Trees Regressor (Regression Tree Ensembles)	$R^2 = \frac{SCR}{SCE + SCR}$
		- Random Forest Regressor (Regression Tree Ensembles)	$MSE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}$
		-Redes Neuronales Artificiales (BPNN)	$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(y_i - x_i)^2}{n}}$
			$MAE = \frac{\sum_{i=1}^{n}\lvert y_i - x_i\rvert}{n}$
Dependiente: Propiedades mecánicas del suelo	Capacidad de soporte (CBR)	-Granulometría -Límites de consistencia -Tipo de suelo SUCS -Departamento	Factor de correlación de Spearman para CBR $^{p} = 1 - \frac{6\sum_{i=1}^{n}(d_i)^2}{n(n^2-1)}$
	Máxima densidad seca (MDD)	-Granulometría -Límites de consistencia -Tipo de suelo SUCS -Departamento	Factor de correlación de Spearman para MDS $^{p} = 1 - \frac{6\sum_{i=1}^{n}(d_i)^2}{n(n^2-1)}$
	Optimo contenido de humedad (OMC)	-Granulometría -Límites de consistencia -Tipo de suelo SUCS -Departamento	Factor de correlación de Spearman para OCH $^{p} = 1 - \frac{6\sum_{i=1}^{n}(d_i)^2}{n(n^2-1)}$

4.2. Tipo y Nivel de investigación

4.2.1. Tipo de investigación

El tipo de investigación de la presente investigación es aplicada, ya que basándose en los resultados de la investigación básica, pura o fundamental está orientada a resolver los problemas sociales de una comunidad, región o país, como los problemas de la geotecnia y otros (Ñaupas Paitán et al., 2018, pág. 136). La investigación aplicada se enfoca en resolver problemas

específicos y aplicar técnicas científicas para desarrollar soluciones, en este caso se buscó aumentar la eficiencia del modelo predictivo de las propiedades mecánicas del suelo mediante la evaluación y comparación de diferentes técnicas de regresión de Machine Learning y seleccionando el más adecuado para cada ensayo.

4.2.2. Nivel de investigación

El nivel de investigación es descriptivo, ya que se enfoca en detallar y documentar las características y propiedades del fenómeno estudiado sin intentar modificarlo o manipularlo (Hernández Sampieri et al., 2014, pág. 92). La investigación descriptiva proporciona una visión detallada del contexto y de los factores involucrados, permitiendo una comprensión profunda del problema. En este caso, se describen las propiedades mecánicas de los suelos y su relación con las variables estudiadas utilizando técnicas de Machine Learning.

4.3. Diseño de Investigación

El diseño de investigación es no experimental, porque en esta investigación no se manipuló las variables estudiadas, sino que se observa y analiza su relación con las variables dependientes en su entorno natural. En estos diseños, se busca describir y analizar relaciones y correlaciones entre variables sin intervenir directamente. En este caso, se describen las propiedades mecánicas de los suelos y su relación con las variables estudiadas utilizando técnicas de Machine Learning, y es de tipo transeccional, porque en esta investigación se recogieron y obtuvieron los datos de manera documental en un momento determinado del tiempo para la predicción de las propiedades mecánicas de los suelos. (Hernández Sampieri et al., 2014, págs. 152,155).

4.4. Población y Muestra

4.4.1. Población

La población estuvo compuesta por las muestras existentes de suelo a nivel mundial. Las muestras de suelo son obtenidas de ensayos de laboratorio, resultados de investigaciones pasadas, expedientes técnicos de obras de construcción, bases de datos y otros tipos. Por ello, la población del presente estudio fue de tipo infinita.

4.4.2. Muestra

El tipo de muestra es no probabilística intencional o por conveniencia, ya que se recogieron una cantidad de muestras de suelos que fueron escogidos intencionalmente de diferentes departamentos del Perú para obtener una muestra balanceada de los diferentes tipos de suelos del Perú. Las muestras de suelo se obtuvieron de los expedientes técnicos de obras de pavimentación y otros tipos que fueron realizados en el Perú. Asimismo, también se descargó información del Sistema Electrónico de Contrataciones del Estado (SEACE 3.0) y expedientes técnicos aprobados por el Organismo Supervisor de las Contrataciones del Estado (OSCE), donde, a pesar de que se calculó según el muestreo para poblaciones infinitas de 384 muestras de suelo, se logró recolectar 3294 muestras de suelo de todo el Perú.

Muestreo para poblaciones infinitas, donde:

$$n = \frac{Z^2 * P * Q}{E^2}\ ;\ \frac{1.96^2 * 0.5 * 0.5}{0.05^2} = 384 \text{ muestras de suelo } (\text{Se logró } 3294)$$

n **=** *Tamaño de Muestra.*

Z **=** *Parámetro estadístico que depende del Nivel de Confianza, en este caso será 1.96.*

E **=** *Error de estimación máximo aceptado 5%.*

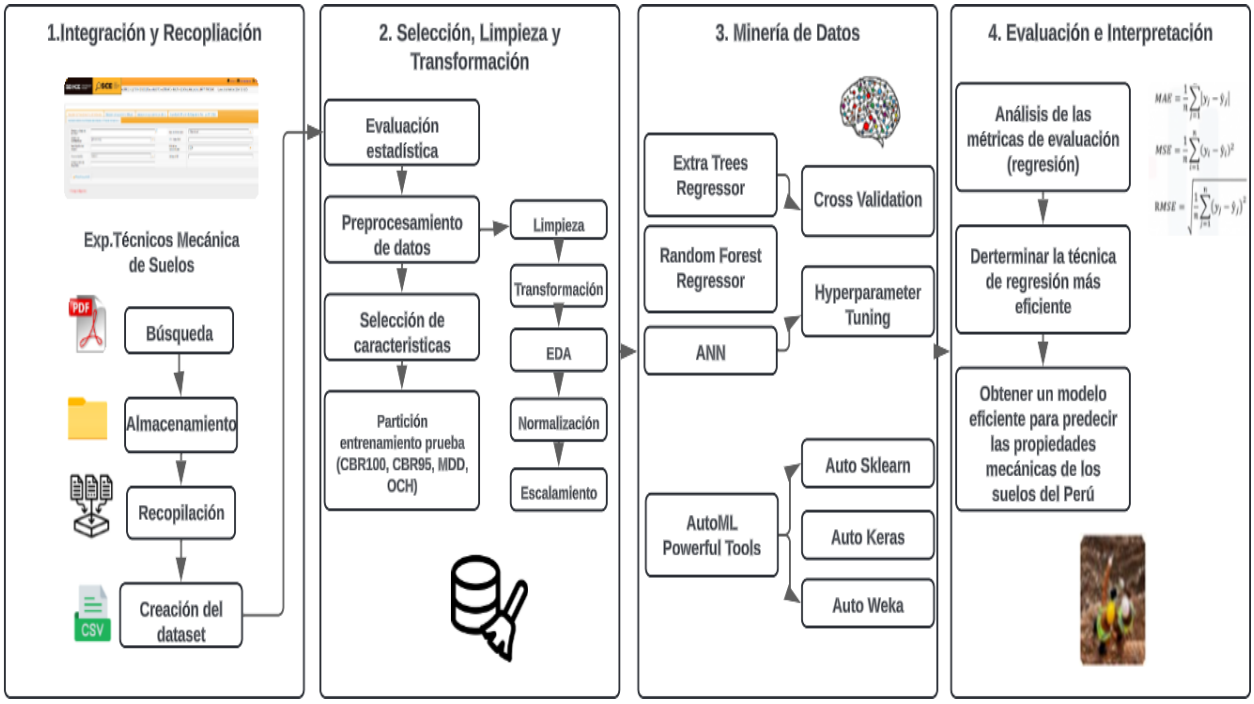
P **=** *Probabilidad de que ocurra el evento estudiado (éxito) es 0.5.*

Q *= (1-P) = Probabilidad de que no ocurra el evento estudiado es 0.5.*

4.5. Procedimiento de la Investigación

En esta investigación, se implementó la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD, por sus siglas en inglés: Knowledge Discovery in Databases). Así como también, se propuso un pipeline que sigue los principios, etapas y procesos de la metodología KDD aplicados y adaptados al presente trabajo de investigación. Este enfoque metodológico comprendió varios procesos y cuatro etapas, las cuales se detallan a continuación en la Figura 29.

Figura 29 Proceso de la solución desarrollada con la metodología KDD



4.5.1. Etapa I: Integración y Recopilación

En esta primera etapa, primeramente, se inició con la búsqueda de expedientes técnicos de mecánica de suelos y estudios geotécnicos de proyectos de construcción principalmente de carreteras, pavimentación, puentes, murros de contención y taludes en el Sistema Electrónico de Contrataciones del Estado (SEACE 3.0) y expedientes técnicos aprobados por el Organismo Supervisor de las Contrataciones del Estado (OSCE). Así como también se realizó de la búsqueda de informes de tesis de investigaciones pasadas donde incluyeran datos de expedientes técnicos de estudios definitivos de mecánica de suelos realizados en diferentes departamentos del Perú.

Posteriormente, se procedió a filtrar y almacenar todos los expedientes técnicos recopilados guardándolos en carpetas con nombres de departamentos del Perú y cada expediente técnico con un meta código como el Código de Único de Inversión (CUI) o el código del Sistema Nacional de Inversión Pública SNIP, cualquiera de los 2 que estuviera disponible, para su fácil ubicación y corrobación de los datos recolectados. Dicha forma de trabajo permitió una organizada búsqueda y filtración de expedientes técnicos que contengan los datos buscados, así como también, su fácil ubicación y descarga de archivos en el sistema después de dejar descargar en un determinado

departamento, para luego, ubicarse donde se dejó y volver a buscar y descargar los archivos (ver en el Anexo 1 la carpeta con los expedientes técnicos).

Después de descargar todos los expedientes técnicos filtrados de todos los departamentos del Perú se procedió a realizar la recopilación manual de los datos de los expedientes técnicos e informes utilizando el instrumento de recolección de datos en formato Excel (ver Anexo 2). Finalmente, se creó el set de datos con una cantidad de 3294 muestras de suelo o calicatas de estudios definitivos que fueron recolectados y codificados con el Código de Único de Inversión (CUI) o el código del Sistema Nacional de Inversión Pública (SNIP) para su verificación en dicho estudio (ver Anexo 1). A continuación, se presenta el diccionario de datos utilizados para el presente trabajo:

Tabla 04 Diccionario de datos de las variables de entrada

VARIABLES DE ENTRADA	DESCRIPCIÓN	TIPO DE DATO	DOMINIO
DEPARTAMENTO	Departamento del cual se obtuvo la muestra de suelo	Texto	Valores: Los 25 departamentos del Perú
SUCS	Clasificación de los suelos según el sistema SUCS	Texto	Valores: Los valores según la clasificación SUCS
GRAVA	Cantidad de grava de la muestra de suelo en porcentaje	Decimal	Valores: 0-100%
ARENA	Cantidad de arena de la muestra de suelo en porcentaje	Decimal	Valores: 0-100%
FINOS	Cantidad de finos de la muestra de suelo en porcentaje	Decimal	Valores: 0-100%
Límite Líquido (LL)	Contenido de agua a partir del cual el suelo pasa a estado líquido	Decimal	Valores: 0-100%
Límite Plástico (LP)	Contenido de agua a partir del cual el suelo pasa a estado plástico	Decimal	Valores: 0-100%
Índice de Plasticidad (IP)	Diferencia entre el límite líquido y el límite plástico	Decimal	Valores: 0-100%

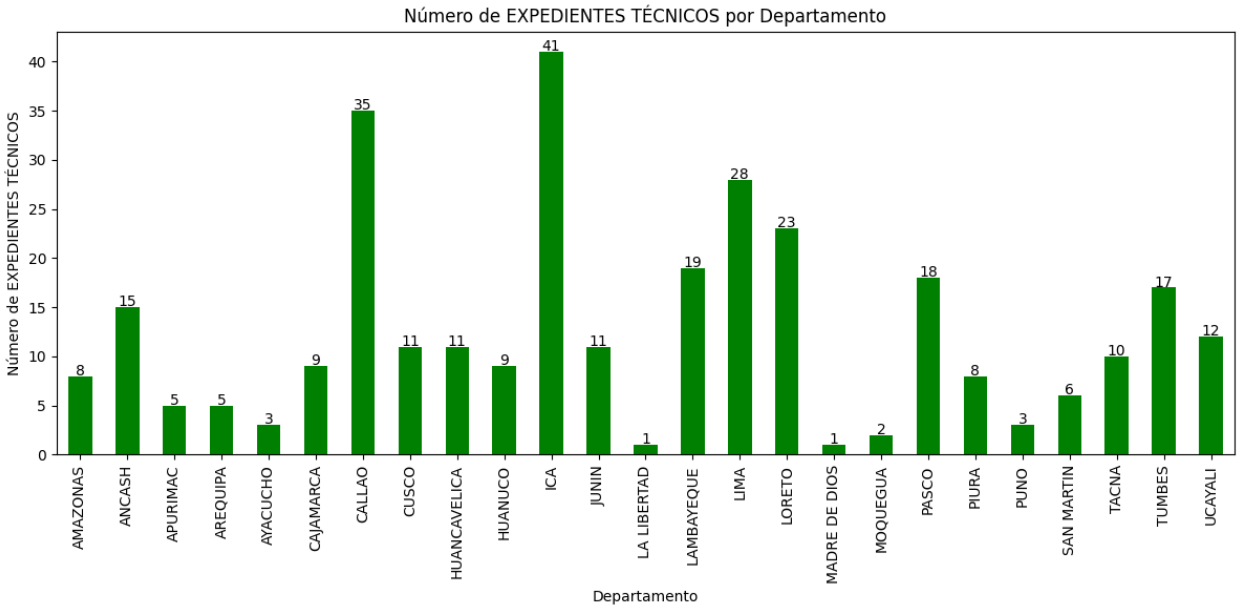
Tabla 05 Diccionario de datos de las variables de salida

VARIABLES DE ENTRADA	DESCRIPCIÓN	TIPO DE DATO	DOMINIO
Máxima Densidad Seca (MDD)	Máxima densidad seca del suelo en g/cm³	Decimal	Valores: Dependiendo del tipo de suelo (usualmente 0-3

			g/cm³)
Óptimo Contenido de Humedad (OMC)	Contenido de humedad óptimo del suelo en porcentaje	Decimal	Valores: 0-100%
CBR95	California Bearing Ratio al 95% de compactación	Decimal	Valores: 0-100%
CBR100	California Bearing Ratio al 100% de compactación	Decimal	Valores: 0-100%

A continuación, en la Figura 30 se muestra en un gráfico de barras el resumen de la cantidad de expedientes técnicos de estudios definitivos de mecánica de suelos recopilados por departamento en todo el Perú.

Figura 30 Cantidad de expedientes técnicos recopilados por departamentos del Perú

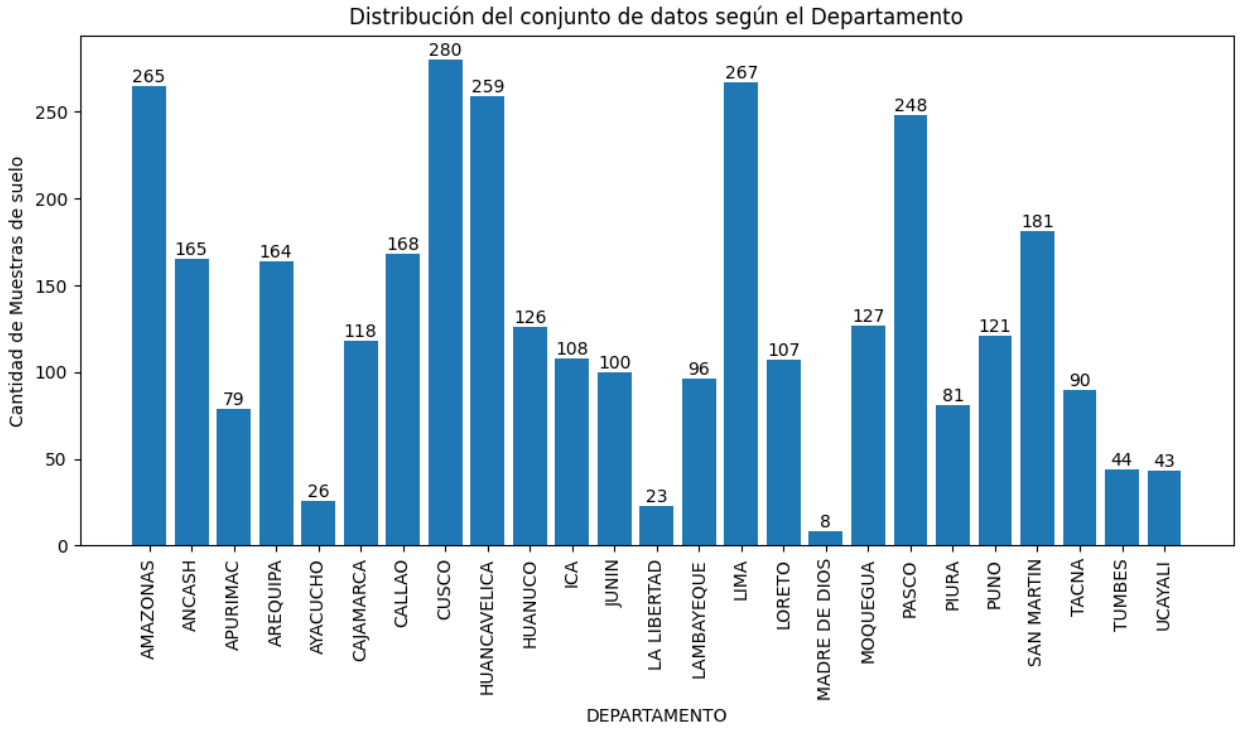


Se puede observar que se descargaron en total 311 expedientes técnicos o estudios definitivos de mecánica de suelos de proyectos de construcción del Sistema Electrónico de Contrataciones del Estado (SEACE 3.0) y expedientes técnicos aprobados por el Organismo Supervisor de las Contrataciones del Estado (OSCE) en el periodo de los años 2019, 2020, 2021, 2022 y 2023. De los gobiernos regionales, capitales y principales ciudades de cada departamento del Perú. Se pudo comprobar que muchos departamentos o gobiernos regionales, provincias y distritos y en ese orden (como La Libertad y Madre de Dios) no suelen subir sus estudios

definitivos a dicho sistema en comparación con otras ciudades probablemente más desarrolladas con políticas digitales (como el Callao e Ica).

En el siguiente gráfico de barras de la Figura 31 se muestra un resumen de la cantidad de muestras de suelos o calicatas recolectadas de los diferentes expedientes técnicos o estudios definitivos de mecánica de suelos de proyectos de construcción por departamentos del Perú. Se puede observar que se recolectaron 3294 muestras de suelo o calicatas los cuales se distribuyen en pequeñas muestras de rango de 8 a 280 muestras de suelos para cada uno de los 25 departamentos del Perú. Cabe mencionar que dicha distribución de recolección de muestras de los suelos permitió cubrir los 9 tipos de suelos y sus 15 grupos según el sistema de clasificación de suelos SUCS.

Figura 31 Cantidad de muestras de suelo recopilados por departamentos del Perú



El cuadro también muestra la distribución de la cantidad de muestras de suelo por departamento en el Perú. La mayoría de los departamentos tienen entre 100 y 200 muestras, con Amazonas, Cusco, Huancavelica, Lima y Pasco como los que más poseen. Por otro lado, Ayacucho, La Libertad, Madre de Dios y Ucayali son los que menos tienen. La información del cuadro puede ser útil para identificar los departamentos que necesitan más estudios, investigación,

políticas y estándares de almacenamiento de datos de estudios de mecánicas de suelos del Perú. Esta información también puede ser utilizada para planificar futuros estudios, investigaciones y políticas públicas relacionadas con la mecánica de los suelos en el Perú y el almacenamiento correcto de los datos de estos estudios.

4.5.2. Etapa II: Selección, Limpieza y Transformación

4.5.2.1. Evaluación estadística

A. Resultados estadísticos descriptivos

A partir de la etapa anterior se consiguió el siguiente set de datos donde se muestran 15 atributos: SUCS, GRAVA, ARENA, FINOS, LL, LP, IP, OMC, MDD, CBR100, CBR95, CALICATA, DEPARTAMENTO, METACODIGO, CUI/SNIP. Donde los atributos de entrada para el modelo de Machine Learning fueron GRAVA, ARENA, FINOS, LL, LP, IP y 2 atributos adicionales propuestos DEPARTAMENTO y SUCS. Los atributos de salida fueron OMC, MDD, CBR100, CBR95. Los atributos CALICATA, METACODIGO, CUI/SNIP fueron atributos adicionales irrelevantes para el modelo, ya que solo sirvieron para organizar y facilitar la recolección de datos e identificar la procedencia de estos. A continuación, en la siguiente Figura 32 se muestra un resumen del set de datos con los atributos mencionados y sus valores respectivos en la herramienta de desarrollo Google Colaboratory.

Figura 32 Resumen del set de datos inicial

	SUCS	GRAVA	ARENA	FINOS	LL	LP	IP	OMC	MDD	CBR100	CBR95	CALICATA	DEPARTAMENTO	METACODIGO	CUI/SNIP
0	SC	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	59.3	48.20	C-01	AMAZONAS	5.2019.13.	294299.0
1	SC-SM	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	55.4	43.50	C-02	AMAZONAS	5.2019.13.	294299.0
2	GP	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	55.0	44.50	C-03	AMAZONAS	5.2019.13.	294299.0
3	GP	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	82.0	58.00	C-04	AMAZONAS	5.2019.13.	294299.0
4	ML	NaN	NaN	NaN	28.4	25.6	2.8	14.44	1.720	10.2	9.69	C-01	AMAZONAS	5.2020.5.	2449158.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3289	GM	51.4	27.1	21.5	42.0	32.0	10.0	11.10	1.915	48.7	31.30	C-177-M2	PASCO	NaN	NaN
3290	SM	22.8	51.0	26.2	36.0	29.0	7.0	14.20	1.781	38.5	25.70	C-185-M1	PASCO	NaN	NaN
3291	SM	28.1	46.6	25.3	38.0	31.0	7.0	18.80	1.616	23.3	15.50	C-185-M2	PASCO	NaN	NaN
3292	GP-GM	48.7	42.4	8.9	33.0	25.0	8.0	10.70	1.866	64.8	43.60	C-193-M1	PASCO	NaN	NaN
3293	GM	51.8	34.2	14.0	35.0	30.0	5.0	9.40	2.185	48.2	38.30	C-201-M1	PASCO	NaN	NaN

3294 rows × 15 columns

De las 3294 muestras de suelo o calicatas que se utilizaron para desarrollar modelos predictivos de las propiedades mecánicas de los suelos en la Figura 33 se muestra la distribución de los tipos de suelos en el conjunto de datos. Los tres tipos de suelo que conforman el set de datos son grava, arena y finos (materiales limo arcillosos). La grava es el tipo de suelo más abundante, con un 58,38%. La arena son el segundo tipo de suelo más abundante, con un 30,37%. Los finos es el tipo de suelo menos abundante, con un 11,25%.

Figura 33 Distribución de los tipos de suelos del conjunto de datos

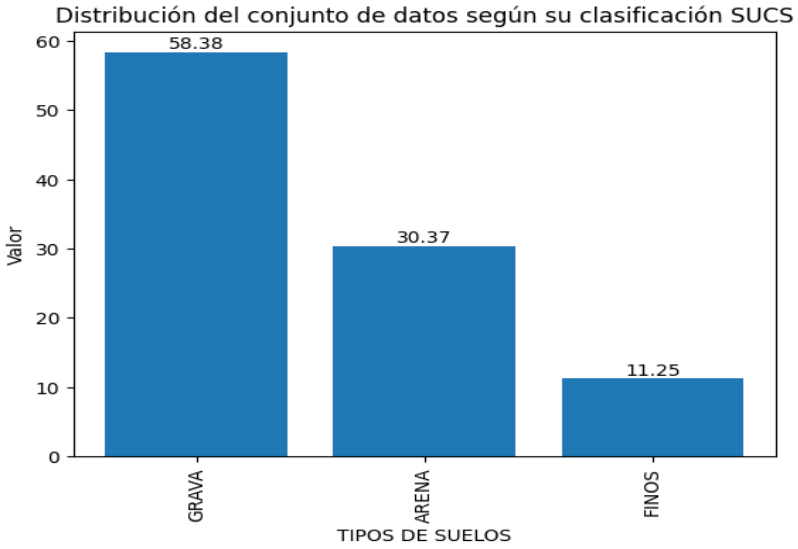
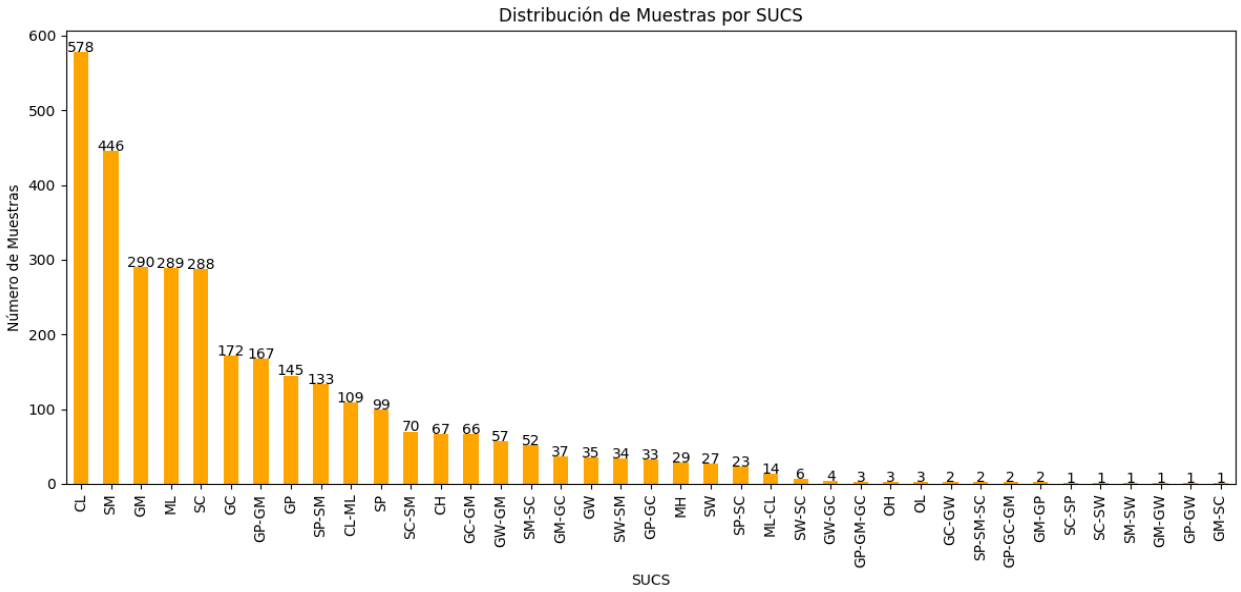


Figura 34 Distribución de los tipos de suelos según la clasificación SUCS del conjunto de datos



Para una exploración más exhaustiva del conjunto de datos de los tipos de suelos del Perú según el sistema de clasificación SUCS se muestra en la figura 34 un resumen detallado de todos los tipos de suelos que existen en el set de datos. Cabe recalcar que según el sistema de clasificación SUCS se propone 9 tipos de suelos y 15 grupos de suelos. En el set de datos de suelos del Perú recolectado se supera esta clasificación incluso formándose otros grupos con algunos de los 15 grupos de clasificación, incluso grupos de 3 grupos, dejando en evidencia que los suelos del Perú son muy ricos y diversos.

Figura 35 Resumen estadístico descriptivo del conjunto de datos

	GRAVA	ARENA	FINOS	LL	LP	IP	OMC	MDD	CBR100	CBR95	CUI/SNIP
count	3034.000000	3034.000000	3040.000000	3280.000000	3280.000000	3280.000000	2885.000000	2884.000000	2965.000000	2994.000000	2.733000e+03
mean	25.804243	39.526207	34.657604	23.516692	14.973131	8.544486	10.934324	1.931792	30.122226	20.213312	2.266310e+06
std	24.797726	21.754349	27.618116	14.864244	10.688493	7.798035	4.451546	0.200966	25.762607	17.270853	4.726292e+05
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.610000	1.125000	0.000000	0.080000	1.743670e+05
25%	0.862500	23.832500	11.800000	16.970000	0.000000	1.630000	7.800000	1.798000	10.800000	7.100000	2.196451e+06
50%	19.425000	37.200000	26.590000	25.200000	18.380000	7.000000	10.180000	1.922000	21.700000	14.910000	2.386552e+06
75%	46.337500	50.890000	55.442500	33.000000	22.500000	13.800000	13.200000	2.098250	41.800000	27.575000	2.478824e+06
max	100.000000	99.660000	100.000000	80.050000	79.120000	45.700000	106.000000	2.900000	131.300000	90.400000	2.603055e+06

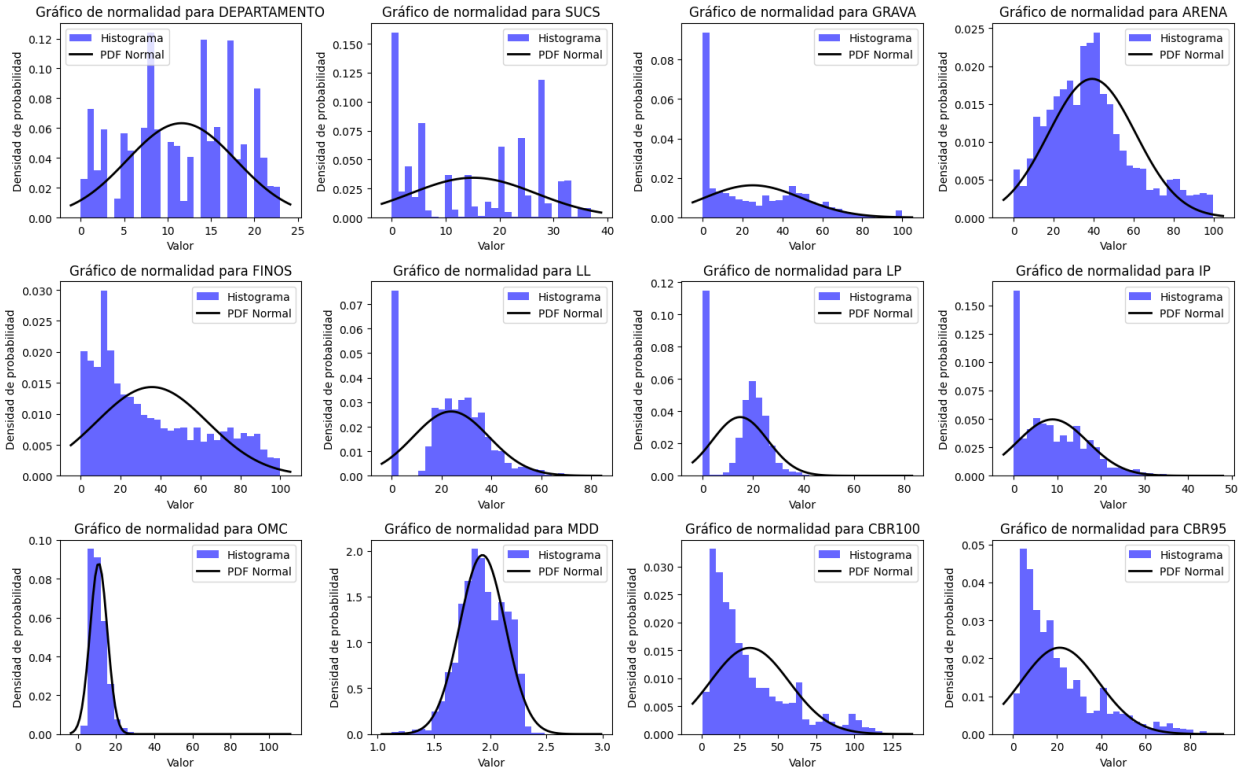
La figura anterior (Figura 35) titulada “Resumen estadístico descriptivo del conjunto de datos” muestra las estadísticas descriptivas para los atributos numéricos del set de datos. La tabla muestra las propiedades físicas y mecánicas de los diferentes tipos de suelo. Para cada atributo numérico del set de datos, se muestra la cantidad de datos (count), la media, la desviación estándar, el dato mínimo, el percentil 25, la mediana (percentil 50), el percentil 75 y el dato máximo. Dichas estadísticas descriptivas ayudaron a realizar el análisis exploratorio y comprender mejor el set de datos.

B. Prueba gráfica de normalidad

En la Figura 36 se muestra el gráfico de normalidad para cada variable del conjunto de datos. Donde para cada diagrama el eje horizontal muestra los valores de la variable, mientras que el eje vertical representa la frecuencia o probabilidad de que cada valor se presente en el conjunto de datos. La línea normal es la curva simétrica que indica la distribución normal teórica para la variable, mientras que las barras representan la frecuencia o probabilidad real de cada valor

observado. La comparación entre la distribución de las barras y la línea normal permite determinar si la variable se ajusta a una distribución normal. Se puede observar un gráfico de normalidad para la variable "MDD". Las barras se ajustan razonablemente a la línea normal, lo que indica que la variable "MDD" se distribuye aproximadamente de forma normal. Sin embargo, se observa una ligera curtosis positiva y una asimetría, lo que indica que la distribución es un poco más levantada que la normal teórica lo que podría sugerir que no se distribuye de forma normal. Y así se puede realizar un análisis de normalidad gráfica para cada variable y sugerir que ninguna variable se distribuye de forma normal. Sin embargo, es importante tener en cuenta que el gráfico de normalidad no es una prueba definitiva de normalidad. Se recomienda realizar otras pruebas estadísticas, como la prueba de Shapiro-Wilk, para confirmar la normalidad de cada variable.

Figura 36 Diagrama de normalidad para el análisis de normalidad de cada variable



C. Resultado de la prueba de Kolmogorov – Smirnov

En la siguiente tabla (Tabla 06) se analizó la distribución normal del conjunto de datos obtenidos con la técnica estadística de Kolmogorov – Smirnov. Luego, para determinar la normalidad se analizó según la significancia obtenida por el software SPSS para cada variable del

conjunto de datos, lo cual también se comprobó usando la librería “stats” de Python, mostrando los mismos resultados.

En la siguiente tabla obtenida por el software SPSS se pudo observar las pruebas de normalidad obtenidas como Kolmogorov – Smirnov y Shapiro-Wilk. Para los datos obtenidos se realizó la prueba de Kolmogorov – Smirnov ya que los datos superan el límite de la cantidad de datos de 50 que es el apropiado para dicha prueba. En la siguiente tabla se observó la prueba realizada de forma conjunta para todas las variables de entrada y salida mostrando resultados generales de la aplicación de dicha prueba.

Tabla 06 Prueba de normalidad conjunta del conjunto de datos con el software SPSS

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
DPTMNTO	,846	3293	,000	,945	3293	3,747
SUCS	,820	3293	,000	,885	3293	4,203
Grava	,732	3293	,000	,903	3293	1,568
Arena	,981	3293	,000	,960	3293	3,702
Finos	,942	3293	,000	,920	3293	1,111
LL	,797	3293	,000	,936	3293	1,848
LP	,709	3293	,000	,862	3293	,000
IP	,721	3293	,000	,907	3293	7,992
OMC	,997	3293	,000	,937	3293	3,907
MDD	,920	3293	,000	,988	3293	5,102
CBR100	,960	3293	,000	,877	3293	4,203
CBR95	,974	3293	,000	,864	3293	,000

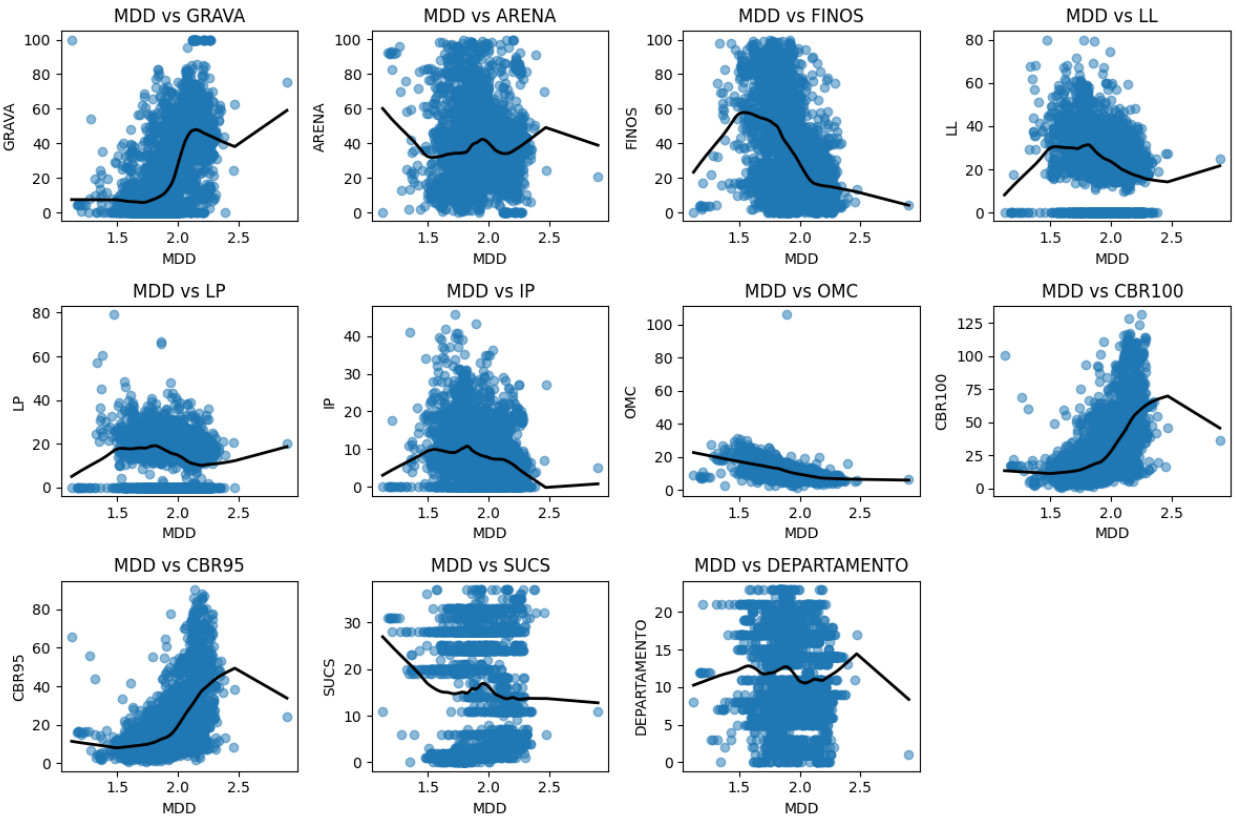
a. Corrección de significación de Lilliefors

Se puede concluir, del cuadro de prueba de normalidad anterior, que mediante la prueba de Kolmogorov – Smirnov los datos de las variables de entrada y salida no tienen una distribución normal ya que la significancia obtenida de 0,000 no supera la significancia del 0.050 que es el adecuado para que los datos sigan una distribución normal. En consecuencia, para correlacionar dichas variables no se puede utilizar la correlación de Pearson ya que no siguen una distribución normal y no tienen una relación lineal, por ello para correlacionar dichas variables y para la selección de características se utilizará la correlación de Spearman.

D. Resultado de los diagramas de dispersión

De la Figura 37 el diagrama de dispersión permite visualizar la relación entre la variable de salida MDD y las variables de entrada. En este gráfico, cada punto representa un dato, donde su posición en el eje vertical indica el valor de la variable de entrada y su posición en el eje horizontal indica el valor de MDD. La tendencia de los puntos puede revelar patrones o correlaciones entre MDD y las variables de entrada. Del análisis de correlación, se obtuvo que la variable MDD tuvo una buena correlación con las variables CBR100, CBR95 y Grava. Cabe mencionar que para modelos de Machine Learning no lineales las correlaciones de las variables no necesariamente tienen que ser lineales.

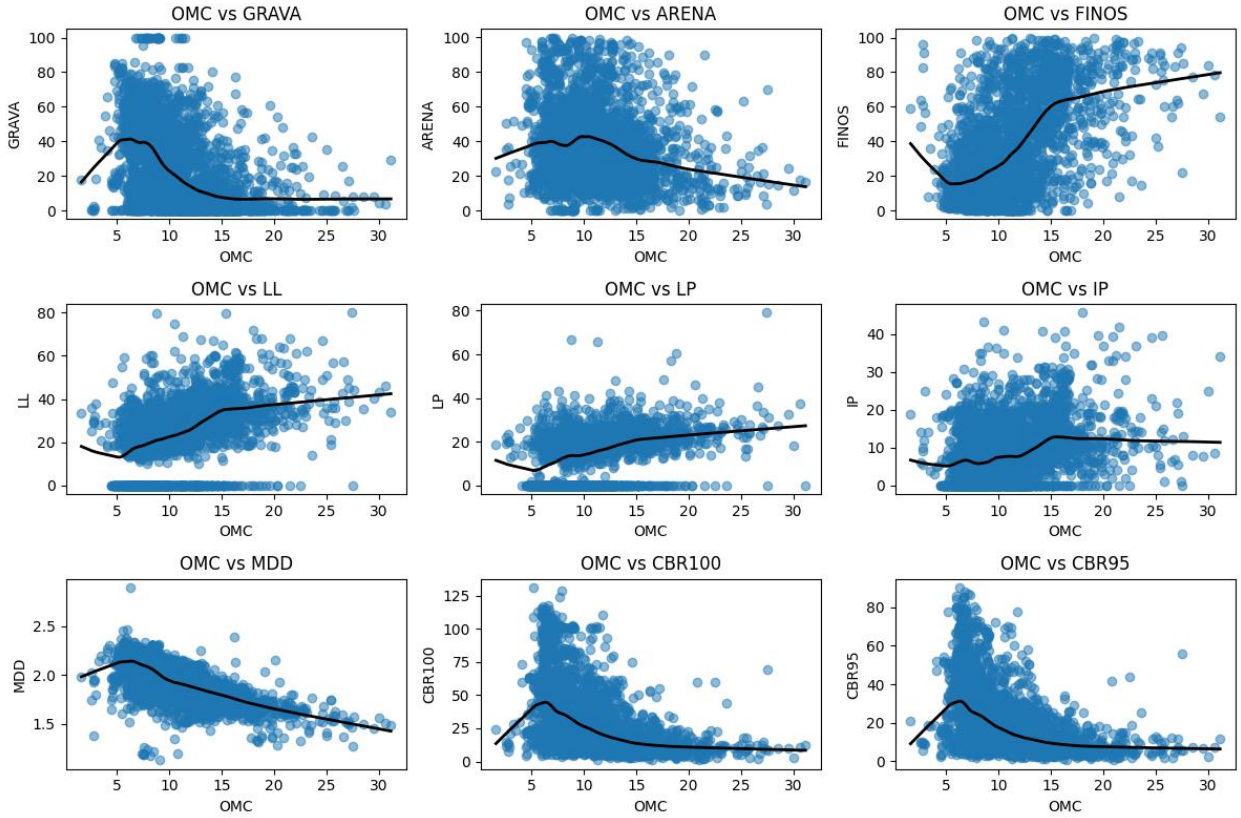
Figura 37 Diagrama de dispersión para la variable MDD con las variables de entrada



De la Figura 38 el diagrama de dispersión permite visualizar la relación entre la variable de salida OMC y las variables de entrada. En este gráfico, cada punto representa un dato, donde

su posición en el eje vertical indica el valor de la variable de entrada y su posición en el eje horizontal indica el valor de OMC. La tendencia de los puntos puede revelar patrones o correlaciones entre OMC y las variables de entrada. Del análisis de correlación, se obtuvo que la variable OMC tuvo una buena correlación con las variables FINOS, LL, LP y IP. Cabe mencionar que para modelos de Machine Learning no lineales las correlaciones de las variables no necesariamente tienen que ser lineales.

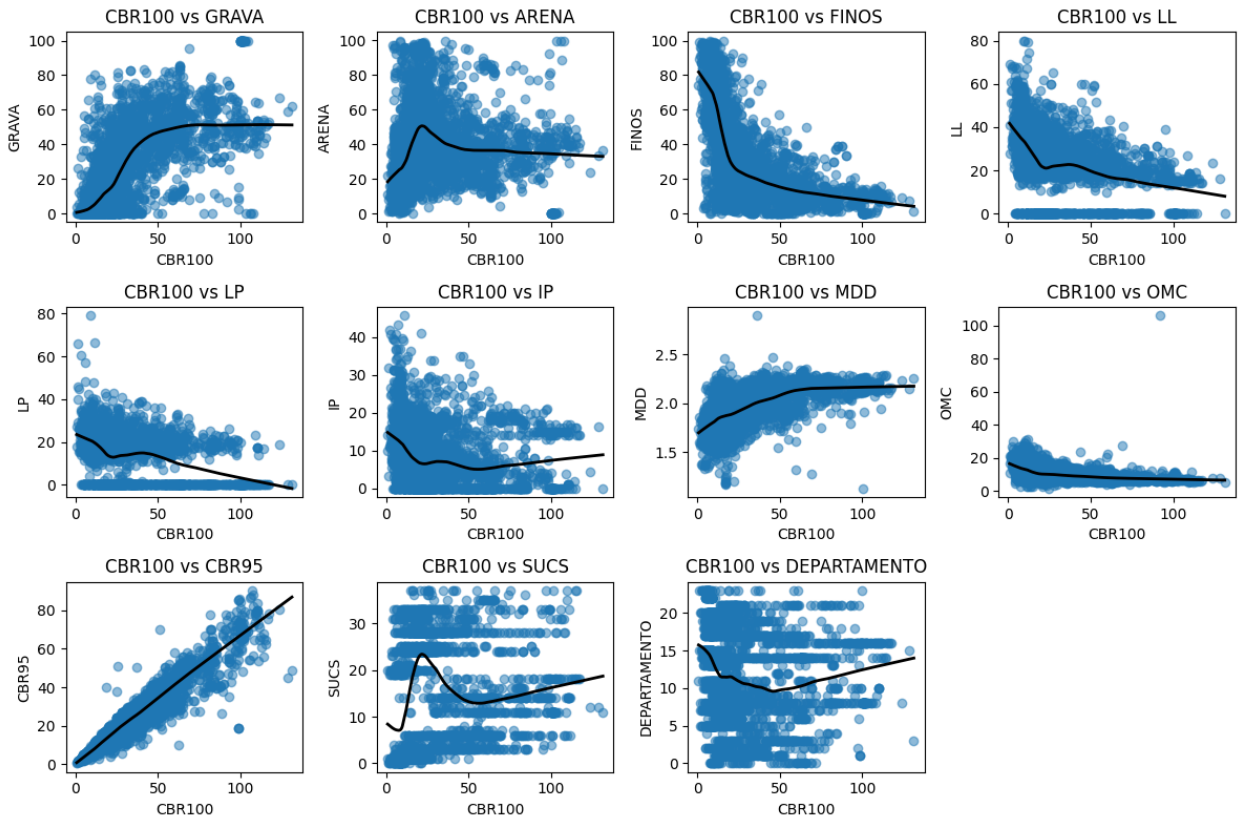
Figura 38 Diagrama de dispersión para la variable OMC con las variables de entrada



De la Figura 39 el diagrama de dispersión permite visualizar la relación entre la variable de salida CBR100 y las variables de entrada. En este gráfico, cada punto representa un dato, donde su posición en el eje vertical indica el valor de la variable de entrada y su posición en el eje horizontal indica el valor de CBR100. La tendencia de los puntos puede revelar patrones o correlaciones entre CBR100 y las variables de entrada. Del análisis de correlación, se obtuvo que la variable CBR100 tuvo una buena correlación con las variables MDD, CBR95 y Grava. Cabe

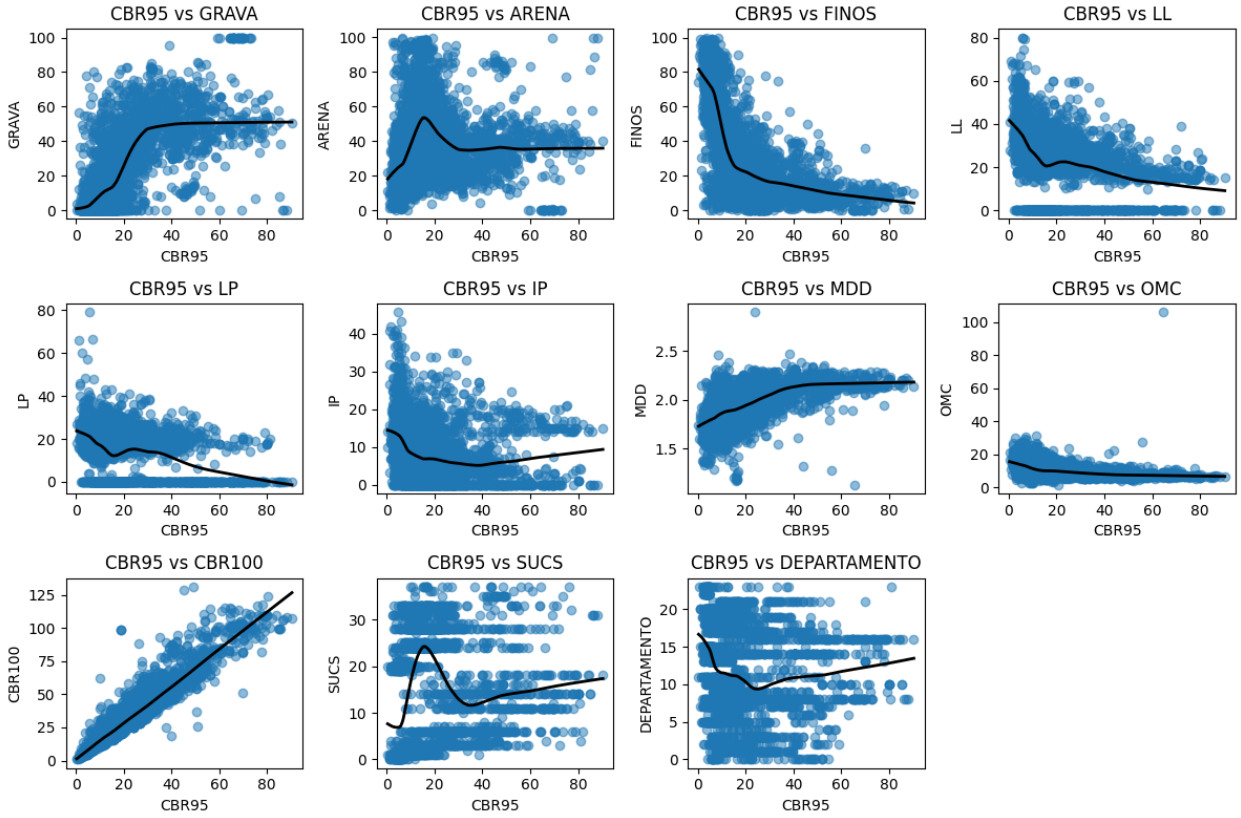
mencionar que para modelos de Machine Learning no lineales las correlaciones de las variables no necesariamente tienen que ser lineales.

Figura 39 Diagrama de dispersión para la variable CBR100 con las variables de entrada



De la Figura 40 el diagrama de dispersión permite visualizar la relación entre la variable de salida CBR95 y las variables de entrada. En este gráfico, cada punto representa un dato, donde su posición en el eje vertical indica el valor de la variable de entrada y su posición en el eje horizontal indica el valor de CBR95. La tendencia de los puntos puede revelar patrones o correlaciones entre CBR95 y las variables de entrada. Del análisis de correlación, se obtuvo que la variable CBR95 tuvo una buena correlación con las variables MDD, CBR100 y Grava. Cabe mencionar que para modelos de Machine Learning no lineales las correlaciones de las variables no necesariamente tienen que ser lineales.

Figura 40 Diagrama de dispersión para la variable CBR95 con las variables de entrada

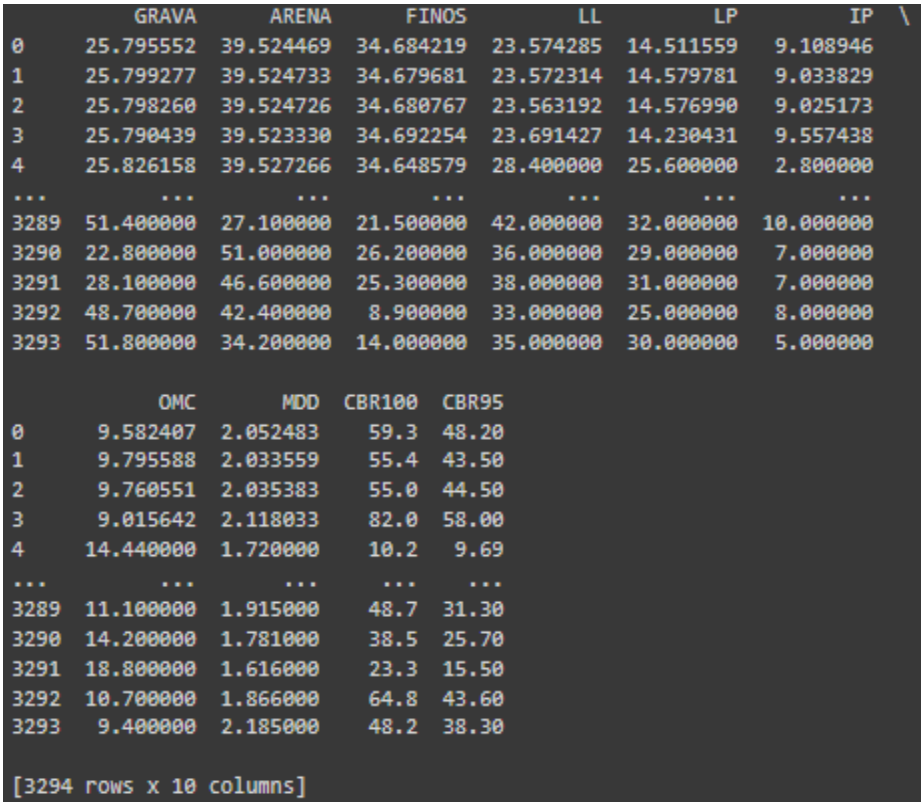


4.5.2.2. Preprocesamiento de Datos

A. Resultados de la imputación de datos faltantes

Se utilizó la librería de Python “fancyimpute” para realizar la imputación de los datos faltantes, dicha librería permite completar los datos faltantes del set de datos mediante una regresión en base a los datos no faltantes del set de datos utilizando técnicas avanzadas de aprendizaje automático (Ver código fuente en el Anexo 01.). En la Figura 41 se muestra los resultados después de aplicar la imputación de datos con el método de regresión de la librería “fancyimpute” de Python. Se puede observar que los valores faltantes fueron completados con dicha técnica.

Figura 41 Resultado del set de datos después de aplicar la librería de Python fancyimpute



	GRAVA	ARENA	FINOS	LL	LP	IP	\
0	25.795552	39.524469	34.684219	23.574285	14.511559	9.108946	
1	25.799277	39.524733	34.679681	23.572314	14.579781	9.033829	
2	25.798260	39.524726	34.680767	23.563192	14.576990	9.025173	
3	25.790439	39.523330	34.692254	23.691427	14.230431	9.557438	
4	25.826158	39.527266	34.648579	28.400000	25.600000	2.800000	
...	...	...	...	...	...	...	
3289	51.400000	27.100000	21.500000	42.000000	32.000000	10.000000	
3290	22.800000	51.000000	26.200000	36.000000	29.000000	7.000000	
3291	28.100000	46.600000	25.300000	38.000000	31.000000	7.000000	
3292	48.700000	42.400000	8.900000	33.000000	25.000000	8.000000	
3293	51.800000	34.200000	14.000000	35.000000	30.000000	5.000000	

	OMC	MDD	CBR100	CBR95
0	9.582407	2.052483	59.3	48.20
1	9.795588	2.033559	55.4	43.50
2	9.760551	2.035383	55.0	44.50
3	9.015642	2.118033	82.0	58.00
4	14.440000	1.720000	10.2	9.69
...	...	...	...	...
3289	11.100000	1.915000	48.7	31.30
3290	14.200000	1.781000	38.5	25.70
3291	18.800000	1.616000	23.3	15.50
3292	10.700000	1.866000	64.8	43.60
3293	9.400000	2.185000	48.2	38.30

[3294 rows x 10 columns]

B. Resultados de la codificación de las variables nominales

Se codificaron las variables nominales DEPARTAMENTO y SUCS asignando un código numérico único a cada valor. Esto facilita el análisis estadístico, permite comparaciones entre categorías y mejora la precisión de los modelos de aprendizaje automático. La codificación permite visualizar la relación entre las variables mediante diagramas de dispersión, histogramas y mapas de calor. Se pueden realizar pruebas estadísticas y utilizar modelos de aprendizaje automático para analizar los datos. Es importante elegir la codificación adecuada y documentarla cuidadosamente. Existen diferentes técnicas de codificación, como la binaria, ordinal y one-hot, la elección dependerá del tipo de variable y del análisis que se va a realizar. En este caso, se utilizó la codificación ordinal aplicando la función LabelEnconder de la librería preprocessing de Python, ya que son datos que se utilizaron para entrenar un modelo de regresión y permitió relacionar con el formato de las otras variables. En la Figura 42 y Figura 43 se pueden observar los valores codificados y valores reales para cada valor de las variables DEPRATEMNTO y SUCS en formato nominal y numérico respectivamente.

Figura 42 Resultado de la codificación de la variable DEPARTAMENTO

	DEPARTAMENTO_real	DEPARTAMENTO_codificado
1	AMAZONAS	0
2	ANCASH	1
3	APURIMAC	2
4	AREQUIPA	3
5	AYACUCHO	4
6	CAJAMARCA	5
7	CALLAO	6
8	CUSCO	7
9	HUANCAVELICA	8
10	HUANUCO	9
11	ICA	10
12	JUNIN	11
13	LA LIBERTAD	12
14	LAMBAYEQUE	13
15	LIMA	14
16	LORETO	15
17	MADRE DE DIOS	16
18	MOQUEGUA	17
19	PASCO	18
20	PIURA	19
21	PUNO	20
22	SAN MARTIN	21
23	TACNA	22
24	TUMBES	23
25	UCAYALI	24

Figura 43 Resultado de la codificación de la variable SUCS

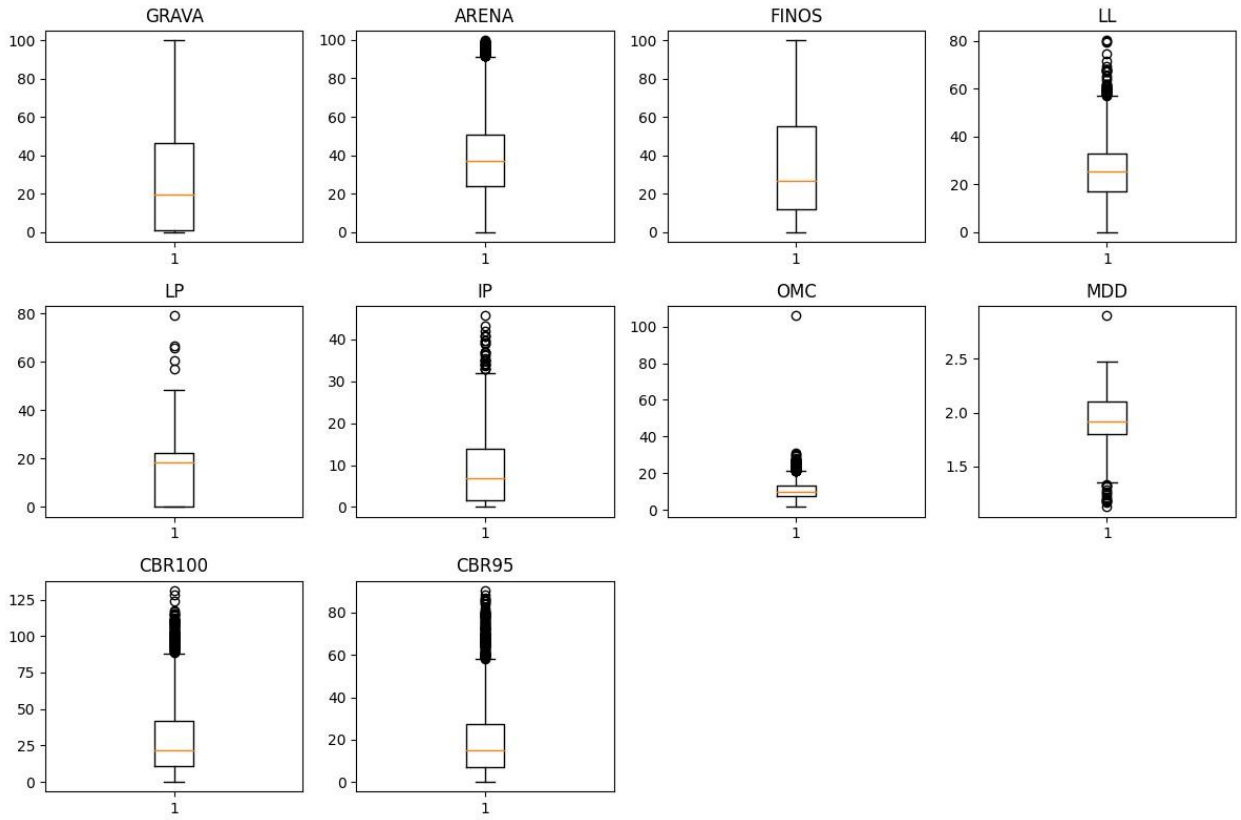
	SUCS_real	SUCS_codificado
1	SC	25
2	SC-SM	26
3	GP	11
4	ML	21
5	SM	29
6	GM	6
7	GC	3
8	GP-GC	12
9	GP-GM	14
10	SP	32
11	GM-GC	7
12	CL	1
13	CL-ML	2
14	MH	20
15	SP-SM	34
16	SP-SC	33
17	CH	0
18	ML-CL	22
19	GC-GM	4
20	GW-GM	19
21	SW-SM	38

22	SM-SC	30
23	GW-GC	18
24	SC-SP	27
25	GC-GW	5
26	SC-SW	28
27	SW	36
28	GW	17
29	GP-GM-GC	15
30	SP-SM-SC	35
31	OH	23
32	GP-GC-GM	13
33	GM-GP	8
34	SW-SC	37
35	OL	24
36	SM-SW	31
37	GM-GW	9
38	GP-GW	16
39	GM-SC	10

C. Resultados de la limpieza de datos

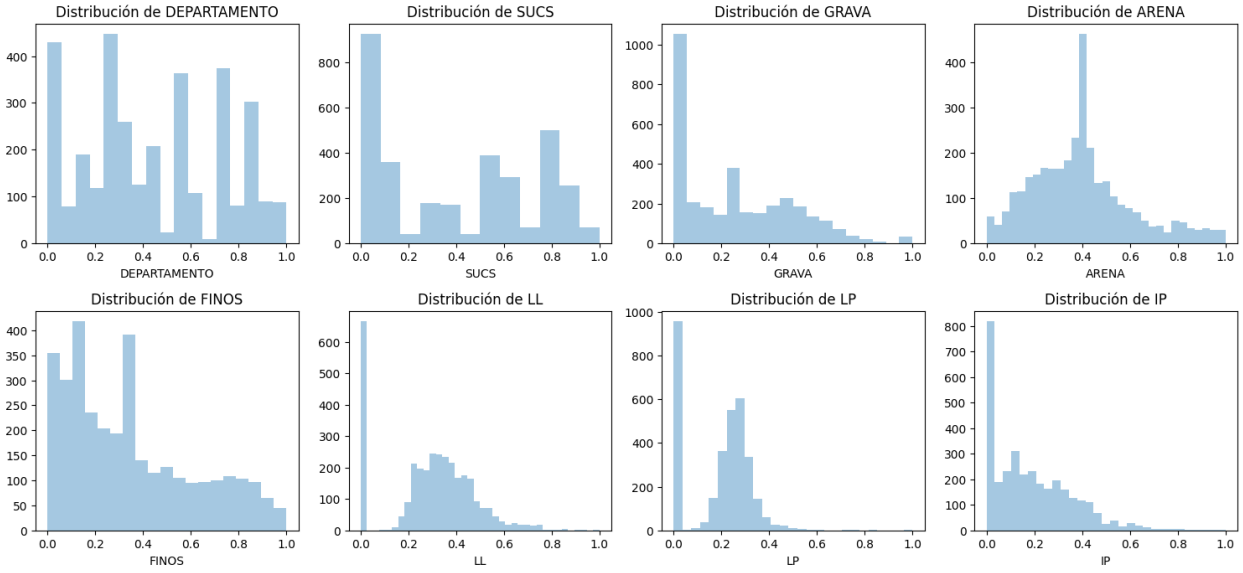
Así como también se realizó la limpieza de los datos para la generación de los modelos de Machine Learning. La limpieza de datos consistió en eliminar datos nulos, corregir errores y eliminar valores atípicos. En la Figura 44 se muestra un diagrama de caja y bigotes para cada una de las variables del conjunto de datos. Se muestra la mediana, el rango intercuartil, los valores mínimo y máximo, y los valores atípicos. La longitud de la caja y los bigotes indica la dispersión de los datos, mientras que la posición de la mediana dentro de la caja indica la tendencia central. La comparación de diagramas de cajas para los diferentes conjuntos de datos permite visualizar las diferencias en la distribución de las variables. Los diagramas de caja y bigotes fue una herramienta útil para visualizar la distribución de los datos y detectar valores atípicos. El análisis de la imagen muestra que se pudieron eliminar algunos valores atípicos en algunas de las variables, como es el caso de la variable OMC que muestra un único valor atípico de 100 muy alejado de la tendencia, el cual pudo haber sido un error de tipeo. La limpieza y tratamiento de valores atípicos permitió mejorar la calidad del conjunto de datos para el rendimiento del modelo. Es importante conocer los valores de los sets de datos para no confundir los datos atípicos con datos que se recolectaron en menor cantidad. Se eliminaron valores atípicos que podrían afectar significativamente el análisis utilizando la técnica de la desviación estándar.

Figura 44 Diagrama de cajas para el tratamiento de datos atípicos para cada variable



Para la normalización de las características del conjunto de datos se utilizó la transformación Min-Max. La transformación Min-Max ayuda a escalar la distribución de las variables del set de datos, que a menudo tienen límites claros y conocidos. Son sensibles a datos atípicos, por ello, primero se reconocieron los límites del set de datos y se eliminaron los datos atípicos que posiblemente se generaron por un error de tipeo. No se utilizaron otras normalizaciones como la normalización logarítmica ya que estos eliminarían los datos atípicos y muchos de los datos atípicos pertenecen a al rango de los valores de cada variable. El conjunto de datos se ha escalado mediante una transformación Min-Max para corregir las diferencias de escala en la distribución de las variables. Esta transformación es útil cuando se trabaja con técnicas sensibles a la escala, como las redes neuronales y los algoritmos de optimización basados en gradientes. En la Figura 45 se muestran histogramas de las variables después de la transformación, evidenciando cómo la transformación Min-Max mejora la distribución y escala los valores entre cero y uno.

Figura 45 Diagrama distribución de la transformación Min-Max para cada variable por primera vez



4.5.2.3. Selección de características y partición de datos

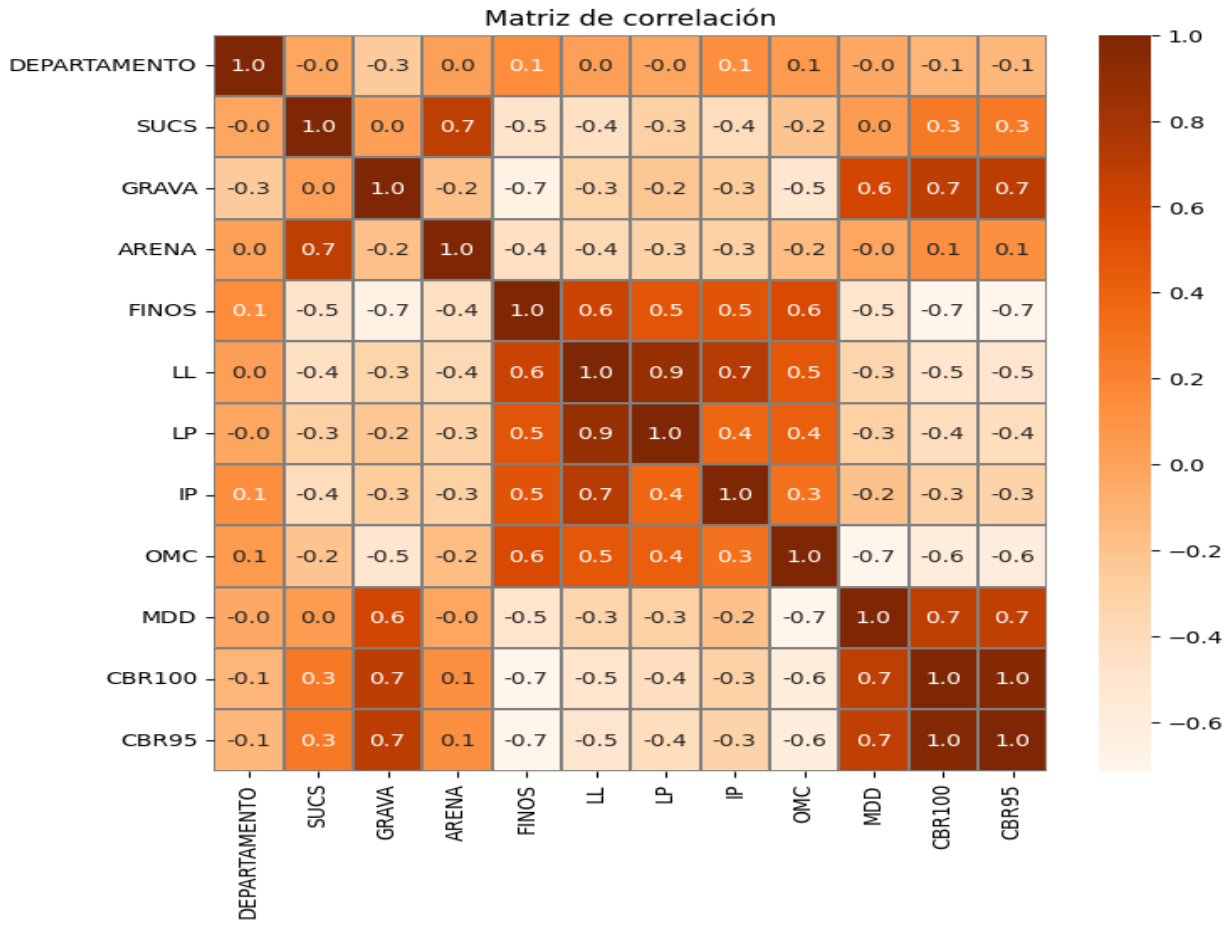
A. Selección de características: correlación de Spearman

Se utilizó el método de correlación de Spearman en las características o variables de entrada del modelo (GRAVA, ARENA, FINOS, LL, LP, IP) para seleccionar aquellas características que tienen una mejor correlación con las variables de salida (OMC, MDD, CBR100 y CBR95). En la Figura 46 las correlaciones de las características del modelo revelan relaciones significativas con las variables de salida. Para la variable de salida OMC destacan correlaciones positivas moderadas entre características como FINOS, LL, LP y IP las cuales indican una relación lineal directa, mientras que GRAVA, ARENA exhiben correlaciones negativas moderadas los cuales indican una relación lineal inversa. Para las variables de salida MDD, CBR100 y CBR95 esto todo lo contrario, donde destacan correlaciones positivas moderadas entre características como GRAVA, ARENA las cuales indican una relación lineal directa, mientras que FINOS, LL, LP y IP exhiben correlaciones negativas moderadas los cuales indican una relación lineal inversa. Así mismo, se observan pocas correlaciones de valor cero, estas correlaciones indican que no existe ninguna relación lineal con las variables objetivo por ende no aportan valor al modelo.

Estos hallazgos proporcionan una visión detallada de cómo cada característica se relaciona entre sí y con la variable de salida, lo que informa sobre la importancia de seleccionar las

características más relevantes para el modelo predictivo. En este caso, como se trata de un modelo de regresión con Machine Learning donde las correlaciones de las variables no siempre son lineales se tomarán tanto las correlaciones negativas, positivas y neutras que serán consideradas para el entrenamiento del modelo. De este modo, con la misma lógica se decidió tomar en cuenta las características SUCS y DEPARTAMENTO como propuesta para compararlos con modelos que no incluyan dichas características, es decir comparar con modelos que solo consideren las propiedades físicas y mecánicas de los suelos con modelos con características adicionales como el que se propone en esta investigación. Con la adición de estás dos características, se concluye que las características mas representativas para la variable objetivo OMC es Finos y Límite Líquido, para las variables objetivos MDD, CBR100 y CBR95 son Grava y SUCS.

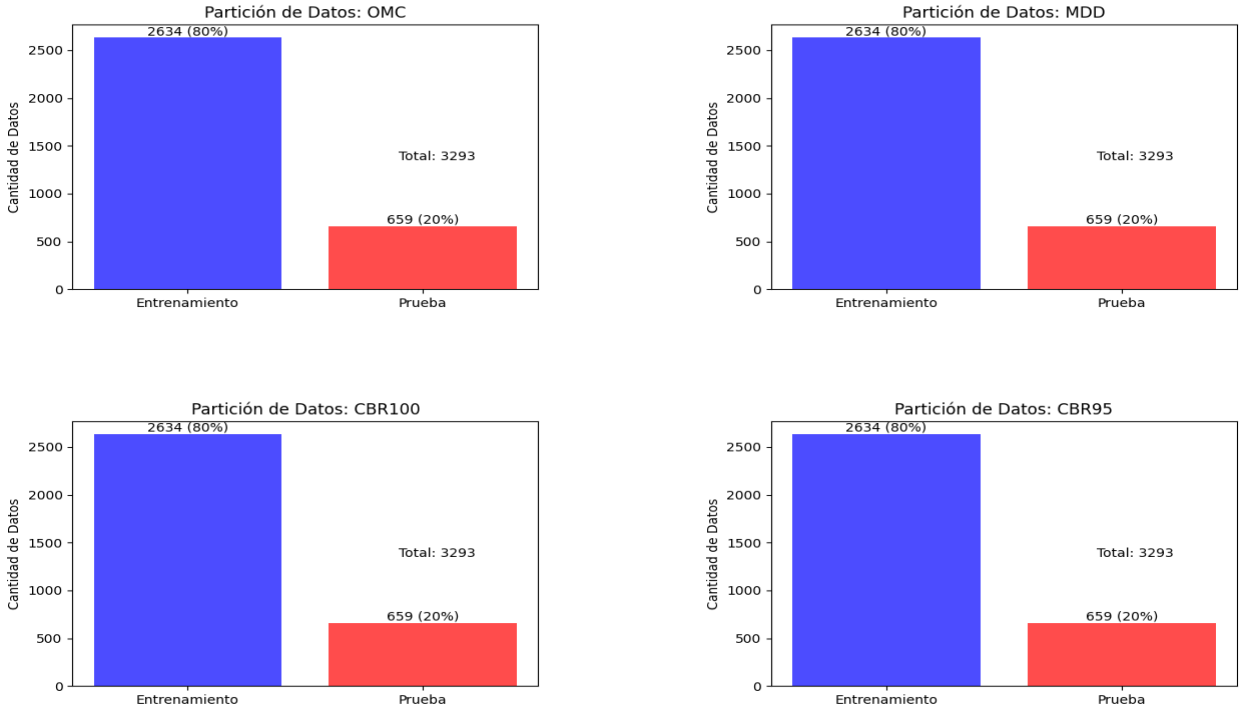
Figura 46 Matriz de correlación de las variables de entrada y salida



B. Partición de datos: entrenamiento y prueba

Posteriormente, se procedió a realizar la partición de los datos de entrenamiento y prueba. Esta división es fundamental en el proceso de entrenamiento y evaluación de modelos de aprendizaje automático, ya que nos permite evaluar el rendimiento del modelo en datos no vistos y comprobar su capacidad de generalización. Primero, los datos se dividieron en características (x) y etiquetas (y). Las características representan las variables independientes, mientras que las etiquetas son las variables dependientes que queremos predecir. Luego, los datos se dividieron nuevamente, esta vez en conjuntos de entrenamiento con un 80% de los datos y prueba con un 20% de los datos. El conjunto de entrenamiento se utiliza para ajustar el modelo, mientras que el conjunto de prueba se reserva para evaluar el rendimiento del modelo en datos no vistos. Además, como el problema implica múltiples variables objetivo (OMC, MDD, CBR95 y CBR100), se realizaron divisiones horizontales adicionales para cada una de estas variables objetivo, tanto en el conjunto de entrenamiento como en el conjunto de prueba. Esto asegura que podamos entrenar y evaluar modelos específicos para cada variable objetivo. En la Figura 47 se muestra gráficamente la partición de los datos de entrenamiento y prueba para cada variable.

Figura 47 Partición de datos entrenamiento y prueba para cada variable



4.5.3. Etapa III: Minería de Datos

En esta sección se aplican diversas técnicas de minería de datos y Machine Learning como Extra Trees Regressor, Random Forest Regressor y Redes Neuronales Artificiales a los diferentes conjuntos de datos de las variables objetivos de CBR95, CBR100, MDD y OMC. Así como también se utilizaron técnicas y herramientas de selección y validación de modelos de Machine Learning como Validación Cruzada, Grid SearchCV, Randomized SearchCV y Auto-ML para mejorar la eficiencia de los diferentes modelos generados.

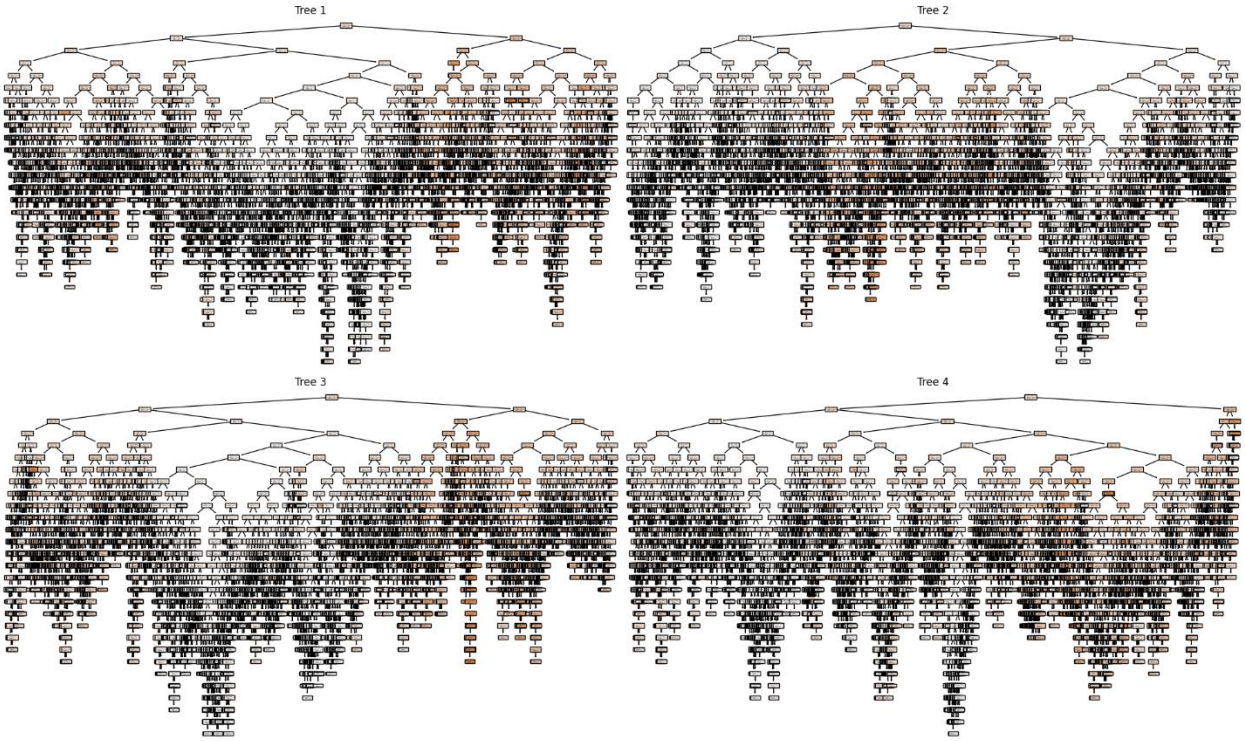
4.5.3.1. Técnica Extra Trees Regressor

La técnica Extra Trees Regressor fue aplicada a los diferentes conjuntos de datos de entrenamiento y prueba con el objetivo de modelar y predecir la variable objetivo. El modelo Extra Trees se ajustó al conjunto de datos de las variables objetivos de CBR95, CBR100, MDD y OMC utilizando los diferentes parámetros e hiperparametros de cada algoritmo o técnica de Machine Learning.

A. Extra Trees Regressor para predecir la variable objetivo CBR95

En la Figura 48 se presenta 4 de 100 árboles del modelo completo de árboles generado mediante la técnica de Extra Trees Regressor. Estos árboles extensos y complejos ilustran la profundidad de la minería de datos aplicada. En cada árbol del bosque de árboles se implementan reglas de decisión para predecir la variable objetivo. Donde cada nodo hoja revela una variable de importancia dentro del modelo, acompañada de su regla de decisión correspondiente, lo que brindó una visión detallada de cómo se estructuraron las predicciones del modelo en función de las características o variables de entrada del modelo para predecir el valor del ensayo de CBR95 en cada árbol del bosque de árboles.

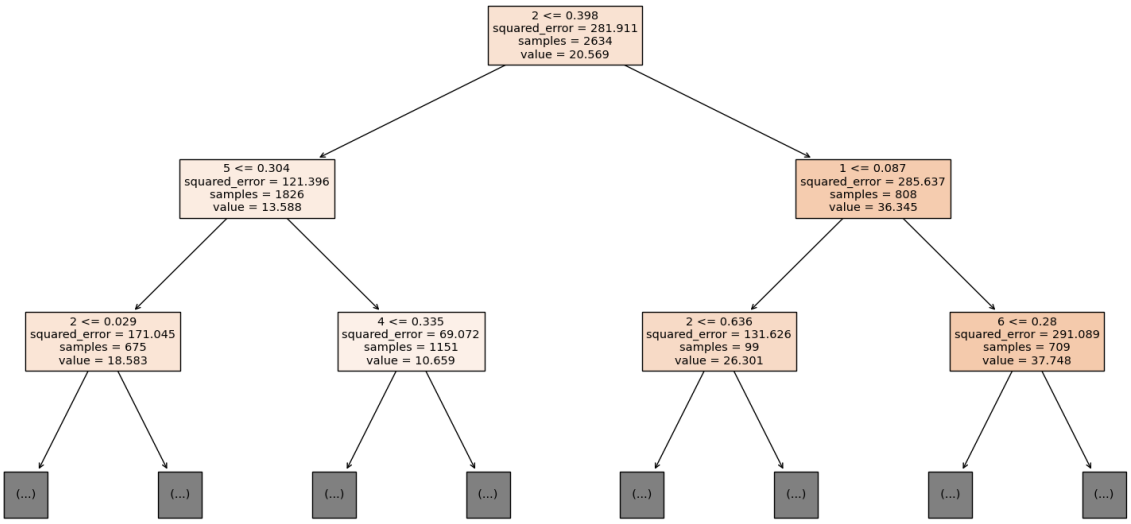
Figura 48 Propiedad Mecánica CBR95: Modelo de árboles utilizando Extra Trees Regressor



Para un mejor entendimiento de cada árbol del bosque de árboles, en la Figura 49 se presenta el modelo de un árbol de la técnica Extra Trees Regressor, con una profundidad máxima de 2 hojas. En este diagrama, se destaca la variable más representativa en el nodo raíz, que en este caso es la variable 2 (Variable SUCS), con una condición de que debe ser menor o igual a 0.398. Además, se visualizan el error cuadrático, que asciende a 281.911, y el valor correspondiente al CBR95, que alcanza los 20.569.

En las siguientes hojas de este árbol de decisión, se observan nuevas reglas de decisión para las otras variables. Cada regla de decisión sucesiva contribuye a disminuir el error cuadrático y ajustar progresivamente el valor de la variable objetivo, CBR95. Este proceso se repite hasta alcanzar el nodo hoja final, que representa el valor de salida del modelo predictivo en función de las variables de entrada al árbol de decisión.

Figura 49 Propiedad Mecánica CBR95: Modelo de un árbol de profundidad de 2 hojas utilizando Extra Trees Regressor



Para generar dichos árboles y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 07 los siguientes hiperparámetros utilizados en la técnica de regresión Extra Trees Regressor:

Tabla 07 Hiper parámetros utilizados Extra Trees Regressor para predecir el CBR95

Descripción	***Valor***
Bootstrap	False
Ccp_alpha	0.0
Criterion	Squared_error
Max_depth	None
Max_features	1.0
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_samples_leaf	1
Min_samples_split	2
Min_weight_fraction_leaf	0.0
N_stimators	100
N_jobs	None
Oob_score	False
Random_state	42
Verbose	0
Warm_start	False

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 63.3174 (%), un Root Mean Squared Error de 7.9572 (%), Mean Absolute Error de 4.9235 (%) y R-Squared de 0.7750.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

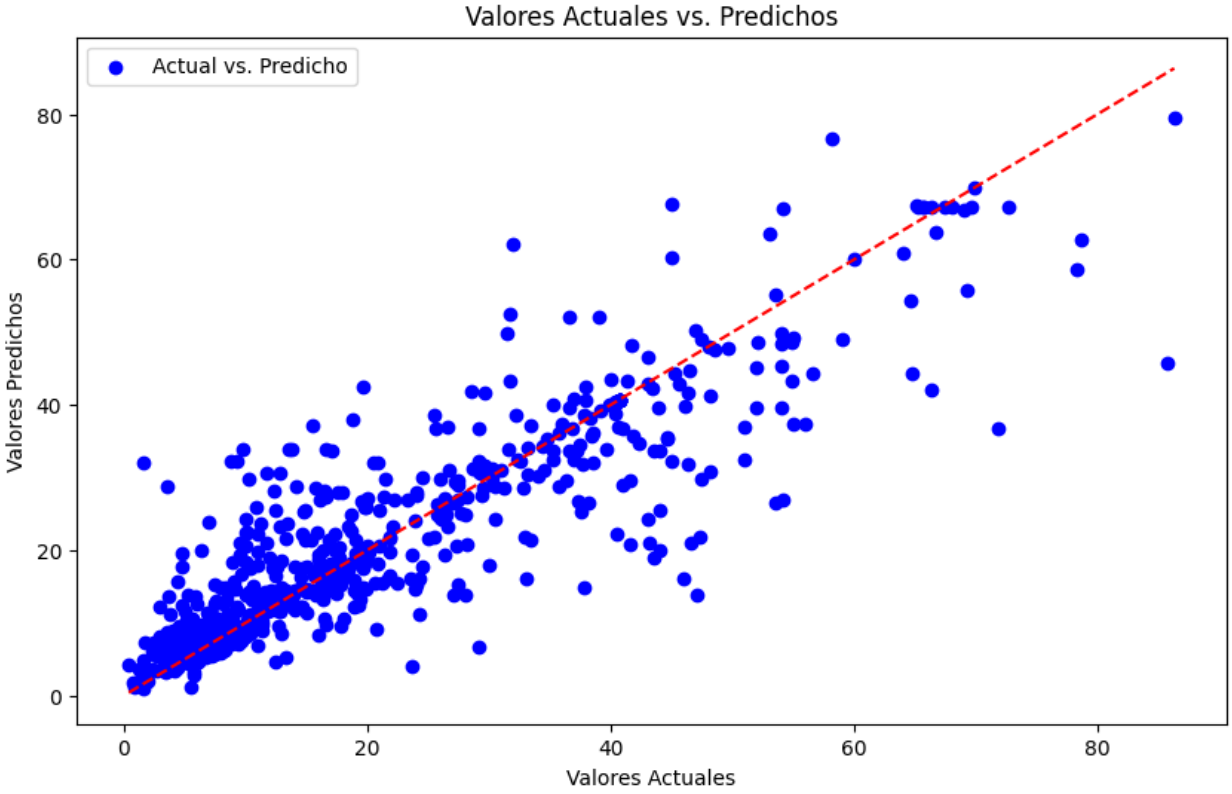
Tabla 08 Hiper parámetros con Grid SearchCV de Extra Trees Regressor para predecir el CBR95

Descripción	***Valor***
Max_depth	None
Max_features	sqrt
Min_samples_leaf	1
Min_samples_split	5
N_stimators	300
Random_state	42

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 62.1625 (%), un Root Mean Squared Error de 7.8843 (%), Mean Absolute Error de 5.0304 (%) y R-Squared de 0.7791.

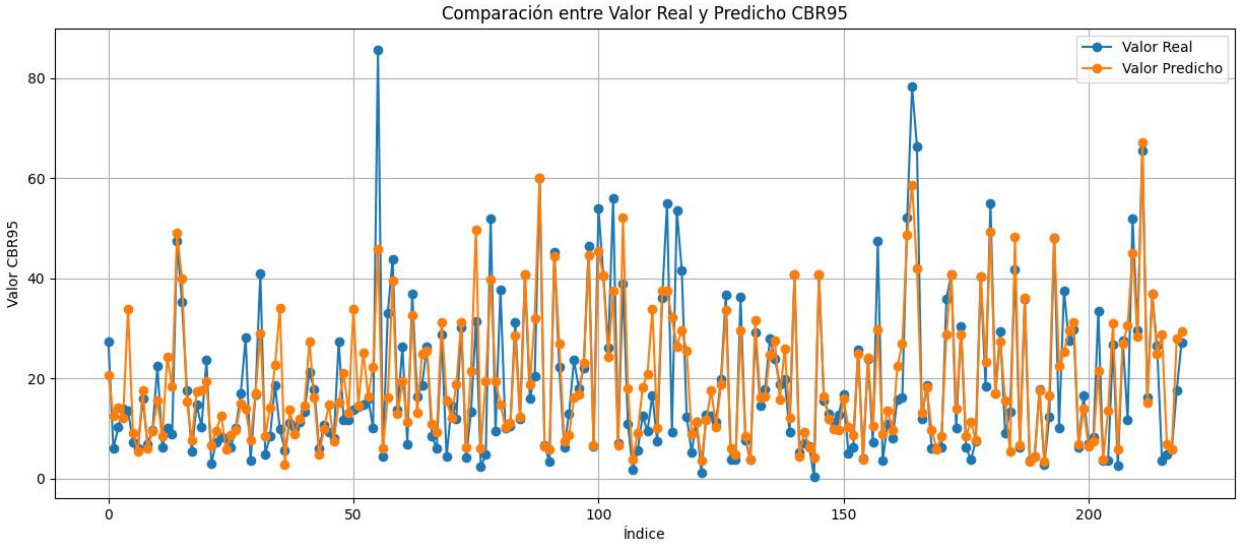
A continuación, en la Figura 50 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el CBR95 con la técnica Extra Trees Regressor se ajusta a la línea de regresión de los valores reales y predichos.

Figura 50 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Extra Trees Regressor para predecir el CBR95



Finalmente, se muestra el diagrama de líneas en la Figura 51, donde se puede verificar la similitud entre los valores reales y valores predichos del CBR95 para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

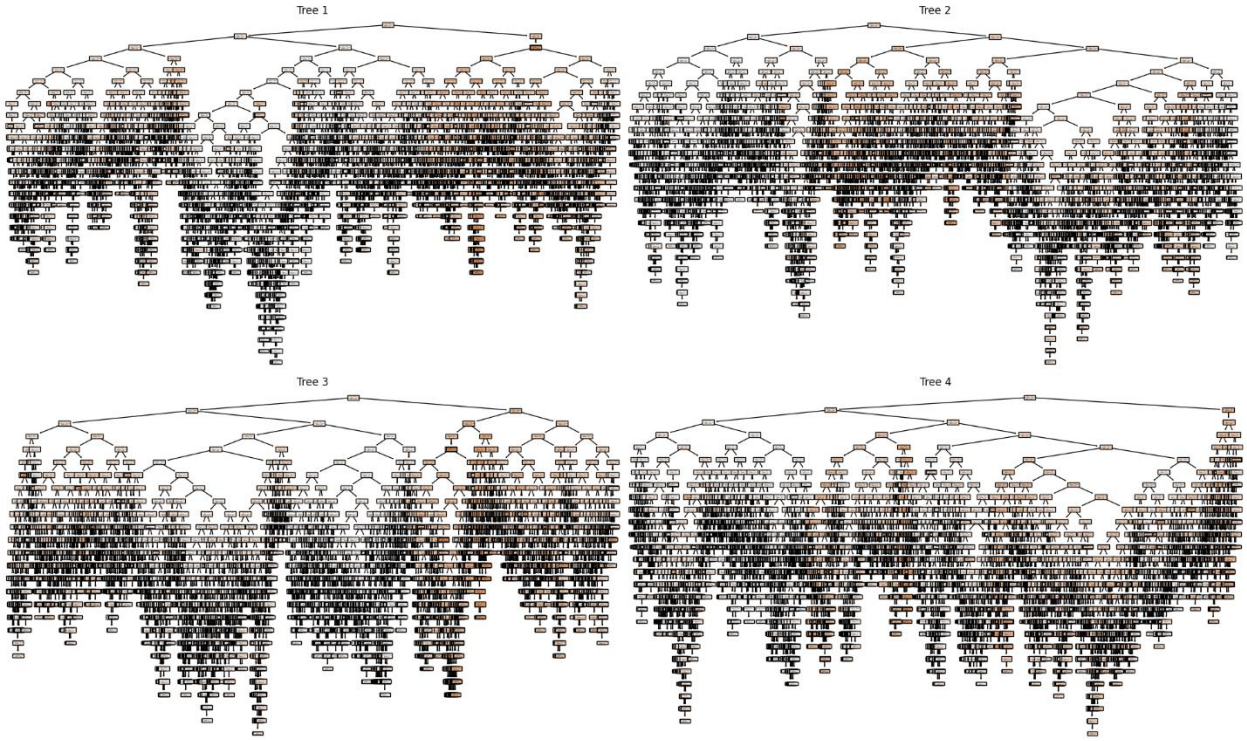
Figura 51 Diagrama de líneas de los valores Reales vs. Predichos utilizando Extra Trees Regressor para predecir el CBR95



B. Extra Trees Regressor para predecir la variable objetivo CBR100

En la Figura 52 se presenta 4 de 100 árboles del modelo completo de árboles generado mediante la técnica de Extra Trees Regressor. Estos árboles extensos y complejos ilustran la profundidad de la minería de datos aplicada. En cada árbol del bosque de árboles se implementan reglas de decisión para predecir la variable objetivo. Donde cada nodo hoja revela una variable de importancia dentro del modelo, acompañada de su regla de decisión correspondiente, lo que brindó una visión detallada de cómo se estructuraron las predicciones del modelo en función de las características o variables de entrada del modelo para predecir el valor del ensayo de CBR100 en cada árbol del bosque de árboles.

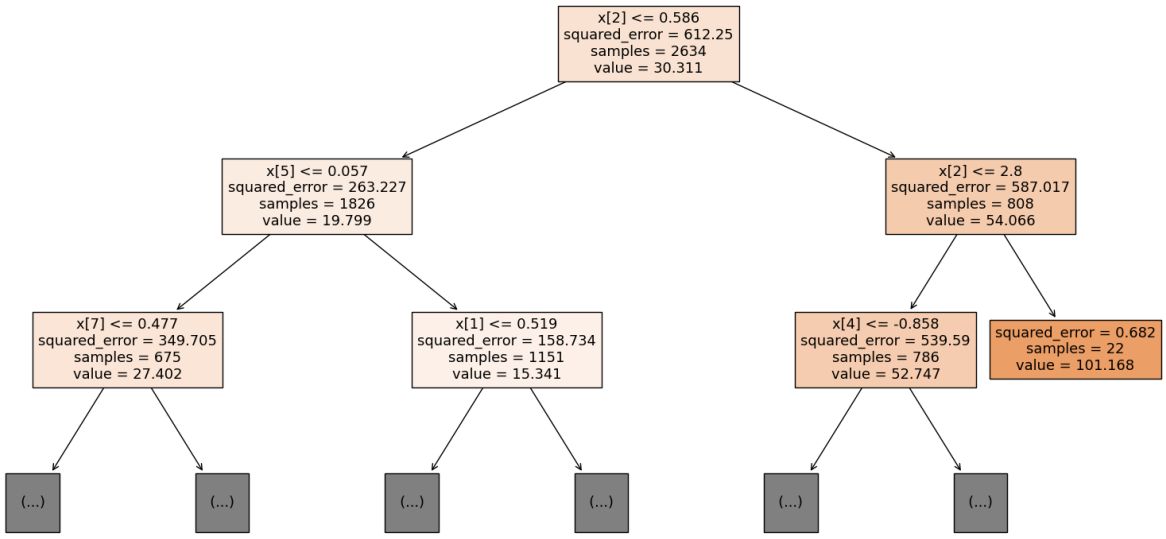
Figura 52 Propiedad Mecánica CBR100: Modelo de árboles utilizando Extra Trees Regressor



Para un mejor entendimiento de cada árbol del bosque de árboles, en la Figura 53 se presenta el modelo de un árbol de la técnica Extra Trees Regressor, con una profundidad máxima de 2 hojas. En este diagrama, se destaca la variable más representativa en el nodo raíz, que en este caso es la variable 2 (Variable SUCS), con una condición de que debe ser menor o igual a 0.586. Además, se visualizan el error cuadrático, que asciende a 281.911, y el valor correspondiente al CBR100, que alcanza los 20.569.

En las siguientes hojas de este árbol de decisión, se observan nuevas reglas de decisión para las otras variables. Cada regla de decisión sucesiva contribuye a disminuir el error cuadrático y ajustar progresivamente el valor de la variable objetivo, CBR100. Este proceso se repite hasta alcanzar el nodo hoja final, que representa el valor de salida del modelo predictivo en función de las variables de entrada al árbol de decisión.

Figura 53 Propiedad Mecánica CBR100: Modelo de un árbol de profundidad de 2 hojas utilizando Extra Trees Regressor



Para generar dichos árboles y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 09 los siguientes hiperparámetros utilizados en la técnica de regresión Extra Trees Regressor:

Tabla 09 Hiper parámetros utilizados Extra Trees Regressor para predecir el CBR100

Descripción	***Valor***
Bootstrap	False
Ccp_alpha	0.0
Criterion	Squared_error
Max_depth	None
Max_features	1.0
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_samples_leaf	1
Min_samples_split	2
Min_weight_fraction_leaf	0.0
N_stimators	100
N_jobs	None
Oob_score	False
Random_state	42
Verbose	0
Warm_start	False

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 163.2917 (%), un Root Mean Squared Error de 12.7785 (%), Mean Absolute Error de 7.6829 (%) y R-Squared de 0.7572.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

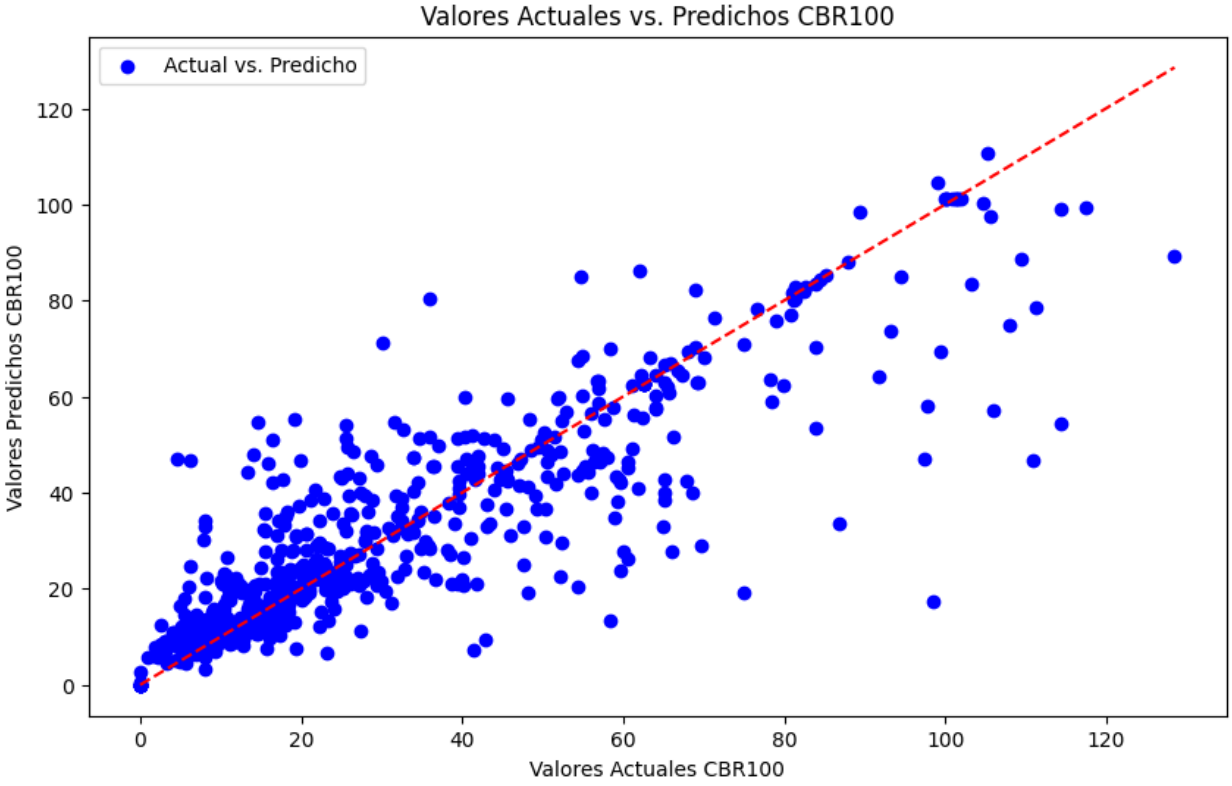
Tabla 10 Hiper parámetros con Grid SearchCV de Extra Trees Regressor para predecir el CBR100

Descripción	***Valor***
Max_depth	None
Max_features	sqrt
Min_samples_leaf	1
Min_samples_split	5
N_stimators	300
Random_state	42

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 163.4903 (%), un Root Mean Squared Error de 12.7863 (%), Mean Absolute Error de 7.7973 (%) y R-Squared de 0.7569.

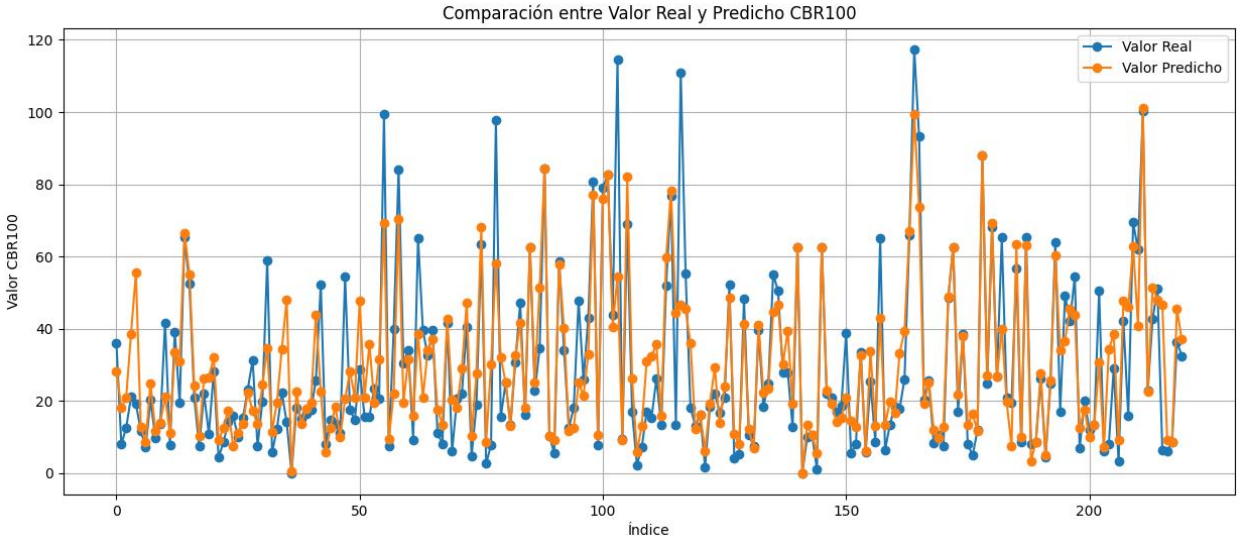
A continuación, en la Figura 54 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el CBR100 con la técnica Extra Trees Regressor se ajusta a la línea de regresión de los valores reales y predichos.

Figura 54 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Extra Trees Regressor para predecir el CBR100



Finalmente, se muestra el diagrama de líneas en la Figura 55, donde se puede verificar la similitud entre los valores reales y valores predichos del CBR100 para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

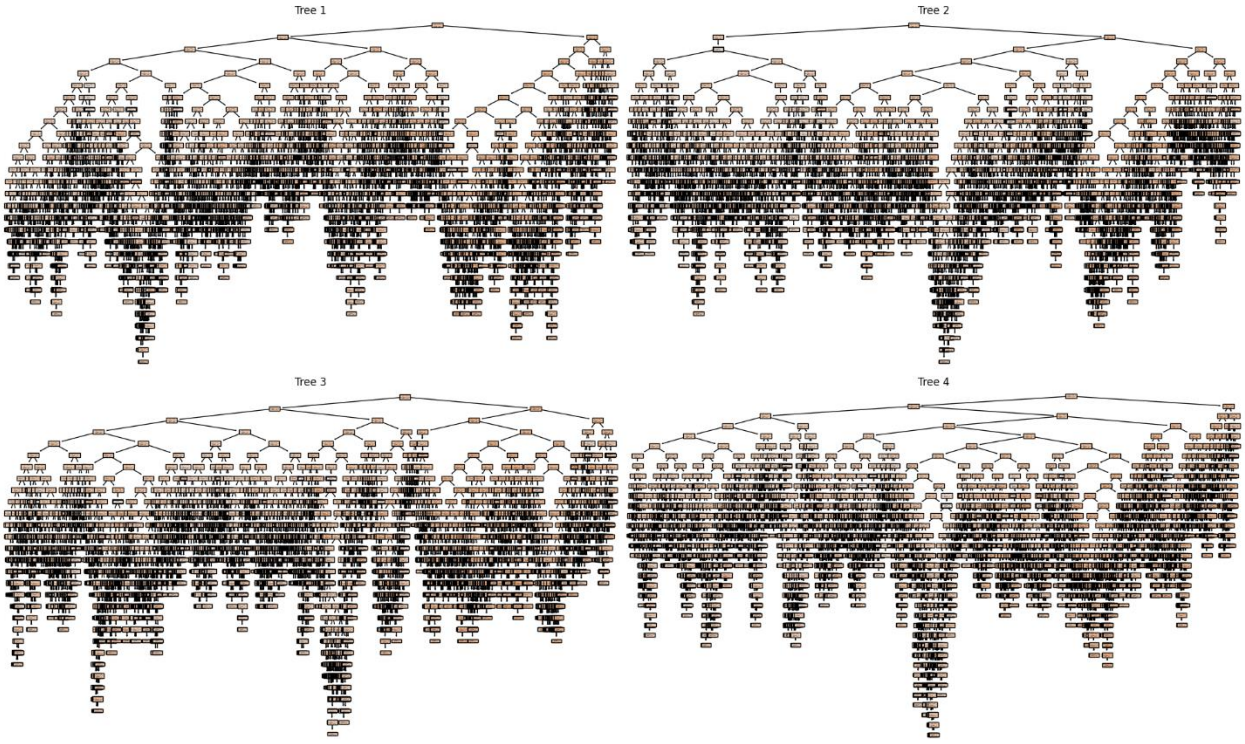
Figura 55 Diagrama de líneas de los valores Reales vs. Predichos utilizando Extra Trees Regressor para predecir el CBR100



C. Extra Trees Regressor para predecir la variable objetivo MDD

En la Figura 56 se presenta 4 de 100 árboles del modelo completo de árboles generado mediante la técnica de Extra Trees Regressor. Estos árboles extensos y complejos ilustran la profundidad de la minería de datos aplicada. En cada árbol del bosque de árboles se implementan reglas de decisión para predecir la variable objetivo. Donde cada nodo hoja revela una variable de importancia dentro del modelo, acompañada de su regla de decisión correspondiente, lo que brindó una visión detallada de cómo se estructuraron las predicciones del modelo en función de las características o variables de entrada del modelo para predecir el valor del ensayo de MDD en cada árbol del bosque de árboles.

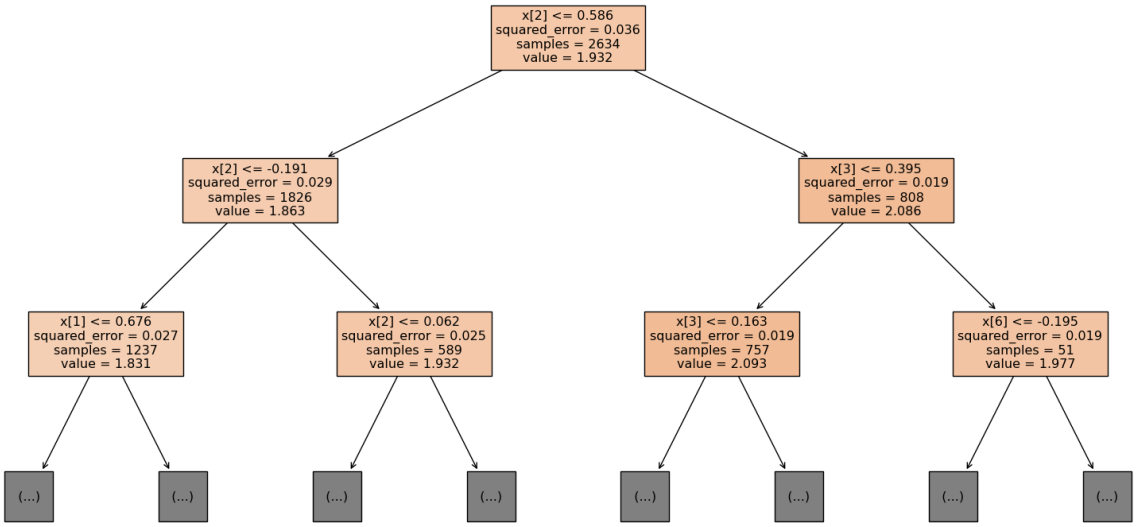
Figura 56 Propiedad Mecánica MDD: Modelo de árboles utilizando Extra Trees Regressor



Para un mejor entendimiento de cada árbol del bosque de árboles, en la Figura 57 se presenta el modelo de un árbol de la técnica Extra Trees Regressor, con una profundidad máxima de 2 hojas. En este diagrama, se destaca la variable más representativa en el nodo raíz, que en este caso es la variable 2 (Variable SUCS), con una condición de que debe ser menor o igual a 0.398. Además, se visualizan el error cuadrático, que asciende a 281.911, y el valor correspondiente al MDD, que alcanza los 20.569.

En las siguientes hojas de este árbol de decisión, se observan nuevas reglas de decisión para las otras variables. Cada regla de decisión sucesiva contribuye a disminuir el error cuadrático y ajustar progresivamente el valor de la variable objetivo, MDD. Este proceso se repite hasta alcanzar el nodo hoja final, que representa el valor de salida del modelo predictivo en función de las variables de entrada al árbol de decisión.

Figura 57 Propiedad Mecánica MDD: Modelo de un árbol de profundidad de 2 hojas utilizando Extra Trees Regressor



Para generar dichos árboles y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 11 los siguientes hiperparámetros utilizados en la técnica de regresión Extra Trees Regressor:

Tabla 11 Hiper parámetros utilizados Extra Trees Regressor para predecir el MDD

Descripción	***Valor***
Bootstrap	False
Ccp_alpha	0.0
Criterion	Squared_error
Max_depth	None
Max_features	1.0
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_samples_leaf	1
Min_samples_split	2
Min_weight_fraction_leaf	0.0
N_stimators	100
N_jobs	None
Oob_score	False
Random_state	42
Verbose	0
Warm_start	False

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 0.0162 (g/cm3), un Root Mean Squared Error de 0.1275 (g/cm3), Mean Absolute Error de 0.0835 (g/cm3) y R-Squared de 0.5782.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

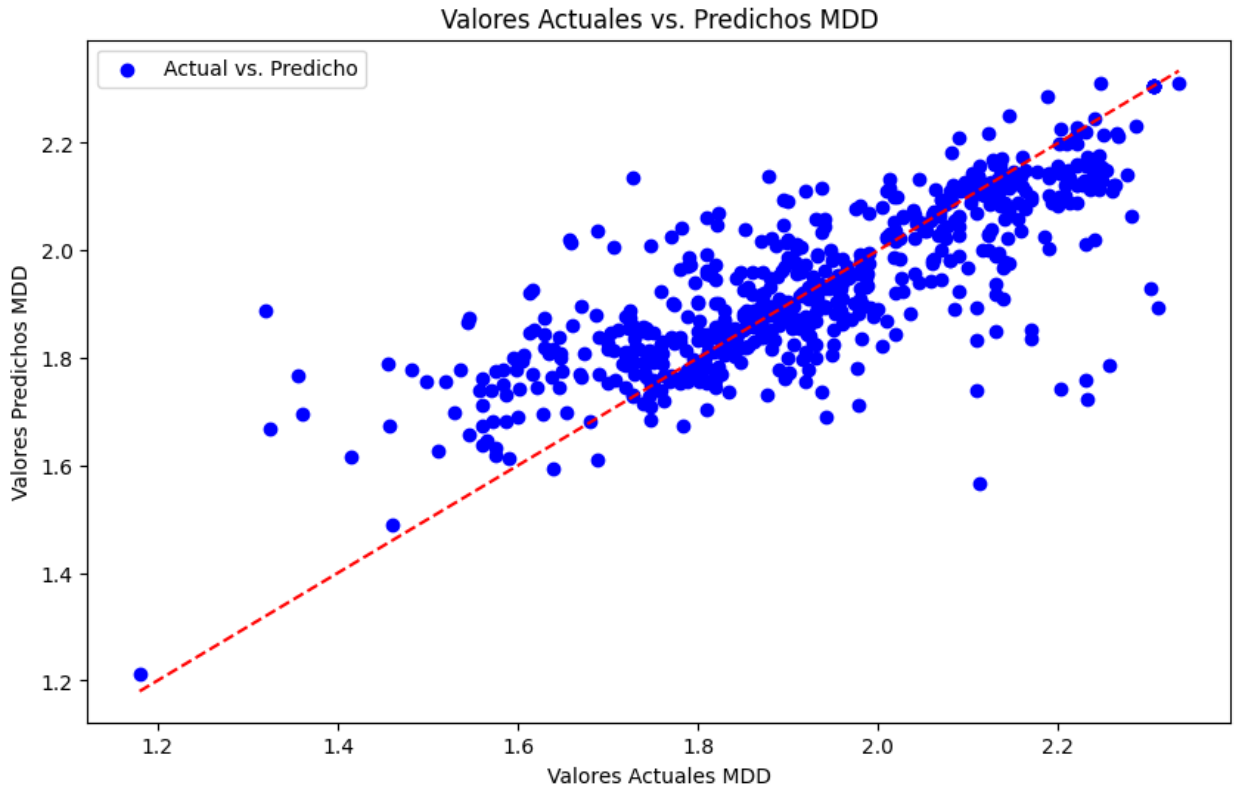
Tabla 12 Hiper parámetros con Grid SearchCV de Extra Trees Regressor para predecir el MDD

Descripción	***Valor***
Max_depth	None
Max_features	sqrt
Min_samples_leaf	1
Min_samples_split	5
N_stimators	300
Random_state	42

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 0.0142 (g/cm3), un Root Mean Squared Error de 0.1194 (g/cm3), Mean Absolute Error de 0.0820 (g/cm3) y R-Squared de 0.6298.

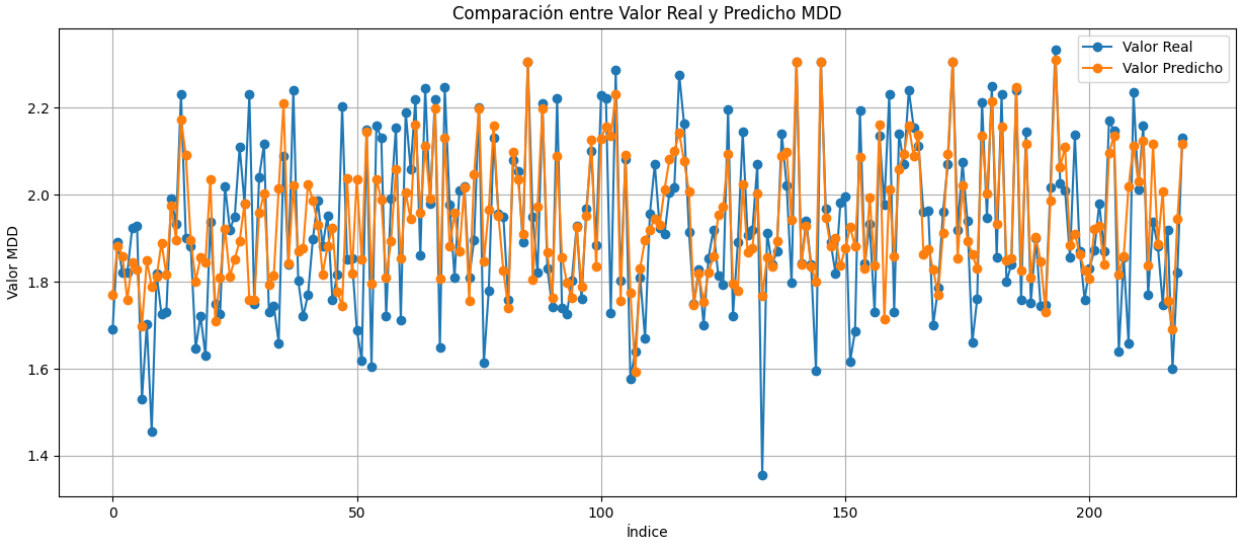
A continuación, en la Figura 58 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el MDD con la técnica Extra Trees Regressor se ajusta a la línea de regresión de los valores reales y predichos.

Figura 58 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Extra Trees Regressor para predecir el MDD



Finalmente, se muestra el diagrama de líneas en la Figura 59, donde se puede verificar la similitud entre los valores reales y valores predichos del MDD para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

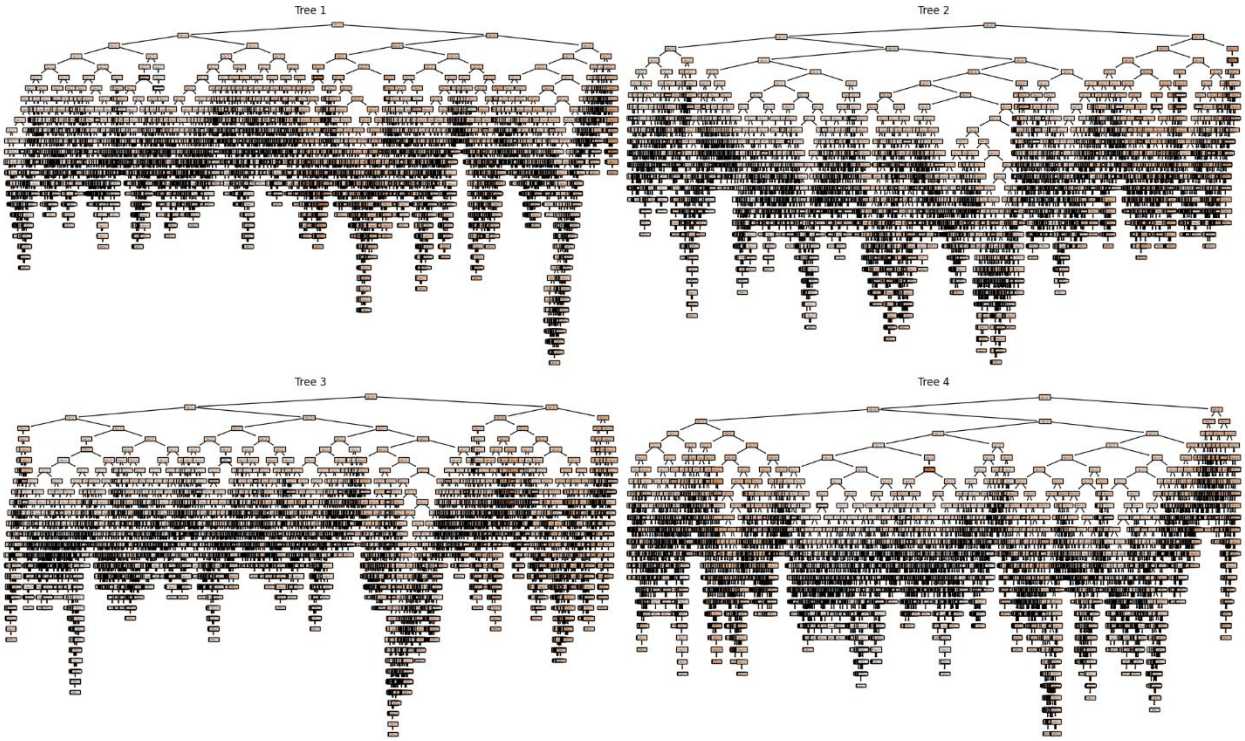
Figura 59 Diagrama de líneas de los valores Reales vs. Predichos utilizando Extra Trees Regressor para predecir el MDD



D. Extra Trees Regressor para predecir la variable objetivo OMC

En la Figura 60 se presenta 4 de 100 árboles del modelo completo de árboles generado mediante la técnica de Extra Trees Regressor. Estos árboles extensos y complejos ilustran la profundidad de la minería de datos aplicada. En cada árbol del bosque de árboles se implementan reglas de decisión para predecir la variable objetivo. Donde cada nodo hoja revela una variable de importancia dentro del modelo, acompañada de su regla de decisión correspondiente, lo que brindó una visión detallada de cómo se estructuraron las predicciones del modelo en función de las características o variables de entrada del modelo para predecir el valor del ensayo de OMC en cada árbol del bosque de árboles.

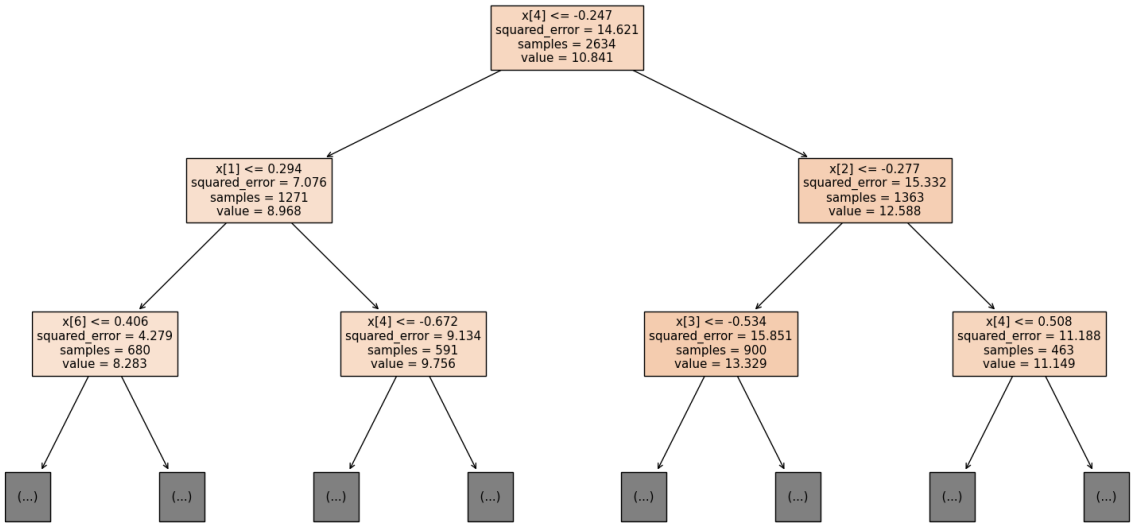
Figura 60 Propiedad Mecánica OMC: Modelo de árboles utilizando Extra Trees Regressor



Para un mejor entendimiento de cada árbol del bosque de árboles, en la Figura 61 se presenta el modelo de un árbol de la técnica Extra Trees Regressor, con una profundidad máxima de 2 hojas. En este diagrama, se destaca la variable más representativa en el nodo raíz, que en este caso es la variable 2 (Variable SUCS), con una condición de que debe ser menor o igual a 0.398. Además, se visualizan el error cuadrático, que asciende a 281.911, y el valor correspondiente al OMC, que alcanza los 20.569.

En las siguientes hojas de este árbol de decisión, se observan nuevas reglas de decisión para las otras variables. Cada regla de decisión sucesiva contribuye a disminuir el error cuadrático y ajustar progresivamente el valor de la variable objetivo, OMC. Este proceso se repite hasta alcanzar el nodo hoja final, que representa el valor de salida del modelo predictivo en función de las variables de entrada al árbol de decisión.

Figura 61 Propiedad Mecánica OMC: Modelo de un árbol de profundidad de 2 hojas utilizando Extra Trees Regressor



Para generar dichos árboles y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 13 los siguientes hiperparámetros utilizados en la técnica de regresión Extra Trees Regressor:

Tabla 13 Hiper parámetros utilizados Extra Trees Regressor para predecir el OMC

Descripción	***Valor***
Bootstrap	False
Ccp_alpha	0.0
Criterion	Squared_error
Max_depth	None
Max_features	1.0
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_samples_leaf	1
Min_samples_split	2
Min_weight_fraction_leaf	0.0
N_stimators	100
N_jobs	None
Oob_score	False
Random_state	42
Verbose	0
Warm_start	False

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 8.5054 (%), un Root Mean Squared Error de 2.9164 (%), Mean Absolute Error de 1.7624 (%) y R-Squared de 0.4966.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

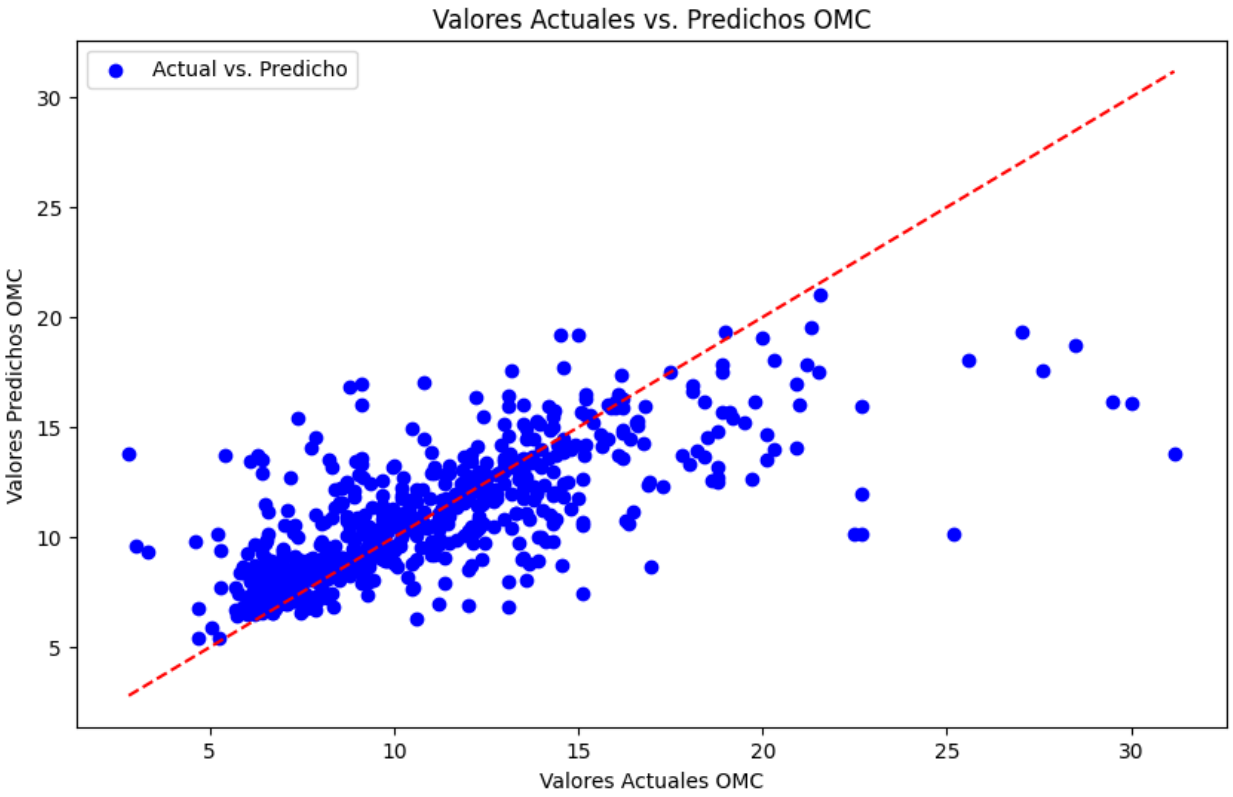
Tabla 14 Hiper parámetros con Grid SearchCV de Extra Trees Regressor para predecir el OMC

Descripción	***Valor***
Max_depth	None
Max_features	sqrt
Min_samples_leaf	1
Min_samples_split	5
N_stimators	300
Random_state	42

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 7.7777 (%), un Root Mean Squared Error de 2.7888 (%), Mean Absolute Error de 1.7296 (%) y R-Squared de 0.5397.

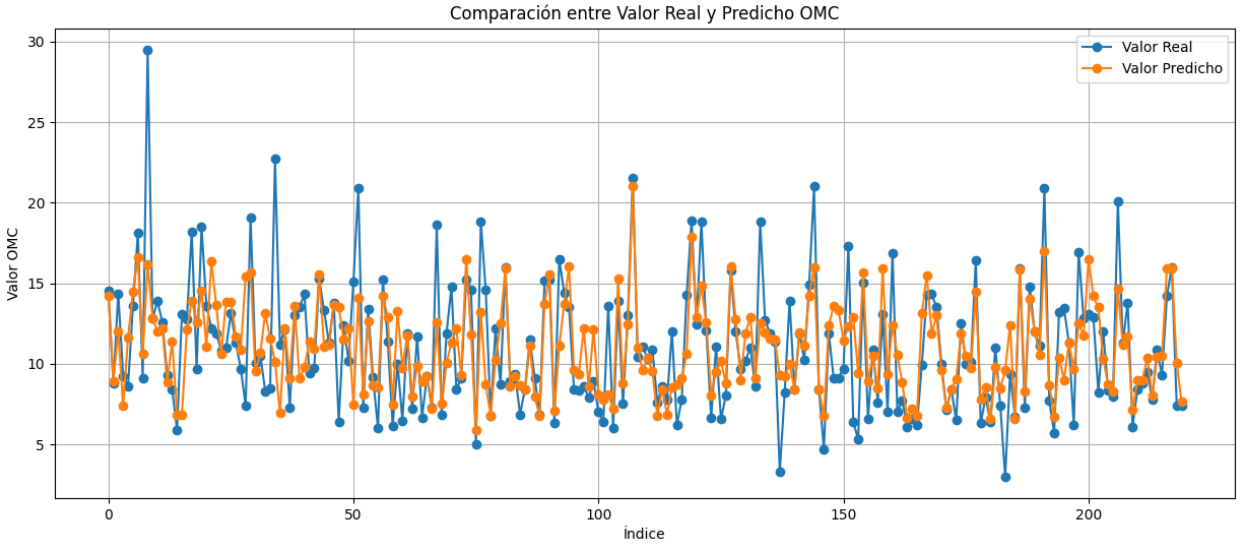
A continuación, en la Figura 62 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el OMC con la técnica Extra Trees Regressor se ajusta a la línea de regresión de los valores reales y predichos.

Figura 62 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Extra Trees Regressor para predecir el OMC



Finalmente, se muestra el diagrama de líneas en la Figura 63, donde se puede verificar la similitud entre los valores reales y valores predichos del OMC para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

Figura 63 Diagrama de líneas de los valores Reales vs. Predichos utilizando Extra Trees Regressor para predecir el OMC



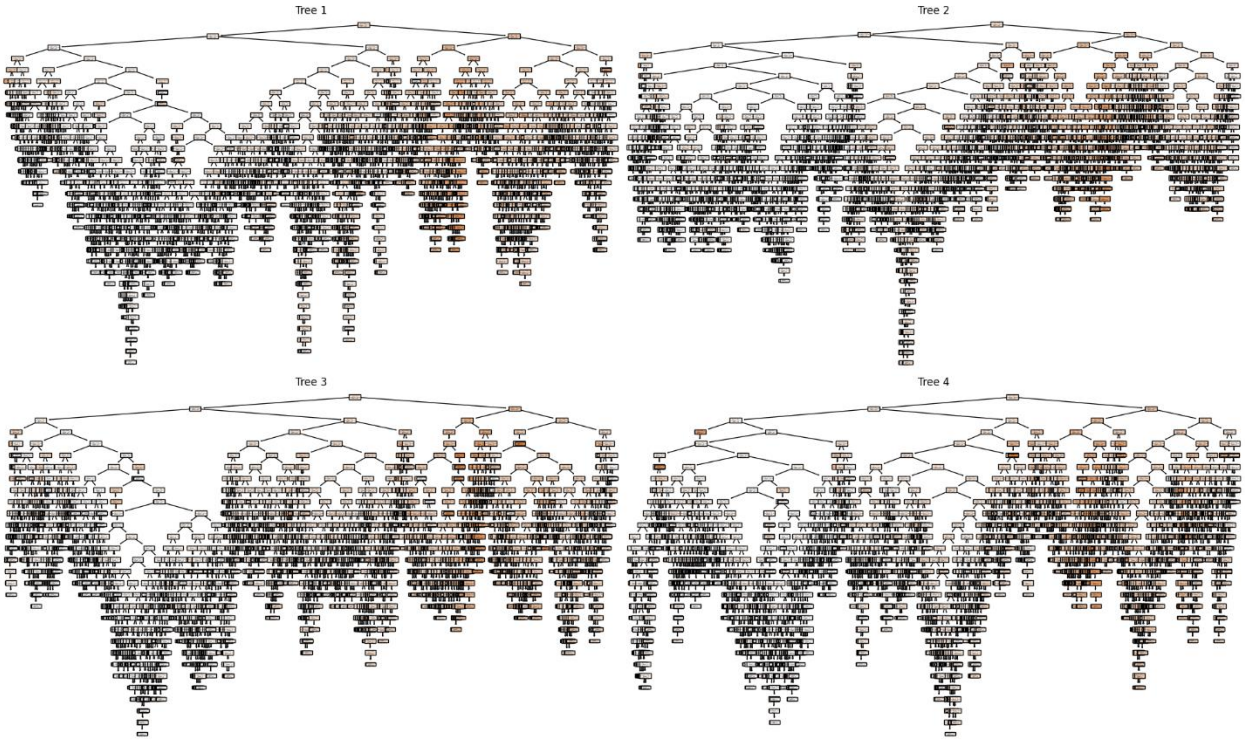
4.5.3.2. Técnica Random Forest Regressor

La técnica Random Forest Regressor fue aplicada a los diferentes conjuntos de datos de entrenamiento y prueba con el objetivo de modelar y predecir la variable objetivo. El modelo Random Forest se ajustó al conjunto de datos de las variables objetivos de CBR95, CBR100, MDD y OMC utilizando los diferentes parámetros e hiperparametros de cada algoritmo o técnica de Machine Learning.

A. Random Forest Regressor para predecir la variable objetivo CBR95

En la Figura 64 se presenta 4 de 100 árboles del modelo completo de árboles generado mediante la técnica de Random Forest Regressor. Estos árboles extensos y complejos ilustran la profundidad de la minería de datos aplicada. En cada árbol del bosque de árboles se implementan reglas de decisión para predecir la variable objetivo. Donde cada nodo hoja revela una variable de importancia dentro del modelo, acompañada de su regla de decisión correspondiente, lo que brindó una visión detallada de cómo se estructuraron las predicciones del modelo en función de las características o variables de entrada del modelo para predecir el valor del ensayo de CBR95 en cada árbol del bosque de árboles.

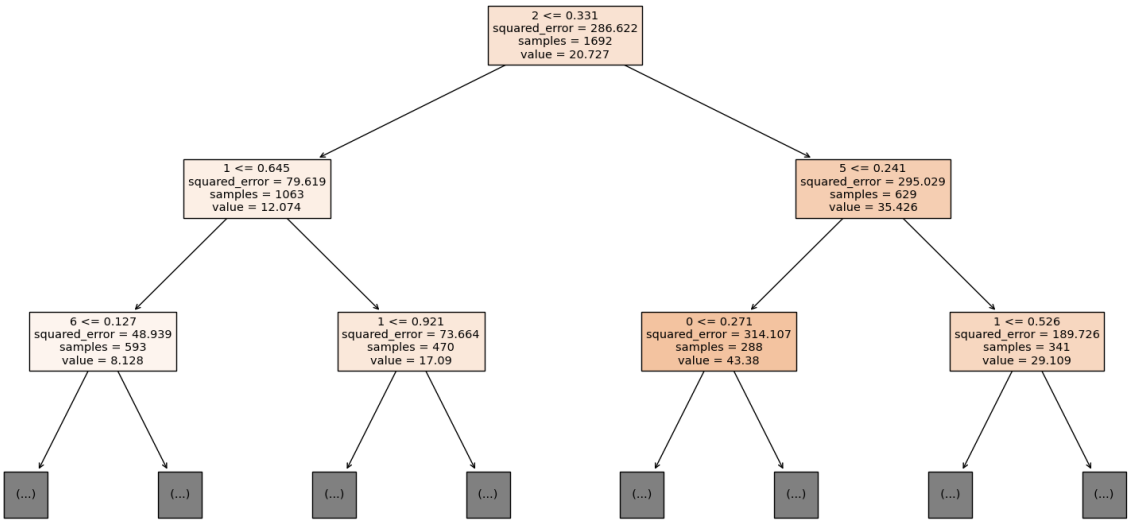
Figura 64 Propiedad Mecánica CBR95: Modelo de árboles utilizando Random Forest Regressor



Para un mejor entendimiento de cada árbol del bosque de árboles, en la Figura 65 se presenta el modelo de un árbol de la técnica Random Forest Regressor, con una profundidad máxima de 2 hojas. En este diagrama, se destaca la variable más representativa en el nodo raíz, que en este caso es la variable 2 (Variable SUCS), con una condición de que debe ser menor o igual a 0.398. Además, se visualizan el error cuadrático, que asciende a 281.911, y el valor correspondiente al CBR95, que alcanza los 20.569.

En las siguientes hojas de este árbol de decisión, se observan nuevas reglas de decisión para las otras variables. Cada regla de decisión sucesiva contribuye a disminuir el error cuadrático y ajustar progresivamente el valor de la variable objetivo, CBR95. Este proceso se repite hasta alcanzar el nodo hoja final, que representa el valor de salida del modelo predictivo en función de las variables de entrada al árbol de decisión.

Figura 65 Propiedad Mecánica CBR95: Modelo de un árbol de profundidad de 2 hojas utilizando Random Forest Regressor



Para generar dichos árboles y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 15 los siguientes hiperparámetros utilizados en la técnica de regresión Random Forest Regressor:

Tabla 15 Hiper parámetros utilizados Random Forest Regressor para predecir el CBR95

Descripción	***Valor***
Bootstrap	False
Ccp_alpha	0.0
Criterion	Squared_error
Max_depth	None
Max_features	1.0
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_samples_leaf	1
Min_samples_split	2
Min_weight_fraction_leaf	0.0
N_stimators	100
N_jobs	None
Oob_score	False
Random_state	42
Verbose	0
Warm_start	False

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 63.4443 (%), un Root Mean Squared Error de 7.9651 (%), Mean Absolute Error de 5.0974 (%) y R-Squared de 0.7745.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

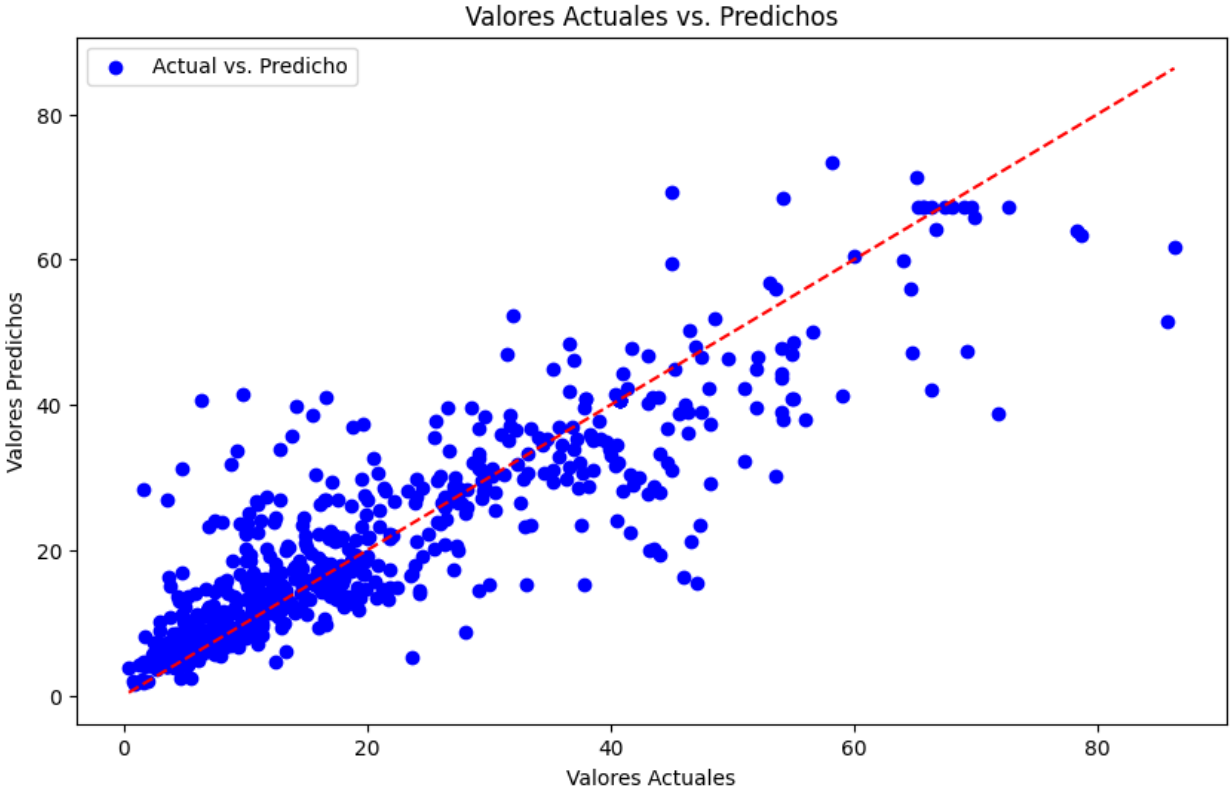
Tabla 16 Hiper parámetros con Grid SearchCV de Random Forest Regressor para predecir el CBR95

Descripción	***Valor***
Max_depth	None
Max_features	sqrt
Min_samples_leaf	1
Min_samples_split	5
N_stimators	300
Random_state	42

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 63.4443 (%), un Root Mean Squared Error de 7.9651 (%), Mean Absolute Error de 5.0974 (%) y R-Squared de 0.7745.

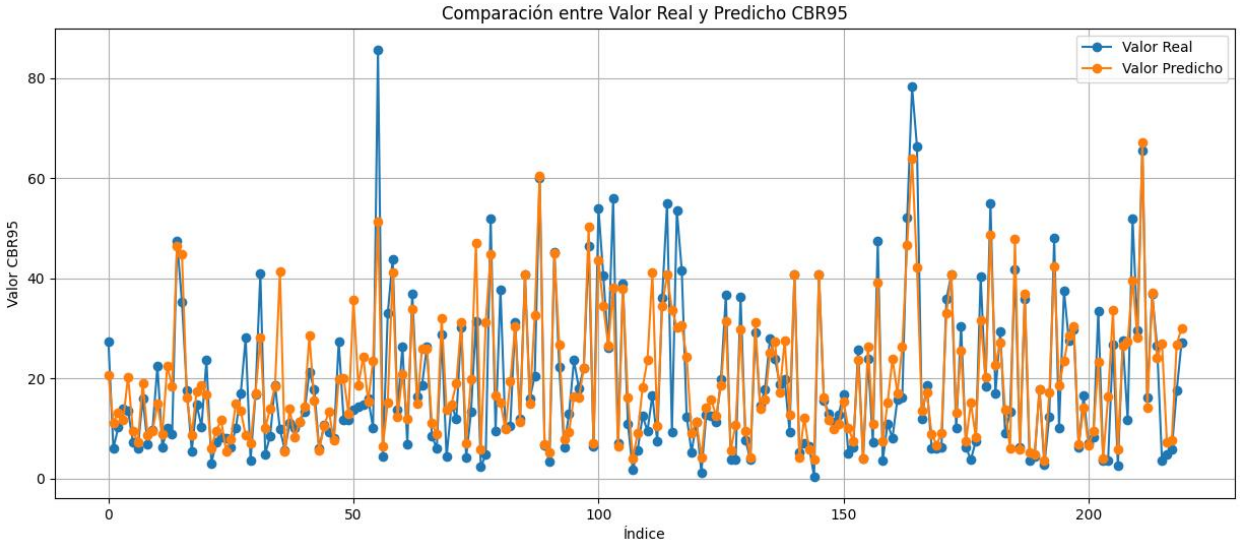
A continuación, en la Figura 66 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el CBR95 con la técnica Random Forest Regressor se ajusta a la línea de regresión de los valores reales y predichos.

Figura 66 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Random Forest Regressor para predecir el CBR95



Finalmente, se muestra el diagrama de líneas en la Figura 67, donde se puede verificar la similitud entre los valores reales y valores predichos del CBR95 para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

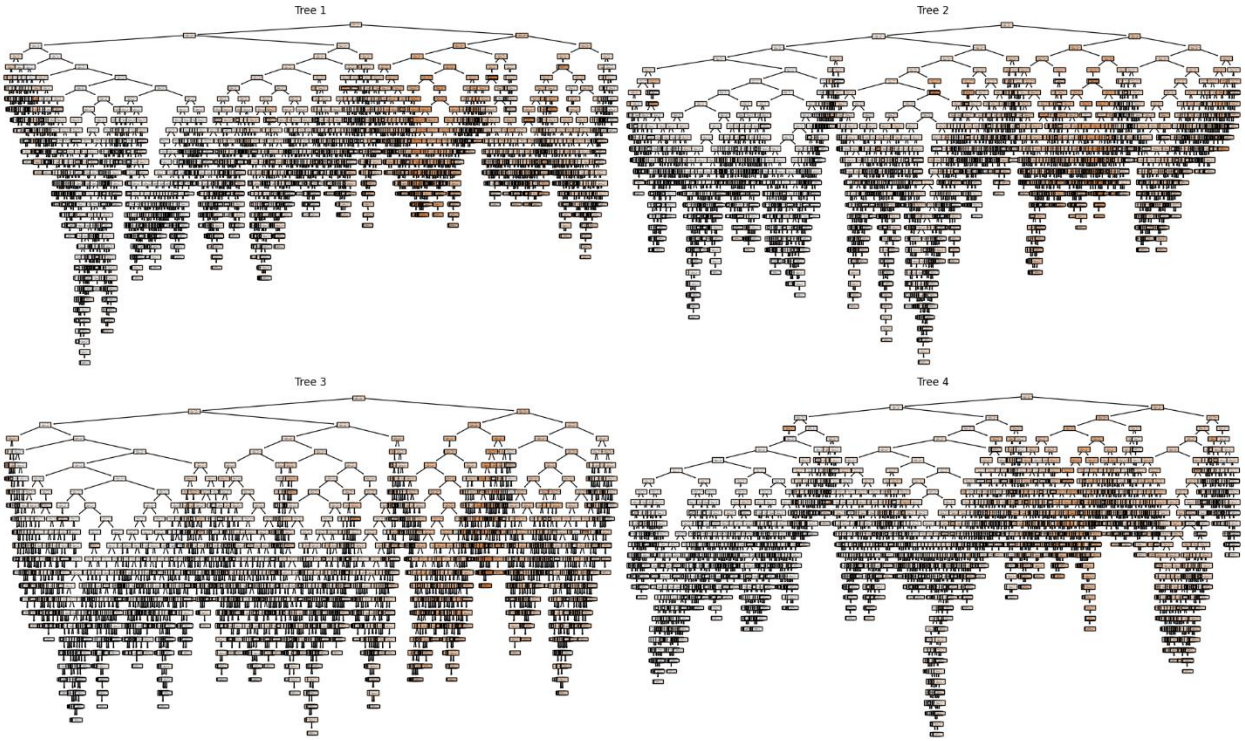
Figura 67 Diagrama de líneas de los valores Reales vs. Predichos utilizando Random Forest Regressor para predecir el CBR95



B. Random Forest Regressor para predecir el CBR100

En la Figura 68 se presenta 4 de 100 árboles del modelo completo de árboles generado mediante la técnica de Random Forest Regressor. Estos árboles extensos y complejos ilustran la profundidad de la minería de datos aplicada. En cada árbol del bosque de árboles se implementan reglas de decisión para predecir la variable objetivo. Donde cada nodo hoja revela una variable de importancia dentro del modelo, acompañada de su regla de decisión correspondiente, lo que brindó una visión detallada de cómo se estructuraron las predicciones del modelo en función de las características o variables de entrada del modelo para predecir el valor del ensayo de CBR100 en cada árbol del bosque de árboles.

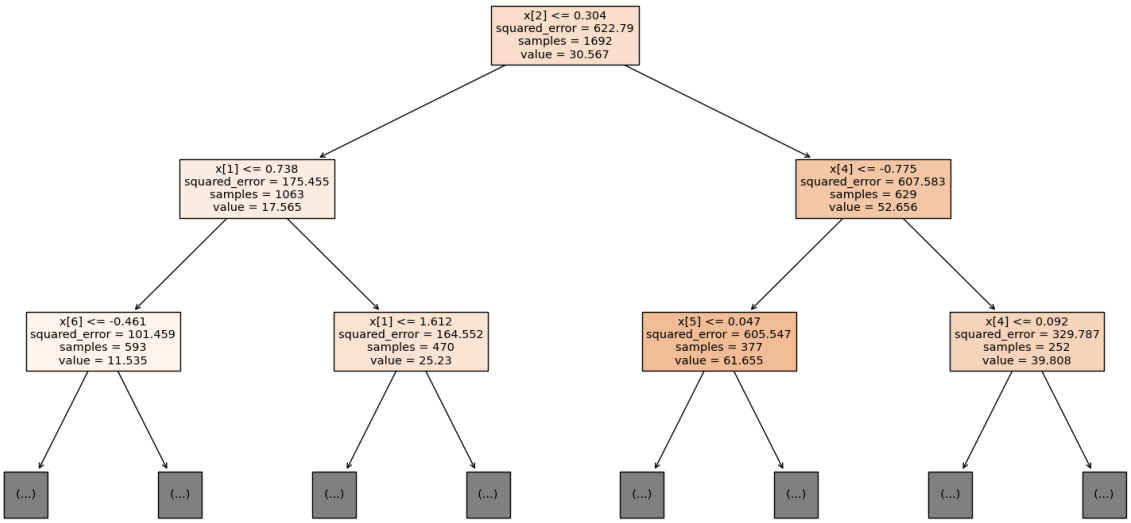
Figura 68 Propiedad Mecánica CBR100: Modelo de árboles utilizando Random Forest Regressor



Para un mejor entendimiento de cada árbol del bosque de árboles, en la Figura 69 se presenta el modelo de un árbol de la técnica Random Forest Regressor, con una profundidad máxima de 2 hojas. En este diagrama, se destaca la variable más representativa en el nodo raíz, que en este caso es la variable 2 (Variable SUCS), con una condición de que debe ser menor o igual a 0.398. Además, se visualizan el error cuadrático, que asciende a 281.911, y el valor correspondiente al CBR100, que alcanza los 20.569.

En las siguientes hojas de este árbol de decisión, se observan nuevas reglas de decisión para las otras variables. Cada regla de decisión sucesiva contribuye a disminuir el error cuadrático y ajustar progresivamente el valor de la variable objetivo, CBR100. Este proceso se repite hasta alcanzar el nodo hoja final, que representa el valor de salida del modelo predictivo en función de las variables de entrada al árbol de decisión.

Figura 69 Propiedad Mecánica CBR100: Modelo de un árbol de profundidad de 2 hojas utilizando Random Forest Regressor



Para generar dichos árboles y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 17 los siguientes hiperparámetros utilizados en la técnica de regresión Random Forest Regressor:

Tabla 17 Hiper parámetros utilizados Random Forest Regressor para predecir el CBR100

Descripción	***Valor***
Bootstrap	False
Ccp_alpha	0.0
Criterion	Squared_error
Max_depth	None
Max_features	1.0
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_samples_leaf	1
Min_samples_split	2
Min_weight_fraction_leaf	0.0
N_stimators	100
N_jobs	None
Oob_score	False
Random_state	42
Verbose	0
Warm_start	False

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 163.0446 (%), un Root Mean Squared Error de 12.7688 (%), Mean Absolute Error de 7.9889 (%) y R-Squared de 0.7575.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

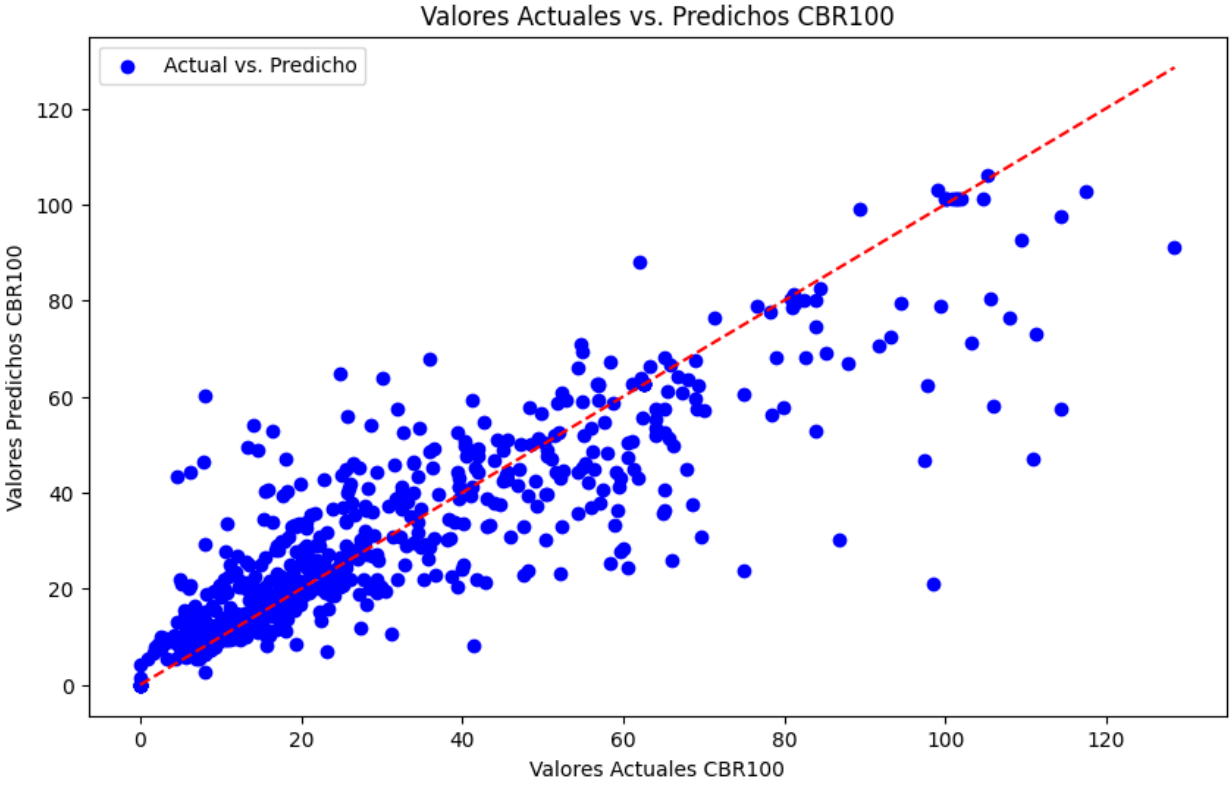
Tabla 18 Hiper parámetros con Grid SearchCV de Random Forest Regressor para predecir el CBR100

Descripción	***Valor***
Max_depth	None
Max_features	sqrt
Min_samples_leaf	1
Min_samples_split	5
N_stimators	300
Random_state	42

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 163.0446 (%), un Root Mean Squared Error de 12.7688 (%), Mean Absolute Error de 7.9889 (%) y R-Squared de 0.7575.

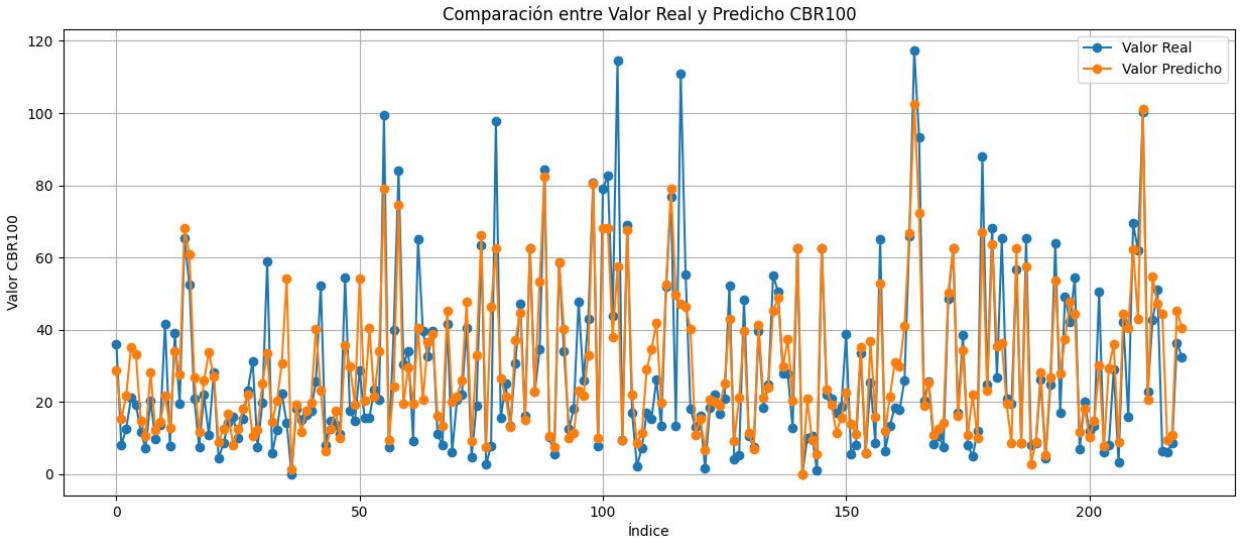
A continuación, en la Figura 70 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el CBR100 con la técnica Random Forest Regressor se ajusta a la línea de regresión de los valores reales y predichos.

Figura 70 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Random Forest Regressor para predecir el CBR100



Finalmente, se muestra el diagrama de líneas en la Figura 71, donde se puede verificar la similitud entre los valores reales y valores predichos del CBR100 para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

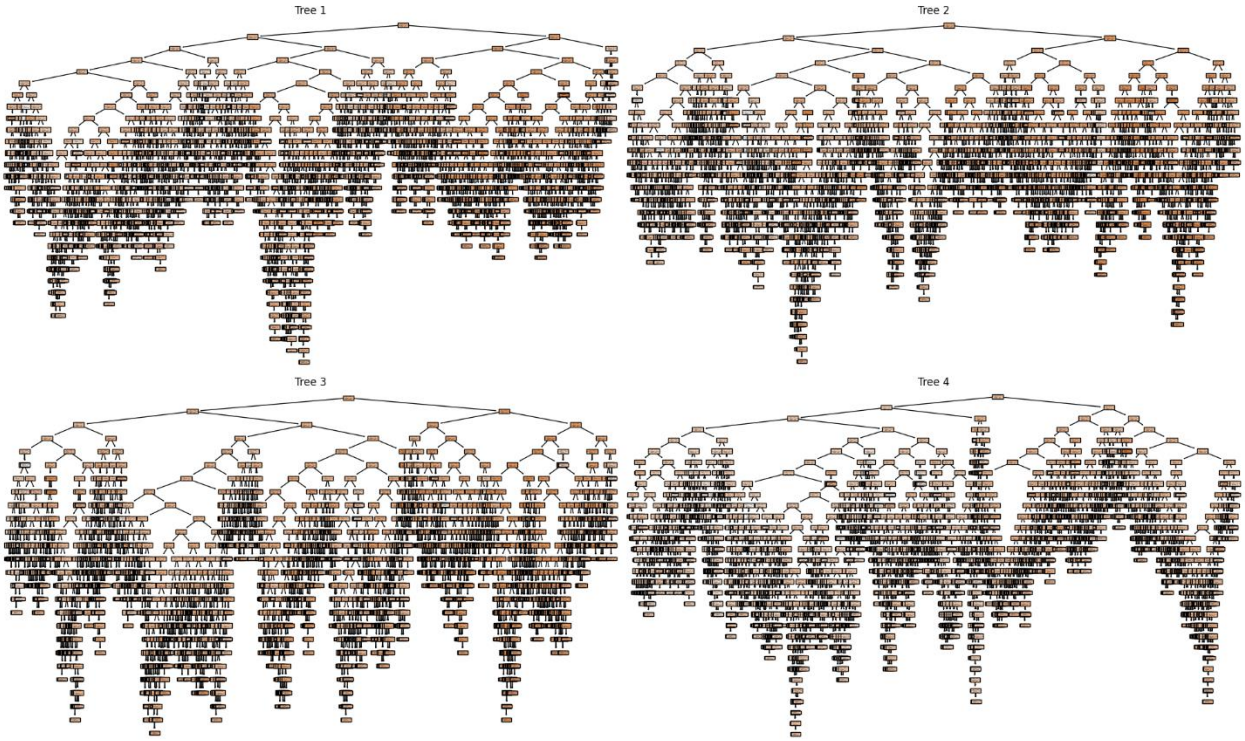
Figura 71 Diagrama de líneas de los valores Reales vs. Predichos utilizando Random Forest Regressor para predecir el CBR100



C. Random Forest Regressor para predecir la variable objetivo MDD

En la Figura 72 se presenta 4 de 100 árboles del modelo completo de árboles generado mediante la técnica de Random Forest Regressor. Estos árboles extensos y complejos ilustran la profundidad de la minería de datos aplicada. En cada árbol del bosque de árboles se implementan reglas de decisión para predecir la variable objetivo. Donde cada nodo hoja revela una variable de importancia dentro del modelo, acompañada de su regla de decisión correspondiente, lo que brindó una visión detallada de cómo se estructuraron las predicciones del modelo en función de las características o variables de entrada del modelo para predecir el valor del ensayo de MDD en cada árbol del bosque de árboles.

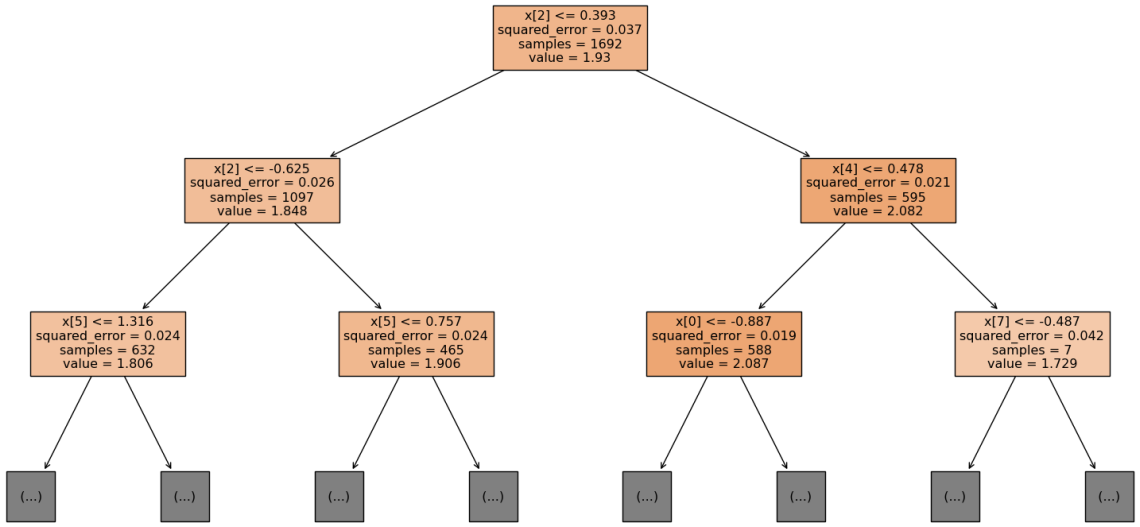
Figura 72 Propiedad Mecánica MDD: Modelo de árboles utilizando Random Forest Regressor



Para un mejor entendimiento de cada árbol del bosque de árboles, en la Figura 73 se presenta el modelo de un árbol de la técnica Random Forest Regressor, con una profundidad máxima de 2 hojas. En este diagrama, se destaca la variable más representativa en el nodo raíz, que en este caso es la variable 2 (Variable SUCS), con una condición de que debe ser menor o igual a 0.398. Además, se visualizan el error cuadrático, que asciende a 281.911, y el valor correspondiente al MDD, que alcanza los 20.569.

En las siguientes hojas de este árbol de decisión, se observan nuevas reglas de decisión para las otras variables. Cada regla de decisión sucesiva contribuye a disminuir el error cuadrático y ajustar progresivamente el valor de la variable objetivo, MDD. Este proceso se repite hasta alcanzar el nodo hoja final, que representa el valor de salida del modelo predictivo en función de las variables de entrada al árbol de decisión.

Figura 73 Propiedad Mecánica MDD: Modelo de un árbol de profundidad de 2 hojas utilizando Random Forest Regressor



Para generar dichos árboles y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 19 los siguientes hiperparámetros utilizados en la técnica de regresión Random Forest Regressor:

Tabla 19 Hiper parámetros utilizados Random Forest Regressor para predecir el MDD

Descripción	***Valor***
Bootstrap	False
Ccp_alpha	0.0
Criterion	Squared_error
Max_depth	None
Max_features	1.0
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_samples_leaf	1
Min_samples_split	2
Min_weight_fraction_leaf	0.0
N_stimators	100
N_jobs	None
Oob_score	False
Random_state	42
Verbose	0
Warm_start	False

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 0.0145 (g/cm3), un Root Mean Squared Error de 0.1204 (g/cm3), Mean Absolute Error de 0.0827 (g/cm3) y R-Squared de 0.6238.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

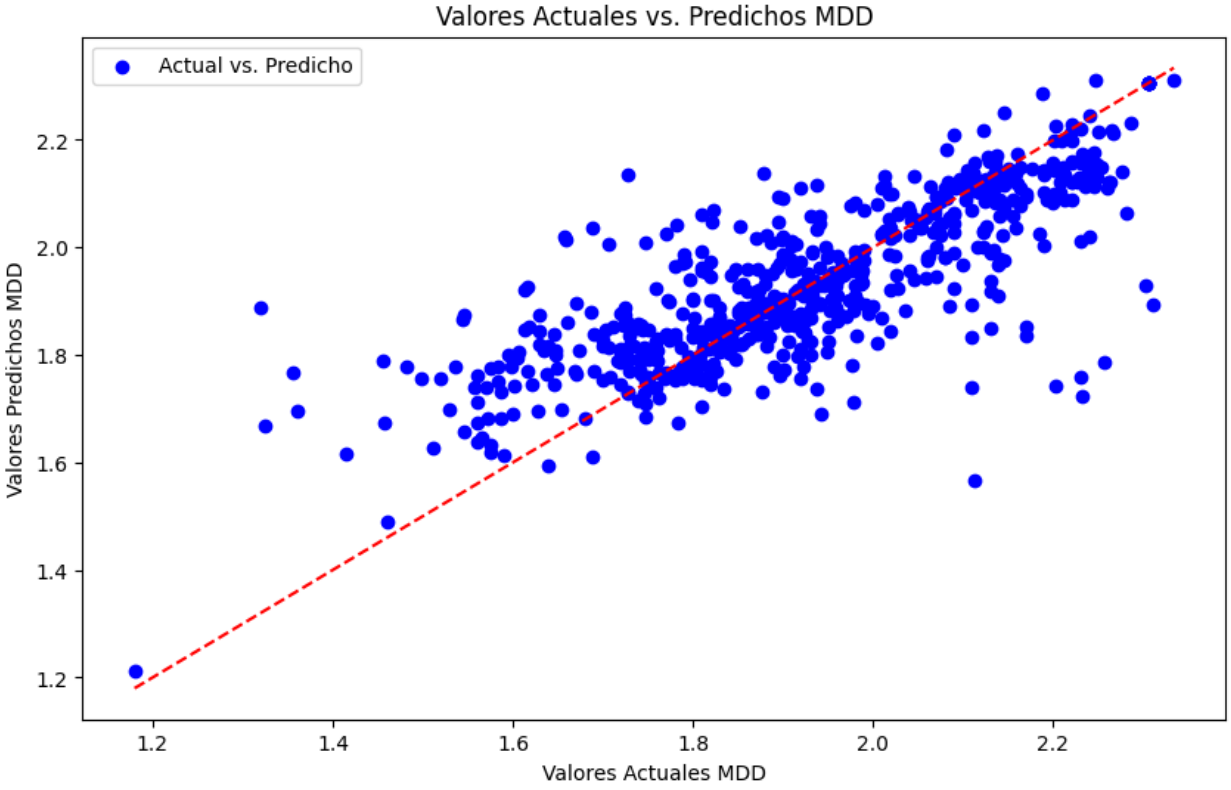
Tabla 20 Hiper parámetros con Grid SearchCV de Random Forest Regressor para predecir el MDD

Descripción	***Valor***
Max_depth	None
Max_features	sqrt
Min_samples_leaf	1
Min_samples_split	5
N_stimators	300
Random_state	42

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 0.0141 (g/cm3), un Root Mean Squared Error de 0.1188 (g/cm3), Mean Absolute Error de 0.0825 (g/cm3) y R-Squared de 0.6335.

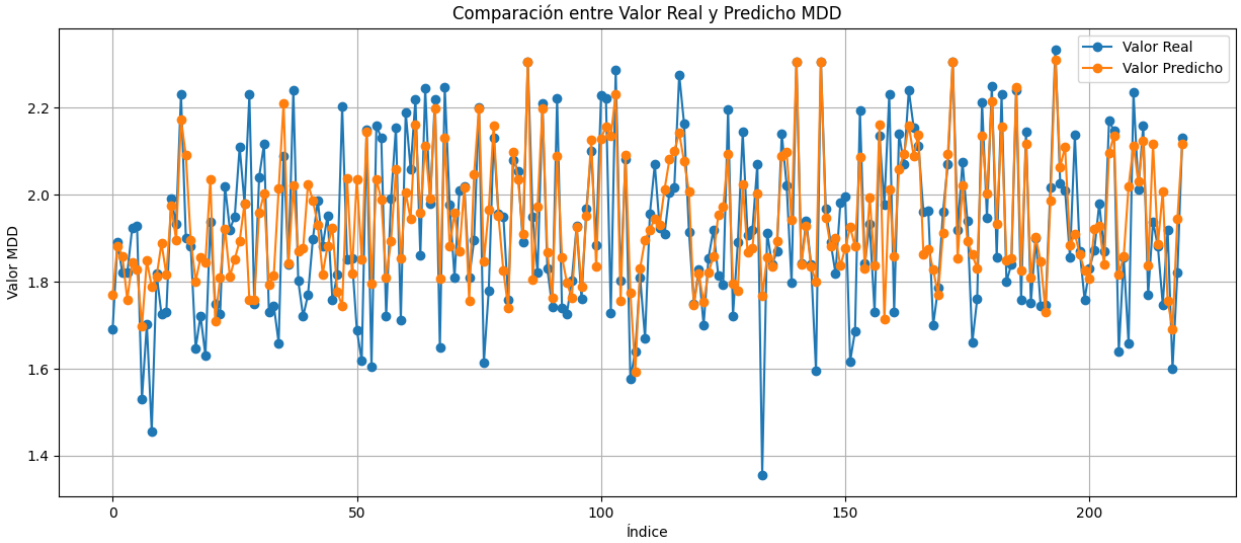
A continuación, en la Figura 74 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el MDD con la técnica Random Forest Regressor se ajusta a la línea de regresión de los valores reales y predichos.

Figura 74 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Random Forest Regressor para predecir el MDD



Finalmente, se muestra el diagrama de líneas en la Figura 75, donde se puede verificar la similitud entre los valores reales y valores predichos del MDD para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

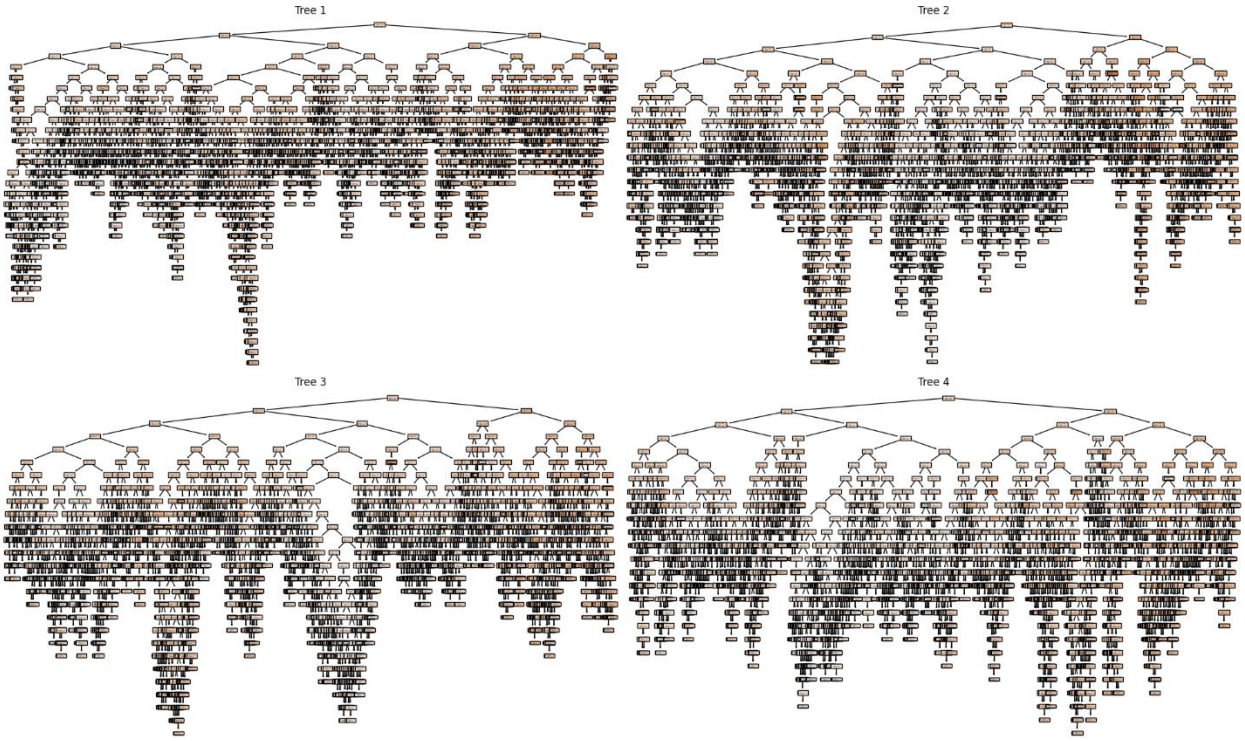
Figura 75 Diagrama de líneas de los valores Reales vs. Predichos utilizando Random Forest Regressor para predecir el MDD



D. Extra Trees Regressor para predecir la variable objetivo OMC

En la Figura 76 se presenta 4 de 100 árboles del modelo completo de árboles generado mediante la técnica de Random Forest Regressor. Estos árboles extensos y complejos ilustran la profundidad de la minería de datos aplicada. En cada árbol del bosque de árboles se implementan reglas de decisión para predecir la variable objetivo. Donde cada nodo hoja revela una variable de importancia dentro del modelo, acompañada de su regla de decisión correspondiente, lo que brindó una visión detallada de cómo se estructuraron las predicciones del modelo en función de las características o variables de entrada del modelo para predecir el valor del ensayo de OMC en cada árbol del bosque de árboles.

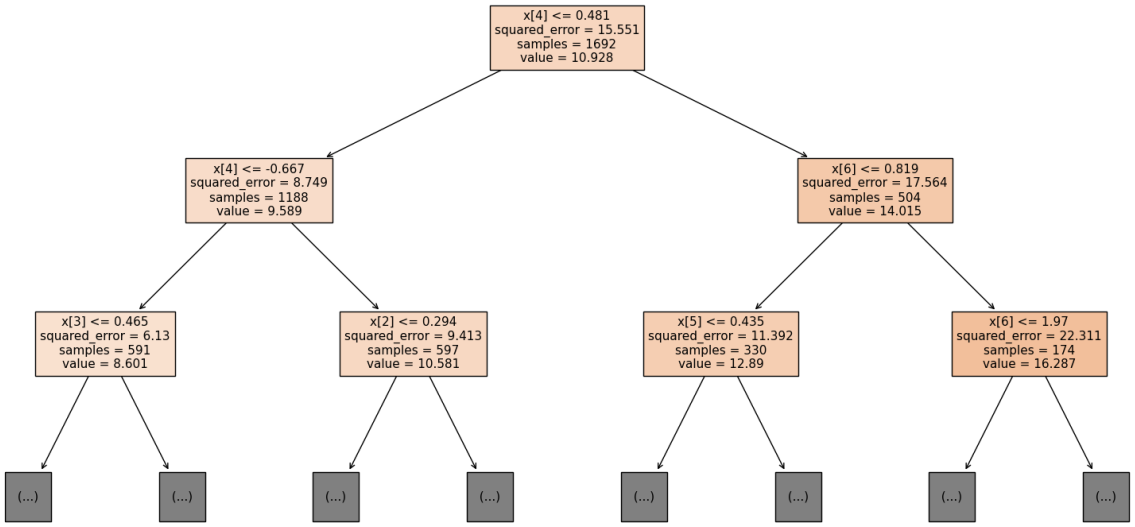
Figura 76 Propiedad Mecánica OMC: Modelo de árboles utilizando Random Forest Regressor



Para un mejor entendimiento de cada árbol del bosque de árboles, en la Figura 77 se presenta el modelo de un árbol de la técnica Random Forest Regressor, con una profundidad máxima de 2 hojas. En este diagrama, se destaca la variable más representativa en el nodo raíz, que en este caso es la variable 2 (Variable SUCS), con una condición de que debe ser menor o igual a 0.398. Además, se visualizan el error cuadrático, que asciende a 281.911, y el valor correspondiente al OMC, que alcanza los 20.569.

En las siguientes hojas de este árbol de decisión, se observan nuevas reglas de decisión para las otras variables. Cada regla de decisión sucesiva contribuye a disminuir el error cuadrático y ajustar progresivamente el valor de la variable objetivo, OMC. Este proceso se repite hasta alcanzar el nodo hoja final, que representa el valor de salida del modelo predictivo en función de las variables de entrada al árbol de decisión.

Figura 77 Propiedad Mecánica OMC: Modelo de un árbol de profundidad de 2 hojas utilizando Random Forest Regressor



Para generar dichos árboles y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 21 los siguientes hiperparámetros utilizados en la técnica de regresión Random Forest Regressor:

Tabla 21 Hiper parámetros utilizados Random Forest Regressor para predecir el OMC

Descripción	***Valor***
Bootstrap	False
Ccp_alpha	0.0
Criterion	Squared_error
Max_depth	None
Max_features	1.0
Max_leaf_nodes	None
Max_samples	None
Min_impurity_decrease	0.0
Min_samples_leaf	1
Min_samples_split	2
Min_weight_fraction_leaf	0.0
N_stimators	100
N_jobs	None
Oob_score	False
Random_state	42
Verbose	0
Warm_start	False

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 8.1513 (%), un Root Mean Squared Error de 2.8550 (%), Mean Absolute Error de 1.8104 (%) y R-Squared de 0.5176.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

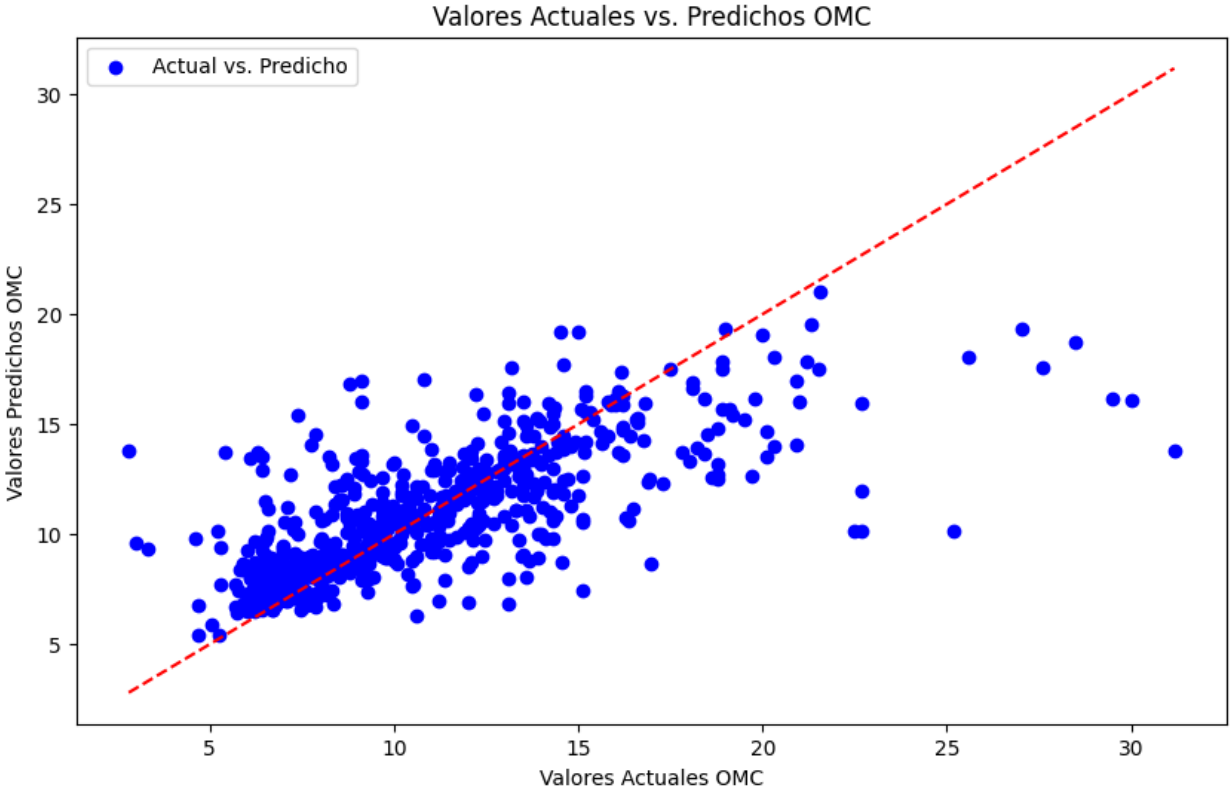
Tabla 22 Hiper parámetros con Grid SearchCV de Random Forest Regressor para predecir el OMC

Descripción	***Valor***
Max_depth	None
Max_features	sqrt
Min_samples_leaf	1
Min_samples_split	5
N_stimators	300
Random_state	42

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 8.0046 (%), un Root Mean Squared Error de 2.8292 (%), Mean Absolute Error de 1.7991 (%) y R-Squared de 0.5262.

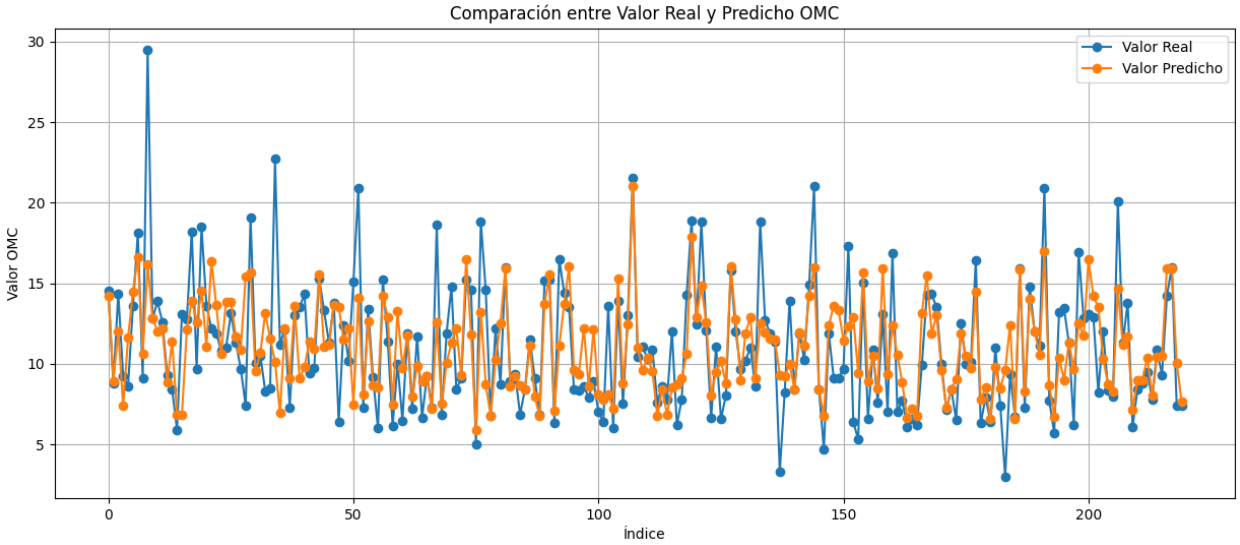
A continuación, en la Figura 78 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el OMC con la técnica Random Forest Regressor se ajusta a la línea de regresión de los valores reales y predichos.

Figura 78 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Random Forest Regressor para predecir el OMC



Finalmente, se muestra el diagrama de líneas en la Figura 79, donde se puede verificar la similitud entre los valores reales y valores predichos del OMC para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

Figura 79 Diagrama de líneas de los valores Reales vs. Predichos utilizando Random Forest Regressor para predecir el OMC



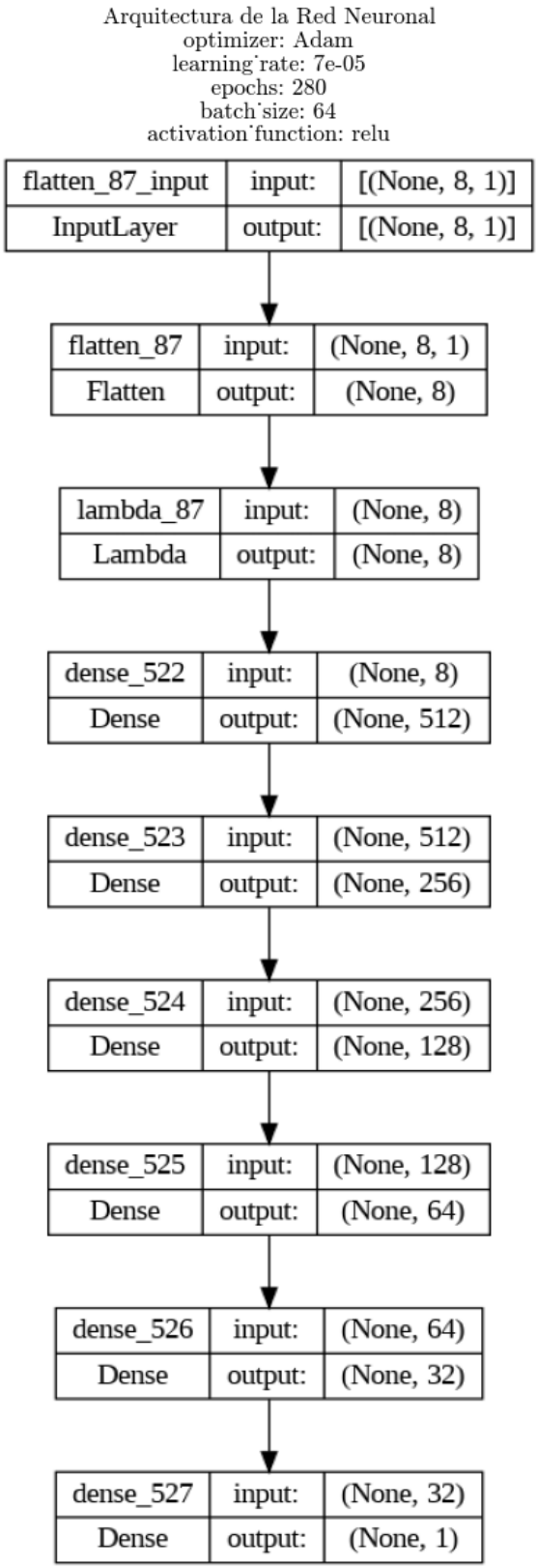
4.5.3.3. Técnica Redes Neuronales Artificiales (BPNN)

La técnica Redes Neuronales Artificiales (BPNN) fue aplicada a los diferentes conjuntos de datos de entrenamiento y prueba con el objetivo de modelar y predecir la variable objetivo. El modelo Redes Neuronales Artificiales se ajustó al conjunto de datos de las variables objetivos de CBR95, CBR100, MDD y OMC utilizando los diferentes parámetros e hiperparametros de cada algoritmo o técnica de Machine Learning.

A. Redes Neuronales Artificiales para predecir la variable objetivo CBR95

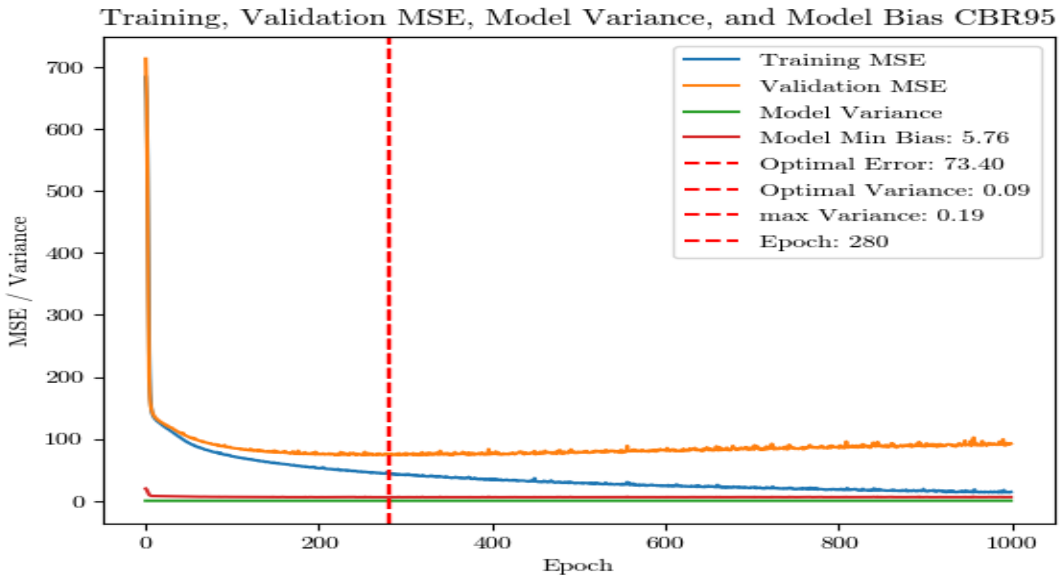
En la Figura 80 se presenta el modelo de la arquitectura de las redes neuronales generado mediante la técnica de Redes Neuronales Artificiales (BPNN). Donde, se puede observar, en primera instancia, los hiperparametros utilizados y sus respectivos valores. Valores como el optimizador que en este caso se utilizó el optimizador Adam, Taza de aprendizaje que fue de 0.00007, el número de épocas que fue de 280, el tamaño de lote que fue de 64, la función de activación que fue relu. Así como también, se puede observar gráficamente la representación de cada capa de dicha red neuronal en cada recuadro con sus respectivos valores de entrada, salida y el tipo de red neuronal utilizados.

Figura 80 Propiedad Mecánica CBR95: Hiperparámetros y Arquitectura del modelo utilizando Redes Neuronales Artificiales



En el siguiente gráfico la Figura 81 se muestra el diagrama al entrenar el modelo de redes neuronales artificiales en relación con los valores del MSE de entrenamiento y validación, varianza del modelo, bias(sesgo) del modelo y las épocas de entrenamiento. Se puede observar que el MSE de validación y el MSE de entrenamiento disminuyen a medida que aumenta el número de épocas. Lo que evidencia que el modelo está aprendiendo y mejorando su rendimiento a lo largo del tiempo. La varianza del modelo se mantiene relativamente constante y aumentando mínimamente durante todo el proceso de entrenamiento, por otro lado, el bias del modelo se mantiene relativamente constante y disminuyendo mínimamente durante todo el proceso de entrenamiento. Así como también, se muestra el error óptimo MSE (73.40) y la varianza óptima (0.09) para un tiempo o época de 280. Esto significa que el modelo logra su optimo rendimiento cuando el error alcanza el mínimo valor, se disminuye el bias y se aumenta la varianza del modelo hasta lograr un valor en equilibrio con relación a estos 3 valores. En este caso, se utilizó el MSE de validación mínimo, ya que como se puede visualizar el MSE de entrenamiento continúa disminuyendo mínimamente hasta lograr el equilibrio del error mínimo con la varianza y bias del modelo para una época significativamente grande, que por cuestiones de eficiencia no se consideró dicho valor, sino más bien el valor de las épocas de entrenamiento para el MSE mínimo de la validación. Lo que permitirá al modelo predecir valores que no se vieron en el set de datos.

Figura 81 Propiedad Mecánica CBR95: Diagrama Bias-Variance Tradeoff utilizando Redes Neuronales Artificiales



Para los anteriores hiperparámetros, arquitectura de red neuronal y después del análisis del diagrama Bias-Variance Tradeoff se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 81.3957 (%), un Root Mean Squared Error de 9.0219 (%), Mean Absolute Error de 5.7619 (%) y R-Squared de 0.7108.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

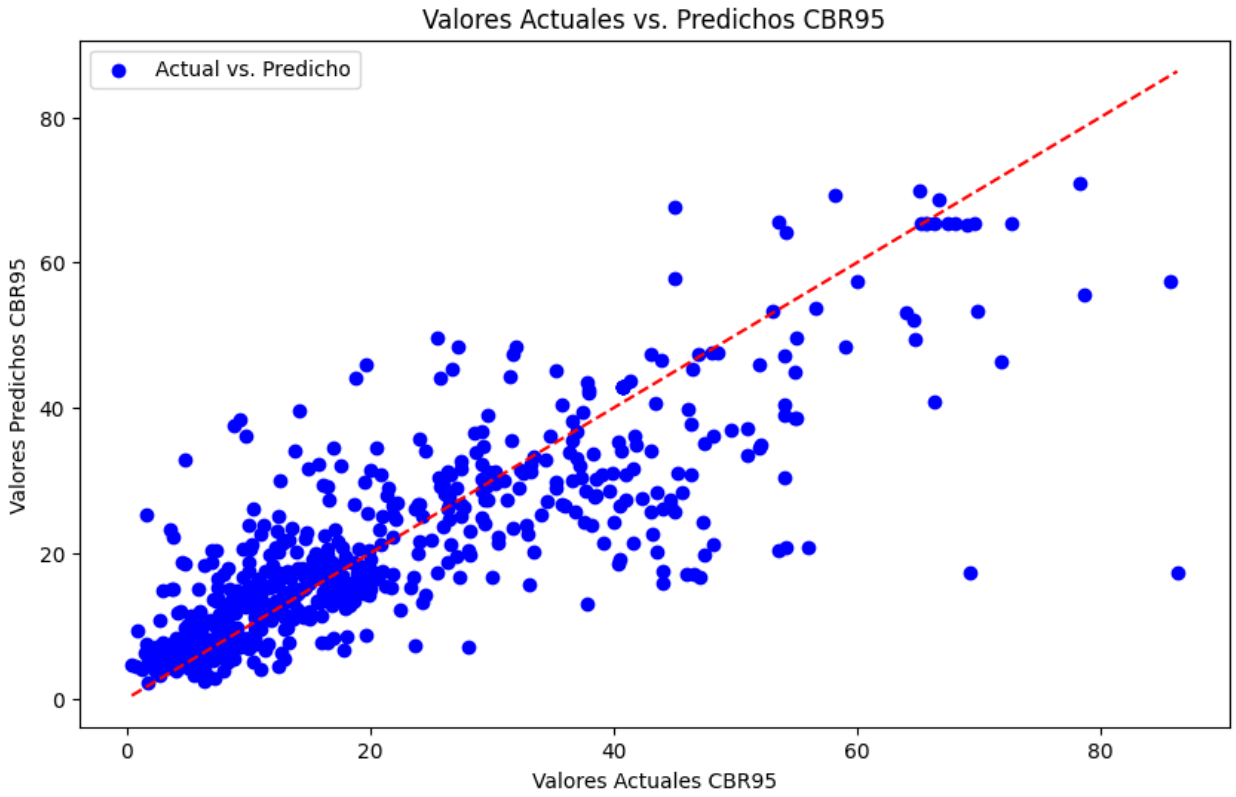
Tabla 23 Hiperparámetros con Grid SearchCV de Redes Neuronales Artificiales para predecir el CBR95

Descripción	***Valor***
validation_split	0.3
optimizer	adam
metrics	mae
loss	mean_squared_error
epochs	200
batch_size	32

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 98.0474 (%), un Root Mean Squared Error de 9.9018 (%), Mean Absolute Error de 6.1997 (%) y R-Squared de 0.6516.

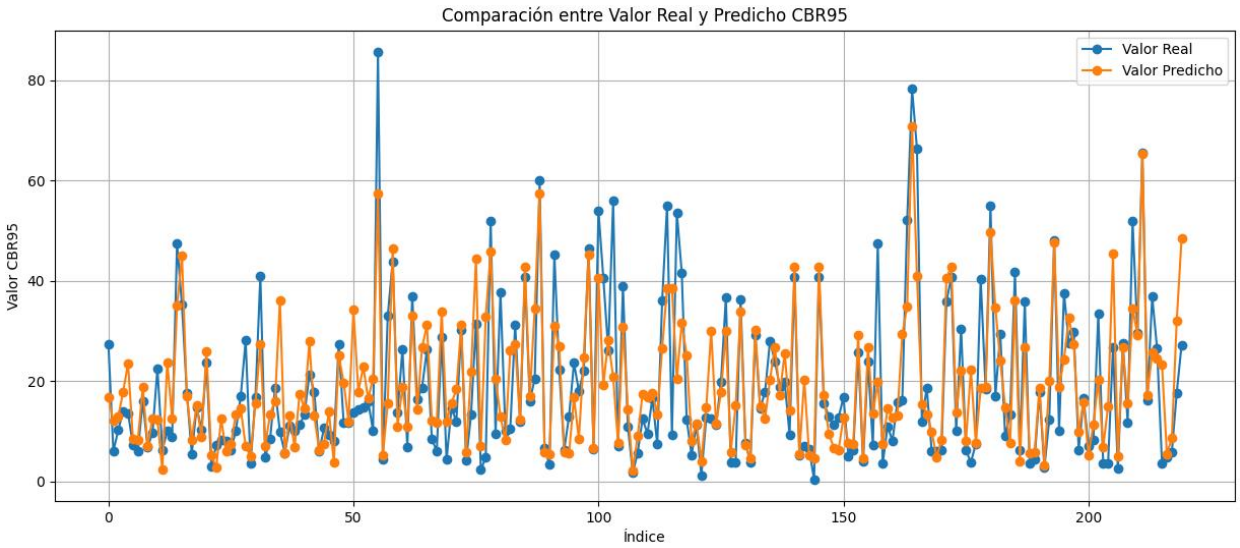
A continuación, en la Figura 82 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el CBR95 con la técnica Redes Neuronales Artificiales se ajusta a la línea de regresión de los valores reales y predichos.

Figura 82 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Redes Neuronales Artificiales para predecir el CBR95



Finalmente, se muestra el diagrama de líneas en la Figura 83, donde se puede verificar la similitud entre los valores reales y valores predichos del CBR95 para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

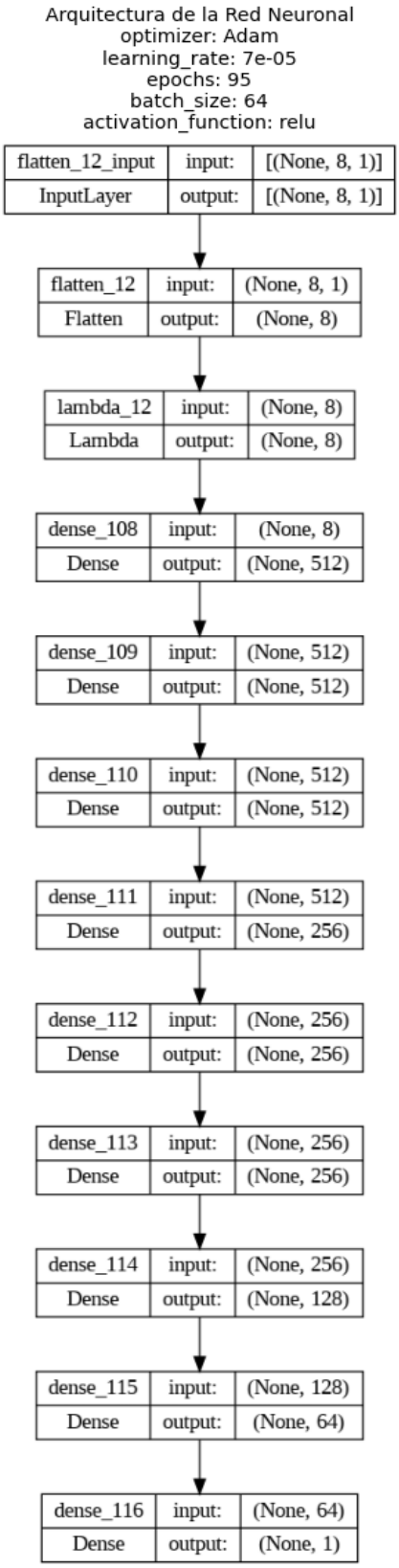
Figura 83 Diagrama de líneas de los valores Reales vs. Predichos utilizando Redes Neuronales Artificiales para predecir el CBR95



B. Redes Neuronales Artificiales para predecir la variable objetivo CBR100

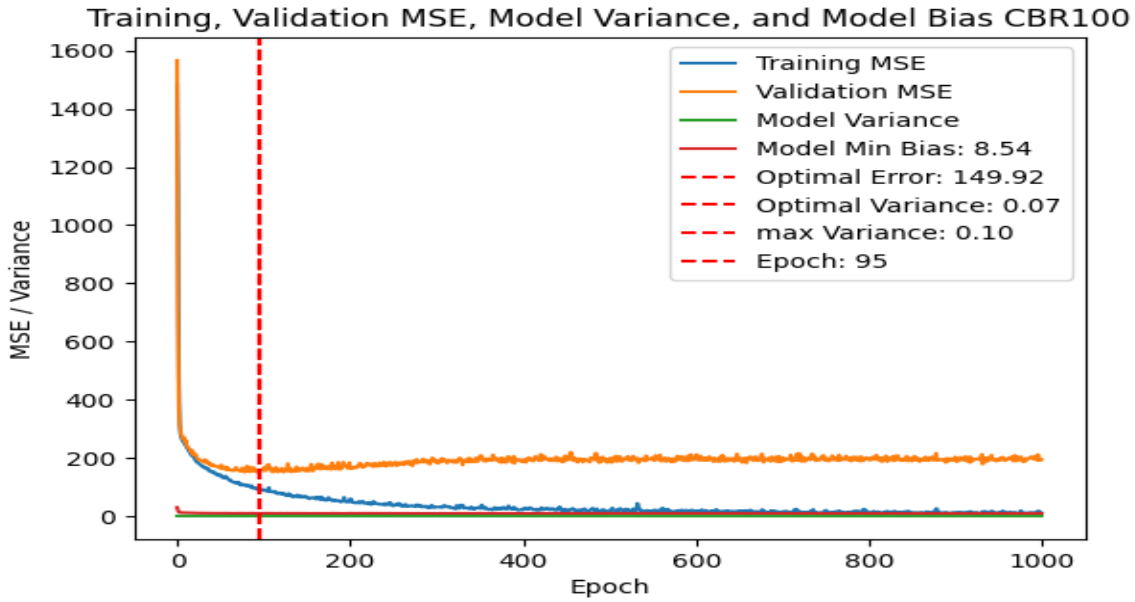
En la Figura 84 se presenta el modelo de la arquitectura de las redes neuronales generado mediante la técnica de Redes Neuronales Artificiales (BPNN). Donde, se puede observar, en primera instancia, los hiperparametros utilizados y sus respectivos valores. Valores como el optimizador que en este caso se utilizó el optimizador Adam, Taza de aprendizaje que fue de 0.00007, el número de épocas que fue de 95, el tamaño de lote que fue de 64, la función de activación que fue relu. Así como también, se puede observar gráficamente la representación de cada capa de dicha red neuronal en cada recuadro con sus respectivos valores de entrada, salida y el tipo de red neuronal utilizados.

Figura 84 Propiedad Mecánica CBR100: Hiperparámetros y Arquitectura del modelo utilizando Redes Neuronales Artificiales



En el siguiente gráfico la Figura 85 se muestra el diagrama al entrenar el modelo de redes neuronales artificiales en relación con los valores del MSE de entrenamiento y validación, varianza del modelo, bias(sesgo) del modelo y las épocas de entrenamiento. Se puede observar que el MSE de validación y el MSE de entrenamiento disminuyen a medida que aumenta el número de épocas. Lo que evidencia que el modelo está aprendiendo y mejorando su rendimiento a lo largo del tiempo. La varianza del modelo se mantiene relativamente constante y aumentando mínimamente durante todo el proceso de entrenamiento, por otro lado, el bias del modelo se mantiene relativamente constante y disminuyendo mínimamente durante todo el proceso de entrenamiento. Así como también, se muestra el error óptimo MSE (149.92) y la varianza óptima (0.07) para un tiempo o época de 95. Esto significa que el modelo logra su optimo rendimiento cuando el error alcanza el mínimo valor, se disminuye el bias y se aumenta la varianza del modelo hasta lograr un valor en equilibrio con relación a estos 3 valores. En este caso, se utilizó el MSE de validación mínimo, ya que como se puede visualizar el MSE de entrenamiento continúa disminuyendo mínimamente hasta lograr el equilibrio del error mínimo con la varianza y bias del modelo para una época significativamente grande, que por cuestiones de eficiencia no se consideró dicho valor, sino más bien el valor de las épocas de entrenamiento para el MSE mínimo de la validación. Lo que permitirá al modelo predecir valores que no se vieron en el set de datos.

Figura 85 Propiedad Mecánica CBR100: Diagrama Bias-Variance Tradeoff utilizando Redes Neuronales Artificiales



Para los anteriores hiperparámetros, arquitectura de red neuronal y después del análisis del diagrama Bias-Variance Tradeoff se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 206.4105 (%), un Root Mean Squared Error de 14.3669 (%), Mean Absolute Error de 9.0309 (%) y R-Squared de 0.6930.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

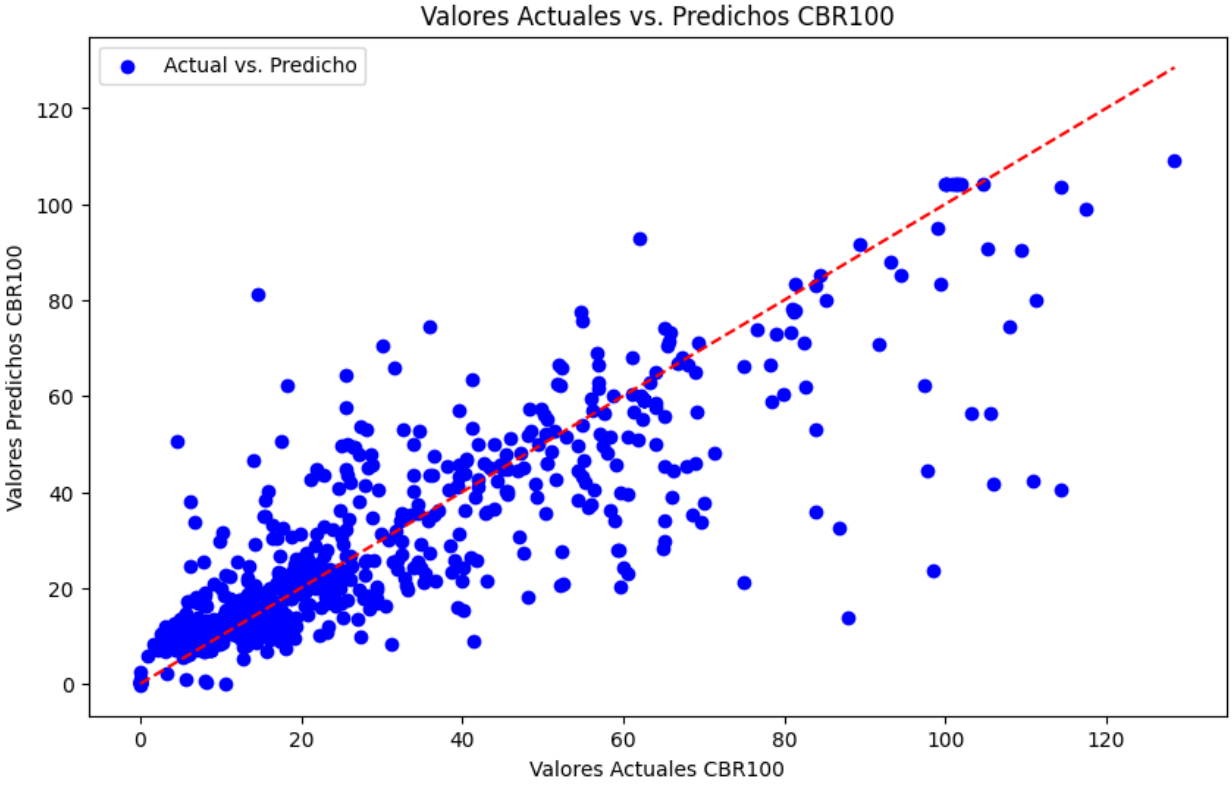
Tabla 24 Hiper parámetros con Grid SearchCV de Redes Neuronales Artificiales para predecir el CBR100

Descripción	***Valor***
validation_split	0.2
optimizer	sgd
metrics	mse
loss	mean_squared_error
epochs	100
batch_size	32

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 225.7351 (%), un Root Mean Squared Error de 15.0244 (%), Mean Absolute Error de 9.4575 (%) y R-Squared de 0.6643.

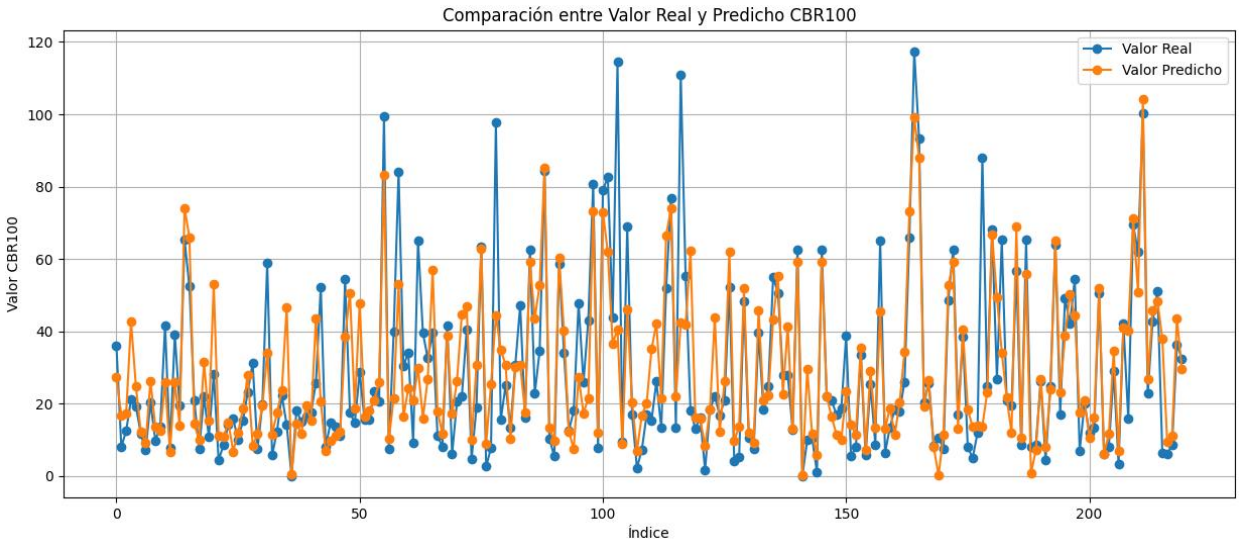
A continuación, en la Figura 86 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el CBR100 con la técnica Redes Neuronales Artificiales se ajusta a la línea de regresión de los valores reales y predichos.

Figura 86 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Redes Neuronales Artificiales para predecir el CBR100



Finalmente, se muestra el diagrama de líneas en la Figura 87, donde se puede verificar la similitud entre los valores reales y valores predichos del CBR100 para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

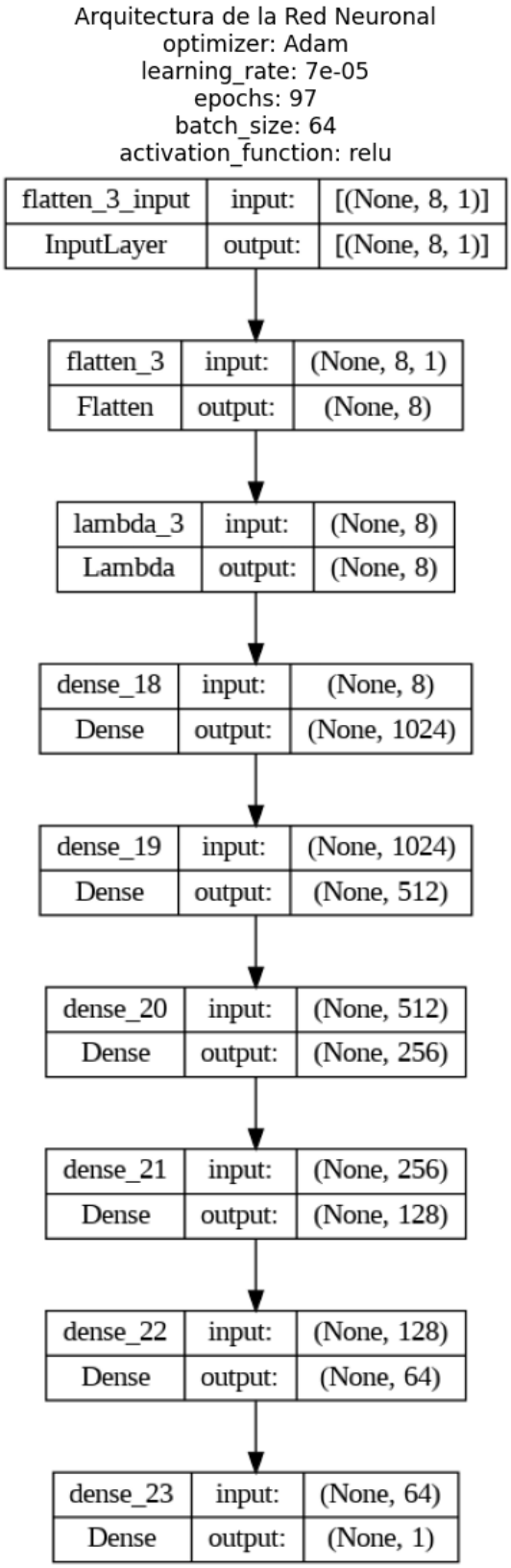
Figura 87 Diagrama de líneas de los valores Reales vs. Predichos utilizando Redes Neuronales Artificiales para predecir el CBR100



C. Redes Neuronales Artificiales para predecir la variable objetivo MDD

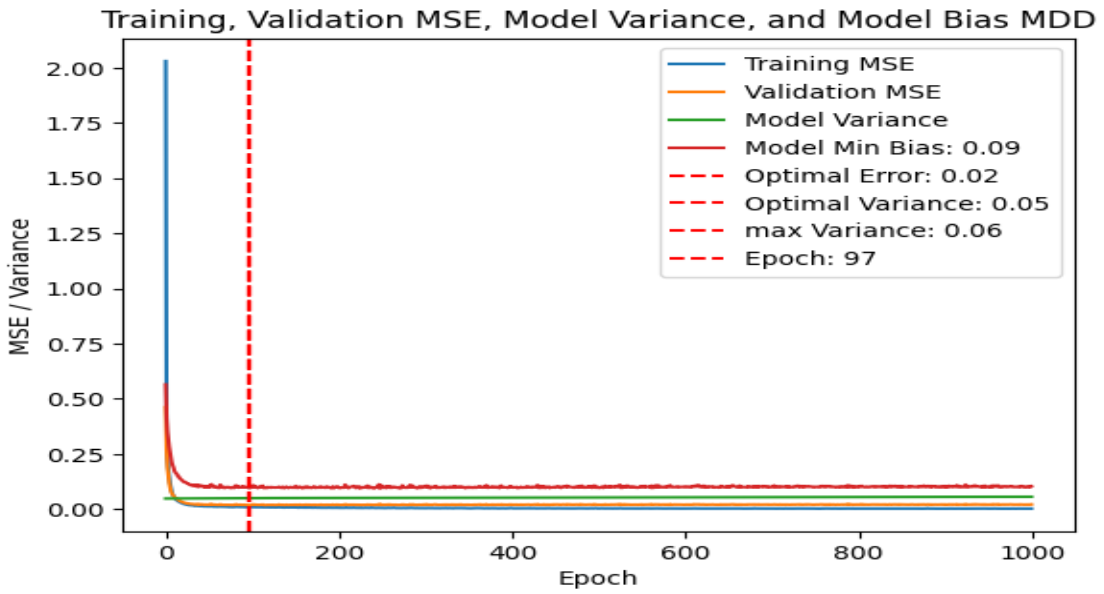
En la Figura 88 se presenta el modelo de la arquitectura de las redes neuronales generado mediante la técnica de Redes Neuronales Artificiales (BPNN). Donde, se puede observar, en primera instancia, los hiperparametros utilizados y sus respectivos valores. Valores como el optimizador que en este caso se utilizó el optimizador Adam, Taza de aprendizaje que fue de 0.00007, el número de épocas que fue de 97, el tamaño de lote que fue de 64, la función de activación que fue relu. Así como también, se puede observar gráficamente la representación de cada capa de dicha red neuronal en cada recuadro con sus respectivos valores de entrada, salida y el tipo de red neuronal utilizados.

Figura 88 Propiedad Mecánica MDD: Hiperparámetros y Arquitectura del modelo utilizando Redes Neuronales Artificiales



En el siguiente gráfico la Figura 89 se muestra el diagrama al entrenar el modelo de redes neuronales artificiales en relación con los valores del MSE de entrenamiento y validación, varianza del modelo, bias(sesgo) del modelo y las épocas de entrenamiento. Se puede observar que el MSE de validación y el MSE de entrenamiento disminuyen a medida que aumenta el número de épocas. Lo que evidencia que el modelo está aprendiendo y mejorando su rendimiento a lo largo del tiempo. La varianza del modelo se mantiene relativamente constante y aumentando mínimamente durante todo el proceso de entrenamiento, por otro lado, el bias del modelo se mantiene relativamente constante y disminuyendo mínimamente durante todo el proceso de entrenamiento. Así como también, se muestra el error óptimo MSE (0.02) y la varianza óptima (0.05) para un tiempo o época de 97. Esto significa que el modelo logra su optimo rendimiento cuando el error alcanza el mínimo valor, se disminuye el bias y se aumenta la varianza del modelo hasta lograr un valor en equilibrio con relación a estos 3 valores. En este caso, se utilizó el MSE de validación mínimo, ya que como se puede visualizar el MSE de entrenamiento continúa disminuyendo mínimamente hasta lograr el equilibrio del error mínimo con la varianza y bias del modelo para una época significativamente grande, que por cuestiones de eficiencia no se consideró dicho valor, sino más bien el valor de las épocas de entrenamiento para el MSE mínimo de la validación. Lo que permitirá al modelo predecir valores que no se vieron en el set de datos.

Figura 89 Propiedad Mecánica MDD: Diagrama Bias-Variance Tradeoff utilizando Redes Neuronales Artificiales



Para los anteriores hiperparámetros, arquitectura de red neuronal y después del análisis del diagrama Bias-Variance Tradeoff se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 0.0185 (g/cm3), un Root Mean Squared Error de 0.1361 (g/cm3), Mean Absolute Error de 0.0984 (g/cm3) y R-Squared de 0.5194.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

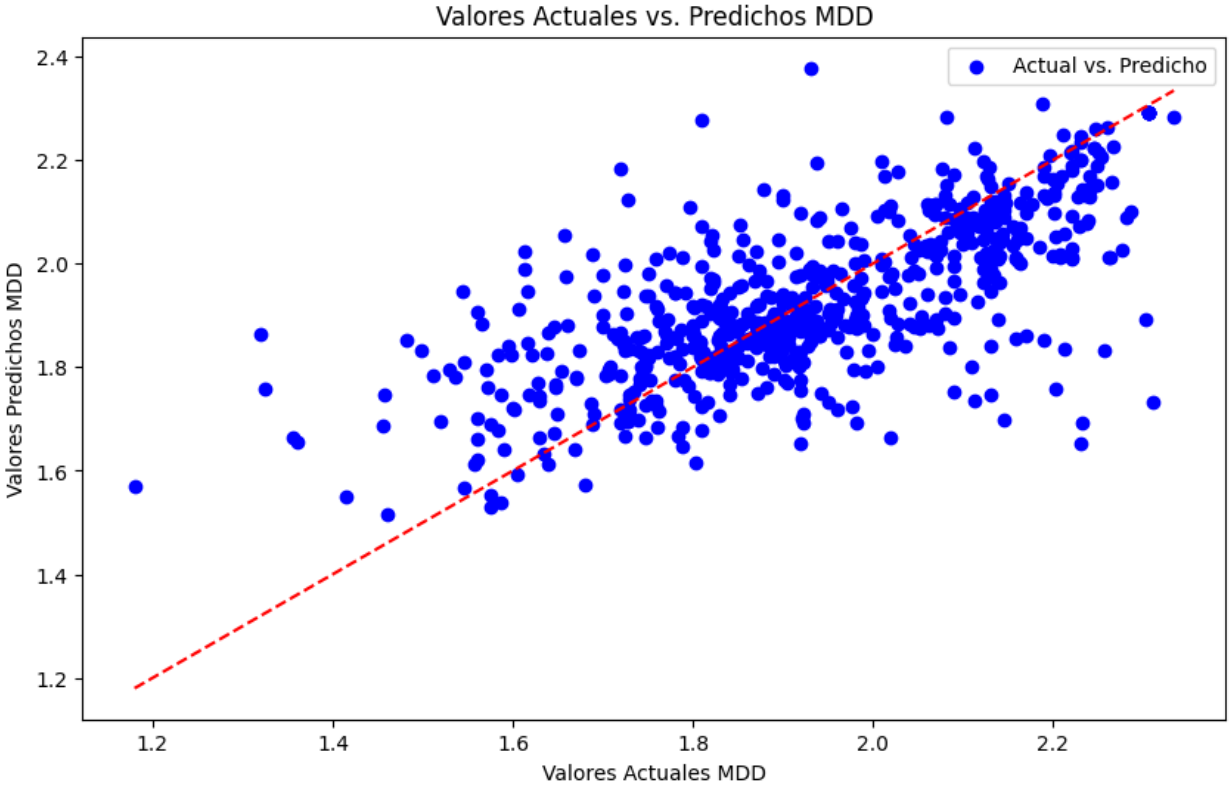
Tabla 25 Hiperparámetros con Grid SearchCV de Redes Neuronales Artificiales para predecir el MDD

Descripción	***Valor***
validation_split	0.1
optimizer	sgd
metrics	mse
loss	mean_squared_error
epochs	50
batch_size	128

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 0.0200 (g/cm3), un Root Mean Squared Error de 0.1414 (g/cm3), Mean Absolute Error de 0.1023 (g/cm3) y R-Squared de 0.4801.

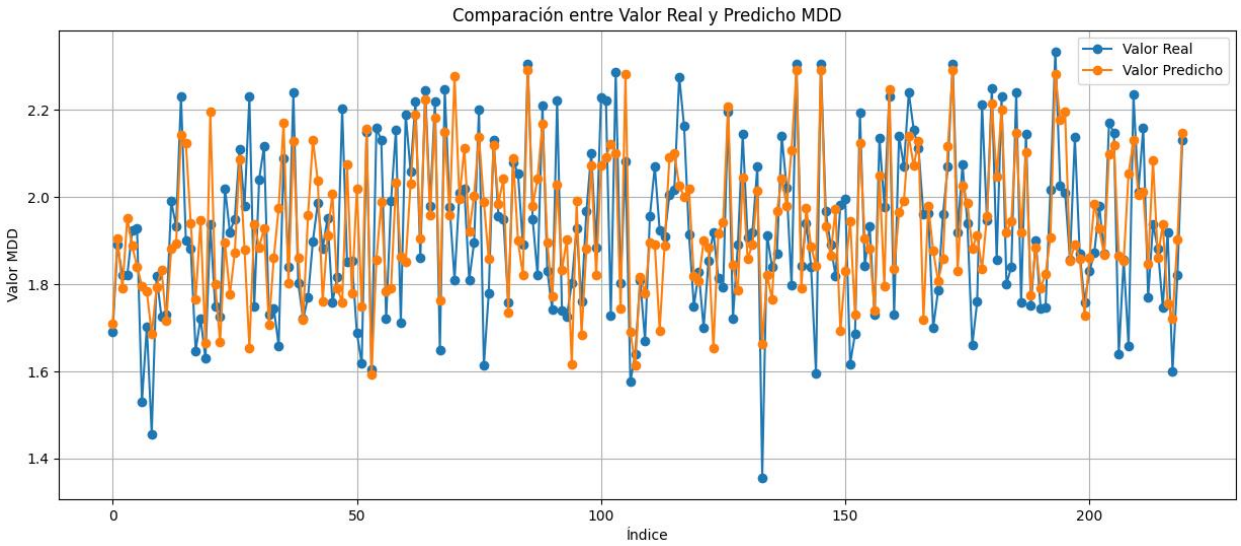
A continuación, en la Figura 90 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el MDD con la técnica Redes Neuronales Artificiales se ajusta a la línea de regresión de los valores reales y predichos.

Figura 90 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Redes Neuronales Artificiales para predecir el MDD



Finalmente, se muestra el diagrama de líneas en la Figura 91, donde se puede verificar la similitud entre los valores reales y valores predichos del MDD para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

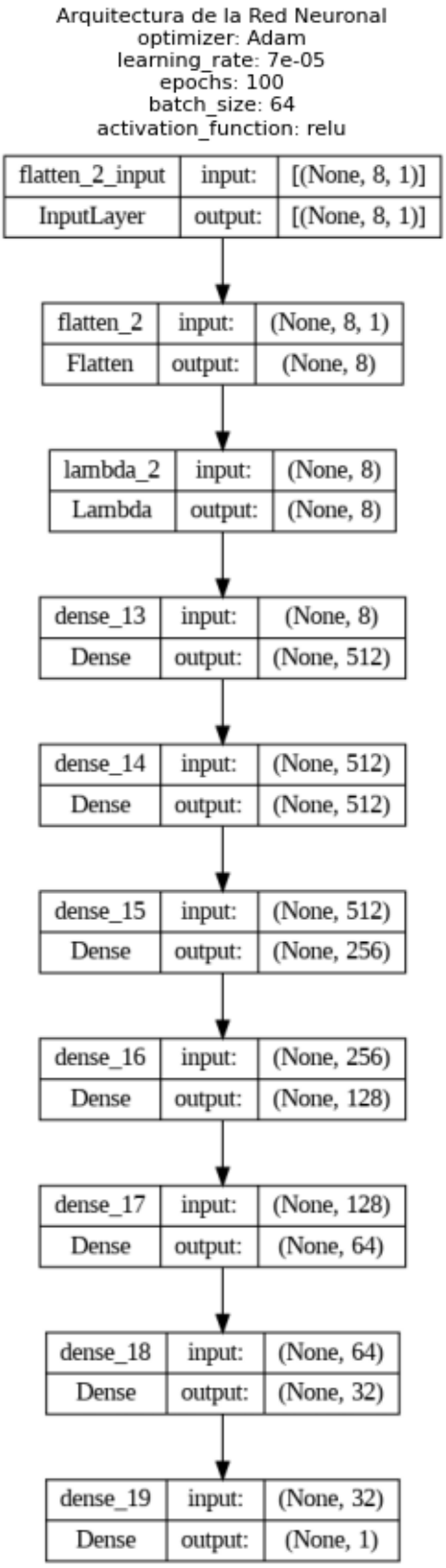
Figura 91 Diagrama de líneas de los valores Reales vs. Predichos utilizando Redes Neuronales Artificiales para predecir el MDD



D. Redes Neuronales Artificiales para predecir la variable objetivo OMC

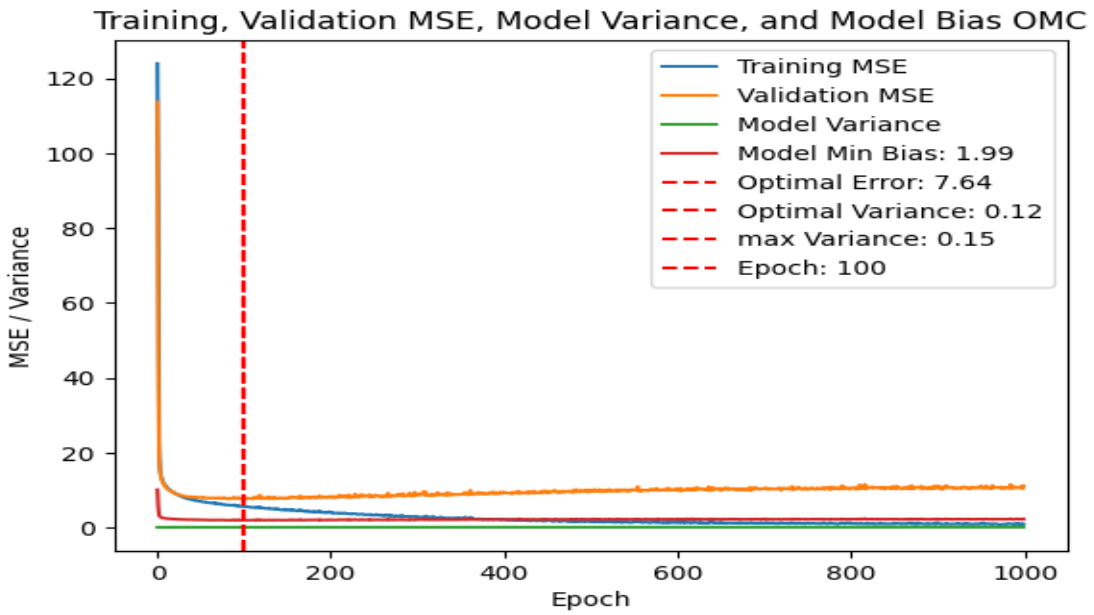
En la Figura 92 se presenta el modelo de la arquitectura de las redes neuronales generado mediante la técnica de Redes Neuronales Artificiales (BPNN). Donde, se puede observar, en primera instancia, los hiperparametros utilizados y sus respectivos valores. Valores como el optimizador que en este caso se utilizó el optimizador Adam, Taza de aprendizaje que fue de 0.00007, el número de épocas que fue de 100, el tamaño de lote que fue de 64, la función de activación que fue relu. Así como también, se puede observar gráficamente la representación de cada capa de dicha red neuronal en cada recuadro con sus respectivos valores de entrada, salida y el tipo de red neuronal utilizados.

Figura 92 Propiedad Mecánica OMC: Hiperparámetros y Arquitectura del modelo utilizando Redes Neuronales Artificiales



En el siguiente gráfico la Figura 93 se muestra el diagrama al entrenar el modelo de redes neuronales artificiales en relación con los valores del MSE de entrenamiento y validación, varianza del modelo, bias(sesgo) del modelo y las épocas de entrenamiento. Se puede observar que el MSE de validación y el MSE de entrenamiento disminuyen a medida que aumenta el número de épocas. Lo que evidencia que el modelo está aprendiendo y mejorando su rendimiento a lo largo del tiempo. La varianza del modelo se mantiene relativamente constante y aumentando mínimamente durante todo el proceso de entrenamiento, por otro lado, el bias del modelo se mantiene relativamente constante y disminuyendo mínimamente durante todo el proceso de entrenamiento. Así como también, se muestra el error óptimo MSE (7.64) y la varianza óptima (0.12) para un tiempo o época de 100. Esto significa que el modelo logra su optimo rendimiento cuando el error alcanza el mínimo valor, se disminuye el bias y se aumenta la varianza del modelo hasta lograr un valor en equilibrio con relación a estos 3 valores. En este caso, se utilizó el MSE de validación mínimo, ya que como se puede visualizar el MSE de entrenamiento continúa disminuyendo mínimamente hasta lograr el equilibrio del error mínimo con la varianza y bias del modelo para una época significativamente grande, que por cuestiones de eficiencia no se consideró dicho valor, sino más bien el valor de las épocas de entrenamiento para el MSE mínimo de la validación. Lo que permitirá al modelo predecir valores que no se vieron en el set de datos.

Figura 93 Propiedad Mecánica OMC: Diagrama Bias-Variance Tradeoff utilizando Redes Neuronales Artificiales



Para los anteriores hiperparámetros, arquitectura de red neuronal y después del análisis del diagrama Bias-Variance Tradeoff se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 9.1967 (%), un Root Mean Squared Error de 3.0326 (%), Mean Absolute Error de 2.0249 (%) y R-Squared de 0.4557.

Así como también, se utilizó la herramienta Grid SerarchCV para mejorar y optimizar la eficiencia de dicho modelo. En algunos casos, ajustar los hiperparámetros del modelo puede mejorar su eficiencia sin sacrificar el rendimiento. Esto implica encontrar la configuración óptima de los hiperparámetros que permita al modelo lograr resultados satisfactorios con menos recursos. Para cuestiones de replicabilidad de dicho modelo la herramienta Grid SerarchCV arrojo los siguientes Hiper parámetros:

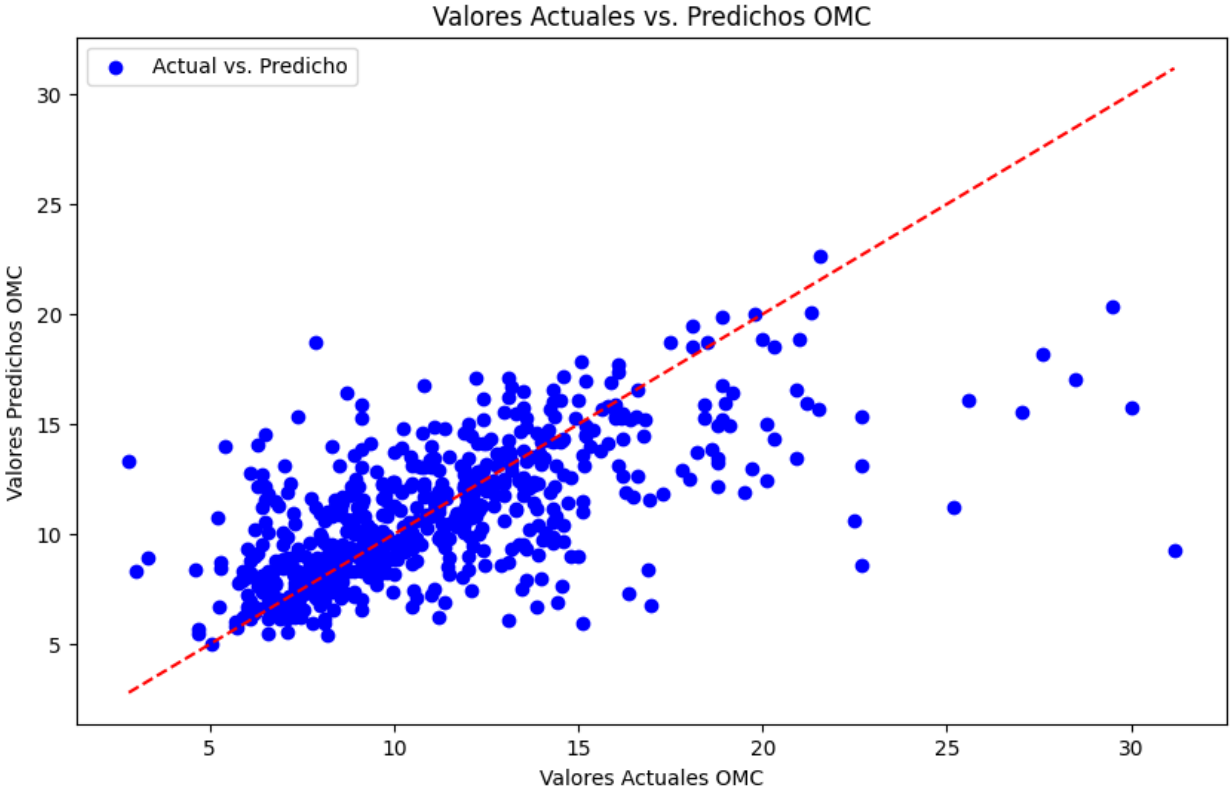
Tabla 26 Hiperparámetros con Grid SearchCV de Redes Neuronales Artificiales para predecir el OMC

Descripción	***Valor***
validation_split	0.3
optimizer	adam
metrics	mse
loss	mean_squared_error
epochs	200
batch_size	128

Después del ajuste de hiperparámetros con la herramienta Grid SerarchCV se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 12.0976 (%), un Root Mean Squared Error de 3.4781 (%), Mean Absolute Error de 2.2583 (%) y R-Squared de 0.2840.

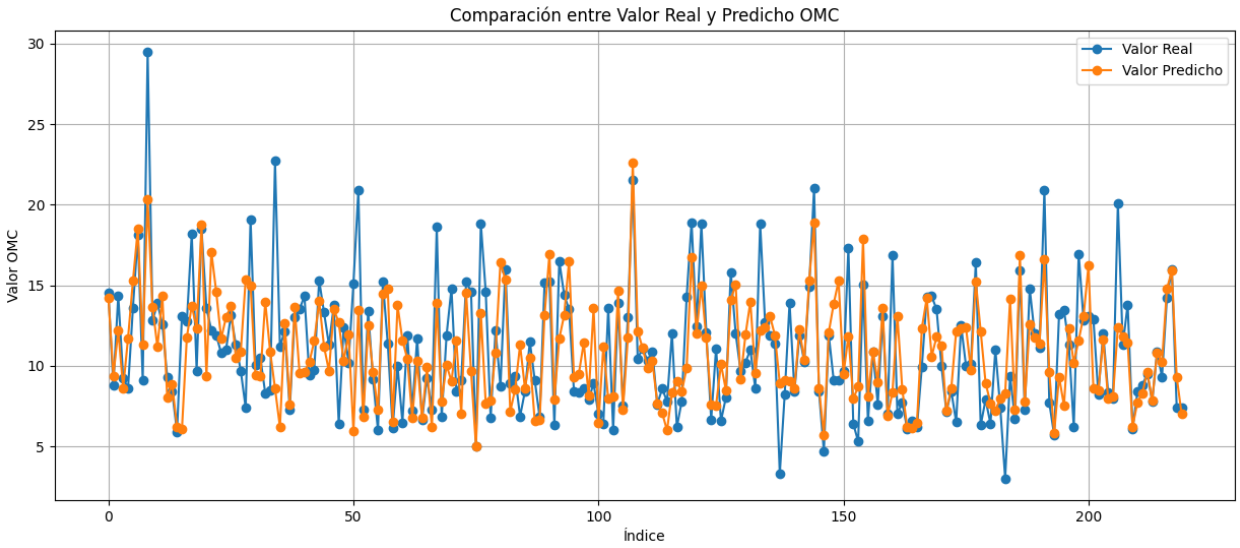
A continuación, en la Figura 94 se muestra el diagrama de dispersión de los valores actuales o reales vs. los valores predichos, donde se pude apreciar una correlación positiva moderada entre los valores reales y los valores predichos. Así como también, se puede verificar que el modelo para predecir el OMC con la técnica Redes Neuronales Artificiales se ajusta a la línea de regresión de los valores reales y predichos.

Figura 94 Diagrama de dispersión de los valores Actuales vs. Predichos utilizando Redes Neuronales Artificiales para predecir el OMC



Finalmente, se muestra el diagrama de líneas en la Figura 95, donde se puede verificar la similitud entre los valores reales y valores predichos del OMC para una cierta cantidad de datos o índice de datos del conjunto de prueba. Así como también, se puede verificar que las líneas de los valores reales y los valores predichos siguen patrones similares y se asemejan gráficamente.

Figura 95 Diagrama de líneas de los valores Reales vs. Predichos utilizando Redes Neuronales Artificiales para predecir el OMC



4.5.3.4. Herramientas de Auto Machine Learning

La herramienta Auto Machine Learning fue utilizada como parte fundamental de la metodología de este estudio para automatizar los procesos de minería de datos, selección y ajuste de modelos de aprendizaje automático. Donde se emplearon varias herramientas de Auto Machine Learning, entre las que se incluyen AutoKeras y AutoWeka. Estas herramientas ofrecieron una amplia gama de funcionalidades, desde la normalización, selección automática de modelos hasta la validación cruzada y la optimización de hiperparámetros. Lo cual permitió agilizar el proceso de modelado, reduciendo significativamente el tiempo necesario para explorar diferentes configuraciones de modelos e hiperparámetros liberando recursos de búsqueda manual e iterativa. En este estudio se demostró que la herramienta Auto Machine Learning es fundamental en el desarrollo de modelos de Machine Learning, ya que proporcionó una manera eficiente y efectiva de construir modelos de aprendizaje automático de alta calidad, lo que contribuyó significativamente a los resultados obtenidos en la presente investigación.

A. Auto Machine Learning para predecir la variable objetivo CBR95

En la Figura 96 se presenta la mejor arquitectura de red neuronal después de ejecutar la herramienta de Auto Machine Learning, específicamente Auto-Keras para el conjunto de datos de la variable objetivo CBR95. Se puede observar que la herramienta de Auto-ML arroja el mejor modelo después de entrenar 134.401 parámetros y compararlos entre ellos. Para luego, mostrar la mejor arquitectura de red neuronal descrita en el recuadro, donde se tiene una arquitectura compleja con 7 capas, incluyendo capas de entrada (con 8 neuronas), de normalización (8 neuronas), capas densas ReLU (con 256 y 512 neuronas) y una capa de salida (con 1 neurona). La capa de normalización permitió convertir las características en un formato flotante de 32 bits para una mejor operacionalización y ajuste de pesos de la red neuronal. Así mismo, La función de activación ReLU en las capas ocultas permitió a la red aprender relaciones no lineales entre las características y el valor de la variable objetivo CBR95.

Figura 96 Propiedad Mecánica CBR95: Mejor arquitectura de red neuronal utilizando Auto-Keras

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 8)	0
cast_to_float32 (CastToFloat32)	(None, 8)	0
dense (Dense)	(None, 256)	2,304
re_lu (ReLU)	(None, 256)	0
dense_1 (Dense)	(None, 512)	131,584
re_lu_1 (ReLU)	(None, 512)	0
regression_head_1 (Dense)	(None, 1)	513

Total params: 134,401 (525.00 KB)
Trainable params: 134,401 (525.00 KB)
Non-trainable params: 0 (0.00 B)

Para Auto-Keras generar dicha arquitectura de red neuronal y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 27 los siguientes hiperparámetros utilizados con dicha herramienta de Auto-ML:

Tabla 27 Hiperparámetros utilizados con la herramienta Auto-Keras para predecir el CBR95

Descripción	***Valor***
max_trials	50
tuner	greedy
overwrite	True
seed	42

epochs	50

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 91.3795 (%), un Root Mean Squared Error de 9.5592 (%), Mean Absolute Error de 6.1334 (%) y R-Squared de 0.6753.

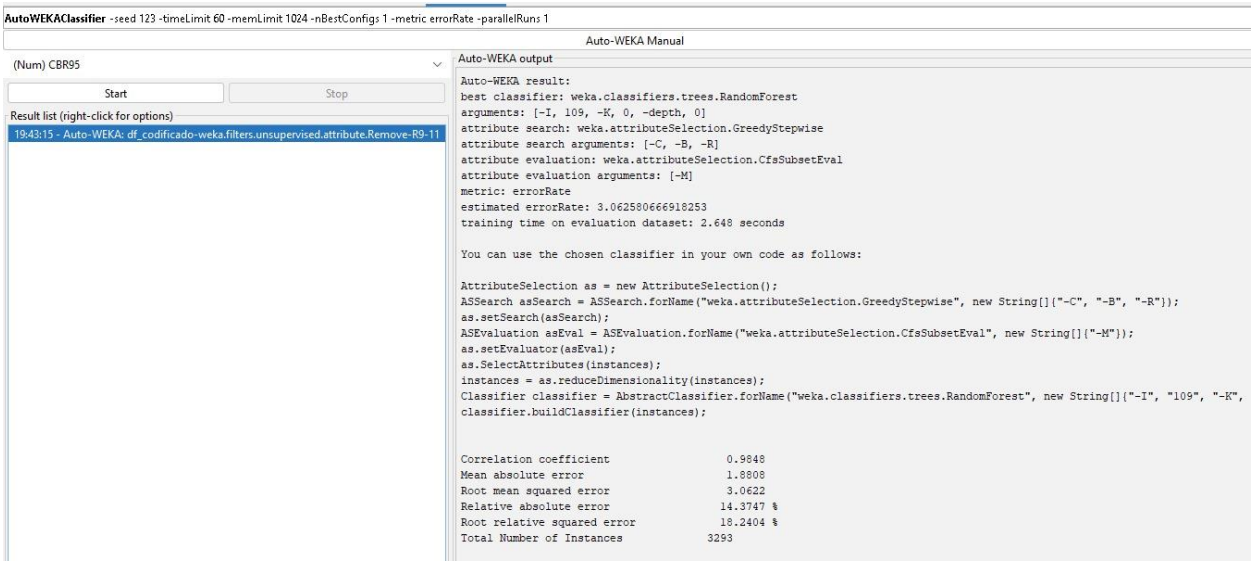
Por otro lado, en la Figura 97 se presenta los hiperparámetros obtenidos después de ejecutar la herramienta de Auto Machine Learning, específicamente Auto-WEKA para el conjunto de datos de la variable objetivo CBR95. Donde se puede observar que se identificó el mejor modelo que es Random Forest después de entrenar 3293 instancias o modelos. Así como también arrojó los hiperparámetros que se utilizaron para generar dicho modelo (Tabla 28). Auto-WEKA empleó técnicas de optimización para seleccionar la configuración óptima del modelo, maximizando una métrica de rendimiento, en este caso se utilizó la tasa de error, para luego compararlos entre ellos e identificar el mejor modelo.

Tabla 28 Hiperparámetros del mejor modelo generado por la herramienta Auto-WEKA para predecir el CBR95

Descripción	***Valor***
-I (N_stimators)	109
-K (Max_features)	0
-depth (Max_depth)	0
attribute search	weka.attributeSelection.GreedyStepwise
attribute search arguments	categorical (-C), binary (-B) or real (-R)
attribute evaluation	weka.attributeSelection.CfsSubsetEval
attribute evaluation arguments	-M (Merit)
metric	errorRate

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 9.3771 (%), un Root Mean Squared Error de 3.0622 (%), Mean Absolute Error de 1.8808 (%) y R-Squared de 0.9848.

Figura 97 Propiedad Mecánica CBR95: Resultados de la herramienta Auto-WEKA



Para Auto-WEKA generar dicho modelo y para cuestiones de replicabilidad se muestran en la Tabla 29 los siguientes hiperparámetros utilizados como entrada de la herramienta Auto-WEKA:

Tabla 29 Propiedad Mecánica CBR95: Hiperparámetros de entrada utilizados por la herramienta Auto-WEKA

Descripción	***Valor***
batchSize	100
debug	False
doNotCheckCapabilities	False
memLimit	1024
metric	errorRate
nBestConfigs	1
numDecimalPlaces	2
ParallelRuns	1
seed	123
timeLimit	60

B. Auto Machine Learning para predecir la variable objetivo CBR100

En la Figura 98 se presenta la mejor arquitectura de red neuronal después de ejecutar la herramienta de Auto Machine Learning, específicamente Auto-Keras para el conjunto de datos de

la variable objetivo CBR100. Se puede observar que la herramienta de Auto-ML arroja el mejor modelo después de entrenar 34.433 parámetros y compararlos entre ellos. Para luego, mostrar la mejor arquitectura de red neuronal descrita en el recuadro, donde se tiene una arquitectura compleja con 10 capas, incluyendo capas de entrada (con 8 neuronas), de normalización (8 neuronas), capas densas ReLU y dropout (con 128 y 256 neuronas) y una capa de salida (con 1 neurona). La capa de normalización permitió convertir las características en un formato flotante de 32 bits para una mejor operacionalización y ajuste de pesos de la red neuronal. Así mismo, La función de activación ReLU en las capas ocultas permitió a la red aprender relaciones no lineales entre las características y el valor de la variable objetivo CBR100 y la función de dropout permitió apagar aleatoriamente un porcentaje de las neuronas con el objetivo de que las neuronas no caigan en el sobre ajuste.

Figura 98 Propiedad Mecánica CBR100: Mejor arquitectura de red neuronal utilizando Auto-Keras

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 8)	0
cast_to_float32 (CastToFloat32)	(None, 8)	0
dense (Dense)	(None, 128)	1,152
re_lu (ReLU)	(None, 128)	0
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 256)	33,024
re_lu_1 (ReLU)	(None, 256)	0
dropout_1 (Dropout)	(None, 256)	0
dropout_2 (Dropout)	(None, 256)	0
regression_head_1 (Dense)	(None, 1)	257

Total params: 34,433 (134.50 KB)
Trainable params: 34,433 (134.50 KB)
Non-trainable params: 0 (0.00 B)

Para Auto-Keras generar dicha arquitectura de red neuronal y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 30 los siguientes hiperparámetros utilizados con dicha herramienta de Auto-ML:

Tabla 30 Hiperparámetros utilizados con la herramienta Auto-Keras para predecir el CBR100

Descripción	***Valor***
max_trials	50
tuner	random
overwrite	True
seed	42
epochs	50

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 214.3554 (%), un Root Mean Squared Error de 14.6408 (%), Mean Absolute Error de 9.4795 (%) y R-Squared de 0.6812.

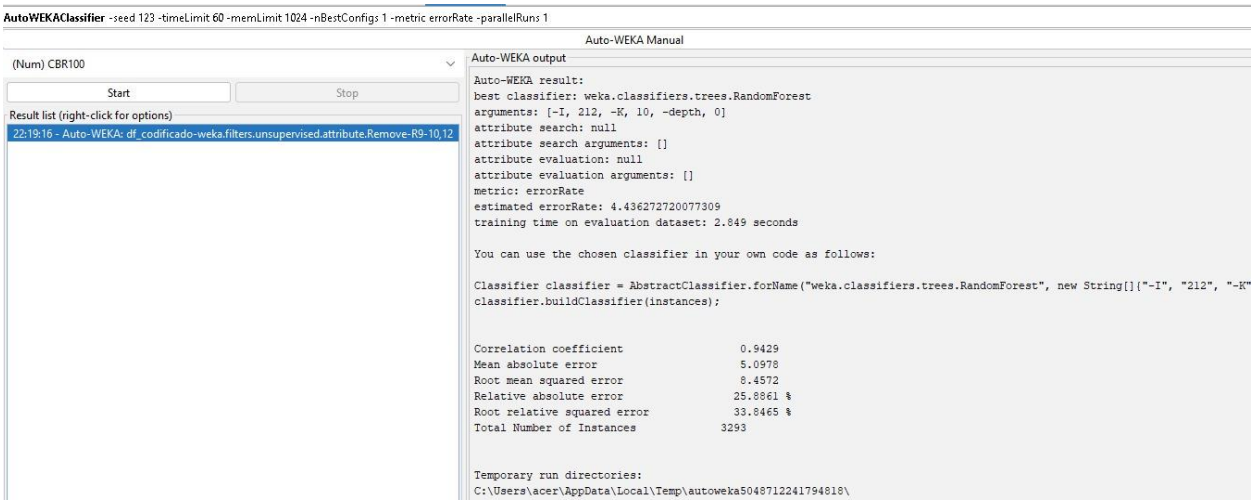
Por otro lado, en la Figura 99 se presenta los hiperparámetros obtenidos después de ejecutar la herramienta de Auto Machine Learning, específicamente Auto-WEKA para el conjunto de datos de la variable objetivo CBR100. Donde se puede observar que se identificó el mejor modelo que es Random Forest después de entrenar 3293 instancias o modelos. Así como también arrojó los hiperparámetros que se utilizaron para generar dicho modelo (Tabla 31). Auto-WEKA empleó técnicas de optimización para seleccionar la configuración óptima del modelo, maximizando una métrica de rendimiento, en este caso se utilizó la tasa de error, para luego compararlos entre ellos e identificar el mejor modelo.

Tabla 31 Hiperparámetros del mejor modelo generado por la herramienta Auto-WEKA para predecir el CBR100

Descripción	***Valor***
-I (N_stimators)	212
-K (Max_features)	10
-depth (Max_depth)	0
attribute search	null
attribute search arguments	-
attribute evaluation	null
attribute evaluation arguments	-
metric	errorRate

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 71.5242 (%), un Root Mean Squared Error de 8.4572 (%), Mean Absolute Error de 5.0978 (%) y R-Squared de 0.9429.

Figura 99 Propiedad Mecánica CBR100: Resultados de la herramienta Auto-WEKA



Para Auto-WEKA generar dicho modelo y para cuestiones de replicabilidad se muestran en la Tabla 32 los siguientes hiperparámetros utilizados como entrada de la herramienta Auto-WEKA:

Tabla 32 Propiedad Mecánica CBR100: Hiperparámetros de entrada utilizados por la herramienta Auto-WEKA

Descripción	***Valor***
batchSize	100
debug	False
doNotCheckCapabilities	False
memLimit	1024
metric	errorRate
nBestConfigs	1
numDecimalPlaces	2
ParallelRuns	1
seed	123
timeLimit	60

C. Auto Machine Learning para predecir la variable objetivo MDD

En la Figura 100 se presenta la mejor arquitectura de red neuronal después de ejecutar la herramienta de Auto Machine Learning, específicamente Auto-Keras para el conjunto de datos de la variable objetivo MDD. Se puede observar que la herramienta de Auto-ML arroja el mejor modelo después de entrenar 1.377 parámetros y compararlos entre ellos. Para luego, mostrar la mejor arquitectura de red neuronal descrita en el recuadro, donde se tiene una arquitectura

compleja con 9 capas, incluyendo capas de entrada (con 8 neuronas), de normalización (8 neuronas), capas densas ReLU y dropout (con 32 y 32 neuronas) y una capa de salida (con 1 neurona). La capa de normalización permitió convertir las características en un formato flotante de 32 bits para una mejor operacionalización y ajuste de pesos de la red neuronal. Así mismo, La función de activación ReLU en las capas ocultas permitió a la red aprender relaciones no lineales entre las características y el valor de la variable objetivo MDD y la función de dropout permitió apagar aleatoriamente un porcentaje de las neuronas con el objetivo de que las neuronas no caigan en el sobre ajuste.

Figura 100 Propiedad Mecánica MDD: Mejor arquitectura de red neuronal utilizando Auto-Keras

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 8)	0
cast_to_float32 (CastToFloat32)	(None, 8)	0
dense (Dense)	(None, 32)	288
re_lu (ReLU)	(None, 32)	0
dropout (Dropout)	(None, 32)	0
dense_1 (Dense)	(None, 32)	1,056
re_lu_1 (ReLU)	(None, 32)	0
dropout_1 (Dropout)	(None, 32)	0
regression_head_1 (Dense)	(None, 1)	33

Total params: 1,377 (5.38 KB)
Trainable params: 1,377 (5.38 KB)
Non-trainable params: 0 (0.00 B)

Para Auto-Keras generar dicha arquitectura de red neuronal y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 33 los siguientes hiperparámetros utilizados con dicha herramienta de Auto-ML:

Tabla 33 Hiper parámetros utilizados con la herramienta Auto-Keras para predecir el MDD

Descripción	***Valor***
max_trials	50
tuner	greedy
overwrite	True
seed	42
epochs	50

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 0.0185 (g/cm3), un Root Mean Squared Error de 0.1362 (g/cm3), Mean Absolute Error de 0.1022 (g/cm3) y R-Squared de 0.5187.

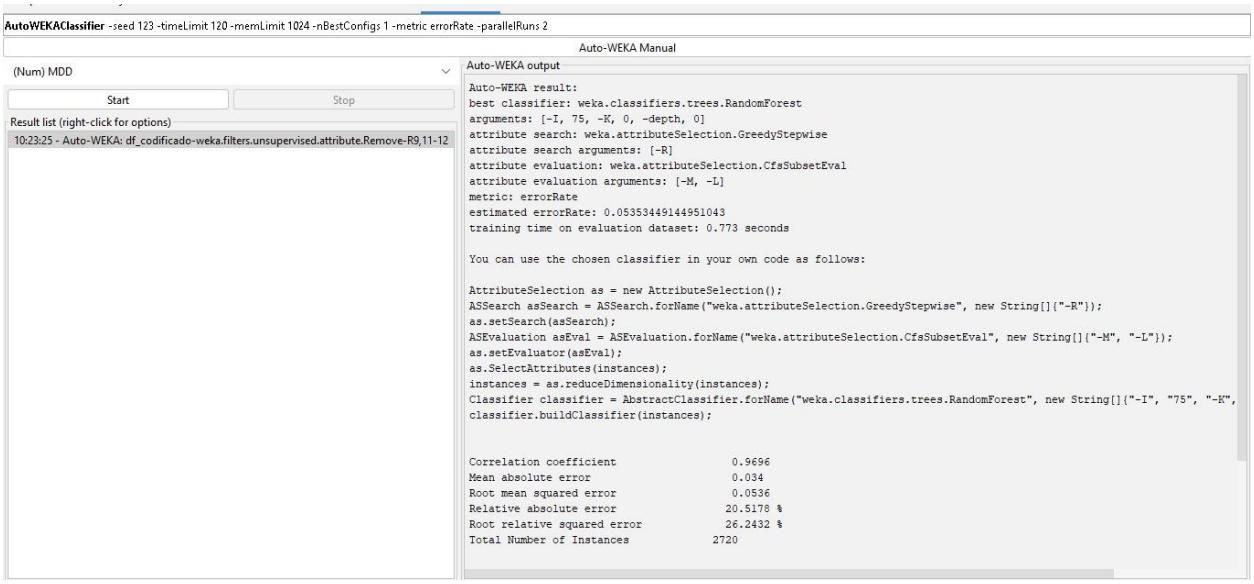
Por otro lado, en la Figura 101 se presenta los hiperparámetros obtenidos después de ejecutar la herramienta de Auto Machine Learning, específicamente Auto-WEKA para el conjunto de datos de la variable objetivo MDD. Donde se puede observar que se identificó el mejor modelo que es Random Forest después de entrenar 2720 instancias o modelos. Así como también arrojó los hiperparámetros que se utilizaron para generar dicho modelo (Tabla 34). Auto-WEKA empleó técnicas de optimización para seleccionar la configuración óptima del modelo, maximizando una métrica de rendimiento, en este caso se utilizó la tasa de error, para luego compararlos entre ellos e identificar el mejor modelo.

Tabla 34 Hiperparámetros del mejor modelo generado por la herramienta Auto-WEKA para predecir el MDD

Descripción	***Valor***
-I (N_stimators)	75
-K (Max_features)	0
-depth (Max_depth)	0
attribute search	weka.attributeSelection.GreedyStepwise
attribute search arguments	real (-R)
attribute evaluation	weka.attributeSelection.CfsSubsetEval
attribute evaluation arguments	-M (Merit), -L
metric	errorRate

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 0.0003 (g/cm3), un Root Mean Squared Error de 0.0536 (g/cm3), Mean Absolute Error de 0.0340 (g/cm3) y R-Squared de 0.9696.

Figura 101 Propiedad Mecánica MDD: Resultados de la herramienta Auto-WEKA



Para Auto-WEKA generar dicho modelo y para cuestiones de replicabilidad se muestran en la Tabla 35 los siguientes hiperparámetros utilizados como entrada de la herramienta Auto-WEKA:

Tabla 35 Propiedad Mecánica MDD: Hiperparámetros de entrada utilizados por la herramienta Auto-WEKA

Descripción	***Valor***
batchSize	100
debug	False
doNotCheckCapabilities	False
memLimit	1024
metric	errorRate
nBestConfigs	1
numDecimalPlaces	2
ParallelRuns	2
seed	123
timeLimit	120

D. Auto Machine Learning para predecir la variable objetivo OMC

En la Figura 102 se presenta la mejor arquitectura de red neuronal después de ejecutar la herramienta de Auto Machine Learning, específicamente Auto-Keras para el conjunto de datos de la variable objetivo OMC. Se puede observar que la herramienta de Auto-ML arroja el mejor modelo después de entrenar 793.601 parámetros y compararlos entre ellos. Para luego, mostrar la

mejor arquitectura de red neuronal descrita en el recuadro, donde se tiene una arquitectura compleja con 12 capas, incluyendo capas de entrada (con 8 neuronas), de normalización (8 neuronas), capas densas ReLU y dropout (con 512, 512 y 1024 neuronas) y una capa de salida (con 1 neurona). La capa de normalización permitió convertir las características en un formato flotante de 32 bits para una mejor operacionalización y ajuste de pesos de la red neuronal. Así mismo, La función de activación ReLU en las capas ocultas permitió a la red aprender relaciones no lineales entre las características y el valor de la variable objetivo OMC y la función de dropout permitió apagar aleatoriamente un porcentaje de las neuronas con el objetivo de que las neuronas no caigan en el sobre ajuste.

Figura 102 Propiedad Mecánica OMC: Mejor arquitectura de red neuronal utilizando Auto-Keras

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 8)	0
cast_to_float32 (CastToFloat32)	(None, 8)	0
dense (Dense)	(None, 512)	4,608
re_lu (ReLU)	(None, 512)	0
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 512)	262,656
re_lu_1 (ReLU)	(None, 512)	0
dropout_1 (Dropout)	(None, 512)	0
dense_2 (Dense)	(None, 1024)	525,312
re_lu_2 (ReLU)	(None, 1024)	0
dropout_2 (Dropout)	(None, 1024)	0
regression_head_1 (Dense)	(None, 1)	1,025

Total params: 793,601 (3.03 MB)
Trainable params: 793,601 (3.03 MB)
Non-trainable params: 0 (0.00 B)

Para Auto-Keras generar dicha arquitectura de red neuronal y para cuestiones de replicabilidad de dicho modelo se muestran en la Tabla 36 los siguientes hiperparámetros utilizados con dicha herramienta de Auto-ML:

Tabla 36 Hipeparámetros utilizados con la herramienta Auto-Keras para predecir el OMC

Descripción	***Valor***
max_trials	50
tuner	greedy
overwrite	True
seed	42
epochs	50

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 9.1904 (%), un Root Mean Squared Error de 3.0315 (%), Mean Absolute Error de 1.9983 (%) y R-Squared de 0.4561.

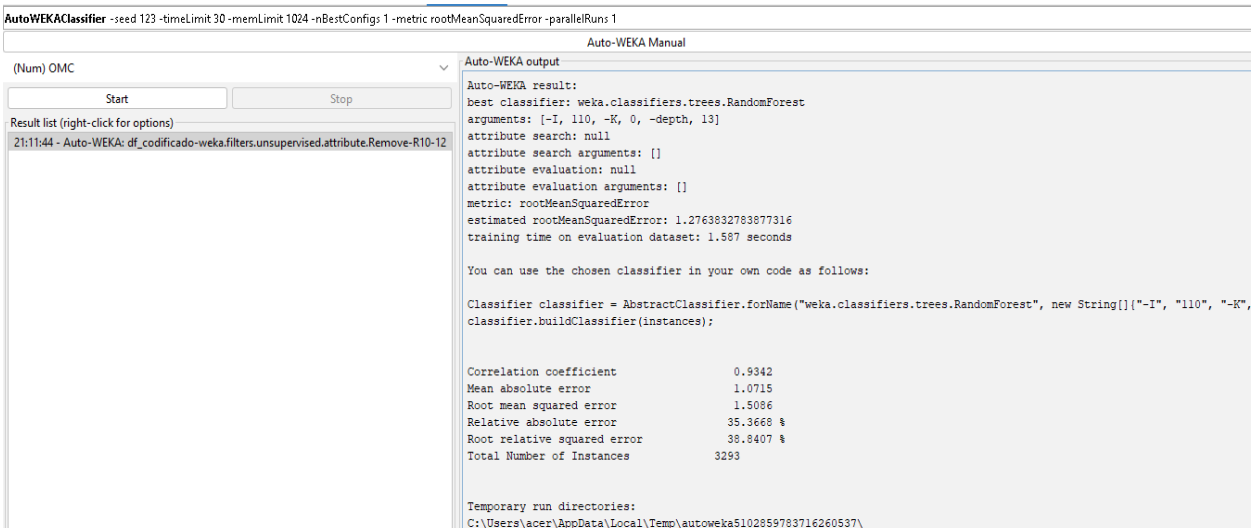
Por otro lado, en la Figura 103 se presenta los hiperparámetros obtenidos después de ejecutar la herramienta de Auto Machine Learning, específicamente Auto-WEKA para el conjunto de datos de la variable objetivo OMC. Donde se puede observar que se identificó el mejor modelo que es Random Forest después de entrenar 3293 instancias o modelos. Así como también arrojó los hiperparámetros que se utilizaron para generar dicho modelo (Tabla 37 Auto-WEKA empleó técnicas de optimización para seleccionar la configuración óptima del modelo, maximizando una métrica de rendimiento, en este caso se utilizó la Raíz del Error Cuadrático Medio, para luego compararlos entre ellos e identificar el mejor modelo.

Tabla 37 Hiperparámetros del mejor modelo generado por la herramienta Auto-WEKA para predecir el OMC

Descripción	***Valor***
-I (N_stimators)	110
-K (Max_features)	0
-depth (Max_depth)	13
attribute search	null
attribute search arguments	-
attribute evaluation	null
attribute evaluation arguments	-
metric	rootMeanSquaredError

Para los anteriores hiperparámetros se obtuvo los resultados de las métricas de evaluación. Donde se obtuvo un Mean Squared Error de 2.2759 (%), un Root Mean Squared Error de 1.5086 (%), Mean Absolute Error de 1.0715 (%) y R-Squared de 0.9342.

Figura 103 Propiedad Mecánica OMC: Resultados de la herramienta Auto-WEKA



Para Auto-WEKA generar dicho modelo y para cuestiones de replicabilidad se muestran en la Tabla 38 los siguientes hiperparámetros utilizados como entrada de la herramienta Auto-WEKA:

Tabla 38 Propiedad Mecánica OMC: Hiperparámetros de entrada utilizados por la herramienta Auto-WEKA

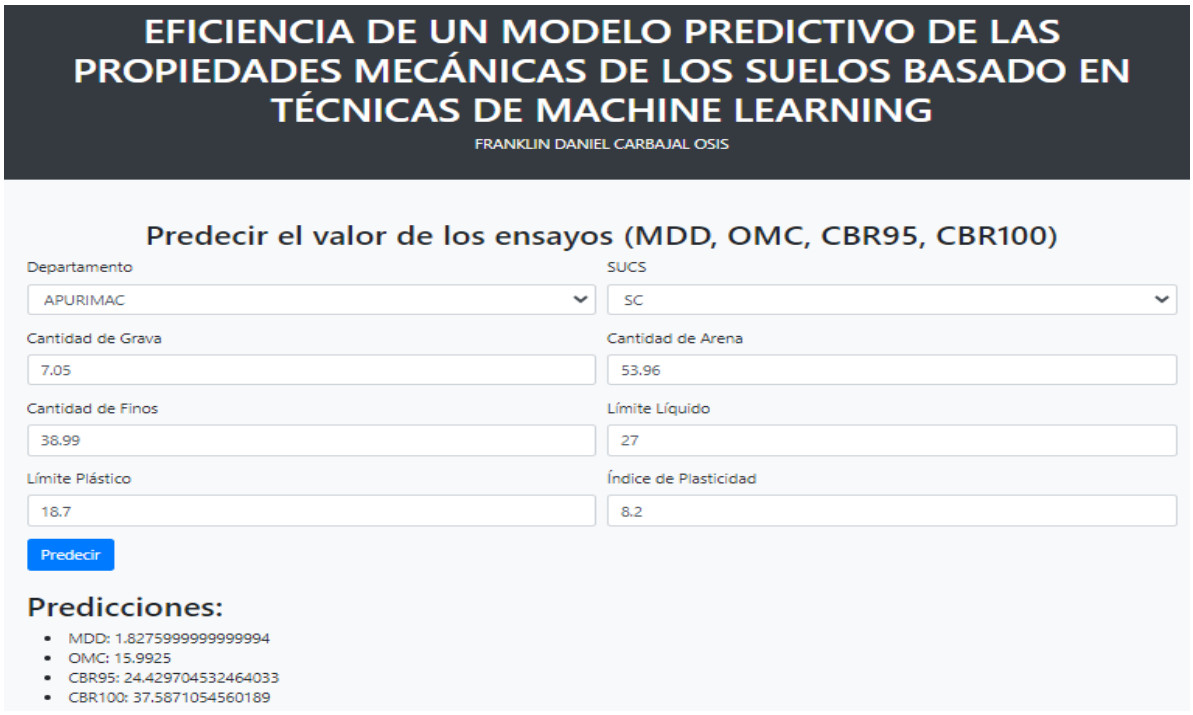
Descripción	***Valor***
batchSize	100
debug	False
doNotCheckCapabilities	False
memLimit	1024
metric	rootMeanSquaredError
nBestConfigs	1
numDecimalPlaces	2
ParallelRuns	1
seed	123
timeLimit	30

4.5.4. Etapa IV: Evaluación e Interpretación

En esta etapa se implementó las métricas de evaluación de Machine Learning donde se evaluó e interpretó las predicciones del modelo predictivo, para luego identificar cuál de las técnicas de Machine Learning fue el más adecuado para la predicción de las propiedades mecánicas del suelo. Se aplicó la ficha de registro de datos para el rendimiento de los modelos (Ver Anexo 3) que se muestran en el siguiente capitulo como resultado de la comparación de los modelos.

Así mismo, de forma adicional se implementó una aplicación para predecir las propiedades mecánicas de los suelos con los modelos desarrollados anteriormente en cada etapa. Donde se puede observar que dichos modelos son funcionales y aplicables mediante el ingreso de datos a través de una interfaz de usuario. Dicha interfaz de usuario permitió predecir valores no antes vistas en el set de datos, cabe recalcar que la precisión de las predicciones depende de las métricas de evaluación anteriormente aplicadas y analizadas en cada modelo. En la Figura 104 se puede apreciar la predicción de una muestra de suelo del departamento de Apurímac con sus respectivos valores de entrada del modelo y como resultado muestra los valores de salida para cada propiedad mecánica del suelo.

Figura 104 Aplicación para predecir las propiedades mecánicas de los suelos



4.6. Técnicas e instrumentos de Recolección de Datos

Se utilizaron las técnicas para el recojo de información que son: Recopilación de datos de Bases de datos, publicaciones científicas e informes. Por consiguiente, se utilizó el instrumento: Formatos de recopilación de base de datos y Análisis de datos.

4.7. Técnicas estadísticas

- Test de Kolmogorov – Smirnov (mayor 50)

  La prueba de Kolmogorov-Smirnov es un método estadístico utilizado para comparar una distribución empírica (obtenida a partir de datos experimentales) con una distribución teórica. Esta prueba evalúa la diferencia entre las dos distribuciones y determina si existe una diferencia significativa o no. En otras palabras, la prueba de Kolmogorov-Smirnov compara un conjunto de datos reales con un modelo hipotético para ver si los datos se ajustan al modelo o no (Cordova Valentin & Mori Montalvo, 2021). En la presente investigación se utilizó para probar la normalidad de las variables en la etapa de selección, limpieza y transformación de la metodología KDD.

- Coeficiente de correlación Spearman (P)

  Si de la prueba Kolmogorov-Smirnov, las variables no se ajustan a una distribución normal, se utilizará la correlación de Spearman. En la presente investigación se utilizó para correlacionar las variables de entrada del modelo con las variables a predecir de dicho modelo en la etapa de selección, limpieza y transformación de la metodología KDD.

$$^{p} = 1 - \frac{6\sum_{i=1}^{n}(d_i)^2}{n(n^2-1)}$$

  Donde:

  d_i: es la diferencia entre los rangos de las observaciones en las dos variables para el i-ésimo par de datos.

  n: Es el número de observaciones.

- Coeficiente de Determinación (R-Squared)

Como ya se vio en el marco teórico mide el ajuste por el modelo de regresión. Un valor alto (cercano a 1) indica un buen ajuste del modelo, mientras que un valor bajo (cercano a 0) indica un mal ajuste. En la presente investigación se utilizó en la etapa de evaluación e interpretación de la metodología KDD.

- Error cuadrático medio (MSE)

  Como ya se vio en el marco teórico mide la diferencia promedio al cuadrado entre los valores predichos y los reales. Un MSE bajo indica un mejor ajuste del modelo. En la presente investigación se utilizó en la etapa de evaluación e interpretación de la metodología KDD.

- Raíz cuadrada del error cuadrático medio (RMSE)

  Como ya se vio en el marco teórico es la raíz cuadrada del MSE. Representa la desviación estándar de los errores predichos. Un RMSE bajo indica un mejor ajuste del modelo. En la presente investigación se utilizó en la etapa de evaluación e interpretación de la metodología KDD.

- Error absoluto medio (MAE)

  Como ya se vio en el marco teórico mide la diferencia promedio absoluta entre los valores predichos y los reales. Un MAE bajo indica un mejor ajuste del modelo. En la presente investigación se utilizó en la etapa de evaluación e interpretación de la metodología KDD.

CAPÍTULO V
RESULTADOS

5.1. Resultados de la comparación de modelos

En la Tabla 39 se presentan los resultados comparativos de los diferentes modelos de regresión desarrollados en la presente investigación, sin incluir los modelos desarrollados con las herramientas de Auto-ML. Donde se incluyen los modelos de las técnicas de Extra Trees Regressor, Random Forest Regressor y Redes Neuronales Artificiales. Mediante los algoritmos de evaluación de métricas de modelos de regresión de Python se pudo calcular las métricas estadísticas anteriormente detalladas en el marco teórico. Métricas como el Error Cuadrático Medio (MSE), Raíz del Error Cuadrático Medio (RMSE), Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R-Squared), que sirvieron para evaluar el rendimiento de cada técnica de predicción para cada propiedad mecánica del suelo (CBR95, CBR100, MDD y OMC).

Se ha utilizado las comparaciones de las métricas de evaluación tanto del modelo base como los resultados de la búsqueda de hiperparámetros con validación cruzada (GridSearchCV). Los resultados indican que, en general, los modelos de ensamble como Random Forest Regressor y Extra Trees Regressor superan a las Redes Neuronales Artificiales en términos de precisión. Esta diferencia puede variar según el ajuste de hiperparámetros y las condiciones de entrenamiento, ya que en el caso de las Redes Neuronales Artificiales se necesitó más tiempo de entrenamiento y diferentes experimentos con arquitecturas e hiperhiperparámetros. A diferencia de los conjuntos de árboles que necesitan poco tiempo de entrenamiento y pocas configuraciones de hiperparámetros. Por último, las herramientas GridSearchCV permite validar y aumentar el rendimiento de los modelos mediante el ajuste de hiperparámetros, pero de igual manera requiere una gran cantidad de tiempo. A diferencia de utilizar técnicas avanzadas de investigaciones recientes como los desafíos de Neural Architecture Search (NAS) y Combined Algorithm Selection and Hyperparameter Optimization (CASH) de Auto-ML, que, además, de utilizar GridSearchCV hacen uso de técnicas innovadoras como el morfismo de red, la optimización bayesiana y entre otras técnicas que permiten ahorrar tiempo y recursos en el desarrollo y búsqueda del mejor modelo.

Tabla 39 Resultados de las métricas estadísticas del rendimiento de los modelos

PROPIEDAD	MSE	RMSE	MAE	R-SQUARED
CBR95 (%)	(Model, GridSCV)	(Model, GridSCV)	(Model, GridSCV)	(Model, GridSCV)
Extra Trees Regressor	63.3174, **62.1625**	7.9572, **7.8843**	4.9235, **5.0304**	0.7750, **0.7791**
Random Forest Regressor	**63.4443, 63.4443**	**7.9651, 7.9651**	**5.0974, 5.0974**	**0.7745, 0.7745**
Redes neuronales artificiales	**81.3957**, 98.0474	**9.0219**, 9.9018	**5.7619**, 6.1997	**0.7108**, 0.6516
CBR100 (%)				
Extra Trees Regressor	**163.2917**, 163.4903	**12.7785**, 12.7863	**7.6829**, 7.7973	**0.7572**, 0.7569
Random Forest Regressor	**163.0446, 163.0446**	**12.7688, 12.7688**	**7.9889, 7.9889**	**0.7575, 0.7575**
Redes neuronales artificiales	**206.4105**, 225.7351	**14.3669**, 15.0244	**9.0309**, 9.4575	**0.6930**, 0.6643
MDD (g/cm3)				
Extra Trees Regressor	0.0162, **0.0142**	0.1275, **0.1194**	0.0835, **0.0820**	0.5782, **0.6298**
Random Forest Regressor	0.0145, **0.0141**	0.1204, **0.1188**	0.0827, **0.0825**	0.6238, **0.6335**
Redes neuronales artificiales	**0.0185**, 0.0200	**0.1361**, 0.1414	**0.0984**, 0.1023	**0.5194**, 0.4801
OMC (%)				
Extra Trees Regressor	8.5054, **7.7777**	2.9164, **2.7888**	1.7624, **1.7296**	0.4966, **0.5397**
Random Forest Regressor	8.1513, **8.0046**	2.8550, **2.8292**	1.8104, **1.7991**	0.5176, **0.5262**
Redes neuronales artificiales	**9.1967**, 12.0976	**3.0326**, 3.4781	**2.0249**, 2.2583	**0.4557**, 0.2840

Así mismo, los resultados indican que las técnicas con mejor rendimiento con respecto a las métricas de evaluación de los modelos de las propiedades mecánicas de los suelos son: La técnica Extra Trees Regressor para predecir el valor del CBR95, la técnica Random Forest Regressor para predecir el valor del CBR100, técnica Random Forest Regressor para predecir el valor de la MDD y la técnica Extra Trees Regressor para predecir el valor del OMC. Donde, se puede evidenciar que los errores de cada mejor técnica y su respectivo modelo es el más cercano al valor de cero y el más mínimo en comparación con las otras técnicas, por otro lado, el coeficiente de determinación es el más cercano al valor de uno y el máximo en comparación con las otras técnicas. Des está manera, se determinó las mejores técnicas y sus respectivos modelos que serán comparados con los modelos utilizando las técnicas de Auto-ML para escoger la técnica y modelo de Machine Learning más eficiente.

En la siguiente tabla (Tabla 40) se comparan las eficiencias de las técnicas de Machine Learning y sus respectivos modelos utilizando técnicas avanzadas de Auto-ML como Neural Architecture Search (NAS) y Combined Algorithm Selection and Hyperparameter Optimization (CASH) que ya vienen implementadas en las herramientas de Auto-Keras y Auto- WEKA. Para medir la eficiencia de dichas técnicas y de sus respectivos modelos, primero se midió la eficacia

de los rendimientos de los modelos que se muestran en la Tabla 39. Para luego, que mediante el uso de las herramientas de Auto-ML se comparan las eficiencias de las técnicas y sus respectivos modelos con las técnicas que no utilizaron dichas herramientas avanzadas de Auto-ML.

Tabla 40 Resultados de las métricas estadísticas del rendimiento de los modelos

PROPIEDAD	MSE	RMSE	MAE	R-SQUARED
CBR95 (%)				
Extra Trees Regressor	**62.1625**	**7.8843**	**5.0304**	**0.7791**
Random Forest Regressor	63.4443	7.9651	5.0974	0.7745
Redes neuronales artificiales	81.3957	9.0219	5.7619	0.7108
Redes neuronales artificiales (Auto-Keras)	91.3795	9.5592	6.1334	0.6753
Random Forest Regressor (Auto-WEKA)	**9.3771**	**3.0622**	**1.8808**	**0.9848**
CBR100 (%)				
Extra Trees Regressor	163.2917	12.7785	7.6829	0.7572
Random Forest Regressor	**163.0446**	**12.7688**	**7.9889**	**0.7575**
Redes neuronales artificiales	206.4105	14.3669	9.0309	0.6930
Redes neuronales artificiales (Auto-Keras)	214.3554	14.6408	9.4795	0.6812
Random Forest Regressor (Auto-WEKA)	**71.5242**	**8.4572**	**5.0978**	**0.9429**
MDD (g/cm3)				
Extra Trees Regressor	0.0142	0.1194	0.0820	0.6298
Random Forest Regressor	**0.0141**	**0.1188**	**0.0825**	**0.6335**
Redes neuronales artificiales	0.0185	0.1361	0.0984	0.5194
Redes neuronales artificiales (Auto-Keras)	0.0185	0.1362	0.1022	0.5187
Random Forest Regressor (Auto-WEKA)	**0.0003**	**0.0536**	**0.0340**	**0.9696**
OMC (%)				
Extra Trees Regressor	**7.7777**	**2.7888**	**1.7296**	**0.5397**
Random Forest Regressor	8.0046	2.8292	1.7991	0.5262
Redes neuronales artificiales	9.1967	3.0326	2.0249	0.4557
Redes neuronales artificiales (Auto-Keras)	9.1904	3.0315	1.9983	0.4561
Random Forest Regressor (Auto-WEKA)	**2.2759**	**1.5086**	**1.0715**	**0.9342**

Después, de comparar los rendimientos de las diferentes técnicas y sus respectivos modelos se puede evidenciar que la técnica de Random Forest Regressor utilizando la herramienta Auto-WEKA es superior significativamente a todas las técnicas utilizadas. En cuestiones de eficiencia se puede evidenciar, que, para cada propiedad mecánica, la técnica Random Forest Regressor

utilizando la herramienta Auto-WEKA es significativamente superior a la técnica Random Forest Regressor sin utilizar la herramienta de Auto-ML. Por otro lado, se presenta lo contrario para la técnica de Redes Neuronales Artificiales utilizando la herramienta Auto-Keras. Esto se debe a que la herramienta Auto-Keras necesita más tiempo y capacidad de procesamiento para encontrar la arquitectura neuronal optima. Todo lo contario se pudo evidenciar en la herramienta Auto-WEKA ya que utiliza una combinación de técnicas y algoritmos de optimización (CASH) para encontrar la mejor técnica disminuyendo u optimizando una métrica de evaluación.

5.2. Resultados de los objetivos específicos

5.2.1. Resultado del objetivo específico 1

O1: Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para para predecir el resultado del ensayo de CBR (California Bearing Ratio).

En la siguiente tabla se muestra los resultados de las métricas de evaluación de las técnicas de Machine Learning:

Tabla 41 Resultados para el objetivo específico 1 CBR95

Métricas de evaluación	**CBR95: Algoritmos de Regresión**		
	Extra Trees Regressor	**Redes neuronales artificiales**	**Random Forest Regressor (Auto-WEKA)**
MSE (%)	62.1625	81.3957	**9.3771**
RMSE (%)	7.8843	9.0219	**3.0622**
MAE (%)	5.0304	5.7619	**1.8808**
R-SQUARED	0.7791	0.7108	**0.9848**

Tabla 42 Resultados para el objetivo específico 1 CBR100

Métricas de evaluación	**CBR100: Algoritmos de Regresión**		
	Extra Trees Regressor	**Redes neuronales artificiales**	**Random Forest Regressor (Auto-WEKA)**

MSE (%)	163.2917	206.4105	**71.5242**
RMSE (%)	12.7785	14.3669	**8.4572**
MAE (%)	7.6829	9.0309	**5.0978**
R-SQUARED	0.7572	0.6930	**0.9429**

La técnica de regresión de Machine Learning que mejor resultado tuvo es **Random Forest Regressor** para predecir el resultado del ensayo de CBR95 y CBR100 con un Mean Squared Error de **9.3771** (%) **y 71.5242** (%), un Root Mean Squared Error de **3.0622** (%) **y 8.4572** (%), Mean Absolute Error de **1.8808** (%) **y 5.0978** (%), R-Squared de **0.9848 y 0.9429**. Así mismo, a través del Error Absoluto Medio (MAE) igual a **1.8808** (%) **y 5.0978** (%) que expresa la existencia de un grado de error promedio MUY BAJO y R-SQUARED igual a **0.9848 y 0.9429** como coeficiente de determinación EXCELENTE. Esto indica que la técnica de regresión de Machine Learning de mejor rendimiento fue **Random Forest Regressor** y según los resultados y análisis de las tablas 39, 40, 41 y 42 aumenta la eficiente del modelo para para predecir el resultado del ensayo de CBR. Cabe mencionar, que el valor del ensayo de CBR95 tiene mejor rendimiento en comparación al valor del ensayo de CBR100, esto es conveniente, ya que en condiciones normales del suelo se utiliza más el ensayo de CBR95 y el CBR100 en casos teóricos o de comparación.

5.2.2. Resultado del objetivo específico 2

O2: Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para para predecir el resultado del ensayo de MDD (Máxima densidad seca).

En la siguiente tabla se muestra los resultados de las métricas de evaluación de las técnicas de Machine Learning:

Tabla 43 Resultados para el objetivo específico 2

Métricas de evaluación	**MDD: Algoritmos de Regresión**		
	Extra Trees Regressor	**Redes neuronales artificiales**	**Random Forest Regressor (Auto-WEKA)**
MSE (g/cm3)	0.0142	0.0185	**0.0003**
RMSE (g/cm3)	0.1194	0.1361	**0.0536**

MAE (g/cm3)	0.0820	0.0984	**0.0340**
R-SQUARED	0.6298	0.5194	**0.9696**

La técnica de regresión de Machine Learning que mejor resultado tuvo es **Random Forest Regressor** para predecir el resultado del ensayo de MDD con un Mean Squared Error de **0.0003** (g/cm3), un Root Mean Squared Error de **0.0536** (g/cm3), Mean Absolute Error de **0.0340** (g/cm3), R-Squared de **0.9696**. Así mismo, a través del Error Absoluto Medio (MAE) igual a **0.0340** (g/cm3) que expresa la existencia de un grado de error promedio MUY BAJO y R-SQUARED igual a **0.9696** como coeficiente de determinación EXCELENTE. Esto indica que la técnica de regresión de Machine Learning de mejor rendimiento fue **Random Forest Regressor** y según los resultados y análisis de las tablas 39, 40 y 43 aumenta la eficiente del modelo para para predecir el resultado del ensayo de MDD.

5.2.3. Resultado del objetivo específico 3

O3: Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para predecir el resultado del ensayo de OMC (Contenido óptimo de humedad).

En la siguiente tabla se muestra los resultados de las métricas de evaluación de las técnicas de Machine Learning:

Tabla 44 Resultados para el objetivo específico 3

Métricas de evaluación	**OMC: Algoritmos de Regresión**		
	Extra Trees Regressor	**Redes neuronales artificiales (Auto-Keras)**	**Random Forest Regressor (Auto-WEKA)**
MSE (%)	7.7777	9.1904	**2.2759**
RMSE (%)	2.7888	3.0315	**1.5086**
MAE (%)	1.7296	1.9983	**1.0715**
R-SQUARED	0.5397	0.4561	**0.9342**

La técnica de regresión de Machine Learning que mejor resultado tuvo es **Random Forest Regressor** para predecir el resultado del ensayo de OMC con un Mean Squared Error de **2.2759** (%), un Root Mean Squared Error de **1.5086** (%), Mean Absolute Error de **1.0715** (%), R-Squared de **0.9342**. Así mismo, a través del Error Absoluto Medio (MAE) igual a **1.0715** (%) que expresa la existencia de un grado de error promedio MUY BAJO y R-SQUARED igual a **0.9342** como coeficiente de determinación EXCELENTE. Esto indica que la técnica de regresión de Machine Learning de mejor rendimiento fue **Random Forest Regressor** y según los resultados y análisis de las tablas 39, 40 y 44 aumenta la eficiente del modelo para para predecir el resultado del ensayo de OMC.

5.3. Resultado del objetivo general

5.3.1. Objetivo general

OG: Aumentar la eficiencia de un modelo predictivo de las propiedades mecánicas de los suelos basado en técnicas de Machine Learning.

Según los resultados obtenidos de los objetivos 1, 2 y 3 la técnica de regresión de Machine Learning que mejor resultado tuvo en la predicción de los resultados de los ensayos de CBR, MDD y OMC fue **Random Forest Regressor**. Esto indica que la eficiencia del modelo predictivo de las propiedades mecánicas de los suelos aumenta con las técnicas de Machine Learning basado en Random Forest Regressor.

CAPITULO VI

DISCUSIÓN

Los resultados de este estudio, donde se utilizó 3294 datos de muestras de suelos y que se centró en técnicas como Enssemble Trees (Extra Treess Regressor y Random Forest Regressor) y Redes Neuronales Artificiales, muestran que Random Forest Regressor, especialmente cuando se utiliza técnicas avanzadas de Auto-ML, supera a las otras técnicas en eficiencia para predecir las propiedades mecánicas de los suelos, ya que, además, de utilizar GridSearchCV hacen uso de técnicas innovadoras como el morfismo de red, la optimización bayesiana y entre otras técnicas que ya vienen implementadas en la herramienta Auto-WEKA. Logrando un valor de R-Squared de 0.9342 para OMC, 0.9696 para MDD y 0.9848 para CBR. En comparación con el estudio de Li et al. (2024) donde utilizó 168 datos de muestras de suelos y diferentes técnicas, como Máquina de Vectores de Soporte (SVM), Redes Neuronales Artificiales (ANN), Bosques Aleatorios (RF) y Árboles de Impulso de Gradiente Extremo (XGBoost), y encontró que XGBoost ofrecía las predicciones más precisas para MDD y OMC, logrando un valor de R-Squared de 0.9234 para OMC y 0.9098 para MDD. Se verifica que los resultados de la presente investigación superan significativamente a los resultados de los autores Li et al. (2024). Ambos estudios respaldan la efectividad de las técnicas de Ensemble Trees Regressions utilizando las configuraciones y optimizadores adecuados. Así mismo, en el rendimiento de las Redes Neuronales Artificiales, se respalda la importancia de considerar la variabilidad en los conjuntos de datos y las configuraciones de la arquitectura neuronal de los modelos, así como la necesidad de optimizar y explorar continuamente diferentes técnicas para mejorar la precisión de las predicciones.

En comparación con el estudio de Gowda et al. (2024), quienes investigaron el desarrollo de modelos predictivos de CBR utilizando 2191 datos de muestras de suelos y técnicas estadísticas y de Soft Computing. En la presente investigación se utilizaron 3294 datos de muestras de suelos y se muestra diferencias significativas en cuanto a las técnicas empleadas y los resultados obtenidos. Mientras que Gowda et al. se enfocan en el uso de técnicas como la regresión lineal multivariable (MLR), las redes neuronales artificiales (ANN) y el sistema de inferencia neuro-difuso adaptativo (ANFIS). En el presente estudio se evalúa modelos de Tree Ensembles y redes neuronales artificiales, junto con técnicas avanzadas de Auto-ML como Neural Architecture Search (NAS) y Combined Algorithm Selection and Hyperparameter Optimization (CASH).

Aunque ambos estudios reconocen la eficacia de las técnicas de Soft Computing en la predicción precisa del CBR, con un resultado sobresaliente con un valor de R-Squared de 0.81 obtenido por ANFIS en el estudio de Gowda et al., y en este trabajo por Random Forest Regressor con un valor de R-Squared de 0.9848 obtenido. Se demuestra la superioridad de los resultados obtenidos en el presente trabajo, así como también, las diferencias en las técnicas específicas y enfoques, demuestra la diversidad de métodos disponibles y la necesidad de evaluar múltiples opciones para encontrar la solución más adecuada en cada contexto de investigación.

Los resultados de la presente investigación coinciden con los hallazgos de Ofrikhter et al. (2022) en términos del uso de técnicas de Machine Learning para predecir las propiedades mecánicas del suelo. Pudiéndose verificar que la precisión de las técnicas de Machine Learning es superior a la de los métodos estadísticos conocidos. A diferencia de la presente investigación donde se utilizaron las técnicas de Ensamble Trees como el Random Forest Regressor y el Extra Trees Regressor que superan a las Redes Neuronales Artificiales, en la predicción de propiedades mecánicas del suelo como el CBR, la MDD y la OMC. Ofrikhter et al., se enfocó en el uso de las Redes Neuronales Artificiales para predecir las propiedades mecánicas de Cohesión Específica y el Ángulo de Fricción de un suelo arcilloso con un valor de MAPE de 15.33% y 9.1% respectivamente. Así mismo, al tratarse de diferentes propiedades mecánicas del suelo en ambos estudios y tener el mismo tipo de propiedades físicas. Para futuros estudios de investigación se podrían investigar las relaciones entre estas propiedades mecánicas, en consecuencia, tener modelos completos interrelacionados que abarquen más propiedades mecánicas del suelo y no tratarlas por separado y/o aislado.

Tanto la presente investigación como la de Taffese y Abegaz (2021) se centran en el uso de técnicas de inteligencia artificial para predecir propiedades mecánicas del suelo, donde utilizan 408 datos de muestras de suelos y la presente investigación 3294 datos de muestras de suelos. Taffese y Abegaz (2021) utilizan modelos de Machine Learning para predecir propiedades mecánicas del suelo, como el contenido de humedad óptimo (OMC), la densidad seca máxima (MDD) con un valor de MSE igual a 13.23 (%) y 21.201 (g/cm3) respectivamente, y en la presente investigación con un valor de MSE igual a 0.0003 (%) y 2.2759 (g/cm3). Se puede observar que los resultados de la presente investigación son significativamente superiores ya que poseen un

error demasiado bajo. En términos de los modelos utilizados, mientras que la presente investigación se enfoca en comparar diferentes técnicas de Machine Learning, como Random Forest Regressor y Redes Neuronales Artificiales, la investigación de Taffese y Abegaz se centra en modelos de Ensemble of Regression Trees, los cuales muestran un rendimiento superior a los modelos de Redes Neuronales Artificiales en su estudio. Sin embargo, ambos estudios llegan a la conclusión de que los modelos basados en inteligencia artificial pueden mejorar la predicción de propiedades del suelo, lo que tiene implicaciones importantes para la planificación y construcción de infraestructuras de manera más eficiente y económica.

La presente investigación, al igual que la de Diaz et al. (2021), se suma al creciente interés en la aplicación de técnicas de Machine Learning en la geotecnia y mecánica de suelos. Ambos estudios se enfocan en correlacionar propiedades del suelo, pero mientras que la presente investigación se centra en la correlación y predicción de parámetros mecánicos como el contenido de humedad óptimo (OMC), la máxima densidad seca (MDD) y la capacidad portante del suelo (CBR), el estudio de Diaz et al. se enfoca en correlacionar y predecir parámetros físicos como el límite líquido obtenido por diferentes métodos. Ambos estudios demuestran la eficacia de las técnicas de Machine Learning para mejorar la precisión de las predicciones, lo que subraya la importancia de estas herramientas para optimizar la predicción de las propiedades del suelo de manera más precisa y eficiente. Además, ambos estudios destacan la necesidad de utilizar una gran cantidad de conjuntos de datos variados para garantizar la robustez y la aplicabilidad de los modelos en diferentes regiones geográficas.

El presente trabajo de investigación se alinea con el enfoque de Nguyen et al. (2020), quienes también exploraron el potencial de las redes neuronales artificiales (ANN) para predecir las propiedades mecánicas del suelo. Mientras que su estudio se centró en predecir la propiedad mecánica del Ángulo de Fricción Interna del Suelo con resultados de un valor de R = 0,8885, RMSE = 0,0442 (rad) y MAE = 0,0328 (rad) para un conjunto de datos de 145 muestras. La presente investigación se enfoca en la predicción de parámetros mecánicos como el óptimo contenido de humedad, la máxima densidad seca y la capacidad portante del suelo con 3294 datos de muestras de suelos. Ambos estudios resaltan la importancia de los datos en la construcción de modelos predictivos confiables, destacando la necesidad de recopilar datos adicionales para

mejorar la precisión de los algoritmos. Este respaldo en los hallazgos resalta la importancia y efectividad de las técnicas de Machine Learning para tratar desafíos en geotecnia y mecánica de suelos.

La presente investigación se relaciona con el estudio de Pham et al. (2019) en el uso de inteligencia artificial para predecir propiedades mecánicas del suelo mediante las propiedades físicas del suelo. Ambos estudios comparten el objetivo general de desarrollar modelos predictivos utilizando diferentes técnicas de inteligencia artificial y datos de suelos. Mientras que su investigación se centró en modelos como Redes Neuronales Artificiales (ANN), Sistema de Inferencia Difusa basado en Red Adaptativa (ANFIS) y Máquina de Vector de Soporte (SVM) para predecir el Coeficiente de Compresión (Cc), la presente investigación se enfocó en la predicción de propiedades como el CBR, el OMC y la MDD utilizando técnicas como Extra Trees Regressor, Random Forest Regressor y Redes Neuronales Artificiales. La presente investigación da soporte a Pham et al., en la importancia de la técnica de preprocesamiento del análisis de multicolinealidad en los datos, que consiste en identificar los parámetros más representativos para la predicción precisa, así como la necesidad de técnicas de preprocesamiento de datos para mejorar la calidad de las predicciones que se aplicaron en el presente estudio. Estos hallazgos dan importancia al valor de las técnicas de inteligencia artificial en la predicción de propiedades mecánicas del suelo.

El presente trabajo de investigación guarda similitudes con la tesis de Valderrama Purizaca (2020) en el uso de técnicas de Machine Learning para predecir las propiedades mecánicas del suelo, aunque difieren en el conjunto de datos, ya que en la presente investigación se utilizó 3294 datos de muestras de suelos de todos los departamentos del Perú y dicho autor utilizó solamente 285 datos de muestras de suelos del departamento de Lambayeque. Así como también, difieren en las propiedades específicas analizadas, ya que además de utilizar las propiedades físicas de análisis granulométrico y límites de consistencia, en el presente estudio se incluyeron 2 variables adicionales como la clasificación de los suelos (SUCS) y el departamento de cada ensayo, los cuales fueron codificados en variables numéricas y se comprobó que tienen más correlación con las variables objetivo que algunas propiedades físicas. Mientras que Valderrama Purizaca adicionó las propiedades como el contenido de humedad y el contenido de sales. Así mismo, Valderrama

Purizaca se enfocó en probar una gran cantidad de modelos de redes neuronales artificiales con la técnica de selección y validación de modelos de GridSearchCV, en la presente investigación, se utilizaron las técnicas como Extra Trees Regressor, Random Forest Regressor y Redes Neuronales Artificiales, donde se demuestra la eficiencia de dichos modelos con el uso de técnicas avanzadas de Auto-ML que además de incluir el GridSearchCV, incluyen técnicas avanzadas y procesos de optimización para no generar y buscar una gran cantidad de configuraciones arquitecturas e hiperparámetros de las redes neuronales artificiales. Los mejores modelos basados en redes neuronales artificiales de dicho autor alcanzaron un MAE, para el MDD (0.035 g/cm3), el OMC (1.13 %), el CBR95 (1.65 %) y le CBR100 (2.79 %). En el presente trabajo se obtuvo los mejores modelos basado en Random Forest Regressor con un MAE, para el MDD (0.034 g/cm3), el OMC (1.07 %), el CBR95 (1.88 %) y le CBR100 (5.09%). Se verifica que los resultados de la presente investigación son mínimamente superiores en la predicción de las propiedades de OMC Y MDD y mínimamente inferiores en la predicción de las propiedades de CBR95 Y CBR100. Además, que para esos resultados dicho autor solo considera una pequeña cantidad de muestras de suelo del departamento de Lambayeque, en el presente estudio se pudo minimizar el error para una gran cantidad de datos de tipos de suelos de todo el Perú, que obviamente generaliza mejor las predicciones. Estos hallazgos respaldan la eficacia de las técnicas de Machine Learning en la predicción de propiedades del suelo, lo que tiene implicaciones significativas en las predicciones de las propiedades mecánicas del suelo de todos los departamentos del Perú.

El presente trabajo de investigación guarda similitudes con la tesis de Manrique Dávila (2019) en el uso de técnicas de Machine Learning para predecir las propiedades mecánicas del suelo, aunque difieren en el conjunto de datos, ya que en la presente investigación se utilizó 3294 datos de muestras de suelos de todos los departamentos del Perú de 311 expedientes técnicos de estudios y proyectos definitivos de construcción del Sistema Electrónico de Contrataciones del Estado (SEACE 3.0) y dicho autor utilizó solamente 222 datos de las propiedades mecánicas y físicas del suelo de 5 expedientes técnicos de estudios y proyectos definitivos realizados por Provias Nacional. Así como también, difieren en las propiedades específicas analizadas, ya que además de utilizar las propiedades físicas de análisis granulométrico y límites de consistencia, en el presente estudio se incluyeron 2 variables adicionales como la clasificación de los suelos (SUCS) y el departamento de cada ensayo. Mientras que Manrique Dávila no adicionó ninguna

variable. Así mismo, Manrique Dávila se enfocó en utilizar el software NeuralTools y probar una cantidad de modelos de redes neuronales artificiales, en la presente investigación, se utilizaron las técnicas como Extra Trees Regressor, Random Forest Regressor y Redes Neuronales Artificiales, donde se demuestra la eficiencia de dichos modelos con el uso de técnicas avanzadas de Auto-ML que además de incluir el GridSearchCV, incluyen técnicas avanzadas y procesos de optimización para no generar y buscar una gran cantidad de configuraciones arquitecturas e hiperparámetros de las redes neuronales artificiales. Los mejores modelos basados en redes neuronales artificiales de dicho autor alcanzaron un R-Squared de 0.8445 para OMC, 0.9265 para MDD, 0.8229 para CBR95 y 0.7317 para CBR100. En el presente trabajo se obtuvo los mejores modelos basado en Random Forest Regressor con un R-Squared de 0.9342 para OMC, 0.9696 para MDD, 0.9848 para CBR95 y 0.9429 para CBR100. Se verifica que los resultados de la presente investigación superan significativamente a los resultados de la autora Manrique Dávila. Además, tanto la presente investigación como la tesis de Manrique Dávila resaltan la importancia de la confiabilidad de los datos utilizados, lo que respalda la robustez de los modelos de Machine Learning en la predicción de las propiedades mecánicas del suelo para aplicaciones de geotecnia, mecánica de suelos, en ingeniería civil, estudio de taludes y diseño de infraestructuras.

CONCLUSIONES

- Se logró Aumentar la eficiencia de los modelos predictivos de las propiedades mecánicas de los suelos con las técnicas de Machine Learning basado en Random Forest Regressor. Además, se destaca el papel crucial de herramientas avanzadas de Auto-ML, como Auto-WEKA, en la mejora del rendimiento de los modelos.
- Se logró determinar que la técnica de Random Forest Regressor de Machine Learning es la que aumenta la eficiencia del modelo para para predecir el resultado del ensayo de CBR (California Bearing Ratio), con un Mean Squared Error de 9.3771 (%), un Root Mean Squared Error de 3.0622 (%), Mean Absolute Error de 1.8808 (%), R-Squared de 0.9848. Además, se concluye que tanto Extra Trees Regressor como Random Forest Regressor presentan un rendimiento superior para predecir el valor del CBR en comparación con Redes Neuronales Artificiales. Por último, la herramienta Auto-WEKA mejora aún más la eficiencia de Random Forest Regressor, convirtiéndolo en la técnica más eficiente para esta tarea.
- Se logró Determinar que la técnica de Random Forest Regressor de Machine Learning es la que aumenta la eficiencia del modelo para para predecir el resultado del ensayo de MDD (Máxima densidad seca), con un Mean Squared Error de 0.0003 (g/cm3), un Root Mean Squared Error de 0.0536 (g/cm3), Mean Absolute Error de 0.0340 (g/cm3), R-Squared de 0.9696. Además, se concluye que tanto Extra Trees Regressor como Random Forest Regressor presentan un rendimiento superior para predecir el valor del MDD en comparación con Redes Neuronales Artificiales. Por último, la herramienta Auto-WEKA mejora aún más la eficiencia de Random Forest Regressor, convirtiéndolo en la técnica más eficiente para esta tarea.
- Se logró Determinar que la técnica de Random Forest Regressor de Machine Learning es la que aumenta la eficiencia de un modelo para predecir el resultado del ensayo de OMC (Contenido óptimo de humedad), con un Mean Squared Error de 2.2759 (%), un Root Mean Squared Error de 1.5086 (%), Mean Absolute Error de 1.0715 (%), R-Squared de 0.9342. Además, se concluye que tanto Extra Trees Regressor como Random Forest Regressor presentan un rendimiento superior para predecir el valor del OMC en comparación con Redes Neuronales Artificiales. Por último, la herramienta Auto-WEKA mejora aún más la eficiencia de Random Forest Regressor, convirtiéndolo en la técnica más eficiente para esta tarea.

RECOMENDACIONES

- Se recomienda realizar la recolección de datos de investigaciones anteriores, expedientes técnicos, base de datos gubernamentales e institucionales, como el SEACE 3.0.
- Se recomienda guardar en una base de datos todos los ensayos de geotecnia y mecánica de suelos que se realizan en una universidad para contribuir en investigaciones futuras.
- Se recomienda profundizar en conocimientos de ciencia de datos, minería de datos, inteligencia artificial, aprendizaje máquina, geotecnia y mecánica de suelos.
- Aplicar técnicas de tratamiento de datos y preprocesamiento de datos para una buena predicción de las propiedades mecánicas del suelo.
- Se recomienda realizar la comparación de las diferentes técnicas o algoritmos de la inteligencia artificial que mejor rendimiento tengan en base a la literatura.
- Se recomienda utilizar el entorno de desarrollo Google Colaboratory como herramienta tecnológica de ciencia de datos e inteligencia artificial, ya que brinda un entorno amigable, librerías preinstaladas, acceso gratuito de recursos de hardware y permite la programación en la nube.
- Se sugiere utilizar el lenguaje de programación Python u otros que permitan realizar todas las etapas de ciencia de datos e inteligencia artificial. Así como también, analizar las ventajas y desventajas del lenguaje de programación con relación al propósito de la investigación.
- Se recomienda utilizar la metodología KDD ya que incluye etapas y procesos que permiten abordar el problema de Machine Learning de una forma ordenada y estructurada de principio a fin.
- Para investigaciones futuras, realizar investigaciones en el que se incluyan más propiedades mecánicas y físicas de los suelos para la predicción de propiedades similares o diferentes al presente trabajo de investigación.
- Para investigaciones futuras, investigar las correlaciones de las diferentes propiedades mecánicas de los suelos con propiedades no vistas en el presente trabajo y lograr integrar modelos más completos que incluyan más propiedades mecánicas del suelo.
- Para investigaciones futuras, se recomienda desarrollar una aplicación digital para predecir las propiedades mecánicas de los suelos del Perú que se ajuste a los requerimientos de los clientes de la industria de la construcción y geotecnia.

REFERENCIAS BIBLIOGRÁFICAS

Alaiad, A., Migdady, A., Al-Khatib, R., Alzoubi, O., Zitar, R., & Abualigah, L. (2023). *AutoKeras approach: A robust automated deep learning network for diagnosis disease cases in medical images.* Journal of Imaging, 9(3), 64. https://doi.org/10.3390/jimaging9030064

Araujo Navarro, W. S. (2014). Ecuaciones de correlación de CBR con propiedades índice de suelos. *Tesis para optar el Título de Ingeniero Civil.* Universidad de Piura, Piura. https://pirhua.udep.edu.pe/bitstream/handle/11042/2192/ICI_208.pdf?sequence=1

Centeno Franco, A. (2019). Deep Learning. *Trabajo fin de grado.* Universidad de Sevilla, Sevilla. https://idus.us.es/handle/11441/90004

Cordova Valentin, K. H., & Mori Montalvo, A. F. (2021). Modelo matemático para la predicción de la Capacidad de Soporte (CBR) en suelos expansivos estabilizados con cenizas de cáscara de arroz y cal a partir de sus propiedades índice y de compactación. *Tesis para optar el título profesional de Ingeniero Civil.* Universidad Peruana de Ciencias Aplicadas, Lima. Retrieved Diciembre 4, 2022, from http://hdl.handle.net/10757/657405

Crespo Villalaz, C. (2004). *Mecanica de suelos y cimentaciones.* Mexico: LIMUSA, 2004. 968-18-6489-1.

Dark, S. (2019). *Aprendizaje Automático La guía para Principiantes para comprender el Aprendizaje Automático* (1 ed.). (K. Gill, Ed.) España: Editorial Karanvir Gill. https://es.b-ok.lat/book/11343378/11dca3

Das, B. M. (2012). *Fundamentos de Ingenieria de Cimentaciones* (Séptima ed.). Cengage Learning Editores SA de CV. https://bookshelf.vitalsource.com/books/9786074818239

Day, R. W. (2010). *Foundation Engineering Handbook: Design and Construction with the 2009 International Building Code* (Segunda ed.). New York: The McGraw-Hill Companies, Inc. https://www.accessengineeringlibrary.com/content/book/9780071740098

Diaz, E., Pastor, J., Á., R., & Tomas, R. (2021). *Machine learning techniques for relating liquid limit obtained by Casagrande cup and fall cone test in low-medium plasticity fine grained soils.* Eng. Geol. 294. https://doi.org/106381

E.050. (2015). *Norma Técnica Peruana: Suelos y Cimentaciones.* Lima: MACRO, 2015.

El Peruano. (2022). *Por qué estudiar los suelos de las áreas urbanas.* Diario Oficial del Estado del Perú. Retrieved Diciembre 2, 2022, from https://www.elperuano.pe/noticia/199148-por-que-estudiar-los-suelos-de-las-areas-urbanas

El-Amir, H., & Hamdy, M. (2020). *Deep Learning Pipeline: Building a Deep Learning Model with TensorFlow.* Apress. https://doi.org/10.1007/978-1-4842-5349-6

Galán Cortina, V. (2015). Aplicación de la metodología CRISP-DM a un proyecto de minería de datos en el entorno universitario. *Trabajo Fin de Grado, Universidad Carlos III de Madrid, Escuela Politécnica Superior, Ingeniería en Informática.* https://e-archivo.uc3m.es/rest/api/core/bitstreams/714c5452-962e-44cf-993f-ebb3088d4aa5/content

Gao, J. (2009). *Digital Analysis of Remotely Sensed Imagery* (Primera ed.). New York: The McGraw-Hill Companies. https://www.accessengineeringlibrary.com/content/book/9780071604659

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (Segunda ed.). O'Reilly Media.

Gonzáles, L. (2019). *Machine Learning con Python - Aprendizaje Supervizado* (1 ed.). Caracas, Venezuela: Editorial Santillana. https://escuela.aprendeia.com/cursos/machine-learning-con-python-aprendizaje-supervisado/#tab-course-section__overview

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* Massachusetts Institute of Technology.

Gopal, M. (2019). *Applied Machine Learning. 1st ed.* New York: McGraw-Hill Education. Retrieved Setiembre 15, 2022, from https://www.accessengineeringlibrary.com/content/book/9781260456844

Gowda, S., Kunjar, V., Gupta, A., Kavitha, G., Shukla, B., & Sihag, P. (2024). *Prediction of the Subgrade Soil California Bearing Ratio Using Machine Learning and Neuro-Fuzzy Inference System Techniques: A Sustainable Approach in Urban Infrastructure Development.* Urban Science, 8(1), 4. https://doi.org/10.3390/urbansci8010004

Hernández Sampieri, R., Fernández Collado, C., & Bapstista Lucio, P. (2014). *Metodología de la Investigación* (Sexta ed.). México DF, México: INTERAMERICA EDITORES. Retrieved 10 06, 2021, from

https://academia.utp.edu.co/grupobasicoclinicayaplicadas/files/2013/06/Metodología-de-la-Investigación.pdf

Hottong, M., Sperling, M., & Müller, C. (2024). *Using Auto-ML on Synthetic Point Cloud Generation.* Applied Sciences, 14(2), 742. https://doi.org/10.3390/app14020742

INEI. (2021). *Producto Bruto Interno creció en el segundo trimestre de 2021 impulsado por la demanda interna.* Retrieved Diciembre 2, 2022, from https://m.inei.gob.pe/prensa/noticias/producto-bruto-interno-crecio-en-el-segundo-trimestre-de-2021-impulsado-por-la-demanda-interna-13069/

Instituto Geofísico del Perú. (2018). *IGP realiza estudios en suelos de Apurímac tras rajaduras en paredes de viviendas.* IGP Instituto Geofísico del Perú. Retrieved Diciembre 2, 2022, from https://www.gob.pe/institucion/igp/noticias/74318-igp-realiza-estudios-en-suelos-de-apurimac-tras-rajaduras-en-paredes-de-viviendas

Jin, H., Song, Q., & Hu, X. (2019). *Auto-Keras: An efficient neural architecture search system.* ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1808-1816. https://doi.org/arXiv:1806.10282

Juárez Badillo, E., & Rico Rodríguez, A. (2012). *Mecánica de suelos I: Fundamentos de la mecánica de suelos.* México: Limusa.

Konasani, V. R., & Kadre, S. (2021). *Machine Learning and Deep Learning Using Python and TensorFlow.* New York: McGraw Hill. Retrieved Setiembre 20, 2022, from https://www.accessengineeringlibrary.com/content/book/9781260462296

Kotthoff, L., Thornton, C., Hoos, H., Hutter, F., & Leyton-Brown, K. (2017). *Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA.* Journal of Machine Learning Research, 18(1), 1-5.

La Cámara. (2022). *Sector construcción acumula crecimiento de 0,74% entre enero y mayo 2022.* Revista Digital de la Cámara del Comercio de Lima. Retrieved Diciembre 2, 2022, from https://lacamara.pe/sector-construccion-acumula-crecimiento-de-074-entre-enero-y-mayo-2022/#:~:text=Contenido%20Patrocinado-,Sector%20construcci%C3%B3n%20acumula%20crecimiento%20de%200%2C74%25%20entre%20enero%20y,impulso%20de%20la%20inversi%C3%B3n%20priva

Li, B., You, Z., Ni, K., & Wang, Y. (2024). *Prediction of Soil Compaction Parameters Using Machine Learning Models.* Applied Sciences (2076-3417), 14(7), 2716. https://doi.org/10.3390/app14072716

LINEA DE ESTADÍSTICA UPC. (2017). *CUADERNO DE TRABAJO-MA444-2017-01*. Repositorio académico UPC. https://repositorioacademico.upc.edu.pe/handle/10757/270366

Manrique Dávila, M. (2019). Predicción de propiedades mecánicas de suelos utilizando redes neuronales artificiales. *Tesis de grado.* Universidad Continental. https://hdl.handle.net/20.500.12394/7191

Marsh McLennan. (2021, Octubre 4). *El Futuro de la Construcción*. Retrieved Noviembre 20, 2022, from Marsh: https://www.marsh.com/co/industries/construction/insights/the-future-of-construction.html

MTC. (2016). *Manual de ensayos de materiales.* Lima. Retrieved Diciembre 04, 2022, from https://portal.mtc.gob.pe/transportes/caminos/normas_carreteras/documentos/manuales/Manual%20Ensayo%20de%20Materiales.pdf

Nguyen, T.-A., Ly, H.-B., & Pham, B. T. (2020). *Backpropagation Neural Network-Based Machine Learning Model for Prediction of Soil Friction Angle.* Hindawi Mathematical Problems in Engineering Volume 2020, Article ID 8845768, 11 pages. https://doi.org/10.1155/2020/8845768

NTP 339.127. (2006). *Método de ensayo para determinar el Contenido de* . Lima.

NTP 339.128. (2006). *Método de Ensayo para el análisis granulométrico.* Lima.

NTP 339.129. (2006). *Método de ensayo para determinar el límite líquido, límite plástico e índices de plasticidad del suelo.* Lima.

NTP 339.145. (2006). *Metodo de ensayo de CBR (Relacion de Soporte de California) de suelos compactados en el laboratorio.* Lima.

Ñaupas Paitán, H., Palacios Vilela, J. J., Valdivia Dueñas, M. R., & Romero Delgado, H. E. (2018). *Metodología de la investigación Cuantitativa - Cualitativa y Redacción de la Tesis* (Quinta ed.). Bogotá, Colombia: Ediciones de la U. Retrieved 10 2021, 06, from https://corladancash.com/wp-content/uploads/2020/01/Metodologia-de-la-inv-cuanti-y-cuali-Humberto-Naupas-Paitan.pdf

Ofrikhter, I., Ponomaryov, A., Zakharov, A., & Shenkman, R. (2022). *Estimation of soil properties by an artificial neural network.* Magazine of Civil Engineering. 2022. 110(2). Article No. 11011. https://doi.org/10.34910/MCE.110.11

Pham, B., Nguyen, M., Dao, D., Prakash, I., Ly, H.-B., Le, T.-T., & Ho, L. (2019). *Development of artificial intelligence models for the prediction of Compression Coefficient of soil: An application of Monte Carlo sensitivity analysis.* Sci. Total Environ. 679, 172–184. https://doi.org/10.1016/j.scitotenv.2019.05.061

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning* (2 ed.). Birmingham, Reino Unido: Editorial Packt Publishing . https://books.google.com.pe/books?id=_plGDwAAQBAJ&printsec=frontcover&dq=Python+Machine+Learning+Second+Edition+Machine+Learning+and+Deep+Learning+with+Python,+scikit-learn,+and+TensorFlow&hl=es&sa=X&redir_esc=y#v=onepage&q=Python%20Machine%20Learning%20S

Russell, R. (2018). *Machine Learning Guía Detallada para Implementar Algoritmos de Machine Learning con Python* (1 ed.). España: CreateSpace Independent Publishing Platform. https://www.amazon.com/-/es/Rudolph-Russell/dp/1720933685

Taffese, W., & Abegaz, K. (2021). *Artificial Intelligence for Prediction of Physical and Mechanical Properties of Stabilized Soil for Affordable Housing.* UNIVERSIDAD DE CÓRDOBA. Córdoba: Appl. Sci. 2021, 11, 7503. https://doi.org/10.3390/app11167503

Valderrama Purizaca, F. (2020). DESARROLLO DE UN MODELO DE REDES NEURONALES ARTIFICIALES PARA PREDECIR LAS PROPIEDADES MECÁNICAS DEL SUELO. *Tesis.* Universidad Señor de Sipán. https://hdl.handle.net/20.500.12802/8395

Zhang, P., Yin, Z., Jin, Y., & al., L. X. (2021). *Modelling the mechanical behaviour of soils using machine learning algorithms with explicit formulations.* Acta Geotech. 17, 1403–1422 (2022). https://doi.org/10.1007/s11440-021-01170-4

ANEXOS

Anexo 01. Código Fuente del Proyecto

Para acceder al código fuente y al conjunto de datos utilizados en la presente investigación, puedes contactar al correo electrónico proporcionado o consultar el siguiente enlace:

https://drive.google.com/drive/folders/1W5LPwN4bezJwl7pVBI6Ijm74RVs-ACRs?usp=sharing

Anexo 02. Ficha de registro de datos para el conjunto de datos

Los resultados de los ensayos de laboratorio fueron registrados en la siguiente tabla:

Muestra	Contenido de humedad (%)	Contenido de sales (%)	Análisis granulométrico (% que pasa)			Límites de Atterberg			Clasificación		Compactación		CBR	
			3 in	N° 4	N° 200	LL	LP	PI	SUCS	AASHTO	MDD	OMC	al 95%	al 100%
	1													
	2													
…														

Se da el formato que se crea conveniente para ingresar al programa Google Collaboratory. En este caso se utilizó la siguiente tabla:

Muestra	H	S	G	A	F	LL	PL	PI	SUCS	AASHTO	MDD	OMC	CBR95	CBR100
1														
2														
…														

Fuente: (Valderrama Purizaca, 2020)

Anexo 03. Ficha de registro de datos para el rendimiento de los modelos

PROPIEDAD	MSE	MAE	R
Propiedad 1 del suelo			
Algoritmo 1			
Algoritmo 2			
Algoritmo 3			
Propiedad 2 del suelo			
Algoritmo 1			
Algoritmo 2			
Algoritmo 3			
Propiedad 3 del suelo			
Algoritmo 1			
Algoritmo 2			
Algoritmo 3			

Fuente: (Valderrama Purizaca, 2020)

Anexo 04. Matriz de consistencia de investigación

<table>
<tr><th>PROBLEMA GENERAL</th><th>OBJETIVOS GENERAL</th><th>VARIABLES/DIMENSIONES</th><th>METODOLOGIA</th></tr>
<tr><td>¿De qué manera se puede aumentar la eficiencia de un modelo predictivo de las propiedades mecánicas de los suelos basado en técnicas de Machine Learning?</td><td>Aumentar la eficiencia de un modelo predictivo de las propiedades mecánicas de los suelos basado en técnicas de Machine Learning.</td><td rowspan="3">VARIABLE 1
Técnicas de Machine Learning
Dimensiones:
Modelos predictivos

VARIABLE 2
Propiedades mecánicas del suelo
Dimensiones:
1. Capacidad de soporte (CBR).
2. Máxima densidad seca (MDD).
3. Optimo contenido de humedad (OMC).</td><td rowspan="3">TIPO: Aplicada
Nivel: Investigación Descriptiva
Diseño: No experimental
Método: Cuantitativo
Población y muestra:
Población:
Muestras existentes de suelo a nivel mundial.
Muestra:
3294 muestras de suelo del Perú.
Tipo de muestra:
El tipo de muestra es no probabilística intencional.
Técnicas:
Recopilación de datos de Bases de datos, publicaciones científicas e informes.
Instrumentos:
Análisis de datos.
Formatos de recopilación de datos.
Método de análisis de datos:
Estadístico descriptivo e inferencial con apoyo de SPSS, MINITAB y Excel.</td></tr>
<tr><td>PROBLEMAS ESPECIFICOS</td><td>OBJETIVOS ESPECIFICOS</td></tr>
<tr><td>- ¿Cuál de las técnicas de Regresión de Machine Learning aumenta la eficiencia de un modelo para predecir el resultado del ensayo de CBR (California Bearing Ratio)?
- ¿Cuál de las técnicas de Regresión de Machine Learning aumenta la eficiencia de un modelo para predecir el resultado del ensayo de MDD (Máxima densidad seca)?
- ¿Cuál de las técnicas de Regresión de Machine Learning aumenta la eficiencia de un modelo para predecir el resultado del ensayo de OMC (Contenido óptimo de humedad)?</td><td>- Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para para predecir el resultado del ensayo de CBR (California Bearing Ratio).
- Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para para predecir el resultado del ensayo de MDD (Máxima densidad seca).
- Determinar la técnica de Regresión de Machine Learning que aumenta la eficiencia de un modelo para predecir el resultado del ensayo de OMC (Contenido óptimo de humedad).</td></tr>
</table>

Anexo 05. Capturas de pantalla adicionales de la aplicación de la metodología KDD

Figura 01. Búsqueda de expedientes técnicos en la plataforma SEACE 3.0

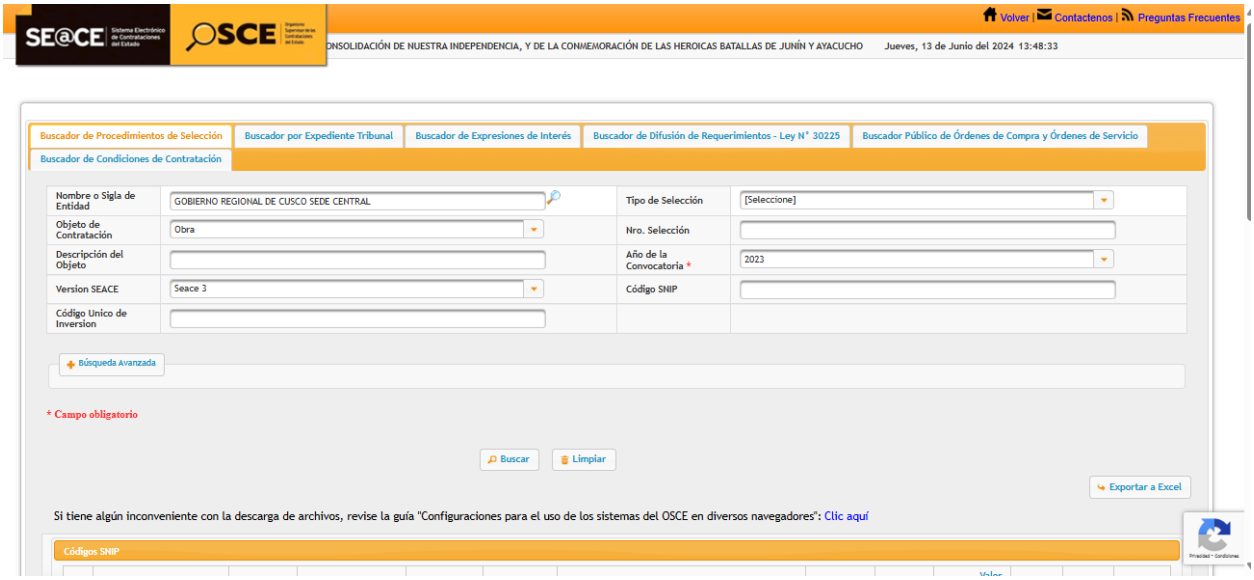
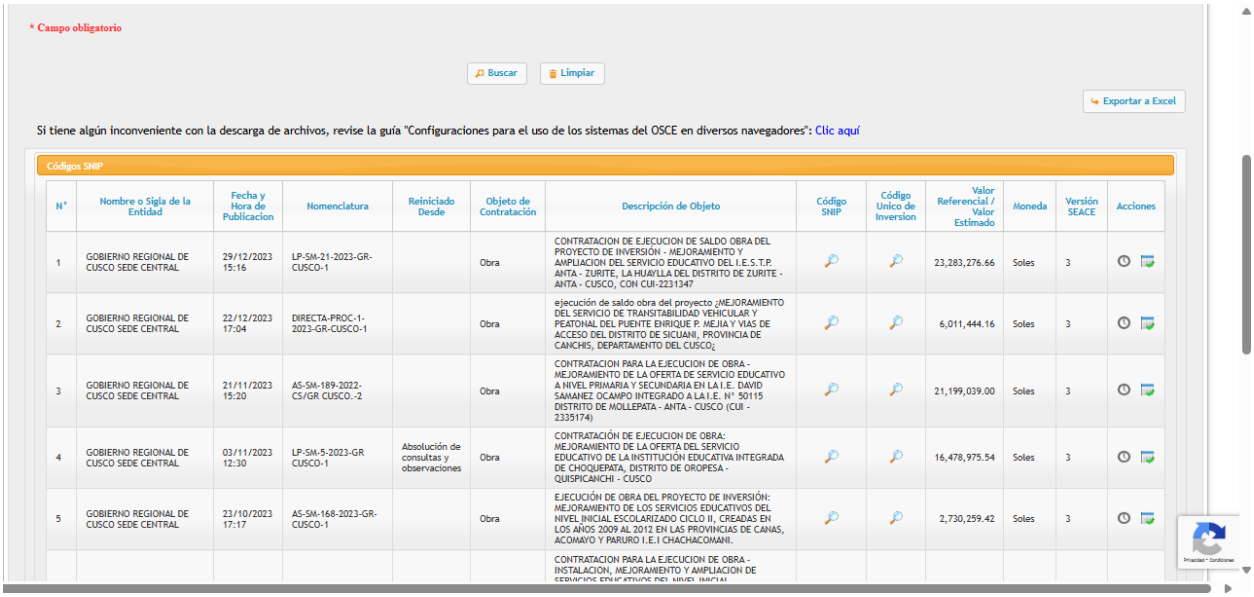


Figura 02. Resultados de la búsqueda de expedientes técnicos en la plataforma SEACE 3.0



N°	Nombre o Sigla de la Entidad	Fecha y Hora de Publicacion	Nomenclatura	Reiniciado Desde	Objeto de Contratación	Descripción de Objeto	Código SNIP	Código Unico de Inversion	Valor Referencial / Valor Estimado	Moneda	Versión SEACE	Acciones
1	GOBIERNO REGIONAL DE CUSCO SEDE CENTRAL	29/12/2023 15:16	LP-SM-21-2023-GR-CUSCO-1		Obra	CONTRATACION DE EJECUCION DE SALDO OBRA DEL PROYECTO DE INVERSIÓN - MEJORAMIENTO Y AMPLIACION DEL SERVICIO EDUCATIVO DEL I.E.S.T.P. ANTA - ZURITE, LA HUAYLLA DEL DISTRITO DE ZURITE - ANTA - CUSCO, CON CUI-2231347			23,283,276.66	Soles	3	
2	GOBIERNO REGIONAL DE CUSCO SEDE CENTRAL	22/12/2023 17:04	DIRECTA-PROC-1-2023-GR-CUSCO-1		Obra	ejecución de saldo obra del proyecto ¿MEJORAMIENTO DEL SERVICIO DE TRANSITABILIDAD VEHICULAR Y PEATONAL DEL PUENTE ENRIQUE P. MEJIA Y VIAS DE ACCESO DEL DISTRITO DE SICUANI, PROVINCIA DE CANCHIS, DEPARTAMENTO DEL CUSCO;			6,011,444.16	Soles	3	
3	GOBIERNO REGIONAL DE CUSCO SEDE CENTRAL	21/11/2023 15:20	AS-SM-189-2022-CS/GR CUSCO.-2		Obra	CONTRATACION PARA LA EJECUCION DE OBRA - MEJORAMIENTO DE LA OFERTA DE SERVICIO EDUCATIVO A NIVEL PRIMARIA Y SECUNDARIA EN LA I.E. DAVID SAMANEZ OCAMPO INTEGRADO A LA I.E. N° 50115 DISTRITO DE MOLLEPATA - ANTA - CUSCO (CUI - 2335174)			21,199,039.00	Soles	3	
4	GOBIERNO REGIONAL DE CUSCO SEDE CENTRAL	03/11/2023 12:30	LP-SM-5-2023-GR CUSCO-1	Absolución de consultas y observaciones	Obra	CONTRATACIÓN DE EJECUCION DE OBRA: MEJORAMIENTO DE LA OFERTA DEL SERVICIO EDUCATIVO DE LA INSTITUCIÓN EDUCATIVA INTEGRADA DE CHOQUEPATA, DISTRITO DE OROPESA - QUISPICANCHI - CUSCO			16,478,975.54	Soles	3	
5	GOBIERNO REGIONAL DE CUSCO SEDE CENTRAL	23/10/2023 17:17	AS-SM-168-2023-GR-CUSCO-1		Obra	EJECUCIÓN DE OBRA DEL PROYECTO DE INVERSIÓN: MEJORAMIENTO DE LOS SERVICIOS EDUCATIVOS DEL NIVEL INICIAL ESCOLARIZADO CICLO II, CREADAS EN LOS AÑOS 2009 AL 2012 EN LAS PROVINCIAS DE CANAS, ACOMAYO Y PARURO I.E.I CHACHACOMANI.			2,730,259.42	Soles	3	
						CONTRATACION PARA LA EJECUCION DE OBRA - INSTALACION, MEJORAMIENTO Y AMPLIACION DE						

Figura 03. Recopilación de expedientes técnicos en carpetas por cada departamento del Perú

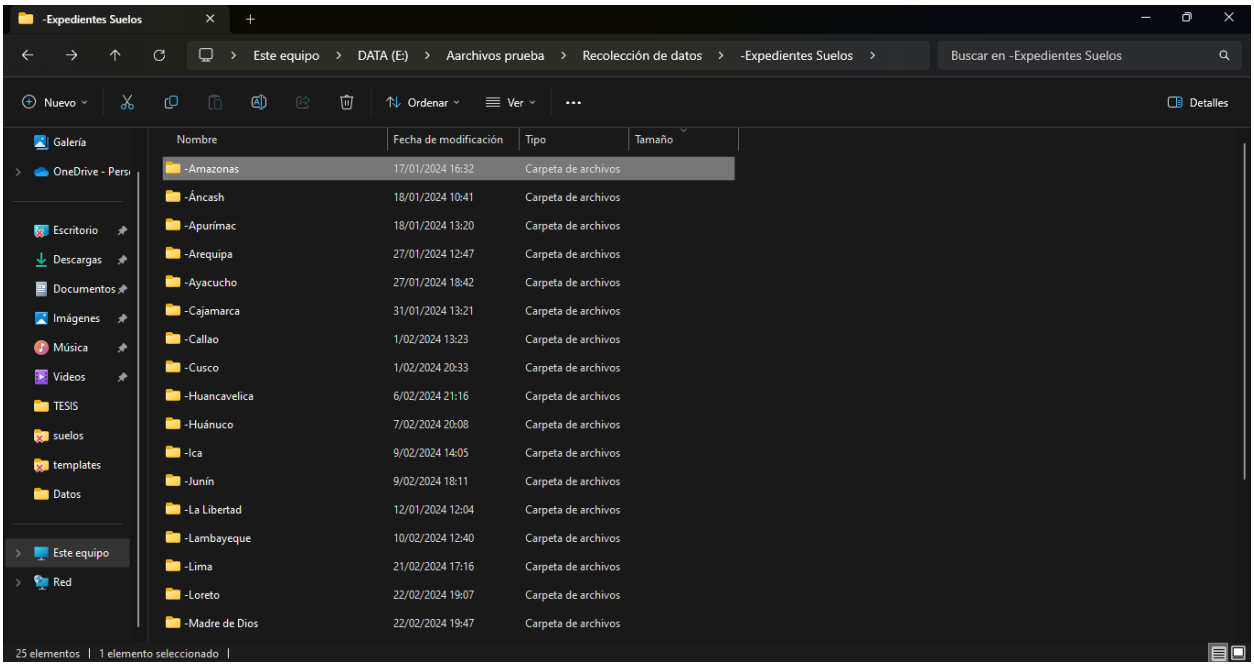


Figura 04. Recopilación de expedientes técnicos de cada carpeta de los 311 expedientes técnicos

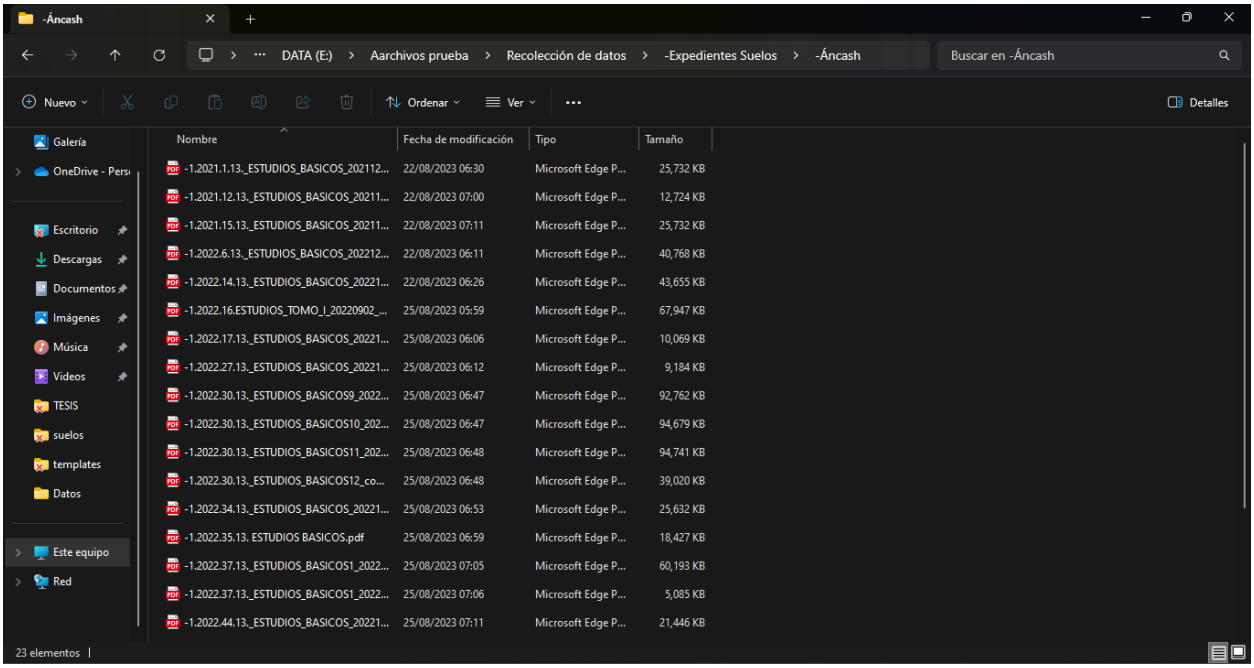


Figura 05. Muestra de un expediente técnico de 329 hojas

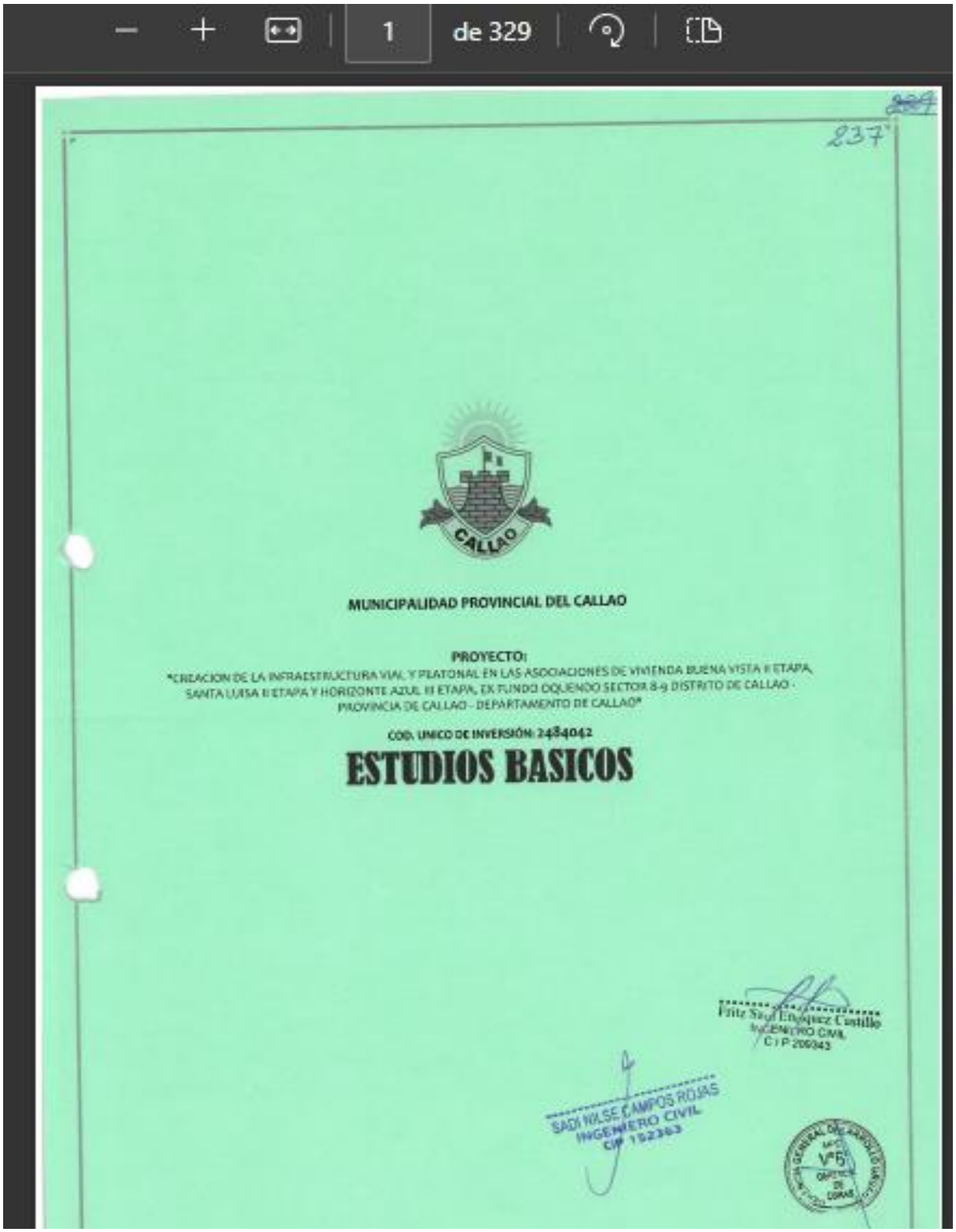


Figura 06. Muestra de un expediente técnico de los datos del análisis granulométrico

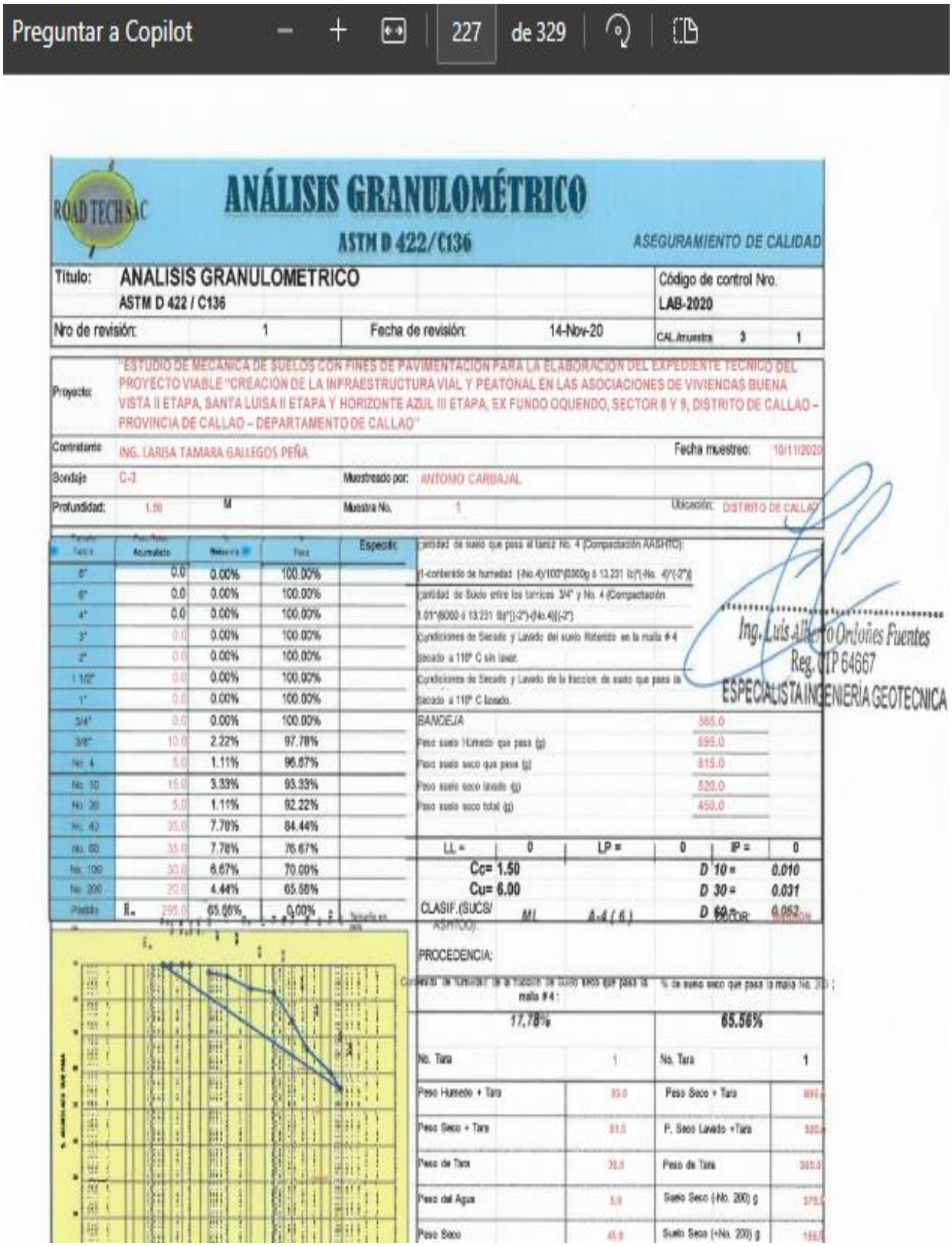


Figura 07. Muestra de un expediente técnico de los datos del OMC, MDD Y CBR

LABORATORIO DE MECÁNICA DE SUELOS, CONCRETO Y PAVIMENTOS

DG INGENIEROS S.A.C

ENSAYO DE CALIFORNIA BEARNING RATIO

A. DATOS GENERALES

1. Obra: "REPARACIÓN DE CARRETERAS DE ACCESO: EN LA VIA VECINAL CRUCE LAS CAÑAS – CHONZA LAGUNA, DISTRITO DE COPALLIN, PROVINCIA BAGUA, DEPARTAMENTO AMAZONAS"
2. Ubicación: CRUCE LAS CAÑAS – CHONZA LAGUNA — MATERIAL : IN SITU
3. Solicitante: ING. SAUL VASQUEZ RAMIRES — FECHA: 24/12/2019

DATOS DEL PROCTOR	
Humedad Optima (%)	8.60
Máxima Densidad Seca (gr/cm³)	2.158
0.95% M. D. S.	2.050
Tipo de Suelo (SUCS)	

DATOS DEL C.B.R.	
C.B.R. : 01"	58.00
C.B.R. al 95% de M.D.S. (%)	22.00
C.B.R. : 02"	54.72
C.B.R. al 95% de M.D.S. (%)	20.20

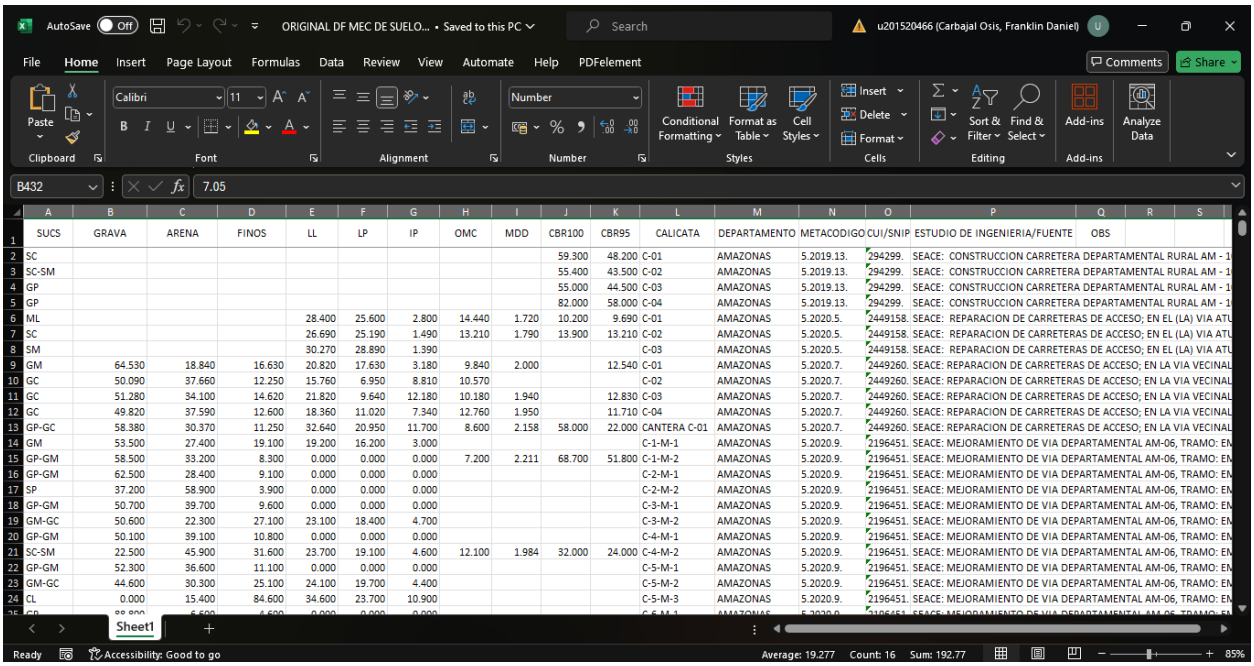
DGEOLAB

Nilton Cabrera Tuppia
INGENIERO CIVIL

DGEOLAB

Bach. Wilson S. Torrillo Delgado

Figura 08. Muestra del Excel de las 3294 muestras de suelos con sus respectivas características



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	SUCS	GRAVA	ARENA	FINOS	LL	LP	IP	OMC	MDD	CBR100	CBR95	CALICATA	DEPARTAMENTO	METACODIGO	CUI/SNIP	ESTUDIO DE INGENIERIA/FUENTE	OBS		
2	SC									59.300	48.200	C-01	AMAZONAS	5.2019.13.	294299.	SEACE: CONSTRUCCION CARRETERA DEPARTAMENTAL RURAL AM - 1			
3	SC-SM									55.400	43.500	C-02	AMAZONAS	5.2019.13.	294299.	SEACE: CONSTRUCCION CARRETERA DEPARTAMENTAL RURAL AM - 1			
4	GP									55.000	44.500	C-03	AMAZONAS	5.2019.13.	294299.	SEACE: CONSTRUCCION CARRETERA DEPARTAMENTAL RURAL AM - 1			
5	GP									82.000	58.000	C-04	AMAZONAS	5.2019.13.	294299.	SEACE: CONSTRUCCION CARRETERA DEPARTAMENTAL RURAL AM - 1			
6	ML				28.400	25.600	2.800	14.440	1.720	10.200	9.690	C-01	AMAZONAS	5.2020.5.	2449158.	SEACE: REPARACION DE CARRETERAS DE ACCESO; EN EL (LA) VIA ATU			
7	SC				26.690	25.190	1.490	13.210	1.790	13.900	13.210	C-02	AMAZONAS	5.2020.5.	2449158.	SEACE: REPARACION DE CARRETERAS DE ACCESO; EN EL (LA) VIA ATU			
8	SM				30.270	28.890	1.390					C-03	AMAZONAS	5.2020.5.	2449158.	SEACE: REPARACION DE CARRETERAS DE ACCESO; EN EL (LA) VIA ATU			
9	GM	64.530	18.840	16.630	20.820	17.630	3.180	9.840	2.000		12.540	C-01	AMAZONAS	5.2020.7.	2449260.	SEACE: REPARACION DE CARRETERAS DE ACCESO; EN LA VIA VECINAL			
10	GC	50.090	37.660	12.250	15.760	6.950	8.810	10.570				C-02	AMAZONAS	5.2020.7.	2449260.	SEACE: REPARACION DE CARRETERAS DE ACCESO; EN LA VIA VECINAL			
11	GC	51.280	34.100	14.620	21.820	9.640	12.180	10.180	1.940		12.830	C-03	AMAZONAS	5.2020.7.	2449260.	SEACE: REPARACION DE CARRETERAS DE ACCESO; EN LA VIA VECINAL			
12	GC	49.820	37.590	12.600	18.360	11.020	7.340	12.760	1.950		11.710	C-04	AMAZONAS	5.2020.7.	2449260.	SEACE: REPARACION DE CARRETERAS DE ACCESO; EN LA VIA VECINAL			
13	GP-GC	58.380	30.370	11.250	32.640	20.950	11.700	8.600	2.158	58.000	22.000	CANTERA C-01	AMAZONAS	5.2020.7.	2449260.	SEACE: REPARACION DE CARRETERAS DE ACCESO; EN LA VIA VECINAL			
14	GM	53.500	27.400	19.100	19.200	16.200	3.000					C-1-M-1	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
15	GP-GM	58.500	33.200	8.300	0.000	0.000	0.000	7.200	2.211	68.700	51.800	C-1-M-2	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
16	GP-GM	62.500	28.400	9.100	0.000	0.000	0.000					C-2-M-1	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
17	SP	37.200	58.900	3.900	0.000	0.000	0.000					C-2-M-2	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
18	GP-GM	50.700	39.700	9.600	0.000	0.000	0.000					C-3-M-1	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
19	GM-GC	50.600	22.300	27.100	23.100	18.400	4.700					C-3-M-2	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
20	GP-GM	50.100	39.100	10.800	0.000	0.000	0.000					C-4-M-1	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
21	SC-SM	22.500	45.900	31.600	23.700	19.100	4.600	12.100	1.984	32.000	24.000	C-4-M-2	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
22	GP-GM	52.300	36.600	11.100	0.000	0.000	0.000					C-5-M-1	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
23	GM-GC	44.600	30.300	25.100	24.100	19.700	4.400					C-5-M-2	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			
24	CL	0.000	15.400	84.600	34.600	23.700	10.900					C-5-M-3	AMAZONAS	5.2020.9.	2196451.	SEACE: MEJORAMIENTO DE VIA DEPARTAMENTAL AM-06, TRAMO: EM			

Figura 09. Muestra de la carpeta compartida en Google Drive de la codificación utilizada y los respectivos sets de datos

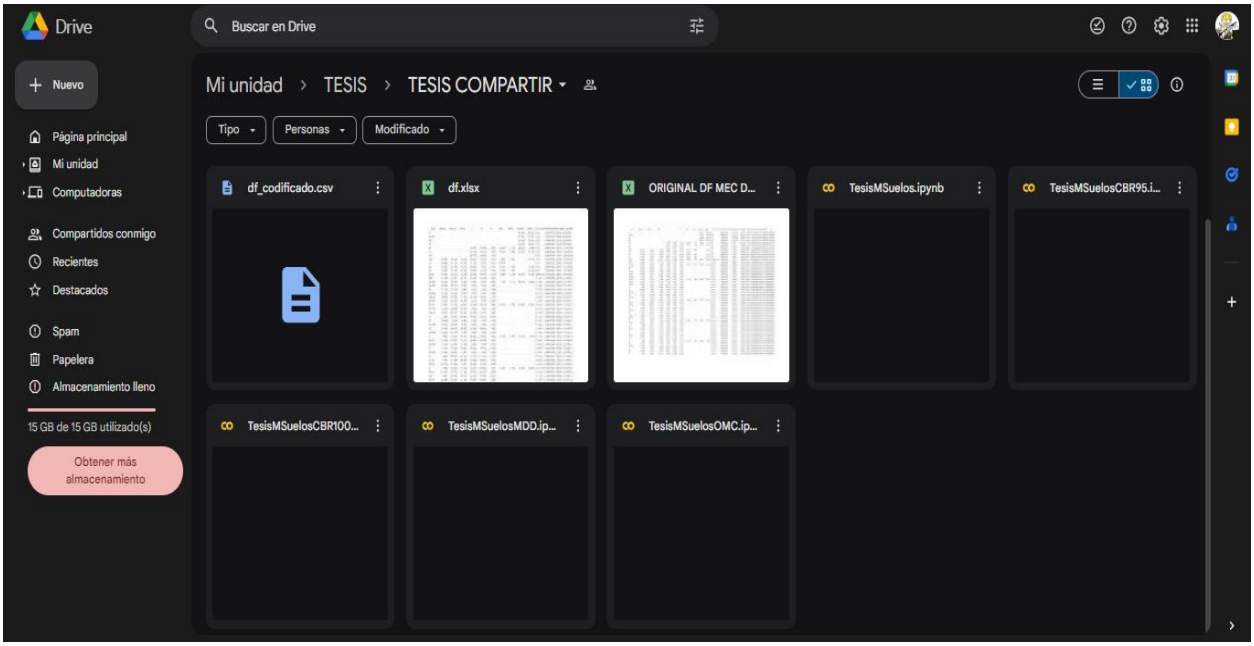


Figura 10. Muestra del archivo donde se encuentra el código para el preprocesamiento de datos

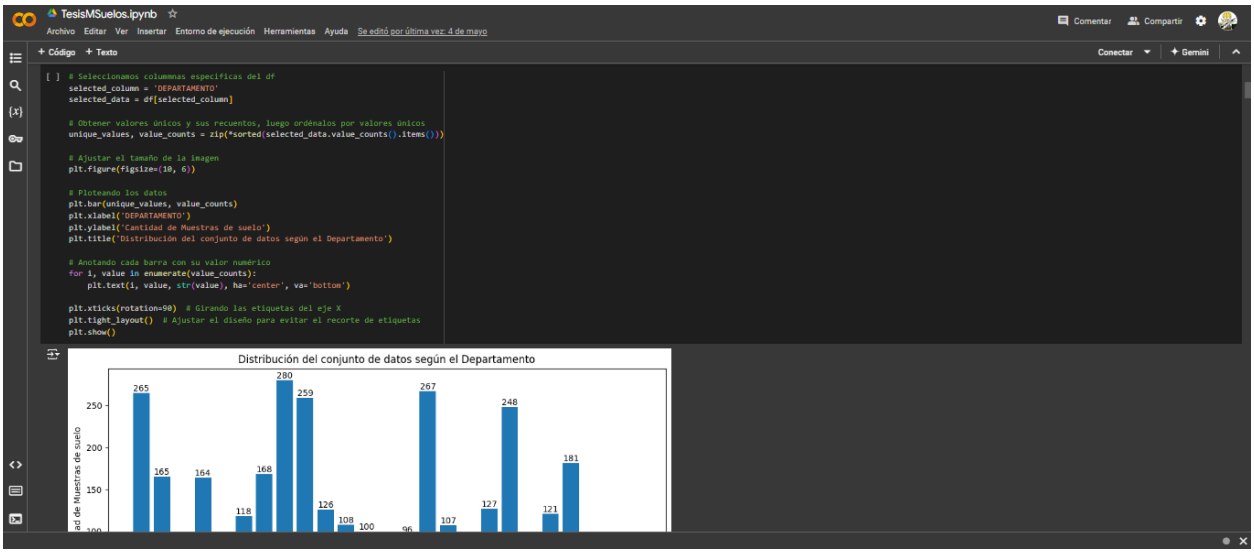


Figura 11. Muestra del archivo donde se encuentra el código para la minería de datos de la variable de salida CBR95, el cual es similar a las otras variables de salida

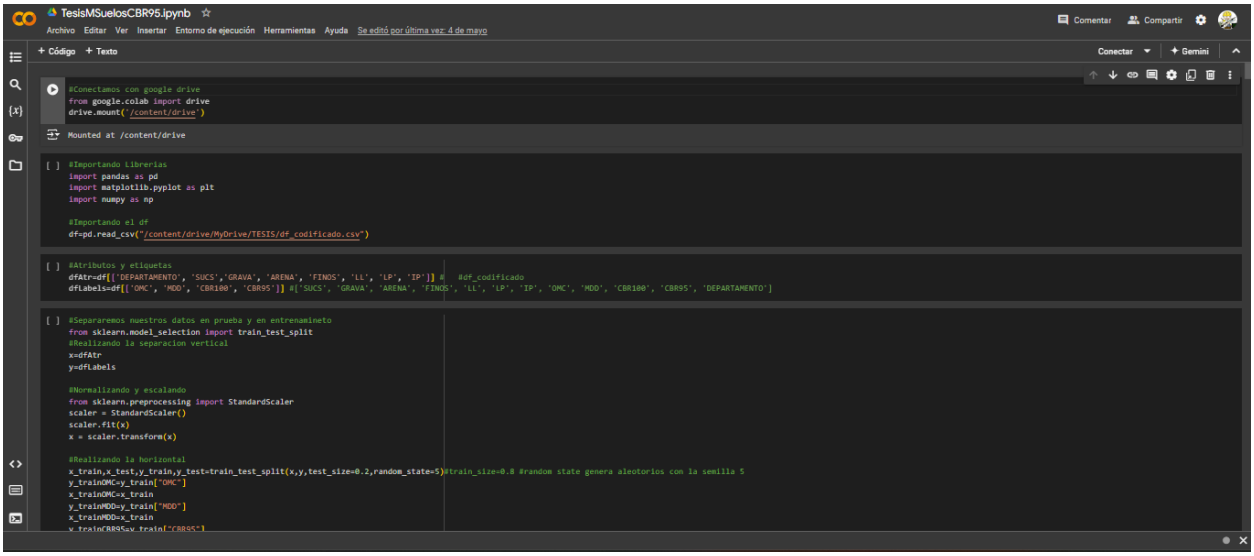


Figura 12. Muestra de la carpeta archivos de la aplicación para predecir las propiedades mecánicas de los suelos desarrollado con Flask, Html, JavaScript y Python.

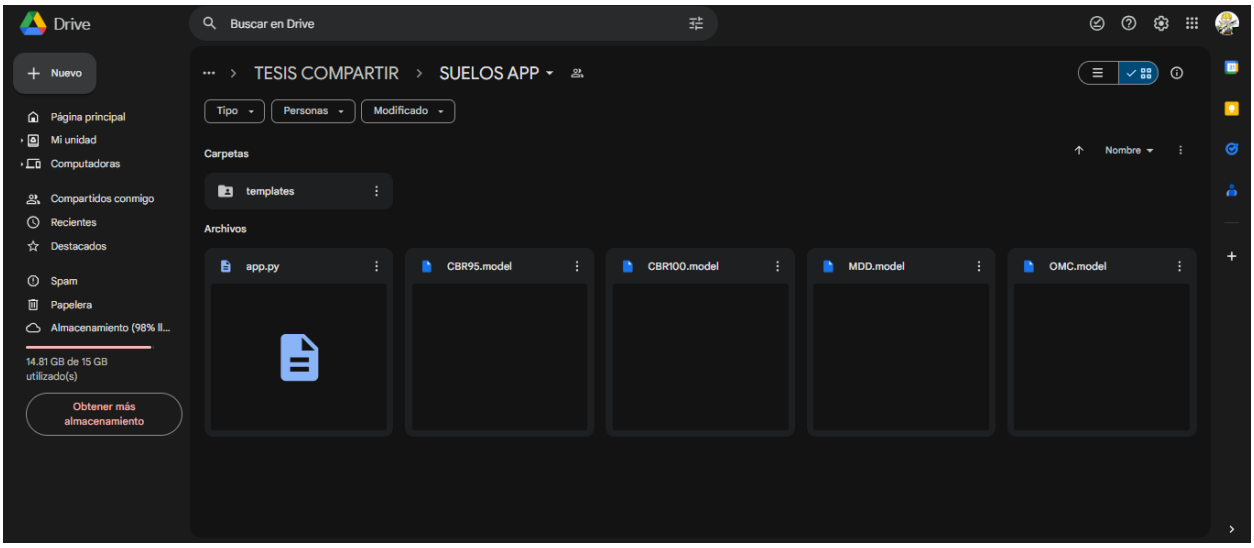


Figura 13. Muestra del archivo app.py donde se muestra parte del código fuente con Flask y Python.

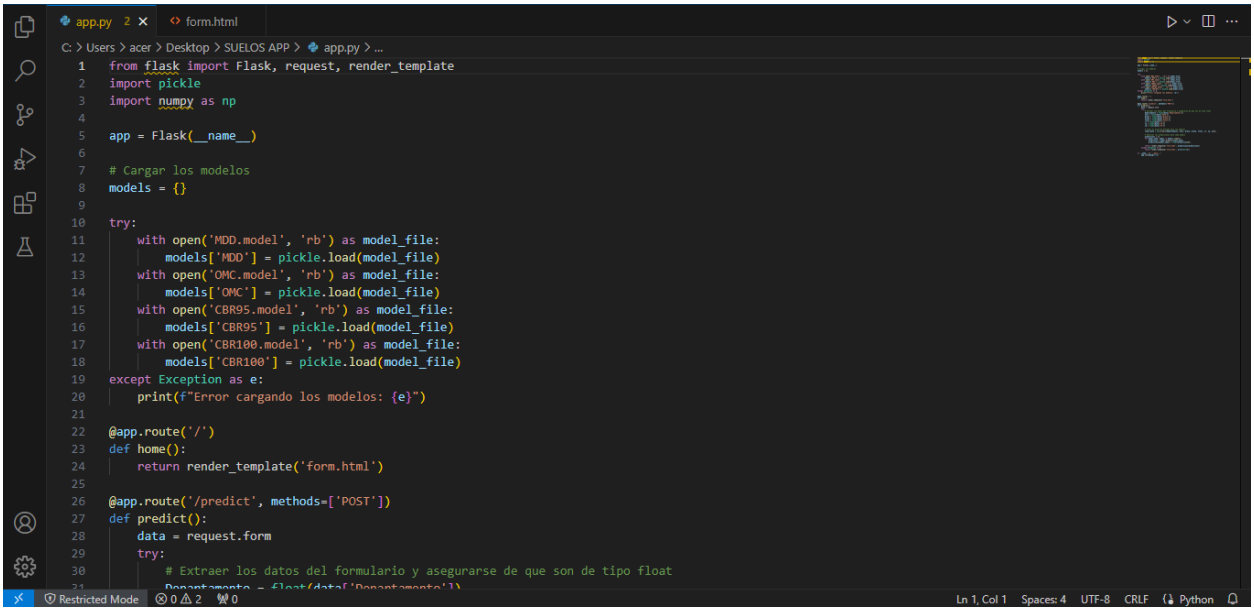


Figura 14. Muestra del archivo form.html donde se muestra parte del código fuente con html, css, Bootstrap y JavaScript.

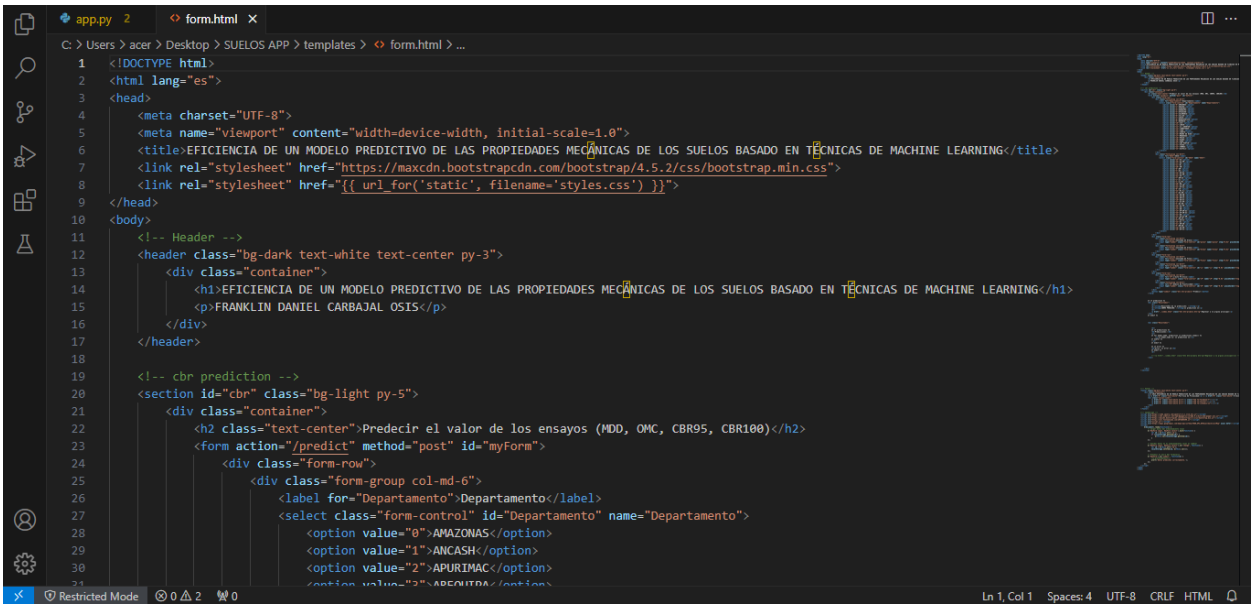


Figura 15. Muestra de la interfaz de usuario desarrollada para predecir las propiedades mecánicas de los suelos del departamento de Apurímac.

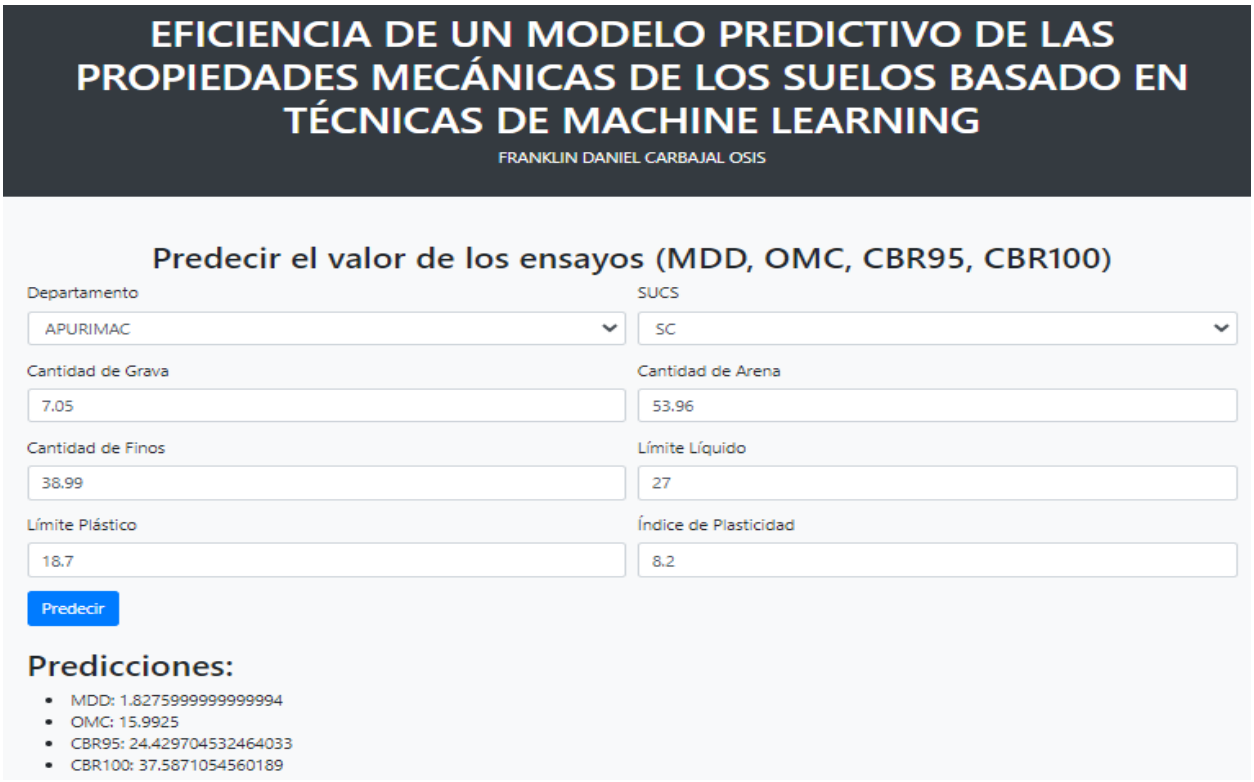


Figura 16. Muestra de la interfaz de usuario desarrollada para predecir las propiedades mecánicas de los suelos del departamento de Puno.