



UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

**DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA
INFORMACIÓN**

**MODELO DE CLASIFICACIÓN DE LITOLOGÍAS EMPLEANDO
APRENDIZAJE AUTOMÁTICO**

**TESIS PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN**

PRESENTA:

ROBERTO KARLO GOMEZ COLLADO

BAJO LA DIRECCIÓN DE:

DR. OSCAR ALBERTO CHÁVEZ BOSQUEZ

CUNDUACÁN, TABASCO, A: SEPTIEMBRE 2026



UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

**DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA
INFORMACIÓN**

**MODELO DE CLASIFICACIÓN DE LITOLOGÍAS EMPLEANDO
APRENDIZAJE AUTOMÁTICO**

**TESIS PARA OBTENER EL GRADO DE:
MAESTRO EN CIENCIAS DE LA COMPUTACIÓN**

PRESENTA:

ROBERTO KARLO GOMEZ COLLADO

BAJO LA DIRECCIÓN DE:

DR. OSCAR ALBERTO CHÁVEZ BOSQUEZ

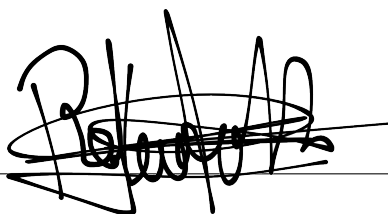
CUNDUACÁN, TABASCO, A: SEPTIEMBRE 2026

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán el día treinta del mes de Septiembre del año 2026, el que suscribe **Roberto Karlo Gomez Collado**, alumno del Programa de **Maestro en Ciencias de la Computación** con número de matrícula **232H21002**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención de Grado de Maestría y titulada **Modelo de clasificación de litologías empleando aprendizaje automático**, dirigida por el Dr. Oscar Alberto Chávez Bosquez

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco a 30 de Septiembre de 2026.



Estudiante: Roberto Karlo Gomez Collado

Cunduacán, Tabasco, a 19 de agosto de 2026
Oficio No. 1743/2026/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Roberto Karlo Gómez Collado

Egresado de la Maestría en Ciencias de la Computación

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional **"Modelo de clasificación de litologías empleando aprendizaje automático"**, para presentar examen y obtener el Grado de Maestro en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente


Dra. Laura Beatriz Vidal Turrubiates
Directora



C.c. p. Dr. Eduardo Hernández de la Cruz. - Encargado del despacho de la Coordinación de Posgrado.
Archivo

DRA.*LBVT/EHC

Av. Universidad s/n, Zona de la Cultura, Col. Magisterial,
Villahermosa, Centro, Tabasco, Mex. C.P. 86040.
Tel (993) 358 15 00 e-Mail: rectoria@ujat.mx

Carta de Cesión de Derechos

Villahermosa, Tabasco a 30 de Septiembre de 2026.

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Modelo de clasificación de litologías empleando aprendizaje automático.**

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención de que tenemos el derecho al reconocimiento como autores de la obra, y a la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR



Estudiante: Roberto Karlo Gomez Collado

TESTIGOS



Dr. Oscar Alberto Chávez Bosquez



Dr. José Hernández Torruco

Índice general

Índice de Figuras	IV
Índice de Tablas	V
Resumen	VI
Abstract	VII
1. Generalidades	1
1.1. Introducción	1
1.2. Planteamiento del problema	2
1.2.1. Definición del problema	2
1.2.2. Delimitación de la investigación	2
1.3. Preguntas de investigación e hipótesis	3
1.4. Objetivo general	3
1.5. Objetivos específicos	3
1.6. Justificación	4
1.7. Metodología utilizada	4
2. Marco teórico	6
2.1. Conceptos y teorías fundamentales de la investigación	6
2.1.1. Aprendizaje automático	6
2.1.2. La clasificación: una tarea del aprendizaje supervisado	7
2.1.3. Tipos de problemas de clasificación	7

2.1.4. Manejo de Desbalance de Clases	7
2.1.5. Algoritmos de clasificación	8
2.1.5.1. Support Vector Machine	8
2.1.5.2. Multi-Layer Perceptron (MLP)	9
2.1.5.3. Random Forest	10
2.1.6. Evaluación de modelos de clasificación	11
2.1.7. Partición y validación de datos	13
2.1.7.1. Partición de Datos	13
2.1.7.2. Validación Cruzada	13
2.2. Literatura relacionada	14
2.2.1. Estudios comparativos de algoritmos de clasificación	14
2.2.2. Enfoques metodológicos alternativos y limitaciones	16
2.3. Marco tecnológico	17
2.3.1. Bibliotecas	17
3. Modelo de clasificación de litologías	19
3.1. Descripción de los datos	19
3.1.1. Litologías	22
3.2. Tratamiento de datos faltantes	24
3.2.1. Tratamiento de datos atípicos	26
3.3. Selección y construcción de características	27
3.3.1. Eliminación de características no relevantes	27
3.3.2. Conjunto final de características	29
3.4. Procedimiento experimental	29
4. Experimentos y Resultados	31
4.1. Configuraciones de los modelos	32
4.1.1. Random Forest	32
4.1.2. Multi-Layer Perceptron	32
4.1.3. Support Vector Machine	33
4.2. Fase 1	33

4.2.1. Fase 2	34
4.3. Análisis de confusiones	36
4.4. Discusión	37
5. Contribuciones, conclusiones y trabajos futuros	39
5.1. Conclusiones	39
5.2. Contribuciones	41
5.3. Trabajos Futuros	42
Anexo	43
Bibliografía	44

Índice de figuras

3.1. Matriz de valores faltantes	25
4.1. F1 score por clase (fase 1)	34
4.2. F1-Score por clase para SVM (fase 1 vs fase 2)	35

Índice de tablas

2.1. Principales métricas para clasificación binaria	12
2.2. Matriz de confusión multiclase (3 clases) y ejemplo de cálculo para la clase 1. . . .	12
2.3. Enfoques para clasificación de litologías	16
3.1. Tabla de atributos litológicos	22
3.2. Distribución de litologías en el pozo 25_7-2	23
3.3. Matriz de correlación de Spearman	28
4.1. Frecuencia de mediciones por clase en entrenamiento y prueba	31
4.2. Configuración de Random Forest	32
4.3. Configuración de Support Vector Machine	33
4.4. F1-score por clase (fase 1)	33
4.5. F1-score por clase (fase 2)	35
4.6. Confusiones en clase 3	36
4.7. Confusiones de la clase 19	37

Resumen

El presente trabajo compara el desempeño de tres algoritmos de aprendizaje automático: *Support Vector Machine* (SVM), *Random Forest* (RF) y *Multi-Layer Perceptron* (MLP) en la clasificación de litologías. Los datos empleados son registros geofísicos del pozo petrolero 25_7-2, provenientes de la base de datos abierta de Geolink. Se emplea validación cruzada estratificada y partición de datos para evaluar el desempeño global y la capacidad de cada modelo para manejar clases desbalanceadas.

El proceso de clasificación se estructuró en dos fases, la estandarización de características se aplicó únicamente para SVM y MLP mediante *StandardScaler*, mientras que RF se entrenó con los datos sin normalizar debido a su efectividad ante diferencias de escala. En la primera fase, se evaluaron los clasificadores con su configuración predeterminada. En la Fase 2, se emplearon diferentes métodos de búsqueda según el algoritmo: *RandomizedSearchCV* con 50 combinaciones para RF, *GridSearchCV* con 30 combinaciones para MLP, y *BayesSearchCV* con 50 iteraciones para SVM.

RF alcanzó el mejor desempeño con configuración predeterminada ($F1_{\text{macro}} = 0.83$), mientras que SVM requirió optimización, mejorando de 0.74 a 0.80. Este contraste evidencia que RF es inherentemente robusto ante configuraciones no ajustadas, mientras que SVM depende de la calibración de sus hiperparámetros C y γ para el problema específico. MLP mostró estabilidad después de optimización (+0.01), indicando que su arquitectura básica ya está cerca de su capacidad máxima.

La hipótesis se confirma, ya que RF presenta mayor robustez inicial y no requiere normalización, factores que facilitan su aplicación directa en problemas de clasificación litológica.

Palabras clave: Geophysical logs, oil wells, Python, classification models.

Abstract

This paper compares the performance of three machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), and Multi-Layer Perceptron (MLP) in lithology classification. The data used are geophysical logs from oil well 25_7-2, obtained from the open database Geolink. Stratified cross-validation and data partitioning are used to evaluate the overall performance and the ability of each model to handle unbalanced classes.

The classification process was structured in two experimental phases for each of the three algorithms evaluated. Feature standardization was applied only to SVM and MLP using StandardScaler, while RF was trained with the unnormalized data due to its invariance to scale differences. In the first phase, the classifiers were evaluated with their default configuration. In Phase 2, different search methods were employed depending on the algorithm: RandomizedSearchCV with 50 combinations for RF, GridSearchCV with 30 combinations for MLP, and BayesSearchCV with 50 iterations for SVM.

RF achieved the best performance with the default configuration ($F1_{\text{macro}} = 0.83$), while SVM required optimization to become competitive, improving from 0.74 to 0.80. This contrast demonstrates that RF is inherently robust to untuned configurations, while SVM critically depends on the calibration of its hyperparameters C and γ for the specific problem. MLP showed stability with marginal changes after optimization (+0.01), suggesting that its basic architecture is already operating near its maximum capacity.

The hypothesis is confirmed, since RF presents greater initial robustness and does not require normalization, factors that facilitate its direct application in lithological classification problems.

Keywords: Geophysical logs, oil wells, Python, classification models.

Capítulo 1

Generalidades

1.1. Introducción

La clasificación de litologías mediante registros geofísicos ha mejorado con la incorporación de técnicas de aprendizaje automático. Este cambio se traduce en procesar e interpretar con mayor eficacia los múltiples registros geofísicos que caracterizan las propiedades físicas del subsuelo, incluyendo mediciones de resistividad, densidad, radiactividad natural y propiedades acústicas, entre otros.

Los algoritmos de aprendizaje automático tienen la capacidad de identificar patrones complejos en datos multivariados, superando las limitaciones de los métodos estadísticos tradicionales para distinguir entre litologías con propiedades físicas similares. Esta capacidad es valiosa en la clasificación automatizada de litologías, tarea fundamental en la caracterización de yacimientos petroleros, donde la correcta identificación de rocas reservorio, sello y generadoras es crítica para decisiones de exploración y desarrollo.

El diseño de estos modelos debe abordar dos aspectos fundamentales: el desbalance natural entre diferentes tipos de litologías y la validación de su desempeño mediante estrategias apropiadas de partición de datos.

1.2. Planteamiento del problema

1.2.1. Definición del problema

La clasificación automatizada de litologías a partir de registros geofísicos de pozos petroleros presenta desafíos específicos relacionados con la naturaleza multivariada de los datos y el desbalance natural entre tipos de rocas. Los registros de resistividad, densidad, rayos gamma, tiempo de tránsito y calibre proporcionan información complementaria que debe ser integrada para discriminar entre litologías con propiedades físicas similares, como lutitas vs margas o calizas vs calizas arcillosas.

La distribución natural de litologías refleja procesos geológicos específicos: las rocas reservorio (areniscas) y generadoras (lutitas orgánicas) suelen estar subrepresentadas frente a rocas sello y de relleno, generando un desbalance que afecta el rendimiento de los clasificadores tradicionales. Este desbalance puede ser severo en secuencias sedimentarias reales, donde la correcta identificación de las litologías menos frecuentes resulta crítica para las decisiones de exploración y desarrollo en la industria petrolera.

Aunque algoritmos como *Support Vector Machine* (SVM), *Random Forest* (RF) y *Multi-Layer Perceptron* (MLP) son capaces de manejar datos multivariados con desbalance de clases, no se encontró consenso sobre cuál resulta más efectivo para la clasificación de litologías sedimentarias con desbalance severo. El presente trabajo compara el desempeño de estos tres algoritmos utilizando datos de registros geofísicos del pozo petrolero 25_7-2, proveniente de la base de datos abierta de Geolink. Se emplea validación cruzada estratificada y partición de datos para evaluar el desempeño global y la capacidad de cada modelo para manejar clases desbalanceadas.

1.2.2. Delimitación de la investigación

El análisis se realizó utilizando registros geofísicos directos del pozo 25_7-2 sin acceso a variables petrofísicas interpretadas como porosidad efectiva o saturación de agua (S_w), ni a parámetros del lodo de perforación. El estudio se enfoca en la clasificación litológica basada principalmente en mediciones físicas primarias, delimitando el alcance a la comparación metodológica de algoritmos de aprendizaje automático bajo condiciones de desbalance de clases.

1.3. Preguntas de investigación e hipótesis

Con base en el planteamiento del problema descrito, se formulan las siguientes preguntas de investigación que guían el desarrollo de este trabajo:

1. ¿Los algoritmos con configuraciones por defecto pueden manejar adecuadamente el desbalance de clases, o requieren ajustes específicos en sus hiperparámetros para lograr un desempeño competitivo?
2. ¿Cómo influye la complejidad del espacio de hiperparámetros de cada algoritmo en su desempeño final tras el proceso de optimización?

Hipótesis de investigación:

RF presentará mayor robustez frente al desbalance de clases que SVM y MLP, dado que no requiere normalización ni estandarización de datos al ser invariante a los rangos de las características, mientras que SVM y MLP dependen de distancias geométricas y combinaciones lineales que son sensibles a las diferencias de escala entre variables, requiriendo preprocesamiento adicional.

1.4. Objetivo general

Comparar SVM, RF y MLP en la clasificación de litologías del pozo 25_7-2 con desbalance de clases, evaluando sus configuraciones por defecto y el efecto de la optimización de hiperparámetros.

1.5. Objetivos específicos

1. Analizar los registros geofísicos del pozo 25_7-2, identificar la distribución de litologías presentes y preparar los datos mediante limpieza de valores faltantes, tratamiento de outliers y selección de características con base en su relevancia geológica.
2. Entrenar y optimizar SVM, RF y MLP mediante búsqueda de hiperparámetros sobre el conjunto de entrenamiento.

3. Comparar el desempeño de los tres algoritmos en el conjunto de prueba mediante *F1-score macro*, evaluando el impacto de la optimización y analizando confusiones en clases minoritarias.

1.6. Justificación

La selección de algoritmos de aprendizaje automático para clasificación litológica requiere considerar su capacidad para manejar desbalance de clases, aspecto crítico en datos geológicos donde litologías de interés económico suelen estar subrepresentadas. Este trabajo proporciona una comparación controlada de tres algoritmos bajo las mismas condiciones de datos, generando evidencia que puede orientar decisiones metodológicas en proyectos similares.

También se evalúa el impacto de la optimización de hiperparámetros versus configuraciones por defecto en este contexto específico, proporcionando información práctica sobre el esfuerzo de ajuste requerido por cada algoritmo.

1.7. Metodología utilizada

1. Análisis exploratorio y preprocesamiento de datos. Se analizará la distribución de litologías en el pozo 25_7-2, se identificarán valores faltantes y outliers, y se eliminará información no relevante para clasificación (profundidad, diámetro de broca). Se tratarán valores faltantes mediante imputación por media estratificada cuando representen menos del 1 % de los datos.
2. Selección de características. Se seleccionarán registros geofísicos útiles para distinguir litologías y que estén completos en el pozo. Se eliminarán características redundantes mediante análisis de correlación de Spearman. Los datos se normalizarán con `StandardScaler` para SVM y MLP; RF se entrenará sin normalización.

3. Partición estratificada de datos. Se dividirá el conjunto en entrenamiento (80 %) y prueba (20 %) mediante partición estratificada. El conjunto de prueba se reservará exclusivamente para la evaluación final de los modelos.
4. Entrenamiento con configuraciones por defecto. Se entrenarán SVM, RF y MLP con sus hiperparámetros por defecto de scikit-learn sobre el conjunto de entrenamiento para establecer una línea base de comparación.
5. Optimización de hiperparámetros. Se optimizarán los hiperparámetros mediante búsqueda bayesiana para SVM (kernel RBF, C y gamma), RandomizedSearch para RF, y GridSearch para MLP, empleando validación cruzada estratificada 5-fold sobre el conjunto de entrenamiento.
6. Evaluación y comparación. Se evaluarán los modelos con configuraciones por defecto y optimizadas en el conjunto de prueba mediante *F1-score macro*. Se compararán los tres algoritmos entre sí y se analizarán las confusiones específicas en clases minoritarias.

Capítulo 2

Marco teórico

El capítulo anterior estableció que el objetivo de este trabajo es comparar SVM, RF y MLP para clasificación de litologías con desbalance de clases. Este capítulo desarrolla conceptos teóricos necesarios: fundamentos de aprendizaje automático supervisado, descripción de los algoritmos seleccionados, técnicas para manejar el desbalance de clases y métricas de evaluación apropiadas.

2.1. Conceptos y teorías fundamentales de la investigación

2.1.1. Aprendizaje automático

El aprendizaje automático es una rama de la inteligencia artificial que desarrolla algoritmos capaces de aprender patrones a partir de datos. Estos algoritmos ajustan sus parámetros conforme procesan más información, adquiriendo capacidad para hacer predicciones sobre datos nuevos y reconocer estructuras en observaciones no vistas previamente (Mitchell, 1997). Existen dos tipos principales de aprendizaje automático, según cómo los modelos aprenden de los datos (Russell y Norvig, 2020):

- Supervisado: Entrena modelos a partir de datos con ejemplos de entrada y sus salidas conocidas, para predecir correctamente la etiqueta de nuevas observaciones.
- No supervisado: Analiza datos no etiquetados para identificar patrones y agrupar observaciones similares en clústeres.

2.1.2. La clasificación: una tarea del aprendizaje supervisado

La clasificación consiste en asignar una etiqueta de clase Y a un conjunto de observaciones en función f de un conjunto de características X , lo que puede expresarse como:

$$f : X \rightarrow Y \quad (2.1)$$

Esta expresión formaliza todos los problemas de clasificación, donde X corresponde al espacio de características y Y al espacio de salidas discretas, que es el conjunto de clases posibles. El objetivo de la clasificación es aprender una función f que mapee de manera óptima cada observación $x \in X$ a su clase correspondiente $y \in Y$.

2.1.3. Tipos de problemas de clasificación

Los problemas de clasificación pueden ser de tipo binario o multiclase. En clasificación binaria, la variable objetivo Y puede asumir únicamente dos posibles valores o etiquetas, por ejemplo, "positivo." "negativo". En contraste, la clasificación multiclase involucra tres o más etiquetas posibles para Y .

La extensión de algoritmos binarios a problemas multiclase se realiza típicamente mediante dos estrategias principales (Bishop, 2006):

One-vs-Rest (OvR): Entrena K clasificadores binarios, donde cada uno distingue una clase específica del resto. Para una nueva observación, se selecciona la clase cuya confianza del clasificador es mayor.

One-vs-One (OvO): Construye $\frac{K(K-1)}{2}$ clasificadores binarios, uno para cada par de clases. La clasificación final se determina mediante votación mayoritaria.

La elección entre OvR y OvO depende del número de clases, el tamaño del conjunto de datos y las características del algoritmo empleado.

2.1.4. Manejo de Desbalance de Clases

El desbalance de clases surge cuando las clases no están representadas equitativamente en el conjunto de datos (Chawla et al., 2002). Cuantitativamente, se considera desbalance cuando

la clase mayoritaria representa más del 60-70 % de las observaciones, o cuando la razón entre la clase más frecuente y la menos frecuente supera 3:1 (He y García, 2009).

Bajo este tipo de condiciones, los algoritmos tienden a favorecer la predicción de las clases más frecuentes, maximizando la precisión global a costa del rendimiento en clases minoritarias (Japkowicz y Stephen, 2002). Métricas como accuracy pueden alcanzar valores altos simplemente prediciendo la clase mayoritaria, sin reflejar adecuadamente el desempeño en clases minoritarias.

En clasificación de litologías, el desbalance refleja procesos geológicos naturales donde formaciones raras como lutitas con alto contenido orgánico o areniscas con porosidad excepcional pueden representar menos del 5-10 % del registro, pero constituyen horizontes de interés económico crítico.

2.1.5. Algoritmos de clasificación

Los datos geofísicos de pozos presentan alta correlación entre algunas variables, ruido en las mediciones y relaciones no lineales complejas entre registros y litologías. Se evaluaron tres algoritmos con diferentes aproximaciones: SVM encuentra fronteras de decisión óptimas mediante optimización de márgenes, Multi-Layer Perceptron modela relaciones no lineales a través de capas ocultas, y RF construye múltiples árboles mediante votación.

2.1.5.1. Support Vector Machine

SVM es un algoritmo de aprendizaje supervisado desarrollado por Vapnik y sus colaboradores (Cortes y Vapnik, 1995). Su objetivo es encontrar un hiperplano óptimo que separe las clases en el espacio de características, maximizando el margen entre las observaciones más cercanas de diferentes clases. El hiperplano de separación se define matemáticamente como:

$$\mathbf{w}^T \mathbf{x} + b = 0$$

donde $\mathbf{w}$ es el vector de pesos normal al hiperplano y b el término de sesgo. La clasificación se realiza evaluando en qué lado del hiperplano se encuentra una nueva observación:

Si $\mathbf{w}^T \mathbf{x} + b > 0 \rightarrow \text{Clase A}$

Si $\mathbf{w}^T \mathbf{x} + b < 0 \rightarrow \text{Clase B}$

Los vectores de soporte son las observaciones más próximas al hiperplano de separación y definen completamente la frontera de decisión.

Para problemas no linealmente separables, SVM emplea el truco del kernel (kernel trick), que transforma implícitamente los datos mediante una función kernel $K(\mathbf{x}_i, \mathbf{x})$, permitiendo mapear los datos a un espacio de mayor dimensionalidad donde la separación lineal es posible. Los kernels más utilizados incluyen el lineal, polinomial, radial basis function (RBF) y sigmoide.

En clasificación multiclase, SVM se extiende mediante estrategias como one-vs-one (OvO) y one-vs-rest (OvR). La estrategia OvO entrena un clasificador binario para cada par de clases, requiriendo $n(n-1)/2$ modelos para n clases, mientras que OvR entrena n clasificadores binarios, cada uno distinguiendo una clase específica del resto.

2.1.5.2. Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron es un tipo de red neuronal artificial *feedforward* compuesta por múltiples capas de neuronas interconectadas, donde la información fluye en una sola dirección desde la capa de entrada hasta la capa de salida (Rumelhart et al., 1986). Supera las limitaciones del perceptrón simple para resolver problemas no linealmente separables mediante capas ocultas y funciones de activación no lineales.

La arquitectura de un MLP consiste en una capa de entrada, una o más capas ocultas, y una capa de salida. Cada neurona realiza una combinación lineal de entradas seguida de una transformación no lineal:

$$y = f \left(\sum_{i=1}^n w_i x_i + b \right)$$

donde w_i son los pesos sinápticos, x_i las entradas, b el sesgo y f la función de activación. Las funciones de activación más comunes incluyen sigmoide, tangente hiperbólica (tanh) y ReLU

(*Rectified Linear Unit*). Para clasificación multiclase, la capa de salida emplea la función softmax, que convierte las salidas en una distribución de probabilidad sobre las clases.

El entrenamiento del MLP se realiza mediante el algoritmo de retropropagación (*backpropagation*), que ajusta iterativamente los pesos para minimizar una función de pérdida utilizando gradiente descendente.

2.1.5.3. Random Forest

RF es un algoritmo de aprendizaje por ensambles desarrollado por Breiman que combina múltiples árboles de decisión (Breiman, 2001). Su fundamento radica en la sabiduría colectiva: aunque un árbol individual puede cometer errores, la combinación de muchos árboles tiende a producir predicciones más robustas y precisas.

El algoritmo construye un bosque de árboles de decisión mediante dos mecanismos de aleatoriedad. Primero, cada árbol se entrena con una muestra bootstrap diferente del conjunto de datos original, proceso conocido como bagging (bootstrap aggregating). Segundo, en cada nodo de división, solo se considera un subconjunto aleatorio de características para determinar la mejor división, introduciendo diversidad adicional entre los árboles.

La función matemática central para clasificación se basa en el voto mayoritario:

$$\hat{y} = \text{modo}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_B(\mathbf{x})\}$$

donde $h_i(\mathbf{x})$ representa la predicción del i -ésimo árbol y B es el número total de árboles en el bosque. Para obtener probabilidades de clase, se calcula la proporción de votos:

$$P(y = c|\mathbf{x}) = \frac{1}{B} \sum_{i=1}^B I(h_i(\mathbf{x}) = c)$$

donde $I(\cdot)$ es la función indicadora.

Cada árbol individual se construye recursivamente mediante divisiones binarias que maximizan la pureza de los nodos resultantes. Los criterios más comunes para medir la impureza incluyen el índice de Gini y la entropía. El índice de Gini se define como:

$$\text{Gini}(t) = 1 - \sum_{i=1}^c p_i(t)^2$$

donde $p_i(t)$ es la proporción de muestras de la clase i en el nodo t .

RF proporciona medidas de importancia de características, calculadas mediante la reducción promedio de impureza que cada característica aporta a través de todos los árboles del bosque, permitiendo identificar las variables más relevantes para la clasificación.

2.1.6. Evaluación de modelos de clasificación

Un modelo de aprendizaje automático puede evaluarse de distintas maneras, dependiendo tanto del tipo de problema como de la naturaleza de los datos disponibles. En la clasificación binaria, donde únicamente existen dos clases, se presentan cuatro posibles escenarios:

- *TP (True Positive)*: instancia positiva predicha correctamente.
- *FP (False Positive)*: instancia negativa predicha como positiva.
- *FN (False Negative)*: instancia positiva predicha como negativa.
- *TN (True Negative)*: instancia negativa predicha correctamente.

A partir de estos escenarios se derivan las métricas básicas de clasificación binaria (ver Tabla 2.1). Estas métricas pueden extenderse de manera natural a la clasificación multiclase. En este caso, cada clase se trata de forma individual como “positiva” frente a todas las demás consideradas “negativas” (estrategia *one-vs-rest*), lo que permite calcular Precision, Recall y F1 por clase, y posteriormente combinar estos valores mediante agregaciones macro, micro o ponderadas para obtener una medida global del desempeño del modelo.

Para clasificación multiclase, el *F1-score* proporciona una medida equilibrada del desempeño por clase. A partir de los valores de $F1_i$ para cada clase, se define el *F1-score macro*:

$$F1_{\text{macro}} = \frac{1}{K} \sum_{i=1}^K F1_i,$$

donde K es el número de clases. El *F1-score macro* es una métrica global que considera todas las clases por igual, independientemente de su proporción en el conjunto de datos. Para

Métrica	Definición conceptual	Fórmula
Precisión (<i>Precision</i>)	Proporción de instancias predichas como positivas que son realmente positivas. Mide la confiabilidad de las predicciones positivas.	$Precision = \frac{TP}{TP+FP}$
Sensibilidad (<i>Recall</i>)	Proporción de instancias realmente positivas que son correctamente identificadas por el modelo. Mide la cobertura de los positivos reales.	$Recall = \frac{TP}{TP+FN}$
F1-score	Media armónica entre Precision y Recall, proporcionando una medida equilibrada entre confiabilidad y cobertura.	$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$

Tabla 2.1. Principales métricas para clasificación binaria

complementar estas métricas, se utiliza la *matriz de confusión multiclase*, de tamaño $K \times K$, que permite visualizar los aciertos y errores entre todas las clases; en la fila j y columna k se muestra el número de instancias cuya clase real es j y que fueron predichas como k . A continuación se presenta un ejemplo con 3 clases:

Real / Predicha	Clase 1	Clase 2	Clase 3
Clase 1	C_{11}	C_{12}	C_{13}
Clase 2	C_{21}	C_{22}	C_{23}
Clase 3	C_{31}	C_{32}	C_{33}

Tabla 2.2. Matriz de confusión multiclase (3 clases) y ejemplo de cálculo para la clase 1.

A partir de esta matriz se calculan los elementos para cada clase:

- $TP_i = C_{ii}$: instancias de la clase i correctamente clasificadas.
- $FP_i = \sum_{k \neq i} C_{ki}$: instancias de otras clases clasificadas incorrectamente como i .
- $FN_i = \sum_{j \neq i} C_{ij}$: instancias de la clase i clasificadas incorrectamente como otra clase.
- TN_i : suma de todas las demás celdas que no están en la fila ni columna de i .
- $Soporte_i = \sum_{j=1}^K C_{ij}$: número total de instancias reales de la clase i .

Ejemplo para la clase 1:

- $TP_1 = C_{11}$,

- $FP_1 = C_{21} + C_{31}$,
- $FN_1 = C_{12} + C_{13}$,
- $TN_1 = C_{22} + C_{23} + C_{32} + C_{33}$,
- $Soporte_1 = C_{11} + C_{12} + C_{13}$.

2.1.7. Partición y validación de datos

2.1.7.1. Partición de Datos

El conjunto de datos disponible se divide en subconjuntos independientes para entrenar y evaluar el modelo:

- Conjunto de entrenamiento (*training set*): Utilizado para entrenar el modelo, permitiendo que el algoritmo ajuste sus parámetros internos.
- Conjunto de validación (*validation set*): Empleado para evaluar el modelo durante el entrenamiento y seleccionar hiperparámetros.
- Conjunto de prueba (*test set*): Reservado exclusivamente para la evaluación final del modelo, proporcionando una estimación imparcial de su rendimiento en datos no vistos.

2.1.7.2. Validación Cruzada

La validación cruzada es una técnica para estimar el desempeño de un modelo cuando se dispone de conjuntos de datos limitados (Stone, 1974). El método más común es la validación cruzada k -fold, que divide el conjunto de datos en k subconjuntos (folds) de tamaño aproximadamente igual.

El procedimiento consiste en entrenar el modelo k veces, utilizando cada vez $k - 1$ folds para entrenamiento y el fold restante para validación. Esto produce k estimaciones independientes del rendimiento, cuyo promedio proporciona una evaluación más estable y confiable que una sola partición entrenamiento-validación.

Para problemas de clasificación con desbalance de clases, la validación cruzada estratificada preserva la proporción de cada clase en todos los folds, asegurando que cada subconjunto sea representativo de la distribución global de clases.

2.2. Literatura relacionada

Diversos estudios han abordado la clasificación de litologías con aprendizaje automático, con enfoques que varían en número de clases, grado de desbalance, algoritmos empleados y estrategias de validación. A continuación se revisan trabajos comparables al presente estudio y se discuten enfoques alternativos identificados.

2.2.1. Estudios comparativos de algoritmos de clasificación

Merembayev et al. (2018) compararon cinco algoritmos de aprendizaje automático (RF, Logistic Regression, Gradient Boosting, KNN y XGBoost) en la clasificación de seis litologías del depósito Moyunkum en Kazajistán, utilizando datos de 42 pozos con registros de resistividad y potencial espontáneo, complementados con características derivadas mediante gradientes de primera y segunda derivada. RF alcanzó el mayor accuracy (0.70) superando a Logistic Regression (0.68), Gradient Boosting, XGBoost (0.65) y KNN (0.53). Los autores reconocen que estos resultados no son satisfactorios, atribuyendo el bajo desempeño a la complejidad inherente de las transiciones graduales entre tipos de roca en formaciones sedimentarias (Merembayev et al., 2018). Este trabajo evidencia la dificultad del problema de clasificación litológica, incluso con algoritmos ensemble robustos, particularmente en contextos geológicos con fronteras difusas entre clases.

Al-Anazi y Gates (2010) compararon tres enfoques metodológicamente distintos (SVM, Linear Discriminant Analysis y Probabilistic Neural Network) para clasificar 3 litologías (*clean sand*, *shaly sand*, *sandy shale*) en areniscas heterogéneas del Medio Oriente, utilizando datos de 2 pozos. Empleando 9 registros geofísicos reducidos a 4 características mediante *fuzzy ranking*, SVM con kernel sigmoidal alcanzó aproximadamente 70 % de accuracy en validación cruzada entre pozos, superando significativamente a LDA y PNN, especialmente en la correcta identificación de la clase intermedia *shaly sand*. A diferencia del presente trabajo, que evalúa el modelo mediante

partición estratificada dentro del mismo pozo, Al-Anazi y Gates emplearon validación cruzada entre dos pozos diferentes, estrategia orientada a evaluar transferibilidad geográfica del modelo (Al-Anazi y Gates, 2010).

Zhang et al. (2022) realizaron una comparación exhaustiva de siete algoritmos en la clasificación de 3 litologías del Junggar Basin, China, utilizando un dataset de 4,156 muestras con desbalance moderado (relación aproximada 2:1) y sin aplicar técnicas de rebalanceo. XGBoost alcanzó el mejor desempeño con F1-score de 0.876, seguido por RF (0.849) y SVM (0.835), empleando validación cruzada y optimización bayesiana de hiperparámetros para cada algoritmo. El presente trabajo difiere sustancialmente en el grado de desbalance (10:1 vs 2:1) y la complejidad del problema (8 clases vs 3), factores que influyen significativamente en la efectividad relativa de los clasificadores (Zhang et al., 2022).

Mohamed et al. (2019) realizaron el estudio comparable al presente trabajo, comparando cinco algoritmos (SVM, KNN, RF, MLP y K-means) en la clasificación de 9 litologías de 8 pozos del Anadarko Basin, Kansas, utilizando 5 registros geofísicos. Con validación mediante un pozo ciego completamente independiente del conjunto de entrenamiento, SVM alcanzó el mejor desempeño (F1-score=0.49, accuracy en capas adyacentes=0.89), superando a RF (F1=0.45) y Multi-Layer Perceptron (F1=0.42). Los autores concluyeron que métodos supervisados simples con optimización apropiada de hiperparámetros pueden superar a arquitecturas de redes neuronales más complejas cuando el desbalance de clases no se mitiga artificialmente mediante técnicas de remuestreo o ajuste de pesos. Aunque su estrategia de validación evalúa transferibilidad entre ubicaciones geográficamente distintas, el presente trabajo adopta un enfoque más conservador mediante partición estratificada dentro del mismo pozo, evitando asumir homogeneidad en las distribuciones de características geofísicas entre pozos de diferentes contextos estructurales (Mohamed et al., 2019).

Rohit et al. (2024) compararon cinco algoritmos en clasificación de 4 litologías del Volve Field, Noruega, utilizando datos de 8 pozos. Decision Tree alcanzó accuracy de 0.98, seguido por SVC (0.92) y AdaBoost (0.88). La menor complejidad del problema (4 litologías vs 8-9 en otros estudios) y ausencia de análisis de gap train/test limitan la comparabilidad con trabajos en problemas más complejos (Rohit et al., 2024).

La Tabla 2.3 sintetiza las características metodológicas y los resultados principales de los

trabajos revisados.

Autor (Año)	Región	Clases	Desbalance	Mejor algoritmo	Métrica	Manejo desbalance
Merembayev (2018)	Kazakhstan	6	N.R.	<i>Random Forest</i>	Acc=0.70	N.R.
Al-Anazi (2010)	Middle East	3	Sí	SVM (sigmoid)	Acc≈0.70	No
Zhang (2022)	China	3	Sí (2:1)	XGBoost	F1=0.876	Opt. bayesiana
Mohamed (2019)	Kansas (Anadarko)	9	Sí	SVM	F1=0.49	No
Rohit (2024)	Noruega (Volve)	4	Sí	Decision Tree	Acc=0.98	No

Tabla 2.3. Enfoques para clasificación de litologías

2.2.2. Enfoques metodológicos alternativos y limitaciones

Durante la revisión bibliográfica se identificaron otros enfoques para clasificación de litologías que presentan diferencias metodológicas que impiden comparación directa con este trabajo.

Ren et al. (2019) diseñaron un método híbrido que integra redes neuronales artificiales con conocimiento geológico experto, específicamente patrones de asociación vertical de litologías observados en núcleos. Mediante un modelo de Permanencia de Ratios, los autores combinaron probabilidades condicionales calculadas con redes neuronales y probabilidades derivadas de análisis estadístico de secuencias estratigráficas, alcanzando 91 % de accuracy en pozos del campo Gudong, China. Aunque este enfoque demuestra la utilidad de incorporar conocimiento del dominio geológico, no permite comparación directa con clasificadores puros de aprendizaje automático, ya que la integración de probabilidades condicionales mediante el modelo de permanencia de ratios constituye una etapa adicional de procesamiento que modifica fundamentalmente la arquitectura del sistema (Ren et al., 2019).

Otros trabajos identificados reportan aplicaciones de algoritmos ensemble modernos (CatBoost, LightGBM) y arquitecturas de aprendizaje profundo (redes convolucionales, redes recurrentes) con resultados prometedores. Sin embargo, la disponibilidad limitada de información metodológica detallada (incluyendo número preciso de clases litológicas, cuantificación del grado de desbalance y estrategias de validación empleadas) impidió su inclusión en el análisis comparativo. Estos trabajos refuerzan la tendencia observada hacia la preferencia por métodos supervisados sobre métodos no supervisados, así como la importancia de la optimización de hiperparámetros en problemas de clasificación litológica con desbalance de clases.

La revisión bibliográfica reveló que la mayoría de estudios utiliza accuracy como métrica principal de evaluación. Esta métrica resulta engañosa en presencia de desbalance severo, donde un clasificador que predice únicamente la clase mayoritaria puede alcanzar valores artificialmente elevados sin capacidad discriminatoria real. El presente trabajo emplea F1-score macro como métrica principal, proporcionando una evaluación más apropiada que pondera equitativamente el desempeño en todas las clases independientemente de su frecuencia.

2.3. Marco tecnológico

Este trabajo abarcó el ciclo completo de un proyecto de ciencia de datos: carga y exploración de registros geofísicos, preprocesamiento, entrenamiento de modelos y análisis de resultados. La implementación se realizó en Python 3.x utilizando scikit-learn como biblioteca principal para los algoritmos de clasificación.

Scikit-learn proporcionó las implementaciones de SVM, RF y MLP, además de herramientas para preprocesamiento, validación cruzada y evaluación. El manejo de datos geofísicos requirió bibliotecas adicionales que se describen a lo largo de esta sección.

2.3.1. Bibliotecas

Para trabajar con aprendizaje automático se seleccionó Scikit-learn, considerada un estándar en trabajos académicos. Esta biblioteca de código abierto proporciona implementaciones eficientes y con documentación detallada de los algoritmos de clasificación utilizados en este trabajo: SVM, RF y MLP, así como herramientas para preprocesamiento, validación y evaluación de modelos. A continuación, se presentan otras bibliotecas utilizadas:

■ Procesamiento y análisis de datos

- *NumPy*: operaciones numéricas sobre arreglos multidimensionales, operaciones matriciales, cálculos estadísticos y transformaciones de datos.
- *Pandas*: manipulación de datos estructurados mediante DataFrames, incluyendo el manejo de valores faltantes.

- *SciPy*: funciones estadísticas, como el cálculo de la correlación de Pearson y Spearman.

■ Visualización

- *Matplotlib*: gráficos estáticos como matrices de confusión y curvas de aprendizaje.
- *Seaborn*: gráficos estadísticos avanzados, como heatmaps y distribuciones.
- *Missingno*: visualización de datos faltantes.

■ Optimización de hiperparámetros

- *Scikit-optimize*: permite la búsqueda bayesiana de hiperparámetros.
- *GridSearchCV* / *RandomizedSearchCV* (Scikit-learn): proporcionan búsquedas exhaustivas o estocásticas de configuraciones óptimas de hiperparámetros.

Capítulo 3

Modelo de clasificación de litologías

Este capítulo describe la implementación del modelo de clasificación de litologías. Se presenta la caracterización de los datos geofísicos del pozo 25_7-2, el preprocesamiento aplicado y la estrategia experimental para comparar Support Vector Machine, Random Forest y Multi-Layer Perceptron.

3.1. Descripción de los datos

Los datos usados en esta investigación provienen del pozo petrolero 25_7-2, proporcionado por Geolink bajo licencia Creative Commons 4.0. Los registros geofísicos miden propiedades físicas del subsuelo a intervalos regulares de profundidad, incluyendo mediciones eléctricas, acústicas y radioactivas, junto con parámetros operacionales de perforación. Cada profundidad registrada cuenta con una etiqueta de clasificación litológica que sirve como variable objetivo para el entrenamiento supervisado de los modelos.

Técnicamente, los registros geofísicos constituyen mediciones continuas de propiedades físicas del subsuelo que permiten caracterizar las formaciones rocosas sin necesidad de extracción de núcleos. La disponibilidad de múltiples registros en cada pozo proporciona información complementaria que facilita la discriminación entre litologías. A continuación se describen los datos disponibles sobre el pozo 25_7-2.

Parámetros operacionales de **profundidad** y **calibración**:

- DEPT (*Depth*): Profundidad medida desde la superficie en metros. Constituye la variable de indexación espacial de todos los demás registros.
- BS (*Bit Size*): Diámetro nominal de la broca en pulgadas. Sirve como referencia para calcular DCAL y evaluar las condiciones del agujero perforado.
- CALI (*Caliper*): Diámetro del pozo medido en pulgadas. Identifica zonas de derrumbe, cavernas o ensanchamiento.
- DCAL (*Differential Caliper*): Diferencia entre CALI y BS. Valores positivos indican ensanchamiento del pozo; valores cercanos a cero sugieren formaciones competentes o presencia de revoque en zonas permeables (Glover, 2000).

Registros de densidad y porosidad:

- RHOB (*Bulk Density*): Densidad total de la formación medida en g/cm³. Fundamental para identificar litologías, ya que diferentes tipos de roca presentan densidades características. (London, 2000).
- DRHO (*Density Correction*): Factor de corrección aplicado a la medición de densidad para compensar efectos de rugosidad del pozo.
- NPHI (*Neutron Porosity*): Porosidad medida mediante dispersión de neutrones, expresada en m³/m³. Sensible al contenido de hidrógeno en la formación, principalmente agua y hidrocarburos en el espacio poroso. (Kansas Geological Survey, 2017a).
- PEF (*Photoelectric Factor*): Factor fotoeléctrico medido en b/e (barns por electrón). Relacionado con el número atómico promedio de la formación, útil para discriminar litologías.

Registros de radioactividad:

- GR (*Gamma Ray*): Radiactividad natural de la formación medida en unidades API. Elevada en lutitas y arcillas debido a su contenido de potasio, torio y uranio; baja en areniscas y carbonatos limpios (2017b).

- SGR (*Spectral Gamma Ray*): Rayos gamma espectral que descompone la señal total en contribuciones individuales, medido en gAPI.
- THOR (*Thorium*): Concentración de torio medida en ppm (partes por millón).
- URAN (*Uranium*): Concentración de uranio medida en ppm. Elevada en lutitas ricas en materia orgánica (2017b).

Registros acústicos:

- DTC (*Compressional Sonic*): Tiempo de tránsito de ondas compresionales a través de la formación, medido en $\mu\text{s}/\text{ft}$. Inversamente relacionado con la velocidad sísmica y la compactación de la roca.

Registros de resistividad:

- RDEP (*Deep Resistivity*): Resistividad profunda de la formación medida en $\text{ohm}\cdot\text{m}$. Refleja las propiedades eléctricas de la roca alejadas de la zona afectada por el lodo de perforación.
- RMED (*Medium Resistivity*): Resistividad a profundidad media de investigación, expresada en $\text{ohm}\cdot\text{m}$.
- RSHA (*Shallow Resistivity*): Resistividad somera medida en $\text{ohm}\cdot\text{m}$. Afectada por la invasión de filtrado de lodo.
- RMIC (*Micro Resistivity*): Resistividad de la zona lavada inmediatamente adyacente a la pared del pozo, medida en $\text{ohm}\cdot\text{m}$.

Otros registros:

- SP (*Spontaneous Potential*): Potencial eléctrico natural generado por diferencias de salinidad entre el lodo de perforación y las formaciones permeables, medido en mV.
- RHOM (*Matrix Density*): Densidad de la matriz rocosa en g/cm^3 , calculada asumiendo porosidad cero.

3.1.1. Litologías

El conjunto de datos presenta once litologías sedimentarias distintas, codificadas numéricamente del 2 al 19, representativas de ambientes deposicionales típicos de cuencas petroleras.

Abbreviation	Lithology Attribute	Color	Attribute
2	Sand	LightYellow	Silty Sand
3	CrossBedded Sand	Gold	Cross Bedded
5	Silt	LightSalmon	Sst
6	Silt	DarkYellow	Silt
7	Shale	DarkGreen	Shaly Silt
11	Limestone	SlateBlue	Silty Shale
12	Chalk	Dodger Blue	Limestone
13	Chalk	DeepSkyBlue	Argillaceous Limestone
16	Sandy Limestone	Cyan	Marlstone
18	Coal	Black	Calcareous Cement
19	Extrusive	Cyan	Cinerite

Tabla 3.1. Tabla de atributos litológicos

Las litologías identificadas corresponden a:

- Sand (2): Areniscas, rocas sedimentarias clásticas de grano medio compuestas predominantemente por cuarzo. Constituyen las principales rocas reservorio en sistemas petroleros clásticos.
- Cross-Bedded Sand (3): Areniscas con estratificación cruzada, indicativas de ambientes de alta energía como canales fluviales o dunas eólicas.
- Silt (5): Limolitas de grano fino, con tamaño de partícula intermedio entre arcillas y arenas.
- Silt (6): Limolitas con características distintivas de registro que justifican su clasificación como categoría separada del código 5.
- Shale (7): Lutitas, rocas sedimentarias de grano muy fino ricas en minerales arcillosos. Actúan como rocas sello y, cuando son ricas en materia orgánica, como rocas generadoras de hidrocarburos.
- Limestone (11): Calizas, rocas sedimentarias carbonáticas compuestas principalmente por calcita. Pueden constituir importantes reservorios en sistemas carbonatados.

- Chalk (12): Calizas de grano muy fino, porosas y poco consolidadas, formadas por acumulación de restos de microorganismos calcáreos.
- Chalk (13): Calizas micríticas con propiedades petrofísicas diferenciables del código 12.
- Sandy Limestone (16): Calizas arenosas, rocas de composición mixta carbonatada-clástica resultantes de ambientes transicionales.
- Coal (18): Carbón, roca sedimentaria orgánica formada por acumulación de materia vegetal. Horizonte guía estratigráfico en secuencias continentales.
- Extrusive (19): Rocas volcánicas extrusivas, indicativas de actividad ígnea durante la deposición sedimentaria.

La distribución de estas litologías en el pozo 25_7-2 refleja el desbalance natural característico de secuencias sedimentarias reales. El análisis cuantitativo de la distribución litológica revela varios aspectos relevantes para el diseño del modelo de clasificación:

LITHOLOGY_GEOLINK	25_7-2
2	3553
3	459
5	2616
6	1782
7	136
11	2310
12	3667
13	2930
16	124
18	178
19	344
NaN	5821

Tabla 3.2. Distribución de litologías en el pozo 25_7-2

Desbalance entre clases: Existe una clara predominancia de limolitas (códigos 5 y 6) y calizas (códigos 11, 12 y 13), que en conjunto representan más del 60 % de las muestras etiquetadas. En contraste, litologías como lutitas (código 7), calizas arenosas (código 16) y carbón (código 18) se encuentran significativamente subrepresentadas, con frecuencias inferiores al 1 % del total.

Clases con baja representación: Lutitas (código 7: 136 muestras), calizas arenosas (código 16: 124 muestras) y carbón (código 18: 178 muestras) representan menos del 3 % del conjunto

de datos. Su baja frecuencia plantea desafíos para la evaluación estadística robusta con métricas por clase.

Datos sin etiquetar: El pozo presenta 5821 registros sin clasificación litológica (NaN), representando aproximadamente el 25 % del total de mediciones. Estos datos corresponden a intervalos no interpretados, zonas de calidad de registro deficiente o formaciones cuya clasificación presenta ambigüedad.

Esta caracterización cuantitativa evidencia que el problema de clasificación litológica enfrenta un desbalance severo entre clases mayoritarias y minoritarias, con litologías de interés geológico subrepresentadas en el conjunto de datos. Este desbalance natural justifica la evaluación de la capacidad intrínseca de los algoritmos de clasificación para manejar distribuciones desiguales sin aplicar técnicas específicas de balanceo.

3.2. Tratamiento de datos faltantes

El conjunto de datos presenta dos categorías de valores faltantes que requieren tratamiento diferenciado: ausencia de etiquetas litológicas y datos faltantes en registros geofísicos.

Eliminación de muestras sin etiqueta litológica:

El pozo 25_7-2 contiene intervalos significativos sin clasificación litológica (NaN en LITHOLOGY_GEOLINK), representando aproximadamente el 25 % del total de registros. El análisis de la Matriz 3.1 revela que estos intervalos se concentran en profundidades someras, correspondientes a secciones fuera del intervalo objetivo de exploración. Dado que el entrenamiento supervisado requiere etiquetas de clase conocidas, todas las muestras sin clasificación litológica fueron excluidas del conjunto de datos.

Tratamiento de datos faltantes en registros geofísicos:

Posterior a la eliminación de muestras sin etiqueta, se evaluó la presencia de datos faltantes en los registros geofísicos restantes del pozo 25_7-2. El análisis reveló que varios registros presentan disponibilidad parcial, lo que determinó su exclusión o tratamiento según el porcentaje de valores faltantes.

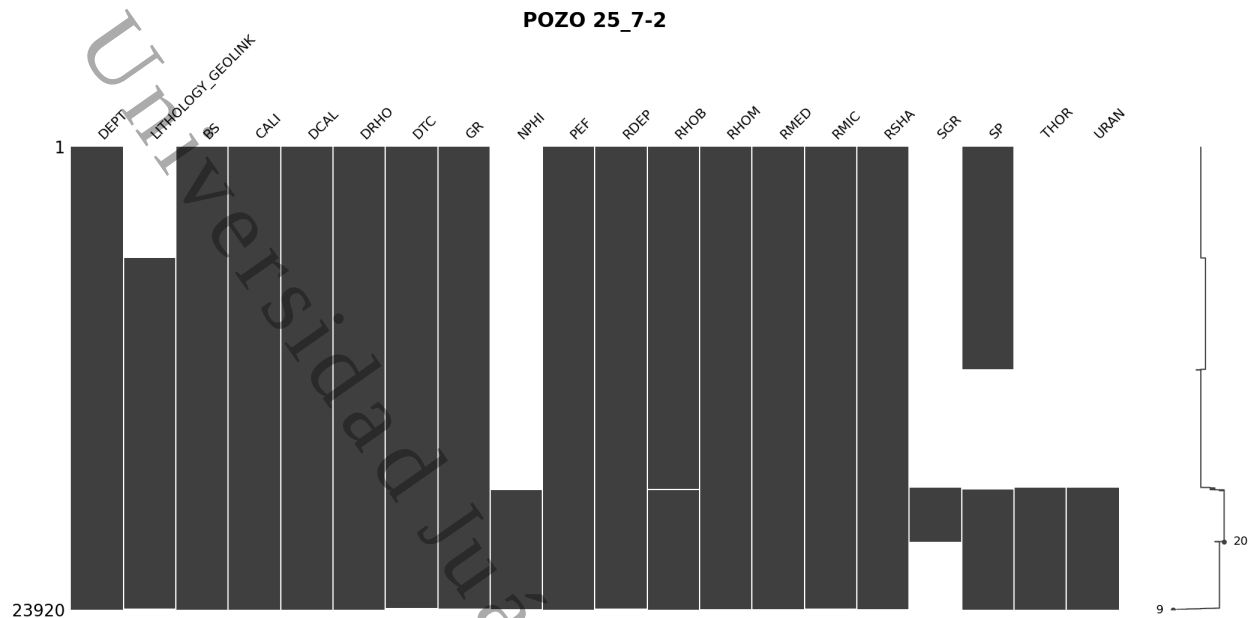


Figura 3.1. Matriz de valores faltantes

Imputación de valores faltantes minoritarios

Después de eliminar las instancias sin etiquetas de clase, se cuantificaron los valores faltantes remanentes en cada registro geofísico. El análisis reveló que el registro de potencial espontáneo (SP) presenta 34.3 % de valores faltantes en el pozo 25_7-2. Debido a que esta proporción compromete la confiabilidad de cualquier método de imputación, se decidió excluir SP del conjunto final de características.

Los registros restantes presentan valores faltantes en proporciones inferiores al 1 % del total de muestras: RHOB con 54 valores faltantes, DTC con 9 valores faltantes y GR con 8 valores faltantes. Para estos casos, se aplicó imputación mediante la media estratificada por clase litológica. Este enfoque se fundamenta en el principio de que diferentes litologías presentan propiedades petrofísicas características distintas; por ejemplo, la densidad promedio de calizas difiere sistemáticamente de la densidad promedio de areniscas. La imputación por clase preserva esta estructura natural de los datos, evitando la introducción de sesgos que resultarían del uso de una media global.

Exclusión de litologías con representación insuficiente

Se excluyeron del análisis tres litologías (códigos 7, 16 y 18) debido a su representación insuficiente para evaluación estadística robusta. Con la partición estratificada 80/20, estas clases

generan conjuntos de prueba de aproximadamente 27, 25 y 36 muestras respectivamente, cantidades insuficientes para cálculo confiable de métricas por clase que presentarían alta varianza debido al tamaño muestral reducido.

Esta decisión metodológica no compromete la validez del estudio, dado que las ocho litologías restantes constituyen una muestra representativa de ambientes deposicionales sedimentarios y mantienen un desbalance significativo entre clases (razón 1:10 entre clases mayoritarias y minoritarias), preservando el desafío de clasificación multiclase desbalanceada que motiva esta investigación.

3.2.1. Tratamiento de datos atípicos

La identificación de valores atípicos en registros geofísicos requiere considerar tanto criterios estadísticos como la validez física de las mediciones. A diferencia de muchos dominios donde los outliers representan errores de medición, en datos petrofísicos los valores extremos pueden corresponder a litologías específicas o condiciones geológicas particulares. Por esta razón, la estrategia adoptada se fundamentó en evaluar caso por caso la naturaleza de los valores extremos, considerando su validez geológica y la proporción de datos afectados.

Se realizó un análisis de rangos para cada registro geofísico de interés. La mayoría de los registros presentaron rangos consistentes con valores esperados; por ejemplo, rayos gamma (GR) exhibió valores entre 12.2 y 191.0 gAPI, y tiempo de tránsito compresional (DTC) mostró rangos entre 51.7 y 139.4 $\mu\text{s}/\text{ft}$. Sin embargo, dos registros presentaron valores anómalos que requirieron análisis detallado: resistividad profunda (RDEP) y densidad de matriz (RHOM).

Densidad de matriz (RHOM)

El registro de densidad de matriz (RHOM) presentó 270 valores anómalos ($<2.635 \text{ g}/\text{cm}^3$, 1.53 % del total), incluyendo 267 valores negativos físicamente imposibles (códigos de error entre -999.9 y $-0.498 \text{ g}/\text{cm}^3$) y 3 valores de $1.103 \text{ g}/\text{cm}^3$ anómalamente bajos para densidad de matriz rocosa. Estos valores fueron reemplazados por NaN y posteriormente imputados mediante media estratificada por clase litológica.

Resistividad profunda (RDEP)

El registro de resistividad profunda presentó valores máximos de 508.66 ohm·m. El análisis de

la distribución ordenada reveló un salto discontinuo significativo de 98.37 ohm·m entre los valores de 79.39 y 177.77 ohm·m, indicando una brecha natural en los datos. Se identificaron 8 instancias con valores superiores a 79.39 ohm·m, todas concentradas en un intervalo de profundidad de aproximadamente 1 metro (4815.08-4816.14 m), correspondientes principalmente a areniscas con estratificación cruzada (código 3).

Aunque valores de resistividad de este orden de magnitud son geológicamente válidos para formaciones altamente resistivas como areniscas con hidrocarburos o muy consolidadas, su concentración en un intervalo localizado de 1 metro representa una anomalía geológica específica no representativa de las condiciones generales del pozo. Este intervalo, aunque geológicamente interesante, no constituye un volumen significativo en términos de caracterización de yacimientos.

Dado que estas 8 instancias representan únicamente el 0.045 % del conjunto de datos, su eliminación se justifica por múltiples razones: primero, no afecta la representatividad estadística de ninguna clase litológica. Segundo, evita que algoritmos como Random Forest generen divisiones especializadas en condiciones excepcionales de frecuencia extremadamente baja. Tercero, previene que modelos basados en distancias (SVM, MLP) asignen peso desproporcionado a valores extremos durante la normalización. El rango resultante de RDEP (0.48-79.39 ohm·m) mantiene valores geológicamente consistentes con la variabilidad natural de resistividades en formaciones sedimentarias y permite que los modelos aprendan patrones representativos del yacimiento.

3.3. Selección y construcción de características

La selección de características identifica el subconjunto de registros geofísicos más relevantes para discriminar entre litologías, eliminando variables redundantes, no informativas o con problemas de calidad.

3.3.1. Eliminación de características no relevantes

Variables de indexación y parámetros operacionales

La profundidad (DEPT), el diámetro de broca (BS) y los registros de calibre (CALI, DCAL) fueron excluidos por no constituir propiedades intrínsecas de las formaciones rocosas. DEPT funciona únicamente como variable de indexación espacial sin relación causal con la litología.

BS, CALI y DCAL representan parámetros operacionales de perforación determinados por consideraciones ingenieriles o condiciones mecánicas del pozo, sin valor predictivo directo sobre las propiedades geológicas de la formación.

Características redundantes

El factor de corrección de densidad (DRHO) fue eliminado dado que RHOB representa la densidad total ya corregida, incorporando los ajustes por efectos de rugosidad del pozo y espesor de mudcake.

Multicolinealidad en registros de resistividad

El análisis de correlación de Spearman reveló multicolinealidad extrema entre RDEP, RMED y RSHA, con coeficientes superiores a 0.98 (ver Tabla 3.3). Esta alta correlación refleja que, en las formaciones estudiadas, la invasión de filtrado de lodo es limitada, resultando en mediciones de resistividad virtualmente idénticas a diferentes profundidades de investigación. Se conservó únicamente resistividad profunda (RDEP) por representar las propiedades eléctricas de la formación virgen con mínima contaminación por filtrado de lodo. RMIC fue excluido por representar la resistividad de la zona inmediatamente adyacente a la pared del pozo, fuertemente afectada por la invasión de filtrado de lodo, proporcionando información menos representativa de las propiedades eléctricas de la formación que RDEP.

	RDEP	DTC	GR	PEF	RHOM	RHOB	RMIC	RSHA	RMED
RDEP	1.00	-0.69	-0.18	-0.49	0.19	0.59	0.72	0.98	0.98
DTC	-0.69	1.00	0.57	0.43	-0.42	-0.65	-0.70	-0.74	-0.72
GR	-0.18	0.57	1.00	0.05	-0.13	-0.08	-0.28	-0.26	-0.22
PEF	-0.49	0.43	0.05	1.00	-0.36	-0.42	-0.48	-0.48	-0.46
RHOM	0.19	-0.42	-0.13	-0.36	1.00	0.50	0.30	0.21	0.18
RHOB	0.59	-0.65	-0.08	-0.42	0.50	1.00	0.59	0.61	0.60
RMIC	0.72	-0.70	-0.28	-0.48	0.30	0.59	1.00	0.75	0.74
RSHA	0.98	-0.74	-0.26	-0.48	0.21	0.61	0.75	1.00	0.99
RMED	0.98	-0.72	-0.22	-0.46	0.18	0.60	0.74	0.99	1.00

Tabla 3.3. Matriz de correlación de Spearman

Registros con indisponibilidad o problemas de calidad

Registros como SGR, NPHI, THOR, URAN y RMIC fueron excluidos debido a alta proporción de valores faltantes en el pozo 25_7-2. El registro de potencial espontáneo (SP) presenta 34.3% de valores faltantes, comprometiendo la confiabilidad de cualquier método de imputación. El re-

gistro de factor fotoeléctrico (PEF) presenta 12 % de valores físicamente imposibles (>1000 b/e), cuando el rango esperado para formaciones sedimentarias es 1-10 b/e, indicando problemas instrumentales específicos de esa herramienta. Los rangos observados en los demás registros son consistentes con valores esperados para formaciones sedimentarias, confirmando la confiabilidad de los sensores utilizados.

3.3.2. Conjunto final de características

El proceso de selección resultó en un conjunto de cinco características para el entrenamiento de los modelos de clasificación:

- DTC (*Compressional Sonic*): Tiempo de tránsito de ondas compresionales, relacionado con porosidad y compactación de la formación.
- GR (*Gamma Ray*): Radiactividad natural, discriminador primario entre litologías arcillosas y limpias.
- RDEP (*Deep Resistivity*): Resistividad eléctrica de la formación virgen, sensible a porosidad, saturación y composición mineralógica.
- RHOB (*Bulk Density*): Densidad total de la formación, fuertemente correlacionada con litología y porosidad.
- RHOM (Matrix Density): Densidad de la matriz rocosa, útil para discriminación entre carbonatos y clásticos.

3.4. Procedimiento experimental

El proceso de clasificación se estructuró en dos fases experimentales para cada uno de los tres algoritmos evaluados (SVM, Random Forest y MLP). El pseudocódigo (ver 1) presenta el flujo metodológico completo que se aplicó de manera consistente a los tres clasificadores.

La estandarización de características (línea 4) se aplicó únicamente para SVM y MLP mediante StandardScaler, mientras que Random Forest se entrenó con los datos sin normalizar debido a su invarianza ante diferencias de escala. En la Fase 2 (líneas 15-19), se emplearon diferentes

métodos de búsqueda según el algoritmo: RandomizedSearchCV con 50 combinaciones para Random Forest, GridSearchCV con 30 combinaciones para MLP, y BayesSearchCV con 50 iteraciones para SVM.

Algorithm 1 Clasificador de litologías

- 1: **Preparación de Datos**
 - 2: Cargar datos preprocesados
 - 3: Seleccionar características (x) y variable objetivo (y)
 - 4: Estandarizar características (x) si el algoritmo lo requiere
 - 5:
 - 6: **División de Datos**
 - 7: División estratificada: 80 % entrenamiento y 20 % prueba
 - 8:
 - 9: **Fase 1: Evaluación con Configuración Predeterminada**
 - 10: Inicializar el modelo con parámetros por defecto
 - 11: Ajustar el modelo en el conjunto de entrenamiento
 - 12: Evaluar el desempeño en el conjunto de prueba
 - 13:
 - 14: **Fase 2: Optimización y Evaluación**
 - 15: Definir espacio de hiperparámetros según el algoritmo
 - 16: Seleccionar método de búsqueda apropiado
 - 17: Aplicar validación cruzada estratificada de 5 pliegues
 - 18: Entrenar nuevo modelo con los mejores hiperparámetros encontrados
 - 19: Evaluar el desempeño en el conjunto de prueba
 - 20:
 - 21: **Comparación de Resultados**
 - 22: Comparar métricas (F1-score macro) entre Fase 1 y Fase 2
 - 23: Analizar matrices de confusión por clase
-

Capítulo 4

Experimentos y Resultados

Este capítulo presenta los resultados experimentales obtenidos al aplicar Support Vector Machine, Random Forest y Multi-Layer Perceptron en la clasificación de litologías del pozo 25_7-2. Se evalúa el desempeño de cada algoritmo en dos fases: configuración por defecto (Fase 1) y tras optimización de hiperparámetros (Fase 2), utilizando F1-score macro y análisis de confusiones por clase.

La Tabla 4.1 muestra la distribución de clases en los conjuntos de entrenamiento y prueba tras la partición estratificada 80/20. Esta distribución revela un desbalance significativo, con las clases 3 y 19 presentando menos de 100 instancias en el conjunto de prueba.

Clase	Entrenamiento	Prueba	Total
2	2842	711	3553
3	361	90	451
5	2092	523	2615
6	1426	356	1782
11	1848	462	2310
12	2933	734	3667
13	2344	586	2930
19	275	69	344
Total	14121	3531	17652

Tabla 4.1. Frecuencia de mediciones por clase en entrenamiento y prueba

4.1. Configuraciones de los modelos

Esta sección describe las configuraciones y estrategias de optimización empleadas para cada uno de los algoritmos de clasificación evaluados. Se presentan los hiperparámetros predeterminados utilizados en la Fase 1, los rangos explorados durante la optimización (indicados como L.I. para límite inferior y L.S. para límite superior), y los valores optimizados encontrados para la Fase 2. La configuración óptima de hiperparámetros se realizó mediante validación cruzada estratificada de 5 pliegues en combinación con una técnica específica de búsqueda para cada algoritmo.

4.1.1. Random Forest

Para este modelo se implementó búsqueda aleatoria con 50 combinaciones de hiperparámetros. Los resultados se resumen en la Tabla (4.2).

Parámetro	Valor predeterminado	L.I.	L.S.	Valor optimizado
n_estimators	100	100	300	278
max_depth	None	10	30	25
min_samples_split	2	5	15	5
min_samples_leaf	1	3	10	3
max_features	sqrt	2	5	3

Tabla 4.2. Configuración de Random Forest

4.1.2. Multi-Layer Perceptron

Para este modelo se empleó búsqueda exhaustiva evaluando 30 combinaciones de hiperparámetros. Se exploraron cinco configuraciones de arquitectura (hidden_layer_sizes): (50,), (100,), (100, 50), (100, 100) y (100, 50, 25); tres valores para el parámetro de regularización alpha: 0.0001, 0.001 y 0.01; y dos valores para la tasa de aprendizaje inicial (learning_rate_init): 0.001 y 0.01. Los parámetros activation y solver se mantuvieron fijos en sus valores predeterminados (relu y adam respectivamente), mientras que max_iter se estableció en 1000 para evitar problemas de convergencia. La configuración óptima encontrada correspondió a hidden_layer_sizes=(100, 50), alpha=0.001 y learning_rate_init=0.01.

4.1.3. Support Vector Machine

Para este modelo se ejecutó búsqueda bayesiana durante 50 iteraciones, explorando un espacio bidimensional de hiperparámetros: el factor de regularización C en el rango de 0.1 a 100 y γ en el rango de 0.001 a 10. Ambos parámetros se muestrearon con distribución log-uniforme para explorar eficientemente varios órdenes de magnitud. El kernel se mantuvo fijo en RBF (Radial Basis Function). Los resultados se resumen en la Tabla (4.3).

Parámetro	Valor predeterminado	L.I.	L.S.	Valor optimizado
C	1.0	0.1	100	8.4
γ	scale	0.001	10	2.0
kernel	rbf	—	—	rbf

Tabla 4.3. Configuración de Support Vector Machine

4.2. Fase 1

Tras completar la primera fase para los modelos con su configuración predeterminada, se recolectaron los puntajes F1 por clase, como se muestra en Tabla (4.4).

Clase	SVM (train)	SVM (test)	RF (train)	RF (test)	MLP (train)	MLP (test)
2.0	0.89	0.88	1.00	0.92	0.91	0.90
3.0	0.35	0.30	1.00	0.60	0.54	0.44
5.0	0.79	0.75	1.00	0.80	0.81	0.78
6.0	0.86	0.80	1.00	0.84	0.87	0.83
11.0	0.81	0.79	1.00	0.90	0.84	0.83
12.0	0.83	0.82	1.00	0.89	0.84	0.85
13.0	0.87	0.88	1.00	0.92	0.88	0.90
19.0	0.66	0.67	1.00	0.79	0.73	0.71
F1-macro	0.76	0.74	1.00	0.83	0.80	0.78

Tabla 4.4. F1-score por clase (fase 1)

Random Forest obtuvo los mejores puntajes, tanto en rendimiento global (0.83) como en las clases minoritarias 3 (0.6) y 19 (0.79). En contrase, SVM obtuvo desempeño más bajo en ambos aspectos (ver Gráfica 4.1). Mientras que MLP (0.78) es un punto intermedio entre SVM y RF.

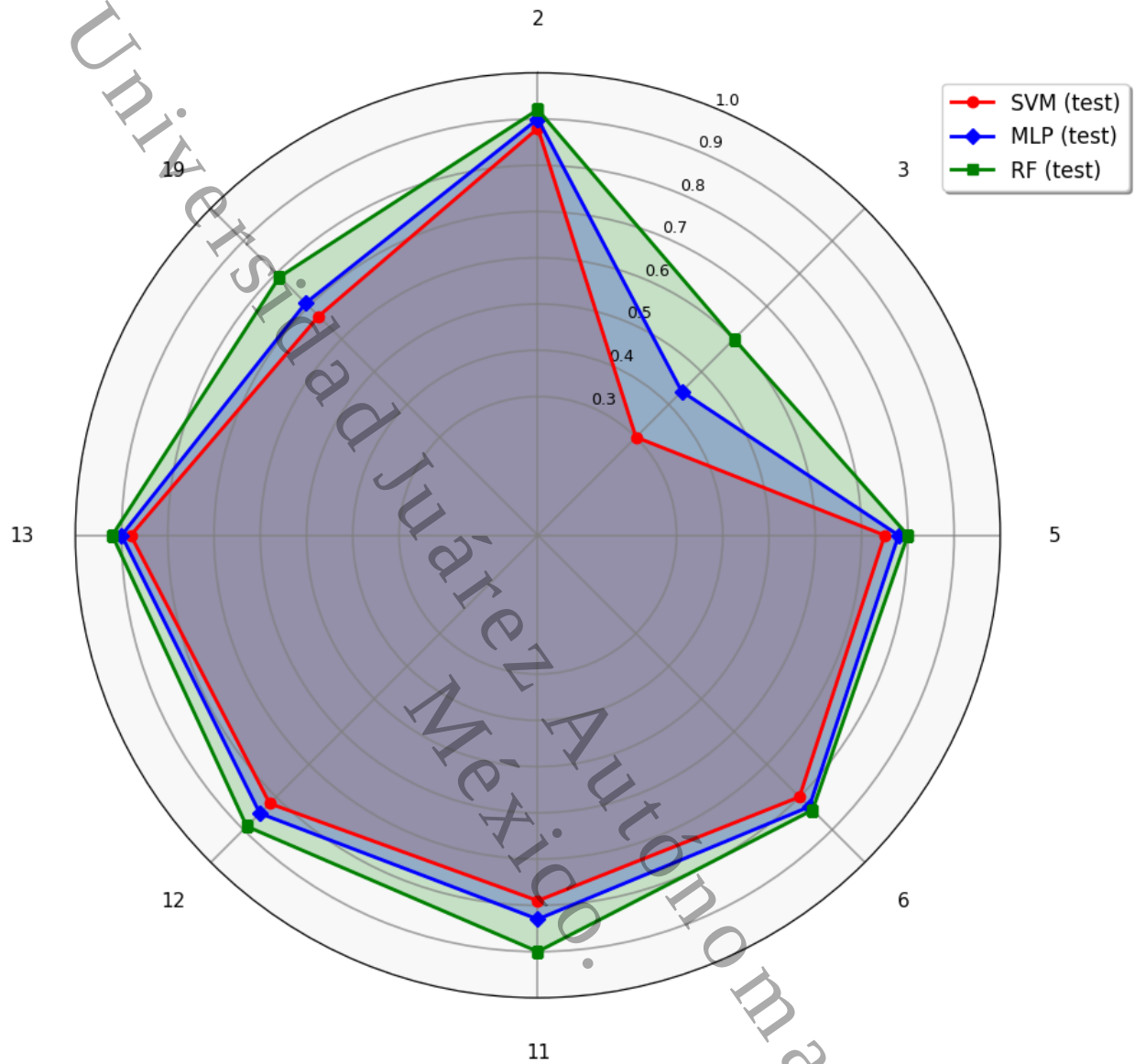


Figura 4.1. F1 score por clase (fase 1)

4.2.1. Fase 2

Tras completar la búsqueda de hiperparámetros para cada modelo, se obtuvieron los resultados mostrados en la Tabla (4.5).

Clase	SVM (train)	SVM (test)	RF (train)	RF (test)	MLP (train)	MLP (test)
2.0	0.95	0.92	0.99	0.91	0.92	0.89
3.0	0.88	0.57	0.98	0.62	0.63	0.55
5.0	0.85	0.79	0.99	0.79	0.88	0.77
6.0	0.91	0.83	0.99	0.83	0.89	0.82
11.0	0.93	0.82	1.00	0.89	0.85	0.84
12.0	0.84	0.86	1.00	0.88	0.86	0.86
13.0	0.87	0.89	0.99	0.92	0.86	0.89
19.0	0.79	0.74	0.97	0.77	0.79	0.73
F1-macro	0.88	0.80	0.99	0.82	0.85	0.79

Tabla 4.5. F1-score por clase (fase 2)

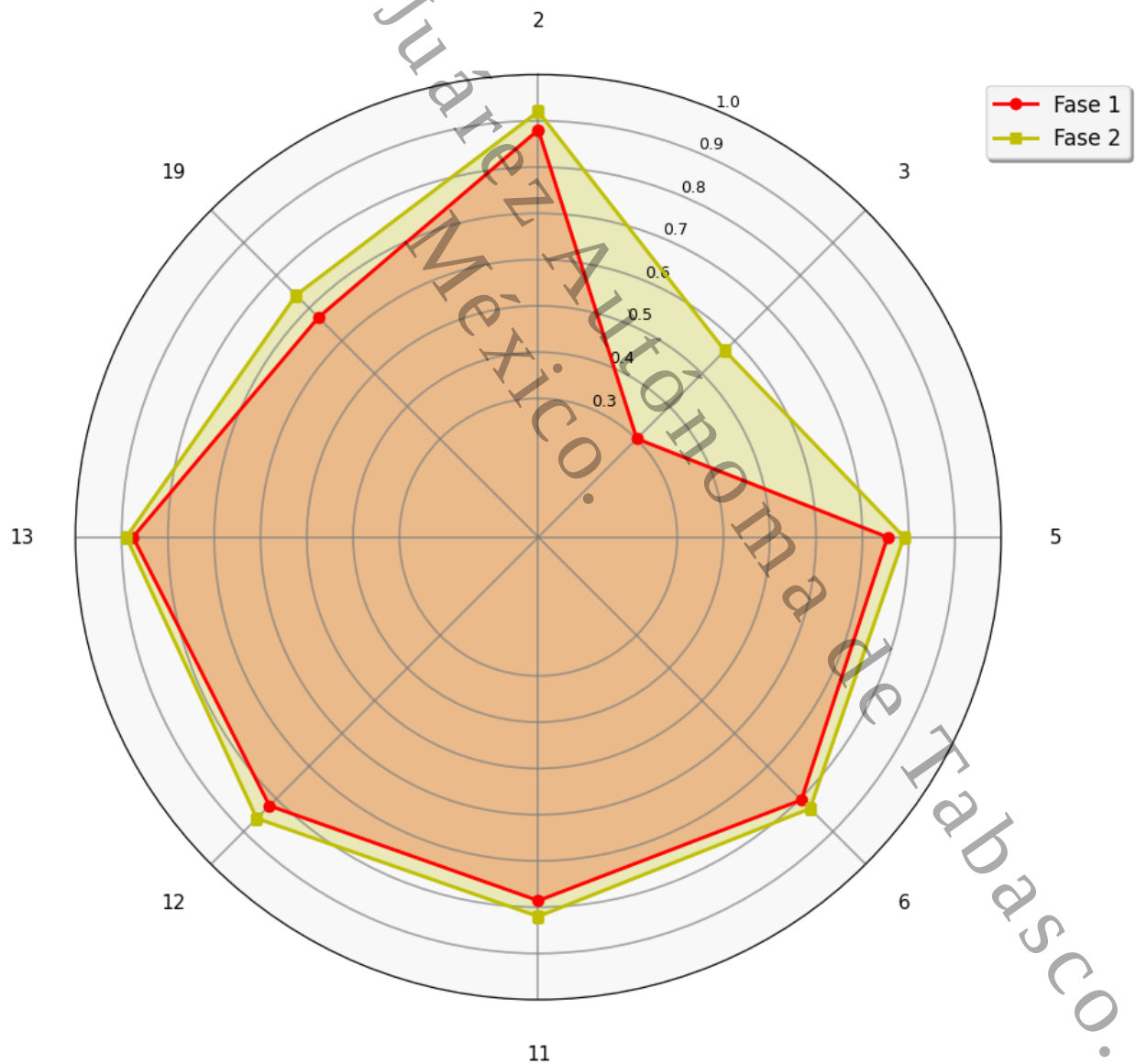


Figura 4.2. F1-Score por clase para SVM (fase 1 vs fase 2)

Support Vector Machine experimentó las mayores mejoras tras la optimización de hiperparámetros (ver gráfico 4.2). El F1-score macro aumentó de 0.74 a 0.80 (6 puntos porcentuales). En la clase 3, el F1 aumentó de 0.3 a 0.57 (27 pp) y para la clase 19 de 0.67 a 0.74 (7 pp). En las demás clases, los cambios fueron mínimos.

Los modelos MLP y RF presentaron cambios marginales que no justifican el costo de optimización. Incluso, algunos puntajes F1 por clase disminuyeron ligeramente ($\approx 1\%$). La optimización fue efectiva únicamente para SVM, cuya configuración predeterminada resultó inadecuada para el problema.

4.3. Análisis de confusiones

Desde el análisis exploratorio de datos, se había previsto que las clases con soporte crítico presentarían desafíos en la clasificación. En el conjunto de prueba, las clases 3 y 19 tienen 90 y 69 instancias, respectivamente, mientras que las clases más representadas, como 2 y 12, tienen 3553 y 3667 instancias. En el caso más crítico, esto representa una proporción 1:10 (90/734) entre la clase 3 y 12.

El caso más importante de confusión es entre la clase 3 y 2 (3 $\rightarrow$ 2), es decir, Cross-Bedded Sand clasificado como Sand (ver Tabla 4.6). Esto tiene sentido ya que geológicamente sus propiedades físicas son similares con matices que solo pueden discriminarse con otro tipo de estudios.

En este contexto, el MLP (fase 2) redujo esta confusión de 46.7% a 15.6%, logrando el mayor número de instancias (56) clasificadas correctamente para la clase 3. El SVM (fase 1) tuvo el peor rendimiento en esta clase, clasificando correctamente 17 de las 90 instancias.

Modelo	Clase 2 $\times$	Clase 3 $\checkmark$	Clase 5 $\times$	F1 de clase 3
SVM (Fase 1)	54 (60.0 %)	17 (18.9 %)	19 (21.1 %)	0.30
SVM (Fase 2)	22 (24.4 %)	46 (51.1 %)	22 (24.4 %)	0.57
RF (Fase 1)	25 (27.8 %)	47 (52.2 %)	18 (20.0 %)	0.60
RF (Fase 2)	30 (33.3 %)	43 (47.8 %)	17 (18.9 %)	0.62
MLP (Fase 1)	42 (46.7 %)	32 (35.6 %)	16 (17.8 %)	0.44
MLP (Fase 2)	14 (15.6 %)	56 (62.2 %)	20 (22.2 %)	0.55

Tabla 4.6. Confusiones en clase 3

La clase 19 (Silty Sand) también presenta patrones de confusión significativos (ver Tabla 4.7), pero con una estructura diferente. A diferencia de la clase 3, su principal confusión se da con la clase 5 (Sandy Shale), mientras que la confusión con la clase 2 (Sand) es mínima. Este fenómeno puede explicarse por las propiedades texturales intermedias que presenta la arenisca limosa entre arena y lutita arenosa. El Random Forest (fase 1) obtuvo el mejor rendimiento para esta clase, clasificando correctamente 56 instancias (81.2 %), seguido por el MLP (fase 2) con 54 instancias (78.3 %).

Modelo	Clase 2 $\times$	Clase 5 $\times$	Clase 19 $\checkmark$	F1 de clase 19
SVM (Fase 1)	2 (2.9 %)	23 (33.3 %)	44 (63.8 %)	0.67
SVM (Fase 2)	2 (2.9 %)	16 (23.2 %)	50 (72.5 %)	0.74
RF (Fase 1)	3 (4.3 %)	9 (13.0 %)	56 (81.2 %)	0.79
RF (Fase 2)	3 (4.3 %)	11 (15.9 %)	54 (78.3 %)	0.77
MLP (Fase 1)	3 (4.3 %)	15 (21.7 %)	50 (72.5 %)	0.71
MLP (Fase 2)	2 (2.9 %)	13 (18.8 %)	54 (78.3 %)	0.73

Tabla 4.7. Confusiones de la clase 19

En ambas tablas, se omitieron los casos de confusión con solo 1 o 2 instancias.

4.4. Discusión

Random Forest alcanzó el mejor desempeño con configuración predeterminada ($F1_{\text{macro}} = 0.83$), mientras que *Support Vector Machine* requirió optimización para volverse competitivo, mejorando de 0.74 a 0.80. Este contraste evidencia que RF es inherentemente robusto ante configuraciones no ajustadas, mientras que SVM depende críticamente de la calibración de sus hiperparámetros C y γ para el problema específico. *Multi-Layer Perceptron* mostró estabilidad con cambios marginales tras optimización (+0.01), sugiriendo que su arquitectura básica ya opera cerca de su capacidad máxima.

La optimización benefició a SVM, con mejora de 6 puntos porcentuales en $F1_{\text{macro}}$ y 32 pp en la clase 3 (de 19 % a 51 % de instancias correctas). Aunque Random Forest alcanzó el mejor $F1$ -score para esta clase tras optimización (0.62), *Multi-Layer Perceptron* logró clasificar correctamente el mayor número absoluto de instancias (56/90), evidenciando diferentes compromisos entre precisión y *recall*. En contraste, RF y MLP presentaron cambios globales menores al 2 %,

indicando que sus configuraciones predeterminadas resultan apropiadas para este problema.

Las confusiones bidireccionales consistentes entre subtipos litológicos como limolitas (5, 6) y calizas (11, 12, 13) a través de los tres algoritmos sugieren que el conjunto de cinco registros geofísicos empleados alcanza un límite de discriminación para ciertas litologías. La coincidencia de estas confusiones independientemente de la arquitectura del modelo indica que tales limitaciones no son algorítmicas, sino inherentes a la información disponible.

Capítulo 5

Contribuciones, conclusiones y trabajos futuros

Este capítulo sintetiza las contribuciones metodológicas y hallazgos principales del trabajo, valida las hipótesis planteadas, responde las preguntas de investigación, y propone direcciones para investigación futura en clasificación automatizada de litologías.

5.1. Conclusiones

Con respecto a la primera pregunta sobre la capacidad de los algoritmos con configuraciones por defecto para manejar el desbalance de clases, la respuesta es sí, Random Forest alcanzó un buen desempeño, tanto a nivel global ($F1\text{-macro} = 0.83$) como en clases minoritarias (ver Tabla 4.4). Random Forest se beneficia de su mecanismo de votación y la construcción de múltiples árboles con muestras bootstrap.

El MLP también mostró desempeño competitivo con configuración por defecto ($F1\text{-macro} = 0.78$), aunque inferior a Random Forest. En contraste, SVM presentó el peor desempeño ($F1\text{-macro} = 0.74$), especialmente problemático en la clase minoritaria 3 (areniscas con estratificación cruzada) donde solo clasificó correctamente el 18.9% de las instancias, obteniendo $F1 = 0.30$.

En conclusión: Random Forest y MLP pueden emplearse efectivamente con configuraciones por defecto en problemas de clasificación de litologías con desbalance, mientras que SVM requiere necesariamente un proceso de optimización de hiperparámetros para alcanzar desempeño

competitivo.

Con respecto a la segunda pregunta, sobre cómo influye la complejidad del espacio de hiperparámetros de cada algoritmo en su desempeño final tras el proceso de optimización, la respuesta es: el análisis comparativo de las fases 1 y 2 revela que la complejidad del espacio de hiperparámetros se relaciona inversamente con la necesidad de optimización. SVM con un espacio bidimensional relativamente simple (C y γ), experimentó las mejoras más significativas tras optimización bayesiana. El incremento de 6 puntos porcentuales en F1-macro y la mejora dramática en la clase 3 (de 18.9 % a 51.1 % en el puntaje F1) demuestran que SVM es altamente sensible a la selección de hiperparámetros, pero una vez optimizados, puede alcanzar desempeño competitivo.

Random Forest, a pesar de explorar un espacio de cinco hiperparámetros ($n_estimators$, max_depth , $min_samples_split$, $min_samples_leaf$, $max_features$) mediante búsqueda aleatoria con 50 iteraciones, presentó cambios marginales. Su F1-macro disminuyó ligeramente de 0.83 a 0.82. Esta estabilidad indica que la configuración predeterminada de Random Forest ya opera cerca de su capacidad óptima, y la complejidad de su espacio de hiperparámetros no se traduce en mejoras sustanciales.

Multi-Layer Perceptron, con exploración exhaustiva de 30 combinaciones que incluyen arquitectura de red ($hidden_layer_sizes$), regularización ($alpha$) y tasa de aprendizaje ($learning_rate_init$), mostró cambios mínimos. Su F1-macro aumentó de 0.78 a 0.79 con variaciones por clase menores al 2 %. Aunque su espacio de configuración es conceptualmente complejo debido a la interacción entre arquitectura y parámetros de entrenamiento, la configuración predeterminada resultó apropiada para este problema específico.

Estos resultados evidencian un patrón contraintuitivo: la complejidad del espacio de hiperparámetros no predice la magnitud de mejora tras optimización. SVM, con el espacio más simple, obtuvo las mejoras más significativas, mientras que RF y MLP, con espacios más complejos, mostraron cambios marginales. Esto sugiere que la necesidad de optimización depende más de qué tan apropiados sean los valores predeterminados para el problema específico que de la dimensionalidad del espacio de configuración.

Con respecto a la hipótesis, los resultados experimentales confirmaron que Random Forest sí demostró mayor robustez con configuración predeterminada (F1-macro = 0.83), superando a

SVM (0.74) y MLP (0.78). Su desempeño en clases minoritarias fue superior para la clase 3 ($F1 = 0.60$), en comparación con SVM ($F1=0.30$) y MLP ($F1=0.44$); para la clase 19 obtuvo el mejor puntaje ($F1=0.79$), superando a MLP ($F1=0.71$) y SVM ($F1=0.67$). Esta superioridad se atribuye efectivamente a su invarianza ante diferencias de escala, ya que las decisiones basadas en umbrales individuales por característica no se ven afectadas por los rangos de los registros geofísicos (RDEP: 0.48-79.39 ohm·m, DTC: 51.7-139.4 $\mu\text{s/ft}$, GR: 12.2-191.0 gAPI).

La hipótesis se confirma: Random Forest presenta mayor robustez inicial y no requiere normalización, factores que facilitan su aplicación directa en problemas de clasificación litológica. Sin embargo, la brecha de desempeño entre algoritmos no es permanente ni insuperable: con optimización apropiada, SVM alcanza desempeño comparable a RF, y MLP mantiene competitividad consistente. Por lo tanto, la elección del algoritmo debe considerar no solo la robustez inherente, sino también los recursos disponibles para optimización y las características específicas del problema geológico.

5.2. Contribuciones

Este trabajo aporta una comparación sistemática de tres algoritmos de aprendizaje automático (SVM, Random Forest y MLP) en la clasificación de litologías con desbalance de clases, utilizando únicamente registros geofísicos del pozo 25_7-2.

Los resultados demuestran que Random Forest alcanza buen desempeño con configuración predeterminada ($F1\text{-macro} = 0.83$), mientras que SVM requiere optimización de hiperparámetros para volverse competitivo (mejora de 0.74 a 0.80). Estas observaciones pueden orientar la selección de algoritmos en proyectos similares de clasificación litológica.

El modelo presenta limitaciones en la clasificación de litologías transicionales, inherentes a la gradación natural de propiedades físicas entre tipos de roca. La validación del modelo se limitó a un solo pozo, por lo que su aplicabilidad en otros contextos geológicos requiere verificación adicional. El modelo podría servir como herramienta de apoyo para caracterización preliminar cuando se dispone de datos etiquetados de pozos cercanos, aunque siempre requeriría validación y ajuste por parte de especialistas en geología.

5.3. Trabajos Futuros

La extensión natural de este trabajo consiste en evaluar la transferibilidad de los modelos a múltiples pozos de uno o varios campos petroleros de la base de datos Geolink. Esta extensión enfrenta desafíos específicos:

- Variabilidad local de las propiedades físicas del subsuelo entre pozos.
- Diversidad litológica dentro de un mismo campo.
- Diferencias en calibración y precisión entre herramientas de perfilaje.
- Disponibilidad parcial de registros geofísicos en diferentes pozos.
- Datos incompletos o de baja calidad en pozos más antiguos.

Adicionalmente, se proponen las siguientes líneas de investigación:

- Incorporar registros geofísicos adicionales (imágenes de pozo, resonancia magnética nuclear) para mejorar la discriminación entre litologías transicionales.
- Evaluar técnicas de balanceo de clases (SMOTE, ajuste de pesos) para mejorar el desempeño en clases minoritarias.
- Implementar estrategias de transfer learning para aprovechar conocimiento de pozos etiquetados en la clasificación de pozos nuevos.
- Validar los resultados de clasificación con geólogos expertos para cuantificar la confiabilidad práctica del modelo.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	Modelo de clasificación de litologías empleando aprendizaje automático
Autor:	Roberto Karlo Gomez Collado
ORCID:	https://orcid.org/0009-0000-2691-030X
Resumen:	El presente trabajo compara el desempeño de tres algoritmos de aprendizaje automático: <i>Support Vector Machine</i> (SVM), <i>Random Forest</i> (RF) y <i>Multi-Layer Perceptron</i> (MLP), en clasificación de litologías. Los datos empleados son registros geofísicos del pozo petrolero 25_7-2, proveniente de la base de datos abierta de Geolink. Se emplea validación cruzada estratificada y partición de datos para evaluar el desempeño global y la capacidad de cada modelo para manejar clases desbalanceadas. El proceso de clasificación se estructuró en dos fases experimentales para cada uno de los tres algoritmos evaluados. La estandarización de características se aplicó únicamente para SVM y MLP mediante <i>StandardScaler</i>, mientras que <i>Random Forest</i> se entrenó con los datos sin normalizar debido a su invarianza ante diferencias de escala. En la primera fase, se evaluaron los clasificadores con su configuración predeterminada. En la Fase 2, se emplearon diferentes métodos de búsqueda según el algoritmo: <i>RandomizedSearchCV</i> con 50 combinaciones para <i>Random Forest</i>, <i>GridSearchCV</i> con 30 combinaciones para MLP, y <i>BayesSearchCV</i> con 50 iteraciones para SVM. <i>Random Forest</i> alcanzó el mejor desempeño con configuración predeterminada ($F1_{macro} = 0.83$), mientras que <i>Support Vector Machine</i> requirió optimización para volverse competitivo, mejorando de 0.74 a 0.80. Este contraste evidencia que RF es inherentemente robusto ante configuraciones no ajustadas, mientras que SVM depende críticamente de la calibración de sus hiperparámetros C y γ para el problema específico. <i>Multi-Layer Perceptron</i> mostró estabilidad con cambios marginales tras optimización (+0.01), sugiriendo que su arquitectura básica ya opera cerca de su capacidad máxima. La hipótesis se confirma, ya que <i>Random Forest</i> presenta mayor robustez inicial y no requiere normalización, factores que facilitan su aplicación directa en problemas de clasificación litológica.
Palabras clave:	registros geofísicos, pozos petroleros, Python, modelos de clasificación.
Referencias citadas:	En la siguiente página se muestran las referencias.

Bibliografía

- Al-Anazi, A., & Gates, I. (2010). On the capability of support vector machines to classify lithology from well logs. *Natural Resources Research*, 19(2), 125-139.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
- Glover, P. W. J. (2000). *Petrophysics: MSc Course Notes* [Section: "Uses of the Caliper Log", p. 92]. https://d1wqtxts1xzle7.cloudfront.net/60661185/_Glover_Paul_W.J._Petrophysicsz-lib.org20190921-75592-rvta6w-libre.pdf
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.
- Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5), 429-449.
- Kansas Geological Survey. (2017a). The Gamma Ray Log [Updated July 31, 2017]. *Kansas Geological Survey, University of Kansas*.
- Kansas Geological Survey. (2017b). Nuclear Logging – Nuclear Porosity Logs [Updated July 31, 2017].
- London, I. C. (2000). *Acquisition And Interpretation of Wireline Logs* (inf. téc.) (PDF; sección "Bulk density log (RHOB)", páginas 42-44). Imperial College London. <https://vdoc.pub/documents/acquisition-and-interpretation-of-wireline-logs-44c6ml6o9p80>
- Merembayev, T., Yunussov, R., & Yedilkhan, A. (2018). Machine learning algorithms for classification geology data from well logging. *2018 14th International Conference on Electronics Computer and Computation (ICECCO)*, 206-212.
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Mohamed, I. M., Mohamed, S., Mazher, I., & Chester, P. (2019). Formation lithology classification: Insights into machine learning methods. *SPE Annual Technical Conference and Exhibition?*, D021S033R005.
- Ren, X., Hou, J., Song, S., Liu, Y., Chen, D., Wang, X., & Dou, L. (2019). Lithology identification using well logs: A method by integrating artificial neural networks and sedimentary patterns. *Journal of Petroleum Science and Engineering*, 182, 106336.

- Rohit, Manda, S. R., Raj, A., Dheeraj, A., Rawat, G. S., & Choudhury, T. (2024). Identification of Lithology from Well Log Data Using Machine Learning. *EAI Endorsed Transactions on Internet of Things*, 10(1), e4. <https://doi.org/10.4108/eetiot.5634>
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088), 533-536.
- Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th). Pearson.
- Stone, M. (1974). Cross-validators choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111-147.
- Zhang, J., He, Y., Zhang, Y., Li, W., & Zhang, J. (2022). Well-Logging-Based Lithology Classification Using Machine Learning Methods for High-Quality Reservoirs Identification: A Case Study of Baikouquan Formation in Mahu Area of Junggar Basin, NW China. *Energies*, 15(10), 3675.