Desarrollo y aplicación de algoritmos de análisis de datos orientados a facilitar la descripción de ambientes sedimentarios aplicados a E&P.

Tesis Doctoral

Lic. Isabel Eugenia Giannoni

Directores

Dr. Augusto Nicolás Varela

Dr. Javier Vásquez

Facultad de Ciencias Naturales y Museo

2025

UNIVERSIDAD NACIONAL DE LA PLATA

CONICET

Y-TEC

YPF TECNOLOGÍA

Índice de Contenidos

Resumen

Las facies sedimentarias, un concepto fundamental en geología, refiere a las características físicas, químicas y biológicas de una unidad de roca sedimentaria que se formó por un proceso sedimentario específico. Estas características incluyen la textura, composición y estructuras sedimentarias (mecánicas y biogénicas) presentes en la roca. Las asociaciones de facies son conjuntos de diferentes facies sedimentarias que están genética y espacialmente relacionadas y permiten a los geólogos reconstruir paleoambientes sedimentarios para así entender los cambios geológicos a lo largo del tiempo. La definición y estudio de facies sedimentarias y sus asociaciones resulta especialmente útil en la exploración y explotación de hidrocarburos y otros recursos, donde la comprensión del paleoambiente sedimentario es fundamental para la toma de decisiones.

Por otro lado, el aprendizaje automático, ha evolucionado hasta convertirse en una presencia omnipresente en la sociedad actual. Gracias a su capacidad para aprender patrones complejos a partir de grandes conjuntos de datos, se ha convertido en la base de aplicaciones actuales; transformando así la manera en que interactuamos con la tecnología y abordamos desafíos en diversos campos. Dentro de la geología, en campos como la sedimentología ha permitido realizar análisis más rápidos, precisos y detallados de las rocas.

El objetivo principal de la presente tesis doctoral es la aplicación de algoritmos incluidos dentro de lo que se denomina aprendizaje automático a imágenes de alta y baja resolución de muestras de testigos coronas, datos geoquímicos y perfiles de resistencia al rayado con el fin de clasificar, inferir y predecir patrones de comportamiento asociados a las facies sedimentarias y sus asociaciones de facies con sus correspondientes procesos y ambientes sedimentarios. Un objetivo secundario, pero no por ello menos importante, de este trabajo doctoral es reducir la subjetividad del observador, mejorando la precisión de las clasificaciones, además de disminuir los tiempos de descripción; permitiendo comparar de manera directa las descripciones de diferentes observadores.

El proceso de análisis consistió en varias etapas, en las que se buscó asegurar la calidad de la información. Inicialmente se realizó la estandarización, homogeneización y el análisis estadístico exploratorio de los datos, así como también la generación de nuevas variables que puedan ser de utilidad.

Como resultado, se obtuvieron tres modelos de aprendizaje automático capaces de predecir la textura granulométrica, la estructura sedimentaria y la composición de las rocas con una exactitud del 65% en el caso de las dos primeras y de 76% para el caso de la composición química. Estas métricas de evaluación muestran la predicción indirecta de las facies sedimentarias a partir de sus características.

Debido que para validar las clasificaciones obtenidas fue necesario la utilización de criterios sedimentológicos, se llegó a la conclusión de que, estas técnicas son herramientas que reducen las subjetividades descriptivas de diferentes observadores respecto de los procesos de análisis sedimentarios.

Abstract

Sedimentary facies, a primary concept in geology, refers to the physical, chemical and biological characteristics of a sedimentary rock unit that was formed by a specific sedimentary process. These characteristics include the texture, composition and sedimentary structures (mechanical and biogenic) present in the rock. Facies associations are assemblages of different sedimentary facies that are genetically and spatially related and allow geologists to reconstruct sedimentary paleoenvironments to understand geologic changes over time. The definition and study of sedimentary facies and their associations is especially useful in the exploration and exploitation of hydrocarbons and other resources, where understanding the sedimentary paleoenvironment is fundamental for decision making.

On the other hand, Machine learning, has evolved to become a ubiquitous presence in today's society. Thanks to its ability to learn complex patterns from large data sets, it has become the basis of current applications; thus, transforming the way we interact with technology and address challenges in various fields. Within geology, in fields such as sedimentology, it has enabled faster, more accurate and detailed analyses of rocks.

The main objective of this PhD thesis is the application of algorithms included in what is called machine learning to high- and low-resolution images of core samples, geochemical data and scratch resistance profiles in order to classify, infer and predict behavioral patterns associated with sedimentary facies and their facies associations with their corresponding sedimentary processes and environments. A secondary objective is to reduce the subjectivity of the observer, improving the accuracy of the classifications and reducing the description times; allowing direct comparison between the descriptions of different observers.

The analysis process consisted of several stages, in which we sought to ensure the quality of the information. Initially, standardization, homogenization and exploratory statistical analysis of the data were carried out, as well as the generation of new variables that could be useful.

As a result, three machine learning models capable of predicting granulometric texture, sedimentary structure and rock composition were obtained with an accuracy of 65% in the case of the first two and 76% for chemical composition. These evaluation metrics show the indirect prediction of sedimentary facies from their characteristics.

Since the use of sedimentological criteria was necessary to validate the classifications obtained, it was concluded that these techniques are tools that reduce the descriptive subjectivities of different observers with respect to sedimentary analysis processes.

Agradecimientos

Creo firmemente que esta tesis no es solo mía, sino que le pertenece a un montón de personas que hicieron que este camino fuera más ameno y me permitiera crecer no solo como profesional sino también como persona. Debo decir (o mejor escribir) que en retrospectiva y haciendo un balance, lo he disfrutado enormemente. Culminar esta etapa representa para mí no solo un logro académico, sino también finalizar un viaje lleno de desafíos y aprendizajes que me deja en la puerta de un nuevo mundo, la ciencia de datos aplicada directamente a las geociencias.

Quiero agradecer en primer lugar a mis directores de tesis, los Dres. Augusto Varela y Javier Vasquez, por confiar en mi para este tema tan poco explorado dentro de la geología. A Juan Pablo Alvarez, jefe de Subsuelos de Y-TEC cuando comencé la tesis, por confiar en la decisión que tomaron Augusto y Javier. La visión a futuro y el empuje de los tres hizo que esta tesis fuera factible. No dejar de nombrar a Martin Sánchez, jefe actual y de los últimos 4 años, 11 meses y semanas, que al tomar el mando de ese barco que era la Misión de Tecnologías de Subsuelo, nos apoyó y nos guio sabiamente.

Tuve la suerte de encontrar en Y-TEC, un excelente grupo personas que me recibieron de la mejor manera, me apoyaron, me enseñaron y me sostuvieron a lo largo de todos estos años. No puedo nombrarlos por miedo a olvidarme de alguien, pero creo que no hay persona que no me haya dado un consejo, una palabra de aliento, un mate para levantar el día ¡Muchas gracias a todos!. Siento la necesidad de hacer una mención especial a la coordinación de Geología y Geoquímica quienes han sido para mí un refugio indispensable. Ojalá todas los grupos de trabajo fueran como el nuestro. Remi, Tincho, Dolo, Gi, Cami, ¡gracias!.

No quiero dejar de agradecer también, a Marian, con quien trabaje codo a codo todo este tiempo y que supo sobrellevar mi obsesiva forma de trabajar; y a Lucre, por ser durante todos estos años, una gran compañera de banco. Por

aguantar el día a día, las reuniones, las esperas entre entrenamiento y entrenamiento con grandes charlas con mates y café.

A toda mi enorme familia, su incondicionalidad, apoyo, amor y compañía han sido vitales a lo largo de este proceso. A Isabel y Marcelo que me dieron la confianza y la libertad para ir tras mis sueños y me ayudan a cumplirlos día a día, los amo. A mis amigos, los de la vida y los de la facultad, por aguantar mis ausencias para trasnochar trabajando; no sé qué sería de mí sin ustedes.

A la gente de EPSLog. S.A, por aportar su base de datos para esta tesis. Por enseñarme sobre la temática y recibirme cálidamente durante mi estancia en su fábrica y centros de análisis de datos. Realmente me sentí parte de lo que es la EPSLog Family, compartiendo momentos hermosos. Fueron un gran apoyo que hicieron que la distancia con mi gente no se sintiera tanto.

Gracias a la Facultad de Ciencias Naturales y Museo, por formarme no solo como geóloga, sino como docente e investigadora. A la Universidad Nacional de La Plata, por una educación pública, gratuita y de calidad. Que sea inclusiva y que esté siempre al alcance de todos.

¡Gracias a todos! 💙

1

Introducción

1.1 Introducción general

Las facies sedimentarias han sido un concepto clave en la conformación de la sedimentología moderna. A partir de su entendimiento y por consecuente, del entendimiento de los procesos sedimentarios que las generan, los geólogos han sido capaces de entender las variaciones de los diferentes parámetros en los distintos ambientes depositacionales (Arche, 2010). La definición y estudio de facies sedimentarias y sus asociaciones de facies resulta especialmente útil en la exploración y explotación de hidrocarburos y otros recursos naturales, donde la comprensión de los procesos sedimentarios y de los paleoambientes sedimentarios son fundamentales para la toma de decisiones como fue planteado por Andrew Miall desde sus trabajos iniciales (Miall, 1977, 1988, 2022).

El aprendizaje automático (ML por sus siglas en inglés), por su parte, es un campo de estudio que en las últimas décadas ha ganado amplia popularidad. Se enfoca en el aprendizaje, la identificación de patrones y el razonamiento para la toma de decisiones, tratando se simular el razonamiento humano (Nilsson, 1998; Posthoff, 2004; Russell & Norvig, 2016; Gómez & Camilion, 2025). Al ser un campo de estudio transversal y poseer diversas aplicaciones, ha generado a través de los años varias subdisciplinas que se aplican en varios campos de generación de conocimientos buscando recrear el funcionamiento del pensamiento humano para resolver problemas complejos (Géron, 2022; Jiang et al., 2022).

Si bien la sedimentología ha avanzado significativamente, la irrupción del ML ha abierto nuevas fronteras en esta disciplina. Los modelos de aprendizaje, capaces de identificar patrones sutiles y complejas relaciones entre variables, ofrecen un potencial sin precedentes para profundizar en el conocimiento de los procesos sedimentarios y mejorar la precisión de las predicciones geológicas (Kanevski et al.,2004; Kanevski et al.,2015; Lary et al.,2016; Castillo Gamarra et al., 2018; Reichstein et al., 2019; Dramsch, 2020; Gómez & Camilion, 2025). Esta combinación de la experiencia geológica y la potencia computacional representa un salto cualitativo en la exploración de recursos naturales.

Impulsada por el desarrollo de nuevos algoritmos, mejoras en las infraestructuras de la nube y el desarrollo de procesadores de alto rendimiento, el ML evoluciona constantemente y cada vez más rápido. Cada año surgen nuevos modelos, que mejoran y optimizan los resultados de los anteriores. Sin embargo, los principios metodológicos y estadísticos en los que se basa la IA continúan vigentes. De esta manera, esta tesis doctoral tiene un componente fuertemente metodológico más allá de los modelos de aprendizaje automático presentados.

1.2 Objetivos

1.2.1 Objetivos generales

El objetivo de la presente tesis doctoral es clasificar, inferir y predecir los patrones de comportamiento asociados a las facies sedimentarias y sus asociaciones de facies con sus correspondientes procesos y ambientes sedimentarios con el fin de reducir la subjetividad del observador, mejorando la precisión de las clasificaciones, además de disminuir los tiempos de descripción; permitiendo comparar de manera directa las descripciones de diferentes observadores. Esto se realizará a partir de la aplicación de algoritmos de aprendizaje automático a imágenes de alta y baja resolución de muestras de testigos coronas, datos geoquímicos y perfiles de resistencia a la rotura.

1.2.2 Objetivos específicos

Para lograr el objetivo general propuesto se plantean los siguientes objetivos específicos:

- Describir la sedimentología e icnológica de los testigos coronas seleccionados a partir de imágenes de corona de baja y alta resolución, curvas geoquímicas de elementos a partir de Fluorescencia de Rayos X (FRX) y curvas de resistencia al rayado.
- Determinar un código de facies estandarizado que, conservando el criterio geológico, pueda a su vez ser utilizado para el análisis de datos automático.
- Homogeneizar y estandarizar los datos geológicos.
- Realizar el análisis exploratorio estadístico de los datos.
- Reconocer los patrones de facies (texturales, composicionales, estructuras sedimentarias tanto mecánicas como biogénicas, petrofísicos y visuales) a través de aprendizaje automático.
- Analizar los diferentes patrones de facies reconocidos y su vinculación con los procesos sedimentológicos que le dieron origen.
- Seleccionar el o los algoritmos adecuados para la resolución de cada una de las problemáticas planteadas.
- Predecir la ocurrencia de patrones de facies y asociaciones de facies a través de algoritmos de clasificación.

1.3 Conceptos teóricos básicos

Dado que este trabajo doctoral es un trabajo de análisis de datos aplicado a un subcampo de la geología como es la sedimentología, se considera un trabajo donde se combinan dos campos de diferentes naturalezas. Es por eso, que los siguientes apartados tienen por objetivos introducir a los lectores no especializados en los conceptos básicos tanto del objeto de estudio (facies sedimentarias), como de las técnicas utilizadas para su análisis (modelos de aprendizaje automático).

1.3.1 Conceptos de Facies Sedimentarias

Las facies sedimentarias son depósitos con características físicas y químicas específicas que se formaron a través de un proceso sedimentario específico. Estas características incluyen la composición presente en la roca, la textura y las estructuras sedimentarias -mecánicas y biogénicas- (Arche, 2010; Reading, 2009; Tucker et al., 2023). Otros autores amplían esta definición sumando atributos como paleocorrientes, trazas fósiles, contenido paleontológico, entre otros (Gressly, 1838; Selley, 1976, Bossi, 2007).

La historia de los conceptos de facies sedimentarias se remonta a los trabajos pioneros de geólogos como Walther, Gressly y Smith en el siglo XIX. Walther (1894) formuló la "ley de la sucesión de facies", la cual establece que las facies depositadas en continuidad lateral también aparecen en sucesión vertical en los registros estratigráficos, bajo condiciones de sedimentación ininterrumpida. En el siglo XX, el concepto de facies sedimentarias fue desarrollado de manera más completa gracias a investigadores como Dunham, Embry y Klovan, Miall y Pettijohn (Arche, 2010). Estos autores buscaron facilitar la comunicación y el estudio de las facies sedimentarias, desarrollando diversos códigos de facies que permiten a los geocientistas clasificar las rocas de manera sistemática. Uno de los sistemas de clasificación más influyentes es el de Dunham (1962), enfocado en las facies carbonáticas. Dunham desarrolló un esquema basado en la textura de la roca y la relación entre los componentes esqueléticos y no esqueléticos, clasificando las rocas carbonáticas en términos como mudstone, wackestone, packstone, grainstone y boundstone. Embry & Klovan (1971) modificaron el esquema de Dunham para incluir categorías adicionales que reflejan mejor la variedad de texturas en los carbonatos fósiles. Más recientemente, Lokier & Al Junaibi (2016) propusieron ajustes adicionales al esquema de Dunham para adaptarlo a las observaciones modernas y técnicas analíticas avanzadas. En el contexto de los sistemas silicoclásticos, Miall (1977, 1978,1996) desarrolló un enfoque basado en las texturas, las estructuras sedimentarias y la geometría de los cuerpos de rocas. Este enfoque permite una interpretación detallada de los ambientes fluviales y sus variaciones espaciales y temporales.

A través del estudio de estas facies y sus asociaciones, los geólogos pueden entender y reconstruir los paleoambientes sedimentarios, proporcionando una visión detallada de los cambios geológicos a lo largo del tiempo (Catuneanu, 2006; Walker, 1992). Las facies sedimentarias son interpretaciones del entorno de depósito de los sedimentos que luego generarán una roca sedimentaria, y su análisis se basa en observaciones de campo y de laboratorio. Este enfoque integral y multiescalar permite a los geólogos realizar interpretaciones ambientales y paleoambientales precisas. En la industria del petróleo, el análisis de facies es crucial para la exploración y desarrollo de reservorios, ya que las características de las facies afectan la porosidad y permeabilidad de las rocas reservorio; así como diferentes características de la roca madre y/o sellos (Walker, 1992; 2006).

Composición de las rocas

La composición de las rocas refiere a los materiales orgánicos (restos de organismos) o inorgánicos (minerales) que componen la roca. Las rocas sedimentarias pueden ser clasificadas de muchas formas distintas. En función de la composición de los materiales que la componen, pueden clasificarse como silicoclásticas, carbonáticas, mixtas y volcaniclásticas. Esta clasificación permite comprender su origen, los ambientes sedimentarios donde se formaron y los procesos sedimentarios que intervinieron en dicha formación.

Las **rocas silicoclásticas** son las que se forman a partir de la erosión, transporte y posterior depositación de detritos mayoritariamente de origen silíceo como cuarzos, feldespatos y micas. Los granos de cuarzo son los más abundantes en estas rocas dado su resistencia química y mecánica, esto hace que además de ser los más abundantes, soporten mejor el transporte llegando a zonas de depositación más lejanas. A su vez, los granos de minerales de la familia de los feldespatos son menos resistentes al transporte y la erosión en comparación con el cuarzo, lo que resulta en que estos granos se conserven en depósitos más cerca de la fuente de su origen y no tanto en depocentros lejanos a dicha fuente. Las rocas que poseen minerales de la familia de las micas, al sufrir erosión se rompen en pequeñas estructuras laminares de muy poco peso

que son transportadas en forma de suspensión. Además de estos minerales, las rocas silicoclásticas pueden contener fragmentos de otras rocas preexistentes, a los que se los denomina fragmentos líticos (Folk, 1980).

Por otro lado, las **rocas carbonáticas** se forman en ambientes cálidos, son producto de procesos químicos, biológicos y biogénicos. Están compuestas mayoritariamente por minerales como calcita, aragonita y dolomita. La calcita es el mineral más abundante, seguido de la aragonita, mientras que la dolomita se forma principalmente a través de procesos de diagénesis. Además de estos minerales, podemos encontrar como componente de una roca carbonática, bioclastos (partes de organismos), oolitas (granos esféricos con capas concéntricas de carbonato formados por precipitación química en agua agitadas), pellets (granos micríticos producidos por la actividad de microorganismos), entre otros, estos componentes que no se generan *in situ* en el lugar de depositación por los que se los llama componentes alóctonos. Químicamente, las rocas carbonáticas están dominadas por óxidos de calcio y magnesio, cuya proporción determina la cantidad relativa de calcita y dolomita. Además, impurezas como sílice, óxidos de hierro y materia orgánica pueden estar presentes, aportando colores y propiedades texturales específicas (Dunham, 1962; Tucker & Wright, 2009).

Dado que los ambientes naturales no son estáticos, sino que evolucionan unos en otros continuamente, los distintos tipos de rocas, sobre todo las silicoclásticas y las carbonáticas pueden mezclarse en ambientes transicionales. A este tipo de rocas que poseen tanto componentes silicoclásticos como carbonáticos, se las denomina rocas mixtas y poseen una subcategoría propia. Esto sucede porque no se puede establecer un límite entre los componentes de ambos tipos de rocas, lo que las hace muy complejas (Pettijohn, 1957; Reading, 1978).

Por último, las **rocas volcaniclásticas** están formadas principalmente por fragmentos volcánicos que se han generado, transportado y depositado a través de procesos volcánicos, sedimentarios o una combinación de ambos. Estos depósitos están conformados por fragmentos vítreos, cristalinos o líticos. Los fragmentos vítreos son las trizas vítreas y los fragmentos pumíceos o pómez y se presentan con distintas formas y grados de vesicularidad, lo que permite

interpretar los procesos eruptivos y de transporte. Los fragmentos cristalinos se componen de cristales desarrollados en el magma previos a la erupción que son liberados de forma independiente o dentro de los fragmentos pumíceos. Los fragmentos líticos son fragmentos de otras rocas que pueden ser parte de porciones de la roca de caja del reservorio magmático (cognatos) o pueden provenir de erupciones previas -accesorios- (Heiken, 1972; 1974; Teruggi et al., 1978; Wohletz, 1983; Sheridan & Marshall, 1983; Petrinovic & D'Elia, 2018).

Textura de las rocas

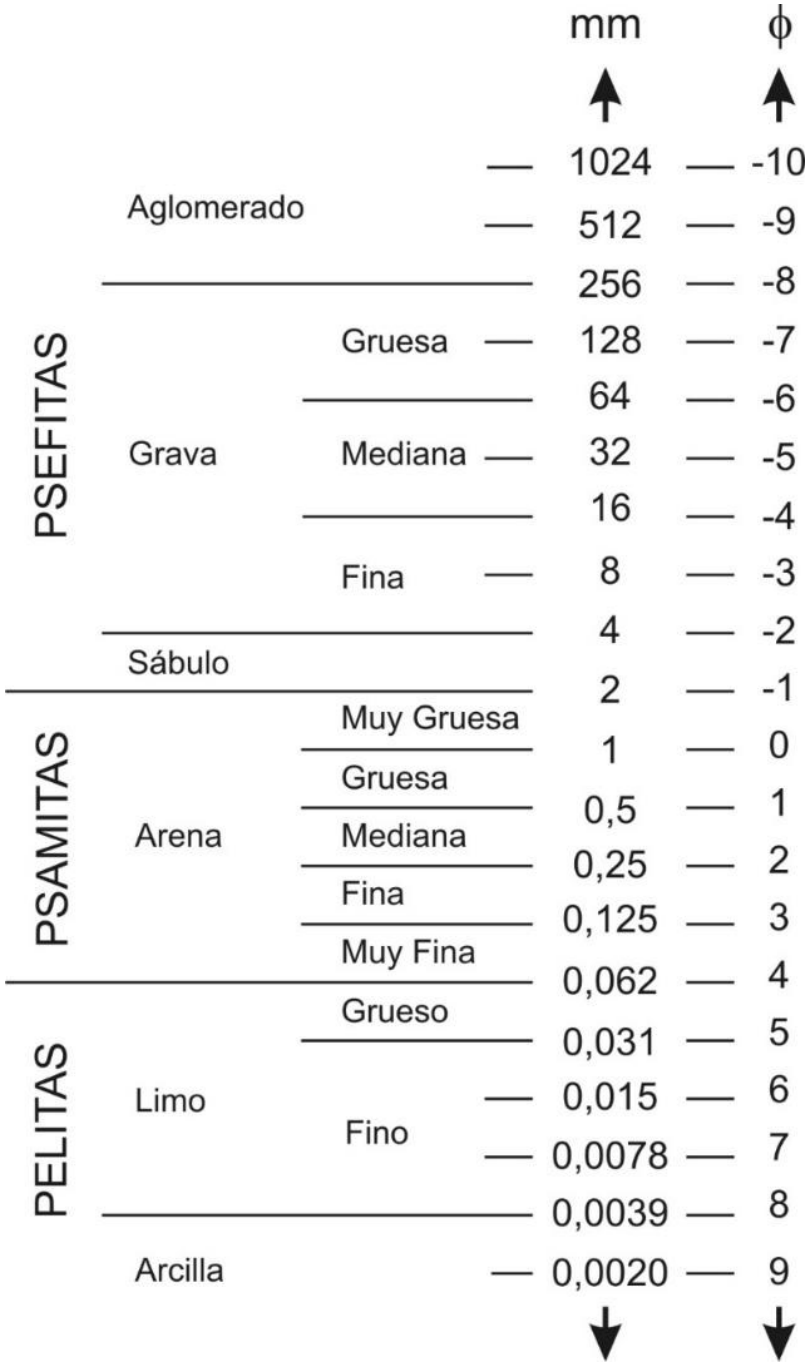
La **textura** de una roca sedimentaria hace referencia a la granulometría o tamaño, la distribución granulométrica, forma y orientación de sus granos y selección. Es directamente responsable de las propiedades físicas que posee la roca y, además, son indicadoras de los procesos sedimentarios y de las condiciones ambientales que predominaban durante la depositación, como por ejemplo la energía y, fluidez del agente de transporte, así como el tiempo involucrado tanto en el transporte como en la depositación. La comprensión de estas características es fundamental a la hora de realizar buenas interpretaciones de los procesos sedimentarios involucrados (Folk, 1980; Blair & McPherson,1999; Blott & Pye, 2008).

El **tamaño de grano** es una característica trascendental a la hora de la descripción y posterior clasificación de las rocas. Comenzó a cobrar importancia cuando autores como Udden (1898), Hopkins (1899), Atterberg (1905), Wentworth (1922), Robinson (1924) y Rubey (1930) propusieron diferentes escalas granulométricas para comenzar a clasificar a los granos que componen la roca. La más aceptada y utilizada hoy en día, es la conocida como "Escala de Udden-Wentworth" modificada por Krumbein en 1934 donde se reconocen 4 tamaños básicos (grava, arena, limo y arcilla) expresados en milímetros con subdivisiones internas y sus equivalentes en número φ (Tabla 1.1). Donde φ se calcula como:

$$\varphi = -log_2 D, \quad (1)$$

siendo D el diámetro del grano expresado en milímetros. En rocas consolidadas, existen diversas formas de conocer el tamaño de grano, a partir de la disgregación del material y posterior tamizado, a través de mediciones con cartillas en imágenes de lupa o a partir mediciones indirectas (2D) en imágenes de microscopía (Scasso & Limarino 1997; Leeder, 2012).

Tabla 1.1: Escala granulométrica de φ



Para realizar un análisis de la textura de grano de las rocas y obtener información representativa, es necesario realizar un análisis estadístico sobre la distribución granulométrica de la roca. La **distribución granulométrica**, se define como la distribución de frecuencia de aparición de los diferentes tamaños de granos que posee la roca. Una vez obtenida dichas curvas de distribución, se pueden calcular diferentes parámetros estadísticos (moda, media, mediana, selección, asimetría y curtosis) que sirven como sustento para interpretar aspectos del ambiente depositacional, los medios de transporte y depositación del sedimento como características del agente de transporte (Visher, 1969; Gao & Collins, 1991, Poizot et al., 2006; 2008).

Dichos parámetros estadísticos pueden calcularse según 2 métodos:

- Método Gráfico: La distribución granulométrica puede graficarse en forma de curva acumulativa de escala aritmética (Fig. 1.1 A), curva acumulativa de escala probabilística -también llamado diagrama de truncamiento- (Fig. 1.1 B), histograma (Fig. 1.1 C) o curva de frecuencia (Fig. 1.1 D) (Scasso & Limarino, 1997).

- Método de momentos: es una forma mucho más exacta de calcular los parámetros estadísticos de la población. Matemáticamente se define como:

$$M(e) = d(e)f, \tag{2}$$

donde M(e) es el momento de la estadística, f resulta de la frecuencia de la clase granulométrica analizada y d(e) es el espaciamiento en unidades de φ entre el punto de rotación y la clase granulométrica correspondiente (Fig. 1.2).

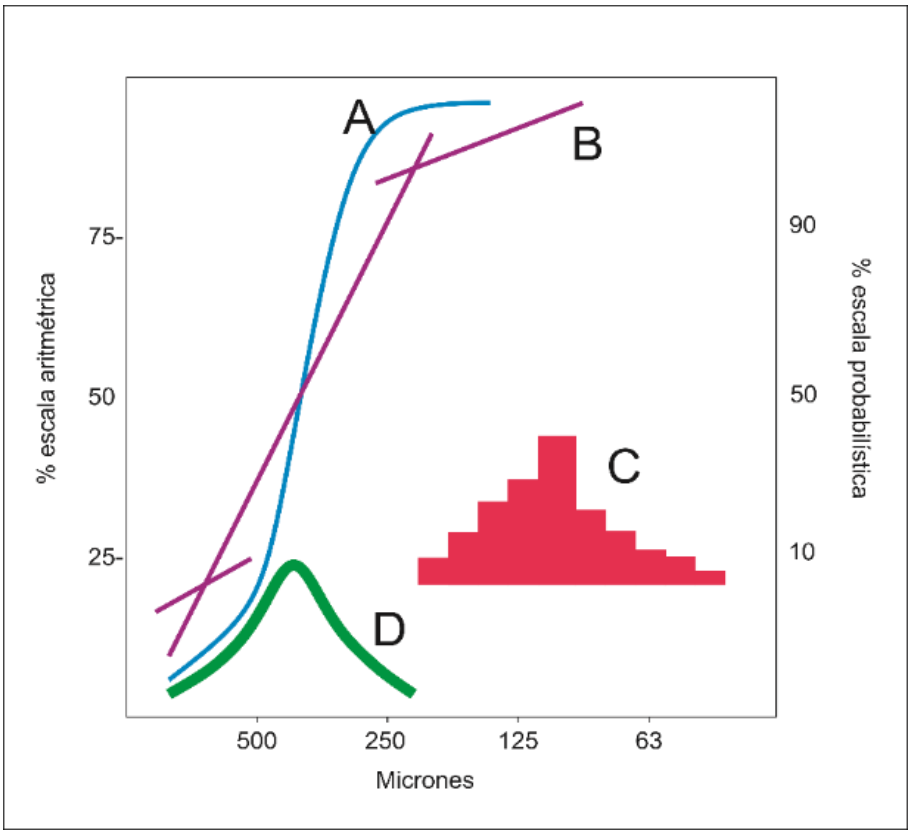


Figura 1.1: Distintos tipos de distribuciones granulométricas de sedimentos modificada de Scasso & Limarino (1997). (A) curva acumulativa de escala aritmética; (B) curva acumulativa de escala probabilística; (C) histograma; (D) curva de frecuencia.

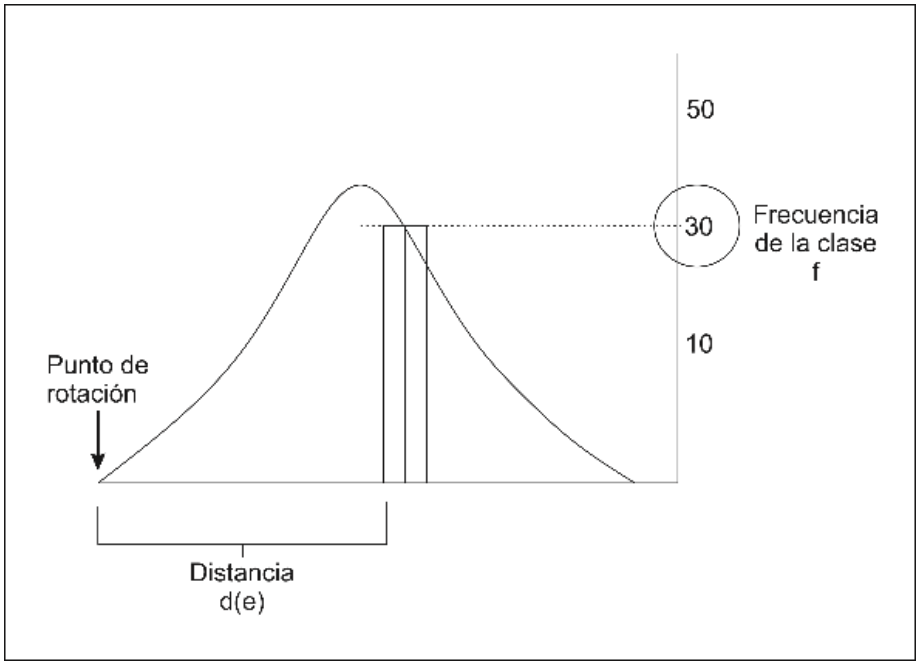


Figura 1.2: ejemplo del cálculo del momento estadístico modificada de Scasso & Limarino (1997).

Diferentes autores han abordado la caracterización de los ambientes sedimentarios a partir de los análisis granulométricos. En este sentido, Moss (1962), Passega (1964), Visher (1969), Glaister & Nelson (1974), Friedman (1979), McManus (1991), entre otros, son partidarios de que a partir de los métodos gráficos antes descriptos se puede llegar a una buena inferencia sobre el agente de transporte. Por ejemplo, Passega (1964) presenta un diagrama textural (diagrama CM) que emplea el diámetro de grano correspondiente al primer percentil versus la mediana del tamaño de grano (Fig. 1.3) para representar el tamaño de grano en sedimentos detríticos. Los puntos graficados dentro de un patrón CM completo indican las condiciones probables de transporte y depositación de los sedimentos. Estos patrones son útiles para identificar diferentes tipos de transporte y depositación, como suspensión uniforme, suspensión graduada, suspensión pelágica, carga de lecho, flujo de turbidez y corrientes de turbidez.

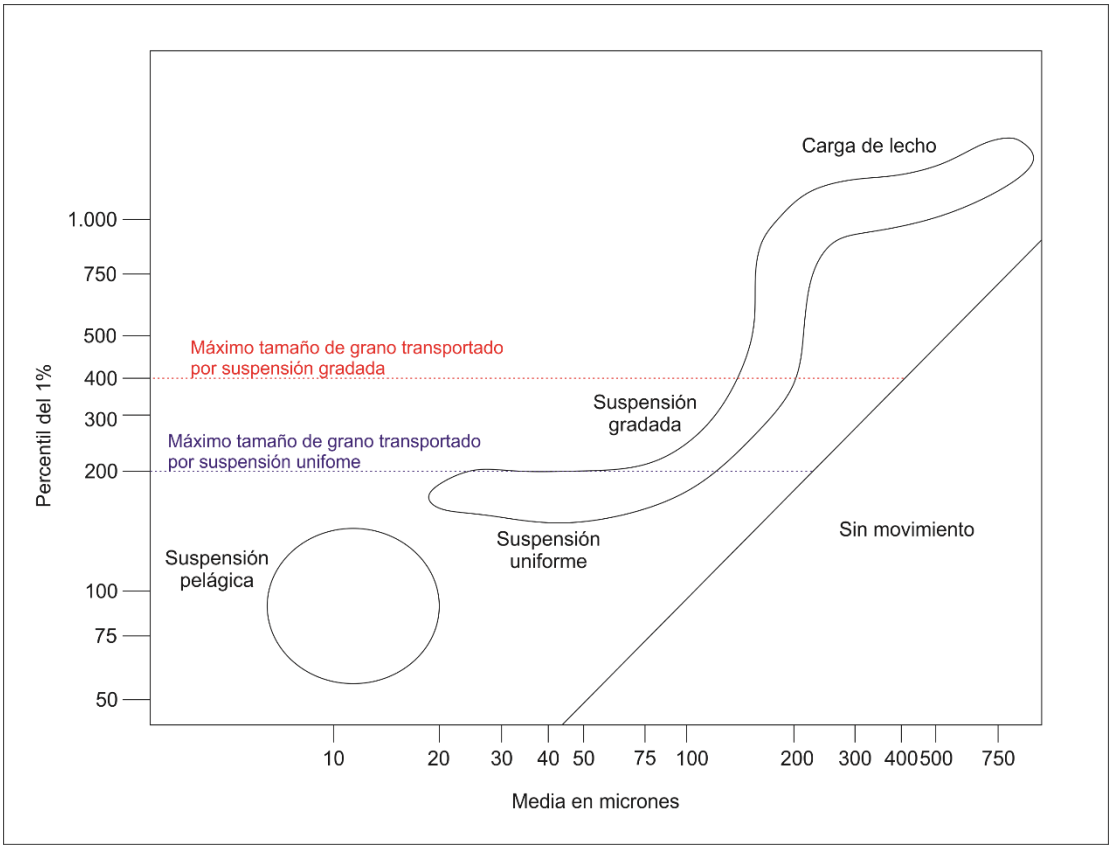


Figura 1.3: Diagrama CM desarrollado por Passega (1964)

Por su parte, Visher (1969), analiza las distribuciones de tamaño de grano en sedimentos clásticos de muestras de ambientes recientes y del registro sedimentario. Demuestra que las distribuciones de tamaño de grano pueden revelar información sobre los mecanismos de transporte y depositación, como la suspensión, el rolido o tracción y la saltación. Este trabajo es fundamental para la sedimentología porque proporciona una metodología para interpretar los diagramas de truncamiento (Fig. 1.4) y relacionarlo con los diferentes tipos de transporte de los sedimentos.

Sin embargo, Syvitski & Murray (1977) y Mc Laren (1981) consideran que no es posible realizar una interpretación satisfactoria de los distintos ambientes depositacionales. Ellos sugieren que la distribución granulométrica es independiente del tipo de transporte y que está fuertemente condicionada por el material detrítico (heredado de la fuente) y de los procesos sedimentarios.

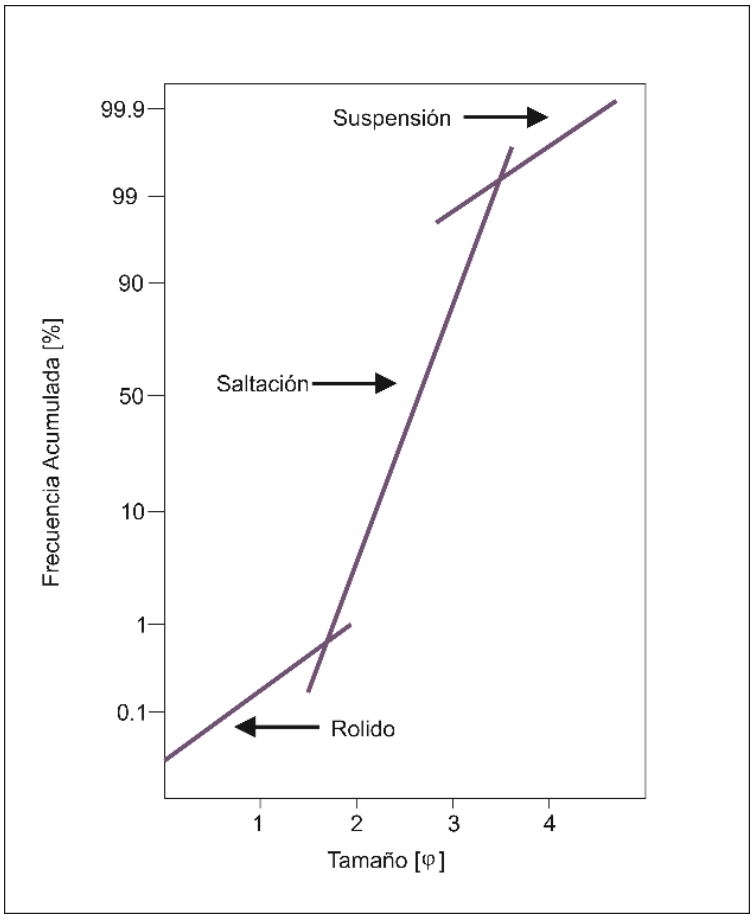


Figura 1.4: Diagrama de truncamiento propuesto por Visher (1969) donde se observan la relación entre la distribución del tamaño de los granos y los procesos deposicionales.

La **forma** que adquieren los clastos o granos durante su proceso de transporte es un aspecto fundamental en la sedimentología, ya que proporciona información sobre la historia del clasto desde su fuente hasta su lugar de depositación. Los estudios de autores como Barrett (1980) y Winklemolen (1982) han sido cruciales para entender cómo el transporte afecta la morfología de los clastos, diferenciando sus características tanto en dos como en tres dimensiones. Este conocimiento es esencial no solo para la clasificación de los clastos, sino también para inferir las condiciones ambientales del pasado, como la energía y la dinámica de los agentes de transporte.

Desde un punto de vista bidimensional, la forma de los clastos puede evaluarse a través de su redondez, la cual se define como la medida de la cantidad de aristas y vértices presentes en el grano. Según Powers (1953), la redondez de un clasto refleja el grado de desgaste que ha sufrido, que a su vez es indicativo de la distancia y el tipo de transporte que ha experimentado. En este sentido, el autor desarrolló una escala comparativa visual para estimar el redondeamiento de los clastos (Fig. 1.5). Los clastos que han sufrido un transporte más largo y continuo tienden a presentar mayor redondez, debido a la abrasión constante y el impacto repetido con otros clastos y superficies en el medio de transporte, mientras que los clastos que han sufrido menor transporte tienden a presentar formas más angulosas.

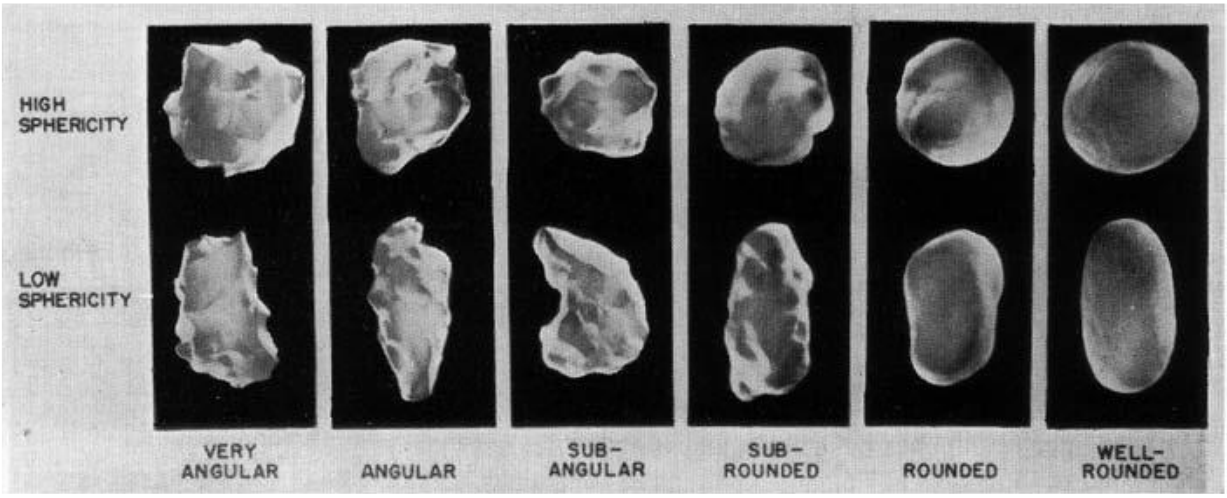


Figura 1.5: Escala comparativa visual original para estimar el redondeamiento de los clastos propuesta por Powers (1953).

La **redondez** de los clastos depende de varios factores, como la dureza del material, la energía del medio de transporte y la duración del mismo (Winklemolen, 1982). En medios de transporte con alta energía los clastos son transportados a grandes distancias y están en constante contacto unos con otros, lo que favorece su redondez. Este proceso de abrasión y desgaste es más notable en clastos de minerales más blandos, mientras que aquellos formados por minerales duros tienden a conservar formas angulares incluso después de largos periodos de transporte (Scasso & Limarino, 1997).

La **esfericidad**, por otra parte, es una medida tridimensional que representa el grado en que la forma de un clasto se asemeja a una esfera (Wadell, 1932). También puede definirse como el cociente entre el volumen de un clasto y el de la menor esfera que lo circunscribe (Krumbein, 1941). Esta característica proporciona información adicional sobre el tipo de transporte y el ambiente depositacional. Clastos con alta esfericidad suelen ser el resultado de transporte prolongado y flujo constante. En contraste, clastos que han experimentado transporte por agentes de baja energía suelen tener baja esfericidad debido a que el transporte es intermitente y con menor abrasión (Scasso & Limarino, 1997).

El tipo de transporte al que se somete un clasto influye de manera determinante en su forma. En sistemas de transporte de alta fluidez, como ríos, los clastos están en constante movimiento, lo que aumenta la frecuencia de colisiones y favorece tanto la redondez como la esfericidad. Este tipo de transporte tiende a producir clastos más redondeados y esféricos, especialmente en aquellos minerales menos resistentes a la abrasión. En contraste, los ambientes de baja fluidez, como los flujos de lodo o ciertos ambientes de deslizamientos gravitacionales, imponen menos movimiento en los clastos, por lo que presentan formas más angulosas y baja esfericidad debido a un menor grado de desgaste (Allen, 1985).

La **viscosidad y fluidez del agente** de transporte también juegan un rol esencial en la morfología de los clastos. La fluidez se refiere a la capacidad del agente de transporte para mover los clastos con facilidad, mientras que la viscosidad determina la resistencia interna al flujo. En ambientes de baja viscosidad (alta fluidez), el transporte de los clastos es más efectivo, ya que los

clastos son libres de moverse por el flujo. Este movimiento continuo promueve el desgaste de las aristas y produce clastos más redondeados. Por otro lado, en medios de alta viscosidad (baja fluidez), el transporte de los clastos es menos efectivo. La alta resistencia al flujo limita la capacidad de los clastos para moverse y rotar, reduciendo el desgaste y la posibilidad de formar clastos redondeados o esféricos (Allen, 1985). En estos casos, los clastos suelen conservar formas angulosas, y presentan baja esfericidad debido al contacto limitado con otros clastos y la baja frecuencia de colisiones.

La **selección** es una característica que refiere a los diferentes tamaños de granos (en φ o en micrones) que componen a la roca. La selección implica la uniformidad en el tamaño de grano, y se ve reflejado claramente en gráficos como histograma y curvas de distribución granulométrica como la forma de la campana de la distribución (Fig. 1.6) (Scasso & Limarino, 1997). Una facies sedimentaria muy bien seleccionada posee tamaños de granos similares entre sí, lo que se evidencia en una curva de tipo leptocúrtica (Fig. 1.6 B) donde la dispersión de tamaños de grano es baja respecto a la dispersión normal (Fig. 1.6 A). Esto es indicativo de que el flujo que dio origen a ese depósito fue un flujo con alta fluidez (baja viscosidad), donde los granos tienen la capacidad de ser transportados diferentes distancias según su tamaño y peso, o que los sedimentos fueron expuestos a retrabajo por parte de un flujo también fluido que fue separando los diferentes tamaños de sedimento con el paso del tiempo (Allen, 1985).

En contraposición con lo anterior, una facies sedimentaria mal seleccionada posee una amplia dispersión en los tamaños de granos. Esta alta variabilidad hace que la curva de distribución siga una forma “amesetada” denominada distribución platicúrtica (Fig. 1.6 C) respecto a la distribución normal (Fig. 1.6 A). Esto es indicativo de que el flujo que dio origen a ese depósito fue un flujo viscoso (baja fluidez), donde los granos no tienen la capacidad moverse libremente. En este tipo de flujos, el agente de transporte no tiene la capacidad de seleccionar la distribución de los tamaños de granos según su tiempo de transporte, tamaño y/o densidad, sino que mueve todos los tamaños de granos al mismo tiempo (Allen, 1985).

En un punto intermedio entre los depósitos bien seleccionados y mal seleccionados, encontramos los depósitos conglomerádicos. En ellos, las

distribuciones granulométricas son características por tener una moda bien identificada, con una cola de otros tamaños granulométricos (Fig. 1.6 D y E). La curva de distribución asimétrica positiva (Fig. 1.6 D), corresponde a depósitos denominados como ortoconglomerados. Estos poseen una moda con una cola de tañamos más gruesos que los correspondientes al tamaño de dicha moda. Por otra parte, encontramos también los denominados paraconglomerados. Estos depósitos se caracterizan por poseer una distribución granulométrica asimétrica negativa, donde la cola es de tamaños inferiores a los del tamaño de la moda (Fig. 1.6 E). Finalmente, encontramos un último tipo de depósitos que se caracterizan por poseer una distribución granulométrica bimodal (Fig. 1.6 F). Estos depósitos se forman por cambios de energía en el agente de transporte generando intercalaciones de distintos tamaños de granos.

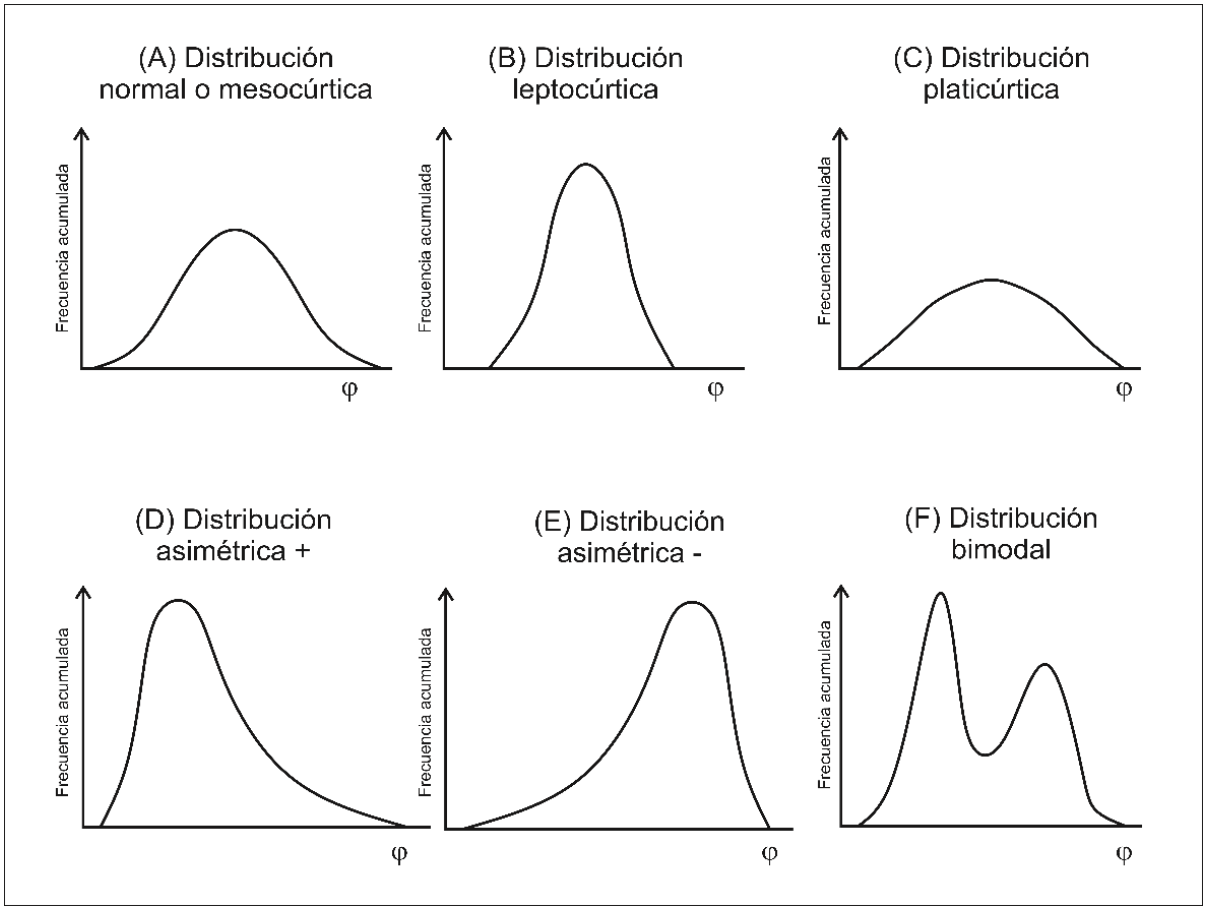


Figura 1.6: Diferentes tipos de distribuciones granulométricas dependiendo de la simetría y la curtosis. Tomada y modificada de *Merodio (1985).*

Estructuras sedimentarias de las rocas

Las **estructuras sedimentarias** (Tabla 1.2) son las disposiciones geométricas tridimensionales que toman los sedimentos durante o poco después de su depositación y brindan información valiosa sobre los procesos mecánicos, químicos y biológicos que actuaron durante ese lapso. Es por eso, que su análisis permite realizar inferencias sobre las características del agente de transporte y sus mecanismos de transporte, las condiciones del régimen de flujo, la dirección y sentido de migración de los sedimentos, las condiciones del sustrato y las condiciones paleoambientales (Allen, 1982; Cheel, 2005; Collinson et al., 2006; Pettijohn, 1957; Ponce et al., 2018). La clasificación de las diferentes estructuras sedimentarias depende de las distintas interpretaciones de los autores. Allen (1982) y Collinson et al. (2006), las clasifican y describen según sean depositacionales, post depositacionales u erosivas, mientras que Ponce (2018) por su parte en “Atlas de estructuras sedimentarias”, lo hace según el origen que tienen las mismas en inorgánicas, formadas por procesos físicos y químicos; y biogénicas, formadas por la actividad de organismos.

Las **estructuras sedimentarias depositacionales** son el resultado de procesos de transporte y acumulación de sedimentos bajo condiciones específicas. Estas estructuras proporcionan información esencial sobre la dinámica y energía del flujo (Allen, 1982; Collinson et al., 2006; Ponce et al., 2018). Entre las estructuras más comunes se encuentran la estructura laminada, la estratificación horizontal, las estratificaciones entrecruzadas, las óndulas (simétricas y asimétricas) y las estructuras de gradación (normal e inversa), cada una resultante de un proceso físico sedimentario específico. Por ejemplo, las estructuras laminares son características de flujos laminares de baja energía, sedimentación tranquila, generalmente asociada con sedimentos de granulometrías finas transportados por suspensión (Brookfield & Silvestro, 1992; Brookfield & Ahlbrandt, 2000). En contraparte, la estratificación horizontal representa capas tractivas en condiciones de alto régimen de flujo (Cheel, 1990; Collinson et al., 2006; Arche, 2010). Las estratificaciones cruzadas (tangencial y/o en artesa) se forman debido a la migración de óndulas en condiciones de flujo activo. Este tipo de estructura es utilizada para interpretar la dirección y el

Tabla 1.2: Estructuras sedimentarias recopiladas de Collinson et al. (2006) y Ponce et al., (2018)

Depositacional	Mecánicas	Estratificación mixta	Flaser
			Ondulítica (Wavy)
			Lentiforme
		Cambios en la energía del flujo	Maciza
			Maciza por bioturbación
			Gradación normal
			Gradación inversa
			Mud drapes
		Flujo de bajo régimen	Estratificación entrecruzada tangencial
			Estratificación entrecruzada en artesa
			Laminación planar
			Óndulas de corrientes
			Óndulas escalonadas
			Óndulas de oscilación
		Flujo de alto régimen	Antidunas
			Estratificación horizontal
			Estratificación entrecruzada de bajo ángulo
		Tormentas	Hummocky y Swaley
	Químicas	Precipitación	Nódulos
			Concreciones
			Crecimiento vertical de cristales
			Laminación
Postdepositacionales	Biológicas	Estructuras de bioturbación	
		Estructuras de biodepositación	
		Estructuras de bioestratificación	
	Disolución	Superficies estilolíticas	
		Estilolitas	
	Deformación	Laminación convoluta	
		Calcos de carga	
		Escape de agua	
		Slumps	
Erosivas	Estructuras erosivas de mayor escala	Estrías de deslizamiento	
		Superficies de deflación	
		Canales	Simples
			Complejos
			Migracion lateral
	Estructuras que quedan registradas en la capa suprayacente	Formas por erosión del agua	Surcos por obstáculos
			Surcos longitudinales
			Marcas de rocas
			Mega flutes
		Formas por erosión del viento	Surcos de deflación
			Ventifactos
	Marcas en el suelo	Marcas de arrastre	Surcos por obstáculos
			Flutes
			Surcos longitudinales
			Gutter cast
	Marcas de objetos	Continuas	Surcos
			Chevrons
		Discontinuas	Prod marks
			Marcas de rebotes
			Marcas de saltación

sentido de las paleocorrientes y la velocidad del flujo (Leclair, 2002; Collinson et al., 2006; Bossi, 2006; Arche, 2010). De manera similar, las óndulas se clasifican como simétricas o asimétricas. Las óndulas simétricas se generan en ambientes dominados por olas, mientras que las asimétricas son características de flujos unidireccionales (Anderson, 1987; Ashley, 1990; Southard & Boguchwal, 1990; Andreotti et al., 2006). La gradación normal e inversa son estructuras que reflejan cambios en la energía del medio durante la depositación. La gradación normal, donde los granos más gruesos se depositan primero y los más finos después, es producto de la desaceleración de los flujos. Por el contrario, la gradación inversa ocurre en flujos densos o gravitacionales, donde las partículas más gruesas son transportadas hacia la parte superior por efecto de Bagnold (Bagnold,1941; Anderson & Hallet, 1986). En ambientes marinos someros afectados por tormentas, se desarrollan estructuras *hummocky* las cuales reflejan la interacción entre el *Fetch* de olas y las corrientes oscilatorias, siendo indicativas de eventos de alta energía (Myrow & Southard, 1996; Midtgaard, 1996). Finalmente, las estructuras de carga se forman cuando sedimentos saturados de agua se deforman debido a inestabilidades gravitacionales, creando estructuras características en depósitos de alta carga sedimentaria (Allen, 1982; Collinson et al., 2006; Ponce et al., 2018).

Las **estructuras sedimentarias erosivas** se forman cuando el sedimento ya depositado es modificado mientras aún no se ha consolidado (Allen, 1982; Collinson et al., 2006; Ponce et al., 2018). Para que las estructuras erosivas se conserven, el sedimento erosionado tiene que ser lo suficientemente cohesivo y resistente como para mantener el relieve que se genera por erosión hasta que quede sepultado por un sedimento nuevo que lo preserve en el registro sedimentario. Las estructuras erosivas a pequeña escala se reconocen casi siempre como relieves en la base del lecho inmediatamente suprayacente a la superficie erosionada. La erosión también se reconoce en secciones verticales por el truncamiento de la estratificación o laminación en el sedimento debajo de la superficie erosionada. Dentro de esta categoría encontramos los canales erosivos, *flutes* o marcas de corrientes, estructuras de corte y relleno. Son indicadores de la dirección de las paleocorrientes, de la presencia o arrastre de obstáculos en los sustratos, y de zonas de derrumbes (Allen, 1982; Collinson et al., 2006; Bossi, 2006; Ponce et al., 2018).

Las **estructuras sedimentarias post depositacionales** surgen cuando los sedimentos, tras su depositación inicial, son sometidos a procesos físicos, químicos o biológicos que alteran su estructura original. Estas modificaciones pueden incluir deformaciones blandas, como pliegues o estructuras de carga, que se generan debido a la inestabilidad mecánica de los sedimentos saturados de agua. Entre las más comunes se encuentran las estructuras por escape de fluidos, donde el agua atrapada en los poros es expulsada rápidamente (Allen, 1982; Collinson et al., 2006; Ponce et al., 2018). También es frecuente la bioturbación, que ocurre cuando organismos vivos remueven y mezclan los sedimentos, destruyendo o modificando las estructuras originales (Buatois & Mangano, 2011). A nivel químico, los procesos de diagénesis temprana pueden formar concreciones, nódulos o recristalizaciones, preservando o distorsionando detalles del sedimento original (Fazio et al., 2007). Estas estructuras no solo registran cambios ambientales y tectónicos posteriores a la sedimentación, sino que también proporcionan información sobre los procesos que afectaron al sedimento antes de su consolidación (Allen, 1982; Collinson et al., 2006; Ponce et al., 2018).

Interpretación del proceso sedimentario de las facies sedimentarias

La textura de una roca sedimentaria, que incluye aspectos como el tamaño de grano, la redondez y el grado de selección, permite deducir la energía del agente de transporte y la distancia de transporte desde la fuente hasta el sitio de depositación (Scasso & Limarino, 1997).

En sistemas de flujos fluidos (agua y aire), el tamaño de los granos es un indicador directo de la energía necesaria para el transporte de los mismos. A mayor tamaño de granos, mayor es la energía necesaria que debe tener el agente de transporte para movilizarlo (Krumbein, 1941; Powers, 1953; Barrett, 1980; Winklemolen, 1982; Allen, 1985; Scasso & Limarino, 1997). Cuando dicho agente de transporte pierde energía, los sedimentos dejan de ser movilizados y se depositan. Durante este proceso de transporte, las partículas sufren choques repetidos entre sí, lo que produce un desgaste progresivo que redondea las aristas y vértices. Si la capacidad de carga del agente desciende cerca de la fuente, los sedimentos se depositan sin sufrir suficiente erosión, lo que se refleja

en una textura con granos individuales poco redondeados (Allen, 1985; Scasso & Limarino, 1997). En entornos donde el flujo tiene suficiente energía, los granos gruesos pueden transportarse largas distancias mediante tracción y saltación.En flujos turbulentos de alta energía, los impactos frecuentes entre partículas aceleran el redondeamiento del sedimento, mientras que, en flujos de baja o moderada energía, que tienden a ser laminares, los granos interactúan menos, manteniendo formas más angulosas y conservando aristas. Estos patrones texturales permiten inferir las condiciones de flujo de los agentes de transporte (Passega, 1964; Visher 1969; Allen,1985; Scasso & Limarino, 1997).

Las partículas más finas, como arcillas y limos, son más ligeras y, por lo tanto, más fáciles de transportar. Este tipo de sedimento se suele transportar por largas distancias y llegar a depositarse en áreas más alejadas de la fuente, como la zona más profunda de los depocentros. Su transporte se realiza comúnmente en suspensión, en el cual existe poca interacción entre los granos.

La textura de la roca, en conjunto con las estructuras sedimentarias, ofrece una comprensión detallada del tipo de ambiente y los procesos sedimentarios implicados en la formación de depósitos específicos. Este análisis permite no solo reconstruir condiciones ambientales pasadas, sino también entender procesos modernos en cuencas actuales, lo cual es aplicable a la exploración de recursos naturales como el petróleo, el gas y el agua subterránea.

1.3.2 Concepto de Aprendizaje automático

El Aprendizaje automático (ML por sus siglas en inglés, *Machine Learning*) es parte de una de las subdisciplinas de la Inteligencia Artificial. Se enfoca en el desarrollo de algoritmos que les permiten a las computadoras tomar decisiones o realizar tareas basadas en el análisis y entendimiento de patrones de comportamiento (Géron, 2022; Posthoff, 2024).

Los distintos modelos de ML se pueden clasificar de varias maneras, la más aceptada y utilizada es la que los subdivide diferenciándolos por un lado en la configuración de los datos y por otro, en el objetivo planteado para su utilización en categorías como aprendizaje supervisado y no supervisado. Existe

una tercer categorización denominada aprendizaje por refuerzo, la cual no será abordada en este trabajo.

Los modelos no supervisados son aquellos en los que el modelo se entrena con datos no etiquetados, es decir, que entre los datos no encontramos la variable respuesta. Aquí, el modelo en lugar de aprender que patrones relacionan los datos de entrada con los datos de salida, simplemente se ocupa de reconocer los patrones y ordenar o agrupar los datos según esos patrones. Dentro de este grupo se incluyen modelos de agrupamiento como *Clustering* o *k-means* y modelos de ordenamiento como por ejemplo Análisis de Componentes Principales (Dangeti, 2017; Géron, 2022).

En los modelos supervisados, se entrena al algoritmo indicándole cual es la variable respuesta u objetivo. Esto permite que reconozca y aprenda los patrones que relacionan los datos de entrada con los datos de salida, de manera que cuando se le den nuevos datos sin etiquetar este pueda reconocer los patrones y devolver la respuesta correcta. Dentro de este grupo se incluyen a los modelos de clasificaciones y regresiones, modelos de Maquinas de Vector Soporte *o Support Vectors Machine* (SVM); un amplio grupo de familia de modelos basados en árboles de decisión y las redes neuronales (Bishop, 2006; Géron, 2022; James et al. 2013).

Al considerar la utilización de ML, debemos tener en consideración los datos que se le entregan al modelo, ya que estos afectan directamente su capacidad para generalizar patrones (Rollinson, 1993; Martín-Fernández et al., 2011; Witten et al., 2011; García et. al, 2016). En el siguiente capítulo se explicarán las etapas que atraviesan los datos para que los distintos modelos de ML puedan detectar de la mejor manera sus patrones de comportamiento.

Modelos de Aprendizaje automático

En el marco de este trabajo de tesis, para generar un modelo que sea capaz de cumplir la tarea planteada de predecir las facies sedimentarias, se decidió utilizar modelos de tipo supervisados, a fin de aprovechar el hecho de que se realizó la descripción de las rocas que componen el conjunto de datos, generando así la variable objetivo. De igual manera, se realizaron pruebas con

algoritmos no supervisados, sin tener en consideración las variables objetivo generadas para probar su eficacia y para observar el comportamiento de los patrones de los datos.

A continuación, detallaremos los modelos de ML supervisados y no supervisados utilizados en esta tesis para la predicción de facies sedimentarias:

Algoritmos de Support Vectors Machine (SVM)

Los algoritmos de SVM fueron desarrollados por Vapnik en la década de los ’90 (Cortes & Vapnik, 1995). Son utilizados para tareas de clasificación y regresión, basándose en la búsqueda de un hiperplano óptimo para separar de la mejor manera posible las distintas clases de la variable objetivo, maximizando el espacio entre ellas (Ai, 2019; Bishop, 2006). Matemáticamente, este hiperplano puede definirse como:

$$\beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + \cdots + \beta_n\chi_n = 0, \quad (3)$$

donde β_0, β_1, …, β_n son los parámetros para los cuales todos los valores de χ cumplen la igualdad. En un espacio de características bidimensional, un hiperplano es una línea que divide al espacio en dos mitades mientras que, en dimensiones superiores, un hiperplano es un subespacio con una dimensión menos que el espacio en el que se encuentre (Fig. 1.7). Los puntos de muestras más cercanas a los hiperplanos son los llamados vectores soporte, y sirven para calcular el margen, o distancia entre ellos y el hiperplano (Harrington, 2012; Ertel, 2017; Géron, 2022).

Cuando los datos no son linealmente separables en su espacio original, puede existir un espacio equivalente en el que si lo sea. En este sentido, se introdujo el concepto de *kernel*. Un *kernel* es una función que reemplaza el producto escalar que se encuentra en la definición de hiperplano y que busca transformar el espacio dimensional donde habitan los vectores β y χ por otro espacio (espacio dimensional del kernel), en el que sí es posible separarlos (Harrington, 2012; Müller et al., 2016).

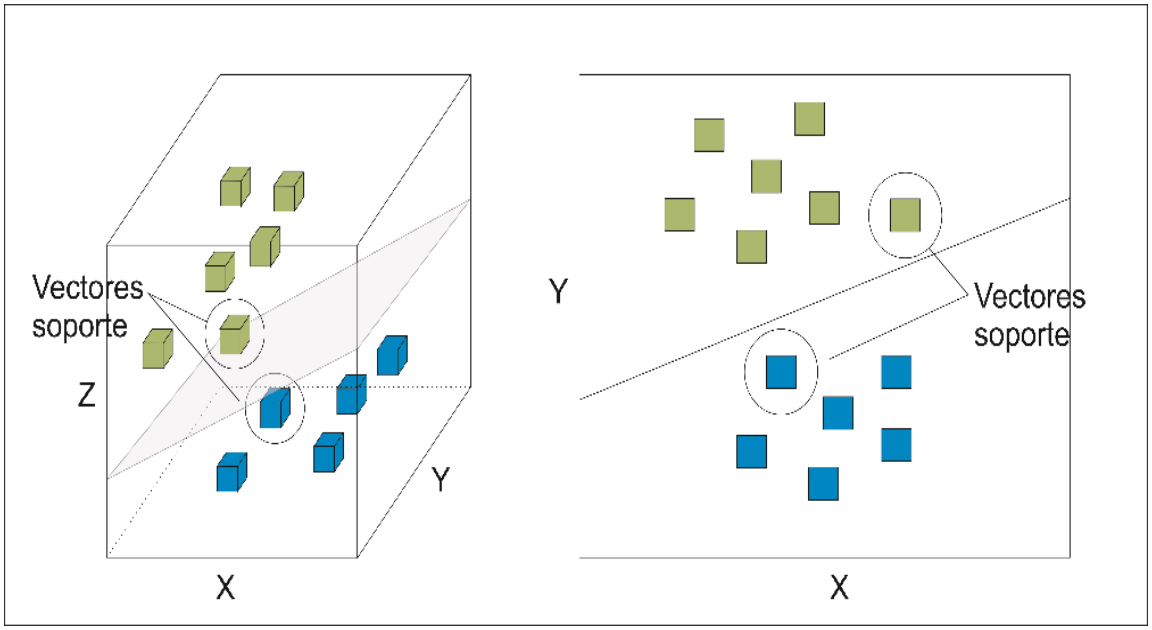


Figura 1.7: Esquema de las SVM en 2D y 3D donde se identifica los vectores soporte de cada grupo.

Algoritmos de Regresión Logística

Introducido por Cox (1958), la regresión logística modela la relación entre una variable dependiente y un conjunto de variables independientes, tanto continuas como categóricas. Esto se logra mediante la función sigmoide, definida como:

$$f(x) = \frac{1}{1+ e^{-x}}, \tag{4}$$

donde x es la variable independiente.

Esta función $f(x)$ tiene una imagen (0,1) para un dominio real. Si x es positivo le corresponde un valor de f(x) mayor a 0.5, y, por el contrario, si x es negativo, le corresponde valores de f(x) menores a 0.5. Si consideramos esto en términos de clasificaciones binarias el primer caso correspondería a una clase, y el segundo caso a otra clase. Con el tiempo, el modelo de regresión logística se generalizó para admitir varias clases directamente, sin tener que entrenar y combinar varios clasificadores binarios. Esto se denomina Regresión Softmax o Regresión Logística Multinomial (Cox, 1958).

Árboles de decisión

Un árbol de decisión es un algoritmo basado en una estructura jerárquica con “nodos”, “ramas” y “hojas”, donde cada nodo representa una "prueba" o condición sobre un atributo, cada rama representa el resultado de la prueba, y cada hoja representa una clase o valor de decisión. El algoritmo funciona dividiendo los datos de entrada en función del valor de una característica específica en regiones más pequeñas y homogéneas según la variable objetivo (Fig. 1.8). El objetivo es encontrar una secuencia de divisiones que maximice la información obtenida sobre la variable de salida (Ertel, 2017).

El proceso de construcción de un árbol de decisión implica varias etapas. Primero, se calcula la mejor división posible. Para ello, el algoritmo selecciona la característica que mejor divide el conjunto de datos en dos o más grupos homogéneos. La medida más común utilizada para esta tarea es la impureza de “Gini” (mide la probabilidad de que un elemento seleccionado al azar sea clasificado incorrectamente si se le asigna una etiqueta aleatoria del conjunto de datos) o la ganancia de información (mide la reducción de la entropía después de dividir el conjunto de datos en base a un atributo) (Gerón, 2019). Luego, según la característica seleccionada, se divide al nodo actual en subnodos y el proceso se repite recursivamente para cada subnodo creado hasta que se alcance el criterio de detención (profundidad máxima del árbol o un número mínimo de muestras en un nodo). Finalmente, a los nodos finales, también llamados nodos hoja, se le asignan una clase basándose en la mayoría de los elementos que se encuentran en ese nodo (Breiman et al., 1984; Criminisi et al., 2011; Ertel, 2017). Los árboles de decisión se utilizan tanto en regresión como en clasificación (Gerón, 2019).

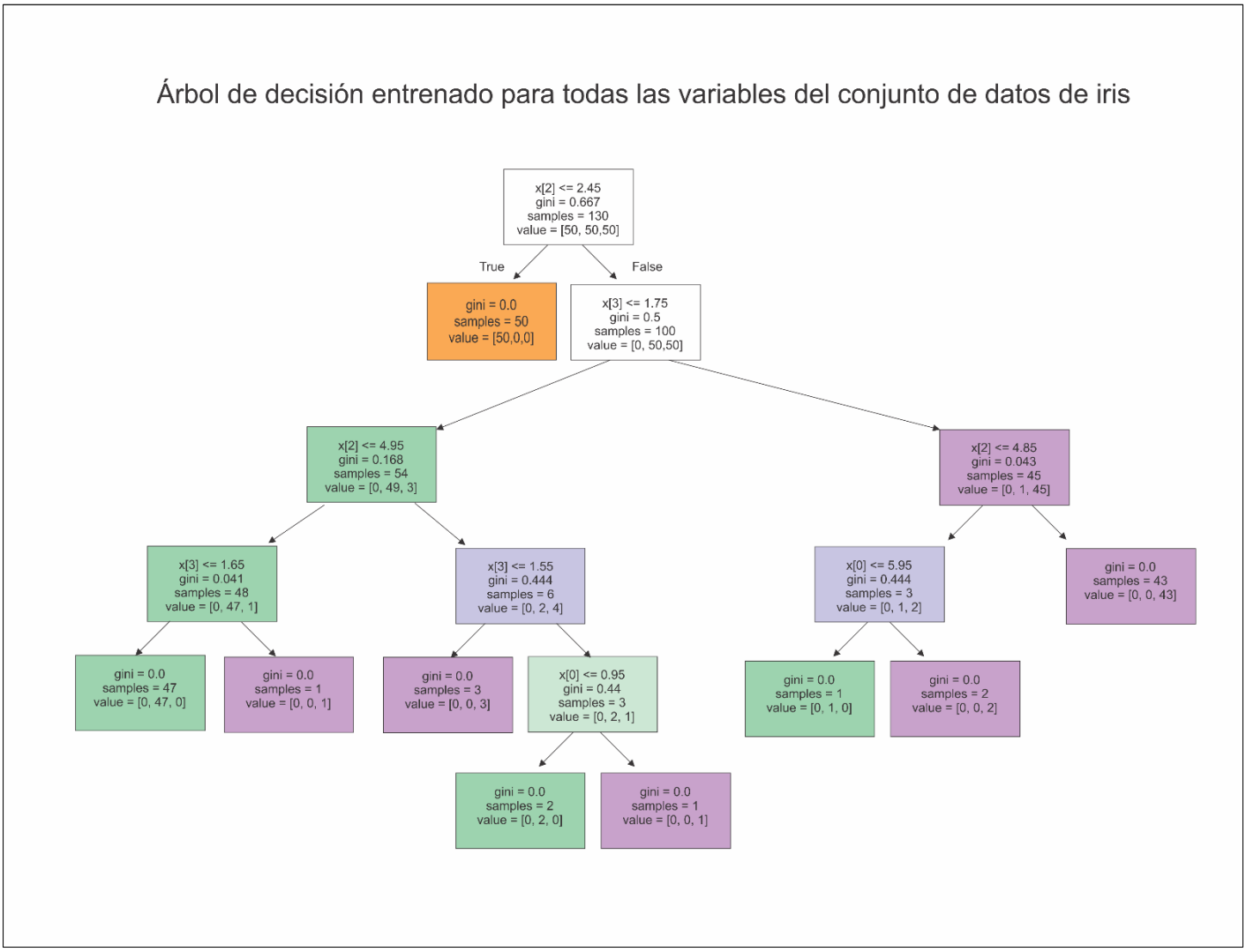


Figura 1.8 Esquema de cómo funcionan los árboles de decisión, tomado de *sckit learn.*

Un ensamble de árboles de decisión es una técnica de ML que combina varios árboles de decisión para mejorar el rendimiento de las predicciones respecto a la de sus componentes individuales *(*Ertel, 2017*)*. Al usar varios árboles de decisión y combinar sus predicciones, los ensambles tienden a ser más robustos y precisos en las predicciones. En las siguientes secciones se dará una pequeña explicación coloquial de las técnicas de ensambles de árboles de decisión que se han utilizado en esta tesis.

Algoritmos basados en arboles de decisión: Randon forest

Los algoritmos de *Random Forest* (RF), tienen sus orígenes en el trabajo de varios investigadores en el siglo XX. En particular, se le atribuye a Breiman et

al. (1984), quienes proporcionaron un marco unificado para construir modelos de decisión, tanto para tareas de clasificación como de regresión, utilizando un enfoque basado en la división recursiva del espacio de entrada (Breiman et al., 1984; Criminisi et al., 2011; Ertel, 2017).

Random Forest es un ensamble de árboles de decisión entrenados de forma independiente (Müller, 2016). Como se muestra en la Figura 1.9, cada árbol es entrenado con una muestra aleatoria del conjunto de datos (mediante el muestreo con reemplazo o *bootstrap*) y, en cada nodo del árbol se selecciona aleatoriamente un subconjunto de *features* para decidir la división del nodo. En el caso de la clasificación, el resultado de cada árbol es contabilizado (Clase X, Y, …, Z), y la clasificación final es aquella que tiene la clase mayoritaria. En el caso de la regresión, la predicción final surge del promedio ponderado de las predicciones individuales de cada árbol. Este proceso de aleatorización reduce la correlación entre los árboles y mejora la precisión del modelo (Breiman et al., 1984; Criminisi et al., 2011; Müller, 2016; Ertel, 2017).

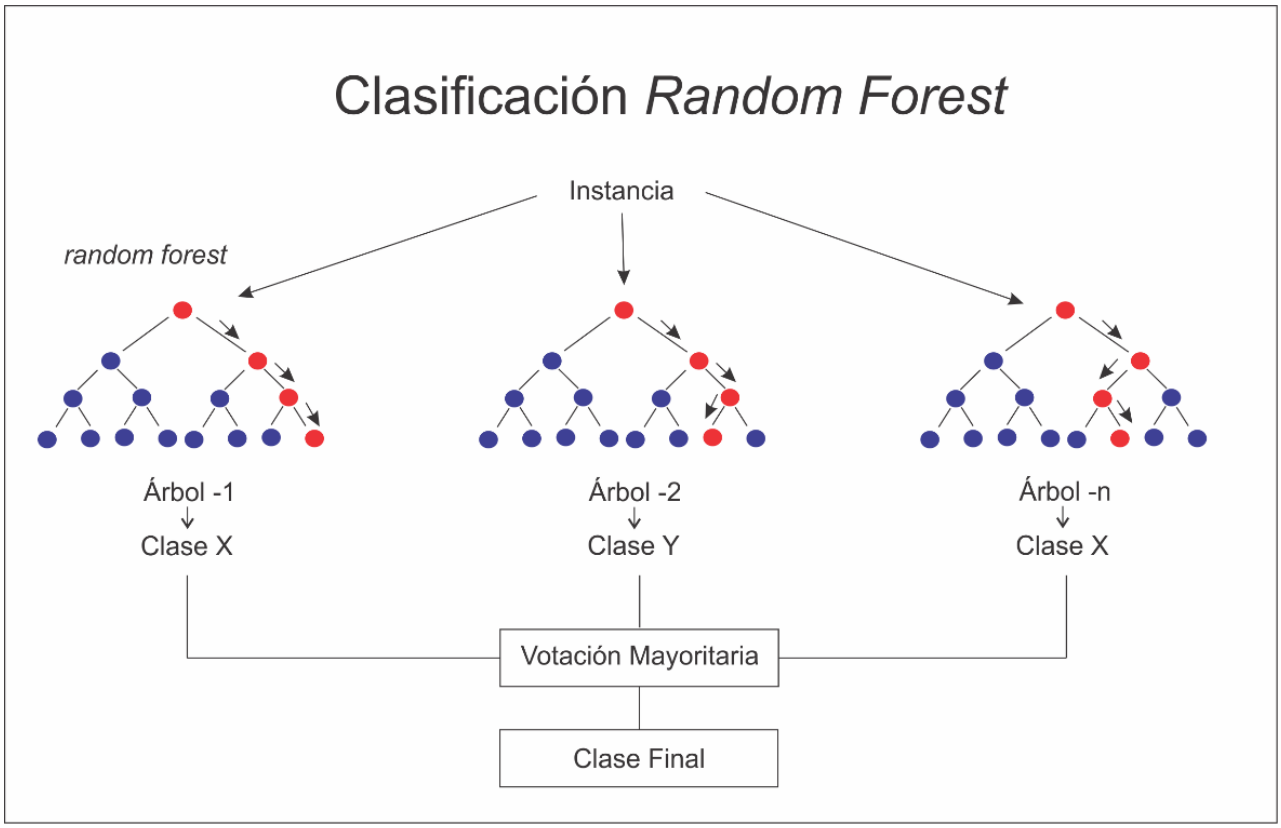


Figura 1.9: Esquema conceptual Random Forest. Tomada y modificada de Khan et al., 2021.

Algoritmos basados en arboles de decisión: Gradient Boosting Machine

Los algoritmos de *Gradient Boosting Machine* (GBM) son utilizados para problemas de clasificación y regresión. Este se construye a partir de la combinación de diferentes arboles de decisión, para mejorar progresivamente la precisión de las predicciones.

El concepto de *boosting* fue introducido por primera vez por Schapire (1990), quien demostró que un conjunto de clasificadores puede ser combinado para formar un clasificador que tenga mejores resultados (Fig. 1.10). Cada clasificador $Y_M(x)$ (rectángulo azul en Fig. 1.10) se entrena de una forma ponderada del conjunto de entrenamiento (flechas azules en Fig. 1.10) en la que las ponderaciones $W_n^{(M)}$ dependen del rendimiento del clasificador base anterior $y_{M-1}(x)$ (flechas verdes en Fig. 1.10). Una vez entrenados todos los clasificadores se combinan (flechas rojas en Fig. 1.10) para obtener el clasificador final $Y_M(x)$. Esta combinación se realiza en el caso del Boosting más utilizado (*Adaboost*) a partir de la función:

$$Y_M(x) = \pm\left(\sum_m^M \alpha_m y_M(x)\right), \tag{5}$$

donde $\alpha_m = \alpha_m(x_n, w_n)$, donde m es un dado árbol del ensamble (m = 1, M) y x_n es un dato con su peso w_n.

El algoritmo GBM, desarrollado por Friedman (2001), es una extensión de esta idea. En lugar de ajustar un solo modelo de árboles de decisión al conjunto de datos, GBM construye múltiples árboles en secuencia, donde cada uno intenta corregir los errores cometidos por los anteriores. Un aspecto clave de GBM es su enfoque iterativo para minimizar la función de pérdida. En cada iteración, el algoritmo calcula el gradiente de la función de pérdida con respecto a las predicciones actuales y ajusta un nuevo modelo para aproximar este gradiente. Este proceso se repite hasta que el modelo alcanza un nivel de precisión deseado o hasta que se cumple un criterio de detención, como un número máximo de iteraciones (Friedman, 2001). GBM también incorpora

técnicas de regularización para evitar el sobreajuste. La regularización se puede lograr mediante la introducción de un parámetro de aprendizaje que escala el impacto de cada nuevo árbol, limitando su contribución al modelo final. Además, se pueden aplicar técnicas como la poda de árboles y la selección aleatoria de subconjuntos de características para mejorar la generalización del modelo (Hastie et al., 2009).

El éxito de GBM se debe en gran parte a su flexibilidad y capacidad para manejar datos complejos. Sin embargo, esta flexibilidad viene acompañada de un costo computacional elevado y de una mayor complejidad en su ajuste. El ajuste adecuado de los hiperparámetros, como la profundidad de los árboles, la tasa de aprendizaje y el número de iteraciones, es crucial para obtener un buen rendimiento del modelo (Chen & Guestrin, 2016).

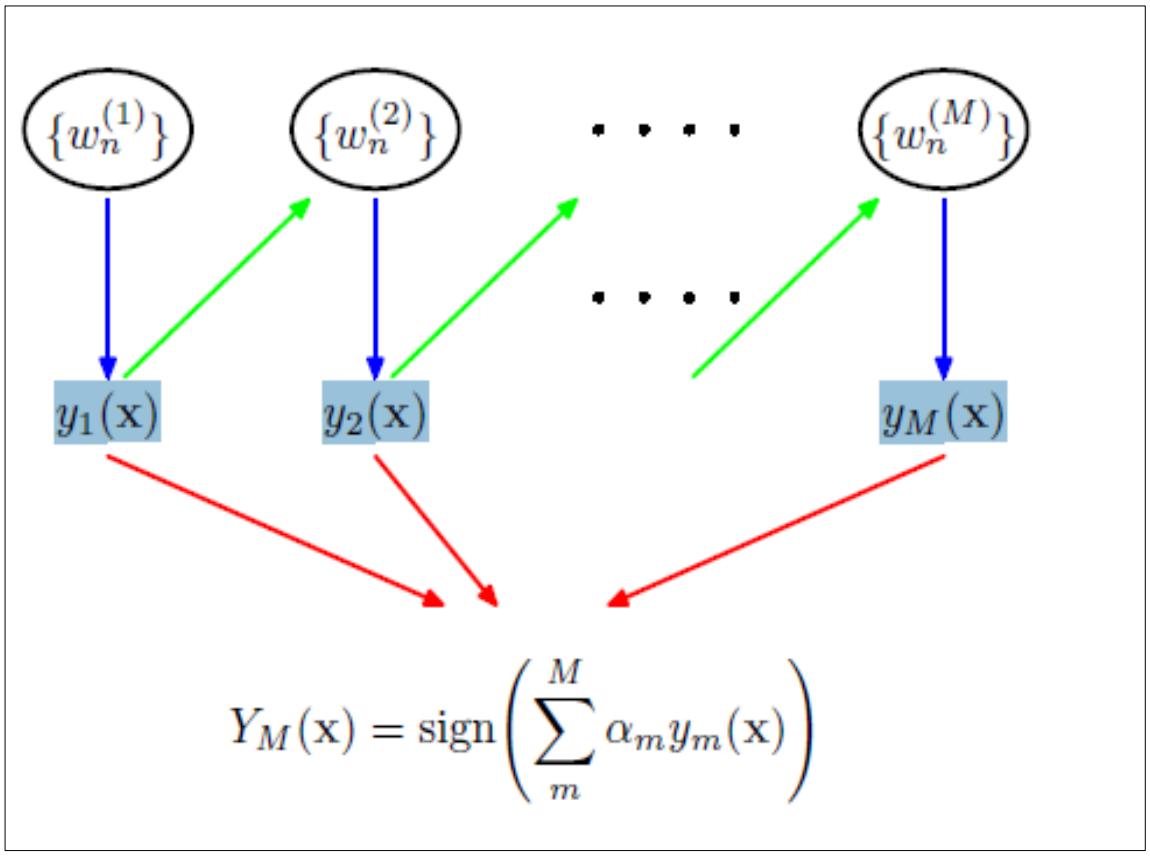


Figura 1.10: Esquema de cómo funciona el Boosting en los entrenamientos del modelo, tomado de Bishop (2006).

Algoritmos de Extreme Gradient Boosting Machine

El algoritmo de *Extreme Gradient Boosting Machine* (XGBM), es una implementación modificada del algoritmo GBM anteriormente descripto. Durante los últimos años, ha ganado gran popularidad en la comunidad de aprendizaje automático debido a su eficiencia, flexibilidad y rendimiento superior en una amplia gama de tareas de predicción con grandes volúmenes de datos complejos (Chen & Guestrin, 2016; Ke et al., 2017; Zheng et al., 2022; Zhong et al., 2020).

XGBoost fue desarrollado por Chen y su equipo, quienes introdujeron una mejora significativa sobre los métodos tradicionales de GBM, abordando problemas de escalabilidad y ofreciendo un conjunto más amplio de funcionalidad. Este algoritmo utiliza la misma idea fundamental que GBM, donde se construyen modelos de árboles de decisión en secuencia para corregir los errores de los modelos anteriores (Chen & Guestrin, 2016). Sin embargo, introduce varias mejoras:

(1) La incorporación de un esquema de regularización más sofisticado que controla la complejidad de los modelos, ayudando a prevenir el sobreajuste. Esta regularización incluye términos como la penalización L1 (*lasso*) y L2 (*ridge*), que ajustan los pesos de las hojas de los árboles para evitar el sobreajuste de modelos (Chen & Guestrin, 2016).
(2) Otra mejora importante es su enfoque de reducción, donde se aplica una tasa de aprendizaje para escalar la contribución de cada nuevo árbol antes de añadirlo al modelo final. Esto permite que el modelo se ajuste de manera más gradual, lo que resulta en un mejor rendimiento. Además, también se añadió "*column subsampling"*, que selecciona aleatoriamente un subconjunto de características para entrenar cada árbol. Este parámetro no solo mejora la generalización del modelo, sino que también reduce el tiempo de entrenamiento (Friedman et al., 2000, Friedman, 2001; Chen & Guestrin, 2016).
(3) También optimiza el uso de la memoria y la eficiencia computacional mediante el uso de una estructura de datos llamada DMatrix, que almacena los datos de manera compacta y permite realizar cálculos

más rápidos. Además, soporta el uso de múltiples núcleos de CPU y GPUs para paralelizar el proceso de entrenamiento, lo que lo hace altamente escalable incluso en conjuntos de datos masivos (Chen & Guestrin, 2016).

Redes Neuronales

Las redes neuronales artificiales (RNAs) comenzaron a desarrollarse en 1943 con el modelo de McCulloch y Pitts, quienes propusieron un marco matemático para describir el funcionamiento de una neurona biológica mediante operaciones lógicas básicas (McCulloch & Pitts, 1943). Se definen como transformaciones geométricas no lineales que combinan una función lineal con una no lineal (Aggarwal, 2018). Cada unidad de una red neuronal recibe entradas, procesan la información y genera salidas que pueden ser utilizadas en diversas aplicaciones (Ertel, 2017; Géron, 2022).

Las redes neuronales están organizadas en capas: una capa de entrada, que recibe los datos de entrada en forma de $x_1, x_2, x_3, \dots, x_n$, una o más capas ocultas, que procesan la información mediante operaciones matemáticas y ponderan los datos de entrada mediante coeficientes $\omega_1, \omega_2, \omega_3, \dots, \omega_n$ conocidos como pesos, y una capa de salida, que proporciona el resultado final de la red. La combinación lineal de estas capas se expresa como:

$$z = \sum_{i=1}^{n} \omega_i \, x_i + b, \tag{6}$$

donde b es el sesgo, termino adicional que ajusta la salida del modelo.

El valor z resultante se transforma mediante una función de activación $f(z)$, que introduce la no linealidad al modelo. Entre las funciones de activación más utilizadas están la sigmoide definida anteriormente en la ecuación 4, que mapea los valores a un rango entre 0 y 1 (Rumelhart et. al., 1986), y la *rectified linear unit*, también conocida como ReLu definida como:

$$f(z) = \max(0, z)\,, \tag{7}$$

(Ertel, 2017; Géron, 2022).

La propagación hacia adelante es el proceso mediante el cual los valores de entrada se procesan a través de las capas para calcular una salida final. Este cálculo se realiza de manera jerárquica, aplicando sucesivamente combinaciones lineales y funciones de activación en cada capa. Para ajustar los parámetros ω_i y b, se utiliza el algoritmo de retropropagación del error, basado en el cálculo del gradiente de la función de costo. Este ajuste se realiza mediante el método de descenso por gradiente:

$$w_i \leftarrow w_i + \Delta w, \tag{8}$$

donde η es la tasa de aprendizaje. Este proceso se repite iterativamente hasta alcanzar un mínimo local de la función de costo J, que mide la discrepancia entre las predicciones $\hat{y}$ y los valores reales y de manera tal que:

$$J = ||\hat{y} - y||^2, \tag{9}$$

(Rumelhart et al., 1986).

K-Means

El algoritmo de *K-Means* es una técnica de agrupamiento ampliamente utilizada en minería de datos y aprendizaje automático para la clasificación no supervisada de datos. Su gran utilización se debe a su simplicidad, eficiencia y aplicabilidad en diversas áreas, ya que se puede utilizar con distintos volúmenes de datos (Harrington, 2012). Este algoritmo busca dividir un conjunto de datos en k grupos o *clusters*, de manera que los elementos dentro de un mismo *cluster* sean más similares entre sí que con los elementos de los otros *clusters* (Fig.1.11) (Müller, et. al,2016; Ertel, 2017).

Para lograr la generación de los grupos, *K-Means* forma los *clusters* siguiendo un proceso iterativo (Fig.1.11) que minimiza la variabilidad dentro de cada grupo. Primero, selecciona el número de centroides iniciales aleatoriamente. Luego, asigna cada dato al grupo cuyo centroide esté más cercano utilizando las distancias como métrica de similitud. Una vez que todos los puntos han sido asignados, se recalculan los centroides como el promedio de todos los puntos dentro de cada clúster. Este proceso de asignación y

actualización se repite hasta que los centroides dejan de cambiar significativamente entre iteraciones o se alcanza un número máximo de iteraciones (Fig. 1.11). De esta manera, *K-Means* logra agrupar los datos en k *clusters* que minimizan la variabilidad interna y maximizan la separación entre grupos (Davis et. al,1986; Harrington, 2012; Müller, et. al,2016).

Uno de los principales desafíos en *K-Means* es determinar el número adecuado de *clusters*. En este trabajo de tesis, se utilizará el método del codo *(Elbow Method),* el cual calcula la suma de los errores cuadráticos dentro de los *clusters* para diferentes números de *clusters* y elige el punto donde la disminución de la suma de los errores se estabiliza, formando un “codo” en la gráfica (Bishop, 2006).

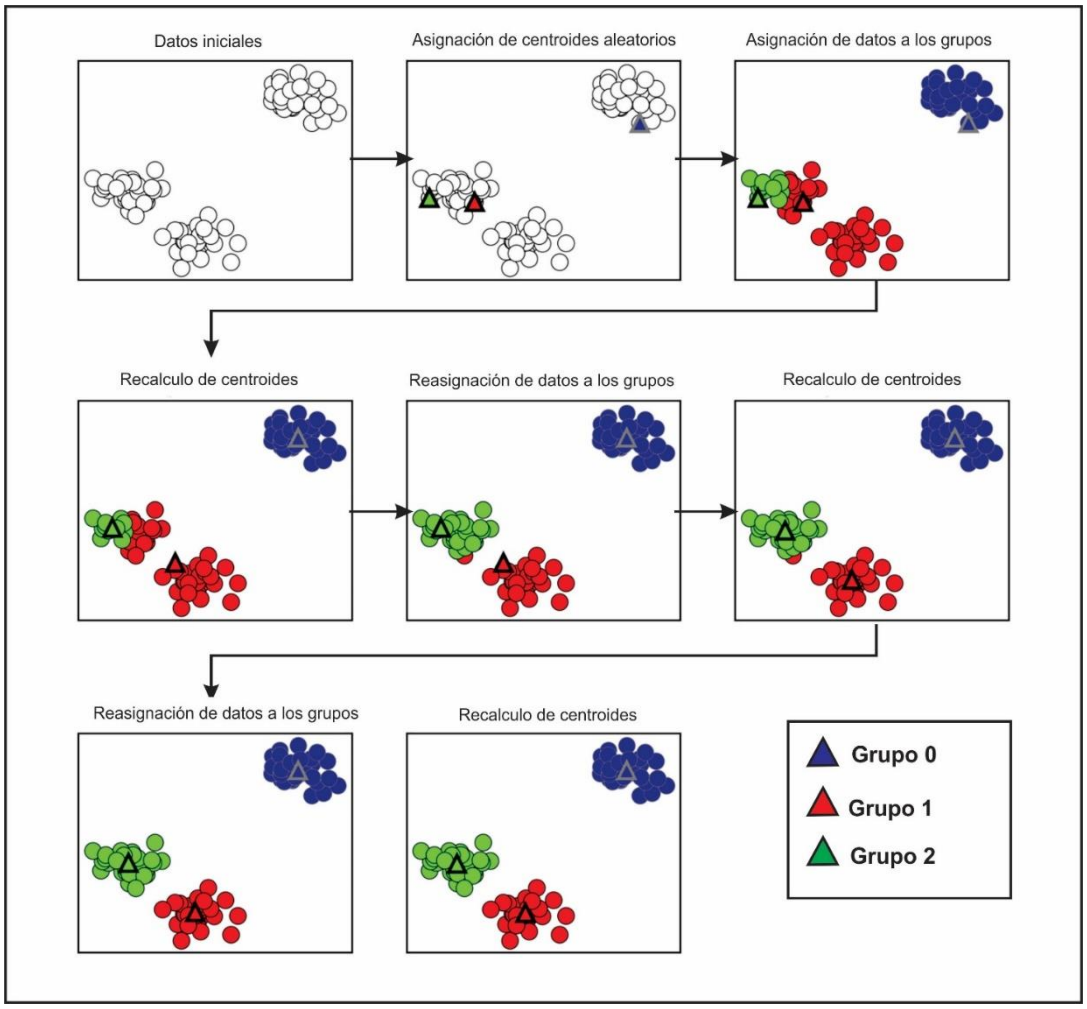


Figura 1.11: Esquema de cómo funciona el proceso de generación de grupos, en el algoritmo *K-Means*. Tomada y modificada de *Müller, et. al (2016*).

2

Estado del arte

2.1 Evolución del concepto de Facies Sedimentarias

El término “facies sedimentarias” fue introducido en la geología por Nicholas Steno en 1669 y modernizado en el siglo XIX por Gressly (1838), refiriendo a la totalidad de los aspectos litológicos y paleontológicos de un determinado cuerpo de roca. Gressly observó que los sedimentos variaban en textura, composición y estructura según las condiciones específicas de su depósito, lo que lo llevó a desarrollar un sistema de clasificación basado en estas variaciones (Gressly, 1838; Walker 2006).

Tras los primeros trabajos de Gressly, el concepto de facies fue combinado con el uniformitarismo propuesto por Lyell entre 1830 y 1833. El uniformitarismo proponía que los procesos observados en la actualidad podían explicar las condiciones pasadas. La conjunción del concepto de facies sedimentaria y de uniformitarismo, definió el concepto utilizado hoy en día de modelo de facies y que permitió interpretar las secuencias estratigráficas como registros de cambios ambientales, facilitando la correlación de secuencias sedimentarias en diferentes localidades. Con base en lo anterior, surgen los primeros trabajos con una mirada integral del concepto de facies (Oomkens y Terwindt, 1960; Evans 1965; Allen, 1963; Bernard et al., 1962; Bouma, 1962; Shearman, 1966; Evans et al.,1969).

La ley de Walther (Walther, 1894) establece que las facies depositadas lateralmente en un mismo ambiente de sedimentación aparecerán superpuestas verticalmente si las condiciones deposicionales permanecen constantes a lo largo del tiempo. Luego Middleton (1973) ofreció una traducción de la definición de Gressly, señalando que diferentes autores lo utilizaban el término de distintas

maneras; y para Middleton (1978) lo redefinió añadiendo terminología empleada por De Raaf et al. (1965). En esta redefinición, las facies son designaciones descriptivas diferenciadas por sus aspectos litológicos, estructurales y orgánicos detectables en el campo, a las que luego se dará una interpretación ambiental. Para Middleton, la clave de la interpretación de las facies consiste en combinar las observaciones realizadas durante la definición de las facies sedimentarias con información comparativa procedente de otras unidades estratigráficas o de ambientes sedimentarios modernos. En este mismo trabajo, Middleton introduce el termino sistema depositacional, definiéndolo como conjuntos de medios sedimentarios y sus productos asociados lateral y verticalmente de forma natural y limitados por discordancias y hiatos (Middleton,1978).

Uno de los estudios clave en esta fase fue el de Harold Reading, quien en 1978 publicó “Sedimentary Environments and Facies”, una obra que se convirtió en una referencia en el campo de la sedimentología. Reading introdujo un enfoque sistemático y clasificatorio para estudiar los ambientes sedimentarios, detallando cómo las características de las facies pueden usarse para inferir el entorno deposicional. Su trabajo ayudó a consolidar un sistema de clasificación para los ambientes sedimentarios que aún se usa ampliamente (Reading, 1978).

En los años posteriores, existieron publicaciones relacionada a los modelos de facies y temas relacionados de autores como Miall (1978; 1984; 1988; 1996; 2014; 2022), Folk (1980), Matthews (1984), Boggs (1987; 2006), Walker (1979;1984;1992; 2006), Scasso & Limarino (1997), Catuneanu (2006), Collinson & Thompson (2006), Tucker & Wright (2009), Leeder (2012). En ellos, se terminó de afianzar la definición actual de facies sedimentarias como una capa de roca con características físicas y químicas específicas que se formó bajo ciertas condiciones de sedimentación. El concepto de asociación de facies como grupo de facies sedimentarias que están relacionadas genética y espacialmente.

Por su origen biológico y los procesos involucrados en su depositación, las rocas sedimentarias carbonáticas tienen diferencias marcadas respecto a las rocas sedimentarias silicoclásticas. Esto conllevó a que se desarrollaran modelos de facies sedimentarios especializados en este tipo de rocas. Dentro de estos, se destacan los modelos de facies desarrollados por Folk (1959) y Dunham (1962), quienes se centraron en la composición y el análisis textural de la roca. La clasificación de Folk se enfoca en los componentes texturales y

mineralógicos de las rocas carbonáticas, permitiendo una descripción de sus características internas, mientras que la clasificación de Dunham, se centra en la relación entre el esqueleto de la roca y el espacio poroso (i.e. fábrica), lo que facilita su interpretación en términos de ambiente de depositación.

Además de los sistemas de clasificación propuestos por Folk (1959) y Dunham (1962), los aportes de Embry & Klovan (1971) resultaron cruciales para la caracterización de las rocas carbonáticas en términos de su ambiente depositacional. Desarrollaron un sistema de clasificación adaptado específicamente para el análisis de carbonatos arrecifales y bioconstrucciones, introduciendo términos nuevos para describir las rocas según el rol de los organismos en la formación de la estructura. Estos nuevos términos, ausentes en los sistemas de Folk (1959) y Dunham (1962), permitieron una interpretación detallada de los ambientes de arrecife, que son complejos y tienen características distintas de los ambientes más uniformes a los que están dirigidas a las clasificaciones previas. Así, la clasificación de Embry & Klovan (1971) complementa los modelos de Folk (1959) y Dunham (1962), ofreciendo un enfoque aplicable a ambientes altamente biodiversos y de alta energía. Estos tres modelos de facies sedimentarias, en conjunto, proveen una metodología integral para la clasificación y análisis de las rocas carbonáticas, abordando tanto sus características texturales como su origen deposicional.

Los modelos de facies de las rocas sedimentarias volcaniclásticas han evolucionado a lo largo de los años, reflejando la complejidad de sus procesos de formación y las diversas fuentes de material volcánico. Fisher (1961) fue pionero en definir y clasificar las rocas volcaniclásticas, identificando tres categorías basadas en el origen de los clastos, epiclásticas, piroclásticas y autoclásticas. Posteriormente, Fisher & Schmincke (1984) ampliaron esta clasificación para incluir a los clastos generados por la interacción entre agua y magma. La introducción de estas clasificaciones en el idioma español, se dieron de la mano de los trabajos de Terruggi et al. (1978) y Mazzoni (1985), quienes tradujeron y adaptaron estas definiciones. En su interpretación, los depósitos volcánicos resedimentados y retrabajados por procesos del ciclo exógeno fueron redefinidos como “piroclásticos reelaborados,” subrayando su transformación y reorganización en condiciones sedimentarias.

En 1987, Cas y Wright propusieron una clasificación alternativa basada no en el origen de los clastos, sino en los mecanismos de transporte y depositación involucrados, sugiriendo un enfoque más funcional para el análisis de estos depósitos. Fisher & Smith (1991), realizan un análisis detallado de los ambientes sedimentarios volcaniclásticos, proporcionando ejemplos de la interacción entre procesos volcánicos y sedimentarios.

McPhie et al. (1993) desarrollaron una clasificación genética, integrando tanto los mecanismos de formación de los clastos como los procesos de transporte y depositación, ofreciendo un esquema más comprensivo y detallado para estudiar estos materiales.

Leyrit & Montenat (2000) y White & Houghton (2006) por su parte, propusieron una nueva perspectiva al identificar cuatro tipos de depósitos según su origen genético: piroclásticos, autoclásticos, hialoclásticos y peperíticos. Estos autores también restringieron el término “volcaniclástico” y recomendaron denominar a los depósitos generados en procesos sedimentarios como “rocas sedimentarias de procedencia volcánica,” aclarando así las diferencias entre materiales volcánicos autóctonos y aquellos que han pasado por procesos sedimentarios antes de su depósito final. Recientemente, Murcia et al. (2013) realizaron una revisión de estas clasificaciones, consolidando el conocimiento acumulado y ofreciendo una visión integradora de los distintos esquemas de clasificación de rocas volcaniclásticas. Por su parte Petrinovic & D’Elia (2018) realizan en su libro “Rocas Volcaniclásticas: Depósitos Procesos y Modelo de Facies: desde el origen hasta las zonas finales de depositación” una importante contribución ya que introduce metodologías actualizadas para la observación y descripción de este tipo de depósitos. Además, presentan detallados modelos de facies que ayudan a entender la distribución y las características de estas rocas desde su formación hasta su depositación final.

A modo de conclusión, aunque cada autor y subcampo de la sedimentología ha desarrollado sus propias metodologías y enfoques, todos convergen en la importancia de describir los procesos sedimentarios. Esta focalización en los procesos sedimentarios permite la formación de asociaciones de facies, que son cruciales para interpretar los paleoambientes sedimentarios. Desde los primeros trabajos hasta las clasificaciones más modernas, el objetivo principal ha sido entender cómo las condiciones depositacionales específicas y

los procesos que las acompañan determinan las características litológicas y paleontológicas de las rocas. Esta comprensión es esencial para la correlación de secuencias sedimentarias y la interpretación de los ambientes en el registro sedimentario, proporcionando una base sólida para la investigación geológica y la exploración de recursos naturales.

2.2 Historia del desarrollo del Machine Learning

La historia del desarrollo del aprendizaje automático ha sido recopilada por Ertel (2017), Skansi (2018) y Abeliuk & Gutiérrez (2021).

Estos autores coinciden que el inicio del desarrollo del aprendizaje automático se dio con Turing (1950), quien pública un trabajo donde se planteó la posibilidad de máquinas inteligentes y donde se formuló el “Test de Turing” para evaluar si una maquina podía imitar el comportamiento humano. Este trabajo fue revolucionario, ya que no solo planteaba la posibilidad de una inteligencia artificial, sino que también abría el camino para el desarrollo de sistemas de aprendizaje automatizado.

Simultáneamente, los científicos comenzaron a estudiar cómo se podían emular los procesos neuronales humanos en máquinas. Es así, como Rosenblatt, psicólogo y científico computacional, publica “Perceptrón” (Rosenblatt, 1957), una red neuronal de una sola capa diseñada para reconocer patrones simples. Dicho trabajo fue arduamente criticado en el libro “*Perceptrons”* publicado por Minsky & Papert (1969), señalando las limitaciones fundamentales del perceptrón, demostrando que las redes neuronales de una sola capa eran incapaces de resolver problemas no lineales como el problema XOR y limitando su aplicabilidad en tareas complejas. La falta de una metodología para entrenar redes neuronales multicapa contribuyó a que muchos investigadores abandonaran este enfoque, y la investigación en redes neuronales quedara rezagada, produciéndose lo que algunos autores llaman “periodo de invierno” o “periodo de estancamiento” (Abeliuk & Gutiérrez, 2021).

Años más tarde, Rumelhart et al. (1986) desarrolló el algoritmo de retropropagación, el cual permite ajustar los pesos de una red neuronal multicapa de manera eficiente haciendo posible que esta aprenda de sus errores en las

predicciones, y minimice la diferencia entre las predicciones y los valores reales. Este desarrollo permitió el entrenamiento de redes neuronales multicapas, superando así las limitaciones de las redes de Rosenblatt. Gracias a la retropropagación, las redes neuronales comenzaron a ser viables para tareas más complejas de reconocimiento de patrones y procesamiento de datos.

Paralelamente al desarrollo de las redes neuronales, surgieron otros enfoques en el aprendizaje supervisado. Los algoritmos como los *Support Vector Machines* (SVM), introducidos por Vapnik en Cortes & Vapnik, (1995), y los árboles de decisión como el "C4.5" de Quinlan (1993), se destacaron como herramientas poderosas para resolver problemas de clasificación y regresión. En particular, el algoritmo C4.5 permitió la creación de árboles de decisión que podían descomponer problemas complejos en decisiones más simples. En conjunto, estos avances ampliaron las aplicaciones prácticas del aprendizaje automático, especialmente en áreas donde los datos eran complejos y requerían un alto nivel de precisión en las predicciones.

Elman (1990) publica el desarrollo de las redes neuronales recurrentes (RNN), las cuales se destacaron en la modelización de secuencias temporales, siendo utilizadas ampliamente en el procesamiento del lenguaje natural y series de tiempo. Posteriormente, LeCun et al. (1998) fue pionero en el desarrollo de las redes neuronales convolucionales (CNN), que demostraron ser altamente efectivas para el reconocimiento de imágenes y video. Ambas arquitecturas de redes hicieron del aprendizaje profundo (*Deep Learning*) la técnica predominante en el ML moderno.

En los últimos años, el aprendizaje por refuerzo ha cobrado gran relevancia ya que permite que los modelos de ML aprendan mediante interacciones repetitivas con su entorno. La tecnología de *transfer learning* y *federated learning* han ganado popularidad, abordando problemas de reutilización de conocimiento y privacidad de los datos, respectivamente. Estas técnicas expanden las posibilidades de aplicaciones del aprendizaje automático, haciéndolo aplicable en contextos más amplios y respetando la privacidad de los datos de los usuarios (Pan & Yang, 2010; McMahan et al., 2017).

2.3 Machine Learning aplicado en las Geociencias

En las últimas décadas, los avances en tecnología han revolucionado diversos campos, y la geociencia no fue la excepción. Esto ha sido posible gracias a la mejora en la ingeniería de las plataformas en la nube, que ha permitido recopilar, almacenar y procesar datos en entornos virtuales y a la mejora de los recursos informáticos para ejecutar modelos a gran escala (Kanevski et al., 2009; Kanevski & Demyanov, 2015). Estos avances han sido fundamentales para el crecimiento del ML como herramienta clave en la exploración, modelado y predicción de fenómenos geológicos y ambientales (Lary et al., 2016; Reichstein et al., 2019; De Iaco et al., 2022).

Numerosos autores han destacado la utilidad del ML para el análisis de datos y, desde algunos años ya es tema de debate en numerosas conferencias, reuniones y congresos (Karpatne et al., 2019; Dramsch, 2020). Karpatne et al. (2019) recopilan los distintos usos del ML en geociencias, enfatizando cuatro grandes categorías de aplicaciones que ofrecen direcciones que los autores denominan prometedoras. Ellas son (1) detección de objetos y eventos, (2) estimación de variables, (3) previsión a largo plazo de variables y (4) relaciones entre los datos geocientíficos. En todas estas categorías, los autores recalcan que los mayores desafíos están atados a la mala calidad de los datos que se obtienen actualmente, así como la limitación en los años de registros digitales. Por su parte, Dramsch (2020) no solo recopila los aportes de diversos autores al ML dentro de las geociencias, sino que realiza una recopilación teórica de los principales modelos.

En lo que respecta a la geología, autores como Ehrlich et al. (1984); Net & Limarino (2000); Hathon et al. (2003); Richa et al. (2006); Larrea et al. (2014); Berrezueta & Kovacs (2017); Budennyy et al. (2017); Berrezueta et al. (2019); Maitre et al. (2019); Rubo et al. (2019); Maerz (2019); Kazak et al. (2021); Zhang & Cai (2021)**;** Petrelli (2024), Gomez & Camilion (2025) entre otros, han trabajado arduamente en el campo del análisis de imágenes para el análisis petrográfico de rocas. En sus trabajos, han logrado delimitar los distintos límites de granos, y han sido capaces de cuantificar la porosidad y la permeabilidad de la imagen, lo que brinda información directa sobre la física de la roca y el transporte de fluidos. Koeshidayatullah et al. (2020) proponen un sistema completamente

automatizado basado en redes neuronales convolucionales profundas (DCNN) para la identificación y cuantificación de componentes petrográficos en rocas carbonatadas. Mediante una base de datos de aproximadamente 4000 imágenes y más de 13.000 objetos etiquetados manualmente, el modelo alcanzó un F1-score del 92% en tareas de clasificación de imágenes y un 84% de precisión en detección de objetos. El estudio demuestra que, a diferencia de los métodos tradicionales de clasificación petrográfica las DCNN permiten una clasificación rápida, reproducible y con un nivel de detalle comparable al análisis humano.

Dentro del subcampo del análisis de suelos, Grimm et al. (2008) aplicó el algoritmo de RF en el contexto de la cartografía digital de suelos, específicamente para la estimación espacial de la concentración y el carbono orgánico del suelo en Panamá. A partir de variables ambientales como topografía, litología, textura del suelo, entre otras, se generaron mapas de alta resolución del carbono orgánico del suelo para diferentes horizontes del perfil edáfico. El estudio evidenció que las variables topográficas fueron las más influyentes en los primeros 10 cm del suelo, mientras que en niveles más profundos predominaron los controles texturales. Este enfoque supera las limitaciones de los métodos convencionales basados en unidades cartográficas, al ofrecer estimaciones espacialmente continuas y cuantificables en términos de error y relevancia de predictores. Pegalajar et al. (2019) proponen un sistema basado en lógica difusa y redes neuronales artificiales para la clasificación del suelo según el color. Este método aborda la subjetividad inherente a la comparación visual con tablas de Munsell, al utilizar imágenes digitales de las muestras tomadas con cámaras convencionales o teléfonos móviles. La lógica difusa permite modelar la incertidumbre en la percepción del color, mientras que las redes neuronales aprenden patrones no lineales complejos para asignar coordenadas HVC (Hue, Value, Chroma) del sistema Munsell. Wadoux et al. (2020) abordan el mapeo digital del suelo, logrando mejorar la precisión en la predicción de propiedades del suelo a diferentes escalas, y señalan que persisten desafíos importantes relacionados con la inclusión del conocimiento pedológico, la validación de modelos y la cuantificación de incertidumbre. Mancini et al. (2020) por su parte, desarrollaron un sistema de clasificación mediante la combinación de equipos electrónicos con algoritmos de Random Forest, logrando resultados de alta precisión en la detección de los colores.

Dicho desarrollo es realmente innovador dado que permite la identificación en campo del color de manera precisa, sin tener que considerar la subjetividad humana por las variaciones de iluminación.

Diversos estudios recientes han demostrado el potencial del ML para la industria de los hidrocarburos. Aliyuda & Howell (2019) propusieron un modelo SVM para la predicción del factor de recuperación en reservorios de la plataforma continental noruega. Utilizando un conjunto de datos compuesto por 93 reservorios y 30 variables geológicas e ingenieriles, el modelo logró una precisión de 76%. El trabajo destaca que, el enfoque del ML permite incorporar eficientemente la complejidad multivariable inherente a los sistemas de reservorios, resultando útil incluso en etapas tempranas del ciclo de vida de los campos. En una línea complementaria, Aliyuda et al. (2020) aplicaron un algoritmo de RF sobre el mismo conjunto de datos con el objetivo de identificar la importancia relativa de diferentes variables en la predicción del desempeño de los reservorios. Los resultados evidenciaron que parámetros geológicos dependientes de la profundidad, como la heterogeneidad estratigráfica, la porosidad promedio, la profundidad del reservorio y el impacto diagenéticos, tienen una alta relevancia en la estimación del RF. Por otro lado, variables relacionadas con el tamaño del campo, como el volumen de roca y el número de pozos, resultaron más influyentes en la predicción de la tasa de producción. Este estudio aporta una perspectiva cuantitativa y jerárquica sobre cómo interactúan los factores geológicos y de desarrollo en el control del desempeño de los yacimientos.

Demyanov et al. (2019) exploran la capacidad de las redes neuronales de Kohonen, también conocidas como mapas autoorganizativos, para reconocer patrones dentro de las estructuras sedimentarias en facies provenientes depósitos fluviales del Rio Paraná (Argentina). Como resultado, demuestran no solo la utilidad de este tipo de modelos de aprendizaje profundo para el reconocimiento de rasgos sedimentológicos, sino también para la interpretación de ambientes sedimentarios.

Sun et al. (2023) desarrollan el conjunto de datos GAN River-I, una herramienta diseñada para el entrenamiento y evaluación de modelos de aprendizaje automático aplicados a la modelación de sistemas fluviales meandriformes. A partir de 25 simulaciones, los autores construyen un conjunto

de imágenes bidimensionales que representan diferentes configuraciones geomorfológicas. GAN River-I permite explorar la incertidumbre geológica en contextos como la exploración de hidrocarburos, la captura y almacenamiento de carbono, la geotermia y la hidrología, consolidándose como un recurso clave para futuras investigaciones en geomodelado asistido por inteligencia artificial.

Al-Anazi & Gates (2010), Yu et al. (2012) Ai et al. (2019), Wang et al. (2019), Li et al. (2020), entre otros, utilizan SVM para predecir diferentes subambientes sedimentarios, litologías y hasta microfacies. En su trabajo, Al-Anazi & Gates (2010), evalúan con SVM registros de pozo en un reservorio de arenisca heterogéneo para predecir diferentes litologías. Por su parte, Yu et al. (2012), investiga el uso del análisis de imágenes combinado con SVM para la clasificación litológica automática utilizando datos de sensores remotos ASTER, modelos digitales de elevación y datos aeromagnéticos. Años más tarde, Ai et al. (2019) analiza índices generados a partir de curvas gamma-ray (GR) de pozos, para clasificar tres subambientes sedimentarios: frente deltaico, planicie deltaica y lago somero-costa. Al mismo tiempo, Wang et al. (2019) proponen el uso de SVM para identificar microfacies deposicionales utilizando registros de pozos de forma automática. Finalmente, Li et al. (2020), investiga el uso de SVM Laplacianas (LapSVM) en combinación con algoritmo de clustering K-means para la identificación de litología a partir de registros de pozos no etiquetados.

Los algoritmos basados en arboles de decisión como *Random Forest (RF)* o *Extreme Gradient Boosting (XGBoost) son* ampliamente utilizados por autores como Harris & Grunsky (2015), Houghton et al. (2023), Nichols et al. (2023), entre otros para realizar mapas litológicos, clasificaciones litológicas o de ambientes sedimentarios. Harris & Grunsky (2015) investigan el uso de algoritmos de RF para la cartografía litológica predictiva en el norte de Canadá a partir de datos geofísicos y geoquímicos, demostrando así que este tipo de algoritmos podría proporcionar una herramienta valiosa para la exploración geológica y la identificación de unidades litológicas en áreas remotas y de difícil acceso. Años más tarde, Nichols et al. (2023) exploran cómo la textura de los sedimentos y la geoquímica pueden utilizarse para clasificar automáticamente los entornos de depositación en un estuario moderno mediante el uso de este tipo de modelos. Durante el mismo año, Houghton et al. (2023) utilizó un modelo

de XGBoost para diferenciar ambientes deposicionales en un estuario en el noroeste de Inglaterra a partir de la utilización de atributos texturales medidos mediante análisis de tamaño de partículas por láser.

Las Redes Neuronales han ganado terreno los últimos años en lo que respecta a las investigaciones científicas dentro de las geociencias. El software llamado CoreBreakout, desarrollado por Meyer et al. (2020), es una herramienta que transforma imágenes de muestras de testigos de rocas en conjuntos de datos registrados por profundidad. Esta utiliza un flujo de trabajo de aprendizaje profundo basado en el algoritmo Mask R-CNN para segmentar las imágenes. El objetivo principal del sofware es digitalizar y estructurar datos de imágenes de testigos coronas, lo que permite el desarrollo de flujos de trabajos más automatizados. Cai et al. (2023), presentó un modelo mejorado de U-Net para la identificación de microfacies sedimentarias a partir de datos de registros de imágenes de pozos. Holden et al. (2023) exploran el uso de redes neuronales convolucionales para predecir propiedades elásticas, volúmenes minerales, y propiedades geomecánicas y de reservorio a partir de datos sísmicos y de pozos en un reservorio no convencional.

Ippolito et al. (2021), Martin et al. (2021), Germay et al. (2023), Eftekhari, et al. (2024), Liu (2024), entre otros, han abordado en sus trabajos de optimización de la predicción de facies sedimentarias a partir de la combinación de distintos modelos de aprendizaje, tanto supervisados como no supervisados, concluyendo que el rendimiento de sus modelos de predicciones o clasificaciones aumenta considerablemente a partir de combinaciones de tipos de modelos. Insua et al. (2015), Xie et al. (2018), Halotel et al. (2020), Ali et al. (2024), entre otros, centraron sus investigaciones en comparar el rendimiento de los diferentes tipos de modelos de clasificación ante determinados datos geológicos.

3

Materiales y Métodos

La implementación de un modelo de *machine learning* sobre un conjunto de datos combina conocimientos de estadística con optimización matemática. Este proceso se compone de distintas etapas que se relacionan y se retroalimentan en conjunto (Fig. 3.1). Generalmente, se inicia con la obtención, recopilación y ordenamiento de los datos, para luego dar paso a la preparación de los mismos asegurando la calidad y relevancia de la información utilizada. Posteriormente, la selección y entrenamiento del modelo adecuado permitirá capturar las complejidades y patrones del problema. Luego, la evaluación de los resultados, la cual es crítica para garantizar que el modelo aplicado sea robusto y generalizable. Finalmente, este ciclo se cierra con la validación y monitoreo de los resultados del modelo en entornos productivos.

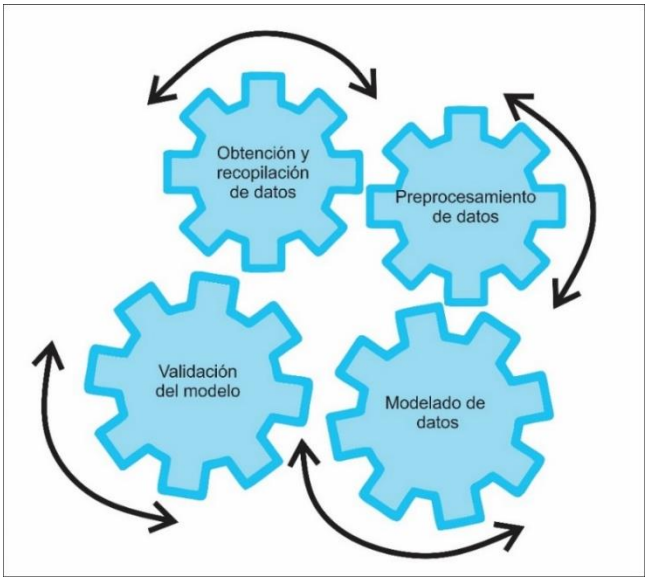


Figura 3.1: Etapas de un análisis de datos clásico

En este capítulo abordaremos los procedimientos resumidos en el párrafo anterior, dividiendo el análisis en cuatro etapas:

1) Obtención y recopilación de datos
2) Procesamiento de los datos
3) Modelado de los datos
4) Validación del modelo

Con el fin brindar mayor claridad en el análisis, en la Figura 3.2 se esquematiza el flujo de trabajo de las cuatro etapas mencionadas, dando detalles específicos de cada instancia en particular.

De esta manera, se explicará el paso a paso de la obtención de los datos y la preparación de los mismos para que estén en condiciones de ser analizados. Luego se explicará el análisis de datos detallando el procesamiento utilizado, antes de dar paso al modelado a fin de cumplir con los objetivos planteados y finalmente validar tanto el modelo entrenado, así como todos los pasos de las etapas anteriores.

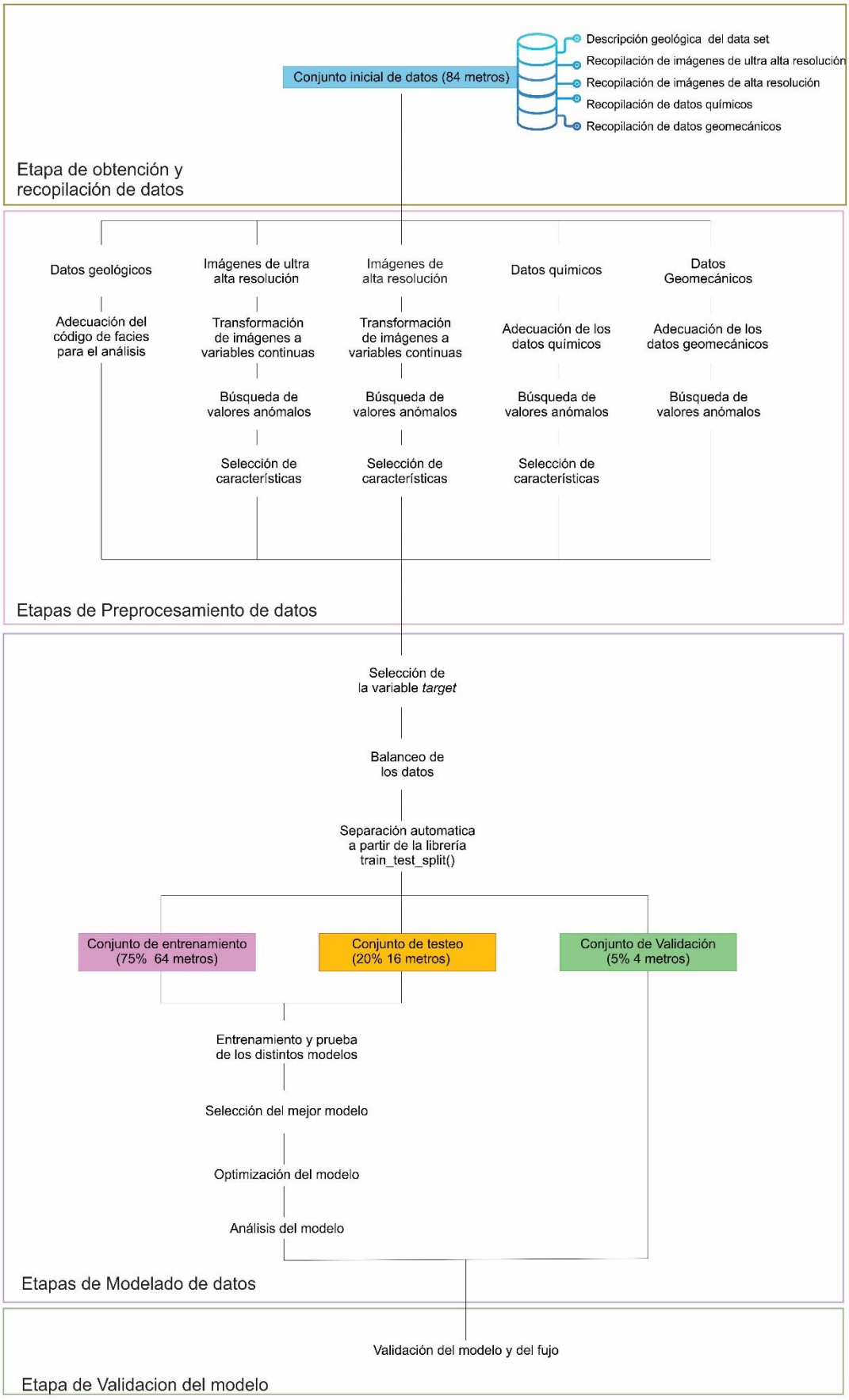


Figura 3.2: Etapas de la metodología utilizada

3.1 Obtención y recopilación de datos

3.1.1 Generación del conjunto de datos

Herramientas utilizadas para la descripción sedimentológica de los datos.

Siguiendo un criterio sedimentológico, se describieron y reconstruyeron las facies sedimentarias de 84 metros de testigos coronas de roca de un conjunto de datos cedido por la empresa EPSLOG S.A en el marco del convenio de colaboración firmado oportunamente con Y-TEC S.A. para la presente tesis doctoral.

El conjunto de datos, al que denominaremos conjunto de datos sintéticos, fue generado de manera tal que las rocas seleccionadas respondan a los diversos procesos y paleoambientes sedimentarios que se dan en la naturaleza. Esto se logró combinando diferentes metros de varios pozos sobre los cuales la compañía EPSLOG S.A realizó servicios (es decir, se compuso un conjunto de datos sintético). Cabe aclarar que, dentro de los 84 metros del conjunto de datos, no existían ejemplos de rocas piroclásticas, y por lo tanto fueron excluidas, por ahora, de los análisis realizados.

La totalidad del conjunto de datos fue analizado con el *Scratch Test Machine* que EPSLOG. S.A diseñó, construye y comercializa (*Scratch Test Machine* Wombat S 1150 MK serie 019). Dicha máquina permite no solo obtener datos de resistencia mecánica al rayado, sino que además a partir de accesorios como el Olympus Vanta VMR Portable XRF series M, logra obtener datos composicionales de fluorescencia de rayos X. También, es posible anexarle una cámara (Basler Ace acA1920) a partir de la cual se obtienen imágenes espectrales de ultra alta resolución (UHRi) con una lente 8gc y de alta resolución (HRi) con una lente 40gc. La denominación gc establecida por el fabricante Basler, es indicativa de que la lente es capaz de obtener imágenes a color.

Para la visualización y descripción se utilizó el *software* CoreDNA (Figs. 3.3 y 3.4). Este software es un desarrollo de la empresa EPSLOG. S.A para el análisis de los datos inicial que extrae el *Scratch Test Machine*, y que permite una visualización triple de las variables. Por un lado, en la pantalla principal (Fig. 3.3), se muestran de manera horizontal y general, los testigos de roca enteros a

partir de las HRi. Debajo de esta, encontramos una segunda pantalla de visualización con las imágenes de detalle a partir de las UHRi. Al costado de ambas visualizaciones, se observa un panel de control con dos solapas (Fig. 3.3). La primera permite intercambiar entre los distintos metros del pozo y realizar ajustes sobre las ventanas de visualizaciones, mientras que la segunda solapa (intercambiable con la anterior) nos permite seleccionar y guardar tanto la textura granulométrica como la estructura sedimentaria para cada tramo de roca seleccionada.

El visor general superior posee una barra de desplazamiento de manera que el usuario pueda recorrer horizontalmente el testigo. Se identifican además sobre esta imagen con puntos verdes los centros de las UHRi para que, al tocar esos puntos, aparezca dicha imagen en el visor subyacente (Fig. 3.3). En ambos visores aparece una línea guía roja, que facilita al usuario la inmediata correlación entre imágenes (Fig. 3.3).

Para la determinación de las estructuras sedimentarias, se utilizó la imagen HRi mostrada en el visor superior, ya que esta permite una visión general a mesoescala de las rocas. Una vez identificada una estructura sedimentaria, se selecciona el tramo de roca que posee dicha estructura y esta se carga en la sopala correspondiente. Para la granulometría, para tamaños menores a 2 mm, la visión más adecuada para su determinación es en el visor de las UHRi, que cuenta con una herramienta de dibujo y medición de vectores lineales, lo que facilitaba una determinación precisa del tamaño de grano, mientras que, para granulometrías mayores a 2 mm, se utiliza el visor general ya que el tamaño de grano es demasiado grande para el visor inferior.

La segunda pantalla (Fig. 3.4), denominada panel de ploteo, se despliega a partir de un botón de activación en el panel superior de la pantalla principal (Fig. 3.3). En esta pantalla accesoria es posible visualizar toda la información numérica en función de la profundidad.

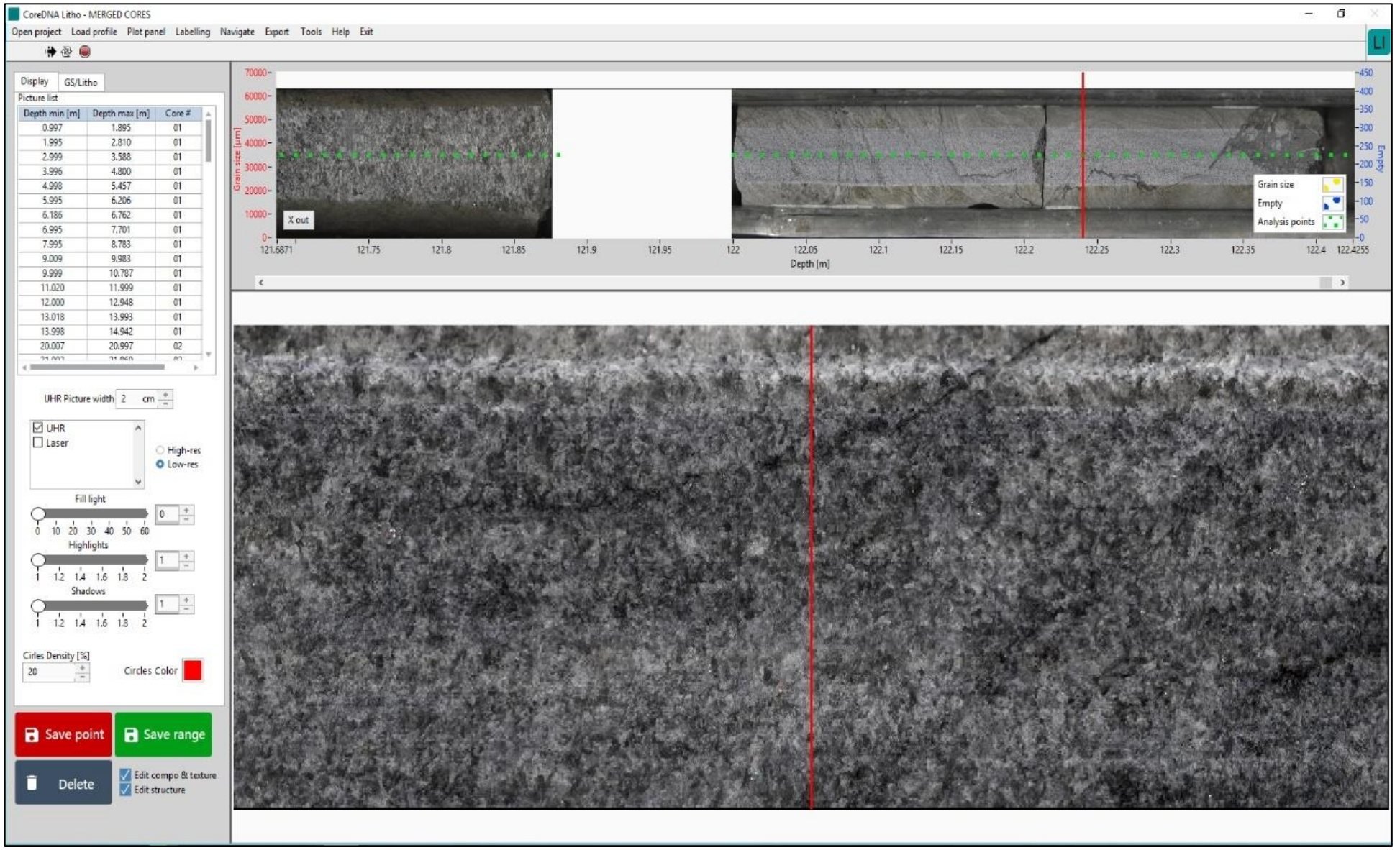


Figura 3.3: Captura pantalla de la primera pantalla del Software Core DNA utilizado para la descripción sedimentológica del *conjunto* de datos. En dicha pantalla se visualizan las imágenes de alta y baja resolución

Figura 3.4: Captura pantalla de la segunda pantalla del Software Core DNA utilizado para la descripción sedimentológica del conjunto de datos. En dicha pantalla se visualizan las concentraciones de los diferentes elementos en el eje y, en función de la profundidad en el eje x.

Las opciones de visualización nos permitieron describir digitalmente en detalle la sedimentología de la roca con una idea precisa de sus características tanto físicas como químicas al proporcionarnos una visión detallada de la textura granulométrica, estructuras sedimentarias mecánicas presentes, estructuras sedimentarias biogénicas (trazas fósiles) y composición química de los testigos de roca en profundidad.

Como resultado de este proceso descriptivo, se obtuvieron las facies sedimentarias de los datos sintéticos, las cuales fueron graficadas en mosaicos para una mejor visualización (Ver Anexo 1). De la misma manera se confeccionó una planilla de cálculo para su posterior análisis estadístico. Para la nomenclatura de dichas facies sedimentarias, se utilizó una denominación basada en Miall (1977, 1988) para rocas silicoclásticas, mientras que para las rocas carbonáticas se utilizó Dunham (1962), ampliada por Embry y Klovan (1971) y modificada luego por Lokier y Al Junaibi (2016). Si bien la base de datos no fue entrenada con datos volcaniclásticos, se realizó la nomenclatura siguiendo las clasificaciones propuestas por Fisher (1961), Fisher & Schmincke (1984), Terruggi et al. (1978) y Mazzoni (1985).

Herramientas utilizadas para la obtención de datos químicos

En lo que respecta a los datos composicionales, los mismos fueron obtenidos a partir de la metodología de Fluorescencia de Rayos X (FRX), la cual se caracteriza por ser una técnica no destructiva de detección simultanea de elementos químicos, lo que permite tener una visión rápida de la composición elemental general de la roca. Dicha técnica se basa en la excitación de los electrones de los átomos de una muestra a partir de un haz de luz o fotón. Este electrón al ser excitado es expulsado de un orbital de mayor energía; luego, el electrón decae a transiciones de energía menores y como resultado de ese proceso el átomo libera energía en el rango de los rayos X. Para cada elemento, la energía liberada en este salto entre orbitales es la misma, por lo que la cantidad de energía liberada puede ser asociada a un elemento químico particular y, en consecuencia, puede ser considerada una huella de identidad de cada elemento. A su vez, el número de fotones emitidos por unidad de tiempo está relacionado con abundancia de cierto elemento en la muestra por lo que, es

posible cuantificar la concentración de elementos específicos (Rollinson, 1993; Larriestra, 2013; Craigie, 2018).

Esta técnica permite identificar y cuantificar la presencia de 38 elementos químicos (Ag; Al; As; Ba; Bi; Ca; Cd; Ce; Co; Cr; Cu; Fe; Hg; K; La; Mg; Mn; Mo; Nb; Nd; Ni; P; Pb; Pr; Rb; S; Sb; Se; Si; Sn; Sr; Th; Ti; U; V; W; Y; Zn; Zr). Además, mide también el error asociado a esa medición, siendo este medido como +/- 3σ (Vanta Family, 2024). Cuando un elemento no es identificado como presente en la muestra, o se encuentra por debajo del límite de detección del instrumento, el instrumento coloca automáticamente un valor de -999.25 identificatorio del error.

Herramientas utilizadas para la obtención de datos geomecánicos

Los datos geomecánicos utilizados, fueron obtenidos a partir de la *Scratch Test Machine*, equipamiento que realiza el ensayo de rayado de manera casi automática. El ensayo de rayado es un método cuasi no destructivo para evaluar las propiedades mecánicas de las rocas mediante la aplicación de una carga controlada que raya la superficie de la muestra. Consiste en empujar a una velocidad constante, una herramienta en forma de cuña sobre la superficie de una roca para generar una ranura. El aparato registra la fuerza que actúa sobre la herramienta al rayar el material con una alta frecuencia de muestreo y una alta precisión y luego devuelve la resistencia a la compresión uniaxial del material (Richard et al. 2012; Germay et al., 2015).

3.2 Procesamiento de datos

3.2.1 Adecuación del código de facies a las nuevas tecnologías

El constante avance y la diversidad de geocientistas que utilizan los distintos códigos de facies, han hecho que no solo se diversificara su interpretación, sino también que sus nomenclaturas no sean siempre respetadas. Dependiendo de la experiencia del descriptor, del objetivo del trabajo y del receptor o demandante es el sesgo o enfoque que se le da al código de facies sedimentarias. Las principales diferencias que se observan en los distintos nombres son:

(1) El idioma: muchos geocientistas prefieren utilizar su idioma nativo para la nomenclatura de facies. Esto puede resultar en grandes diferencias lingüísticas con respecto a los códigos de facies originalmente estructurados y pensados en el idioma inglés.
(2) La cantidad de caracteres utilizados para denotar estructuras sedimentarias: muchos geocientistas utilizan en el nombre de la facies varias estructuras sedimentarias y por ende varios procesos sedimentarios a la vez.
(3) El agregado de características descriptivas en el nombre de la facies sedimentaria que no son propias del proceso sedimentario *sensu stricto*, sino que son producto de procesos post sedimentarios (ej. procesos diagenéticos).

El código de facies sedimentarias resultante de esta adecuación (Anexo II), toma como base a estos códigos internacionalmente aceptados (Dunham, 1962; Embry & Klovan, 1971; Miall, 1977; 1988; Lokier & Al Junaibi; 2016; Fisher, 1961, Fisher & Schmincke, 1984, Terruggi et al.,1978 y Mazzoni,1985), pero introduce cambios a fin de establecer una norma en lo referido a la cantidad de caracteres de la nomenclatura, un orden interno que respeta la naturaleza de la definición de facies sedimentarias y resalta la importancia de las características observadas. Esta adaptación en los códigos de facies permite a su vez una rápida, clara y precisa comunicación de las facies sedimentarias, permitiendo también que sean plausibles de utilizar dentro del ámbito de la ciencia de datos.

La nueva estructura consta de un código interno y otro levemente distinto para la salida gráfica (Fig. 3.5). Esto nace ante la necesidad de que el código interno tenga siempre la misma cantidad de carácteres, en la misma posición para que sea rápidamente analizable por los algoritmos de procesamiento de datos y generar así automáticamente los datos para entrenar los modelos. Mientras que para el operador que leerá finalmente las predicciones del modelo, lo hará viendo el código de la salida gráfica, que sigue las normas de los códigos de facies sedimentarias internacionales. Es así como, el código de facies del modelo (código interno) posee siempre 4 caracteres, el primer y el segundo carácter son siempre en mayúscula y corresponden a la textura de la facies, el tercer carácter es siempre en minúscula y corresponde a la estructura

sedimentaria más predominante (Fig.3.5). Por último, el cuarto carácter también se expresa en minúscula y corresponde a la composición del cuerpo de roca (Fig. 3.5). En contraposición, el código de facies para el operador (código de salida), posee una primera parte correspondiente a la textura, que puede estar compuesta por dos caracteres en mayúscula (esto tiene lugar cuando hay más de una moda en la distribución granulométrica) o por un carácter en mayúscula y un descriptor en subíndice (este subíndice se refiere a la granulometría específica); una segunda parte de largo variable (pudiendo tener de una a tres letras) que corresponde a la estructura sedimentaria predominante, y por último en los casos de que la facies corresponda a composiciones mixtas (mezcla carbonáticas y silicoclásticas) o piroclásticas (líticas, vítreas y cristalinas) se utilizará un último carácter en minúscula para denotar su característica composicional (Fig. 3.5).

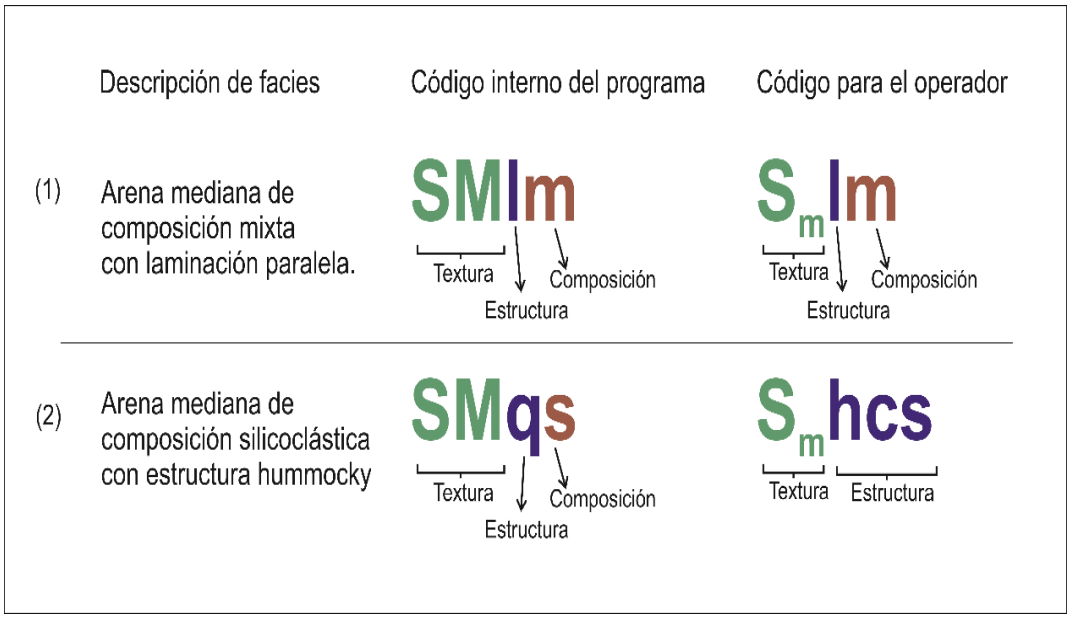


Figura 3.5: Ejemplos de la adecuación del código de facies sedimentarias utilizado para realizar el código interno y el de programa y la salida gráfica del operador.

3.2.2 Metodología para el cálculo de la textura de una imagen

El análisis de imágenes ha sido a lo largo de estos años, un foco de atención en diversas disciplinas. Si bien la primera digitalización y reconstrucción de una imagen se dio en 1920 en el ámbito periodístico, no fue hasta la década de ´60 con la aparición del poder de cómputo necesario, que se desarrollaron las

primeras herramientas de análisis (Gonzalez & Woods, 2009). Desde entonces, estas herramientas se han ido complejizando a medida que las computadoras evolucionan y que la IA avanzaba (Gonzalez & Woods, 2009).

Dentro del ámbito del procesamiento de imágenes, la textura se refiere a la propiedad visual que describe la apariencia superficial de la imagen comparándola con la sensación al tacto que tendría esa imagen (rugosidad, suavidad, aspereza de la superficie). A diferencia del color, que es una propiedad intrínseca de un objeto, la textura se define por la variación espacial de los niveles de gris o el color en cada píxel de una imagen. Esta variación puede ser causada por irregularidades en la superficie del objeto, por la presencia de patrones o por la forma en que la luz incide sobre el objeto al momento de tomar la imagen (Tuceryan & Jain, 1998; Presutti, 2004; Gonzalez & Woods, 2009; Marrón, 2012). La textura de una imagen puede calcularse a partir de una matriz de co-ocurrencia. Dicha matriz, también llamada matriz GLCM (*Grey Level Co-occurrence Matrix),* se define como la distribución de las frecuencias de las intensidades de los niveles de gris de dos píxeles separados por un desplazamiento, *d,* en una o más direcciones.

La figura 3.6 muestra un ejemplo donde d = (1,0) representa la distancia unidad entre dos pixeles ubicados sobre la misma fila y donde A_T es la matriz GLCM resultante (Fig. 3.6). Nótese que por ejemplo para el par (1,1), la frecuencia de ocurrencia de los niveles de grises en la dirección de análisis es de 10 (marcadas en subíndices arcos azules); para el par (1,0), la frecuencia de ocurrencia es 4 (marcadas en subíndices arcos verde); para el par (0,0) es de 6 (marcadas en subíndices arcos rojo) y que el par (0,1) no ocurre, por lo que no tiene frecuencia. Una vez obtenida esta matriz GLCM, esta se simetriza (A_T^{SIM} en Fig. 3.6) (Hall-Beyer, 2017), y luego se transforma a una matriz de probabilidades (P (A_T^{SIM}) en Fig. 3.6) donde cada par posee una probabilidad de ocurrencia ($P_{i,j}$) calculada a partir del número de veces que ocurre cada par i, j, dividido por el número total de frecuencias de pares identificados. Finalmente, con la matriz de probabilidades para cada par de co-ocurrencias, se pueden calcular distintas métricas o parámetros de la imagen realizando cálculos estadísticos, las cuales se listan y explican a continuación:

- Energía: También llamada uniformidad de la imagen mide la homogeneidad local es decir la homogeneidad a cortas distancias entre pixeles, de la imagen a partir de la raíz cuadrada de la suma de las probabilidades de cada par de co-ocurrencias:

$$\sqrt{\sum_{i,j=0}^{N-1} P_{i,j}^2}. \tag{10}$$

- Entropía: Da una idea de cuanta información es esencial si se quisiera comprimir dicha imagen. Matemáticamente se define como:

$$\sum_{i,j=0}^{N-1} -P_{i,j} \ln\left(P_{i,j}\right) . \tag{11}$$

- Contraste: También llamada inercia o varianza de la suma de cuadrados de la imagen. Mide las variaciones locales de la matriz de co-ocurrencia de los niveles de grises de la imagen de una manera exponencial, es decir que mide la diferencia del nivel de gris de un píxel con su píxel vecino. Matemáticamente se define como:

$$\sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2. \tag{12}$$

- Disimilaridad: Similar al contraste, mide las variaciones locales de la matriz de co-ocurrencia de los niveles de grises de la imagen, pero de una manera lineal. Matemáticamente,

$$\sum_{i,j=0}^{N-1} P_{i,j} \,|i-j|. \tag{13}$$

- Homogeneidad: Mide la homogeneidad de la imagen a través de la distribución de elementos de la matriz de co-ocurrencia en su diagonal. En decir, que cuando una imagen es homogénea, la distribución de las probabilidades en la diagonal de la matriz es:

$$\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2} \quad . \tag{14}$$

- Correlación: Mide la probabilidad de aparición conjunta de los pares de pixeles i, j de la imagen, en otras palabras, cuantifica las repeticiones de los pares de pixeles de una imagen. Expresada como:

$$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right], \tag{15}$$

donde la media de i es $\mu_i = \sum_{i,j=0}^{N-1} i(P_{i,j})$; la media de j es $\mu_j = \sum_{i,j=0}^{N-1} j(P_{i,j})$; el desvío estándar de i es $\sigma_i^2 = \sum_{i,j=0}^{N-1} P_{i,j}\,(i-\mu_i)^2$ y, por último, el desvío estándar de j es $\sigma_j^2 = \sum_{i,j=0}^{N-1} P_{i,j}\,(j-\mu_j)^2$.

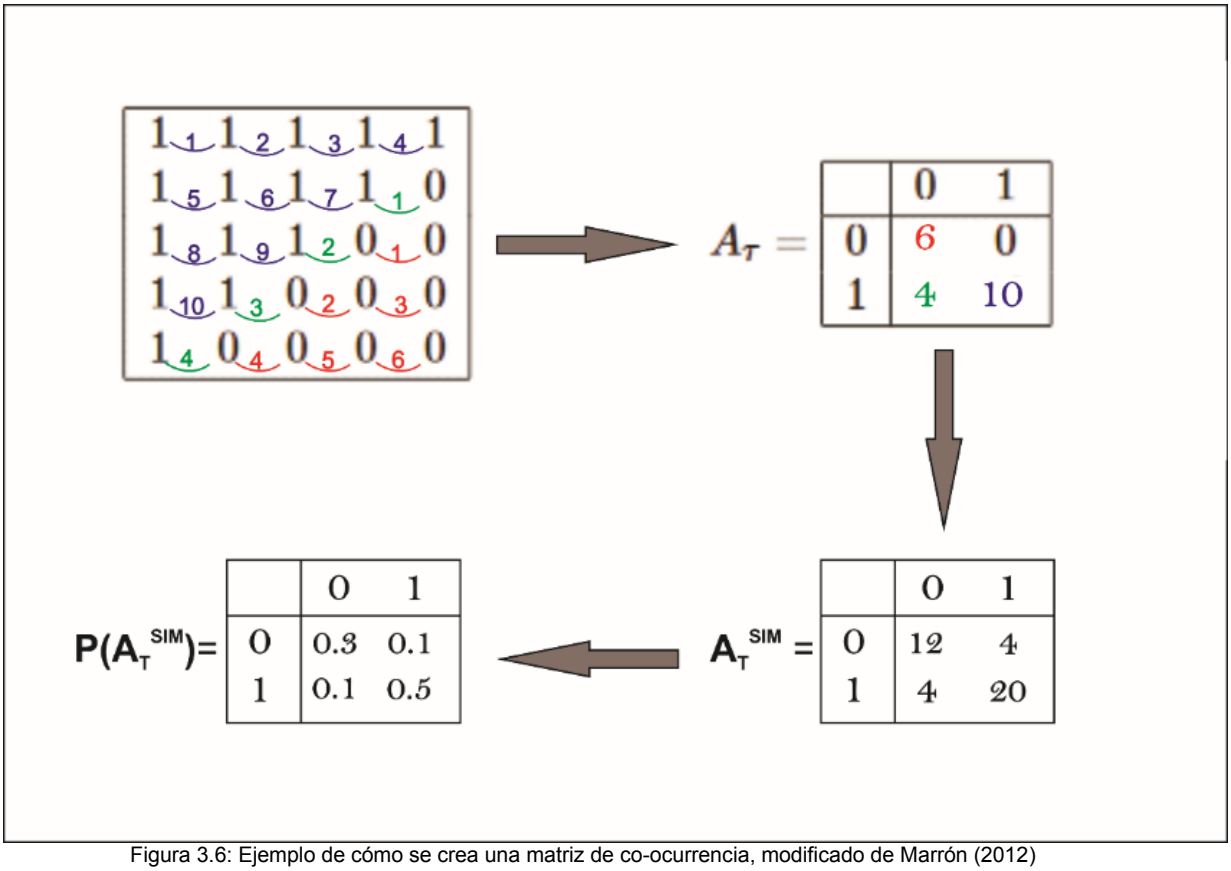


Figura 3.6: Ejemplo de cómo se crea una matriz de co-ocurrencia, modificado de Marrón (2012)

Para el análisis de las imágenes de testigos coronas, tanto de ultra alta resolución (UHRi) como de alta resolución (HRi), se seleccionó este método como forma de pasar de imágenes RGB, cuya dimensión posee 3 capas a una matriz plana unidimensional. Esto no solo facilitó el entendimiento y la observación de patrones antes ocultos en las 3 capas, sino que redujo significativamente el tiempo de procesamiento en lo que respecta al entrenamiento del modelo.

Para probar la factibilidad de esta transformación en relación con las características geológicas, se generaron imágenes de referencia para las texturas granulométricas (Fig. 3.7 A) y las estructuras sedimentarias (Fig. 3.7 B). En el caso de las texturas granulométricas se confeccionó la imagen de referencia a partir de 3 imágenes de UHRi que poseían diferencias marcadas en el tamaño de grano (Fig. 3.7 A). Se seleccionó por un lado un conglomerado, seguido de una limolita gruesa y por último una arenisca mediana. Mientras que, para el caso de las estructuras sedimentarias, la imagen de referencia se confeccionó a partir de imágenes HRi cuyas rocas poseían igual tamaño granulométrico (arena mediana) pero diferente estructura, siendo la primera con estructura tangencial, la segunda con heterolítica flaser y por último una maciza (Fig. 3.7 B).

Dado que esta transformación da como resultado un único valor de textura para cada matriz de co-ocurrencia y que lo que se busca es crear un perfil de comportamiento de estos parámetros en profundidad, se subdividió la imagen en pequeñas sub-imágenes secuenciales en función de la profundidad, que permitió realizar varias matrices de co-ocurrencia y crear así perfiles de comportamiento de las texturas en relación con las imágenes en función de la profundidad (Fig. 3.7 A; B). Para optimizar la cantidad de subdivisiones dentro de las imágenes y optimizar los recursos computacionales se realizó un análisis de los apartamientos de los valores de los parámetros texturales versus ancho de las subdivisiones medido en pixeles. Tomando en consideración que la variación más real posible para los parámetros se da con la ventana mínima posible de 3 pixeles. Se generaron así, curvas de apartamientos tanto para las HRi como para las UHRi (Fig. 3.8 y Fig. 3.9). Al analizarlas, se definió que para las imágenes HRi el ancho óptimo para la transformación de imágenes promediaba los 1.000

píxeles (Fig. 3.8), mientras que, para las imágenes UHRi, el ancho óptimo promediaba los 2.000 pixeles (Fig. 3.9). Además, para que los análisis de frecuencia fueran representativos, se decidió observar la co-ocurrencia de los valores de los pixeles en todas las direcciones posibles de la matriz exceptuando aquellas que generarían una repetición en las comparaciones (0°, 45°, 90°, 135°) comparándolo con el píxel aledaño, es decir a un distanciamiento (0,1).

Una vez corroborado que la metodología a utilizar para la transformación de las imágenes conservaba las características imprescindibles para su clasificación geológica, se procedió a transformar todo el volumen de datos descripto en el apartado 3.1.1

Luego, se analizaron los tramos de las imágenes que contenían faltantes de material. Tanto en las UHRi como en las HRi, el faltante de material está evidenciado por zonas de coloración negra, esto es producto del color de la vaina sobre la cual se dispone el material. Para ello, se tomó una imagen equivalente a 15 cm de largo que contenía faltantes de material (imagen original en Fig. 3.10), y se observó que en esas zonas todas las variables tenían un comportamiento extremo. Para el caso de las variables contraste, disimilaridad y entropía se observó que todas contenían valores de 0, mientras que las variables homogeneidad, energía y correlación tenían valores de 1 (Fig. 3.10). Esto facilitó que las zonas de material faltante fueran fácilmente identificables y posibles de eliminar.

Posteriormente, se realizó un análisis de valores atípicos, descripto en la sección 3.2.4 y un análisis de evaluación del aporte de variabilidad de las distintas variables del conjunto de datos, así como la correlación entre las mismas, que será explicado en el apartado 3.2.5.

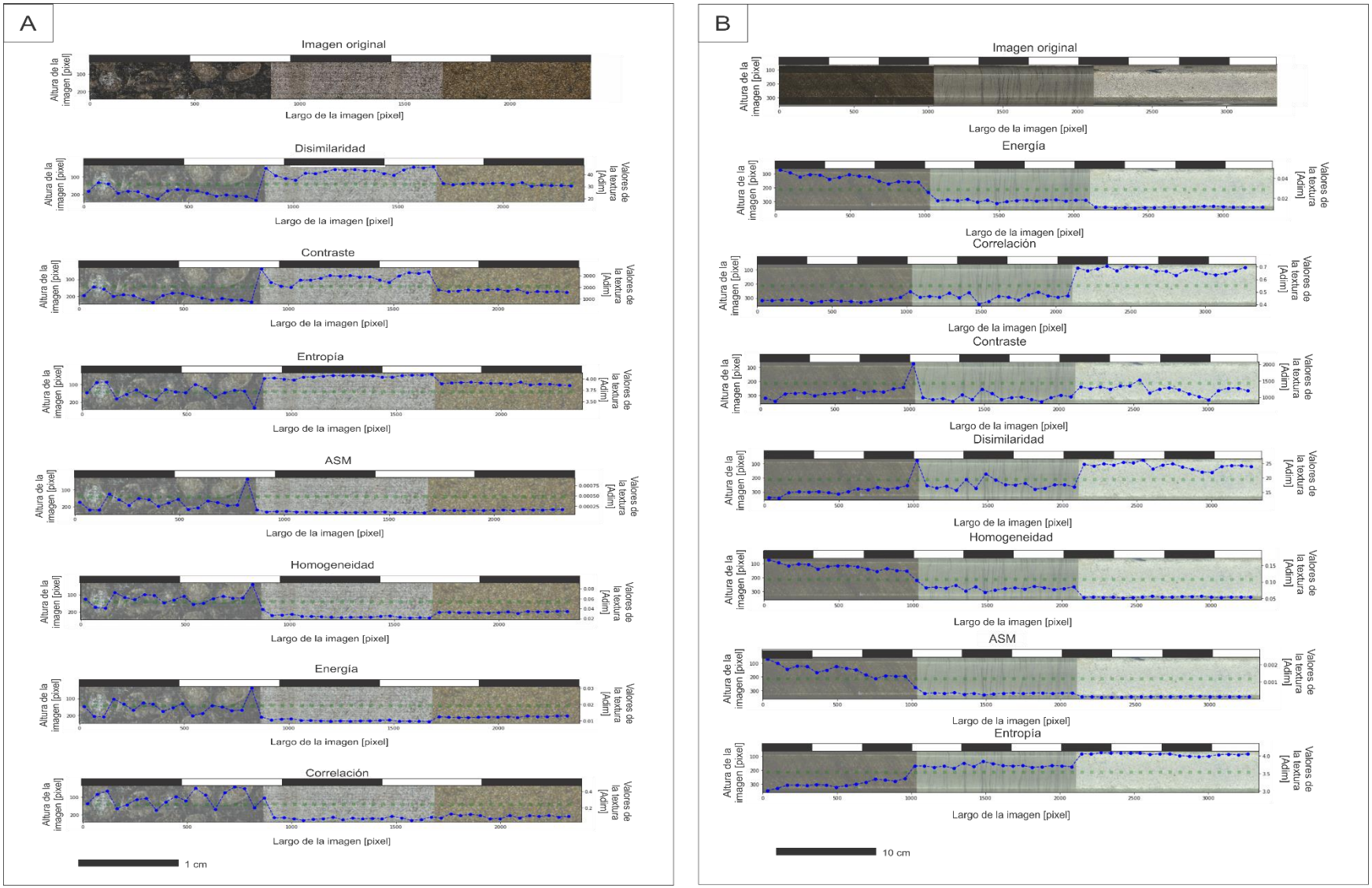


Figura 3.7: Imágenes de referencia para el análisis de la correlación entre las texturas de imagen y las distintas características geológicas. (A) Imagen de referencia para las diferentes tamaños granulométricos y su correlación con las distintas texturas de la imagen. (B) Imagen de referencia para las diferentes Estructuras Sedimentarias y su correlación con las distintas texturas de la imagen.

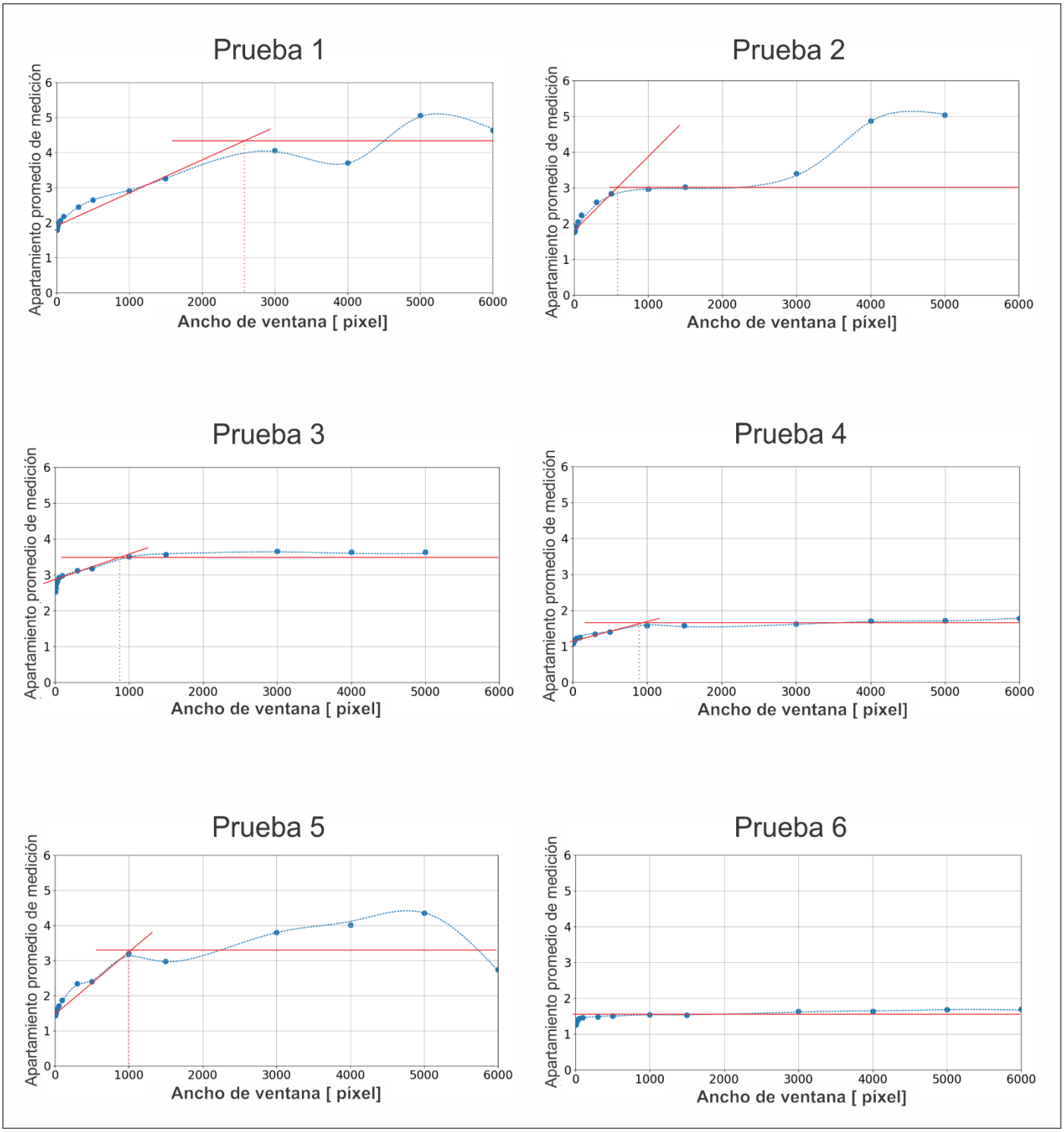


Figura 3.8: Gráficos de Ancho de la ventana vs Apartamiento promedio de las mediciones de texturas de las imágenes para las pruebas realizadas en imágenes de alta resolución.

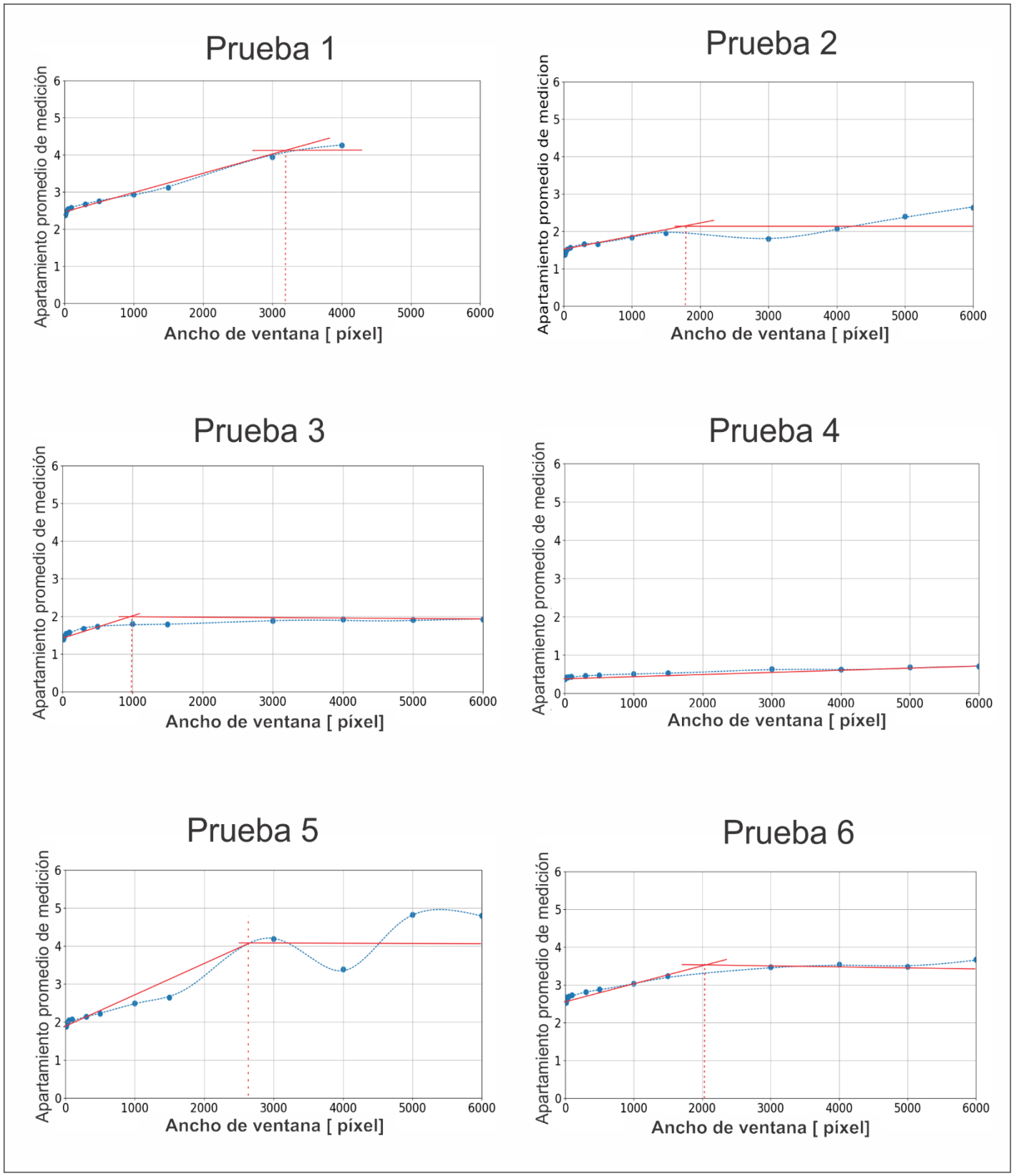


Figura 3.9: Gráficos de Ancho de la ventana vs Apartamiento promedio de las mediciones de texturas de las imágenes para las pruebas realizadas en imágenes de ultra alta resolución.

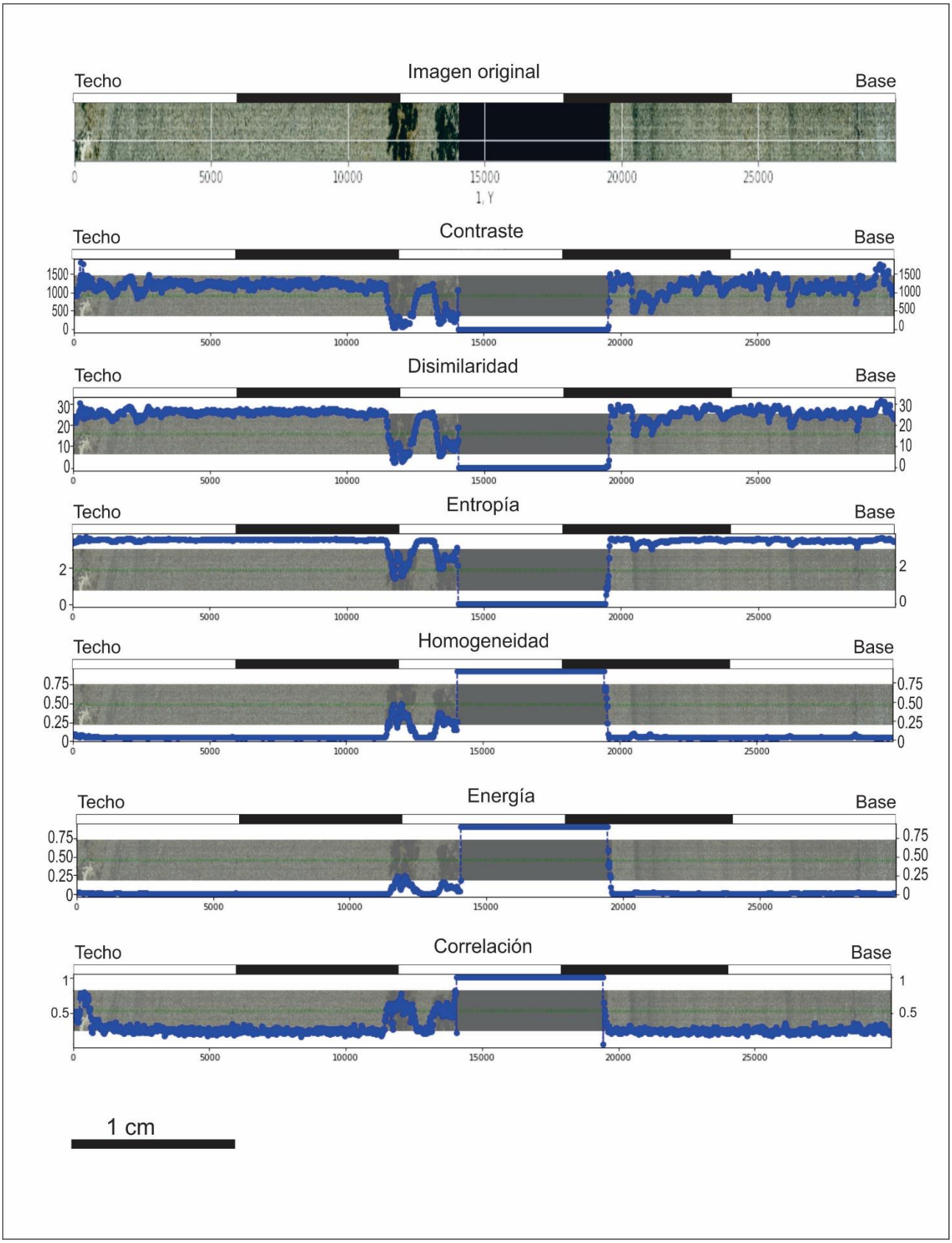


Figura 3.10: Gráfico que ejemplifica el comportamiento de las variables texturales de la imagen en las zonas donde existe faltante del material. Nótese que, en dichas zonas, las variables adquieren un comportamiento extremo (valores en cero y en uno)

3.2.3 Acondicionamiento de los datos químicos

Dentro de los elementos químicos medidos a partir del FRX, y con la idea de conservar la completitud de los datos originales se propuso descartar aquellos elementos químicos con más del 50% de los datos faltantes, es decir, sin datos registrados o con valor del dato debajo del límite de detección. Este umbral se eligió equilibrando la cantidad de datos medidos con la cantidad de *features* (elementos químicos) descartadas por falta de datos.

De esta identificación, se observó que 18 elementos (Ag; As; Bi; Cd; Ce; Co; Hg; La; Mo; Nb; Pr; Sb; Se; Sn; Th; U; V; W) contenían faltantes por encima del umbral del 50%, por lo cual fueron eliminados. Respecto al "valor" de los datos faltantes, se asumió la existencia de los elementos químicos en estos casos, pero con una señal inferior a la detectada por el instrumento, por lo que se asignó el valor de mitad de su error de medición asociado (Alperin, 2013).

Por la forma de medición que tiene esta técnica de FRX, las concentraciones de los elementos presentes en una muestra no son expresadas como partes individuales, sino que son expresadas en forma de proporciones, es decir, como porcentajes (%), partes por millón (ppm) o en casos de muy baja concentración como partes por billón (ppb). Esto hace que la suma de todos los elementos para cada muestra sea un valor constante o semi constante, el cual trae aparejado un gran problema ya que no es posible aplicar las fórmulas de estadística convencional (Alperin, 2013; Grunsky & de Caritat, 2017). En este sentido, se han desarrollo diversas transformaciones para los datos, una de ellas es específica para pruebas que se basan en la distancia como lo es el Análisis de Componentes Principales (PCA), concepto que será explicado en profundidad en el apartado 3.1.5 y es la denominada transformación logcociente centrada (clr), la cual se define como:

$$clr(x) = \left(\ln\frac{x_1}{g(x)}, \ln\frac{x_2}{g(x)}, \dots, \ln\frac{x_D}{g(x)}\right) = (z_1, z_2, \dots, z_D,) = Z, \qquad (16)$$

siendo x_1, x_2 y x_D los datos medidos para cada muestra (vector fila) y $g(x)$ su media geométrica. Como resultado obtendremos una matriz de datos de igual dimensión que la inicial (Aitchison, 1982; Rollinson, 1993; Alperin, 2013; Grunsky & de Caritat, 2017). En el marco de trabajo de esta tesis, en la que se busca

utilizar la mayor cantidad de elementos químicos posibles, se realizó luego de la limpieza y del tratamiento de datos faltantes, la transformación logcociente centrada de los datos. Esto nos permitió realizar el resto del preprocesamiento y el posterior modelado sin la preocupación de observar las relaciones espurias entre las variables y garantizando de esta forma parte de la calidad de los datos.

3.2.4 Acondicionamiento de los datos geomecánicos

La geomecánica de las rocas es una disciplina esencial en la exploración y explotación de hidrocarburos, especialmente en el estudio de las rocas sedimentarias, que al ser depósitos formados por acumulación y compactación de sedimentos a lo largo del tiempo presentan una gran variabilidad en sus propiedades mecánicas. Zoback (2007) enfatiza que la anisotropía mecánica es inherente a las rocas sedimentarias debido a sus diferentes procesos de depositación en capas, el gradiente de presión de los poros y los procesos diagenéticos que afectan a la roca.

El ensayo de rayado es un método cuasi no destructivo para evaluar las propiedades mecánicas de las rocas mediante la aplicación de una carga controlada que raya la superficie de la muestra. Consiste en empujar a una velocidad constante, una herramienta en forma de cuña sobre la superficie de una roca para generar una ranura. El aparato registra la fuerza que actúa sobre la herramienta al rayar el material con una alta frecuencia de muestreo y una alta precisión y luego devuelve la resistencia a la compresión uniaxial del material (Richard et al. 2012; Germay et al., 2015). Una de las ventajas de este método, es la obtención de un perfil en profundidad de la resistencia al rayado de la muestra (curvas en Fig. 3.11). Este registro de alta resolución proporciona información en profundidad sobre las heterogeneidades de los materiales, permitiendo la identificación de regiones geomecánicamente homogéneas.

El conjunto de datos posee originalmente 14 variables, correspondiente a las respectivas mediciones de resistencia al rayado de la roca y la confianza asociada a esas mediciones calculadas automáticamente por el equipo. Estas variables son denominadas de 2 maneras. En el caso de ser mediciones de resistencia al rayado son denominan como “E_STRx”, y en caso de ser

mediciones de confianza se denominan como “E_CONFx”, en ambos casos x corresponde al intervalo de medición en cm. Al adentrarnos en la forma en la que las variables se generan dentro del programa del instrumento, observamos que las variables se construyen a partir de una media móvil de la medición centímetro a centímetro. Esto fue corroborado al graficar el histograma de los datos (Fig. 3.12), se puede observar que los datos tenían la misma distribución de frecuencia. También, se pudo realizar la misma deducción observando el gráfico en profundidad de la resistencia al rayado (Fig. 3.11) ya que se observó que las variables tienen el mismo comportamiento y evolución en profundidad. Por lo expuesto, se decidió trabajar con la primera variable que entrega la máquina (E_STR001) correspondiente a la medición puntual que devuelve el ensayo.

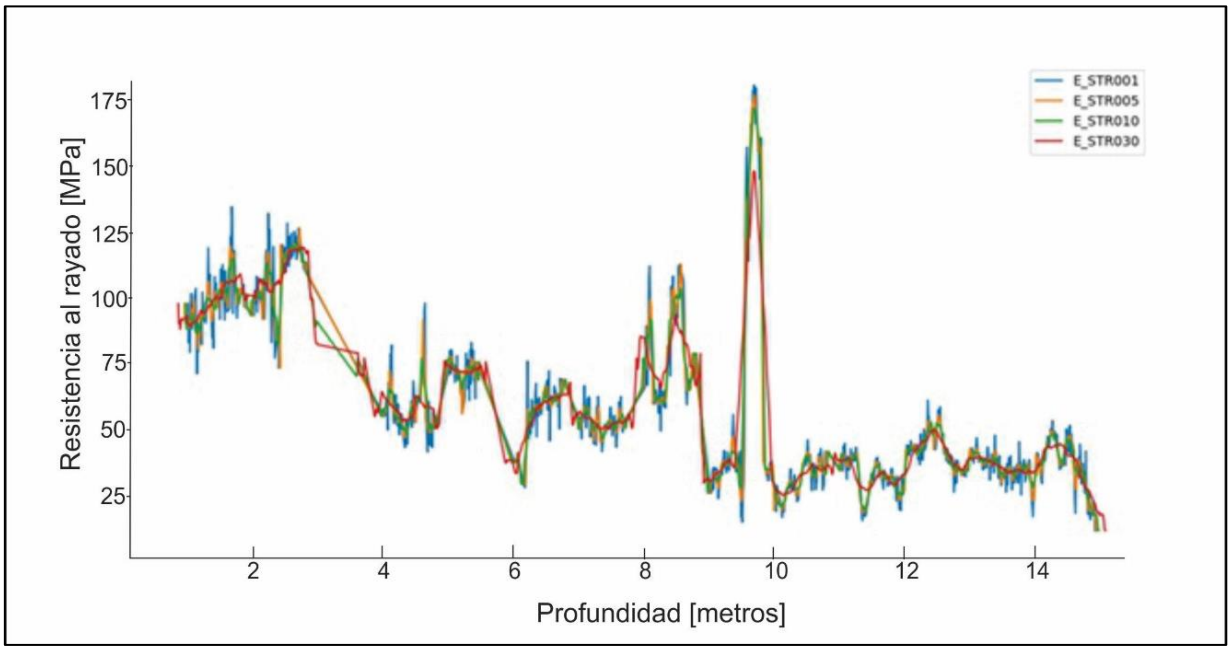


Figura 3.11: Gráficos de perfil en profundidad de la resistencia al rayado de la de *Merge core 1*.

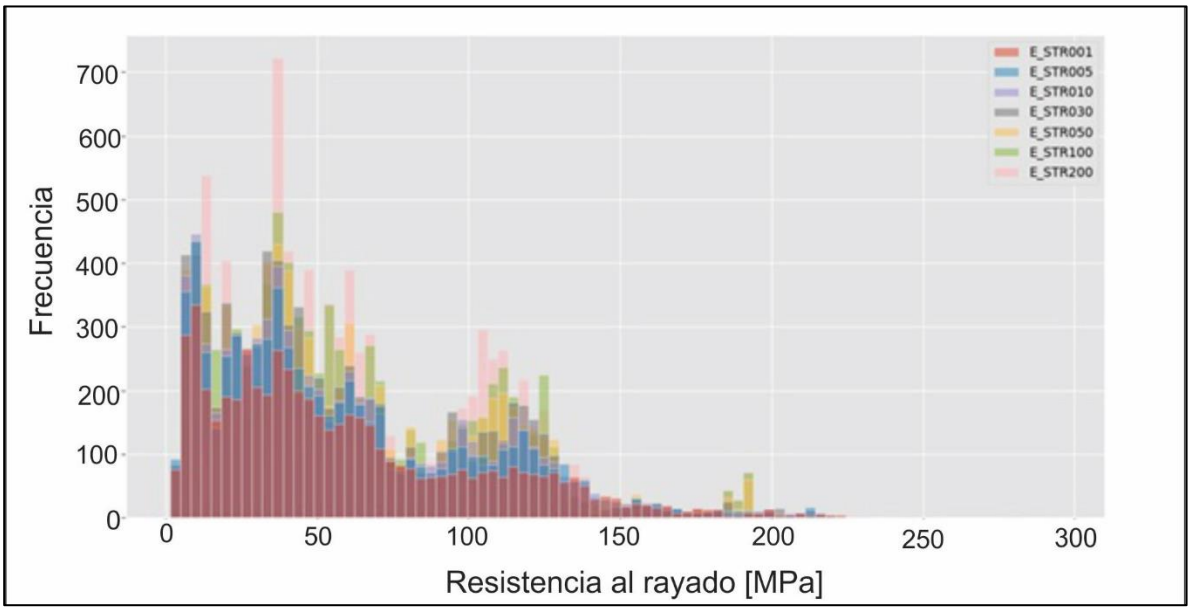


Figura 3.12: Histogramas de frecuencias de las diferentes mediciones que realiza el Scratch Test Machine.

3.2.5 Búsqueda de valores anómalos en los distintos conjuntos de datos.

Aunque todos los conjuntos de datos fueron analizados y/o transformados siguiendo criterios estadísticos y geológicos, estos pueden contener todavía valores que pueden diferenciarse distantes del resto. Estos valores son los llamados valores anómalos o atípicos *(*Alperin, 2013*)*. Asumiendo que el problema es de naturaleza mutivariable, para detectar estos valores anómalos, se utilizó la prueba T2 de Hotellings combinada con la distancia al modelo de cada observación (DmodX) o *Squared Predicction Error* (SPE) que realiza automáticamente el paquete PCA de Python para muestras multivariantes. La prueba T2 Hotellings es un análogo a la prueba T de *Student* para datos multivariados. Esta funciona calculando las distintas pruebas de chi cuadrado en los n componentes superiores, donde se espera que se observe la varianza más alta y, por ende, sea en esos componentes donde se encuentren los valores atípicos. La prueba T2 de Hotellings genera una matriz de valores, que luego se combinan utilizando el método de Fisher, permitiendo identificar los valores de proyecciones que se desvían de su distribución teórica esperada (Hotellings, 1933). En el marco de este trabajo, se decidió analizar los componentes principales de 1 a 5 recomendado por los desarrolladores del paquete utilizado, con un nivel de confianza del 90 %.

Por otro parte, el método DmodX o SPE, calcula la distancia euclidiana entre las observaciones reales y sus proyecciones en el modelo de PCA y genera un ranqueo de estas distancias. Luego, calcula el umbral a partir del cual las muestras superiores son consideradas atípicas.

En el caso del conjunto de datos químicos, como resultado del análisis se observa que varias muestras poseen comportamiento de valores anormales (Fig. 3.13 A1), pero desde un punto de vista geológico no todas pertenecían a valores atípicos. Como se puede observar en la figura 3.13 A1, existen varias muestras por fuera del límite de confianza de la prueba (línea negra punteada Fig. 3.13 A1). Las muestras encerradas en el cirulo rojo pertenecen a 30 muestras que corresponden a profundidades consecutivas, lo cual es llamativo, pero puede deberse a varios factores de origen geológico por lo que en principio no fueron catalogadas como anómalos. Finalmente, se reconocieron 12 muestras de forma aislada e individuales que, con relación a su profundidad no guardan correlación con las de su alrededor, por lo que esas si fueron consideradas valores atípicos. Una vez eliminados esos valores atípicos, se realizó nuevamente el gráfico de dispersión (Fig. 3.13 A2).

En el caso de los datos de UHRi, (Fig. 3.13 B), observamos que los datos originalmente poseen gran cantidad de valores atípicos con respecto al límite de confianza (línea negra punteada Fig. 3.13 B1). Al observar las profundidades de estos 391 puntos, se pudo observar que no correspondían a valores lo suficientemente consecutivos como para asignarle dicha variación a algún proceso geológico, por esta razón se decidió eliminarlos. Una vez realizada esta limpieza, se evaluó nuevamente la dispersión de los datos (Fig. 3.13 B2).

Finalmente, para el caso de los datos de HRi, se observó que 24 muestras poseen comportamiento de valores atípicos por fuera del límite de confianza (Fig. 3.13 C1). Al observar las profundidades, estas no corresponden a valores lo suficientemente consecutivos como para asignarle dicha variación a algún proceso geológico, por esta razón también se decidió eliminarlos. Una vez realizada esta limpieza, se evaluó nuevamente la dispersión de los datos, y se observó como mejoró la dispersión de los datos.

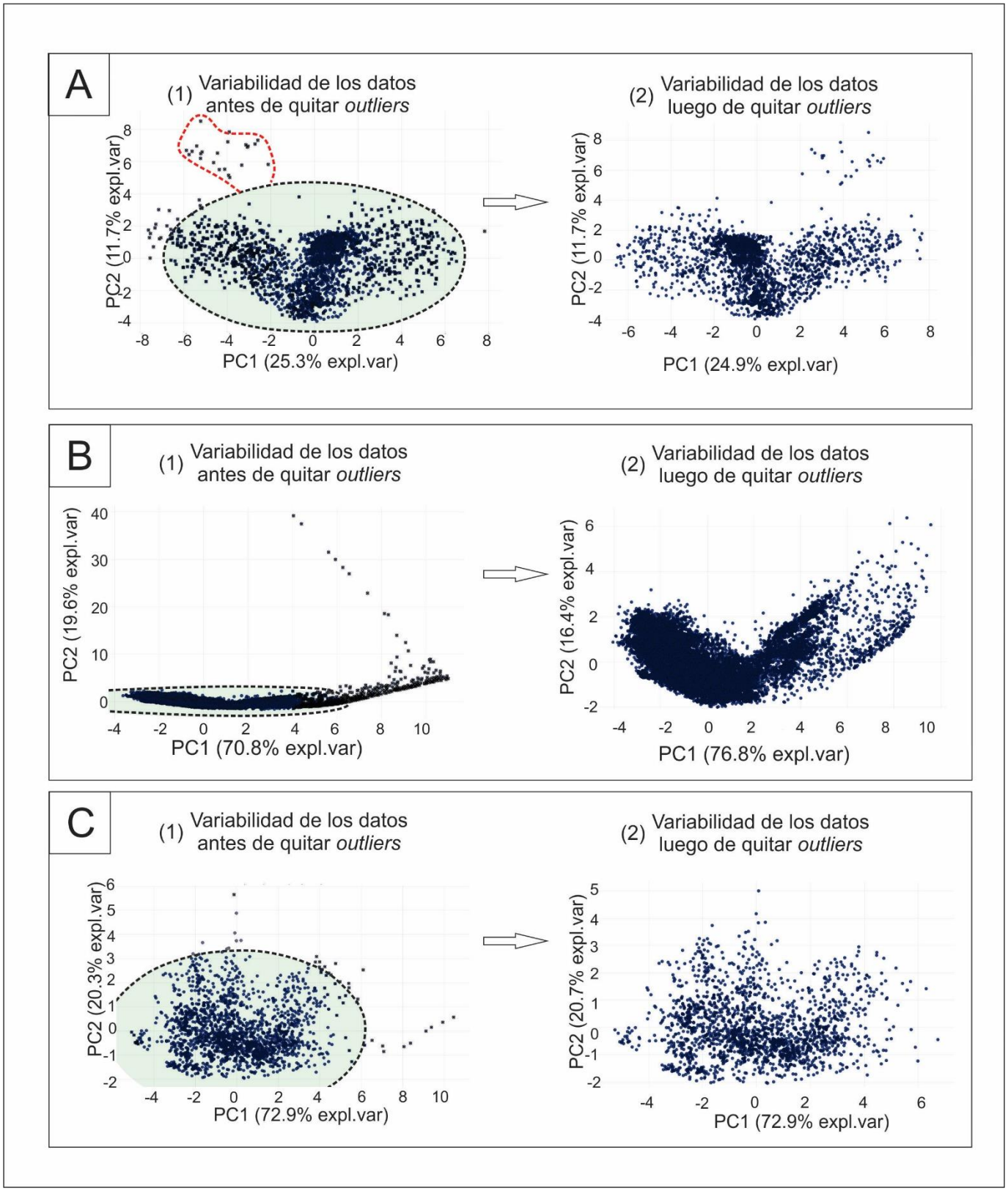


Figura 3.13: Gráficos de dispersión resultantes de los distintos PCA donde se muestran los valores anómalos para las pruebas combinadas de T2 de Hotellings y DmodX. (A1) Gráfico de dispersión para los datos químicos donde se puede observar los valores atípicos. (A2) Gráfico de dispersión para los datos químicos luego de la remoción de los valores atípicos. (B1) Gráfico de dispersión para los datos de UHR_i donde se puede observar los valores atípicos. (B2) Gráfico de dispersión para los datos de UHR_i luego de la remoción de los valores atípicos. (C1) Gráfico de dispersión para los datos de HR_i donde se puede observar los valores atípicos. (C2) Gráfico de dispersión para los datos de HR_i luego de la remoción de los valores atípicos.

3.2.6 Selección de características para reducir la cantidad de variables de los distintos conjuntos de datos.

La selección de variables se realizó para cada conjunto de datos por separado (textura de las imágenes, datos químicos y datos geomecánicos), mediante Análisis de Componentes Principales o PCA (por sus siglas en inglés).

PCA es una técnica estadística multivariada cuyo objetivo es transformar un conjunto de datos de alta dimensionalidad en uno de menor dimensión preservando la mayor cantidad de varianza posible y dando como resultado un conjunto de variables no correlacionadas y ortogonales entre sí, denominadas componentes principales (PC) (Marques et al., 2008, Mishra et al., 2017).Para llevar a cabo el PCA, es importante la estandarización de los datos ya que el análisis es sensible a la escala de las variables.

Dado que la finalidad de este análisis es la reducción de la dimensionalidad, pero conservando las variables originales, se decidió realizar una primera selección de variables a partir de la observación de las colinealidades y de ser necesario, una segunda selección a partir de un ranqueo de las mismas según la varianza que aportaban al conjunto de datos. Para definir este ranqueo, se tomaron las PC que contenían el 90% de la variabilidad, se realizó una suma del valor absoluto del aporte de cada elemento en cada PC, ponderada por la importancia de dicho PC, de modo tal que en este ranqueo se viera reflejado no solo el aporte de variabilidad de cada elemento, sino la importancia del componente principal.

El análisis de componentes principales para los datos químicos (Fig. 3.14) muestra que con 13 componentes principales se conserva el 93.48 % de la variabilidad de los datos (Fig. 3.14 A), el primer componente captura casi el 25 % de ese total, mientras que los restantes no superan el 15%. Al analizar el gráfico de *biplot* del modelo (Fig. 3.14 B), se observó que existían dos grupos de variables que poseían una evidente correlación. El primer grupo, compuesto por Fe y Rb (Fig.3.14 B), se analizó la importancia geológica de cada elemento, y se decidió conservar al Fe por ser un elemento mayoritario de gran importancia en la formación de minerales (Robinson, 2009; Misra, 2012). Mientras que el segundo grupo, está compuesto por Sr, Nd y Mg. En este grupo se observó que el Nd se encuentra exactamente a la mitad de la distancia angular que existe

entre el Sr y el Mg (Fig. 3.14 B), es por lo que se decidió eliminar al Nd, entendiendo que parte de la variabilidad también es explicada por el Mg y otra parte de la variabilidad es explicada por el Sr.

Las restantes variables, si bien se encontraban cerca, no se consideró que había superposición suficiente para optar por una u otra. Es por eso por lo que se realizó el ranqueo de las variables (Tabla 3.1). Se utilizó como umbral de corte a la cantidad de componentes principales que, según el modelo explicaban la variabilidad del conjunto de datos y que a su vez fueron los utilizados para realizar el ranqueo. Teniendo en cuenta entonces, el umbral de corte de 13 variables, se seleccionaron las variables S; Cr; Mn; Mg; Pb; P; Sr; Zn; Ca; K; Al; Ti; Fe como variables para el entrenamiento del modelo siguiendo criterios geoquímicos de ambientes supergénicos (Robinson, 2009; Misra, 2012).

El análisis de componentes principales para los datos de imágenes UHRi (Fig. 3.15) muestra que, con dos componentes principales se conserva el 98.7% de la variabilidad de los datos (Fig. 3.15 A), el primer componente captura casi el 77% de ese total, mientras que el segundo componente captura el 16%. Al analizar el gráfico de *biplot* del modelo (Fig. 3.15 B), se observó que el par de variables contraste y disimilaridad poseen una evidente correlación. Por ello, para decidir sobre cual variable conservar y cual eliminar, así como que variables aportaban más variabilidad al conjunto de datos, se procedió a realizar el ranqueo de las variables (Tabla 3.2). Al analizar este ranqueo, y teniendo en cuenta nuevamente la cantidad de componentes principales que, según el modelo explicaban la variabilidad del conjunto de datos y que a su vez fueron los utilizados para realizar el ranqueo, se seleccionaron las variables Entropía y Disimilaridad como las que más variabilidad aportaban al conjunto de datos, por lo que se decidió que estas finalmente fueran parte del modelado.

Finalmente, el análisis de componentes principales para los datos de imágenes HRi (Fig. 3.16) muestra que, según el PCA, también con 2 componentes principales se conserva el 97.5% de la variabilidad de los datos (Fig. 3.16 A), el primer componente captura el 73% de ese total, mientras que el segundo componente captura el 20.7%. Al analizar el gráfico de *biplot* del modelo (Fig. 3.16 B), se observó que el par de variables contraste y disimilaridad poseen una evidente correlación (más evidente que en los datos de ultra alta resolución).

Por lo tanto, se procedió de igual manera, realizando el ranqueo de las variables (Tabla 3.3). Al analizar este ranqueo, se observó que en este caso eran las variables Energía y Entropía las que más variabilidad aportaban al conjunto de datos, por lo que se decidió que estas finalmente fueran parte del modelado.

Tabla 3.1: Ranqueo de datos de composición quimica

Elemento	Sumatoria de aporte
S	0.183
Cr	0.178
Mn	0.176
Mg	0.176
Pb	0.174
P	0.173
Sr	0.172
Zn	0.169
Ca	0.166
K	0.164
Al	0.164
Ti	0.164
Fe	0.16
Si	0.159
Cu	0.159
Zr	0.153
Ni	0.149
Y	0.148
Ba	0.142

Tabla 3.2: Ranqueo de datos de imágenes de UHRi

Variable	Sumatoria de aporte
Entropía	0.39
Disimilaridad	0.37
Contraste	0.362
Energía	0.359
Homogeneidad	0.356
Correlación	0.329

Tabla 3.3: Ranqueo de datos de imágenes de HRi

Variable	Sumatoria de aporte
Entropía	0.387
Energía	0.371
Disimilaridad	0.37
Contraste	0.354
Homogeneidad	0.353
Correlación	0.238

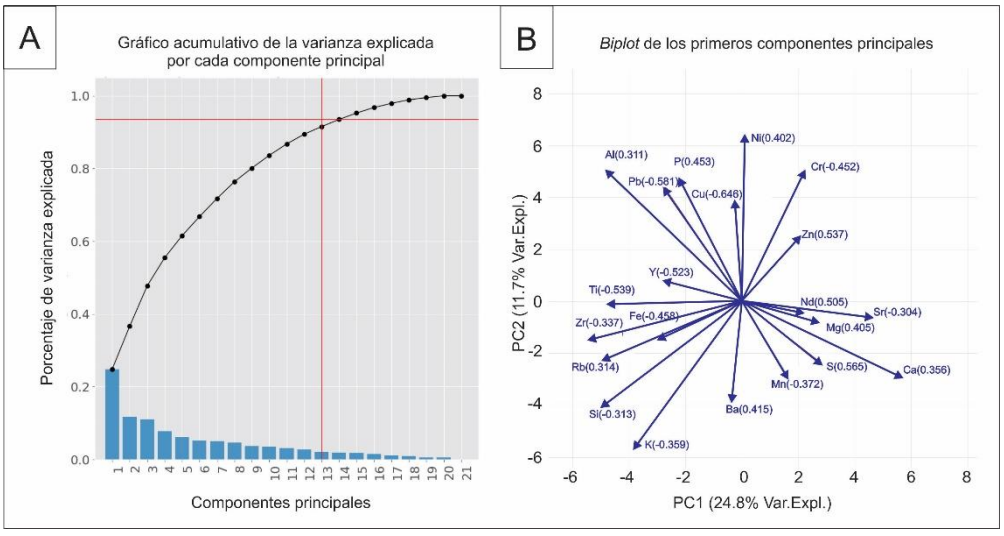


Figura 3.14: Gráficos producto del análisis de PCA de los datos químicos. (A)Gráfico acumulativo de la varianza explicada por cada PC. (B) *Biplot* de PC1 vs PC2 donde se observan los distintos comportamientos de las variables.

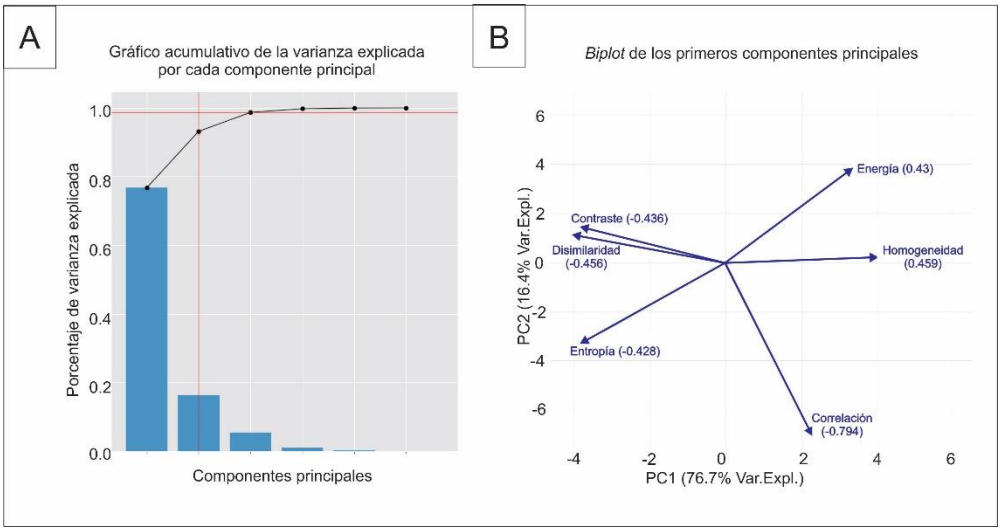


Figura 3.15: Gráficos producto del análisis de PCA de los datos de UHRi. (A)Gráfico acumulativo de la varianza explicada por cada PC. (B) *Biplot* de PC1 vs PC2 donde se observan los distintos comportamientos de las variables.

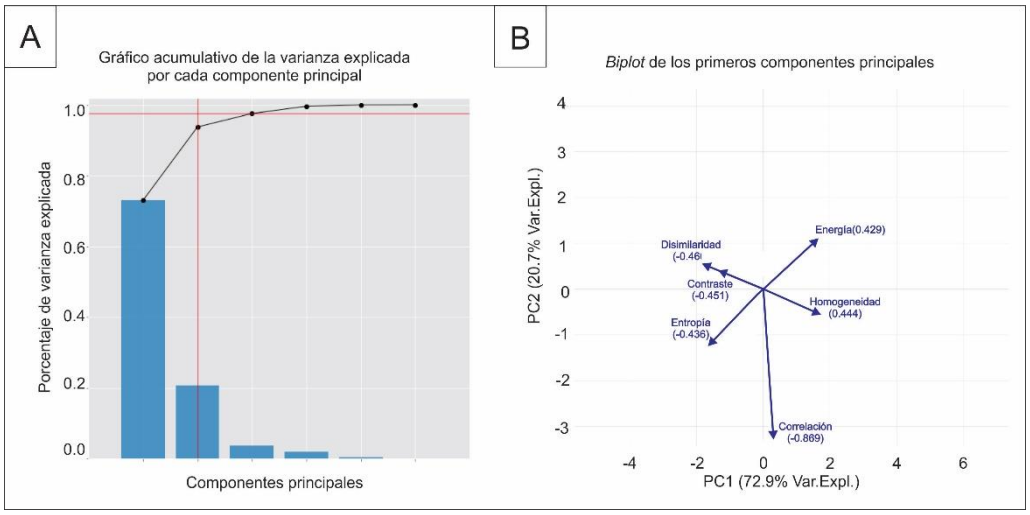


Figura 3.16: Gráficos producto del análisis de PCA de los datos de HRi. (A)Gráfico acumulativo de la varianza explicada por cada PC. (B) *Biplot* de PC1 vs PC2 donde se observan los distintos comportamientos de las variables.

3.3 Modelado de los datos

3.3.1 Selección de la variable Objetivo

Luego de realizar el preprocesamiento donde se ordenaron y limpiaron los distintos conjuntos de datos se procedieron a analizar las distintas variables objetivo, así como las abundancias de las diferentes categorías dentro de cada variable. El objetivo de este preprocesamiento fue establecer si en función de variables, el conjunto de datos contenía igual cantidad de ejemplos (conjunto de datos balanceado) o un número dispar de ejemplos (conjunto de datos desbalanceado). Y, en caso de encontrarnos ante un conjunto de datos desbalanceados, ejecutar una estrategia de balanceo de datos.

Es importante destacar la inclusión dentro de la etapa de modelado a la selección de variable objetivo y el balanceo de los datos. Si bien la selección de la variable objetivo es una definición que se realiza al comienzo del planeamiento del problema, y que depende directamente de los objetivos del mismo; veremos más adelante que la predicción directamente de las facies sedimentarias no es simple. Por ello, se definió que una buena forma de simplificar el problema de la complejidad de la variable objetivo original, es la simplificación del problema a partir de predecir las características que conforman a la facies sedimentaria por separado para luego combinar dichas predicciones. Esto hizo que la variable objetivo cambiara según que característica se quisiera predecir, generando la necesidad de establecerla luego del análisis de preprocesamiento de datos.

Dado que no importa cuál sea la variable objetivo a modelar, el preprocesamiento de los datos (minería, limpieza, exploración e ingeniería de datos) no varió, si no que se realizó una sola vez en el proceso. La etapa de modelado y validación se realizó las veces necesarias, dependiendo de las diferentes opciones de variables objetivo. De esta manera, se definió que la elección de la variable objetivo a modelar y el balanceo de los datos según dicha variable se realizaran luego del preprocesamiento, dentro de la etapa de modelado.

En la búsqueda de predecir las facies sedimentarias, se analizó cuantos ejemplos de cada facies sedimentaria encontrábamos en el conjunto de datos (Tabla 3.4). Como se observa, el conjunto de datos se encuentra muy desbalanceado con respecto a la cantidad de ejemplos de facies sedimentarias, encontrando un máximo de 228 ejemplares de la facies “limo bioturbado” (SSb) y un mínimo de 2 de ejemplos de las facies “arena muy fina con gradación normal” (Svng) y “arena mediana laminada” (Sml). Esto hace imposible cualquier tipo de balanceo y por consecuente el entrenamiento de modelos de aprendizaje automático.

Tabla 3.4: Recuento de ejemplares de cada Facies en el conjunto de datos

Recuento de muestras de facies sedimentarias							
Csl	3		Scb	55		Smt	191
Csm	3		Scm	66		Ssb	228
Dfm	36		Sct	85		Ssm	112
Fsm	61		Sfb	22		Ssr	5
Gfm	100		Sfh	13		Svdm	22
Gft	40		Sfl	18		Svh	31
Gmmm	28		Sflm	12		Svl	62
Gsm	60		Sfm	92		Svm	87
Htb	178		Sfp	3		Svmm	12
Htd	43		Sft	100		Svng	2
Htf	183		Sfm	6		Svp	58
Htw	65		Smb	30		Svt	69
Mcm	151		Sml	2		Svtm	8
Pct	200		Smm	73		Wsm	4

Si, por el contrario, observamos como variable objetivo a las características que definen a las facies sedimentarias, es decir la composición, la textura granulométrica y las estructuras sedimentarias, la cantidad en el número de muestras cambia considerablemente.

En el caso de la composición química, la variable objetivo se simplifica al tipo de roca teniendo estas 7 opciones (Ver Anexo 2); y 3 opciones en nuestro conjunto de datos (Carbonáticas, Silicoclásticas y Mixtas). Si realizamos un recuento de cuantas muestras hay por categorías, observaremos que la abundancia cambia considerablemente (Tabla 3.5).

Tabla 3.5: Recuento de ejemplares de cada composición en el conjunto de datos

Silicoclásticas	2065
Mixtas	82
Carbonáticas	472

El desbalanceo del conjunto de datos, visto desde un punto de vista composicional, corresponde al desbalanceo natural que poseen los tipos de rocas en la corteza terrestre (Folk, 1980; Reading, 1978; Reading, 2009). Sumado al desbalanceo producto del conjunto de datos de las coronas

disponibles para este trabajo de tesis doctoral. De igual manera, desde un punto de vista del análisis de datos, es más factible realizar un balance.

En la textura granulométrica, por su parte, originalmente tenemos 33 posibles variables objetivo (Ver Anexo 2); pero en nuestro conjunto de datos encontramos 14 de esas opciones (Arcilla, Limo, Vaque, Heterolíticas, Arena muy fina, Arena fina, Arena mediana, Arena gruesa, Arena Sabulítica, Arena conglomerádica, Ortoconglomerado fino, Ortoconglomerado mediano, Ortoconglomerado grueso, Paraconglomerado fino, Packstone, Floatstone y Mudstone). Estas texturas granulométricas, en realidad están diferenciadas por composiciones. Si bien eso es correcto, desde un punto de vista granulométrico se podrían establecer paralelismos con los clastos silicoclásticos que están avaladas por el comportamiento hidrodinámico de los bioclastos y/o clastos pumíceos (piroclastos) (Allen, 1984; Fisher & Schmincke, 1984; Kidwell & Holland,1991; Sparks et al., 1997). Tanto bioclastos como clastos pumíceos y silicoclásticos, pueden clasificarse según tamaño, forma y densidad, factores que controlan su transporte y depositación. Allen (1984) demostró que los bioclastos responden a fuerzas hidrodinámicas según su masa, forma (concavidad) y rugosidad, comportamiento comparable al de los clastos silicoclásticos bajo condiciones similares de flujo. Kidwell & Holland (1991) clasificaron rocas bioclásticas según su empaquetamiento y selección de tamaño, criterios que pueden integrarse con las clasificaciones granulométricas utilizadas para sedimentos silicoclásticos. De igual manera, los piroclastos, que incluyen fragmentos volcánicos como ceniza, lapilli y bombas volcánicas, también presentan respuestas hidrodinámicas relacionadas con su tamaño, morfología y densidad (Fisher & Schmincke, 1984), y sus procesos de transporte y depositación están condicionados por la energía del medio (Sparks et al., 1997). Teniendo en cuenta lo anterior, se realizó una tabla de equivalencia (Tabla 3.6) entre las texturas sedimentarias clásticas de las distintas composiciones sedimentarias, llevando todas al dominio de los tamaños granulométricos de las rocas de composición silicoclástica.

Otro punto para tener en cuenta es la escala de observación de este tipo de variable (Fig. 3.17). Como se mencionó anteriormente en el preprocesamiento de las imágenes, está variable se modelará a partir de imágenes de ultra alta

resolución (UHRi) cuya escala es un dato cada 4 milímetros aproximadamente (Fig. 3.17). Si retomamos el concepto de los tamaños de grano y la de la escala

Tabla 3.6: Tabla de equivalencias texturales propuesta entre texturas sedimentarias silicoclásticas y carbonáticas.

Textura Carbonática	Textura Equivalente
Mudstone	Arcilla
Wackestone	Vaque
Packstone	Vaque
Grainstone	Arenas
Floatstone	Paraconglomerado fino
Rudstone	Ortoconglomerado
Bafflestone	Bafflestone
Bindstone	Bindstone
Dolomita	Dolomita
Framestone	Framestone

Textura Piroclástica	Textura Equivalente
Chonita	Arcilla
Toba	Arenas
Lapilli-Toba no soldada	Ortoconglomerado
Lapilli-Toba soldada	Ignimbritas
Lapilli	Arena sabulítica
Aglomerado piroclástico	Ortoconglomerado
Brecha piroclástica	Ortoconglomerado

granulométrica de φ (Tabla 1.1) explicada en el apartado 1.3.2, podemos observar que tendríamos una escala de datos que en caso de granulometría psefíticas o mayores perderíamos resolución, y en el caso de las psamitas medianas a gruesas, nos encontraríamos ante una cantidad de datos poco representativos (Fig. 3.17). En la figura 3.17, de cada sección de análisis (rectángulo amarillo) se obtiene un dato promedio para cada variable de textura de la imagen y estos estarán directamente influenciados por la granulometría y disposición de los granos. Si la roca se encuentra compuesta por granulometrías desde arcilla (0.004 mm) a arenas medianas (0.5 mm; Fig. 3.17 A), el dato obtenido sobre la variabilidad del pixel respecto a sus vecinos estará influenciado por varios granos y sus respectivos contactos (límite de granos). En contraparte, en granulometrías superiores a arenas medianas (>0.5 mm; Fig. 3.17 B), el mismo dato obtenido estará influenciado por pocos granos, llegando en granulometrías muy gruesas a observarse un solo grano por sector.

El objetivo de la predicción está orientado a facilitar la descripción de las facies sedimentarias en yacimientos tipo *shale* y tipo *tight*; y estos yacimientos en conjunto cubren una amplia gama de granulometrías. Teniendo en cuenta que el conjunto de datos posee granulometrías variadas, se estableció arbitrariamente que el mejor espesor de análisis para modelar la textura granulométrica es de 5 cm, ya que de ese modo tendremos representación de

datos en las granulometrías gruesas. Para poder modelar entonces nuestra variable objetivo “textura granulométrica”, teniendo en cuenta el espesor de análisis, se tomó el conjunto de datos correspondiente y se realizaron segmentaciones cada 5 cm de profundidad (Fig. 3.18 A). Una vez delimitados

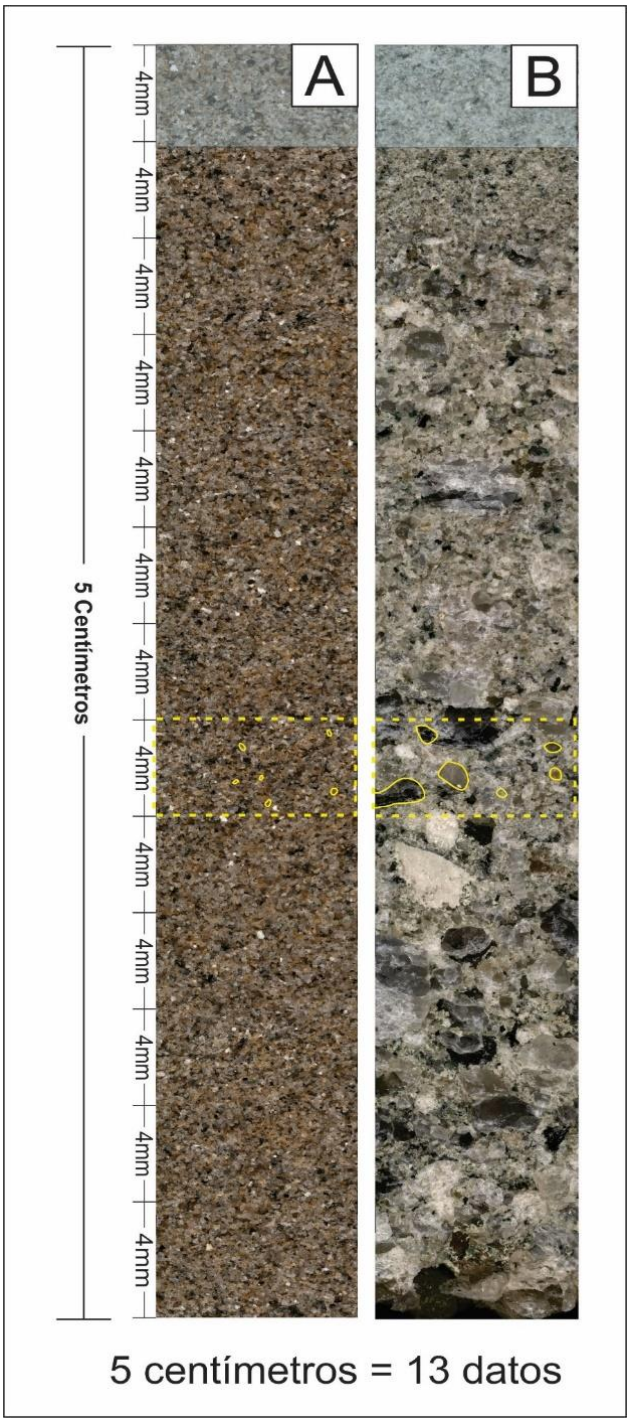


Figura 3.17: Esquema del recorte y análisis de los datos para establecer el espesor del análisis. (A) ejemplo sobre una psamita (B) ejemplo sobre una psefita.

estos 5 cm se tomaron las variables predictoras (columna de disimilaridad y entropía) y se vectorizaron sus valores para que estos quedaran de manera continua en una única fila (Fig. 3.18 B).

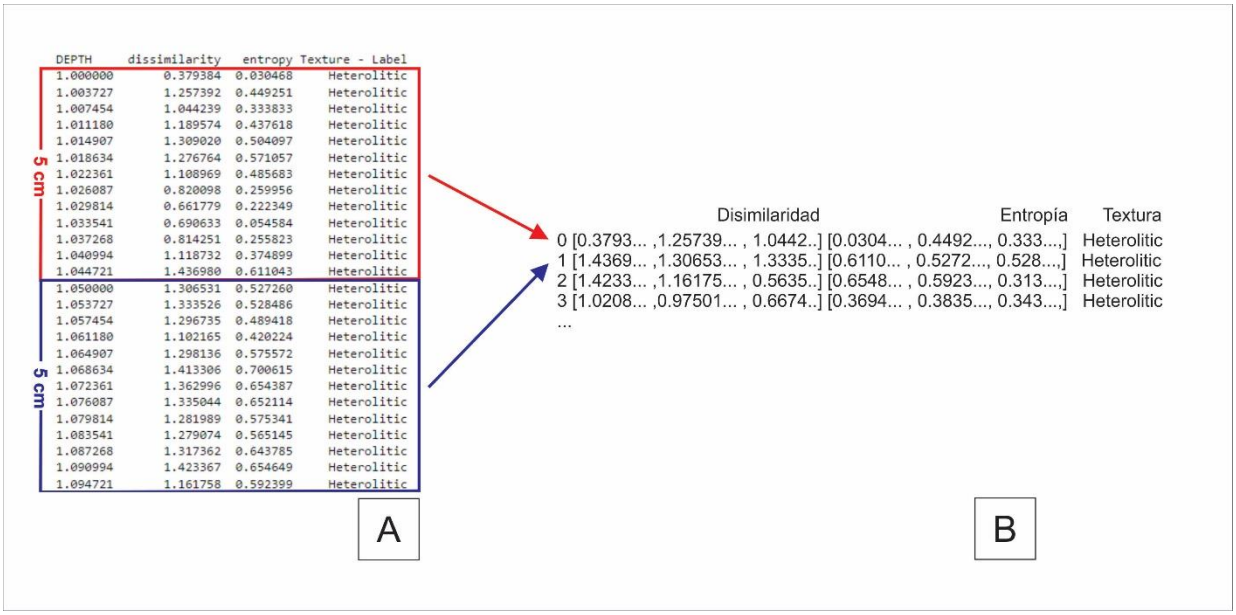


Figura 3.18: Segmentación de la matriz original de datos y posterior vectorización de cada segmento realizado para trabajar con 5 cm de espesor. (A) Matriz original que se segmenta cada 5 cm de profundidad; (B) Matriz resultante con 5 cm de datos por fila.

Luego de realizada está conversión por equivalencia y de establecido el espesor de análisis, realizamos un recuento de cuantas muestras hay por categorías, observaremos que la abundancia de la categoría Heterolítica es la mayor con 234 ejemplares, seguido por la categoría “Arena mediana” con 213 ejemplos (Tabla 3.17). Si bien el conjunto de datos se encuentra desbalanceado, el balanceo de datos es factible dado que todas tienen representatividad suficiente dentro del conjunto de datos para modelar sus patrones de entrenamiento y generar datos sintéticos.

Tabla 3.17: Recuento de ejemplares de cada Textura granulométrica en el conjunto de datos

Heterolítica	234
Vaque	120
Limo	199
Arena - Muy fina	166
Arena - Fina	156
Arena - Mediana	213
Arcilla	90
Arena - Gruesa	137
Ortoconglomerado - Mediano	14
Ortoconglomerado - Fino	81
Arena Sabulítica	24
Paraconglomerado - Fino	50

Para las estructuras sedimentarias, la variable objetivo en el código de facies tiene 24 opciones (Ver Anexo 2). De esas 24 opciones de estructuras sedimentarias, nuestro conjunto de datos presenta 9 (Flaser, Wavy, Bioturbación, Maciza, Estratificación planar, Estratificación entrecruzada tangencial, Laminación, Deformación sinsedimentaria, Estratificación horizontal). Dicha variable será modelada a partir de imágenes de alta resolución (UHRi) cuya escala es un dato cada 3.6 cm aproximadamente. Desde un punto de vista teórico (Collinson et al., 2006), las distintas estructuras sedimentarias poseen diferentes escalas (microescala, mesoescala, macroescala). Utilizar directamente un dato para modelar impediría observar estructuras de mesoescala mayores a 3 cm, como por ejemplo laminaciones cruzadas y flaser. Por eso se determinó que un espesor apropiado para no sobreestimar escalas pequeñas y, a su vez, tener una buena resolución en mesoescala, es de 10 cm. La figura 3.19 muestra dos ejemplos de estructuras. En la figura A se puede observar una estructura heterolítica con estratificación flaser, donde un dato cada 3.3 cm condensa varias capas de la estructura, mientras que en la figura B se observa una estratificación entrecruzada tangencial correspondiente a una

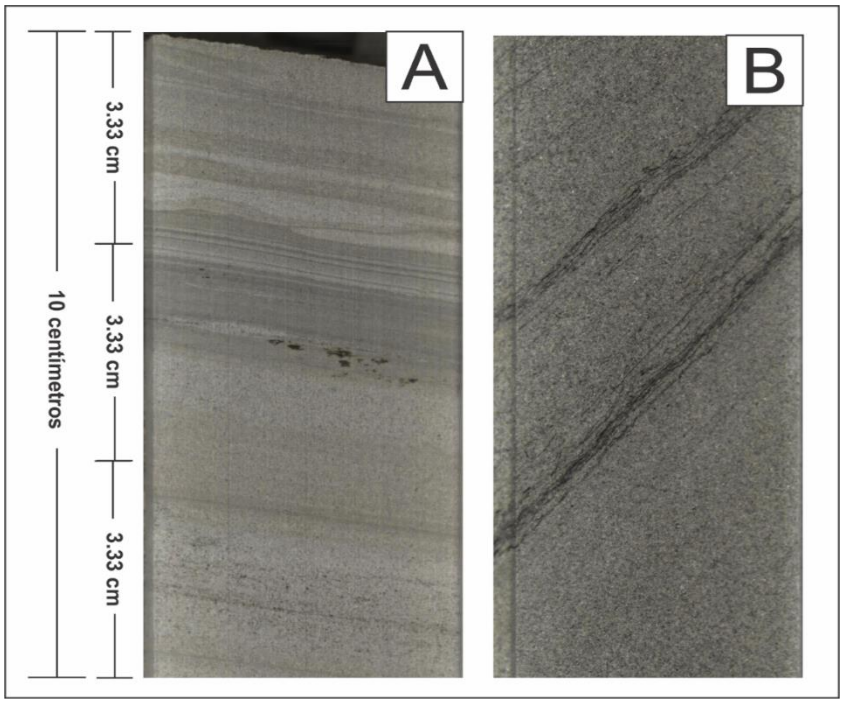


Figura 3.19: Esquema del recorte y análisis de los datos para establecer el espesor del análisis. (A) ejemplo sobre un depósito heterolítico con estratificación flaser (B) ejemplo sobre una estratificación entrecruzada tangencial.

megaóndula 2D, que es de mayor escala (mesoescala). En ambos casos, 3.3 cm de análisis condensan más la estructura interna del set tangencial.

Al igual que como se realizó con el conjunto de datos anterior, se realizaron segmentaciones, pero esta vez cada 10 cm de profundidad (Fig. 3.20 A). Una vez delimitados estos 10 cm se tomaron las variables predictoras (columna de Energía y Entropía) y se vectorizaron sus valores para que estos quedaran de manera continua en una única fila (Fig. 3.20 B).

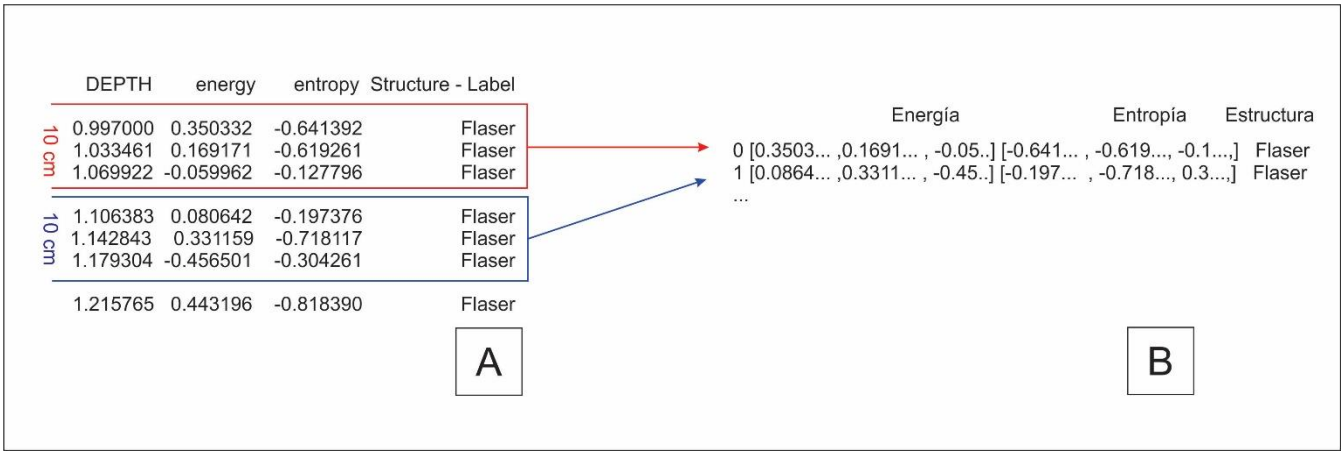


Figura 3.20: Segmentación de la matriz original de datos y posterior vectorización de cada segmento realizado para trabajar con 10 cm de espesor. (A) Matriz original que se segmenta cada 10 cm de profundidad; (B) Matriz resultante con 10 cm de datos por fila.

Si realizamos un recuento de cuantas muestras hay por categorías, observaremos que la abundancia varía considerablemente (Tabla 3.18). Categorías como “estratificación horizontal” o “wavy”, presentan frente al resto de categorías una representación mínima. Sin embargo, se observó que esa cantidad de ejemplos son suficientes para calcular las tendencias y replicarlas con datos sintéticos.

Tabla 3.18: Recuento de ejemplares de cada Estructura sedimentaria en el conjunto de datos

Flaser	26
Wavy	5
Bioturbación	123
Maciza	187
Estratificación planar	10
Estratificación entrecruzada tangencial	158
Laminación	17
Deformación sin sedimentaria	12
Estratificación horizontal	8

3.3.2 Balanceo de datos

El balanceo de datos es el proceso de ajustar la distribución de frecuencias de las clases de la variable objetivo dentro de un conjunto de datos y así evitar sesgos en los modelos. Los conjuntos de datos son desbalanceados cuando la variable objetivo posee diferente cantidad de ejemplos en las distintas clases. Esto lleva a que los modelos de ML generen resultados sesgados, favoreciendo a la clase más frecuente y afectando su capacidad de generalización (Bishop, 2006; Müller et al., 2016).

Si bien existen varias librerías de Python para balancear los datos de manera automática, se optó por realizar un balance manual de los mismos siguiendo criterios geológicos y estadísticos. Para ello, se toman todos los ejemplares de cada categoría de la variable objetivo, se le calcula a cada subgrupo los parámetros estadísticos de tendencia central (media), de dispersión (desvío), el rango de los datos y, en caso de que posean profundidades consecutivas, la tendencia creciente o decreciente del intervalo. Teniendo estos parámetros estadísticos, se crean muestras al azar siguiendo la población de las muestras originales. Esto, por un lado, evita el balanceo de las muestras a partir de sobre muestrear las muestras minoritarias con copias de sí mismas y el submuestreo de las muestras mayoritarias, evitando la perdida de información. De esta manera, se puede generar la cantidad de muestras que se desee sin alterar la naturaleza estadística y geológica de los datos.

3.3.3 Separación de los datos

Con el fin de aplicar un modelo de ML a un conjunto de datos se suele particionar la muestra total en tres conjuntos (1) entrenamiento, (2) prueba, y, por último, (3) validación.

El conjunto de datos de entrenamiento se utiliza para que los modelos de ML puedan captar los rasgos y patrones característicos del conjunto de datos. Los datos de prueba se utilizan para aplicar los modelos ya entrenados y poder resaltar o no los patrones encontrados en el conjunto de entrenamiento (Bishop, 2006; Géron, 2022). Adicionalmente, durante esta etapa de prueba también se realizan una serie de ajustes a los parámetros de los modelos aplicados con el

fin de optimizar y resaltar los patrones encontrados y luego poder generalizar los resultados de los mismos en el conjunto de datos de características similares. En este sentido, se trabaja con dos conceptos, los denominados sobre- o sub-ajuste. Básicamente, el primer término refiere al hecho que los modelos entienden perfectamente los patrones de sólo un conjunto de datos, pero son incapaces de generalizar ese entendimiento. Por otro lado, el subajuste hace referencia al poco entendimiento que un dado modelo toma de un conjunto de datos, por lo tanto, no se podrán distinguir las características más significativas de esos datos (Bishop, 2006; Hastie et.al., 2009; Goodfellow et al., 2016; Müller et al., 2016; Géron, 2022).

Finalmente, el conjunto de validación se utiliza para evaluar de manera objetiva el rendimiento del modelo final, proporcionando una medida de su capacidad para generalizar a datos no evaluados. Esta división asegura que el modelo no solo funcione bien con los datos en los que fue entrenado, sino que también sea robusto y efectivo en situaciones del mundo real (Bishop, 2006; Hastie et.al., 2009; Goodfellow et al., 2016; Müller et al., 2016; Géron, 2022).

Para este trabajo de tesis la separación de los tres conjuntos se realizó de manera automática y aleatoria. En la figura 3.21, se esquematiza el *workflow* de división de grupos realizado, que es una especificación del *workflow* general de trabajo de la figura 3.2. Para ello, se recurrió a la librería *Sckit Learn de Python*, la cual posee una función automática (*train_test_split*) que divide matrices o conjuntos de datos en subconjuntos aleatorios teniendo en consideración una proporción establecida para cada subconjunto. También tiene en cuenta que todas las etiquetas tengan ejemplos en ambos conjuntos de datos asegurando la representatividad de todas opciones posibles de la variable objetivo en todo el proceso de modelado.

Dicha función de separación tiene la limitante que solamente subdivide en dos grupos, no siendo posible la separación automática de los tres grupos al mismo tiempo. Para solventar este problema, se recurrió a hacer la separación de los datos en dos veces consecutivas (Fig. 3.21). Ambas separaciones se realizaron con una proporción 80-20% y respetando, como se mencionó anteriormente la proporcionalidad de las variables a predecir. De esta manera, como resultado de las separaciones se consiguió un conjunto de datos de

entrenamiento de 53.8 metros, uno de testeo de 16.8 metros y finalmente uno de validación de 13.4 metros.

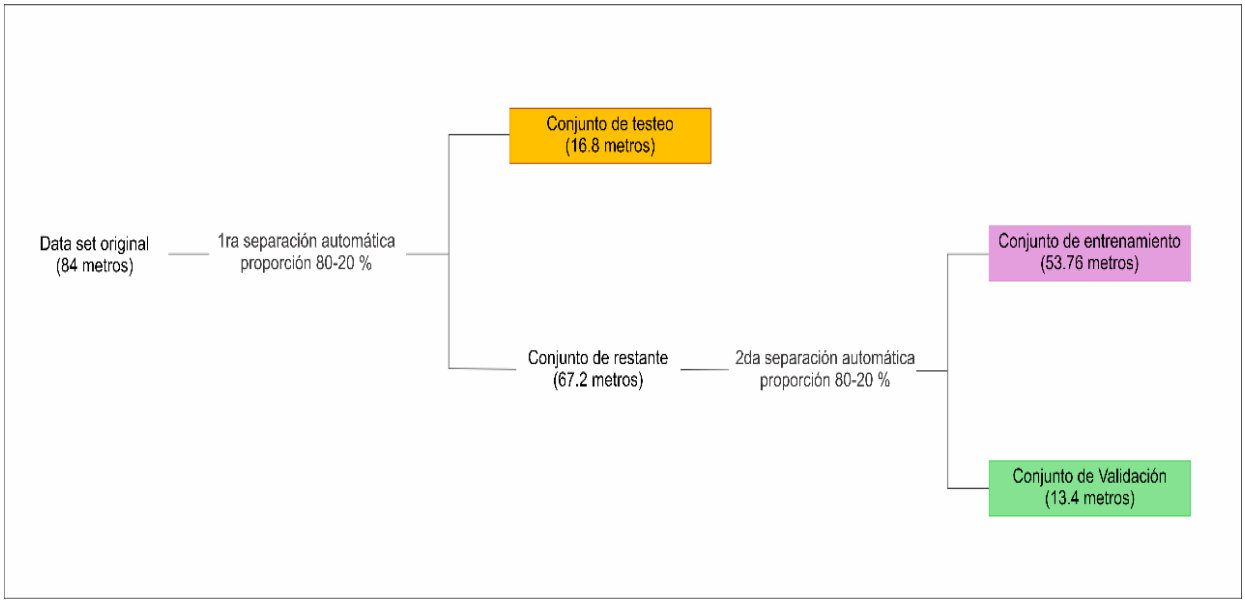


Figura 3.21: Esquema de la separación de los datos originales para obtener 3 subconjuntos

3.3.4 Entrenamiento de los modelos

En lo que respecta al entrenamiento de los diferentes modelos de los datos, se utilizó en su mayoría la librería *scikit-learn,* con excepción del modelo de XGBM que se utilizó la librería *xgboos*. Los modelos utilizados en esta tesis fueron:

- SVM

- Regresión logística

- Modelos basados en árboles (Gradient Boosting Machine (GBM), Random Forest (RF), Extreme GBM (XGBM))

Entrenar un modelo de ML a partir de este tipo de librerías proporciona la ventaja de la simplicidad dado que no se debe codificar por sí mismo el algoritmo matemático/estadístico del método. Los parámetros de un modelo son los que dan forma a la función de mapeo. Los valores de estos parámetros se ajustan según los datos de entrenamiento, inicialmente de manera manual y luego mediante técnicas de optimización. Existen parámetros comunes a todos los modelos (e.g. tolerancia, pesos de determinadas clases, límite de iteraciones,

cantidad de ejemplos por clases, entre otros) y existen también parámetros específicos para cada método.

Para los modelos SVM, la librería *scikit-learn* proporciona un parámetro en el que se debe establecer la función *kernel* con la que se trasformarán los datos. Algunos *kernels* son de manera predeterminada, o bien, dan la opción de configurar manualmente uno. En el marco de esta tesis y, considerando la complejidad de la variable objetivo a predecir, se decidió utilizar el *kernel* no lineal, "Función de base radial" o "RBF", cuya transformación es:

$$\exp(-\gamma \, \|\chi - \chi'\|^2), \tag{17}$$

donde γ debe ser menor que cero y define cuánta influencia tiene un solo ejemplo de entrenamiento. Cuanto mayor sea γ, más cerca deben estar los otros ejemplos para verse afectados. Otro parámetro a tener en cuenta para entrenar un modelo SVM es el parámetro C, común a todos los *kernels*. Un valor bajo de C hace que la superficie de decisión sea uniforme, mientras que un valor alto de C apunta a clasificar todos los ejemplos de entrenamiento correctamente.

En el caso de Regresión logística, la librería ofrece una función para entrenar modelos de Regresión Logística binarios y multiclase. En caso se ser de esta última opción, no es necesario utilizar una función especial, sino que solo se le debe indicar a partir del parámetro "*multi_class*" el tipo de ajuste que se quiere realizar. Entre las opciones dadas existe la opción "uno contra el resto" (u ovr por sus siglas en inglés), en el que el modelo de regresión logística ajusta un clasificador por clase. Para cada clasificador, la clase se ajusta a todas las demás clases. Se eligió este tipo de multiclase dado que es el que mayor interpretabilidad presenta dado que cada clase está representada por un solo clasificador.

Los algoritmos basados en arboles de decisión (RF; GBM; XGBM), por su parte, tienen parámetros comunes a este tipo de modelos que controlan la estructura de los árboles. Entre ellos se destacan *"n_estimator"* o número de árboles que se entrenarán en el modelo*, "max_depth"* o profundidad máxima de cada árbol y "*criterion*" o métrica que se utilizará para evaluar la calidad de las divisiones que se realizan en los datos. En este trabajo, el número de árboles que se entrenarán por modelo, así como la profundidad máxima, fueron

parámetros que utilizaron un valor inicial, los cuales luego fueron optimizados y como criterio de evaluación de la calidad se utilizó en todos los casos el método de impureza de “Gini” explicado en el apartado 1.4.2.

3.3.5 Comparación entre los modelos para la selección

Una vez entrenados los distintos modelos, y para evaluar el rendimiento de los mismos de una forma igualitaria, se utilizaron cuatro métricas. La más conocida y utilizada es la exactitud (*accuracy*), que es la proporción de predicciones correctas sobre el total de predicciones realizadas. Sin embargo, la exactitud no siempre es una métrica adecuada, especialmente en situaciones donde las clases de la variable objetivo están desbalanceadas. En estos casos, métricas como la precisión (*precision*), el *recall* y la medida F1 son más informativas (Sokolova et al., 2009; Géron, 2022). La precisión evalúa la proporción de verdaderos positivos entre todas las predicciones positivas, mientras que el *recall* mide la proporción de verdaderos positivos entre todos los casos positivos reales. La medida F1, que es la media armónica de la precisión y el *recall*, nos da una idea del balance entre ambas métricas y es especialmente útil cuando se requiere un equilibrio entre precisión y exactitud (Sokolova et al., 2009; Géron, 2022). Matemáticamente, las métricas están definidas como:

$$Accuracy = \frac{(Tp + Tn)}{(Tp + Tn + Fp + Fn)}; \tag{18}$$

$$Precisión = \frac{Tp}{(Tp + Fp)}; \tag{19}$$

$$Recall = \frac{Tp}{(Tp + Fn)}; \tag{20}$$

$$F1 = \frac{2 * Presición * Recall}{(Presición + Recall)}; \tag{21}$$

donde Tp es el total de predicciones positivas; Tn es el total de predicciones negativas; Fp es el número de falsos positivos y Fn es el número de falsos negativos.

Como se mencionó antes, se utilizaron estas métricas para comparar el rendimiento de los distintos modelos y seleccionar el que mejor rendimiento demostraba. Para ello, como se verá en el capítulo de Resultados, se generó un cuadro a modo comparativo. Una vez definido el tipo de modelo que mejor predecía las variables objetivo, este se somete a un proceso de optimización de los hiperparámetros, proceso que será descripto en el siguiente apartado.

3.3.6 Optimización del modelo seleccionado

Una vez seleccionado el modelo que mejor rendimiento tiene, debemos optimizar sus hiperparámetros. Los hiperparámetros de un modelo son, los parámetros que deben ser configurados antes del entrenamiento. Cada modelo posee una cantidad definida de hiperparámetros y a su vez estos poseen diferentes opciones que hacen a la configuración de las distintas estrategias de entrenamiento. Los procesos de optimización buscan identificar todas las posibles configuraciones y definir cuál de todas maximiza el rendimiento del modelo en datos no observados. Realizar este ajuste garantiza que el modelo alcance un equilibrio entre el rendimiento y la generalización, evitando el sobre *overfitting* (Hastie et al., 2009; Dangeti, 2017; Géron, 2022). Existen distintas formas de optimizar los hiperparámetros, las más utilizadas son: (1) *Grid Search*, (2) *Random Search*, (3) optimización bayesiana, (4) los basados en poblaciones, y (5) los basados en los gradientes (Kohavi, 1995).

Para este trabajo se utilizó *Grid Search* como forma de optimización, porque, aunque computacionalmente es de alto costo, es exhaustivo y explora todas las combinaciones posibles de hiperparámetros asegurando la mejor opción. La librería *sckit learn* posee una función específica para dicha optimización (GridSearchCV), esta realiza la optimización combinándola con validación cruzada o *cross-validation*. La validación cruzada es una técnica estadística utilizada para evaluar la capacidad predictiva y la generalización de los modelos. Se basa en el principio de dividir el conjunto de datos disponible en múltiples subconjuntos de entrenamiento y prueba e itera varias veces el proceso de entrenamiento del modelo asegurándose que todos los subconjuntos se usen tanto en la etapa de entrenamiento como en la etapa de testeo al menos en una

oportunidad. El rendimiento final del modelo será el rendimiento promedio obtenido a lo largo de todas las iteraciones (Kohavi, 1995). La combinación de *Grid Search* con validación cruzada, aumenta el grado de robustez a la evaluación de los mejores hiperparámetros de los modelos (Hastie et. al, 2009; Géron, 2022).

3.3.7 Análisis del modelo

La interpretabilidad de los modelos se ha convertido en un factor clave en su uso (Van den Broeck et al., 2022). Los modelos lineales, como la regresión lineal, tienen una interpretabilidad inherente debido a la simplicidad de sus relaciones matemáticas. Mientras que, los modelos de caja negra como las redes neuronales o los árboles de decisión ensamblados pueden ser difíciles de interpretar para los investigadores y mucho menos para los geocientistas que no pertenecen a la rama del análisis de datos (Molnar, 2020). Para evaluar el rendimiento del modelo seleccionado, no solo se utilizaron las métricas descriptas oportunamente en el apartado 3.3.5. Sino que también, se utilizaron matrices de confusión y el paquete *Shapley additive explanations* de Python.

Las matrices de confusión (Ting, 2011) son tablas de doble entrada gráficas (Fig. 3.22) que comparan las predicciones realizadas por el modelo con los valores reales de dichas muestras y ofrecen un resumen visual y cuantitativo de como el modelo clasificó correcta o incorrectamente al conjunto de datos.

El paquete *Shapley additive explanations* o SHAP por sus siglas en inglés, es un paquete de Python que se utiliza en la interpretación de modelos de aprendizaje automático (Fig. 3.23) (Shapley, 1953). SHAP proporciona una forma cuantitativa de descomponer la predicción de un modelo en contribuciones

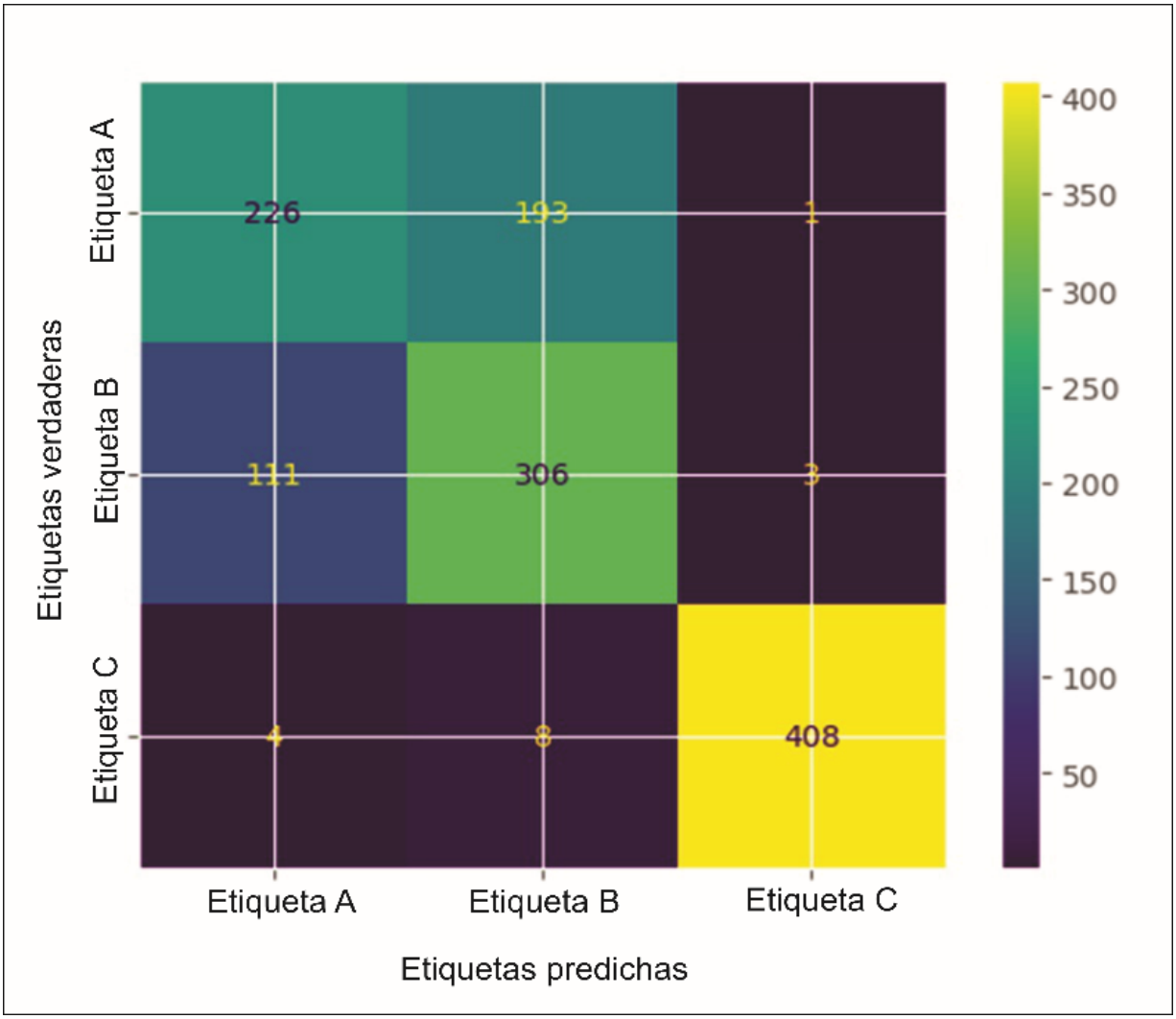


Figura 3.22: Matriz de confusión genérica donde se muestran las predicciones sobre el eje de las x y las etiquetas reales sobre el eje de las y.

individuales de cada variable predictora (Devlin et al., 2019; Madsen et al., 2021; Mosca et al., 2021; Mosca et al., 2022). El paquete SHAP tiene la capacidad para adaptarse a diferentes tipos de modelos, puede ser aplicado tanto a modelos lineales como a modelos no lineales (Mosca et al., 2022).

La teoría que subyace en SHAP está basada en los valores de Shapley, desarrollados por Lloyd Shapley en 1953 en el campo de la teoría de juegos. SHAP puede utilizarse para entender cómo cada variable afecta la predicción de un modelo en términos absolutos y relativos (Shapley, 1953). El cálculo de los valores de Shapley requiere evaluar todas las combinaciones posibles de variables, lo que puede volverse prohibitivo en términos de tiempo y recursos computacionales para modelos grandes o altamente complejos (Van den Broeck

et al., 2022). Si el modelo base no está bien ajustado, las interpretaciones proporcionadas por SHAP pueden ser engañosas (Molnar, 2020; Van den Broeck et al., 2022).

La figura 3.23 muestra un ejemplo genérico de los gráficos SHAP utilizados para analizar los modelos. Este tipo de gráfico en particular es denominado *"summary plot"*, combina la información de importancia global de las variables con la distribución de sus impactos en el modelo, facilitando la interpretación al mostrar cómo influyen las variables predictoras en la salida del modelo. En el eje y se encuentran las variables predictoras ordenadas de arriba hacia abajo según su importancia en el modelo, mientras que en el eje x se representa los valores SHAP que cuantifican la contribución de cada valor específico de una variable a la predicción. Los valores positivos de SHAP indican un aumento en la predicción del modelo, mientras que los valores negativos reflejan una disminución (Lundberg et al., 2020). Cada punto en el gráfico representa un valor de observación para una variable específica. El color indica el valor de la característica, los puntos azules representan valores bajos de la variable, mientras que los rojos indican valores altos. Esta codificación de color permite visualizar cómo los diferentes valores de la variable afectan la predicción. Por ejemplo, en la Variable predictora A, los valores altos (rojos) tienden a asociarse con valores SHAP positivos, lo que sugiere que un aumento en esta variable incrementa la predicción del modelo. En contraste, los valores bajos (azules) están relacionados con impactos negativos, disminuyendo la predicción. Este patrón indica que la variable tiene una relación directa con la salida del modelo (Molnar, 2022).

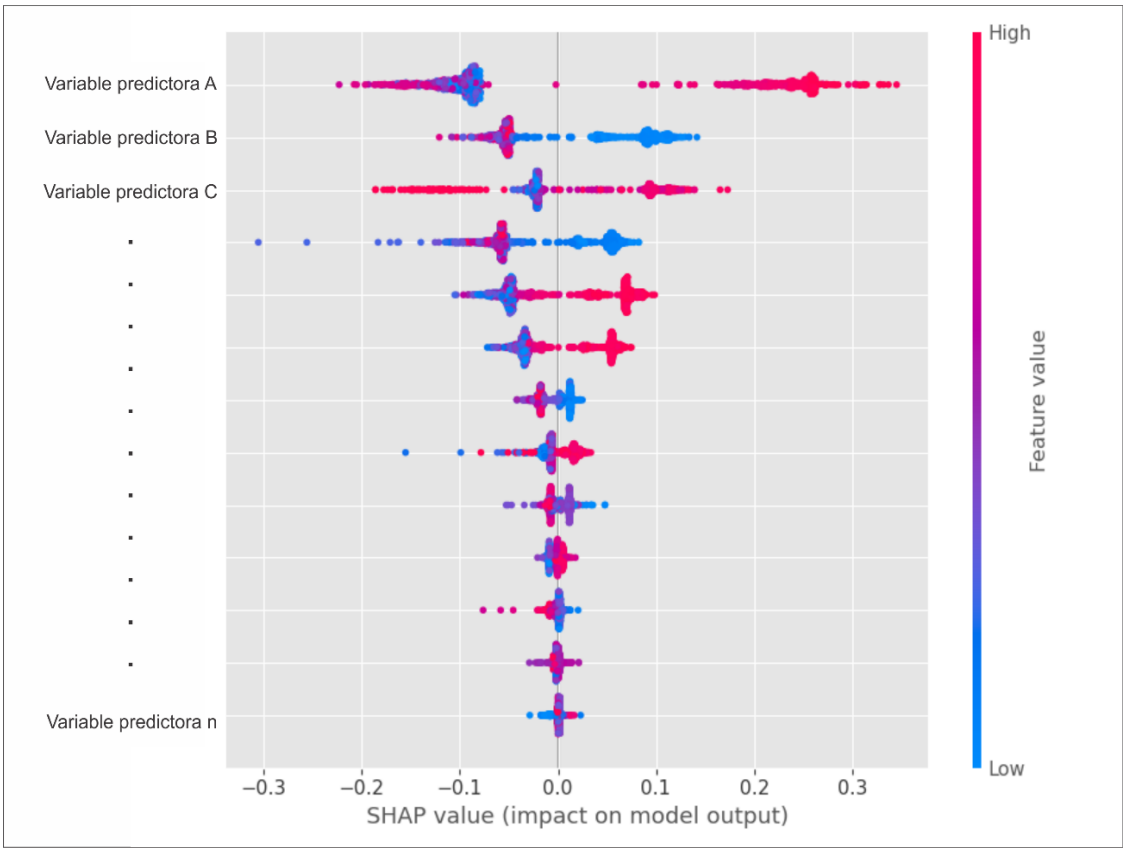


Figura 3.23: Gráfico SHAP para una variable genérica.

4

Resultados

4.1 Resultado sedimentológico y código de facies

Durante el trabajo de esta tesis doctoral, se describió de forma detallada un total de 84 metros de testigos coronas de forma digital siguiendo la metodología explicitada en el 3.1.1 “Generación del conjunto de datos”. Luego se procedió a establecer las facies sedimentarias de dichas descripciones, siguiendo el código de facies explicitado en el apartado 3.2.2 “Adecuación del código de facies a las nuevas tecnologías”. Como resultado, se obtuvieron los 5 mosaicos de los 84 metros de testigos coronas graficados en el Anexo I, los cuales están acompañados de una tabla resumen de las facies sedimentarias establecidas (Anexo I).

4.2 Resultados de algoritmos no supervisados

Se realizó un análisis de agrupamiento o Clustering a partir del algoritmo *k-means cluster*, con el fin de identificar patrones de agrupamiento entre las distintas muestras. Inicialmente se realizó el gráfico de la evolución de la varianza intra - *cluster* o, simplemente el gráfico del codo, para determinar el

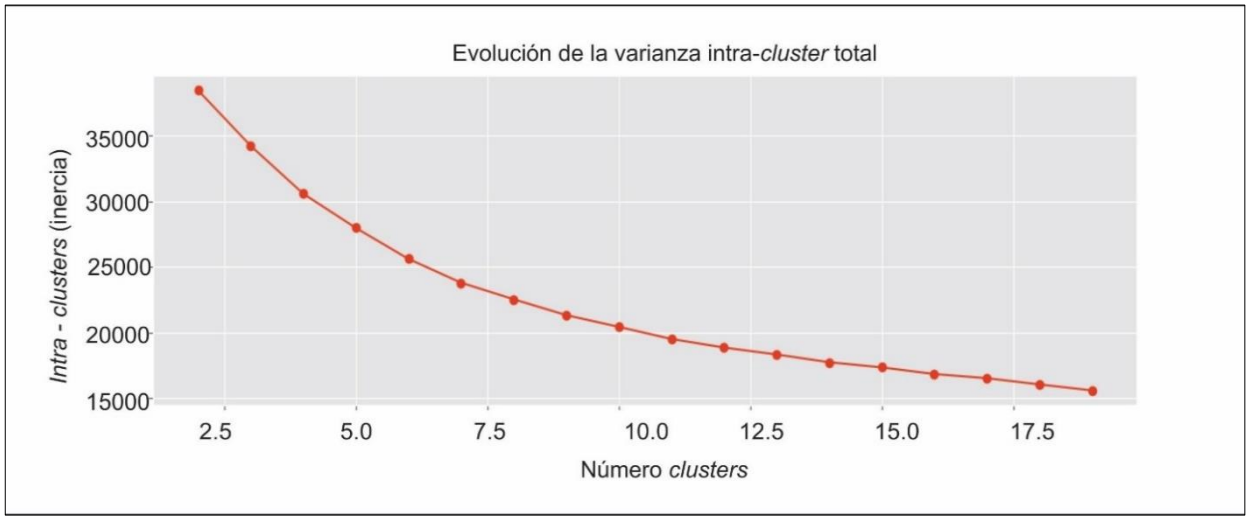


Figura 4.1: Gráfico de varianza intra - *cluster* para los diferentes números de grupos con todas las variables.

número optimo de grupos (Fig. 4.1). No se ve claramente un quiebre en la evolución de la varianza, pero, luego de analizar los cambios relativos entre los grupos consecutivos, se calculó que luego del k=6, los cambios son menores al 5% (Fig. 4.1). Cuando se realizó el *clustering* con los 6 grupos propiamente dichos, se observo en el gráfico 3D que si bien hay en la generalidad una buena diferenciacion de los grupos (Fig. 4.2), estos se encuentran contiguos unos de otros y algunos datos se superponen.

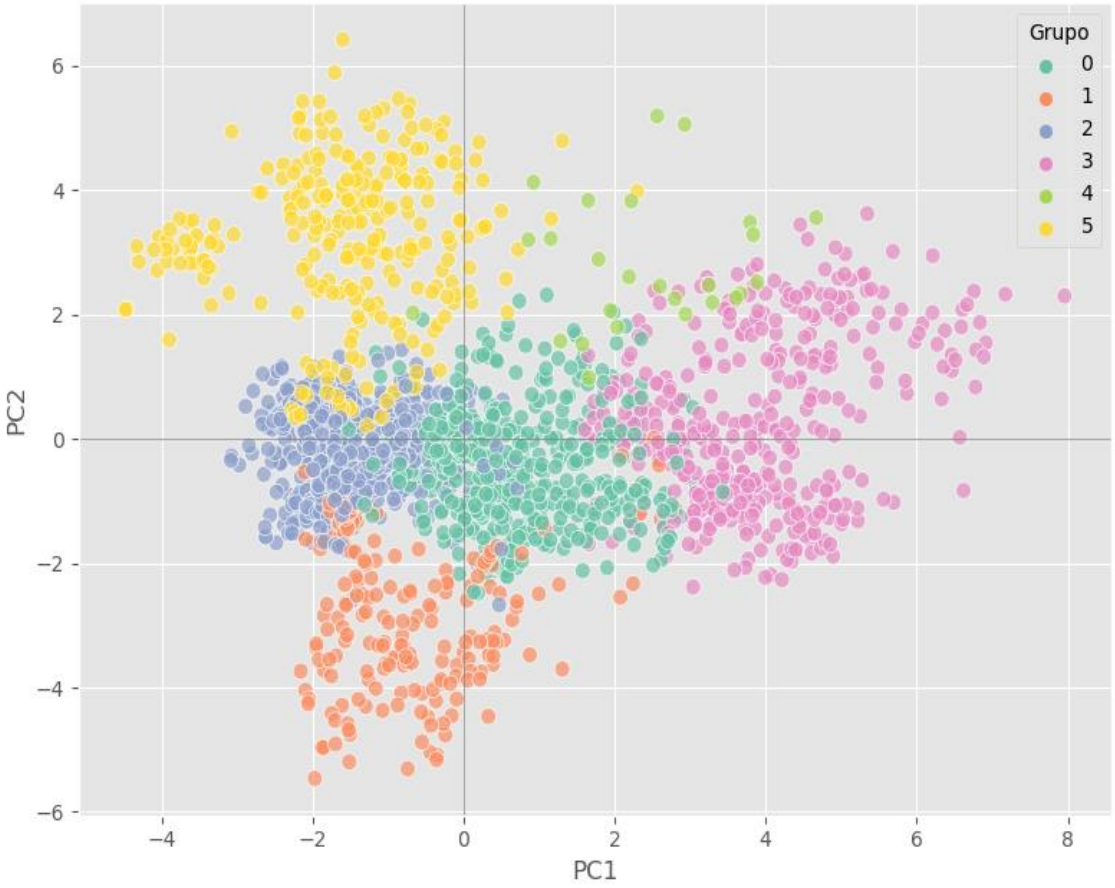


Figura 4.2: Gráfico 3D de los 6 grupos realizados por el *K-means.*

Al ser este método de agrupamiento realizado con datos que en realidad están etiquetados, esto nos permitió realizar una comparativa entre los grupos armados por el algoritmo de agrupamiento y las etiquetas reales (Tabla 4.1). Se pudo observar que ninguno de estos grupos tienen criterio geológico. Desde la vision genética de las facies sedimentrarias y de las caracteristicas

composicionales, texturales y estructurales que componen a las facies sedimentarias estan mezcladas en forma aleatoria (Tabla 4.1).

Tabla 4.1: Facies agrupadas en 6 grupos por *k-means,* En verde se observan las muestras cuya etiquetas reales son silicoclasticas; en rojo las muestras cuya etiquetas reales son carbonáticas y en azul las muestras cuya etiquetas reales son mixtas.

Grupo 0	C_sm; S_sb; S_sm; H_tw; H_td; H_tf; S_vp; S_fp; S_fm; S_vm; $S_v$ng; S_ml; S_fl; S_fh; S_vl; S_vt; S_ft; S_mm; S_mt; S_mb; S_cm; S_ct; S_cb; G_fm; G_ft; P_ct; G_sm; ; M_cm; F_sm; $S_v$mm; $S_v$dm; $S_f$lm; $S_v$tm;$G_m$mm.
Grupo 1	H_tb; S_vt; S_fb; S_ft; S_mt; S_mm; S_ct; S_cb; S_cm; M_cm; G_sm.
Grupo 2	C_sm; C_sl; S_sm; Ssr; S_sb; H_tf; H_tw; H_tb; H_td; W_sm; S_vm; S_vh; S_vp; $S_v$ng; S_vt; S_vl; S_fm; S_fb; S_ft; S_fh; S_fl; S_mm; S_mt; S_mb; S_ct; S_cb; S_cm; G_ft; P_ct; $S_v$dm.
Grupo 3	S_mm; G_ft; S_vm; S_fm; S_cm; $S_f$lm; $S_v$mm; $G_m$mm; $S_v$dm; $S_v$tm P_ct; G_sm; M_cm; F_sm;
Grupo 4	S_vm
Grupo 5	S_ct; S_ft; S_cm; S_mt; S_fb; S_cb; S_pm; S_fm; S_fl ; D_fm; G_fm; G_ft; $G_m$mm

4.3 Resultados de algoritmos supervisados

4.3.1 Modelo Composición química

El modelo de predicción de composición química tiene una resolución de 1 cm de espesor, respetando la distribución espacial original de los datos. La cantidad de datos (N) entre datos originales y sintéticos utilizada para dicho entrenamiento es de 2048 para cada categoría.

Este valor de N responde al mínimo número de muestras sintéticas necesarias para balancear el conjunto de datos debido al alto desequilibrio en cantidad de datos por elemento químico. Este valor N responde a la máxima cantidad de datos medidos de una cierta categoría, tal que todas las categorías con menor cantidad de datos puedan nivelarse a este N, con datos sintéticos complementarios. Se seleccionó luego de evaluar la exactitud de los modelos en distintas pruebas realizadas con distintas cantidades de datos totales, es decir, datos sintéticos + *hard data* (Fig. 4.3). Como se puede observar en la figura 4.3, la exactitud disminuye a razón del 2% a medida que aumenta la cantidad de datos por categoría. Si bien esta disminución, no es importante en términos numéricos, teniendo en cuenta que las otras pruebas contenían el costo computacional adicional de la generación de un mayor número de datos sintéticos, se decidió continuar con el modelo cuyo valor de N es 2048, el cual contiene la menor proporción de datos sintéticos/ datos reales.

Adicionalmente, se puede observar que todos los modelos poseen un rendimiento por encima del 65 % a la hora de clasificar las diferentes

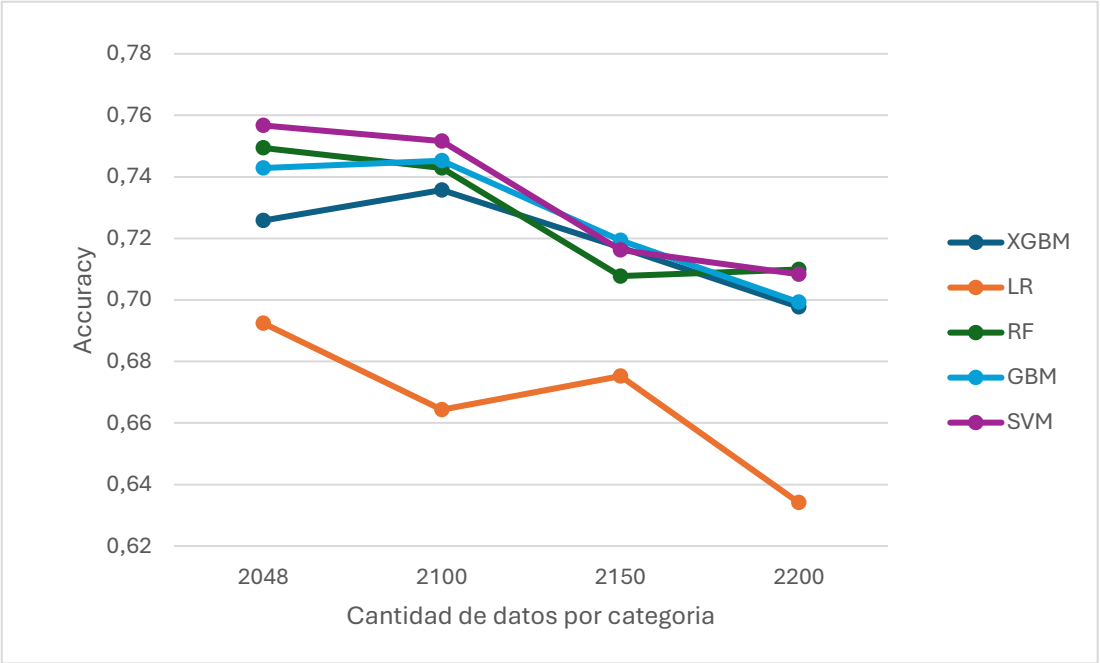


Figura 4.3: Exactitud de los distintos modelos de composición química entrenados en función de la cantidad de muestras por categorías.

composiciones de las rocas. Si bien el de mayor valor de exactitud fue el *Support Machine Vectors* (SVM), el *Random Forest* (RF) obtuvo solo un 0.01 % menos (Fig. 4.3). Al intentar optimizar el modelo de SVM, se reconoció un problema de potencia de cálculo computacional, ya que luego de 72 horas de corrida, los parámetros no se optimizaron. Se decidió entonces optimizar un modelo de la familia a Arboles de Decisión, *Random Forest*, que fue el que segundo mejor rendimiento obtuvo (75%).

Finalmente, luego de optimizar los parámetros se obtuvo un mejor ajuste de 0.76, es decir, del 76%. Evaluar la matriz de confusión obtenida (Fig. 4.4), permitió reconocer que, de las 1229 muestras totales utilizadas para evaluar el modelo, el 74,5% fueron bien predichas (915 ejemplos), y que la mayor concentración de los errores (302 sobre 314, aproximadamente 96% del error total) se encuentra concentrado entre las categorías Carbonáticas y Mixtas (Fig. 4.4). También se obtuvo la precisión para cada categoría siendo 66% para las carbonáticas; 63% para las mixtas y 100% para las silicoclásticas.

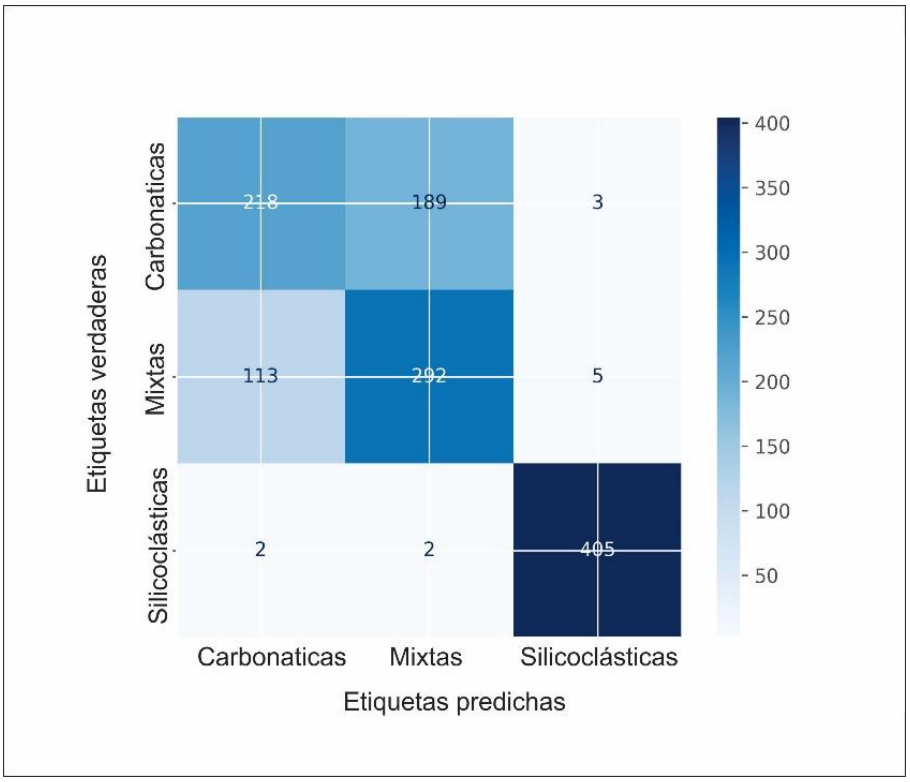


Figura 4.4: Matriz de confusión para el testeo del modelo de composiciones químicas

A la hora de observar la importancia de cada variable predictora se observó la información aportada por la optimización del modelo (Tabla 4.2) y luego se realizaron los gráficos SHAP para para cada categoría (Fig. 1 en Anexo 3).

Tabla 4.2: Importancia de cada predictor luego de la optimización del modelo

Predictor	Importancia
Al	0.22
K	0.14
Ca	0.14
Zn	0.12
Fe	0.1
Sr	0.08
Mg	0.06
Pb	0.05
Ti	0.04
S	0.02
P	0.01
Cr	0.01
Mn	0.01

Luego de la evaluación del modelo y del análisis de la incidencia de cada variable en las diferentes predicciones, se procedió a la realización de la validación del modelo. Como resultado se obtuvo una exactitud del 74% (Fig. 4.5) lo que es consecuente con los resultados mostrados en el testeo.

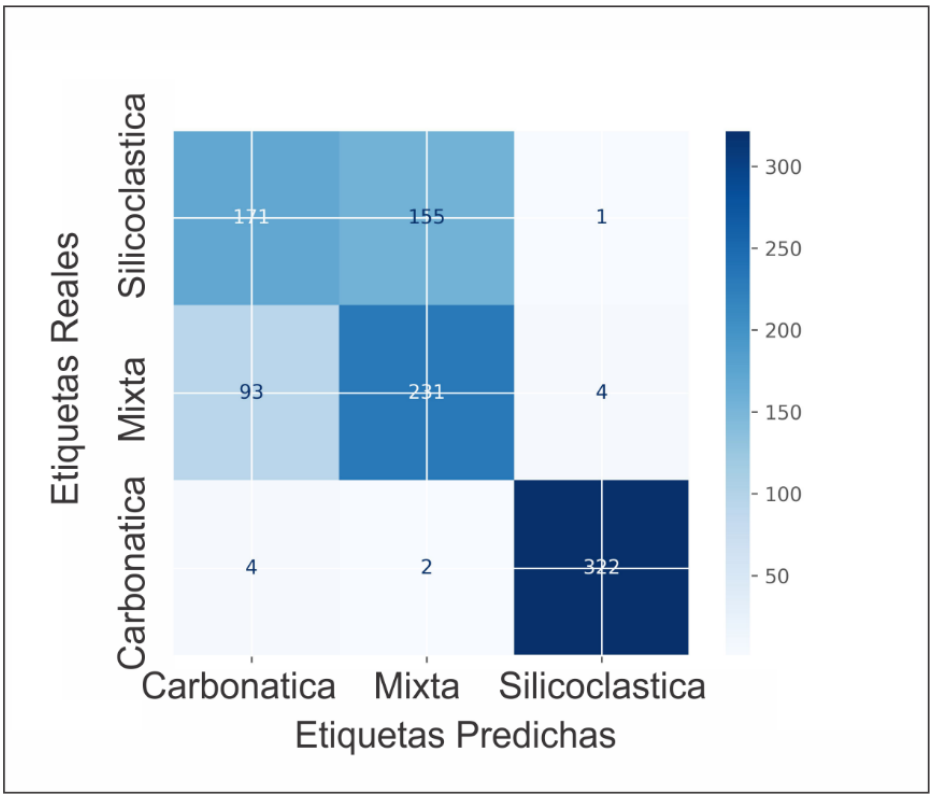


Figura 4.5: Matriz de confusión para la validación del modelo de composiciones químicas

4.3.2 Modelo Textura granulométrica

Como se explicó anteriormente, para el modelo de textura granulométrica, se agruparon los datos en 5 cm de espesor, es decir que cada "dato de entrenamiento" es un vector con los datos de 5 cm de roca (véase Figs. 3.17 y 3.18 en apartado 3.3.1).

La cantidad de datos (N) entre datos originales y datos sintéticos utilizada para dicho entrenamiento es de 234 para cada categoria. Este valor de N, al igual que en el modelo de composición química, responde al mínimo número de muestras sintéticas necesarias para balancear el conjunto de datos.

Inicialmente el valor de N seleccionado según la mejor exactitud mostrada en la figura 4.10 fue de 500, dado que mostraba una mejora del 13% respecto al modelo de menor cantidad de datos por categoria.

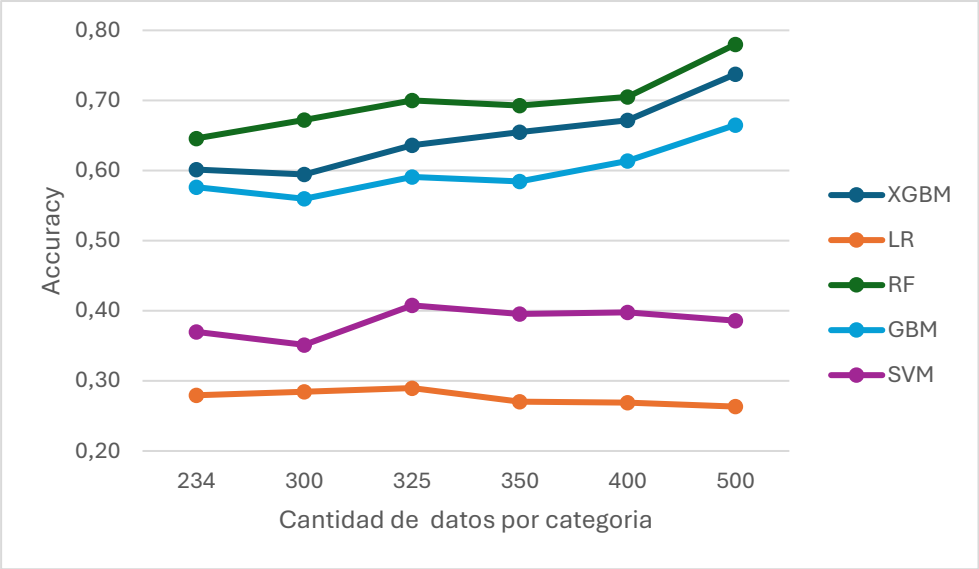


Figura 4.6: Exactitud de los distintos modelos de textura granulométrica entrenados en función de la cantidad de muestras por categorías.

Al optimizar el modelo con 500 datos por categorías se obtuvo la matriz de confusión (Fig. 4.7 A). Al compararla con la matriz de confusión obtenida para el modelo con N = 234 (Fig. 4.7 B) se observó que el modelo cometía los mismos errores, incluso que, en perspectiva de los errores, las predicciones mejoraban levemente en el modelo de menor N. Por este motivo se decidió continuar con el modelo de N=234.

Tal como se hizo con el modelo de composición química, se evaluó el rendimiento de los mismos a partir de la métrica exactitud, se pudo observar que todos los modelos procedentes de la familia de árboles de decisión entrenados poseían un rendimiento superior que el de los modelos de Regresión Logística y SVM a la hora de predecir las diferentes texturas granulométricas de las rocas. El mejor modelo entonces en relación exactitud/datos sintéticos es el RF con una exactitud del 65%.

Al evaluar la matriz de confusión obtenida (Fig. 4.7 B) permitió reconocer que, de las 562 muestras totales utilizadas para evaluar el modelo, 62.1% fueron bien predichas (349 ejemplos), y que la mayor concentración de los errores está

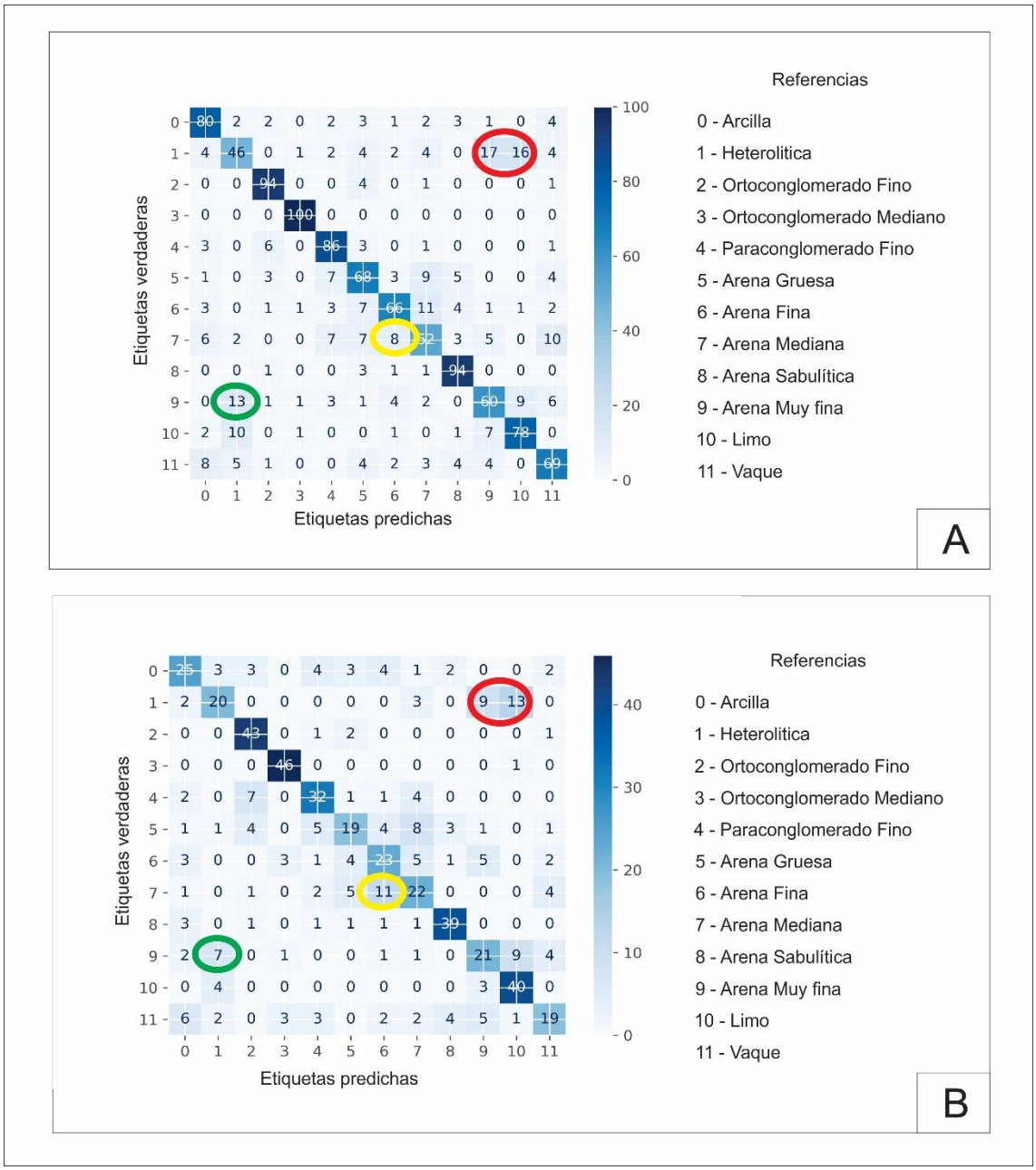


Figura 4.7: Matriz de confusión para el testeo de los modelos de texturas granulométricas con N = 500 (A) y N = 234 (B).

concentrada en tres grupos (Fig. 4.7B). El primer grupo marcado en rojo correspondían a ejemplos de granulometrías Heterolíticas que fueron predichas nueve veces como Arenas muy finas y 13 veces como Limos (Fig. 4.7B). El segundo grupo marcado con amarillo, correspondes a 11 muestras de Arena mediana que para el modelo fueron predichas como Arena finas (Fig. 4.7B). Mientras que el grupo verde son Arenas muy finas predichas como heterolíticas (Fig. 4.7B). A la hora de observar la importancia de cada variable predictora se observó la información aportada por la optimización del modelo (Tabla 4.3), y

luego se realizaron los gráficos SHAP para para cada categoría (Fig. 2 en Anexo 3).

Tabla 4.3: Importancia de cada predictor luego de la optimización del modelo

Predictor	Importancia
Disimilaridad	0.504
Entropía	0.496

Luego de la evaluación del modelo y del análisis de la incidencia de cada variable en las diferentes predicciones, se procedió a la realización de la validación del modelo. Como resultado se obtuvo una exactitud del 65% (Fig. 4.8) similar a los resultados mostrados en el testeo.

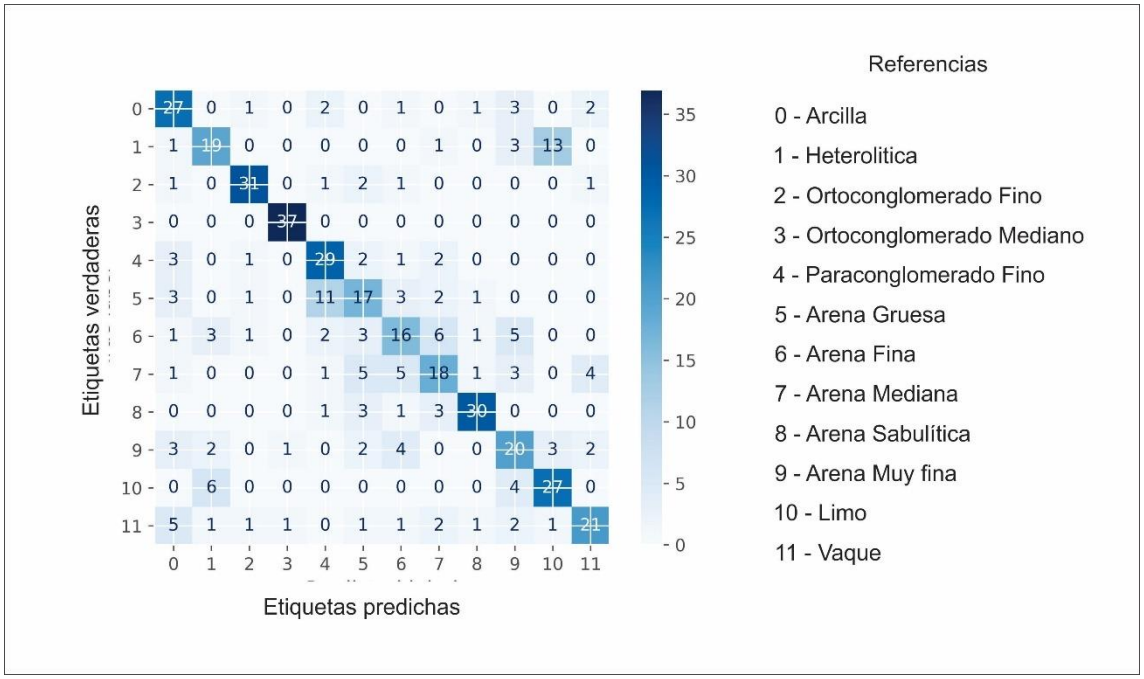


Figura 4.8: Matriz de confusión para la validación del modelo de texturas granulométricas

4.3.3 Modelo Estructura sedimentaria

Para el modelo predictor de las estructuras sedimentarias, se agruparon los datos en 10 cm de espesor, es decir que cada dato de entrenamiento es un vector con los datos de 10 cm de roca (véase Figs. 3.19 y 3.20 en apartado 3.3.1). La cantidad de datos (N) entre datos originales y datos sintéticos utilizada para dicho entrenamiento es de 187 para cada categoria.

Este valor de N, al igual que los modelos entrenados anteriormente, responde al mínimo número de muestras sintéticas necesarias para balancear el conjunto de datos. Inicialmente el valor de N seleccionado según la mejor exactitud mostrada en la figura 4.9 fue de 400, dado que mostraba una mejora del 12% respecto al modelo de menor cantidad de datos por categoría (Fig. 4.9). Al optimizar el modelo con 400 datos por categorías se obtuvo la matriz de confusión (Fig. 4.10 A), pero al compararla con la matriz de confusión obtenida para el modelo con N = 187 (Fig. 4.10 B) se observó que los errores que tenía el modelo con datos sintéticos eran los mismos que los que realizaba el modelo sin la generación de tantos datos sintéticos (círculos rojos, Fig. 4.10). De esta manera, se decidió ponderar la mayor cantidad porcentual de datos reales en la muestra total.

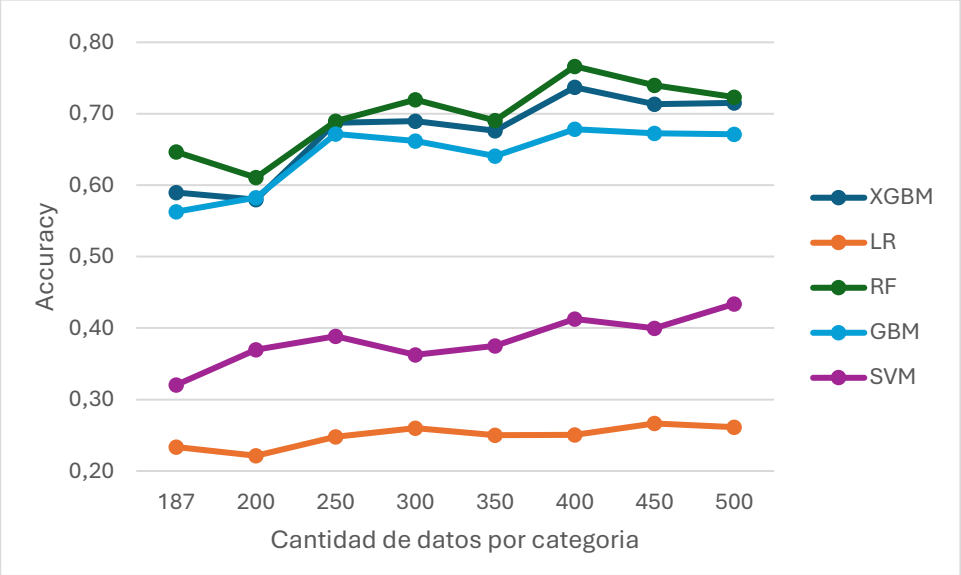


Figura 4.9: Exactitud de los diferentes modelos entrenados en función de los distintos valores de N.

En cuanto a los modelos, se evaluó el rendimiento de los mismos a partir de la métrica exactitud, se pudo observar que como ocurrió con la textura granulométrica, todos los modelos procedentes de la familia de árboles de decisión entrenados poseen un rendimiento muy superior que el de los modelos de Regresión Logística y SVM a la hora de predecir las diferentes texturas granulométricas de las rocas. El mejor modelo entonces en relación

exactitud/datos sintéticos es nuevamente el *Random Forest* con una métrica del 65%.

Al evaluar la matriz de confusión obtenida (Fig. 4.10 B) se reconoció que, de las 333 muestras totales utilizadas para evaluar el modelo, 62,46% fueron bien predichas (208 ejemplos). A diferencia de los modelos anteriores, los errores no se encuentran concentrados en grupos, sino que los mayores errores corresponden a ejemplares de estratificación entrecruzada tangencial, predicha como maciza y viceversa (círculos rojos, Fig. 4.10 B). A la hora de observar la importancia de cada variable predictora se observó la información aportada por

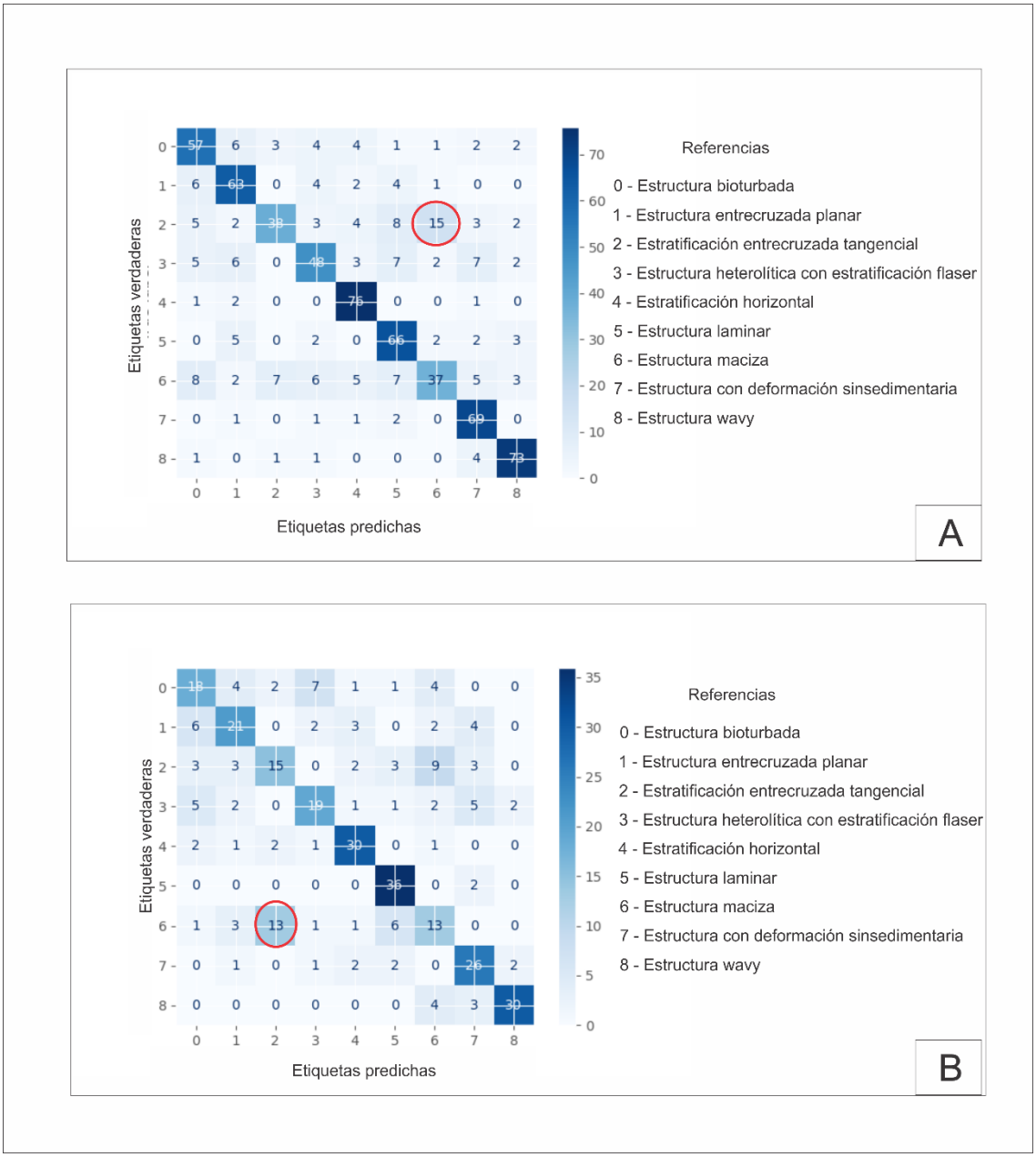


Figura 4.10: Matriz de confusión para el testeo de los modelos de texturas granulométricas con N = 400 (A) y N = 187 (B).

la optimización del modelo (Tabla 4.4), y luego se realizaron los gráficos SHAP para para cada categoría (Fig. 3 en Anexo 3).

Tabla 4.4: Importancia de cada predictor luego de la optimización del modelo

Predictor	Importancia
Energía	0.507
Entropía	0.493

Luego de la evaluación del modelo y del análisis de la incidencia de cada variable en las diferentes predicciones, se procedió a la realización de la validación del modelo (Fig. 4.11). Como resultado, se obtuvo que tanto los resultados, como la importancia de las variables predictoras fueron similares con los resultados mostrados en el testeo (Fig. 4.11).

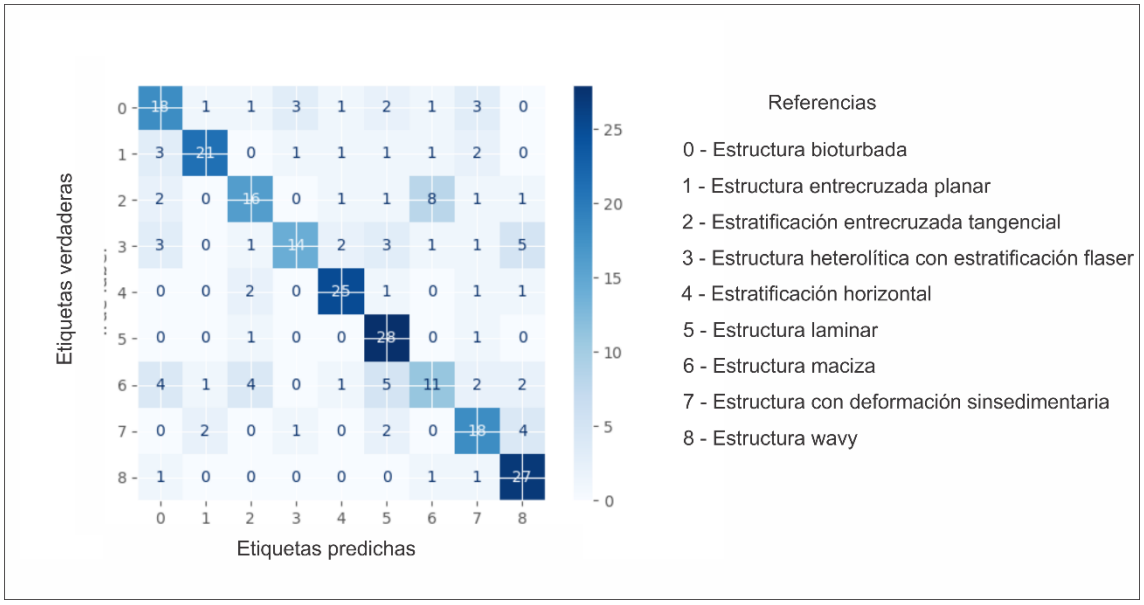


Figura 4.11: Matriz de confusión para la validación del modelo de estructuras sedimentarias

5

Discusiones

5.1 Sobre la descripción geológica

El proceso de descripción sedimentológica constituyó una etapa fundamental para establecer las bases para este tipo de estudio, permitiendo la conformación de un conjunto de datos confiable, homogéneo y estandarizado.

Las descripciones digitales a partir de imágenes de alta y ultra alta resolución (HRi y UHRi respectivamente), conformo un gran desafío dado que no es la manera tradicional de realización. Se observó que las estructuras sedimentarias de mayor escala, como por ejemplo las estratificaciones entrecruzadas tangenciales, planares son visibles a menor escala, mientras que estructuras de menor escala como las estratificaciones mixtas flaser u ondulítica, o las bioturbaciones necesitaron mayor detalle.

En lo que respecta al análisis granulométrico, la determinación del tamaño de grano presentó ciertas dificultades, aunque permitió obtener mediciones precisas. En el caso de las granulometrías finas, la posibilidad de medir a partir de vectores el largo especifico de los diferentes ejes de los clastos, condujo a una clasificación robusta del tamaño de grano. Además, la facilidad del método permitió la realización de numerosas mediciones, disminuyendo posibles sesgos por parte del descriptor. Sin embargo, una limitación importante fue la incapacidad del programa de análisis de imágenes para almacenar automáticamente dichas mediciones, lo que obligaba al descriptor a anotarlas manualmente o memorizarlas para luego calcular los promedios correspondientes. Actualmente, la medición del tamaño granulométrico se realiza de manera visual a partir de cartillas comparadoras o, en el caso de granulometrías muy finas, a partir del micrómetro del microscopio óptico (De Raaf et al., 1965; Kidwell & Holland, 1991; Net &Limarino, 2000; Kietzmann et.

al.,2014; Minisini et. al.,2020). Por este motivo, las mediciones del tamaño de los clastos no son precisas. Sumado a esto, la subjetividad y experiencia del observador que realiza la descripción es una variable importante dado que introduce un sesgo que podría influenciar a los resultados del modelo.

En el caso de las granulometrías gruesas, la descripción digital a partir de imágenes de UHRi queda imposibilitado por razones de escala. Para ello se necesita utilizar las imágenes de HRi las cuales no contaban con las herramientas necesarias para medir la longitud de clasto, por lo que se midieron con el mismo principio, pero por fuera de la herramienta. Esta imposibilidad de medición por parte del software en granulometrías gruesas es sin duda un aspecto a mejorar por parte de la empresa desarrolladora, dado que se debió incurrir a adaptaciones de metodologías clásicas.

Otro factor a considerar en la definición del tamaño de grano es la probabilidad de intersección de corte de los diferentes granos en la imagen de UHRi. Similar a los cortes delgados petrográficos, la obtención de la distribución granulométrica está influenciada por el ángulo en el que los granos fueron cortados y fotografiados (Fig. 5.1A). Este ángulo puede afectar las dimensiones aparentes de los granos y, por lo tanto, la interpretación de su tamaño real (Smith & Yoder 1956; Scasso & Limarino 1997; Taylor et. al, 2022). Cuando la roca es preparada para las mediciones, se nivela con un accesorio especializado (minislab) que asegura una superficie plana y uniforme (Fig.5.1B). Este accesorio puede cortar o arrancar los granos según su tipo y composición (Germay et. al, 2023). En caso de que los granos sean cortados, al igual que en el método de confección de los cortes delgados, la posición de los mismos puede dejar secciones tangenciales o subtangenciales visibles. Las secciones tangenciales pueden presentar los granos como elípticos o alargados, en lugar de su forma real, lo que puede llevar a interpretaciones erróneas sobre su tamaño y distribución (Smith & Yoder, 1956; Scasso & Limarino, 1997; Taylor et. al, 2022).

Si bien existen códigos de facies sedimentarias internacionalmente aceptados que fueron desarrollados y ajustados por diversos autores (Dunham,

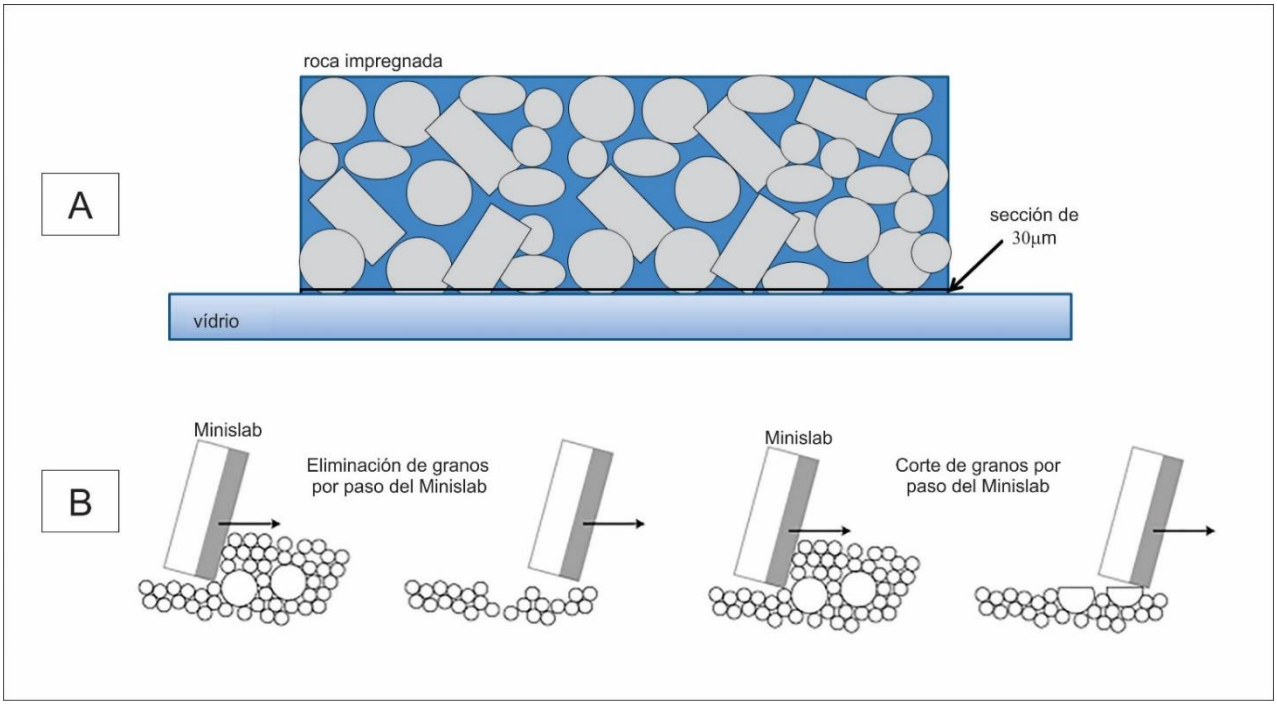


Figura 5.1: Esquema del problema de intersección de corte de los diferentes granos. (A) Intersección de granos en la preparación de un corte delgado, tomado de Taylor et. al, 2022. (B) Intersección de granos en la preparación de la roca por el Minislab, tomado de Germay et. al, 2023.

1962; Embry & Klovan, 1971; Miall, 1977; 1988; Lokier & Al Junaibi; 2016; Fisher, 1961, Fisher & Schmincke, 1984, Terruggi et al.,1978 y Mazzoni,1985), cada geocientista le proporciona su impronta, su idioma, y le introduce modificaciones. Es así como una misma facie sedimentaria, puede poseer diversas nomenclaturas. Por ejemplo, Kietzmann et. al. (2014) describe una facies de Margas laminadas que las codifica como como “Mrh”; Arrouy y colaboradores (2016) codifican una facies de margas laminadas como “Ml”. Sin embargo, en su trabajo Minisini et. al. (2020) nombra con la nomenclatura “Mr” a las margas heterolíticas y no a las laminadas.

En este estudio se buscó respetar esos acuerdos tácitos construidos a lo largo de los años en relación a las nomenclaturas propuestas para cada característica de la roca. Pero también se observó la necesidad de introducir un nuevo código que estandariza la longitud de los mismos para hacerlo entendible a la inteligencia artificial. Dado estas dos caras que parecieran no poder coexistir entre sí, se decidió generar un código interno para el programa y uno más ameno para la salida gráfica.

La combinación de descripciones visuales (a partir de imágenes de HRi y UHRi), con datos químicos y geomecánicos proporcionó un enfoque multidimensional para evaluar las muestras de rocas. Asímismo, la generación de un sistema de codificación de facies interno, donde cada facies sedimentaria se compone de la misma manera y con la misma longitud desde el inicio del trabajo, facilitó la integración de las descripciones geológicas con los algoritmos. Todo esto en su conjunto, permitió generar un conjunto de datos con variabilidad de facies sedimentarias, las cuales sirvieron de referencia para el entrenamiento de modelos de aprendizaje automático.

5.2 Sobre el preprocesamiento de los datos geológicos

Se aplicaron diversas técnicas estadísticas con criterios geológicos que aseguraron la calidad del dato para su modelado. La metodología utilizada en la sustitución de datos faltantes fue una decisión discutida ya que no solo influye en la cantidad de datos a eliminar, sino también en la calidad del dato conservado. En lo que respecta al relleno de datos faltantes, se realizó a partir de la imputación de la media del error instrumental en cada muestra para cada elemento (véase apartado 3.2). De esta manera, se minimiza el impacto de la sensibilidad del instrumento de medición. (Verbovšek, 2011; Henne et al.,2024). Otras opciones de imputación de datos como imputación por promedio o vecinos más cercanos se consideraron poco confiables por el apilamiento de las facies sedimentarias que hace que el promedio general de cada elemento no sea representativo de la facie sedimentaria en sí, y, por otro lado, el cambio abrupto entre facies sedimentarias consecutivas hace que el valor más cercano pueda no provenir de una facies equivalente (Grunsky & de Caritat, 2017).

La detección de valores anómalos puede realizarse a partir de análisis estadísticos básicos (Zar, 1999; Alperin, 2013; Illés & Nagy, 2007), o de metodologías de geoestadística (Bárdossy et al., 1990; Luo et al., 2018; Helwig et al., 2019) cuando la dimensionalidad de los datos es muy grande, es decir, cuando poseemos muchas variables a analizar puede volverse laborioso. En este trabajo de tesis se optó por realizar una búsqueda de valores anómalos de forma multivariada, como lo hizo Lalor et al. (2001), Filzmoser et.al. (2005) y Filzmoser et.al. (2008).

La utilización del análisis PCA para la reducción de dimensionalidad de las variables es ampliamente utilizado en la literatura (de Caritat & Grunsky, 2013; Booker et al., 2022; Guo, et al., 2022; Bashir et al., 2024). Este se ha vuelto un método rápido y robusto para comprender como interactúan las variables entre sí, y sus relaciones. En este trabajo de tesis, se optó por realizar la reducción de variables para cada conjunto de datos por separado identificando y cuantificando relaciones entre las distintas variables.

Por otra parte, la transformación de las imágenes como método para pasar de matrices multidimensionales a datos continuos y representativos de las características de las rocas resultó positivo ya que permitió una forma automatizada de extracción de patrones característicos. Si bien su costo computacional puede ser alto al generar las matrices para luego generar las texturas, este almacenamiento en la memoria es temporal y menor que el entrenamiento de redes neuronales dado que esta última opción no fue posible.

5.3 Sobre los modelos no supervisados

Ippolito et al. (2021); Martin et al. (2021), Germay et al. (2023); Eftekhari, et al. (202) y Liu (2024) utilizaron los resultados de modelos no supervisados como datos de entrada para modelos de algoritmos predictivos. En este trabajo de tesis se realizó un análisis similar.

El algoritmo utilizado (*k-means cluster*) identificó 6 grupos como los grupos óptimos. Si bien este número es solo una referencia en la cantidad de grupos, teniendo en cuenta el conocimiento del problema y un preconcepto de un número de grupos (facies sedimentarias) que podrían formarse, definir 6 grupos no concuerda con la información geológica que se tiene. En esta tesis, y con este conjunto de datos en particular, se concluyó que este algoritmo no es lo suficientemente asertivo para identificar datos correspondientes a mismas facies geológicas. La principal causa de esta discrepancia puede deberse a que las facies sedimentarias son una combinación de características de las rocas, lo que numéricamente se traduce en una alta colinealidad entre los datos (Han et al., 2024). Esto podría dificultar al algoritmo *k-means* y todos los algoritmos que basen su estrategia de clusterización a partir de similitud entre datos (distancia euclídea), la capacidad de distinguir datos con rasgos similares. A partir de esto,

se concluyó que las categorías generadas por el *k-means cluster* no eran representativas de las facies sedimentarias que se buscaban predecir, por lo que no serían de utilidad como insumo para los modelos supervisados.

5.4 Sobre los modelos supervisados

La imposibilidad de entrenar un modelo de aprendizaje automático que prediga directamente las facies sedimentarias puede atribuirse a diferentes motivos, pero en particular, y desde un punto de vista estadístico, se interpreta que puede deberse a:

- La falta de representatividad de las muestras originales.
- Sesgo original de las muestras.

Por otra parte, también puede deberse a la naturaleza geológica del problema. Partiendo de la base de que el concepto de facies sedimentarias es la combinación de texturas granulométricas, su composición y de la estructura sedimentaria que presentan las rocas:

- Existen 5775 categorías objetivos posibles, lo que imposibilita la representatividad de toda la población geológica.
- La colinealidad intrínseca existente dentro de cada categoría, ya que los límites granulométricos y/o composicionales no son limites claramente netos, lo que hace subjetivo la definición de los diferentes tipos.

Aunque el conjunto de datos se confeccionó con la idea de que contenga la mayor representatividad de facies sedimentarias posibles, la inmensa cantidad de combinaciones naturalmente posibles limitó la representatividad de algunas facies. Sumado a esto, el análisis sistemático de testigos coronas en empresas de servicios de la industria del O&G es escaso, debido a que cada sistema petrolero posee características particulares.

De la mano con la representatividad de las facies sedimentarias, se observó que, aunque el problema se subdividió para simplificarlo, el data se seguía estando desbalanceado. Autores como Deng (2017), utiliza SMOTE como técnica de sobre muestreo para balancear las categorias. En cambio, en

la tesis se consideró oportuno realizar un balance de datos manual, utilizando criterios geológicos y estadísticos. A diferencia de los métodos de balanceo propuestos en la literatura como métodos de sub y sobre ajuste (Laurikkala, 2001; Barandela et.al, 2003; Liu, et. al, 2008; Douzas et.al, 2018), este balanceo permitió mantener la distribución estadística original de las muestras ajustándose a las tendencias naturales observadas intrínsecamente en las facies sedimentarias. El procedimiento incluyó el cálculo de estadísticos de tendencia central y de dispersión, así como tendencias en función de las profundidades. Esto evitó las duplicaciones exactas de muestras minoritarias, lo cual podría haber inducido sesgos al modelo. También, se evitó la perdida de información asociada al submuestreo de categorías mayoritarias. Este tipo de generación de datos requiere un entendimiento profundo del concepto de facie sedimentaria, así como de tendencias geoquímicas, granulométricas y estructurales, lo que limita su reproducibilidad en otros trabajos donde el científico de datos no sea geólogo.

Con respecto al modelo de composición química (Fig. 4.3), el rendimiento disminuye a razón del 2% a medida que aumenta la cantidad de datos por categoría. Se decidió continuar con el modelo cuyo valor de N es 2048 ya que tiene el mejor rendimiento y mayor cantidad de datos medidos. La evaluación de la matriz de confusión (Fig. 4.4) permitió reconocer que, de las 1229 muestras totales utilizadas para evaluar el modelo, 915 fueron bien predichas, y que el 96% del error total de las predicciones mal realizadas se encuentra concentrado entre las categorías Carbonáticas y Mixtas. Si bien desde un punto de vista del análisis de datos, tener como respuesta un 76% de exactitud no es un número esperable, si lo observamos desde un punto de vista geológico los errores cometidos por el modelo son los esperables. En la naturaleza geoquímica de rocas, las composiciones mixtas son complejas. Su origen está basado en ambientes sedimentarios donde los procesos de aportes silicoclásticos y carbonáticos ocurren o bien simultáneamente o de manera alterna, lo que genera una estrecha interrelación entre las composiciones de la roca (Boogs, 2009).

A la hora de observar tanto la Tabla 4.1 donde se muestra la importancia de las variables predictoras como los 3 gráficos SHAP (Fig. 1 en Anexo), se reconoce que los elementos más importantes a nivel general para separar las

distintas composiciones son el aluminio, el potasio, el calcio, el zinc, y el hierro, lo cual desde un punto de vista geoquímico es entendible debido a que son los principales elementos formadores de minerales en la corteza terrestre (Misra, 2012). Además, en los gráficos SHAP (Fig. 1 del Anexo 3) se observa que los elementos importantes por cada categoría siguen un criterio geológico, dado que por ejemplo el aluminio (Al) es el elemento más importante para predecir las composiciones silicoclásticas y carbonáticas. En el caso de las composiciones carbonáticas, que una muestra presente valores bajos de aluminio (puntos azules), generan valores SHAP positivos, es decir que tienen alto impacto en las predicciones del modelo para esa categoria. En otras palabras, ante valores bajos de aluminio es altamente probable que el modelo prediga la clase carbonática. Si observamos las predicciones de composiciones mixtas, el aluminio cae en importancia a un segundo lugar, pero mayoritariamente valores altos de aluminio continúan impactando negativamente en las predicciones, lo que responde a que estas rocas mixtas son pobres en aluminio. Esto significa que el modelo es capaz de comprender que, para este conjunto de datos, las rocas carbonáticas y mixtas contienen menor concentración de aluminio que las rocas silicoclásticas.

Por otro lado, el modelo de textura granulométrica presenta una exactitud del 65%. Tal como ocurre con el modelo composicional, aunque en términos del análisis de datos no sea un valor esperable, desde un punto de vista geológico es aceptable dada la complejidad de la variable a predecir. Los errores presentados, están dentro de los errores esperables para descripciones de diferentes geocientistas (Álvarez Trentini & Schwarz, 2016; Álvarez Trentini et al., 2018; Minisini et. al, 2020). En este sentido, por ejemplo, las facies arcilitas y las facies de mudstones presentan similar textura granulométrica, pero en diferentes porcentajes de componentes carbonáticos. Los mudstone presentan más del 60% de carbonato de calcio, mientras que las arcilitas poseen menos del 40%, sin embargo, el proceso sedimentario que las forma es el mismo (Álvarez Trentini & Schwarz, 2016; Álvarez Trentini et al., 2018; Minisini et. al, 2020). El modelo revela tres patrones principales de error: (1) rocas heterolíticas, predichas erróneamente como limos o arenas muy finas; (2) arenas medianas confundidas con arenas finas; y (3) arenas muy finas predichas como limos.

Estos errores, desde un punto de vista geológico, no son considerados como errores graves, sino que son considerados como “esperables”. Las muestras clasificadas como heterolíticas están compuestas por la alternancia de depositación de sedimentos de granulometrías limosas y arenas muy finas o finas, por lo que es esperable que el modelo confunda dichos intervalos con las otras granulometrías. La imposibilidad de separar dicha alternancia se debió a su escaso espesor y a la interrelación de los procesos que las depositan (i.e. decantación a partir de suspensión - tracción). Nótese en la figura 5.2, la similitud entre las imágenes de alta resolución utilizadas para entrenar el modelo, así como el espesor de la interdigitación de los procesos sedimentológicos que genera la roca.

Desde un punto de vista del análisis de datos, estos errores reflejan similitud entre las clases en términos de sus datos (los cual tiene una similitud en su origen geológico), lo que limita la capacidad del modelo para establecer fronteras de decisión precisas. Esta situación podría resolverse redefiniendo las categorías clasificadas como heterolíticas en arenas o limos tratando de cuantificar las proporciones de ambas y seleccionando un umbral de proporciones cuando se encuentran interdigitadas como es el caso de las rocas heterolíticas (aunque esto conllevaría a sumar más heterogeneidad o ruido a las clases limo y arena). Otra posible solución sería incorporar otro tipo de variables predictoras adicionales que puedan capturar mejor las diferencias más sutiles entre los límites de las categorías. Si bien el ACP mostraba que ambas variables eran las que más variabilidad aportaban al conjunto de datos, el replanteamiento de las variables predictoras o la sumatoria de alguna otra podría influenciar positivamente la diferenciación más detallada de este tipo de errores.

Con respecto a los errores de predicción de arenas medianas confundidas con arenas finas y de arenas muy finas predichas como limos, se interpreta que se debe a la superposición de cada categoría (Hastie et al., 2009). La transición granulométrica entre los distintos tamaños de granos es muy puntual considerando los tamaños de la escala φ (tabla 1.1), e incluso muchas veces puede llegar a ser gradual. Las distintas granulometrías no siempre presentan un límite de cambio neto, sino que pueden ir cambiando de manera transicional lo que dificulta su separación incluso en un contexto geológico detallado (Boggs, 2009). La similitud entre los distintos tipos granulométricos en términos de su

tamaño de partícula puede ser engañosa, especialmente si se trabaja con imágenes (Fig. 5.3).

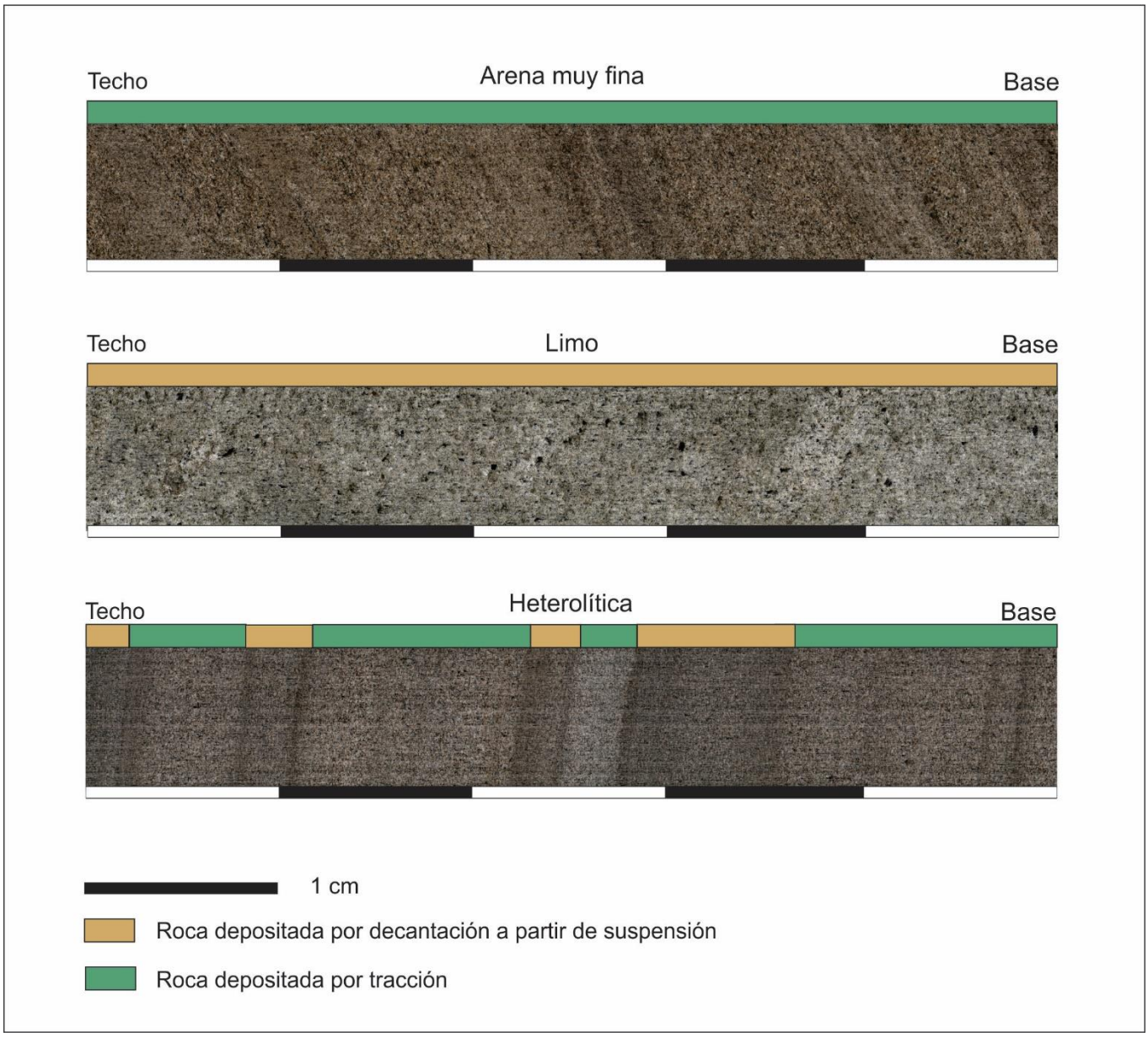


Fig. 5.2: Figura donde se observa la similitud de las imágenes de muestras de roca heterolítica, y rocas de granulometría limo y arena muy fina.

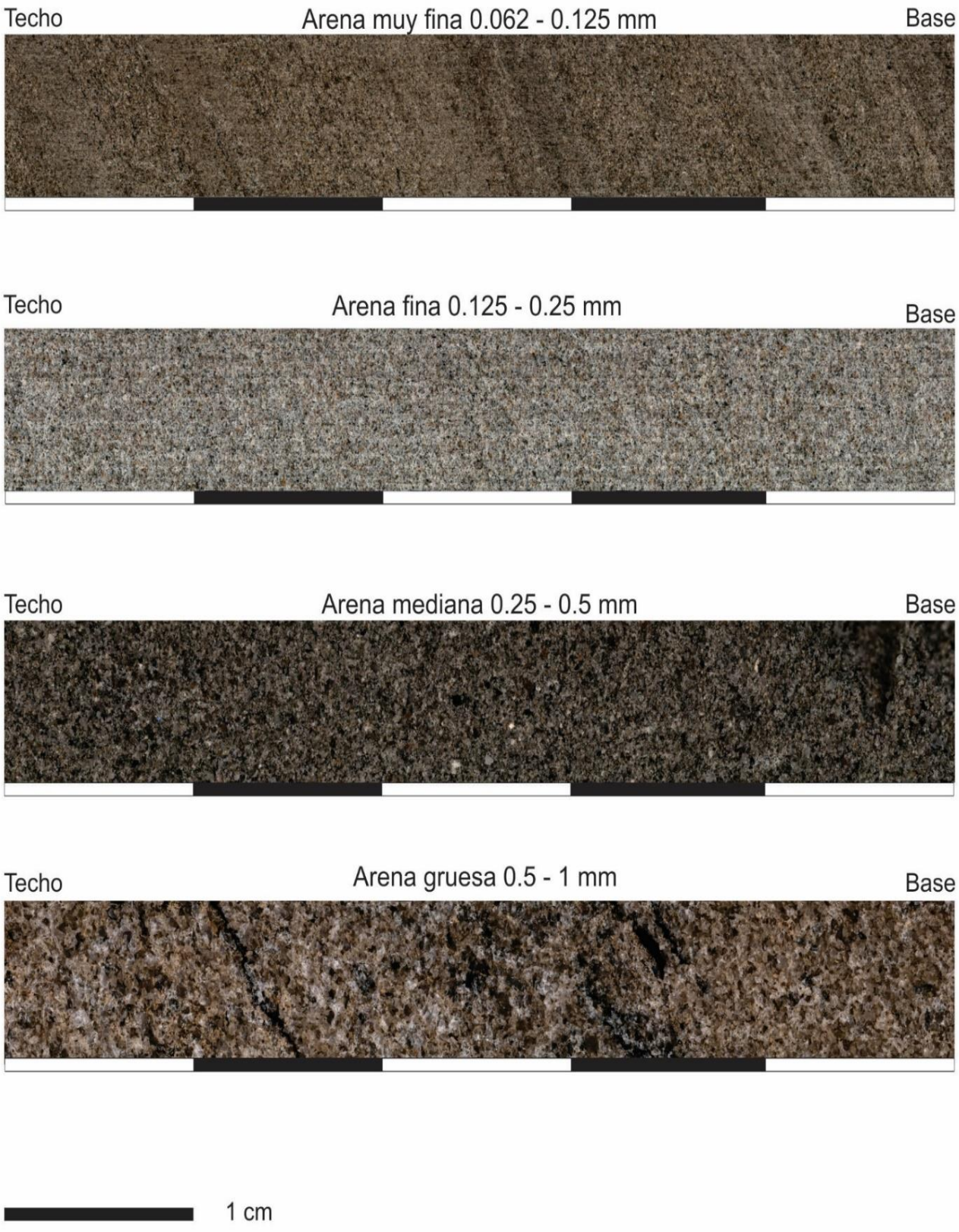


Fig. 5.3: Esquema que muestra la similitud entre los distintos tipos granulométricos en imágenes de UHRi.

Otra posibilidad es que las confusiones observadas podrían también estar relacionadas con la representación insuficiente de estas clases en el conjunto de datos. En este caso se podría aumentar el conjunto de datos inicial de forma tal que las categorías queden más representadas.

Al observar las gráficas de SHAP se observa que en la disimilaridad es más determinante. Esto puede traducirse como que las variaciones locales de la imagen, es decir los cambios cercanos de los valores de pixel son más determinantes en la determinación de la textura de la roca que las tendencias generales captadas en la entropía. Esto tiene sentido si se piensa en los cambios locales como cambios producidos por la detección de los bordes de granos.

El modelo de estructuras sedimentarias destaca confusiones frecuentes entre las estructuras estratificación entrecruzada tangencial con las estructuras macizas. Estos errores fueron catalogados desde un punto de vista geológico, como errores de escala. Las estratificaciones entrecruzadas poseen una escala muy variable, pudiendo estar compuestas de *sets* de pocos centímetros hasta varios metros de espesor. Internamente, cada *set* puede tener una estructura propia que responde al proceso que esté secundando en ese momento al proceso de migración de la ondula/megaóndula primario. Es por eso, que se interpreta que los errores asociados a las muestras con estratificaciones entrecruzadas predichas como macizas, el modelo podría estar observando la estructura interna del *set* ya que la estructura primaria entrecruzada posee más de 10 cm de espesor.

En lo que respecta a las variables predictoras, la energía y la entropía contribuyen en general de igual manera en las predicciones que realizará el modelo, siendo levemente más importante la energía. Desde un punto de vista de las variables, esto es correcto dado que la energía es una forma de medir la homogeneidad de la imagen. Por otro lado, la entropía, mide la cantidad de información o incertidumbre en la imagen, lo que estaría capturando la regularidad o la ausencia de variabilidad que hay en dicha imagen.

Como se ve en las Figs. 4.3, 4.6 y 4.9, el aumento de la cantidad de datos no mejora significativamente el rendimiento de los modelos de textura

granulométrica, estructura sedimentaria. Este comportamiento puede deberse a varias razones:

- Las variables predictoras no contienen suficiente información discriminativa y el modelo no podrá distinguir adecuadamente entre las clases de la variable objetivo, independientemente de la cantidad de datos disponibles (Goodfellow et al., 2016).
- Aumentar el tamaño del conjunto de datos a partir de datos sintéticos ta puede saturar la capacidad del modelo sin mejorar su capacidad de generalización (Domingos, 2012).

En cuanto al rendimiento de los tres modelos de aprendizaje automático en general, es interesante que en todas las ocasiones el RF muestre un rendimiento levemente más alto que el resto de los modelos de la familia de ensamble de árboles de decisión (GBM y XGBM). RF utiliza un enfoque basado en la construcción de múltiples árboles de decisión independientes, combinando sus predicciones mediante votación. Este procedimiento ayuda a reducir el riesgo de sobreajuste, incluso sin un ajuste exhaustivo de hiperparámetros (Breiman, 2001). Por otro lado, los métodos GBM y XGBM, entrenan los árboles secuencialmente, lo que los hace más susceptibles al sobreajuste (Hastie et al., 2009). En conjunto de datos pequeños o medianos, RF tiende a superar a los métodos de *boosting* debido a su capacidad para generalizar mejor con menos datos (Hastie et al., 2009). Además, si los datos contienen relaciones no lineales complejas, RF puede capturarlas eficazmente a través de la aleatorización inherente en su construcción de árboles. Por el contrario, los métodos de *boosting*, al construir árboles más profundos, pueden ajustarse demasiado a relaciones espurias en el conjunto de datos.

Los modelos entrenados en este trabajo de tesis fueron modelos de aprendizaje automático multiclase, los cuales abordan problemas donde un modelo debe predecir múltiples categorías de la variable objetivo simultáneamente. A medida que los modelos multiclase se vuelven más complejos, su interpretabilidad disminuye. Entender cómo las características de entrada afectan múltiples categorías objetivo y cómo estas interactúan entre sí

es una tarea difícil, especialmente en modelos denominados como “caja negra” como las redes neuronales profundas.

Dentro de los enfoques utilizados para entrenar los modelos *multiclase*, la estrategia utilizada en esta tesis es la llamada “*One-vs-Rest” (OvR),* que consiste en transformar el problema de clasificación multiclase en múltiples problemas de clasificación binaria. En cada caso, cada clasificador se entrena para distinguir una clase específica respecto a todas las demás clases combinadas. Esta metodología se utilizó en todos los modelos, ponderando principalmente tres factores en su elección:

(1) simplicidad en la interpretabilidad

(2) escalabilidad en la cantidad de clasificadores binarios (dadas *n* clases, tenemos n clasificadores)

(3) menor tiempo y recursos computacionales respecto a otros métodos de clasificación multiclase como puede ser “One-vs-One” (OvO).

El método OvO ofrece, desde una perspectiva teórica, una mayor precisión respecto al OvR, por la comparación entre sólo dos categorías, independizándose del resto. En el presente estudio, el uso de OvO respecto al OvR, no tiene un alto impacto, al menos en el modelo de composición química, ya que este posee sólo tres categorías. Podría llegar a considerarse más significativo en los otros dos modelos dada la cantidad de categorías (12 en el caso del modelo de textura granulométrica y 9 en el de estructura sedimentaria). Mas allá de esto, asumiendo las incertidumbres del problema geológico en sí mismo, se decidió continuar con el método OvR ponderando los beneficios enumerados en el párrafo anterior.

5.5 Comparación con bibliografía de aplicación de ML en geociencias

Bibliografía de trabajos similares respecto a la aplicación de modelos de ML aplicados a la sedimentología presentan enfoques similares a esta tesis, con resultados disimiles.

Al-Anazi & Gates (2010) y Ai et al. (2019), recurren registros de pozo (gamma ray, curva de densidad, perfil sónico, perfil neutrónico y curvas de resistividad), para modelar distintos tipos de arenas (*"Clean Sand", "Shaly Sand" y "Sandy Shale"*) y subambientes sedimentarios (planicie deltaica, frente deltaico y costa de lago), respectivamente. En ambos trabajos se compararon resultados de diferentes modelos de ML supervisados concluyendo que el SVM era el de mejor rendimiento. En el caso de Al-Anazi & Gates (2010) en su trabajo no exponen métricas claras de rendimiento, sino que hacen referencia a que el SVM supera a otros tipos de modelos con tasas de clasificación mejores. Por su parte, Ai et al. (2019) reportan desempeños del SVM, con exactitud del 70%, aunque los datos utilizados son escasos. En los resultados obtenidos en este trabajo, el SVM posee un rendimiento inferior frente a algoritmos de la familia de los árboles de decisión, como RF o GBM. Esta diferencia puede deberse a el carácter unidimensional de los datos de pozo utilizados por los otros autores.

En contraposición, los datos de esta tesis son de diferente naturaleza (tabulares e imágenes), respecto a los utilizados por los autores citados previamente. En este caso, los modelos basados en árboles, como RF y GBM, superan al SVM para esta tesis.

Demyanov et al. (2019) tienen por objetivo predecir facies sedimentarias a partir de estructuras sedimentarias utilizando mapas autoorganizados (SOM) combinados con métodos de agrupamiento. El enfoque de buscar patrones de comportamiento en los nodos salientes de los mapas resulta metodológicamente complementario al de este trabajo doctoral. Aunque las metodologías son distintas, ambos comparten el uso de datos derivados de testigos de roca. Esta coincidencia sugiere que técnicas como los SOM podrían utilizarse en la tesis para agrupar etiquetas en categorías robustas desde un punto de vista estadístico.

Asimismo, en este trabajo se introduce el uso de valores SHAP, no utilizados por los autores anteriormente mencionados para interpretar el comportamiento de los modelos, lo que representa una ventaja notable. Si bien estos estudios evalúan la precisión de los modelos, no se detienen en la explicación de cómo y por qué se toman las decisiones de clasificación, algo relevante en contextos geológicos donde la interpretabilidad y el criterio geológico son claves para validar resultados.

5.6 Criterios de unificación de los modelos para una única predicción de facies sedimentaria

Para satisfacer el objetivo inicial de predecir las facies sedimentarias, aquí se discute como se debe abordar la unión de los 3 modelos predictivos para conformar una predicción final de una facies sedimentaria.

Al analizar las diferentes clasificaciones existentes para las rocas sedimentarias, observamos que la primera gran diferenciación se basa en el tipo de composición química y/o mineralógica de los sedimentos que conforman a las rocas sedimentarias. Estos diferentes tipos de composición generan los cuatro grupos de rocas sedimentarias: (1) rocas silicoclásticas, (2) rocas carbonáticas, (3) rocas de composición mixta y (4) las rocas volcaniclásticas. Estas son diferenciadas por su contenido en elementos químicos mayoritarios como el aluminio, el silíceo y el carbonato (Craigie, 2018). Si bien cada gran tipo de roca tuvo un desarrollo autónomo respecto al resto en lo que en lo que refiere a la clasificación de facies, todas comparten intrínsecamente la misma idea y estructura basadas en el arreglo textural de los componentes.

Dentro de cada tipo de composición de roca sedimentaria, la siguiente gran diferenciación propuesta por los diversos autores, corresponde a una diferenciación de tipo textural, que responde a aspectos como el tamaño de grano, la forma, la redondez y el grado de selección de los sedimentos. Aunque cada autor le proporcione un nombre distinto, las métricas que utiliza al igual que los límites de separación son los mismos. Un claro ejemplo se reconoce en la clasificación de Dunham (1962) modificada por Embry & Klovan (1971) para rocas carbonáticas, la cual utiliza el término *grainstone* para rocas grano soportadas cuyo tamaño de grano se encuentra entre 63 micrones a 2 milímetros. Si observamos el mismo tipo de roca, pero de composición silicoclástica, Allen (1985) y Pettijohn (1975) las clasifican como areniscas conservando los mismos límites entre las categorías; y si observamos las rocas de composición volcaniclásticas, Terruggi et al. (1978) y Mazzoni (1985), las clasifican como tobas, nuevamente conservando los límites entre los tamaños de granos utilizados en rocas silicoclásticas. Estos aspectos texturales que se utilizan para clasificar las rocas permiten no solo caracterizar los sedimentos disponibles, sino que nos permiten deducir características intrínsecas del agente

de transporte y los procesos deposicionales, como son su energía, fluidez/viscosidad y el tiempo de transporte. Luego de clasificar las rocas en función de su composición y su textura, los diversos autores utilizan la estructura sedimentaria a modo de describir la interacción entre los agentes sedimentarios y el sustrato (estructuras sedimentarias mecánicas); así como también la interacción de los organismos y el sustrato (estructuras sedimentarias orgánicas). Como se mencionó en el apartado 1.3, las estructuras sedimentarias nos brindan información adicional sobre el agente sedimentario y los mecanismos de transporte de sedimentos, las condiciones del régimen de flujo, la dirección y sentido de migración de los sedimentos, las condiciones del sustrato y las condiciones paleoambientales que influenciaron la actividad biológica (Pettijohn, 1957; Bromley,1996; Cheel, 2005; Collinson et al., 2006; Buatois & Mángano, 2011; Ponce et al., 2018).

La unión de los modelos debe reflejar entonces la línea de pensamiento que siguen las clasificaciones internacionalmente aceptadas (Folk 1959; Fisher, 1961; Dunham, 1962; Embry & Klovan, 1971; Terruggi et al. 1978; Miall,1978, 1984, 1988, 2022; Fisher & Schmincke, 1984; Mazzon,1985; Lokier & Al Junaibi, 2016). Es por ello que en la conformación de la facies sedimentaria primero se debe considerar la predicción del modelo de composición, a fin de obtener la generalidad de la composición de la roca. Luego el modelo de textura para caracterizar tanto el sedimento disponible en el entorno para ser movilizado, como las características del agente de transporte que predomina en el ambiente y, finalmente, el modelo de estructura sedimentaria para caracterizar a través de que procesos de transporte los agentes movilizan y depositan el material disponible para la generación de la roca. Es importante destacar que, los modelos fueron entrenados de manera individual, lo que los convierte en independientes uno de otro. Además, cada modelo fue entrenado con un conjunto de datos balanceado, lo que hace que, desde un punto de vista puramente probabilístico, las probabilidades de predecir una u otra característica sean equiprobables dentro cada modelo (Davis & Sampson, 1986; Zar, 1999; Alperin, 2013; Mena, 2016).

Se definieron dos maneras para realizar la unión de los modelos, la primera y más sencilla es de manera determinística. Para ello, se debe tomar la predicción más probable de cada modelo y unirla a la probabilidad del modelo siguiente hasta completar los tres modelos (Fig. 7.1).

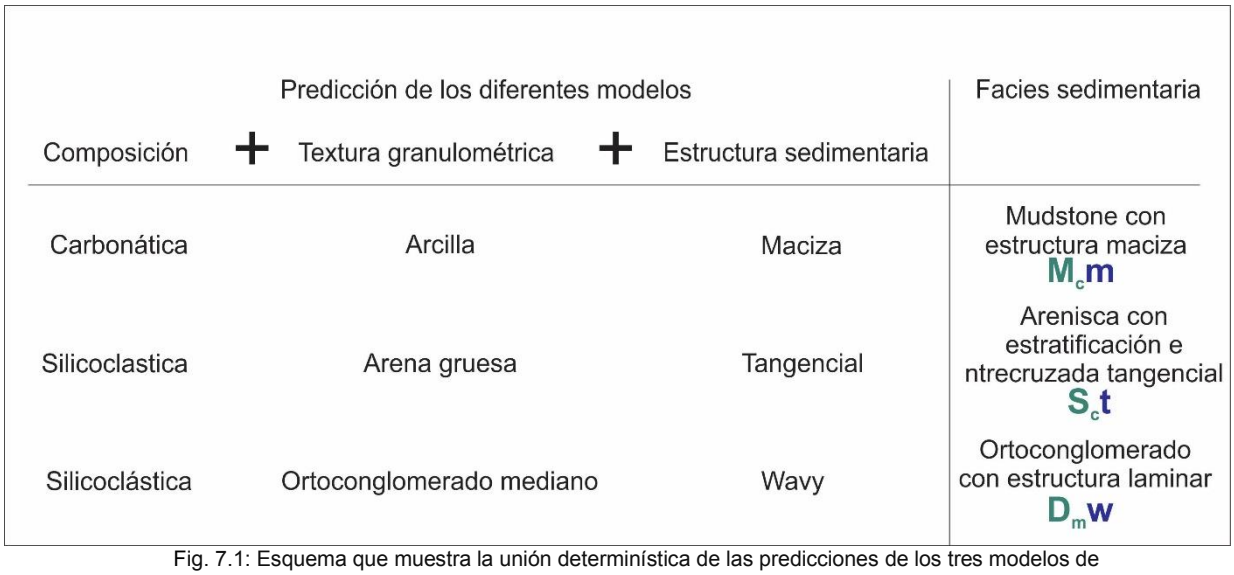


Fig. 7.1: Esquema que muestra la unión determinística de las predicciones de los tres modelos de predicción.

Así como esta forma de unir las predicciones es una forma rápida y sencilla es también, la que contiene mayor potencial de error dado que si un modelo falla en su predicción más probable, puede generar una predicción final de facies sedimentaria que naturalmente puede no formarse. Por ejemplo, una facies que esté formada por un ortoconglomerado mediano, de composición silicoclástica y estructura *wavy* (Fig. 7.1). La estructura *wavy* es una estructura propia de granulometrías finas (desde arcillas hasta arenas finas hasta medianas), en ambientes de energía moderada, mientras que para que se deposite un ortoconglomerado mediano, la energía que debe tener el ambiente es mayor. Además, el tamaño de los granos de un ortoconglomerado no permitiría el proceso físico de la formación de esta estructura de una escala menor al tamaño de grano.

Otra manera de unir las predicciones de los modelos es a partir de un método más probabilístico que tenga en cuenta las probabilidades condicionadas entre las distintas características, ya que las características están condicionadas por los procesos sedimentarios que tienen lugar en la superficie terrestre. Si bien como se mencionó anteriormente los modelos son

independientes entre sí, la relación entre las distintas características de las variables predichas no lo son. Un ejemplo de esto es que la probabilidad de aparición de una estructura por sobre otra, se encuentra condicionada por el tamaño del sedimento que era transportado en ese momento, como ocurre en el ejemplo de la textura *wavy* (Fig. 7.1). Es por lo que, en el marco de este trabajo de tesis, se definió las probabilidades condicionadas entre las distintas características que definen a las facies sedimentarias como:

$$P(facies) = P(textura|composición) * P(estructura|textura)$$

siendo

$$P(textura|composición) = \frac{P(textura) * P(\ composición) * P(textura \cap composición)}{P(composición)}$$

$$P(estructura|textura) = \frac{P(estructura) * P(\ textura) * P(estructura \cap textura)}{P(textura)}$$

En otras palabras, la formula anterior, define que la probabilidad de ocurrencia de una facies sedimentaria concreta está dada por la probabilidad de ocurrencia de una cierta textura sabiendo la composición de la roca, multiplicada por la probabilidad de ocurrencia de una estructura sabiendo cual es la textura de la roca. De esta manera, la variable más independiente y que de cierta manera condiciona al resto de las características geológicas en la unión de los modelos es la composición química y/o mineral de la roca. Este enfoque probabilístico, tiene la ventaja de que brinda una probabilidad de ocurrencia y no un único resultado determinístico. Además, es poco probable que la facies resultante con mayor probabilidad sea una facies errónea (Fig. 7.1). La probabilidad de intersección entre las características geológicas de las facies sedimentarias que son poco probables de generarse a partir de procesos físicos sedimentológicos tendrá una probabilidad de intersección tendiente a cero.

Si bien ambos enfoques tienen pro y contras, la validación de que metodología es mejor dependerá del conjunto de datos y de la calidad de las clasificaciones originales con los que fueron entrenados los modelos. Estas validaciones formarán parte de futuros trabajos.

6

Conclusiones

El presente trabajo ha demostrado el potencial de la aplicación de algoritmos de aprendizaje automático en la caracterización y análisis de facies sedimentarias a partir de datos geológicos tales como perfiles de resistencia al rayado, datos geoquímicos y análisis de imágenes de testigos coronas de roca. El enfoque digital de la descripción redujo la subjetividad en la clasificación, permitió la comparación objetiva entre diferentes muestras y generó una base de datos estructurada que podrá ser utilizada en futuros estudios.

Uno de los principales logros en esta investigación ha sido sortear el problema de la falta de recursos computacionales para el análisis directo de imágenes geológicas a partir de la automatización en la extracción de patrones. La transformación de imágenes en matrices de datos continuos permitió la identificación de texturas y estructuras sedimentarias de manera objetiva y replicable.

El preprocesamiento de datos geoquímicos fue un aspecto crucial en la investigación. La aplicación de transformaciones permitió tratar adecuadamente los datos y corregir los sesgos inherentes a la naturaleza composicional de los datos geoquímicos. Estas transformaciones resultaron fundamentales para extraer patrones de información sin la influencia de restricciones del simplex constante, permitiendo un análisis más preciso de la variabilidad geoquímica en las facies sedimentarias. Asociado a esto, la utilización de algoritmos de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA), proporcionó una representación más eficiente de la información, optimizando la interpretación de los datos sin perder información relevante.

En cuanto a la modelización de facies mediante algoritmos supervisados y no supervisados, se identificaron importantes limitaciones derivadas de la complejidad inherente del concepto de facies sedimentarias. La alta dimensionalidad del problema, junto con la colinealidad entre categorías y la dificultad para obtener conjuntos de datos balanceados, dificultó la clasificación directa de facies. Aunque se implementaron técnicas de balanceo manual para aumentar la representatividad de las clases, la diversidad de combinaciones posibles dentro de los datos geológicos impidió la generación de un modelo completamente generalizable.

Los modelos desarrollados a partir de la simplificación del problema, demostraron ser herramientas de gran utilidad en la descripción y análisis de facies sedimentarias. Se utilizó el algoritmo RF con una exactitud del 65% en el caso del para la textura granulométrica y la estructura sedimentaria y con una exactitud del 76% para la composición química de la roca. Si bien no pueden reemplazar la experiencia del geólogo en la interpretación de ambientes sedimentarios, permiten reducir la subjetividad en la clasificación y optimizar los tiempos de análisis.

La combinación de técnicas de análisis de imágenes con modelos estadísticos abrió nuevas posibilidades para la automatización de la interpretación geológica, optimizando los tiempos de análisis y mejorando la toma de decisiones en proyectos de exploración y producción de recursos naturales. La investigación sugiere la necesidad de ampliar la base de datos con información adicional que incluya una mayor diversidad de facies y una mejor representación estadística de los diferentes tipos de roca. Futuras líneas de trabajo podrían explorar la aplicación de redes neuronales y modelos de *deep learning* en el análisis de imágenes geológicas.

Si bien este método de análisis demostró ser eficiente en la descripción de facies, disminuyendo la subjetividad en la interpretación geológica y permitiendo la comparación sistemática entre diferentes rocas, el rápido avance de la Inteligencia Artificial hace que esté trabajo tenga muchas oportunidades de mejoras. Actualmente, durante la redacción de la presente tesis, la tarea de predicción de facies sedimentarias puede realizarse a partir de modelos más

avanzados que optimizan los recursos y el tiempo de análisis. Por este motivo la tesis se enfocó fuertemente en la metodología de trabajo y procesamiento de datos que constituyen el esqueleto del análisis de datos aplicados, más allá de los modelos utilizados para la clasificación, los cuales evolucionan continuamente.

7

Bibliografía

Ai, X., Wang, H., Sun, B., (2019). Automatic Identification of Sedimentary Facies Based on a Support Vector Machine in the Aryskum Graben, Kazakhstan. *Applied Sciences, 9(21), 4489. https://doi.org/10.3390/app921448*

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *44*(2), 139-160.

Abeliuk, A., & Gutiérrez, C. (2021). Historia y evolución de la inteligencia artificial. *Revista Bits de Ciencia, (21), 14-21.*

Aggarwal, C. C. (2018). Neural networks and deep learning *(Vol. 10, No. 978, p. 3). Cham: Springer.*

Al-Anazi, A., & Gates, I. D. (2010). A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs. *Engineering Geology, 114(3-4), 267-277.*

Ali, M., Zhu, P., Jiang, R., Huolin, M., Ashraf, U., Zhang, H., & Hussain, W., (2024). Data-driven lithofacies prediction in complex tight sandstone reservoirs: A supervised workflow integrating clustering and classification models. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources, 10(1), 70. https://doi.org/10.1007/s40948-024-00787-5*

Aliyuda, K. & Howell, J. (2019). Machine-Learning Algorithm for Estimating Oil-Recovery Factor using a Combination of Engineering and Stratigraphic Dependent Parameters. *Interpretation, 7(3), SE151-SE159. https://doi.org/10.1190/INT-2018-0211.1*

Aliyuda, K., Howell, J., & Humphrey, E. (2020). Impact of Geological Variables in Controlling Oil-Reservoir Performance: An Insight from a Machine-Learning Technique. *SPE Reservoir Evaluation & Engineering, 23 (4), 1314-1327. https://doi.org/10.2118/201196-PA.*

Allen, J. R. L., (1963). The classification of cross-stratified units, with notes on their origin. *Sedimentology, 2, 93-114.*

Allen, J.R.L. (1982). Sedimentary Structures: Their Character and Physical Basis. *Elsevier Scientific Publishing Company.*

Allen, J. R. L. (1984). Experiments on the settling, overturning and entrainment of bivalve shells and related models. *Sedimentology, 31(2), 227-250.*

Allen, J.R. L., (1985). Principles of Physical Sedimentology. *Springer*

Alperin, M., (2013). Introducción al análisis estadístico de datos geológicos. *Editorial de la Universidad Nacional de La Plata.*

Álvarez Trentini, G. & Schwarz, E. (2016). Ciclos carbonáticos-silicoclásticos de alta frecuencia (Fm. Mulichinco, Cuenca Neuquina Central, Argentina): Identificación de cambios de la productividad carbonática en el tiempo y el espacio. VII Congreso Latinoamericano de Sedimentología y XV Reunión Argentina de Sedimentología. La Pampa, Argentina. ISBN: 978-987-42-2083-7. https://sedimentologia.org.ar/ras/XVRAS_libro.pdf

Álvarez Trentini, G., Schwarz, E., Moscariello, A., & De Haller, A. (2018). Nueva metodología de integración entre petrografía óptica avanzada y sistemas QEMSCAN®: su aplicación en sistemas carbonáticos-silicoclásticos. XVI Reunión Argentina de Sedimentología, Rio Negro, Argentina. https://sedimentologia.org.ar/ras/XVIRAS.pdf

Anderson, R.S., (1987). A theoretical model for aeolian impact ripples. *Sedimentology, v. 34, p. 943-956.*

Anderson, R. S., & Hallet, B. (1986). Sediment transport by wind: toward a general model. *Geological Society of America Bulletin, 97(5), 523-535.*

Andreotti, B., Claudin, P., & Pouliquen, O., (2006). Aeolian sand ripples: experimental study of fully developed states. *Physical Review Letters, v. 96, 028001, 4 p.*

Arche, A., (2010). Sedimentología: Del proceso físico a la cuenca sedimentaria (*Vol. 46). Editorial CSIC-CSIC Press.*

Ashley, G.M., (1990). Classification of large-scale subaqueous bed forms: a new look at an old problem. *Journal of Sedimentary Petrology, v. 60, p. 160-172.*

Bagnold, R. A. (1941). The Physics of Blown Sand and Desert Dunes. *Chapman & Hall, 265 p.*

Barandela, R., Sánchez, J. S., Garcıa, V., & Rangel, E. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition, 36(3), 849-851.*

Barrett, P.J., (1980). The Shape of rock particles, a critical review. *Sedimentology 27: 291-304.*

Bashir, R. N., Mzoughi, O., Shahid, M. A., Alturki, N., & Saidani, O. (2024). Principal Component Analysis (PCA) and feature importance-based dimension reduction for Reference Evapotranspiration (ET0) predictions of Taif, Saudi Arabia. C*omputers and Electronics in Agriculture, 222, 109036.*

Bárdossy, A., & Kundzewicz, Z. W. (1990). Geostatistical methods for detection of outliers in groundwater quality spatial fields. *Journal of Hydrology, 115(1-4), 343-359.*

Bernard, H. A., Le Blanc, R. J. y Major, C. F., (1962). Recent and Pleistocene geology of Southwest Texas. *En H. A. Bernard (ed.), Geology of the Gulf Coast of Central Texas, Houston Geological Society, Houston, 175-225.*

Berrezueta, E. & Kovacs, T., (2017). Application of optical image analysis to the assessment of pore space evolution after CO2 injection in sandstones. A case study. *Journal of Petroleum Science and Engineering 159: 679-690.*

Berrezueta, E.; Domínguez-Cuesta, M.J.; Rodríguez-Rey, Á., (2019). Semi-automated procedure of digitalization and study of rock thin section porosity applying optical image analysis tools. *Computers & Geosciences 124: 14-26.*

Bishop C.M., (2006). Pattern recognition and machine learning, *Editorial Springer.*

Blair, T. C., & McPherson, J. G. (1999). Grain-size and textural classification of coarse sedimentary particles. *Journal of Sedimentary Research, 69(1), 6-19.*

Blott, S. J., & Pye, K. (2008). Particle shape: a review and new methods of characterization and classification. *Sedimentology, 55(1), 31-63.*

Booker, N. K., Knights, P., Gates, J. D., & Clegg, R. E. (2022). Applying principal component analysis (PCA) to the selection of forensic analysis methodologies. *Engineering Failure Analysis, 132, 105937.*

Bossi, G.E. (2007). Análisis de Paleocorrientes. *Editorial Magna, San Miguel de Tucumán, 200 pp.*

Bouma, A. H. (1962) Sedimentology of some flysch deposits: a graphic interpretation of depositional systems. *Elsevier, Amsterdam.*

Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., (1984). Classification and Regression Trees (1st Edition). *Editorial Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470*

Bromley, R.G. (1996). Trace fossils: biology, taphonomy and applications. *Chapman and Hall, London, 361 pp.*

Brookfield, M. E.& Silvestro, S. (1992). Eolian System. In Facies Models 4. GEOtext 6, *Geological Association of Canada, St. John's, Newfoundland: 233-264.*

Brookfield, M. E., & Ahlbrandt, T. S. (2000). Eolian sediments and processes (Vol. 38). *Elsevier*.

Buatois, L. A., & Mángano, M. G. (2011). Ichnology: Organism-substrate interactions in space and time. *Cambridge University Press.*

Budennyy, S.; Pachezhertsev, A.; Bukharev, A.; Erofeev, A.; Mitrushkin, D.; Belozerov, B. (2017). Image Processing and Machine Learning Approaches for Petrographic Thin Section Analysis. *In Day 2 Tue, October 17, 2017: SPE: D023S014R005. Moscow, Russia.*

Cai, H., Hu, Y., Zhang, L., Su, M., Yuan, C., & Zhao, Y., (2023). Deep Learning Logging Sedimentary Microfacies via Improved U-Net. *Applied Sciences, 13(19), 10862. https://doi.org/10.3390/app131910862*

Cas, R. A. F., & Wright, J. V., (1987). Volcanic successions: Modern and ancient. *London. Allen & Unwin.*

Catuneanu, O., (2006). Principles of sequence stratigraphy. *Editorial Elsevier.*

Cheel, R.J., (1990). Horizontal lamination and the sequence of bed phases and stratification under upper-flow-regime conditions. *Sedimentology, v. 37, p. 517-529.*

Cheel, R.J., (2005). Introduction to Clastic Sedimentology. *Department of Earth Sciences Ontario University, Canada, 134 pp.*

Chen, T. & Guestrin, C., (2016). XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794).*

Cimarra Muñoz, Daniel (2018). Experimentos de predicción con Gradient Boosting y Random Forest. Proyecto Fin de Carrera / Trabajo Fin de Grado, E.T.S.I. Industriales (UPM).

Clark, M.W., (1981). Quantitative shape analysis: a review. *Journal of Mathematical Geology. 13, 303-320.*

Collinson, J. D., Mountney, N., Thompson, D. B., (2006). Sedimentary structures (3rd Edition). *Editorial Terra publications.*

Cortes, C., & Vapnik, V. (1995). Support-vector networks .Machine Learning, 20 (3), 273-297. http://doi.org/10.1007/BF00994018

Cox, D.R. (1958) The regression analysis of binary sequences (with discussion). J. Roy. Stat. *Soc. B (Methodol.) 20(2), 215–242.*

Craigie, N., (2018). Principles of Elemental Chemostratigraphy. *Editorial Springer International Publishing. https://doi.org/10.1007/978-3-319-71216-1*

Criminisi, A., Shotton, J., Konukoglu, E., (2011). Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114, 5(6), 12.*

Dangeti P., (2017). Statistics for Machine Learning. *Editorial Packt.*

Davis, J. C., & Sampson, R. J. (1986). Statistics and data analysis in geology *(Vol. 646). New York: Wiley.*

De Caritat, P., & Grunsky, E. C. (2013). Defining element associations and inferring geological processes from total element concentrations in Australian catchment outlet sediments: multivariate analysis of continental-scale geochemical data. *Applied Geochemistry, 33, 104-126.*

De Laco, S., Hristopulos, D. T., & Lin, G. (2022). Geostatistics and machine learning. *Mathematical Geosciences, 54(3), 459-465.*

De Raaf, J.F.M., Reading, H.G. & Walker, R.G., (1965). Cyclic sedimentation in the Lower Westphalian of North Devon, England. *Sedimentology, v. 4, p. 1-52*

Demyanov, V., Reesink, A. J. H., & Arnold, D. P. (2019). Can machine learning reveal sedimentological patterns in river deposits?. *Geological Society, London, Special Publications*, *488*(1), 221-235.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) 4171-4186.*

Douzas, G., Bacao, F., Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences, 465, 1-20.*

Dramsch JS (2020). 70 years of machine learning in geoscience in review. *Adv Geophys 61:1-55*

Dunham, R. J., (1962). Classification of carbonate rocks according to depositional textures. *In: Ham, W.E.(Ed). American Association of Petroleum Geologists Memory, 1, 108-121*

Eftekhari, S. H., Memariani, M., Maleki, Z., Aleali, M., Kianoush, P., Shirazy, A., Shirazi, A., & Pour, A. B. (2024). Employing Statistical Algorithms and Clustering Techniques to Assess Lithological Facies for Identifying Optimal Reservoir Rocks: A Case Study of the Mansouri Oilfields, SW Iran. *Minerals, 14(3), 233. https://doi.org/10.3390/min14030233*

Ehrlich, R.; Kennedy, S.K., Crabtree, S.J., Cannon, R.L. (1984). Petrographic Image Analysis, I. Analysis of Reservoir Pore Complexes. *Journal of Sedimentary Petrology. 54 (4): 1365-1378.*

Embry, A. F. & Klovan, J. E., (1971). A late Devonian reef tract on northeastern Banks Island, NWT. *Bulletin of Canadian Petroleum Geology, 19(4), 730-781.*

Ertel, W., (2017). Introduction to Artificial Intelligence. *Editorial Springer. https://doi.org/10.1007/978-3-319-58487-4*

Evans, G., (1965). Intertidal flat sediments and their environments of deposition in the Wash. Q. J. *Geol. Soc. London, 121, 209-245.*

Evans, G.; Schmidt, V.; Bush, P., Nelson, H., (1969). Stratigraphy and geologic history of the sebkha, Abu Dahbi, Persian Gulf. *Sedimentology, 12, 145-159.*

Fazio, A. M., Scasso, R. A., Castro, L. N., Carey, S. (2007). Geochemistry of rare earth elements in early-diagenetic miocene phosphatic concretions of Patagonia, Argentina: Phosphogenetic implications. *Deep Sea Research Part II: Topical Studies in Oceanography*, *54*(11-13), 1414-1432.

Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers & geosciences, 31(5), 579-587.*

Filzmoser, P., Maronna, R., Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis. 52(3), 1694-1711.*

Friedman, G.M., (1979). Address of the retiring President of the International Association of Sedimentologists: differences in size distributions of populations of particles among sands of various origins. *Sedimentology 26, 3–32.*

Friedman, J. H., Hastie, T., Tibshirani, R., (2000). Additive Logistic Regression: A Statistical View of Boosting. *The Annals of Statistics, 28(2), 337-407.*

Friedman, J. H., (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics, 29(5), 1189-1232.*

Fisher, R. V. (1961). Proposed classification of volcaniclastic sediments and rocks. *Geological Society of America Bulletin, 72(9), 1409-1414.*

Fisher, R. V., & Schmincke, H.-U. (1984). Pyroclastic Rocks. Berlin. *Springer-Verlag.*

Fisher, R. & Smith, G.A., (1991) (Eds.) Sedimentation in volcanic settings. SEPM Special Publication, 45.

Folk, R. L. (1959). Practical petrographic classification of limestones. *AAPG Bulletin, 43(1), 1-38.*

Folk, R. (1980). Petrology of sedimentary rocks. *Hemphill. Austin, Texas. USA.: 182 pp.*

Gao, S. & Collins, M., (1991). A Critique of the "McLaren Method" for Defining Sediment Transport Paths: DISCUSSION. *Journal of Sedimentary Research, 61(1).*

García, S. Ramírez-Gallego, S, Luengo J & Herrera, F. (2016). Big Data: Preprocesamiento y calidad de datos. *novática*, *237*(1), 17-20.

Germay, C., Richard, T., Mappanyompa, E., Lindsay, C., Kitching, D., & Khaksar, A. (2015). The continuous-scratch profile: a high-resolution strength log for geomechanical and petrophysical characterization of rocks. *SPE Reservoir Evaluation & Engineering*, *18*(03), 432-440.

Germay, C., Lhomme, T., & Perneder, L. (2023). High-resolution core data and machine learning schemes applied to rock facies identification and classification. *Geological Society of London Special Publications*, *527*(1), SP527-2021.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. " O'Reilly Media, Inc." 864p. ISBN 978-1-09-812246-1. URI: http://10.250.8.41:8080/xmlui/handle/123456789/16253

Glaister, R. P., & Nelson, H. W. (1974). Grain-size distributions, an aid in facies identification. *Bulletin of Canadian Petroleum Geology, 22(3), 203-240.*

Gómez, J. L., & Camilion, E. (2025). Improving the classification of rock samples with diffusion probabilistic machine learning. *Interpretation*, *13*(1), T153-T162.

Gonçalves, Í. G., Kumaira, S., & Guadagnin, F. (2017). A machine learning approach to the potential-field method for implicit modeling of geological structures. *Computers & Geosciences, 103, 173-182. https://doi.org/10.1016/j.cageo.2017.03.015*

Gonzalez, R. C., Woods, R. E., (2009). Digital Image Processing, (3th Edition). *Pearson education india. https://doi.org/10.1117/1.3115362*

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). *Cambridge: MIT press.*

Gressly, A. (1883). Observations géologiques sur le Jura Solenois. N. Denk. allg. *Schweiz Ges. Ges. Naturz., 2:1-112*

Grimm, R., Behrens, T., Märker, M. & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on Barro Colorado Island—digital soil mapping using Random Forests analysis. *Geoderma 146, 102–113 (2008)*

Grunsky, E. C., & de Caritat, P. (2017). Advances in the use of geochemical data for mineral exploration. *In Proceedings of exploration (Vol. 17, pp. 441-456).*

Guo, Q., Wu, W., Massart, D. L., Boucon, C., & de Jong, S. (2002). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems, 61(1-2), 123-132.*

Hall-Beyer, M., (2017). GLCM texture: a tutorial. National Council on Geographic Information and Analysis Remote Sensing Core *Curriculum, 3(1), 75.*

Halotel, J., Demyanov, V., & Gardiner, A. (2020). Value of Geologically Derived Features in Machine Learning Facies Classification. *Mathematical Geosciences, 52(1), 5-29. https://doi.org/10.1007/s11004-019-09838-0*

Han, W., Fu, Z., Xiao, S., Zheng, X., Huang, X., Wang, Y., ... & Yan, D. (2024). Dual-model collaboration consistency semi-supervised learning for few-shot lithology interpretation. *IEEE Transactions on Geoscience and Remote Sensing*.

Harrington, P., (2012). *Machine learning in action. Jeff Bleiel (Ed.) Manning. ISBN 9781617290183.*

Harris, J. R., & Grunsky, E. C. (2015). Predictive lithological mapping of Canada's North using Random Forest classification applied to geophysical and geochemical data. *Computers & Geosciences, 80, 9-25. https://doi.org/10.1016/j.cageo.2015.03.013*

Hastie, T., Tibshirani, R., Friedman, J., (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction (2nd Edition). Springer. 737 pp.*

Hathon, L.A.; Taylor, T.R.; Rambow, F.H.K.; Myers, M.T.; Fanning, D.W. (2003). Rock Properties from 2D Images: Computer Assisted Petrography. *AAPG International Conference, September 21-24, Barcelona, Spain. Article #90017.*

Helwig, Z. D., Guggenberger, J., Elmore, A. C., & Uetrecht, R. (2019). Development of a variogram procedure to identify spatial outliers using a supplemental digital elevation model. *Journal of Hydrology X, 3, 100029.*

Henne, A., Noble, R. R., & Williams, M. (2024). Multi-element geochemical analyses on ultrafine soils in Western Australia–towards establishing abundance ranges in mineral exploration settings. *Geochemistry: Exploration, Environment, Analysis, 24(1), geochem2023-043.*

Holden, T., Kurian, R., Ibrahim, M., Hampson, D., & Downton, J. (2023). Predicting Facies, Rock, and Geomechanical Properties Using Convolutional Neural Networks: A Case Study From an Unconventional Shale Reservoir. Proceedings of the 11th Unconventional Resources Technology Conference. Unconventional Resources Technology Conference, Colorado Convention Center, Denver, Colorado, US. https://doi.org/10.15530/urtec-2023-3862247

Hopkins, C.G., (1899). A plea for a scientific basis for the division of soil particles in mechanical analysis. U.S. *Departamen of Agriculture Bulletin 56: 64:66.*

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24(6), 417.*

Houghton, J., Nichols, T., Griffiths, J., Simon, N., Utley, J., Duller, R., & Worden, R. (2023). Automated Classification of Estuarine Sub-Depositional Environment Using Sediment Texture. *Journal of Geophysical Research: Earth Surface, 128(2), e2022JF006891.*

Hussain, M., Liu, S., Ashraf, U., Ali, M., Hussain, W., Ali, N., & Anees, A. (2022). Application of Machine Learning for Lithofacies Prediction and Cluster Analysis Approach to Identify Rock Type. *Energies, 15(12), 4501. https://doi.org/10.3390/en15124501*

Illés, T., & Nagy, M. (2007). A Mizuno–Todd–Ye type predictor–corrector algorithm for sufficient linear complementarity problems. *European Journal of Operational Research, 181(3), 1097-1111.*

Insua, T. L., Hamel, L., Moran, K., Anderson, L. M., & Webster, J. M. (2015). Advanced classification of carbonate sediments based on physical properties. *Sedimentology, 62(2), 590-606. https://doi.org/10.1111/sed.12168*

Ippolito, M., Ferguson, J., & Jenson, F. (2021). Improving facies prediction by combining supervised and unsupervised learning methods. *Journal of Petroleum Science and Engineering, 200, 108300. https://doi.org/10.1016/j.petrol.2020.108300*

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, No. 1). New York: springer.

Jensen J.R., (1996). Introductory Digital Image Processing (2nd Edition), *Editorial Prentice Hall, pp 316.*

Jiang, Y., Li, X., Luo, H., Yin, S., Kaynak, O., (2022). Quo vadis artificial intelligence? *Discover Artificial Intelligence, 2(1), 4. https://doi.org/10.1007/s44163-022-00022-8.*

Kanevski, M., Parkin, R., Pozdnukhov, A., Timonin, V., Maignan, M., Demyanov, V., & Canu, S. (2004). Environmental data mining and modeling based on machine learning algorithms and geostatistics. *Environmental Modelling & Software, 19(9), 845-855.*

Kanevski, M., Timonin, V., & Pozdnukhov, A. (2009). Machine learning for spatial environmental data: theory, applications, and software. *EPFL press*.

Kanevski, M., & Demyanov, V. (2015). Statistical learning in geoscience modelling: novel algorithms and challenging case studies. *Computers and Geosciences, 85, 1-2.*

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2019). Machine Learning for the Geosciences: Challenges and Opportunities*, IEEE T. Knowl. Data En., 31, 1544–1554.*

Kazak, A.; Simonov, K.; Kulikov, V. (2021). Machine-Learning-Assisted Segmentation of Focused Ion Beam-Scanning Electron Microscopy Images with Artifacts for Improved Void-Space Characterization of Tight Reservoir Rocks. *SPE Journal 26 (04): 1739-1758.*

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T., (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *En Advances in Neural Information Processing Systems (pp. 3149-3157).*

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S., & Raazi, S. M. K. U. R. (2021). Automated Prediction of Good Dictionary Examples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity, 2021(1), 2553199.*

Kidwell, S. M., & Holland, S. M. (1991). Field description of coarse bioclastic fabrics. *PALAIOS, 6(4), 426-434.*

Koeshidayatullah, A., Morsilli, M., Lehrmann, D.J., Al-Ramadan, K., Payne, J.L. (2020). Fully automated carbonate petrography using deep convolutional neural networks. *Marine and Petroleum Geology 122, 104687. https://doi.org/10.1016/j.marpetgeo.2020. 104687.*

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (pp. 1137-1143).*

Krumbein, W.C., (1934). Size frequency distributions od sediments, *Journal of Sedimentary Petrology 4:65-77.*

Krumbein, W. (1941). Measurement and geologic significance of shape and roundness of sedimentary particles. *Journal of Sedimentary Petrology*, *11*, 64-72.

Lalor,G.C., Zhang,C. (2001). Multivariate outlier detection and remediation in geochemical databases. *Science of The Total Environment. 281 (1–3), 99-109.*

Larrea, M.L.; Castro, S.M.; Bjerg, E.A., (2014). A software solution for point counting. Petrographic thin section analysis as a case study. *Arabian Journal of Geosciences. 7 (8): 2981-2989.*

Larriestra, C. (2013). Soft Inorganic Geochemistry: A New Concept for Unconventional Resources Modeling. Search and Discovery Article #80311 (2013), Poster presentation at AAPG Annual Convention and Exhibition, Pittsburgh, Pennsylvania, May 19-22, Posted August 26, 2013. 18 pp.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers, 7(1), 3*-10.

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, *Proceedings 8 (pp. 63-66). Springer Berlin Heidelberg.*

Leclair, S.F., 2002. Preservation of cross-strata due to the migration of subaqueous dunes: an experimental investigation: *Sedimentology, v. 49, p. 1157-1180.*

Leeder, M. R. (1982). Sedimentology: process and product. Springer Science & Business Media. *First Edition.Chapman & Hall, 2-6 Boundary Row, London SE1 8HN, UK. ISBN-13: 978-0-412-53300-6. e-ISBN-13: 978-94-009-5986-6. DOI: 10.1007/978-94-009-5986-6*

Leyrit, H. & Montenat, C., (2000). *(Eds.) Volcaniclastic rocks, from magmas to sediments. Gordon & Beach Science Publishers*

Li, Z., Kang, Y., Feng, D., Wang, X.-M., Lv, W., Chang, J., & Zheng, W. X. (2020). Semi-supervised learning for lithology identification using Laplacian support vector machine. *Journal of Petroleum Science and Engineering, 195, 107510. https://doi.org/10.1016/j.petrol.2020.107510*

Liu, S. (2024). A Grain Size Profile Prediction Method Based on Combined Model of Extreme Gradient Boosting and Artificial Neural Network and Its Application in Sand Control Design. *SPE Journal, 29(06), 2988-3002. https://doi.org/10.2118/219484-PA*

Liu, X. Y., Wu, J., Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2), 539-550.*

Lokier, S. W., Al Junaibi, M., (2016). The petrographic description of carbonate facies: Are we all speaking the same language? *Sedimentology, 63(7), 1843-1885. https://doi.org/10.1111/sed.12293*

Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B.,Katz, R, Himmelfarb J., & Lee, S.I. (2020). From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence, 2(1), 56–67.*

Luo, J., Frisken, S., Machado, I., Zhang, M., Pieper, S., Golland, P., ... & Wells, W. M. (2018). Using the variogram for vector outlier screening: application to feature-based image registration. *International journal of computer assisted radiology and surgery, 13, 1871-1880.*

Madsen, A., Reddy, S., & Chandar, S. (2022). Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys, 55(8), 1-42.*

Maitre, J., Bouchard, K., Bédard, P.L. (2019). Mineral grains recognition using computer vision and machine learning. *Computers and Geosciences 130: 84-93.*

Maerz, S. (2019). Analyzing pore systems through comprehensive digital image analysis (DIA): quantifying pore type geometry, detecting effective porosity and reconstructing pore system evolution. *PhD Thesis, Universität Potsdam: 143 p.*

Mancini, M., Weindorf, D.C., Monteiro, M.E.C., de Faria, A.J.G, Teixeira, A.F.S, de Lima, W, Dias de Lima, F.R, Dijair, T.S.B, D'Auria Marquesa, F., Ribeiroa, D., Silva, S.H.G , Chakrabortyd, S., Curi,N. (2020). From sensor data to Munsell color system: Machine learning algorithm applied to tropical soil color classification via Nix™ Pro sensor. *Geoderma, 375, 114471. https://doi.org/10.1016/j.geoderma.2020.114471*

Marques, W. S., Sial, A. N., De Albuquerque Menor, E., Ferreira, V. P., Freire, G. S. S., De Albuquerque Medeiros Lima, E., Do Amaral Vaz Manso, V., (2008). Principal component analysis (PCA) and mineral associations of litoraneous facies of continental shelf carbonates from northeastern Brazil. *Continental Shelf Research, 28(20), 2709-2717. https://doi.org/10.1016/j.csr.2008.09.005*

Marrón, B. S., (2012). Co-occurrence Matrix and fractal dimension for image segmentation. *Revista de Matemática: Teoría y Aplicaciones, 19(1), 49-63. https://doi.org/10.15517/rmta.v19i1.2104*

Martín-Fernández, J. A., Palarea-Albaladejo, J., Olea, R. A., (2011). Dealing with zeros. Compositional data analysis: Theory and applications, 43-58.

Martin, T., Meyer, R., & Jobe, Z. (2021). Centimeter-Scale Lithology and Facies Prediction in Cored Wells Using Machine Learning. *Frontiers in Earth Science, 9, 659611. https://doi.org/10.3389/feart.2021.659611*

Mazzoni, A. (1985). Las rocas volcánicas: clasificación, descripción y génesis. *Buenos Aires: Asociación Geológica Argentina.*

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*, 115-133.

Mc Laren, P., (1981). An interpretation of trends in grain size measures. *Journal of Sedimentary Petrology 51 (2), 611–624.*

McMahan, H. B., Ramage, D., Talwar, K., & Zhang, L. (2017). Learning differentially private recurrent language models. arXiv preprint arXiv:1710.06963.

McManus, J., (1991). Grain size determination and interpretation. En Tucker, M.(Ed.), "Techniques in Sedimentology" 63-85. Backwell. Oxford.

McPhie, J., Doyle, M., y Allen, R. (1993). Volcanic textures: A guide to the interpretation of textures in volcanic rocks. *Hobart: University of Tasmania*

Mena, M. (2016). *La estadística como herramienta en Ciencias de la Tierra*. *Akadia.1;625. ISBN: 978-987-570-281-3. URI:http://hdl.handle.net/11336/95593*

Merodio, J. C. (1985). *Métodos estadísticos en geología. Asociación Geológica Argentina. Serie B Didáctica y complementaria N° 13. 233 pp.*

Meyer, R. G., Martin, T. P., & Jobe, Z. R. (2020). CoreBreakout: Subsurface core images to depth-registered datasets. *Journal of Open-Source Software, 5(50), 1969.*

Miall, A.D. (1978). Lithofacies Types and Vertical Profile Models en Braided River Deposits: A Summary. In Miall, A.D., Ed., Fluvial Sedimentology, Memoir 5, *Canadian Society of Petroleum Geologists, Calgary, 597-604.*

Miall, A. D. (1978). Tectonic setting and syndepositional deformation of molasse and other nonmarine-paralic sedimentary basins*. Canadian Journal of Earth Sciences, 15(10), 1613-1632.*

Miall, A. D., (1988). Facies Architecture in Clastic Sedimentary Basins. En K. L. Kleinspehn C. Paola (Eds.), New Perspectives in Basin Analysis (pp. 67-81). Springer New York. https://doi.org/10.1007/978-1-4612-3788-4_4

Miall, A. D., & Miall, A. D. (1996). The stratigraphic architecture of fluvial depositional systems. The Geology of Fluvial Deposits: Sedimentary Facies, Basin Analysis, and Petroleum Geology, 251-309.

Miall, A.D. (2022). Stratigraphy: The Modern Synthesis. In: Stratigraphy: A Modern Synthesis (pp. 341-417). Springer Textbooks in *Earth Sciences, Geography and Environment. Springer, Cham. https://doi.org/10.1007/978-3-030-87536-7_7*

Middleton, G. V. (1973). Johannes Walter's Law of Correlation of Facies. *Geol. Soc. Am. Bull., 84,979-988.*

Middleton, G.V., (1978). Facies, in Fairbridge, R.W. and Bourgeois, J.,eds., Encyclopedia of sedimentology: Stroudsburg, *Pennsylvania, Dowden,Hutchinson and Ross, p. 323-325.*

Midtgaard, H. H. (1996). Inner-shelf to lower-shoreface hummocky sandstone bodies with evidence for geostrophic influenced combined flow, Lower Cretaceous, West Greenland. *Journal of Sedimentary Research, 66(2), 343-353.*

Minsky, M., & Papert, S. (1969). An introduction to computational geometry. *Cambridge tiass., HIT, 479(480), 104.*

Mishra, S., Sarkar, U., Taraphder, S., Datta, S., Swain, D., Saikhom, R., Panda, S., Laishram, M., (2017). Principal Component Analysis. *International Journal of Livestock Research, 1. https://doi.org/10.5455/ijlr.20170415115235*

Misra, K. C. (2012). *Introduction to geochemistry: principles and applications.* J*ohn Wiley & Sons. Kula C. Misra (Ed), 464 pp. ISBN 978-1-4443-5095-1 / 978-1-4051-2142-2.*

Molnar, C. (2020). *Interpretable machine learning. A Guide for Making Black Box Models Explainable. ISBN 978-0-244-76852-2. 313 pp. E-book version (epub, mobi) on leanpub.com*

Mosca,E., Wich, M. & Groh, G. (2021). Understanding and interpreting the impact of user context in hate speech detection. *In Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media. p 91–10*

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., & Groh, G. (2022). SHAP-based explanation methods: a review for NLP interpretability. *In Proceedings of the 29th international conference on computational linguistics p. 4593-4603.*

Moss, A. J., (1962). The physical nature of common sandy and pebbly deposits, part I. *American Journal of Science, 260(5), 337-373.*

Müller, A. C., Guido, S., (2016). *Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media, Inc.394 pp.*

Murcia, J., Quidelleur, X., Nomade, S., y Villeneuve, M. (2013). Age constraints on magma genesis and recycling processes in the Central Andes. *Journal of Volcanology and Geothermal Research, 254, 17-35.*

Myrow, P. M., & Southard, J. B. (1996). Tempestite deposition. *Journal of Sedimentary Research, 66(5), 875-887.*

Net, L.I.; Limarino, C.O. (2000). Caracterización y origen de la porosidad en areniscas de la sección inferior del Grupo Paganzo (Carbonífero superior), Cuenca Paganzo, Argentina. *Asociación Argentina de Sedimentología 7 (1-2): 49-72.*

Nichols, T. E., Worden, R. H., Houghton, J. E., Duller, R. A., Griffiths, J., & Utley, J. E. P. (2023). Sediment texture and geochemistry as predictors of sub-depositional environment in a modern estuary using machine learning: A framework for investigating clay-coated sand grains. *Sedimentary Geology, 458, 106530. https://doi.org/10.1016/j.sedgeo.2023.106530*

**Nilsson, N. J., (1998*). Artificial intelligence: a new synthesis. Editorial Morgan Kaufmann. Morgan Kaufmann Publishers, Inc. San Francisco, California.493 pp.*

Oomkens, E. y Terwindt, J.H. J. (1960). Inshore estuarine sediments in the Haringvliet, Netherlands.Geol. *Mijnbouw, 39, 701-710.*

Passega, R., (1964). Grain size representation by C-M patterns as a geological tool. *Journal of Sedimentary Petrology 34: 830-847.*

Pegalajar, M.C., Ruiz, L.G.B., Sánchez-Marañón, M., & Mansilla, L. (2019). A Munsell colour-based approach for soil classification using Fuzzy Logic and Artificial Neural Networks. *Fuzzy Sets and Systems*. Volume 401, 38-54. ISSN 0165-0114. https://doi.org/10.1016/j.fss.2019.11.002

Petrinovic, I. A., & D'Elia, L. (2018). Rocas Volcaniclásticas: Depósitos, Procesos y Modelos de Facies. *Asociación Argentina de Sedimentología, Publicación Especial*, *3*, 184.

Pettijohn, F. J. (1957). Paleocurrents of Lake Superior Precambrian quartzites. *Geological Society of America Bulletin*, *68*(4), 469-480.

Poizot, E., Mear, Y., Thomas, M., & Garnaud, S. (2006). The application of geostatistics in defining the characteristic distance for grain size trend analysis. *Computers & Geosciences*, *32*(3), 360-370.

Poizot, E., Méar, Y., & Biscara, L. (2008). Sediment Trend Analysis through the variation of granulometric parameters: A review of theories and applications. *Earth-Science Reviews*, *86*(1-4), 15-41.

Ponce, J. J., Carmona, N. B., Montagna, A. O., (2018). *Atlas de estructuras sedimentarias inorgánicas y biogénicas: descripción, análisis e interpretación a partir de afloramientos, testigos corona y registro de imágenes de pozo. Ciudad Autónoma de Buenos Aires: Fundación YPF-UNRN 166p.*

Posthoff, C. (2024). *Artificial Intelligence for Everyone. Springer Nature Switzerland.*

Powers, M.C., (1953). A New Roundness Scale for Sedimentary Particles. J*ournal of Sedimentary Research, 23(2), 117-119.*

Presutti, M., (2004). La Matriz de Co-ocurrencia en la clasificación multiespectral: Tutorial para la enseñanza de medidas texturales en cursos de grado universitario. *4ª Jornada de Educação em Sensoriamento Remoto no Âmbito do Mercosul, 11 a 13 de agosto – São Leopoldo, RS, Brasil. p 1-9.*

Petrelli, M. (2024). Machine Learning in Petrology: State-of-the-Art and Future Perspectives. *Journal of Petrology, 65(5), egae036.*

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, Inc. ISBN: 1-55860-238-0. 312 pp.*

Reading, H.G., (1978). *Sedimentary environments and facies. H.G. Reading (Ed.), First Edition. ISBN 978-0-6320-3627-1.*

Reading, H.G., (2009). *Sedimentary environments: Processes, facies and stratigraphy. H.G. Reading (Ed.), Third Edition. John Wiley & Sons. ISBN: 144431369X, 9781444313697. 704 pp.*

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature, 566(7743), 195-204.*

Richa; Mukerji, T.; Mavko, G.; Keehm, Y. (2006). Image analysis and pattern recognition for porosity estimation from thin sections. In SEG Technical Program Expanded Abstracts 2006: *Society of Exploration Geophysicists: 1968-1972*

Richard, T., Dagrain, F., Poyol, E., & Detournay, E. (2012). Rock strength determination from scratch tests. *Engineering Geology*, *147*, 91-100.

Robinson, A. G. (2009). *Inorganic geochemistry: applications to petroleum geology. John Wiley & Sons. ISBN 1444313975, 9781444313970. 264 pp.*

Robinson, G.W., (1924). The forms of mechanical composition curves of soils, clays, and other granular substances. *Journal of Agriculture Science 14: 626-633.*

Rollinson, H.R. (1993). *Using Geochemical Data: Evaluation, Presentation, Interpretation (1st ed.). 384 pp. Routledge, London. https://doi.org/10.4324/9781315845548.*

Rosenblatt, F. (1957). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review, Vol. 65, No. 6, pp. 386–408.*

Rubey, W.W., (1930). Lithologic studies of fine-grained Upper Cretaceous sedimentary rocks of the Black Hills region. *U.S. Geological Survey Professional Paper 165A:1-54.*

Rubin, D. M., & Carter, C. L. (1987). Cross-bedding, bedforms, and paleocurrents. *SEPM Society for Sedimentary Geology.*

Rubo, R.A.; de Carvalho Carneiro, C.; Michelon, M.F.; dos Santos Gioria, R. (2019). Digital petrography: Mineralogy and porosity identification using machine learning algorithms in petrographic thin section images. *Journal of Petroleum Science and Engineering, 183: 106382.*

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323(6088), 533-536.*

Russell, S.J., Norvig, P. (2016). *Artificial intelligence: a modern approach. Pearson.* 946 pp.

Scasso, R. A., Limarino, C. O., (1997). *Petrología y diagénesis de rocas clásticas. Asociación Argentina de Sedimentología.Publicación especial N°1. ISBN: 987-96296-0-4.* 253 pp.

Schapire, R. E., (1990). The Strength of Weak Learnability. *Machine Learning, 5(2), 197-227.*

Seilacher, A., (1964). Biogenic sedimentary structures. *In Approaches to Paleoecology, ed. J. Imbrie, and N. Newell, Chichester, UK: John Wiley & Sons, pp. 296–316.*

Selley (1976). *Medios sedimentarios antiguos. H. Blume edic., Madrid: 251 p.*

Shapley, L. (1953). A Value for n-Person Games. In: Kuhn, H. and Tucker, A., Eds., Contributions to the Theory of Games II, *Princeton University Press, Princeton, 307-317. https://doi.org/10.1515/9781400881970-018*

Shearman, D. J. (1966). Origin of marine evaporates by diagenesis. Trans. Inst. *Mineralogy Metallurgy, 75, 208-215.*

Sheridan, M. F., & Marshall, J. R. (1983). Interpretation of pyroclast surface features using SEM images. *Journal of Volcanology and Geothermal Research*, *16*(1-2), 153-159.

Skansi, S., (2018). Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence. *Springer International Publishing. 191 pp. https://doi.org/10.1007/978-3-319-73004-2*

Smith, J. V., & Yoder Jr, H. S. (1956). Experimental and theoretical studies of the mica polymorphs. *Mineralogical Magazine and Journal of the Mineralogical Society, 31(234), 209-235.*

Sohil, F., Sohali, M. U., Shabbir, J., (2022). *An introduction to statistical learning with applications En R: By Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media. https://doi.org/10.1080/24754269.2021.1980261.*

Sokolova, M., Lapalme, G., (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45(4), 427-437.*

Southard, J.B., & Boguchwal, L.A., (1990). Bed configurations in steady unidirectional water flows. Part 2. Synthesis of flume data: *Journal of Sedimentary Petrology, v. 60, p. 658-679.*

Sparks, R. S. J., Bursik, M. I., Carey, S. N., Gilbert, J. S., Glaze, L. S., Sigurdsson, H., & Woods, A. W. (1997). Volcanic Plumes. John Wiley & Sons.

Sun, C., Demyanov, V., & Arnold, D. (2023). GAN River-I: A process-based low NTG meandering reservoir model dataset for machine learning studies. *Data in Brief*, *46*, 108785.

Syvitski, J. P., & Murray, J. W., (1977). Grain-size distribution using log-probability plots: a discussion. *Bulletin of Canadian Petroleum Geology, 25(3), 683-694.*

Taylor, T.R; Lander, R.H; Bonnell, L.M. (2022). Sandstone petrography, petrology, and modeling. *SEPM Concepts in Sedimentology and Paleontology No. 13. SEPM Society for Sedimentary Geology.*

Terruggi, M. E., Puig, C., Vilela, E. (1978). *Rocas volcánicas y sus productos secundarios. Buenos Aires: Eudeba.*

Ting, K. M., (2011). Confusion matrix, in C. Sammut and G. I. Webb, eds., Encyclopedia of machine learning, 1st ed.: *New York, Springer, 209 p*

Tucker, M. E., & Wright, V. P., (2009). Carbonate sedimentology. John Wiley & Sons. ISBN 1444314165, 9781444314168, 496 pp.

Tucker, M. E. & Jones, S.J., (2023). Sedimentary petrology. (4th. Edition) John Wiley & Sons. 4ta Edición, ilustrada. ISBN 1118786491, 9781118786499. 448 pp.

Udden, J.A, (1998). Mechanical composition of wind deposits. Número 1 de Augustana Library publications, Ill. Lutheran Augustana book concern, printers. Universidad de Michigan. ISBN 0910182000, 9780910182003. 69 pp.

Van den Broeck, G., Lykov, A., Schleich, M., & Suciu, D. (2022). On the tractability of SHAP explanations. Journal of Artificial Intelligence Research, 74, 851-886.

Vanta Family, (2024). X-Ray Fluorescence Analyzer: User's Manual (10-040355-01ES - Rev.2) EVIDENT SCIENTIFIC INC. 48 Woerd Avenue, Waltham, MA 02453, EE. UU.

Visher, G.S., (1969). Grain size distribution and depositional processes. *Journal of Sedimentary Petrology 39: 1074-1106.*

Verbovšek, T. (2011). A comparison of parameters below the limit of detection in geochemical analyses by substitution methods *Primerjava ocenitev parametrov pod mejo določljivosti pri geokemičnih analizah z metodo nadomeščanja. RMZ-Materials and Geoenvironment, 58(4), 393-404.*

Wadell, H., (1932). Volume, shape, and roundness of rock particles. *The Journal of Geology 40 (5). 443-451.*

Wadoux, A.M.J.C., Minasny, B., McBratney, A.B., (2020). Machine learning for digital soil mapping: applications, challenges and suggested solutions. *Earth Sci. Rev. 210, 103359*

Wang, D., Peng, J., Yu, Q., Chen, Y., & Yu, H. (2019). Support Vector Machine Algorithm for Automatically Identifying Depositional Microfacies Using Well Logs. *Sustainability, 11(7), 1919. https://doi.org/10.3390/su11071919*

Walker, R. G. (1992). Facies model: response to sea level change. *Geol. Asso. Canada, 409.*

Walker, R.G., (2006). Facies models revised. En: Posamentier, H.W., Walker, R.G. (Eds.), Facies Models, *Society for Sedimentary Geology, Special Publication, 84: 1-18.*

Walter, J. (1884). Einleitung in die geologie als Historische wissenchaft. Fischer, Köln.

Wentworth, C.K., (1922). A scale of grade and class terms for clastics sediments. *Journal of Geology 30: 377-392.*

White, J. D. L., y Houghton, B. F. (2006). Primary volcaniclastic rocks. *Geology, 34(3), 169-172.*

Winklemolen, A.M., (1982). Critical remarks on grain parameters, with special emphasis on shape. *Sedimentology 29:255-265.*

Witten, I. H., Frank, E., Hall, M. A., (2011). *Data mining: Practical machine learning tools and techniques (3rd Edition). Morgan Kaufmann.*

Xie, Y., Zhu, C., Zhou, W., Li, Z., Liu, X., & Tu, M. (2018). Evaluation of machine learning methods for formation lithology identification: A comparison of tuning processes and model performances. *Journal of Petroleum Science and Engineering, 160, 182-193. https://doi.org/10.1016/j.petrol.2017.10.028*

Yu, L., Porwal, A., Holden, E.-J., & Dentith, M. C. (2012). Towards automatic lithological classification from remote sensing data using support vector machines. *Computers & Geosciences, 45, 229-239. https://doi.org/10.1016/j.cageo.2011.11.019*

Zar, J. H. (1999). *Biostatistical analysis. Pearson Education India.*

Zhang, Z., & Cai, Z. (2021). Permeability prediction of carbonate rocks based on digital image analysis and rock typing using random forest algorithm. *Energy & Fuels, 35(14), 11271-11284.*

Zheng, D., Hou, M., Chen, A., Zhong, H., Qi, Z., Ren, Q., You, J., Wang, H. Ma, C. (2022). Application of machine learning in the identification of fluvial-lacustrine lithofacies from well logs: A case study from Sichuan basin, China. *Journal of Petroleum Science and Engineering, 215, 110610. https://doi.org/10.1016/j.petrol.2022.110610*

Zhong, R., Johnson, R., & Chen, Z., (2020). Generating pseudo density log from drilling and logging-while-drilling data using extreme gradient boosting (XGBoost). *International Journal of Coal Geology, 220, 103416. https://doi.org/10.1016/j.coal.2020.103416*

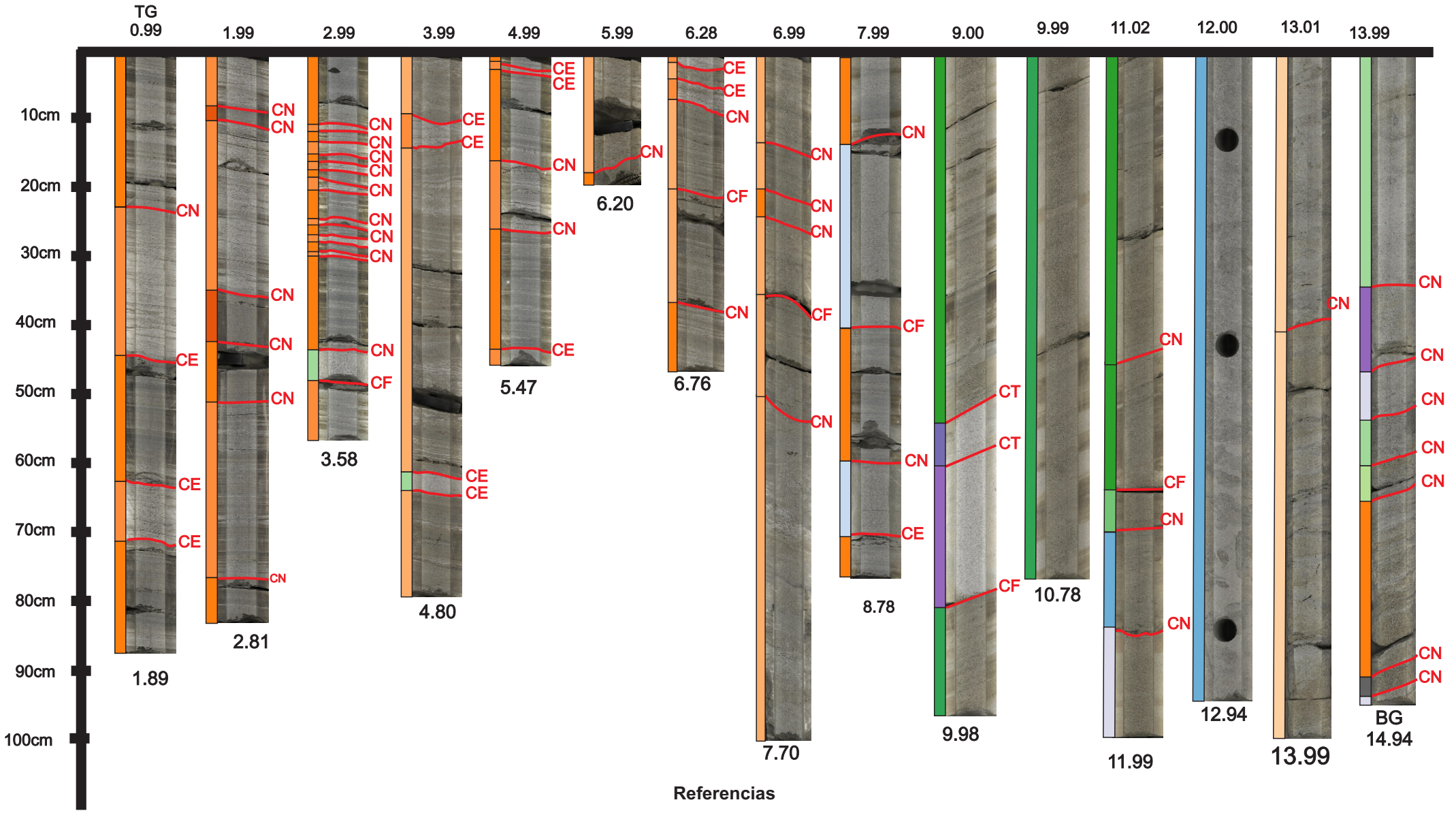
Zhu, X., Goldberg, A.B., (2022). Introduction to semi-supervised learning. *Springer Nature.*

Zoback, M. D. (2007). *Reservoir geomechanics*. Cambridge university press. *490pp.*

ANEXO I

Mosaicos de los testigos coronas sintéticos con las Facies Sedimentarias

Merge Core C01



Referencias

Tipos de Contactos

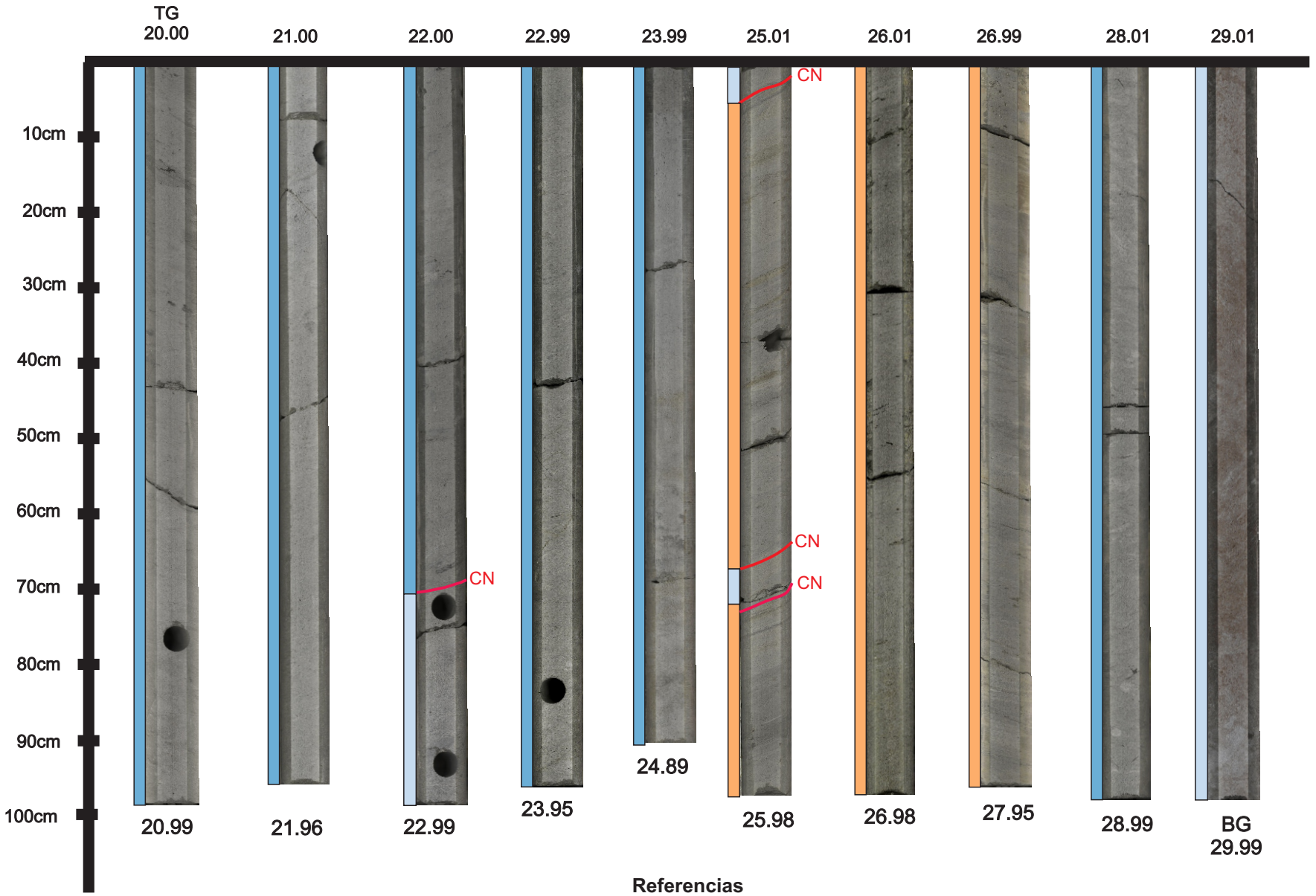
CN	Contactos Netos	CF	Contactos por Fracturas
CT	Contactos Transicionales	CE	Contactos Erosivos

Facies

Htf	Htd	$S_v p$	$S_v m$	$S_l m$	Wsm
Htw	$S_v m$	$S_v h$	Ssb	$S_l p$	$S_m l$
Htb	$S_v ng$	$S_v l$	Ssm	$S_l l$	

BG	Base general
TG	Techo general

Merge Core C02



Referencias

Tipos de Contactos

CN Contactos Netos

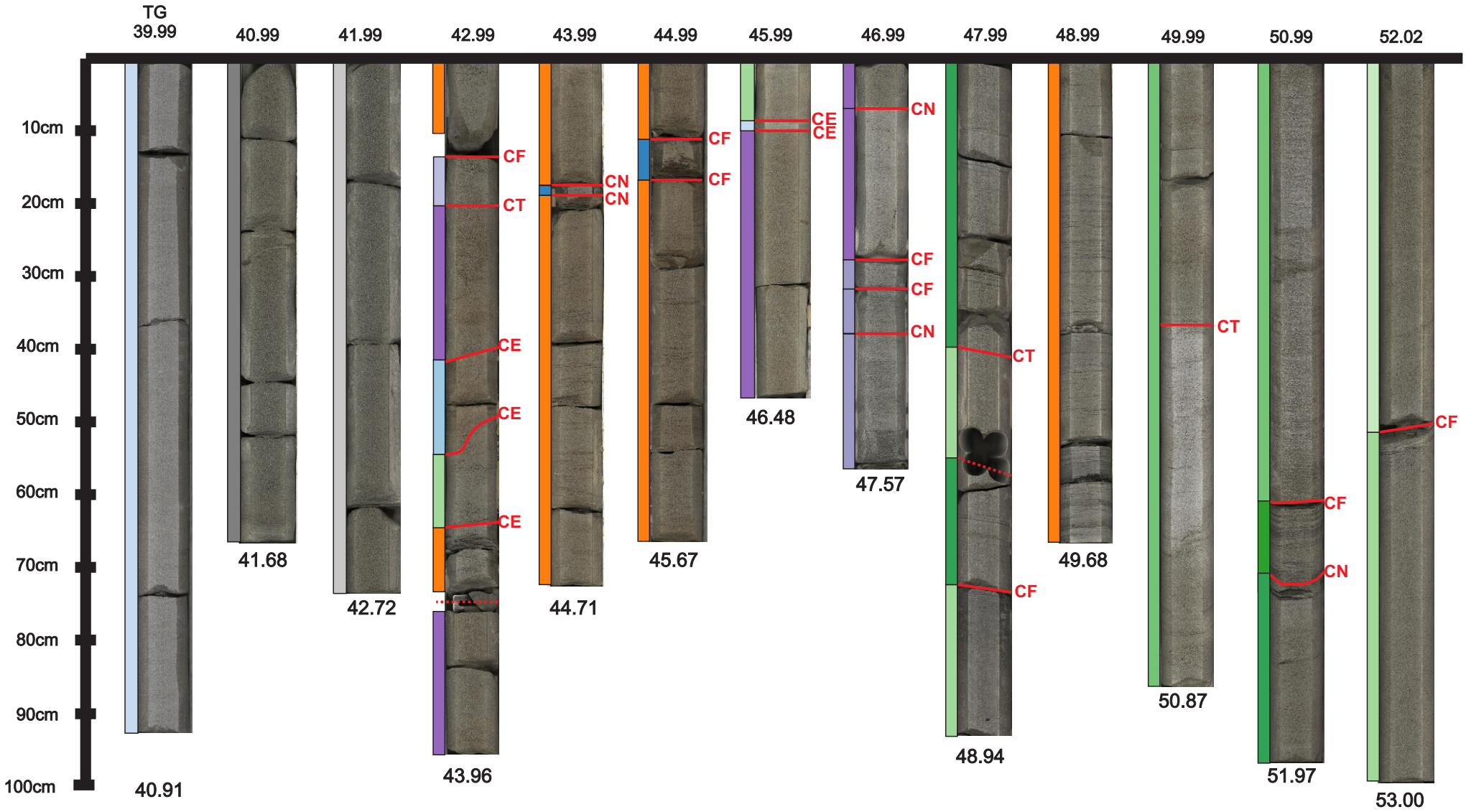
Facies

Ssb Htb Ssm

TG Techo general

BG Base general

Merge Core C03 página 1/3

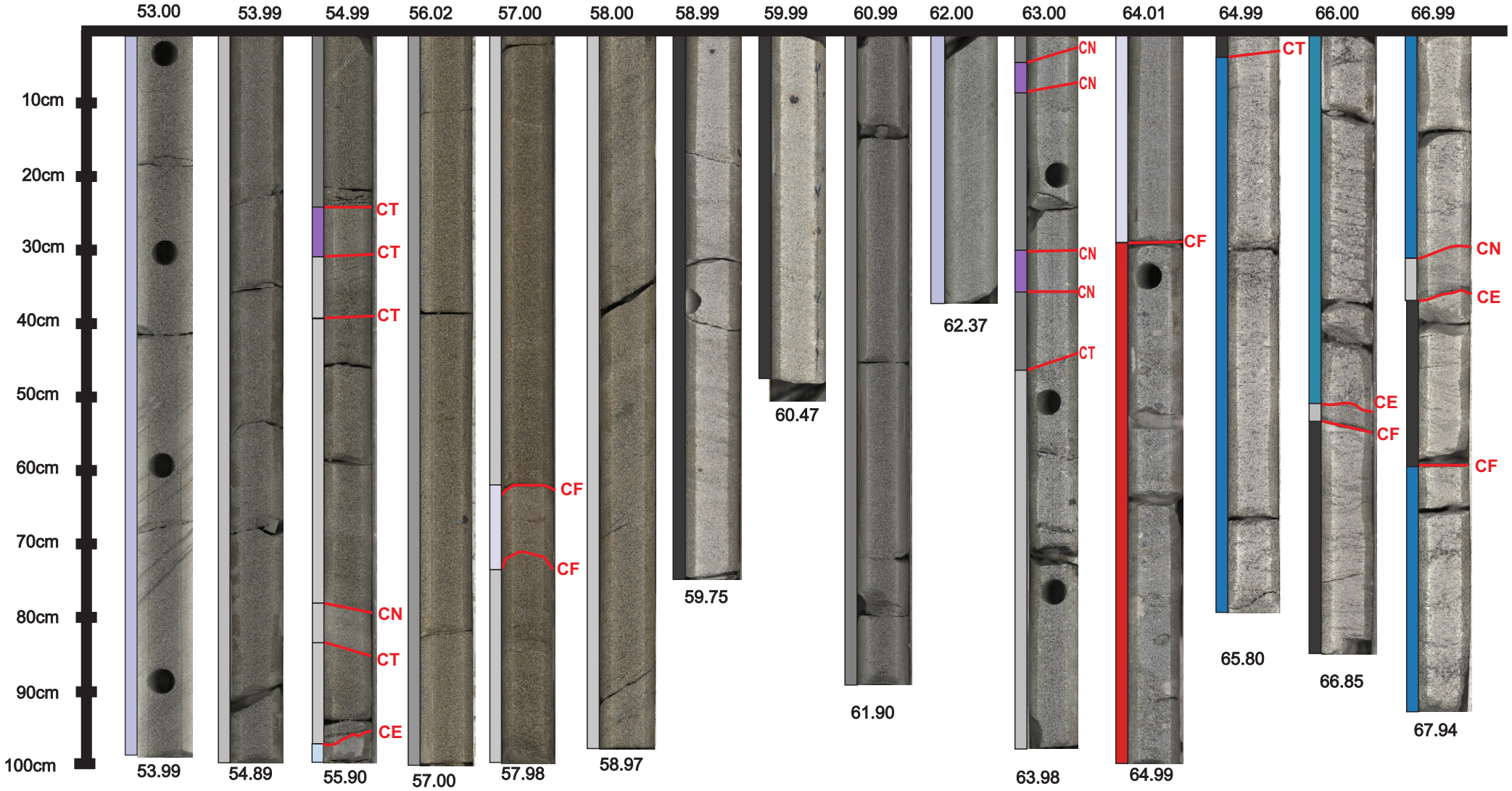


Referencias

BG Base general
TG Techo general
CN Contactos Netos
CF Contactos por Fracturas
CT Contactos Transicionales
CE Contactos Erosivos

Ssl Ssr S_vh S_ft S_fl S_mt S_ct D_fm $G_m$mm
Ssm Htf S_vt S_fm S_fb S_mb S_cb G_fm
Ssb S_vm S_vl S_fh S_mm S_cm S_pm G_ft

Merge Core C03 página 2/3

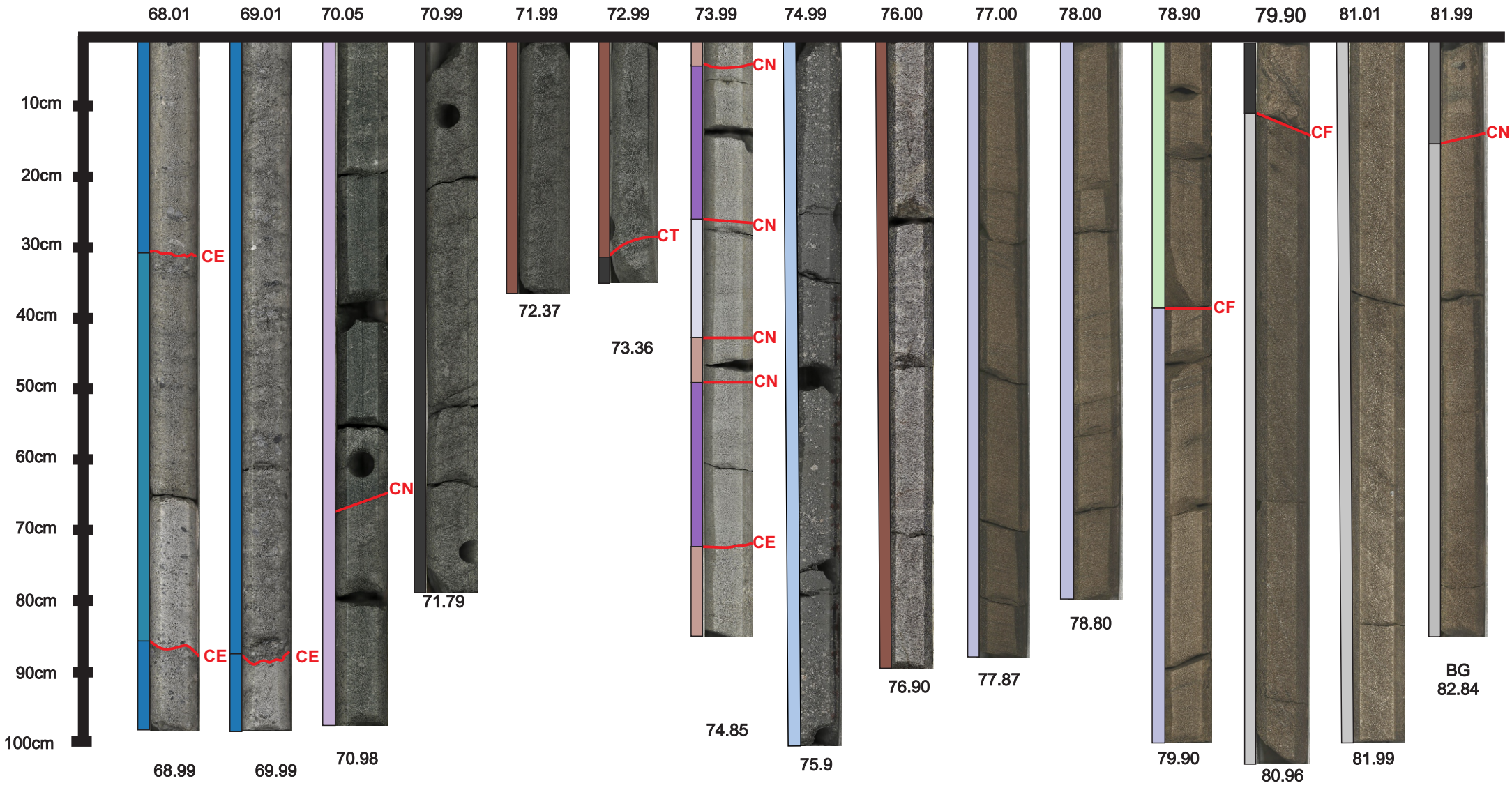


Referencias

BG	Base general	TG	Techo general
CN	Contactos Netos	CF	Contactos por Fracturas
CT	Contactos Transicionales	CE	Contactos Erosivos

Ssl Ssr S_vh S_t S_l S_mt S_ct D_fm $G_m$mm
Ssm Htf S_vt S_fm S_fb S_mb S_cb G_fm
Ssb S_vm S_vl S_fh S_mm S_cm S_pm G_ft

Merge Core C03 página 3/3



Referencias

BG Base general

CN Contactos Netos

CT Contactos Transicionales

TG Techo general

CF Contactos por Fracturas

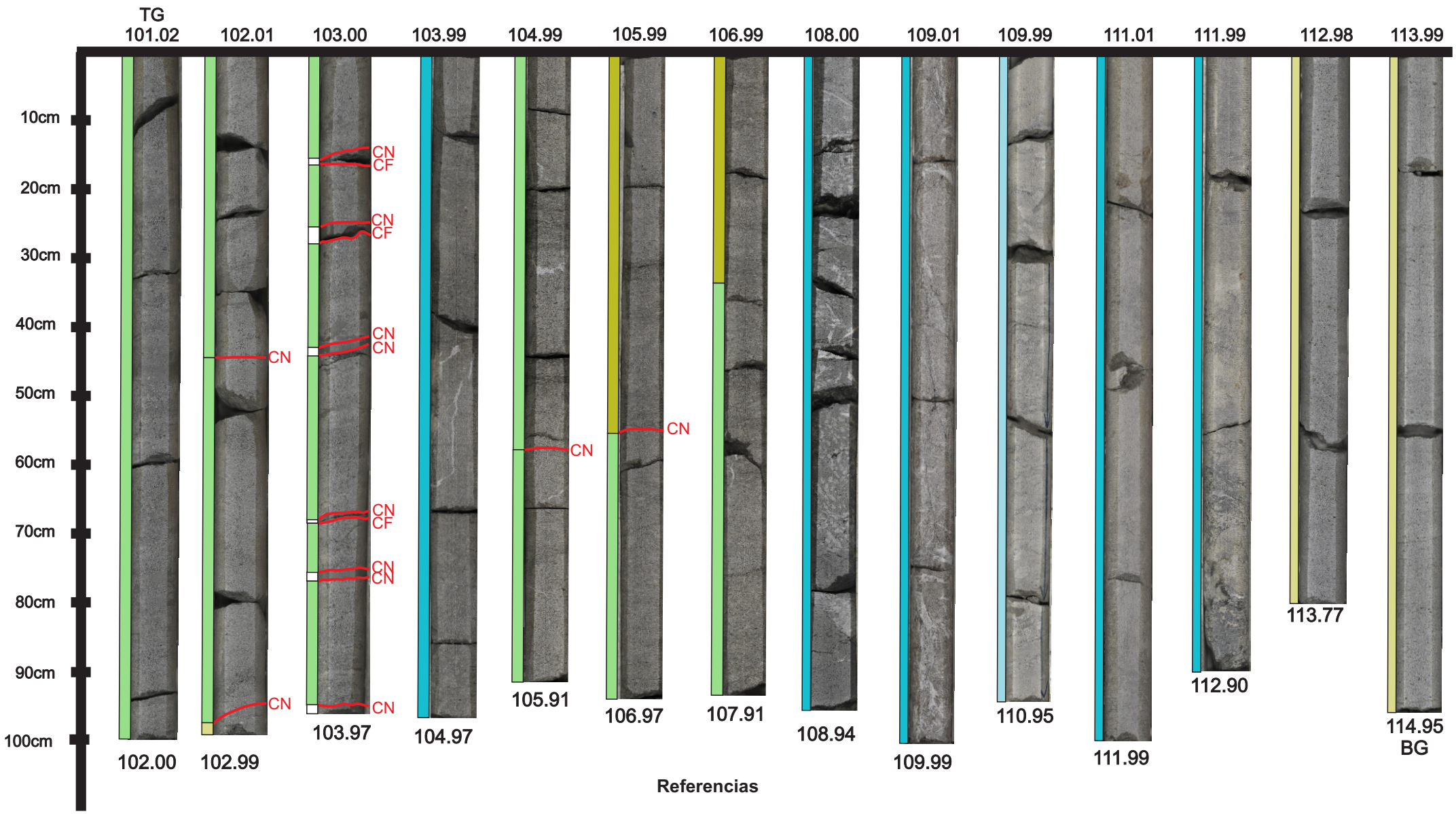
CE Contactos Erosivos

Ssl Ssr $S_v h$ $S_f t$ $S_f l$ $S_m t$ $S_c t$ $D_f m$ $G_m mm$

Ssm Htf $S_v t$ $S_f m$ $S_f b$ $S_m b$ $S_c b$ $G_f m$

Ssb $S_v m$ $S_v l$ $S_f h$ $S_m m$ $S_c m$ $S_p m$ $G_f t$

Merge Core C04



Referencias

Tipos de Contactos

CN Contactos Netos

CF Contactos por Fracturas

Facies

Pct | Gsm | Mcm

Pcl | Fsm | Csm

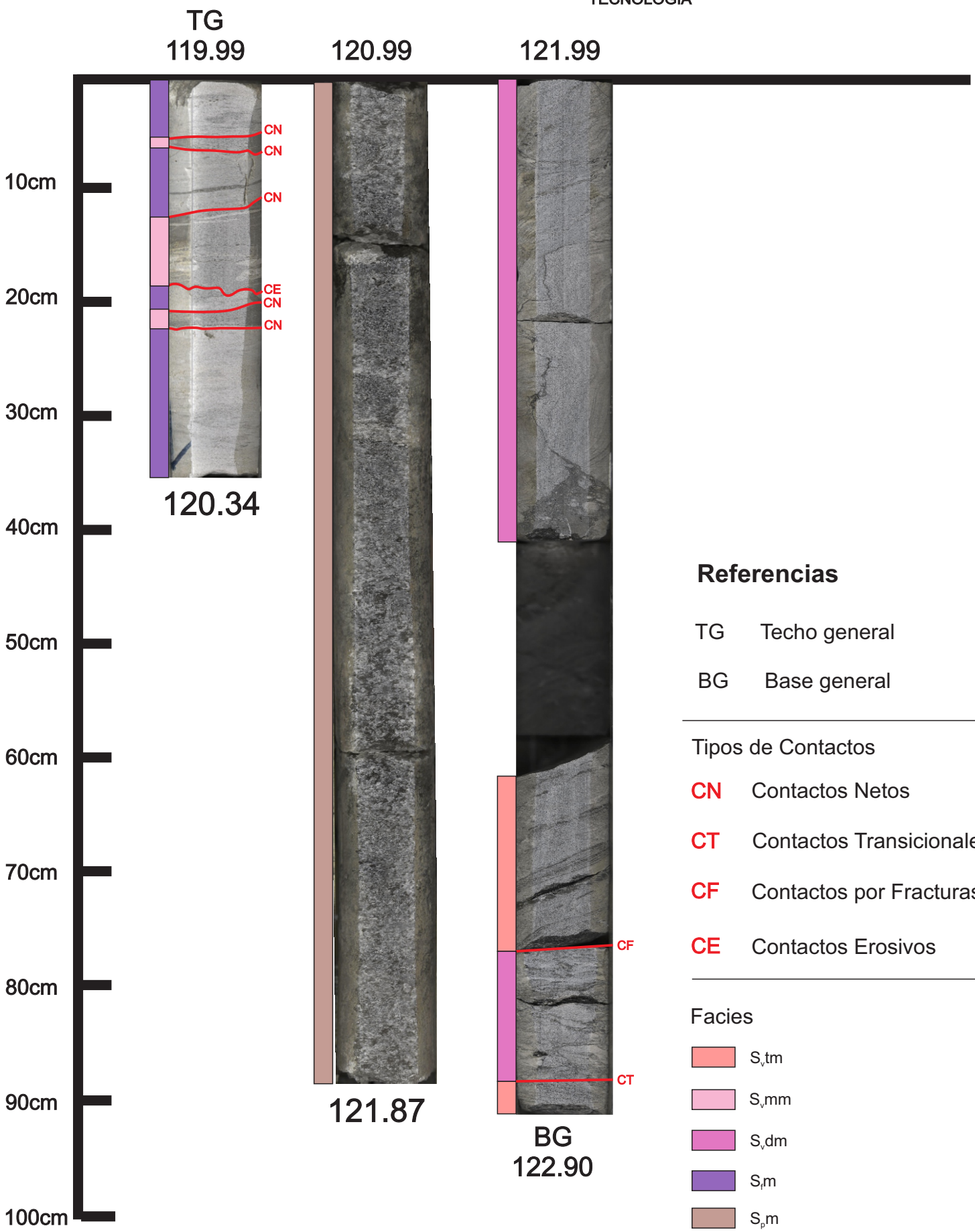
TG Techo general

BG Base general

Merge Core C05

Y-TEC TECNOLOGIA

EPSLOG ENGINEERING



Referencias

TG	Techo general
BG	Base general

Tipos de Contactos

CN	Contactos Netos
CT	Contactos Transicionales
CF	Contactos por Fracturas
CE	Contactos Erosivos

Facies

- $S_v$tm
- $S_v$mm
- $S_v$dm
- S_fm
- S_pm

Tabla 1: Tabla resumen de facies sedimentarias

	Facies (*)	Textura Granulométrica	Estructura Sedimentaria	Composición	Observaciones	Interpretación de procesos sedimentarios
	Csm	Arcilla	Maciza	Silicoclástica	Presencia de restos de valvas y intraclastos de packstone adyacente	Depositación flujo tractivo erosivo
	Ssl	Limosa	Laminar	Silicoclástica		Depositación por decantación a partir de una suspensión en condiciones de baja energía.
	Ssb	Limosa	Bioturbada	Silicoclástica	Presencia *de Planolites isp., Skolicia isp., Chondrites isp., Astherosoma isp., Teichichnus isp., Thalassinoides isp., Rosselia isp., Rhizocorallium isp., Ophiomorpha isp.,* Abundante presencia de materia orgánica en sectores.	Depositación por decantación a partir de una suspensión en condiciones de baja energía.
	Ssr	Limosa	Óndulas escalonadas	Silicoclástica	Presencia de mud drappes delgados	Depositación por corriente unidireccional con sedimentos en suspensión en condiciones de alternancia de tracción, con esporádicos lapsos de decantación.
	Ssm	Limosa	Maciza	Silicoclástica	Presencia de lentes pelíticos y materia orgánica. Alto contenido de Ca en FRX, se observa en UHRi la presencia de cemento.	Depositación por suspensión y/o floculación en condiciones de baja energía.
	Wsm	Vaque	Maciza	Silicoclástica	Presencia de tubos de *Thalassinoides isp., Chondrites isp. y Skolicia isp.*	Depositación por flujos densos en condiciones de baja energía.

(*) Código de facies establecido para la salida gráfica.

Tabla 1 continuación: Tabla resumen de facies sedimentarias

	Htw	Alternancia de textura limosa y arcillosa	Wavy	Silicoclástica		Depósitos dominados por decantación con episodios de alternancia de eventos tractivos.
	Htf	Alternancia de textura limosa y arcillosa	Flaser	Silicoclástica		Depósitos dominados por eventos tractivos con pausas que permiten la decantación.
	Htb	Alternancia de textura limosa y arcillosa	Bioturbada	Silicoclástica	*Phycosiphon isp., Teichichnus isp., Ophiomorpha isp, Planolites isp., Asterosoma isp.?, Thalassinoides isp.* Presencia de valvas cóncavas y convexas en algunos niveles.	Depósitos generados por alternancia de decantación y tracción con abundante actividad biológica que no permite ver el proceso dominante.
	Htd	Alternancia de textura limosa y arcillosa	Deformación sinsedimentaria	Silicoclástica		Depósitos generados por alternancia de decantación y tracción. Presenta tan alto grado de deformación que no permite ver el proceso dominante.
	S_vp	Arena muy fina	Entrecruzada tabular planar	Silicoclástica		Depósitos generados por un flujo tractivo fluido, unidireccional y de bajo régimen.
	S_vh	Arena muy fina	Entrecruzada de bajo ángulo	Silicoclástica		Depósitos generados por un flujo tractivo fluido, unidireccional y de alto régimen.
	S_vl	Arena muy fina	Laminar	Silicoclástica	Presencia de materia orgánica y *Planolites isp.* hacia el techo del pozo, limpia en la base.	Depositación por decantación en condiciones de baja energía.
	S_vm	Arena muy fina	Maciza	Silicoclástica	Presencia de materia orgánica, con *Thalassinoides isp, Planolites isp*. Por sectores cementación castaña.	Depositación por suspensión y/o floculación en condiciones de baja energía.

Tabla 1 continuación: Tabla resumen de facies sedimentarias

	S_vt	Arena muy fina	Entrecruzada tangencial o en artesa	Silicoclástica	Presencia de materia orgánica	Depósitos generados por un flujo tractivo fluido, unidireccional y de bajo régimen. Asociado a migración de ondulas 3D
	S_vng	Arena muy fina	Gradación normal	Silicoclástica	Presencia de abundante materia orgánica	Depósitos generados por corrientes de densidad que experimentan una desaceleración en su velocidad.
	S_fm	Arena fina	Maciza	Silicoclástica	Presencia de *Ophiomorpha isp, Phycosiphon isp., Skolitos isp.*	Depositación por suspensión y/o floculación en condiciones de baja energía.
	S_fp	Arena fina	Entrecruzada tabular planar	Silicoclástica		Depósitos generados por un flujo tractivo fluido, unidireccional y de bajo régimen. Asociado a migración de ondulas 2D
	S_fh	Arena fina	Entrecruzada de bajo ángulo	Silicoclástica		Depósitos generados por un flujo tractivo fluido, unidireccional y de alto régimen.
	S_ft	Arena fina	Entrecruzada tangencial o en artesa	Silicoclástica	Presencia de *Teichichnus isp*	Depósitos generados por un flujo tractivo fluido, unidireccional y de bajo régimen. Asociado a migración de ondulas 2D
	S_fl	Arena fina	Laminar	Silicoclástica	Presencia en sectores de opacos y materia orgánica	Deposito producto de tracción-decantación de bajo régimen de flujo.
	S_fb	Arena fina	Bioturbada	Silicoclástica		Depósito con la fabrica obliterada por actividad biológica.

Tabla 1 continuación: Tabla resumen de facies sedimentarias

	S_{ml}	Arena mediana	Laminar	Silicoclástica		Deposito producto de un flujo tractivo de alto régimen.
	S_{mm}	Arena mediana	Maciza	Silicoclástica	Abundante materia orgánica	Depósitos tractivos producto de corrientes de alta densidad.
	S_{mt}	Arena mediana	Entrecruzada tangencial o en artesa	Silicoclástica	Abundante presencia de materia orgánica y zonas de cemento blanquecino.	Depósitos generados por un flujo tractivo fluido, unidireccional y de bajo régimen. Asociado a migración de ondulas 3D
	S_{mb}	Arena mediana	Bioturbada	Silicoclástica	Presencia de *Arenicolites isp., Rhizocorallium isp., Diplocraterion isp., Skolitos isp.*	Deposito con la fábrica obliterada por actividad biológica.
	S_{cm}	Arena gruesa	Maciza	Silicoclástica		Depósitos tractivos producto de corrientes de alta densidad.
	S_{ct}	Arena gruesa	Entrecruzada tangencial o en artesa	Silicoclástica	Presencia de fitodetritos.	Depósitos generados por un flujo tractivo fluido, unidireccional y de bajo régimen. Asociado a migración de ondulas 3D
	S_{cb}	Arena gruesa	Bioturbada	Silicoclástica	Trazas de *Skolitos isp*. de gran tamaño	Depósito con la fábrica obliterada por actividad biológica.
	S_{pm}	Arena sabulítica	Maciza	Silicoclástica		Flujo fluido de alta energía que deposita que deposita una carpeta tractiva (que deposita solo granulometría gruesas).

Tabla 1 continuación: Tabla resumen de facies sedimentarias

	D_fm	Paraconglomerados fango sostén, compuesto de clastos tamaño fino	Maciza	Silicoclástica	Clastos angulosos de tamaño fino y matriz fangosa	Depósitos producto de flujos densos.
	G_fm	Ortoconglomerados clasto sostén, compuesto de clastos tamaño fino	Maciza	Silicoclástica	Clastos angulosos de tamaño fino. La matriz es clástica con tamaños desde arena muy fina hasta mediana.	Depósitos generados por un flujo tractivo fluido.
	G_ft	Ortoconglomerados clasto sostén, compuesto de clastos tamaño fino	Entrecruzada tangencial o en artesa	Silicoclástica	Clastos angulosos, la matriz es clástica con tamaños desde arena muy fina hasta mediana.	Depósito generado por flujo tractivo fluido. Asociado a migración de ondulas 3D
	Pct	Packstone	Entrecruzada tangencial o en artesa	Carbonática	Matriz granular de mudstone. Las valvas están mayoritariamente en posición cóncava	Depósitos bioclásticos transportados por un flujo tractivo fluido, unidireccional y de bajo régimen.
	Pcl	Packstone	Laminar	Carbonática	Matriz granular de mudstone. Las valvas están mayoritariamente en posición cóncava	Deposito bioclásticos producto de tracción de bajo régimen de flujo.
	Gsm	Grainstone	Maciza	Carbonática	Presencia de valvas convexas	Depósitos bioclásticos acumulados por suspensión
	Fsm	Floatstone	Maciza	Carbonática	Matriz granular de tamaño fino	Depósitos bioclásticos tractivos producto de corrientes de densidad.
	Mcm	Mudstone	Maciza	Carbonática	Restos de valvas dispersas. Trazas de *Thalassinoides isp.* En el último tramo del MC04 (111 a 113 metros) se observa una intensificación de la resistencia al rayado de la roca.	Depósitos por decantación a partir de una suspensión, de productos de actividad biológica en zona fótica.

Tabla 1 continuación: Tabla resumen de facies sedimentarias

	$S_v$tm	Arena muy fina	Entrecruzada tangencial o en artesa	Mixta	Alto contenido de Ca en FRX. Inferidas como mixtas, pero podría ser diagenéticos dado que existe la imposibilidad de diferenciar entre Ca original y Ca diagenéticos.	Depósitos generados por un flujo tractivo fluido, unidireccional y de bajo régimen. Asociado a migración de ondulas 3D
	$S_v$dm	Arena muy fina	Deformación sinsedimentaria	Mixta	Alto contenido de Ca en FRX. Inferidas como mixtas, pero podría ser diagenéticos dado que existe la imposibilidad de diferenciar entre Ca original y Ca diagenéticos.	Depósitos que presentan tan alto grado de deformación sinsedimentaria que no permite ver el proceso primario dominante.
	$S_v$mm	Arena muy fina	Maciza	Mixta	Alto contenido de Ca en FRX. Inferidas como mixtas, pero podría ser diagenéticos dado que existe la imposibilidad de diferenciar entre Ca original y Ca diagenéticos.	Depositación por decantación a partir de suspensión y/o floculación en condiciones de baja energía.
	$G_m$mm	Ortoconglomerado clasto sostén, compuesto por clastos tamaño mediano	Maciza	Mixta	Presencia de bioclastos y clastos silíceos.	Depósito generado por flujo tractivo fluido.

ANEXO II

Código de facies utilizado para la descripción del set de datos y para la realización de las predicciones

Tabla 1: Nomenclaturas del código de facies para las texturas sedimentarias utilizados en el programa.

	Texturas		Nomenclatura del programa	Nomenclatura salida gráfica
Arenas & Mixtas	Arcillita		CS	Cs
	Limolita		SS	Ss
	Vaque		WS	Ws
	Arenisca	Heterolítica	HT	Ht
		Muy fina	SV	S_v
		Fina	SF	S_f
		Media	SM	S_m
		Gruesa	SC	S_c
		Sabulítica	SP	S_p
	Arenisca Conglomerádica		SG	S_g
	Ortoconglomerado	Fina	GF	G_f
		Media	GM	G_m
		Gruesa	GC	G_c
	Paraconglomerado	Fina	DF	D_f
		Media	DM	D_m
		Gruesa	DC	D_c
Carbonáticas	Mudstone		MC	Mc
	Wackestone		WC	Wc
	Packstone		PC	Pc
	Grainstone		GS	Gs
	Floatstone		FS	Fs
	Rudstone		RC	Rc
	Bafflestone		BF	Bf
	Bindstone		BI	Bi
	Dolomita		DC	Dc
	Framestone		FC	Fc
Volcaniclasticas	Brecha piroclástica		BP	Bp
	Aglomerado piroclástico		AP	Ap
	Lapillita		LP	Lp
	Lapillitas - Tufitas Soldada (ignimbritas)		IG	Ig
	Lapillitas – Tufitas No soldada		LT	Lt
	Tufitas		TP	Tp
	Chonitas		CP	Cp

Tabla 2: Nomenclaturas del código de facies para las estructuras sedimentarias utilizados en el programa.

	Estructuras	Nomenclatura del programa	Nomenclatura salida gráfica
Depositacionales	Flaser	f	f
	Ondulítica (Wavy)	o	w
	Lentiforme	l	l
	Masiva	m	m
	Bioturbada	b	b
	Gradación normal	e	ng
	Gradación inversa	g	ig
	Mud drapes	k	md
	Estratificación entrecruzada tangencial	t	t
	Estratificación entrecruzada en artesa	z	t
	Estratificación planar	p	p
	Ondulas de corrientes	r	r
	Ondulas escalonadas	c	c
	Ondulas de oscilación	w	w
	Antidunas	a	a
	Estratificación horizontal	h	h
	Laminación	l	l
	Estratificación entrecruzada de bajo ángulo	y	h
	Laminación parting	x	h
	Hummocky	q	hcs
Postd.	Deformación sinsedimentaria (laminación convoluta, calcos de carga, escape de agua)	d	d
	Deslizamientos	v	d
Erosivas	Intraclastos pelíticos	i	i
	Calcos / Flutes	j	f

Tabla 3: Nomenclaturas del código de facies para las distintas composiciones de las rocas sedimentarias utilizados en el programa.

Composición	**Nomenclatura del programa**
Carbonática	c
Silicoclástica	s
Bioclástica	b
Mixta	m
Glauconítica	g
Oolítica	o
Vítrea	v
Lítica	l
Cristalina	x

ANEXO III

Gráficos SHAP obtenidos durante el modelado de los datos

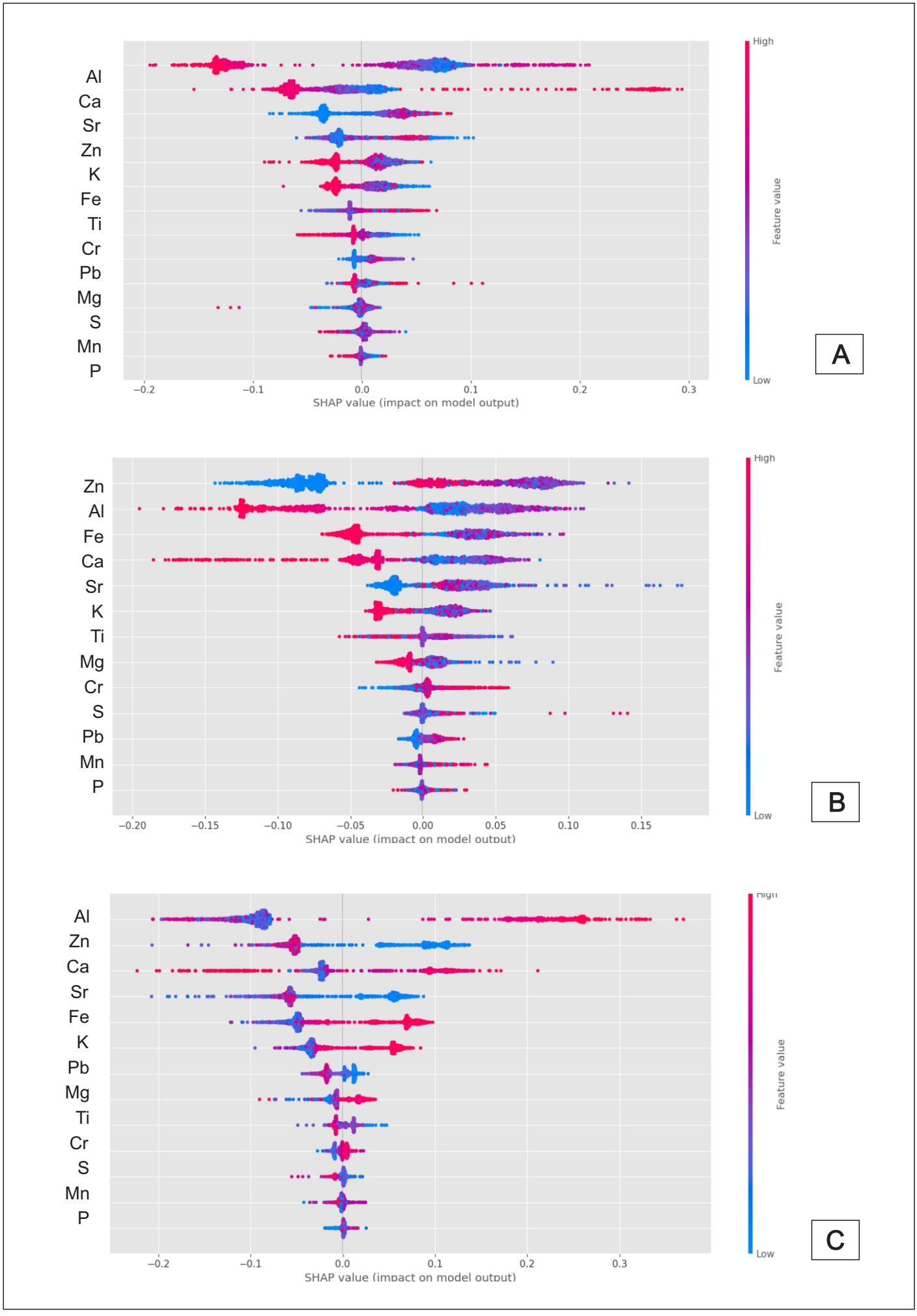


Figura 1: Gráficos de evaluación de la importancia de las diferentes variables composicionales (A) Carbonáticas; (B) Mixtas; y (C) Silicoclásticas.

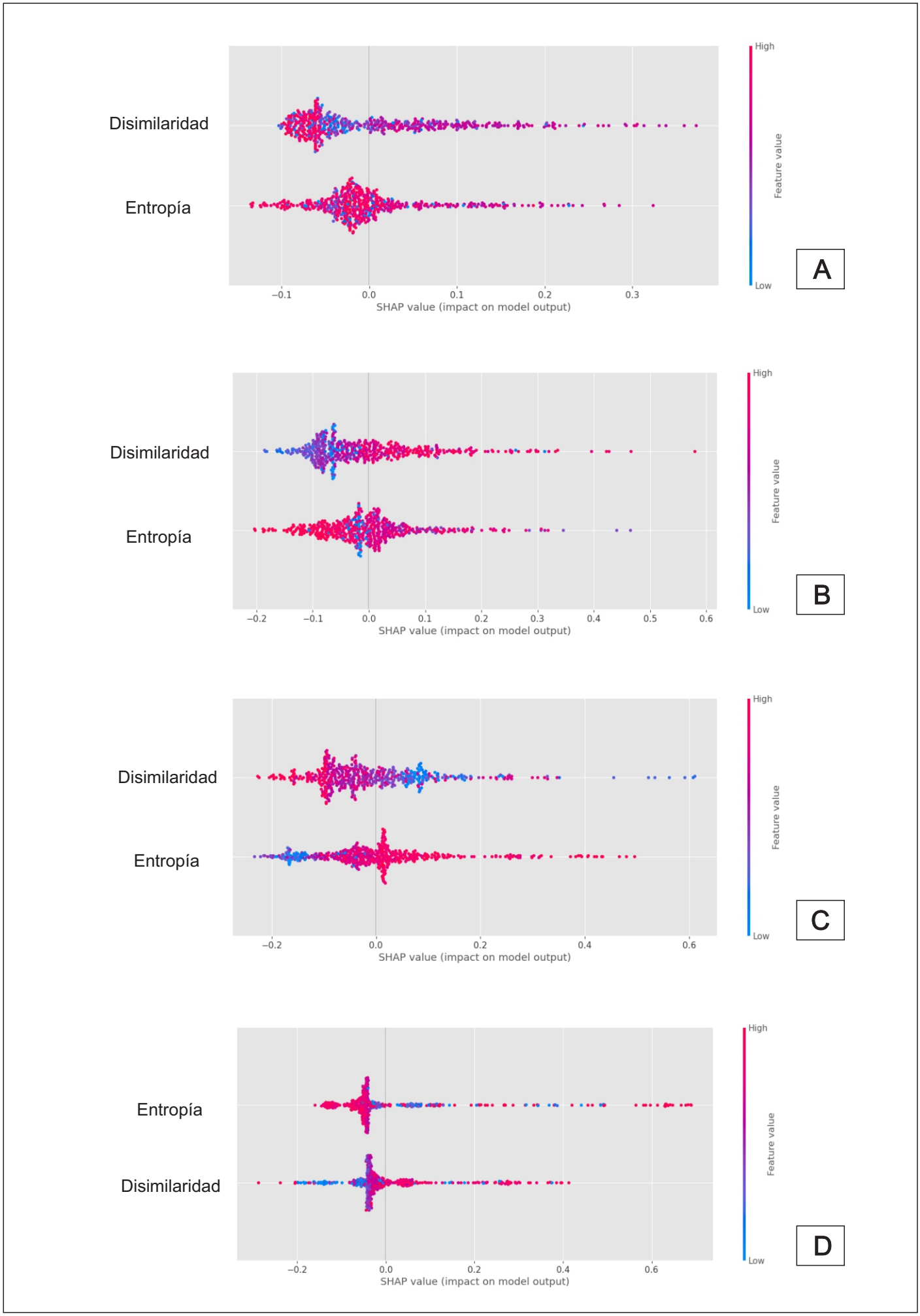


Figura 2: Gráficos de evaluación de la importancia de las diferentes variables de texturas granulométricas (A) Arcillas; (B) Heterolíticas; (C) Ortoconglomerado fino; (D) Ortoconglomerado mediano; (E) Paraconglomerado fino; (F) Arena gruesa; (G) Arena fina; (H) Arena mediana; (I) Arena sabulítica; (J) Arena muy fina; (K) Limo; (L) Vaque.

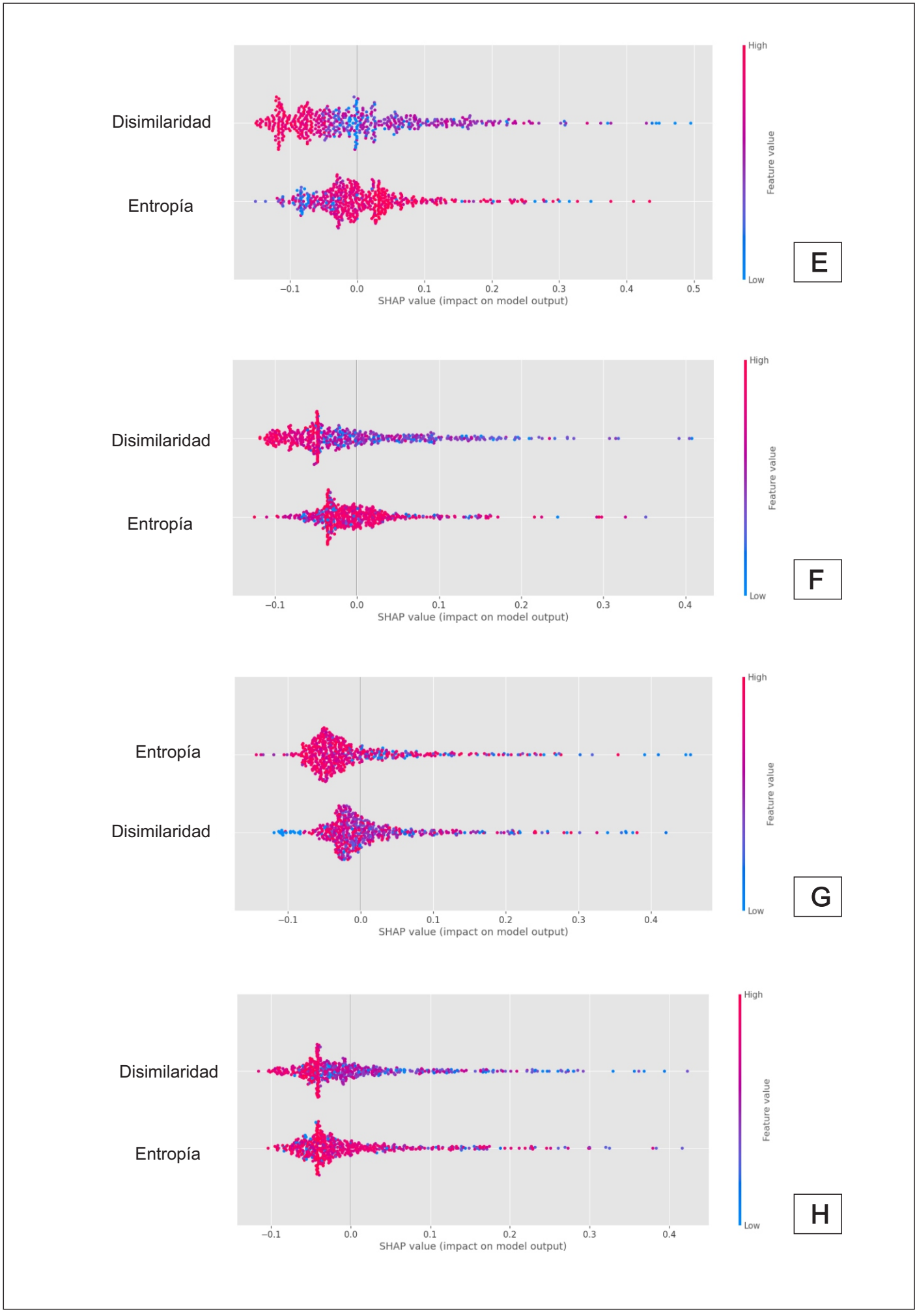


Figura 2 continuación: Gráficos de evaluación de la importancia de las diferentes variables de texturas granulométricas (A) Arcillas; (B) Heterolíticas; (C) Ortoconglomerado fino; (D) Ortoconglomerado mediano; (E) Paraconglomerado fino; (F) Arena gruesa; (G) Arena fina; (H) Arena mediana; (I) Arena sabulítica; (J) Arena muy fina; (K) Limo; (L) Vaque.

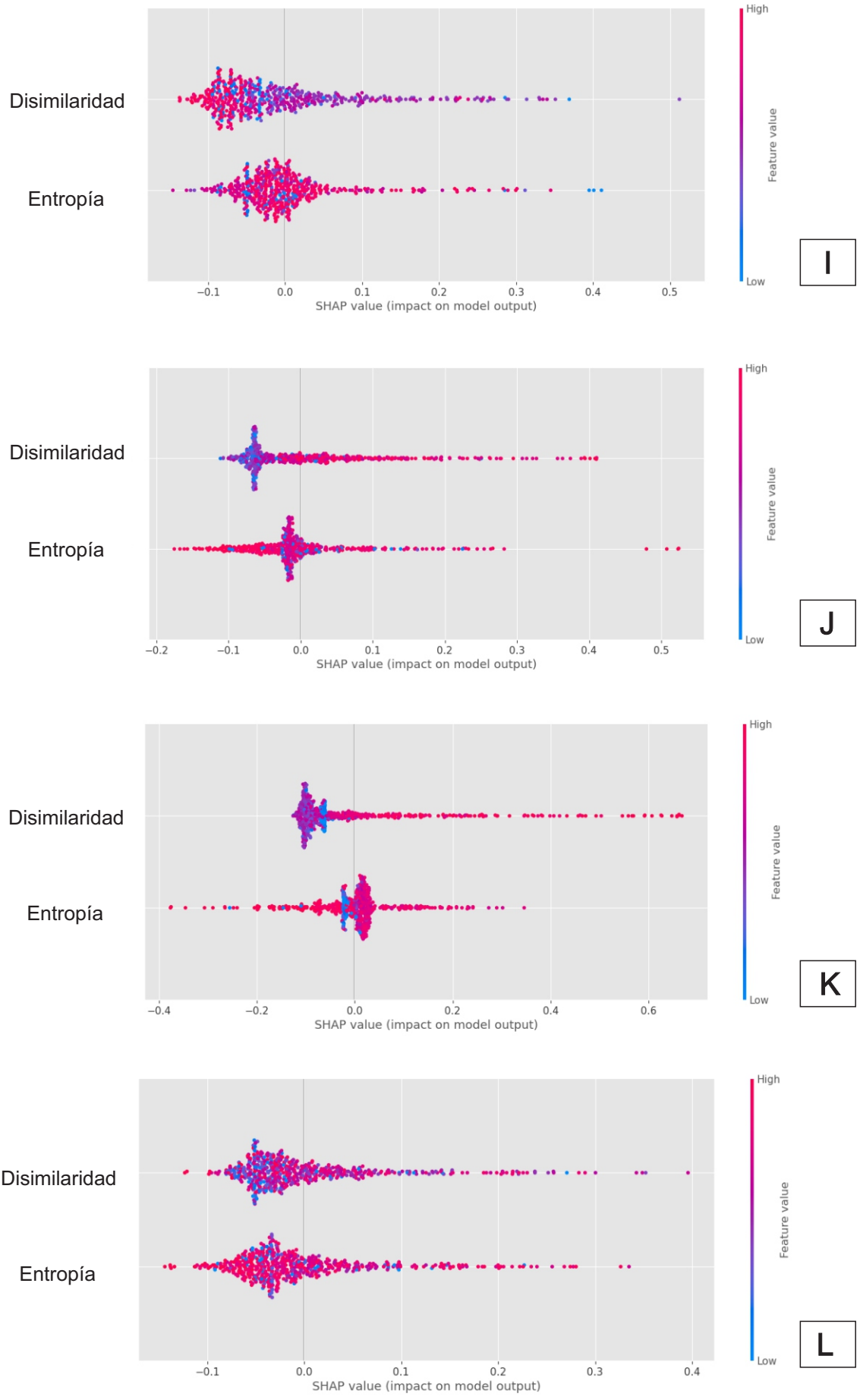


Figura 2 continuación: Gráficos de evaluación de la importancia de las diferentes variables de texturas granulométricas (A) Arcillas; (B) Heterolíticas; (C) Ortoconglomerado fino; (D) Ortoconglomerado mediano; (E) Paraconglomerado fino; (F) Arena gruesa; (G) Arena fina; (H) Arena mediana; (I) Arena sabulítica; (J) Arena muy fina; (K) Limo; (L) Vaque.

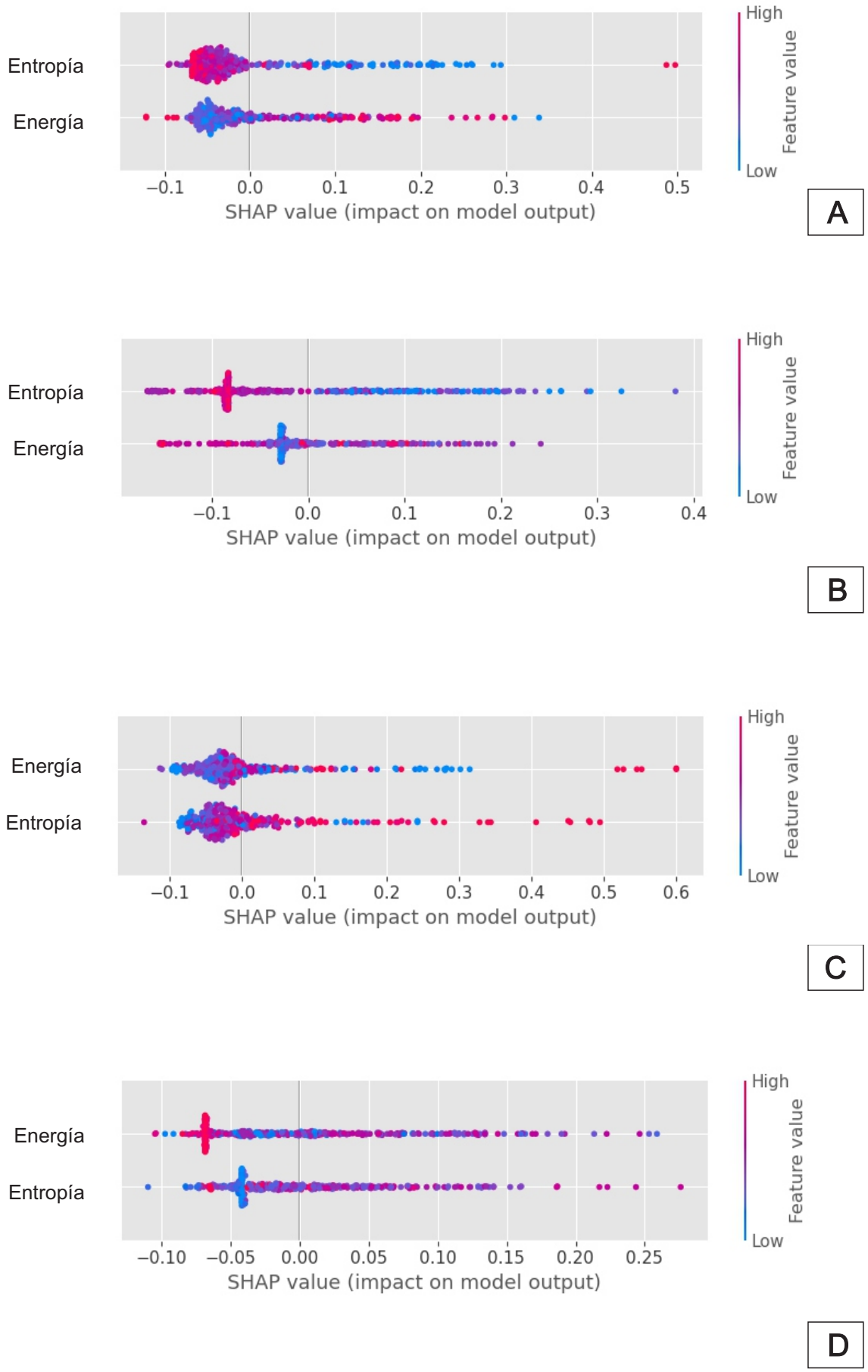


Figura 3: Gráficos de evaluación de la importancia de las diferentes variables de estructuras sedimentarias. (A) Estructura bioturbada; (B) Estructura entrecruzada planar; (C) Estructura entrecruzada tangencial; (D)Estructura flaser; (E) Estructura estratificación horizontal; (F) Estructura laminar; (G) Estructura masiva; (H) Estructura con deformación sinsedimentaria; (I) Estructura wavy.

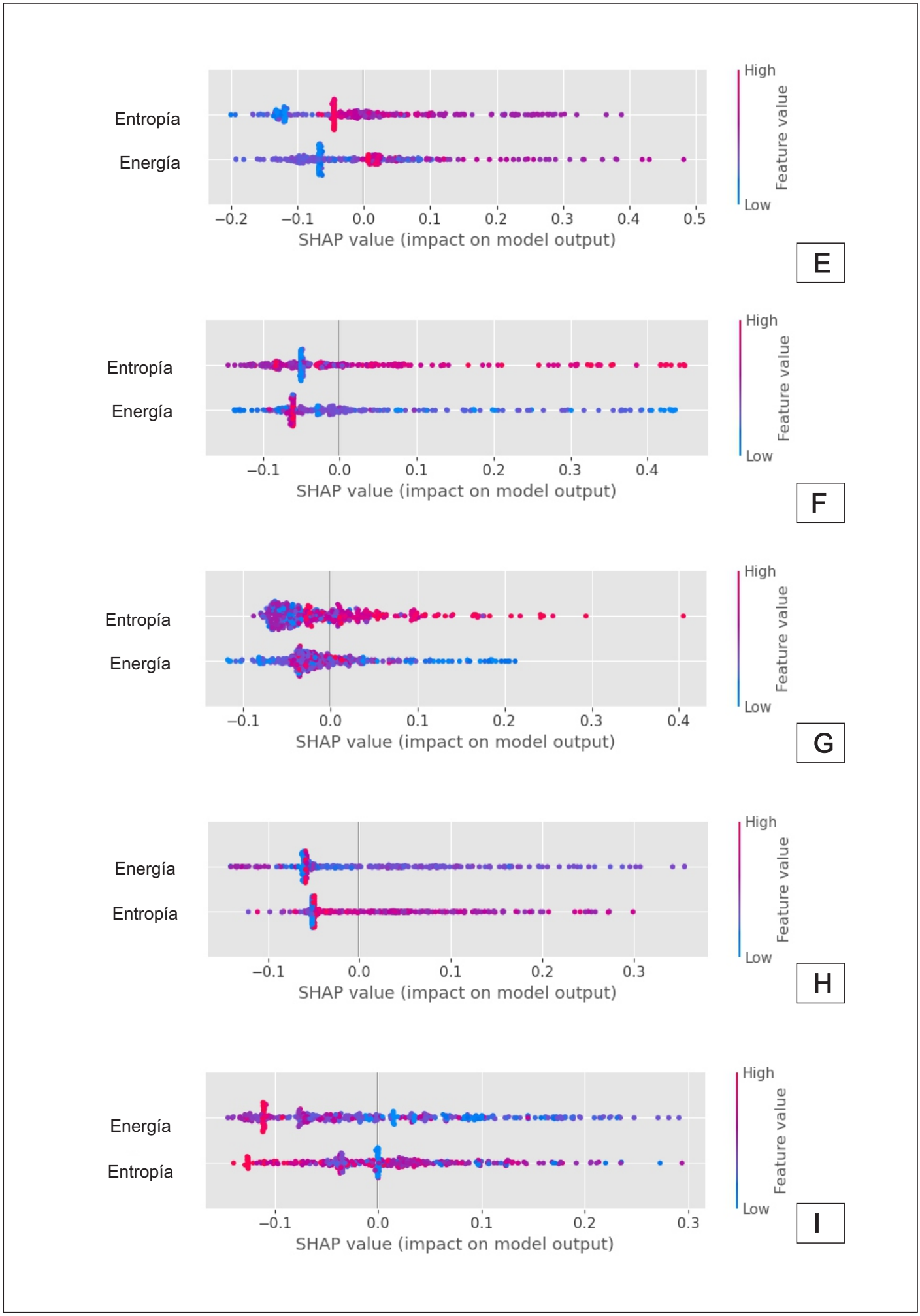


Figura 3 continuación: Gráficos de evaluación de la importancia de las diferentes variables de estructuras sedimentarias. (A) Estructura bioturbada; (B) Estructura entrecruzada planar; (C) Estructura entrecruzada tangencial; (D)Estructura flaser; (E) Estructura estratificación horizontal; (F) Estructura laminar; (G) Estructura masiva; (H) Estructura con deformación sinsedimentaria; (I) Estructura wavy.