UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Clasificación litológica en afloramientos de rocas mediante técnicas de aprendizaje automático usando imágenes satelitales

Elizabeth Rodríguez Arias

Universidad Santo Tomás
Facultad de Estadística
División de Ciencias Económicas y Administrativas
Bogotá, D.C., Colombia
2023

Clasificación litológica en afloramientos de rocas mediante técnicas de aprendizaje automático usando imágenes satelitales

Elizabeth Rodríguez Arias

Trabajo de grado presentado como requisito parcial para optar al título de:
Magister en Estadística Aplicada

Directores:
Juan Carlos Rubriche (MSc.)
Eliseo Tesón Del Hoyo (PhD)

Línea de Investigación:
Aprendizaje automático
Grupo de Investigación:
USTAdistica

Universidad Santo Tomás
Facultad de Estadística
División de Ciencias Económicas y Administrativas
Bogotá, D.C., Colombia
2023

Dedicatoria

A mi amado esposo Eliseo, porque gracias a su permanente e incondicional apoyo he podido hacer este gran sueño realidad.

A mis hermanas Rosy y Joha gracias por su cariño, fortaleza y aliento que me mantuvieron a flote, aun cuando por momentos, las dificultades amenazaban por hundirme.

A mis hijos Alejandro y Eloy les pido perdón por mis continuas ausencias durantes estos meses, los ratos de mal humor y por las cenas improvisadas con comida congelada.

...y en memoria de Don Elías, quien me enseñó que las cosas que merecen la pena requieren esfuerzo y que el trabajo duro siempre trae una dulce recompensa.

Resumen

Aplicar técnicas de aprendizaje de máquinas para realizar caracterización de terreno a partir de imágenes de satélite, supone un gran avance en el estudio de la superficie de la tierra; debido a la heterogeneidad de las estructuras rocosas y las dificultades asociadas a cambios abruptos de topografía, la clasificación convencional tiene un gran componente subjetivo, que en ocasiones puede afectar la precisión del proceso. El objetivo de usar aprendizaje de máquinas para hacer clasificación litológica es generar una herramienta que asista el proceso de clasificación y así mejorar los aspectos que puedan limitar los resultados, como suelen ser, optimizar el tiempo de procesamiento, mejorar la precisión de la clasificación, procesar bases de datos más voluminosas (imágenes con mayor resolución espectral y espacial) y cubrir áreas de estudio más extensas. La metodología está desarrollada sobre la plataforma de análisis geoespacial con procesamiento en la nube *Google Earth Engine* (GEE) y está implementada en cuatro fases principales; la primera es generar variables adicionales para el modelo de clasificación supervisada, mediante técnicas estadísticas de transformación de imagen y mejora espectral, como el cálculo de índices espectrales geológicos y el análisis de componentes principales (PCA), así como incorporar información topográfica mediante la adición del modelo digital de terrero (DEM) y agregando datos de textura por medio de la matriz de coocurrencia de nivel de grises (GLMC), también se incorporan los resultados del algoritmo de clasificación no supervisada *K-means* como una variable predictora. La segunda fase consiste en entrenar los algoritmos de aprendizaje supervisado *Ramdon Forest* (RF), *Support Vector Machine* (SVM), *Classification and Regression Trees* (CART), *Minimum Distance* (MD) y *Naive Bayes* (NB) teniendo en cuenta las variables generadas en la primera fase junto información de referencia de campo o *Ground Thruth*, para reproducir varios modelos de clasificación de los cuales en la tercera fase, se elige el que produzca mejores métricas de desempeño, generando un modelo de clasificación confiable que en una última fase se reproduce sobre una zona de superficie con características y litología desconocidas, aportando así información de base para realizar cartografía geológica sin haber requerido presencia de personal in situ.

Palabras clave: Teledección, machine learning, litología, banda espectral, reflectancia.

Abstract

The application of machine learning techniques to perform terrain characterization from satellite images represents a significant advance in the study of the Earth's surface. Because of the heterogeneity of rock structures and the difficulties associated with abrupt changes in topography, conventional classification possesses a large subjective component, which sometimes can affect the accuracy of the process. The desired result of using machine learning techniques in Lithological classification is to generate a tool that besides assisting the classification process, also helps to improve the aspects that may limit the results, such as optimizing processing time, increasing classification accuracy, larger data processing (images with higher spectral and temporal resolution) and to cover larger study areas. The methodology is developed on the geospatial analysis platform with cloud processing Google Earth Engine (GEE) and it is implemented in four main phases: the first one consists of generating additional variables for the supervised classification model using statistical techniques, image transformation, and spectral enhancement, including the calculation of geological spectral indices and Principal Component Analysis (PCA), as well as incorporating topographic information through addition of Digital Terrain Model (DEM), adding together texture data through the Gray Level Matrix Co-occurrence (GLMC), and finally incorporate an additional band of the k-means algorithm no supervised classification. The second phase incorporates algorithm training under the supervision of Random Forest (RF), Support Vector Machine (SVM), Classification and Regression CART and Minimum Distance MD, taking into account the variables generated in the first phase together with the field reference information or ground thruth to reproduce several classification models that in the third phase, whichever displays the best performance metrics, is chosen leading to a reliable classification model, then, in the last phase, it will be applied to a surface area with unknown characteristics and lithology providing support information to carry out geological cartography without having required presence of people on site.

Key words: Remote sensing, machine learning, algorithm, lithology, spectral resolution, reflectance.

Lista de Figuras

Lista de Tablas

1 Introduction

Realizar discriminación litológica usando imágenes de satélite mediante técnicas de machine learning es una de las tareas más desafiantes para la geología, en especial cuando se trata de grandes áreas geográficas y de litologías complejas con múltiples clases a identificar (Serbouti et al., 2022). Si bien, automatizar el proceso de clasificación brinda grandes beneficios, como más agilidad en tiempo de procesamiento de datos, mejor precisión y mayor extensión de área estudiada; aún cuenta con varias limitaciones, como, lidiar con la similitud espectral de algunas litologías, la baja y media resolución espacial y espectral de imágenes de algunos satélites de libre acceso y que aún se requiere de información de campo, datos importantes a la hora de entrenar los algoritmos de clasificación supervisada; sin embargo existen múltiples estudios en donde se han puesto a prueba diferentes técnicas de fusión de datos y herramientas estadísticas que ayudan a manejar estas dificultades obteniendo resultados inmejorables(Serbouti et al., 2022).

Los sensores de datos abiertos más utilizados para estudios geológicos son ASTER del satélite TERRA y OLI del satélite LANDSAT, gracias a que cuentan con bandas en el espectro infrarrojo de onda corta (SWIR) y el infrarrojo térmico (TIR), en las cuales los minerales exhiben mayor reflectancia facilitando la discriminación litológica de rocas. (Ninomiya, 2004)sin embargo la resolución espacial de estos dos sensores es relativamente baja, lo que ha permitido abrir el panorama del mapeo litológico a otros satélites de datos abiertos de resolución espacial media como Sentinel-2 con el sensor MSI, el cual carece de bandas espectrales del infrarrojo térmico TIR, pero tiene mejor resolución espacial que los otros señores en sus bandas espectrales, dando excelentes resultados en discriminación de litologías, (Serbouti et al., 2022); en vista de estás particularidades, para complementar las propiedades espectrales de cada satélite se han planteado estudios comparativos entre los diferentes sensores, así como pruebas con diferentes algoritmos de clasificación y la adición de información proveniente de técnicas de mejora espectral, (Shirmard et al., 2022), de análisis de terreno y de textura de imagen,(Serbouti et al., 2022).

El método de clasificación más usado para discriminación litológica es la clasificación supervisada, en particular los algoritmos Random Forest (RF), Support Vector Machine (SVM), Clasificación y Regresión (CART) y Mínima Distancia (MD) métodos que pueden ejecutarse

de manera muy eficiente en términos de tiempo y capacidad computacional, en la plataforma de análisis geoespacial con procesamiento en la nube Google Earth Engine (GEE), la cual cuenta con un extenso catálogo de datos corregidos de múltiples satélites para descargar, visualizar y procesar mediante una interfaz de programación que usa lenguaje JavaScript, siendo ésta muy intuitiva y cuenta con infinidad de repositorios e información libre y de acceso general.

Existen múltiples aplicaciones de los sistemas de teledetección; desde sus inicios en los que fueron muy utilizados en el monitoreo del ambiente, hasta su uso más reciente en la clasificación de la superficie de la tierra, por medio de la generación de mapas o "mapeo"; a través de la aplicación de técnicas de clasificación y reconocimiento de patrones, para generar mapas con las características específicas de los objetos de interés, por ejemplo, mapas de cobertura de suelo (tipo y densidad de la vegetación, presencia de árboles, arbustos, bosques, etc.) (Pokhariya et al., 2023)y mapas de uso de la tierra (uso agrícola, residencial, explotación minera, etc.) (Congalton y Green, 2009).

Dentro de las incursiones más novedosas de la teledetección en la geología son los aportes a la cartografía, siendo esta una disciplina transversal a las ciencias de la tierra, que trata de plasmar sobre un mapa de escala determinada, diferentes rasgos geológicos de interés y es quizá, la disciplina de la geología más importante en el sentido de que pretende, mediante una abstracción, plasmar el mayor número de rasgos geológicos posibles que sirven de soporte para posteriores interpretaciones (Hargitai, 2019). Generalmente, la cartografía geológica captura las diferencias litológicas representables a la escala del mapa construido (Quirós, 2009).

En un estudio previo en donde se realizó mapeo geológico usando diferentes algoritmos de clasificación supervisada e imágenes Landsat 7 y datos geofísicos aéreos, se llegó a la conclusión que el algoritmo Ramdon Forest aporta un resultado confiable al combinar información espectral y datos geofísicos (Cracknell y Reading, 2013). En otro estudio se en donde se quiso evaluar el efecto de incorporar información topográfica en la clasificación litológica, se encontró que la precisión de la clasificación aumentaba en un 10 % cuando se tenía en cuenta el modelo de elevación digital como una variable adicional en el entrenamiento de los algoritmos de clasificación, siendo el mejor el método SVM (Ge et al., 2018). Como se puede observar en estos estudios y preliminares, cuando se trata de la aplicación especifica de mapeo litológico, se puede mejorar la precisión de la clasificación incluyendo información complementaria a la extraída de las bandas de las imágenes Satelitales, esto debido a la similitud espectral de las litologías, lo que puede dar lugar a errores de clasificación.

El objetivo principal de este trabajo es encontrar el mejor modelo de clasificación litológica usando los algoritmos CART, Random Forest, Minimun Distance, SVM y NB, junto a la combinación de diferentes variables predictoras que resalten las características de la imagen de base. Así mismo, se pretende facilitar las interpretaciones de áreas donde es muy difícil o físicamente imposible acceder a los afloramientos de roca, ya sea porque son lugares con topografía muy agreste, de difícil accesibilidad o porque se vea comprometida la integridad y seguridad de los analistas al tratar de personarse en ellas.

La metodología planteada para este estudio se desarrolla en cuatro etapas principales que pueden encontrar en el capítulo 4 de este documento. La primera sección consiste en descargar y preprocesar el conjunto de imágenes Sentinel MSI, para generar la geodatabase con la cual se va a trabajar, posterior a eso, se realiza el procesamiento que consiste en la aplicación de técnicas de mejoramiento espectral y de transformación de imagen, que permite generar variables adicionales para incorporar al modelo de clasificación. La segunda etapa es la valoración de los algoritmos de clasificación, para lo cual, primero se han de generar de etiquetas según las clases litológicas y la configuración de datos de entrenamiento y validación de cada modelo, a través de validación cruzada. La tercera etapa consiste en la evaluación de precisión de cada algoritmo y la selección del mejor, que será usado en la última etapa para la caracterización e interpretación de una zona de geología desconocida y comprobar si los resultados son satisfactorios.

2 Marco Teórico

La teoría que enmarca el objeto de estudio del presente trabajo de investigación abarca dos grandes áreas del conocimiento, que se entrecruzan en aplicaciones muy específicas, como en este caso soportar el proceso de cartografía geológica. A grandes rasgos, las dos áreas que soportan el estudio son, aprendizaje automático (Machine Learning) y datos de teledetección (Remote sensing data); de los cuales se derivan múltiples aplicaciones, en este capítulo se hará énfasis en la implementación del aprendizaje automático en imágenes satelitales para realizar clasificación litológica.

2.1. Datos de teledetección satelital –Remote Sensing Data:

La Teledetección o -Remote Sensing - es una disciplina que se enfoca en las metodologías, instrumentos y herramientas computacionales, para estudiar la superficie de la tierra y el medio ambiente, a través de sensores que recolectan la energía emitida y/o reflejada de los objetos, pero que no están en contacto directo con su superficie (Thenkabail, 2016). Existen múltiples tipos de plataformas de sensores que interactúan con el área objetivo a través de ondas electromagnéticas, dentro de los cuales se encuentran, los sensores a bordo de los satélites, plataformas aéreas como aviones, drones, globos o en elementos fijos sobre la tierra como grúas, edificios, antenas (Moser y Zerubia, 2018). En este capítulo se hace énfasis en las imágenes obtenidas por medio de los sensores satelitales.

En la teledetección intervienen una serie de elementos, que interactúan entre sí para hacer posible la observación y monitoreo de forma remota de la superficie de la tierra. El proceso inicia cuando los sensores a bordo de los satélites registran la radiación electromagnética emitida por los objetos y luego la trasmiten de forma electrónica a estaciones de procesamiento, en donde se extrae información relevante del área objetivo, por medio de técnicas y algoritmos, que posteriormente será analizada e interpretada para la respectiva toma decisiones (Thenkabail, 2016).

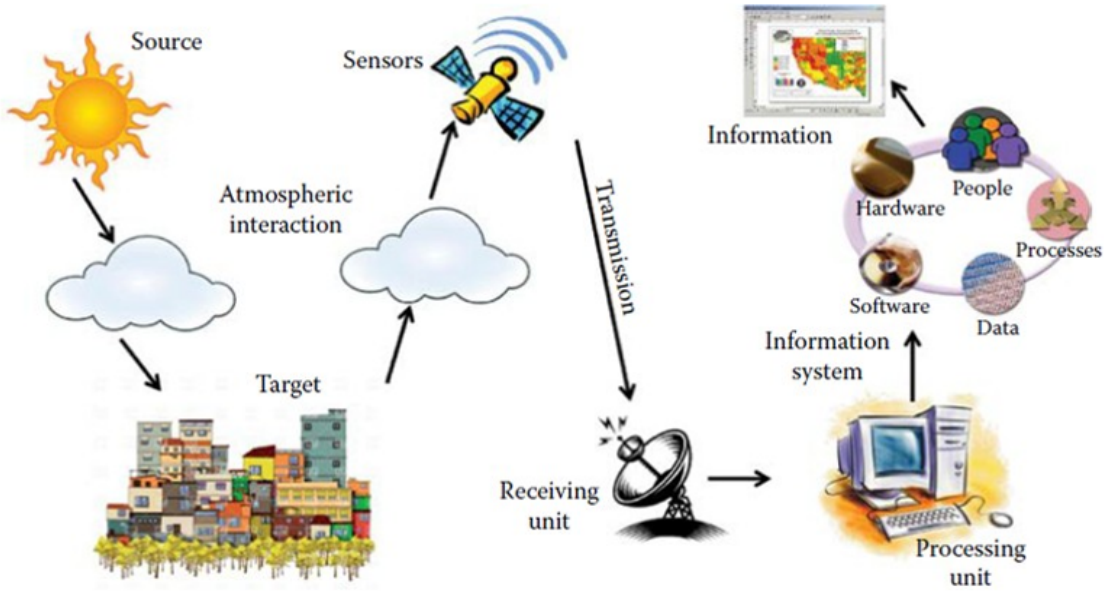

Imagen 2-1: Esquema de la generación de datos de teledetección. Tomado de "Multisensor Data Fusion and Machine Learning for Environmental Remote Sensing", Chan, y Bai, (2018), pag. 2.

En la imagen 2-1 se muestra un esquema con el paso a paso que sigue el proceso de adquisición de las imágenes de satélite y los elementos que el intervienen. A continuación, se describen detalladamente los actores que hacen posible el proceso de teledetección.

2.1.1. Fuente de energía:

En el proceso de teledetección se requiere una fuente de energía, ya sea de origen natural o artificial. El sol es la principal fuente de energía electromagnética natural, la cual es amplia- mente usada por la mayoría de los sensores pasivos. Esta energía emitida por el sol es transmitida en forma de ondas electromagnéticas (Borra et al., 2019). Cada objeto sobre la superficie de la tierra absorbe una cantidad de energía y refleja otra en determinada longitud de onda; gracias al principio de reflexión, se pueden identificar claramente diferentes elementos sobre la superficie de la tierra, como, por ejemplo, cuerpos de agua limpios y contaminados, vegetación, carreteras asfaltadas, cultivos, etc.

Al rango en que se distribuye la radiación se le denomina espectro electromagnético, el cual está dividido básicamente en siete zonas de acuerdo con la longitud de onda, siendo el espectro visible una pequeña porción del amplio espectro. En la imagen 2-2 se puede apreciar el espectro EM y la relación (inversamente proporcional) entre la frecuencia y la energía respecto a la longitud de onda.

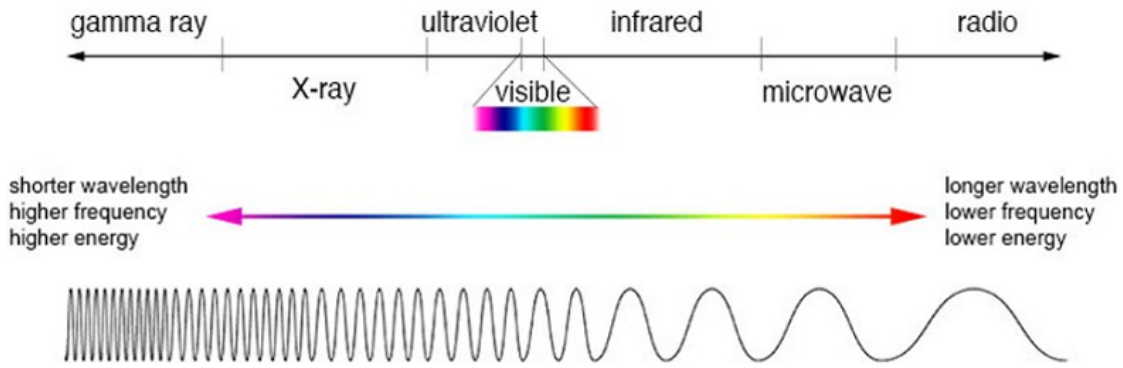

Imagen 2-2: Comparación de longitud de onda, frecuencia y energía del espectro electromagnético. (NASA's Imagine the Universe, tps://imagine.gsfc.nasa.gov/science/toolbox/ emspectrum1.html, accedida marzo, 2020.)

Para la teledetección se usan las regiones del espectro, ultravioleta (UV), visible, infrarrojo (IR), el cual a su vez tiene las regiones de energía, IR emitida e IR termal; y por último el espectro de microondas (Canada Center for Remote Sensing - CCRS).

En el espectro visible (0.4 - 0.7 µm) se encuentran las longitudes de onda de los colores primarios, azul (0.446 - 0.5 µm), verde (0.5 - 0.578 µm), rojo (0.620 - 0.7 µm), y a partir de ellos se puede conseguir toda la gama cromática.

2.1.2. El medio de propagación:

Las ondas electromagnéticas emitidas desde la fuente de energía viajan a través de la atmosfera e interactúan con ella.

Los gases y las partículas presentes en la atmosfera como motas de polvo, vapor de agua, polen y gotas de agua, entre otras, pueden causar cambios de dirección de la radiación electromagnética, generando el fenómeno de dispersión.

Otro fenómeno que se presenta es el de absorción, en el cual las moléculas de la atmosfera absorben parte de la radiación; esto implica que no toda la energía emitida por el sol llega a la superficie de la tierra(Tso y Mather, 2009).

2.1.3. Interacción con el objetivo:

La radiación que logra incidir sobre la superficie es reflejada, transmitida (pasa a través) o absorbida por los objetos (CCRS). La energía reflejada es la que es recolectada por los

sensores de los satélites y es expresada en la imagen en forma de intensidad de brillo.

2.1.4. Sensores:

Los sensores pueden ser activos o pasivos, dependiendo de la fuente que genera la energía que es direccionada al objeto de estudio. En el caso de los sensores pasivos también conocidos como sensores ópticos, requieren de una fuente de energía externa (el sol) que incida sobre superficie de la tierra para ser posteriormente recolectada por estos, por lo tanto, tiene limitada operación en las noches cuando no hay luz solar. Por el contrario, los sensores activos transmiten pulsos de energía electromecánica en forma de ondas microondas (Radar) o laser (Lidar) y reciben de vuelta señales eco, pueden tomar imágenes sin importar la presencia de nubes y durante las 24 horas del día, la desventaja es el alto costo de funcionamiento (Moser y Zerubia, 2018).

2.1.5. Características de las imágenes:

A continuación, se describen las características de las imágenes que son relativas a la resolución del sensor, como son la resolución espacial, la resolución espectral, la resolución temporal y la resolución radiométrica y un ejemplo gráfico en la imagen 2-3.

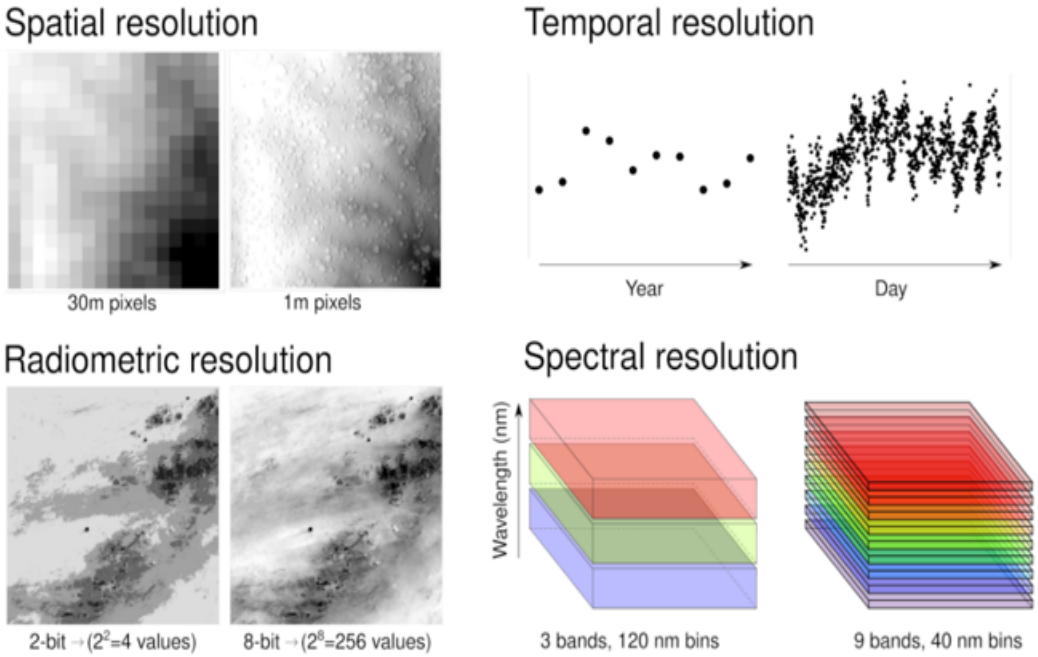

Imagen 2-3: Ejemplos de baja y alta resolución espacial, temporal, radiométrica y espectral. Tomada de "Ten simple rules for working high resolution remote sensing data", Mahood et al, 2023.

- **Resolución espacial:**
  Está asociada al área en la que el sensor puede medir la energía reflejada para discernir la característica más pequeña (Thenkabail, 2016). Va desde un metro hasta varios kilómetros. Cada sensor aporta diferente resolución espacial dependiendo de la banda espectral que se esté analizando.

- **Resolución espectral:**
  Es la capacidad del sensor de definir intervalos de longitud de onda en el espectro electromagnético o el número de bandas que un sensor puede detectar (Thenkabail, 2016). Se tienen así, imágenes pancromáticas (una única banda), multiespectrales (varias bandas, 2, 8, 11, etc.) e hiperespectrales (200 bandas), por ejemplo, el sensor hyperion del satélite.

- **Resolución temporal:**
  Es el tiempo que tarda el satélite en hacer un recorrido completo sobre la superficie de la tierra, denominado también, tiempo de revisita, va desde 99 días hasta 1 día para los satélites comerciales.

- **Resolución radiométrica:**
  Es la sensibilidad del sensor de detectar diferencias muy pequeñas de energía. El valor máximo de brillo dado a cada píxel está afectado por la unidad fundamental de un sistema binario o "bit", los cuales se expresan en potencias de 2.
  Por ejemplo, un sensor con resolución radiométrica de 8 bits tendrá: $2^8 = 256$ valores, esta es la escala de brillo y toma los valores de 0 a 255 (CCRS). A estos valores que se corresponden en una escala de grises se les denomina Digital Number o número digital y representan el valor de la reflectancia en cada píxel.

- **Tamaño:**
  Para explicar el tamaño que puede llegar a tener una imagen satelital, se toma como ejemplo una escena del sensor MSI-2 de Sentinel, esta tiene una dimensión de 290.000 m x 290.000m, con una resolución espacial de 20 m, eso deja una imagen con 290.000/20= 14.500 pixeles, los cuales tienen un DN que requieren 12 bits de almacenamiento en un rango de 0 a 4095; entonces el tamaño promedio de esta imagen es alrededor de 14500*14.500*12= 2 GB, lo que demuestra la gran capacidad de almacenamiento requerida para procesar imágenes satelitales.

Habiendo explicado las características que definen una imagen de satélite, se procede a expresar su composición con base en estos conceptos previos. Una imagen obtenida por un sistema de teledetección es una matriz bidimensional de dimensión $n\ x\ n$ pixeles, en donde cada píxel, no es un componente único, sino que es un vector de dimensión

b, que corresponde al número de bandas espectrales de la imagen, en donde cada píxel de dicho vector tiene un valor asociado denominado *Digital Number-DN*, que indica el nivel de intensidad de energía reflejada por el objeto visualizado en escala de grises (Mahood et al., 2023). Debido a que la imagen es un arreglo matricial, se le denomina y se almacena como dato ráster (Lavender y A., 2023), a continuación, se expresa una imagen satelital en su forma matricial:

$$X_{n,b} = \begin{pmatrix} x_{1,1} & ..... & x_{1,n} \\ \vdots & \ddots & \vdots \\ x_{b,1} & ..... & x_{b,n} \end{pmatrix} \tag{2-1}$$

En donde b es el número de bandas y n el número de pixeles en cada una, a partir de la ecuación 2-1 se puede definir una observación en teledetección como un vector X con b componentes así:

$$X = [X_1, X_2, \cdots, X_b] \tag{2-2}$$

En la ecuación 2-2 X es un vector observado de la matriz de la ecuación 2-1, en donde X_1 representa la reflectancia o el *Digital Number DN* de la banda i - ésima, por tanto, $X_i, \in D_i$ y D= $[0, 1, 2, \ldots .DN]$ en donde DN o escala de grises depende de la resolución radiométrica de la imagen (8 bits o 255 y 32 bits o 4095). En la ecuación 2-3 se define el espacio de observaciones ϕ como el conjunto de todos los valores posibles que puede tomar una observación X:

$$\phi = \bigotimes_{i=1}^{b} D_i \tag{2-3}$$

En donde $\otimes$ denota el producto cartesiano.

2.2. Generalidades del Aprendizaje automático - Machine Learning

El aprendizaje automático o de máquinas, es una de las aplicaciones de la inteligencia artificial, en la que la máquina, se dedica a escrutar ("aprender de") cantidades ingentes de datos, con el objetivo de encontrar o identificar patrones que le permitan hacer predicciones o modelar comportamientos y realimentarse constantemente (Dangeti, 2017). A lo que se le denomina "maquina", es en realidad, una variedad de algoritmos, los cuales son a grandes

rasgos, una secuencia de instrucciones ordenadas (entrada, procesamiento, salida); estos se aplican dependiendo de la necesidad específica que se tenga, ya sea predictiva o descriptiva (James et al., 2013). Por "datos", se hace referencia a cualquier vector/matriz de características que aporte información, por ejemplo, datos numéricos (edad, temperatura, cantidad de llamadas en un mes, etc.), caractéres (discursos, encuestas, opiniones, estados de ánimo, etc.) e imágenes del espectro electromagnético (imágenes de teledetección, fotografías de rostros, animales, ecografías, rayos X, etc.).

2.3. Aprendizaje automático enfocado a aplicaciones de mapeo geológico con imágenes satelitales:

En el presente apartado, se va a profundizar en los enfoques que tienen lugar dentro del marco del estudio propuesto; es decir, se explorarán los algoritmos de aprendizaje automático supervisado y no supervisado más usados para hacer mapeo geológico o discriminación litológica a través de clasificación de imágenes de satélite o teledetección. Dentro de los algoritmos de clasificación supervisada más usados en teledetección se encuentran Ranfom Forest, Classification and Regression trees, Minimum Distance, Naive Bayes y Support Vector Machine (Shirmard et al., 2022), dentro de los algoritmos de clasificación no supervisado se encuentra el K-means.

A continuación, se mencionan las etapas más relevantes e imprescindibles llevadas a cabo durante el proceso de clasificación.

- Adquisición y preprocesamiento de imágenes satelitales.
- Procesamiento y mejora de imágenes satelitales: extracción de características.
- Clasificación.
- Evaluación de precisión de la clasificación.

2.3.1. Adquisición y preprocesamiento de imágenes:

Existen dos tipos de imágenes provenientes de sensores remotos a las que se puede acceder; las que están disponibles gratuitamente a todo el público y las de uso comercial (Lavender y A., 2023); la diferencia radica en que en la segunda opción se debe pagar por su adquisición y en su defensa se puede recalcar que superan en calidad en cuanto a resolución espacial (hasta 1m), espectral (hiperespectrales hasta 220 bandas), radiométrica (16 bits) y temporal (hasta 1 día de tiempo de revisita) respecto las de adquisición libre. De acuerdo con la aplicación o área del estudio, se evalúan las características que deben cumplir las imágenes y se define

la fuente de datos más apropiada, con base en los requerimientos espaciales y espectrales y radiométricos.

- Fuente: Actualmente existen múltiples opciones para acceder a las imágenes de teledetección, la mayoría de ellas son de libre acceso y se encuentran disponibles para ser usadas abiertamente. Para aplicaciones de geología y estudio de la cobertura de la tierra son ampliamente usadas las imágenes multiespectrales de los satélites TERRA, LANSAT y SENTINEL, las cuales son de dominio público y se pueden descargar de las páginas oficiales de la National Aeronautics and Space Administration NASA y de la European Space Agency ESA. A continuación, se relacionan algunos links para la adquisición de imágenes de libre acceso:

  - https://landsat.gsfc.nasa.gov/data/where-to-get-data/
  - https://earth.esa.int/eogateway/activities/pi-community
  - https://www.copernicus.eu/en/copernicus-services
  - https://www.geoportal.org/
  - https://www.usgs.gov/landsat-missions/data/
  - https://oceancolor.gsfc.nasa.gov/
  - https://earthengine.google.com/

En la tabla 2-1, se mencionan algunos de los satélites más relevantes en el ámbito científico (estudio del clima, del ambiente, uso del suelo, geología, agricultura, calidad del agua, estudio de desastres, urbanismo, entre otros), y se describen sus características de resolución espectral (número de bandas y ubicación en el espectro EM), espacial, temporal (tiempo de revisita) y radiométrica, así como los tipos de sensores (pasivos) que llevan a bordo y su fecha de lanzamiento.

Tabla 2-1: Satélites más usados con sus respectivas características

Satelite	Resolucion			
	Espectral	**Espacial**	**Temporal**	**Radiometrica**
LANDSAT 8 sensores: OLI TIRS Lanzamiento :2013	11 Bandas 1,4,8 Luz visible 0.433um 12.5um	30m (visible, NIR,SWIR) 100m (thermal) 15m (Pancromatico)	Ciclo: 16 días Periodo 99 días	8 bits,escala de grises 0 a 255
SENTINEL 2 Sensor: MSI Lanzamiento: 2015	12 Bandas Banda 1 Bandas 2- 3 Bandas 5 -7 Banda 8 Banda 10 Banda 11 - 12	60m (Aerosol) 10 m (Visible) 20 m (Red Edge) 20 m (NIR) 60 m (Vapor de agua) 60 m (SWIR - Cirrus) 20 m (SWIR)	Ciclo: 10 días Periodo 99 días	12 bits,escala de grises 0 a 4095
	Sensor Aater			
TERRA Sensores: ASTER, CERES,MISR MOPITT, MODIS Lanzamiento: 1999	14 Bandas Bandas 1 - 3 Bandas 4 - 9 Banda 10 - 14	15m (SWIR) 30m(NIR) 90m (Thermal)	Ciclo: 16 días Periodo 99 días	8 bits,escala de grises 0 a 255 12 bits, escala de grises 0 a 4095
	Sensor Hyperion			
EO-1 Sensores: Hyperion, ALY, AC Lanzamiento: 2000	Hyperion: 200 Bandas 70 (VNIR) 172(SWIR) ALI:9 Bandas	Hyperion: 30m (VNIR, SWIR) ALI 10m (Pan) 30m (multiesp)	Ciclo: 16 días Periodo 99 días	12 bits,escala de grises 0 a 4095
WORLVEW - 2 Lanzamiento 2009 Very High resolution	8 Bandas	0.5 m (Pancromatico) 2m (multiespectral)	1 a 3 días	16 Bits 0 a 65535

Fuente: Elaboracion Propia

2.3.2. Preprocesamiento:

El procesamiento de una imagen digital es una aparte esencial para la clasificación, en este paso se corrigen características producidas por efectos propios de los sensores, factores atmosféricos o de la propia superficie observada, que de otro modo generarían ruido y errores en la predicción (Cardile et al., 2024). Una vez descargadas las imágenes habiendo especificado la fecha o periodo específico de estudio y acotado por ubicación geográfica de interés, en donde se genera el subconjunto de datos (imagen) a trabajar, se procede con la aplicación de las técnicas necesarias de corrección y/o transformación de la imagen, así como el enmascaramiento de nubes. Las nubes y sus sombras reducen la visión de los sensores ópticos

y bloquean u oscurecen por completo la respuesta espectral de la superficie de la Tierra (Cao et al., 2020). Trabajar con píxeles con nubes afecta significativamente la precisión del reconocimiento de patrones y la clasificación en teledetección (Cardile et al., 2024), por lo tanto, la información proporcionada por los algoritmos de detección y enmascaramiento de nubes es fundamental para excluir las nubes y las sombras de las nubes de los pasos de procesamiento posteriores.

En la tabla 2-2 se mencionan las técnicas que usualmente se aplican para corregir aspectos propios del sensor o del medio que interfieren con la calidad de las imágenes y del procesamiento de estas.

Tabla 2-2: Tipos de correcciones o transformaciones que se le deben realizar a las imágenes.

TIPO		FUENTE	PROSEDIMIENTO
Correcciones	Atmosferica	▪ Geometria del sensor	Metodo de correccion absoluta :calculos de transferencia radial como TRAN atmosferica de resolucion moderada(Chan y Bai, 2018)..
		▪ Gases y aerosoles en a atmosfera	Metodo de correccion Relativa : Comparar imagenes de la misma areá con diferentes fechas con una imagen de referencia (Chan y Bai, 2018).
	Radiometrica	▪ Difucion de la luz del sol	Calibrar la energia irradiada y la señal de salida del sensor(Chan y Bai, 2018).
		▪ Irregularidades del sensor, tropografia	La difucion de los rayos del sol puedenprovocar sombras y aclaramiento de zonas lo cual se corrige con tecnicas de analisis de fourier para extarer componentes de baja frecuencia (Chan y Bai, 2018).
	Geometrica	▪ Curvatura de la tierra	Las distorciones geometricas requieren mapear o georreferenciar los pixeles a tra vez de las transformaciones polinomiales (Chan y Bai, 2018).
Transformaciones	Geometrica	▪ Curvatura de la tierra	Euclidiana : Se cambia la posiscion y orientacon de un objeto pormedio de traslacion, rotacion o refliexion (Chan y Bai, 2018).
		▪ Desplazamiento proveientes del sensor	Alfines: Es un metodo de mapeo lineal en la que los objetos conservan la colinealidad y las relaciones de distancia (Chan y Bai, 2018).
			Proyectivas: se transforman los datos de un sistema de coordenadas a otro mediante informacion de proyeccion (Chan y Bai, 2018).
	Mosaico	▪ Areas de estudio muy amplias	Se fucionan varias imagenes para conseguir mayor cobertura espacial (Chan y Bai, 2018).
	Relleno de Huecos	▪ Nubes y mal funcionamiento instrumental	Espacial: Interpolacion espacial, utiliza el valor de pixeles vecinos para rellenar los valores faltantes (Chan y Bai, 2018).
			Temporal: Rellenana los valores faltantes usando informacion historica
			Hibrido: Enfoque combinando espacial y temporal (Chan y Bai, 2018).

Fuente: Elaboracion Propia

- Software de preprocesamiento y procesamiento: Existen diferentes herramientas geomáticas o software de procesamiento de imágenes para el preprocesamiento y procesamiento de datos de teledetección, algunos son de libre acceso y otros requieren adquirir licencia

para su uso; dentro de las más conocidos licenciados están ENVI, ERDAS y ARGIS, dentro de los de código abierto están QGIS, SNAP y GEE entre otros, la ventaja de este último es que ofrece un inmenso catálogo de imágenes corregidas y todo el procesamiento se genera en la nube lo que disminuye sustancialmente el gasto computacional, de almacenamiento y el tiempo de ejecución. La elección de uno y otro radica primordialmente en los objetivos del estudio y los recursos disponibles para llevarlo a cabo, así como el dominio que se tenga de la herramienta, dado que estas herramientas usan su propio lenguaje y algoritmos intrínsecos (Lavender y A., 2023).

2.3.3. Procesamiento y mejora de imágenes:

Los datos provenientes de la teledetección, generalmente son complejos, redundantes y multidimensionales (Chan y Bai, 2018); para llevar a cabo tareas de clasificación con éxito, se deben realizar procedimientos preliminares, que tienen que ver con extracción y selección de características y reducción de dimensiones; los cuales están orientados a garantizar el mejor desempeño, por medio del análisis y comprensión de la información proporcionada por los datos crudos y la aplicación de técnicas que transforman esos datos en materia prima idónea para los algoritmos de clasificación.

La extracción de características o procesamiento hace parte de la fase exploratoria de los datos, e implica un proceso de búsqueda de información relevante y la condensación en un conjunto de variables más compactas, que facilita la fase de clasificación y toma de decisiones(Chan y Bai, 2018). Para llevar a cabo un proceso de extracción de características exitoso, se hace necesario aplicar técnicas de reducción de dimensionalidad y eliminación de ruido para resaltar los contrastes de las características y facilitar así la interpretación.

El procesamiento de imágenes incluye un análisis descriptivo multivariado, a través de técnicas como análisis de componentes principales *(Principal Component Analysis - PCA)*, análisis discriminante lineal *(Linear Discriminant Analysis – LDA)*, mínima fracción de ruido *(Minimum Noise Fraction- MNF)*, análisis de factor *(Factor Analysis – FA)*, análisis de componentes independientes (Independent Component Analysis – ICA) y análisis de componentes vecinos *(Neighborhood Component Analysis - NCA)*, (Borra et al., 2019). Las transformaciones de imágenes se realizan mediante cálculos complejos entre bandas de imágenes usando algebra matricial y matemáticas más avanzadas, que proporcionan información adicional en unas pocas variables (Cardile et al., 2024), lo que favorece la precisión de la clasificación, a continuación, se mencionan las transformaciones de imagen más comunes y las que se llevan a cabo en este estudio.

1. **Transformación de imagen avanzada basada en pixeles.**
   - Manipulación de imágenes mediante expresiones: la relación de bandas y el cálculo

de índices espectrales consiste en la operación aritmética entre bandas para resaltar o generar características de reflectancia de interés de la superficie a estudiar, por ejemplo, detectar específicamente presencia de agua, vegetación, minerales, entre otros. El índice más conocido en la práctica es el NDVI (Normalize Difference Vegetation Index) y mide la cantidad y la vigorosidad (verdor) de vegetación en una imagen, puede tomar valores entre 1 y -1, en donde más cercano a 1 significa más verdor por tanto más vegetación y más saludable y entre más lejos ausencia de vegetación (Kamusoko, 2019). A continuación, en la tabla 2-3 se muestran algunos índices relevantes para realizar mapeo litológico.

Tabla 2-3: Índices espectrales generados a partir de imágenes Santinel-2

Indice	Formula		Definicion
Indice de Vegetacion (Normalized difference Vgetation index)	NDIV	$\frac{(B8-B4)}{(B8+B4)}$	Permite dfinir y visualizar areas con vegetacion en el mapa y medir la salud de la misma
Razon de minerales arcillosos (Clay Mineral Ration index CRM)	CRM	$\frac{B11}{B12}$	Indica la presecnia de de arcillas y alunita
Razon de arcilla limosa y areniscas (Clay and Sandstoene Ratio)	CSR	$\frac{B11}{B2}$	Indica la presencia de arcillas y areniscas
Hierro Ferrico Fe3+	FOR	$\frac{B4}{B3}$	Resalta la precencia de hierro Ferrico
Razon de óxidos de hierro (iron Oxide Ration)	IOR	$\frac{B4}{B2}$	Permite identificar rocas alteradas, (volcanes) que contiene sulfatos que contienes hierro y que sea han oxidado
Caliza	Caliza	$\frac{B4+B6}{B7+B8}$	Consigue resaltar la presencia de diferentes tipos de caliza

Fuente: Elaboracion Propia

- Manipulación de imágenes mediante álgebra matricial (PCA): El análisis de componentes principales es una técnica estadística multivariada de mejoramiento espectral de imagen que tiene como objetivo disminuir la correlación entre las bandas espectrales y maximizar las diferencias entre ellas, construyendo un nuevo conjunto de ejes ortogonales entre sí, no correlacionados (Gupta, 2018), en donde el eje principal contiene la mayor cantidad de dispersión de las bandas iniciales, y contiene la información de brillo de la imagen (intensidad), los sucesivos contienen menos información y el último suele contener el ruido de la imagen (Prost, 2014).

  El análisis de componentes principales se lleva a cabo en tres pasos principales, el primero es el cálculo de la matriz de varianza-covarianza de la imagen multiespectral, el segundo es la extracción de los valores y los vectores propios de la matriz y el tercero la transformación de las coordenadas de espacio de las características por medio de los valores propios (Tso y Mather, 2009).

  Partiendo de la ecuación 2-1, en donde se expresa a una imagen en su forma matricial, se procede a calcular la matriz de varianza-covarianza aplicando la ecuación 2.4 así:

$$C_{b,b} = \begin{pmatrix} \sigma_{1,1} & ..... & \sigma_{1,b} \\ \vdots & \ddots & \vdots \\ \sigma_{b,1} & ..... & \sigma_{b,b} \end{pmatrix} \tag{2-4}$$

$$\sigma_{ij} = \frac{1}{N-1} \sum_{p=1}^{N} (\mathrm{DN}_{p,i} - \mu_i)(\mathrm{DN}_{p,j} - \mu_j) \tag{2-5}$$

En la ecuación 2.5, se tiene que N es el número de pixeles, $\sigma_{i,j}$, es la covarianza de cada par de diferentes bandas, $\mathrm{DN}_{p,j}$ el número digital de un pixel p y la banda i y $\mathrm{DN}_{p,j}$ el número digital de un pixel p en una banda j, μ_i and μ_j son los promedios de los números digitales de las bandas i y j respectivamente. De la matriz de varianza y covarianza los valores propios (λ) son calculados como las raíces de la ecuación característica: $\det(C - \lambda I) = 0$, donde C, es la matriz de covarianza de las bandas e I es la diagonal de la matriz de identidad(Estornell et al., 2013).

A partir de los valores propios, el porcentaje de la varianza explicada por cada componente principal puede ser obtenido calculando la relación de cada valor propio en relación con la suma de todos ellos. Los componentes principales se expresan en la ecuación 2.6 como sigue:

$$y_b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_b \end{pmatrix} \begin{pmatrix} \omega_{1,1} & \dots & \omega_{1,b} \\ \omega_{2,1} & \dots & \omega_{2,b} \\ \vdots & \ddots & \vdots \\ \omega_{b,1} & \dots & \omega_{b,b} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{2-6}$$

Donde Y es el vector de los componentes principales, W la matriz de transformación y X el vector de los datos originales. Los coeficientes de la matriz W son los vectores propios y proporcionan información de la relación de las bandas con cada componente principal. Los vectores propios pueden ser calculados desde la ecuación matriz para cada valor propio λ_k, según la ecuación 2.7;

$$(C - \lambda_k I)\,\omega_K = 0 \tag{2-7}$$

Donde C es matriz de covarianza, es λ_k es el k- esimo valor propio y I es la diagonal de la matriz de identidad y ω_k es el k vector propio.

Calcular y analizar los componentes principales permite mejorar la interpretación de la imagen y es útil para generar nuevas variables (bandas) para los modelos de clasificación, se recomienda usarlos de manera individual o aplicar técnicas de composición

de color entre los diferentes componentes y así revelar presencia de minerales o alteraciones de terreno que en la imagen original no se evidencian claramente (Marghany, 2022).

2. **Transformación de imagen basada en el vecindario:** las transformaciones de imagen basadas en el vecindario permiten que la información contenida en los pixeles que rodean un píxel focal informe a la función que transforma el valor de este, muy útil cuando se tienen valores outliers; estas funciones disminuyen el efecto de dato atípico (Cardile, et al, 2024). Existen varios tipos de funciones para transformación de imagen mediante vecinos a continuación se mencionan las más relevantes en teledetección:
   - **Mediana:** La mediana es un filtro basado en vecindad que se usa para remover ruido en las imágenes, especialmente cuando hay pixeles atípicos o con valores anormalmente altos o bajos. Esta función crea una cuadricula impar de números a partir del píxel focal y sus vecinos, llamada núcleo o Kernel, la mediana de este núcleo reemplaza el valor central de la imagen de salida (Lavender y A., 2023), esto es muy útil cuando hay pixeles brillantes o ruido ocasionado por el sensor o por errores de medición.
   - **Medida de Textura:** dentro de los parámetros básicos para la interpretación de una imagen se encuentran el tono o color, el cual es la medida relativa de brillo en escala de grises de la imagen; la forma, que se refiere al contorno de un objeto; el tamaño, que está directamente relacionado con la resolución radiométrica de la imagen y finalmente la textura la cual expresa el patrón de relaciones espaciales entre los niveles de gris de los pixeles cercanos (Kamusoko, 2019). Este parámetro es una herramienta importante en el análisis de imágenes y la teledetección, que proporciona medidas cuantitativas de la variabilidad espacial en una imagen para describir la textura de la superficie representada, a continuación, se menciona el método más utilizado en esta aplicación en clasificación litológica (Serbouti et al., 2022).
   - **Matriz de co-ocurrencia de escala de grises Gray-level Co-occurrence Matrices GLCM:** la información de textura de una imagen se define por la relación de adyacencia que tienen los tonos de grises entre si dentro de la imagen, la matriz de co-ocurrencia es una aproximación de la función de densidad probabilística conjunta entre pares de píxeles (Kamusoko, 2019), en la matriz se describe la frecuencia con la que se producen diferentes combinaciones de brillo en los pixeles de manera vertical, horizontal y en diagonal, formando una matriz de $n\ x\ n$ de una imagen con n posibles valores de DN, valores digitales (Cardile et al., 2024), la distribución espacial de cada pixel de la imagen se puede expresar así:

$$P_{n,m}\left(d_\theta, N_g\right) = \{((x,y),(x\ +\ a, y\ +\ b)/I(x,y) = n; I(x\ + a, y\ + b) = m)\}$$

(2-8)

En donde el parámetro d es la distancia entre cada par de pixeles en la imagen, Θ es el ángulo (0, 45, 90 o 135) y Ng el número de niveles de gris o lo que es lo mismo el número digital considerado en la imagen *n, m, con n,m=1,2,3,...Ng*, son los niveles de gris de los dos pixeles separados por el vector de desplazamiento d de coordenadas *(a,b) y (x,y)* son las coordenadas espaciales del pixel que tiene el nivel de gris n y $(x + a, y + b)$ del pixel con nivel de gris m (Serbouti et al., 2022).

Después de calculada la matriz de co-ocurrencia se pueden calcular los índices de textura, por ejemplo, segundo momento angular, contraste, entropía, correlación, suma de cuadrados de varianza, momento de diferencia inverso, suma del promedio, suma de la varianza, suma de la entropía, diferencia de la varianza, diferencia de la entropía e información de medidas de correlación (Haralick et al., 1973). A continuación, se expresan las ecuaciones de los índices más usados en la clasificación litológica, contraste y entropía. (Serbouti et al., 2022).

$$CON = \sum_{i,j=1}^{N_g} (i-j)^2 P_{i,j} \tag{2-9}$$

$$ENT = \sum_{i,j=1}^{N_g} P_{i,j} \log P_{i,j} \tag{2-10}$$

En donde $P_i, jdenotala(i,j) - esimaentradainunamatrizdeco - ocurrenciadeescaladegrisesno$ $al., 1973$).

En el cálculo del contraste se asigna un mayor peso a los $P_i, jqueestánmásalejadosdeladiagonalpri$

2.3.4. Clasificación de imágenes de teledetección:

Cuando se habla de clasificación se refiere a realizar la predicción de una respuesta discreta o cualitativa para una observación, que puede tener desde dos a un número finito de categorías, por ejemplo, determinar si un correo recibido es spam o no, detectar si un cliente va a hacer recompra en un tiempo determinado o no o predecir la incidencia de una enfermedad en individuos, etcétera (Cardile et al., 2024). En teledetección, se le llama clasificación al proceso en que se asigna a cada píxel de una imagen, una etiqueta que obedece a unas clases definidas con anterioridad o asignadas por semejanza de características espectrales (Quirós, 2009), mediante un procedimiento automático efectuado por un algoritmo, de este proceso se obtiene un mapa temático en donde se distingue una distribución de las clases en agrupaciones según la aplicación y/o necesidad del estudio.

Dado el volumen de información que representan las imágenes de teledetección, producto del aumento paulatino y relativamente reciente de la resolución espectral, espacial y temporal de los sensores incorporados a los satélites en órbita; se han desarrollado múltiples técnicas automatizadas de extracción de características y clasificación que proporcionan precisión, celeridad y mayor cobertura a los estudios de la superficie de la tierra a través del procesamiento de imágenes. Dependiendo de la disponibilidad y la cantidad de información previa del terreno de estudio, se pueden aplicar los enfoques de aprendizaje automático, supervisado y no supervisado.

A continuación, se detallan los principios y características propias, el objetivo de estudio y los casos en que aplica usar uno u otro enfoque de aprendizaje automático o *Machine Learning*.

- **Aprendizaje supervisado:** es la herramienta más usada en el análisis cuantitativo de imágenes satelitales; tiene como objetivo segmentar el dominio espectral en regiones que coincidan con las clases inherentes de la superficie estudiada de la tierra (Richards, 2013). Se utiliza el enfoque supervisado, cuando se cuenta con información previa de los elementos en la base de datos con la cual se desea modelar la clasificación o predicción. Los modelos de clasificación se construyen a partir de muestras representativas de las clases de la superficie que se desean discriminar, y que van a servir para entrenar los algoritmos. Para este fin, el analista debe recolectar información (generalmente tomada en sitio -*ground truth* de las clases, tipo y cantidad de estas, generando los polígonos o shape files con datos georreferenciados que serán usados para generar el conjunto de entrenamiento que va a permitir al algoritmo identificar los límites de decisión (Tso y Mather, 2009), realizar la extracción de características y generar la separabilidad espectral entre las clases (Borra et al., 2019).

  La ventaja de usar este enfoque de clasificación radica en que los mapas o las áreas clasificadas se pueden reusar o aplicar en otros escenarios en donde se tengan características semejantes; por otro lado, las desventajas están relacionadas con la cantidad de trabajo que supone conseguir las etiquetas de clase en sitio y generar el conjunto de entrenamiento, lo cual también puede incurrir en errores humanos, afectando significativamente el desempeño del proceso de clasificación (Borra et al., 2019).

  Elegir el esquema de clasificación y definir las clases es el primer paso en este proceso de clasificación, seguido de la selección del conjunto de entrenamiento para cada clase; con esta información se estiman las firmas espectrales de clase (entrenamieto del clasificador) en el espacio de características, se procede a asignar las clases espectrales a todos los pixeles de la imagen y a generar el mapa temático, finalmente se evalúa el desempeño del clasificador (Borra et al., 2019).

Para definir estadísticamente un clasificador se debe partir de la notación matricial de una imagen satelital dada en la ecuación 2-1 y retomando el vector de una observación y el espacio de las observaciones en teledetección de las ecuaciones 2-2 y 2-3 respectivamente y se procede a definir a Ω como el conjunto de las distintas potenciales clases de una imagen así:

$$\Omega = \{\omega_1, \omega_2, ..., \omega_j, \omega_0\} \tag{2-11}$$

En donde j es el numero de clases y *Wo* es la clase de rechazo que se le asigna a las observaciones que no tienen certeza en la clasificación.
Se tiene entonces que para toda observación X hay una función *d(X)* $\in \omega$.
El algoritmo tiene que dividir el espacio de observaciones ω en j regiones disjuntas *(Ro, R2, . . . ,Rj)* y construir una regla que identifique cualquier observación X como perteneciente a la clase i si la región en la que esté ese vector está asociada a la misma clase (Quirós, 2009).

Partiendo de la regla de decisión que debe adoptar el algoritmo es preciso entrenarlo a partir de información conocida generando un conjunto de entrenamiento que contenga las observaciones para cada una de las i clases a discriminar junto con la clase conocida y se puede expresar así:

$$\Im = \{(X_1, c_1), (X_2, c_2), ..., (X_n, c_n)\} \tag{2-12}$$

En donde n es el numero de muestras de entrenamiento tomadas y c_i es la clase verdadera de la observación X_i (Quirós, 2009).

La selección de las áreas de entrenamiento debe ser el resultado de un muestreo que tenga en cuenta el criterio de representatividad y validez estadística. El primero criterio indica que se debe tener en cuenta la densidad del elemento que se quiere clasificar, se sugiere evitar muestrear en zonas de transición de clases para evitar errores de clasificación. El segundo criterio hace referencia al tamaño y a la distribución de las muestras, es importante muestrear sobre toda el área de interés de forma uniforme, sin dejar zonas sin etiquetar y el tamaño, aunque depende del algoritmo que se pretende usar, se puede hacer una generalización para definir el tamaño de la muestra y es que el número de pixeles de entrenamiento este dentro de 10 a 30 veces de numero de las bandas espectrales de la imagen (Tso y Mather, 2009).

Los enfoques de clasificación supervisada parten de una serie de suposiciones a partir de la estructura de la información previa y de los datos; en la mayoría de las aplicaciones de teledetección, se sigue un enfoque paramétrico, en donde se asume que la distribución de probabilidad de las clases sigue una distribución normal multivariada (Quirós, 2009), los métodos más usados dentro de este enfoque son la la distancia mínima (mínimum distance MD) y Nayve Bayes (NB) y se describen a continuación:

1. **Método de mínima distancia:** Es un clasificador lineal, de los más usados en teledetección debido a su versatilidad en sencillez de implementación su fácil, interpretación. Con este método, cuando el algoritmo encuentra un píxel de identidad desconocida lo etiqueta calculando la distancia euclidiana (D_E^2 en la ecuación 2-13) entre el valor del píxel desconocido y cada centroide de clase por turno. Luego asigna la etiqueta del centroide más cercano al píxel observado (Tso y Mather, 2009). Las fronteras de decisión del algoritmo son hiperplanos de ecuación lineal que separan las regiones y que son perpendiculares a la línea que une las medias entre clases (Quirós, 2009).

$$D_E^2 = (x_i \ - \ \mu_j)^2 \tag{2-13}$$

   Donde x_i es el vector observado del i-esimo píxel y μ_j es el vector de la media del j-esim centroide de clase. La dimensión del vector x_i es igual al número de bandas de la imagen de entrada.

2. **Naïve Bayes:** es un algoritmo estadístico que usa el teorema de Bayes para predecir una clase estimando la probabilidad a posterior más alta para cada clase asumiendo que las entradas para cada clase son condicionalmente independientes (Serbouti et al., 2022). Partiendo del teorema de Bayes Ec 2-14, se tiene que: $Pr(T = k|X = x)$; es la probabilidad a posteriori de que una observación $X = x$ pertenezca a la k-ésima clase, π_k representa la probabilidad a priori y $f_k(X)Pr(X|Y = k)$ es la función de densidad X para una observación que viene de la k-ésima clase (James et al., 2013).

$$Pr\left(Y = \ k|\ X = x\right) = \frac{\pi_k f_k\left(x\right)}{\sum_{l=1}^{K} \pi_l f_l\left(x\right)} \tag{2-14}$$

   Estimar y $f_k(X)$ la función de densidad p-dimensional es un desafío, porque no solo se debe considerar la distribución marginal de cada variable predictora, sino que también se ha de tener en cuenta la distribución conjunta, es decir la asociación entre ellas, y caracterizar esa asociación es muy difícil sin hacer alguna suposición que lo simplifique, por eso se procede a asumir convenientemente que las p covariables son independientes dentro de cada clase, es decir no hay asociación entre las variables predictoras (James et al., 2013). Esta suposición se puede expresar de la siguiente manera:

$$f_k(x) = f_{k1}(x_1) \times \ f_{k2}(x_2) \times ... \times f_{kp}(x_{1p}), \tag{2-15}$$

Donde f_{kj} es la función de densidad del j-esimo predictor entre las observaciones en la k-ésima clase. Integrando la ecuación 2-15 en el teorema de bayes, la probabilidad a posteriori se puede expresar como sigue:

$$Pr(Y = \ k|\ X = x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times \ldots \times f_{kp}(x_{1p})}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{l2}(x_2) \times \ldots \times f_{lp}(x_p)} \tag{2-16}$$

El clasificador Naïve Bayes sigue los siguientes pasos para asignar las clases:

- Estimar las densidades de las variables predictoras en cada clase
- Usar el teorema de Bayes en el modelamiento de las probabilidades a posteriori
- Asignación de la clase a la observación que produzca la máxima probabilidad posterior (Borra et al., 2019).

El enfoque no paramétrico no está restringido a suposiciones previas de distribuciones, que es el caso cuando se trabaja con imágenes multi-sensor o multi-temporales (Camps y Bruzzone, 2009). Dentro de los clasificadores no paramétricos, los más comunes son Random Forest (RF), Support Vector Machine (SVM), Classificaction and regression trees (CART). A continuacion se detallan los métodos de clasificación supervisada más usados en teledetección y sus características:

1. **Aprendizaje de árboles de decisión:**

   El aprendizaje basado en árboles de decisión crea un modelo jerárquico, en donde se proponen un conjunto de reglas de decisión sobre las observaciones de entrada, que siguen un marco técnico de capas sobre las que se va generando la subdivisión hasta llegar a conjuntos muy pequeños simulado un árbol (Breiman, 2001) que consta de nodo raíz, ramas, nodos internos y nodos terminales (hojas) (Chan y Bai, 2018). Es esta estructura de árbol, las hojas representan las clases, los atributos están representados por los nodos intermedios. Dependiendo de la naturaleza de la variable respuesta los árboles de decisión se pueden dividir en arboles de regresión (variables continuas) o árboles de clasificación (variables discretas). Dentro de los algoritmos más usados para realizar clasificación en teledetección son *Classsification and Regression Trees CART* y *Random Forest RF*.

   a) Classification and Regression Trees:

   CART es un algoritmo binario que implementa una función de impureza, que a través del índice de Gini examina la bondad de la división de cada nodo

(Breiman et al., 2017). La medición de impurezas del nodo t se puede expresar de la siguiente manera:

$$i\left(t\right)=1-\sum_{j=1}^{K}P_{j}\left(t\right)^{2} \tag{2-17}$$

En donde $P_j(t)$ indica la probabilidad a posteriori de la clase j presente en el nodo t. Esta probabilidad se puede definir como la proporcion entre el numero de muestras de entrenamiento que van en el nodo t etiquetadas como clase j y el total de numero de muestras de entrenamiento con en el nodo t (Chan y Bai, 2018), asi:

$$P_{j}\left(t\right)=\frac{N_{j}\left(t\right)}{N\left(t\right)},\ j=1,2,...,k \tag{2-18}$$

La bondad de la división s, para el nodo t, se puede calcular como sigue:

$$\Delta i(s,t)=i(t)-P_{R}i(t_{R})-P_{L}i(t_{L}) \tag{2-19}$$

Donde P_R y P_L son las proporciones de las muestras en el nodo t que caen al descendiente derecho t_R y al descendiente izquierdo t_L, respectivamente. La bondad de estas divisiones debe maximizarse para lograr la impureza más baja a cada paso hacia la mayor pureza en los nodos terminales. El criterio de parada para el algoritmo CART se define así (Breiman et al., 2017):

$$max_{s}\Delta i(s,t)<\beta \tag{2-20}$$

En donde β es un umbral predeterminado. El árbol dejará de crecer cuando se cumpla el criterio de detención lo que quiere decir que el proceso de entrenamiento del clasificador ha sido completado, indicando en los nodos terminales la clase asignada.

b) Random Forest:

El bosque aleatorio: es un clasificador se construye bajo el concepto de múltiples arboles de decisión (Cardile et al., 2024), por lo que adopta el nombre de "bosque", en donde cada árbol se genera con un vector aleatorio muestreado independientemente del conjunto de entrenamiento de vectores de entrada, entonces puede predecir las clases utilizando un voto mayoritario de acuerdo con la partición de múltiples arboles de decisión a través del índice de Gini y así elegir el umbral de división de las variables de entrada (Breiman, 2001). Debido a el contexto jerárquico de los árboles de decisión, este método es muy útil para la minería de datos y contribuyen a los avances en clasificación de imágenes de teledetección (Tso y Mather, 2009).

2. **Support Vector Machine:**

   Las máquinas de soporte vectorial son un algoritmo que separa o clasifica las observaciones de prueba creando un hiperplano o perímetro de decisión, sobre la distribución en el espacio de las características del conjunto de entrenamiento, maximizando el margen entre dos o más clases (Serbouti et al., 2022). La ubicación del hiperplano o límite de decisión depende sólo de un subconjunto de puntos de datos de entrenamiento más cercanos a él. Este subconjunto de puntos de datos de entrenamiento más cercanos al límite de decisión se conoce como vectores de soporte (Kamusoko, 2019).

   Clasificación con límites de decisión no lineales: Los clasificadores de soporte vectorial teóricamente son de naturaleza binaria, es decir que la frontera de decisión entre dos clases es lineal, sin embargo, en teledetección usualmente se trata de clasificar más de 2 clases, lo que produce límites de decisión no lineales entre clases, en este caso se considera ampliar el espacio de características mediante transformaciones polinomiales llamadas Kernel (James et al., 2013). Partiendo del producto interno de dos observaciones x_i, x_i esta dado por:

$$\langle x_i, x_{i'} \rangle = \sum_{j=0}^{p} x_{ij}, x_{i'j} \tag{2-21}$$

   El clasificador SVM lineal se puede expresar como sigue:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x_i, x_{i'} \rangle \tag{2-22}$$

   En donde hay n parámetros α_i, $i = 1, \ldots, n$, y para estimarlos se requiere $n(n-1)/2$ productos internos $\langle x_i, x_{i'} \rangle$ entre todos los pares de las observaciones de entrenamiento. Ahora en la ecuación 2-22, se va a reemplazar el producto de las dos observaciones por una función generalizada Kernel; entonces se produce la función Kernel lineal, que cuantifica la similitud de dos observaciones usando la correlación de Pearson y se expresa como sigue (James et al., 2013):

$$K(x_i, x_i') = \sum_{j=1}^{P} \langle x_i, x_{i'} \rangle \tag{2-23}$$

Cuando un clasificador de soporte vectorial se combina con un kernel no lineal el clasificador resultante se conoce como máquina de soporte vectorial. La ecuación 2-24 es un kernel polinomial de grado d, que conduce a un límite de decisión más flexible, ya que equivale a instalar un clasificador de vectores en un espacio de dimensiones

superiores que involucra polinomios de grado d (James et al., 2013).

$$K\left(x_i, x_i'\right) = 1 + \sum_{j=1}^{P} \langle x_i, x_{i'} \rangle^d \tag{2-24}$$

Otro ejemplo de kernel no lineal es el kernel radial, donde es una constante positiva. El Kernel radial tiene un comportamiento muy local, en el sentido de que sólo las observaciones de entrenamiento cercanas tienen un efecto en la etiqueta de clase de una observación de prueba (James et al., 2013).

$$K\left(x_i, x_i'\right) = exp\left(-\gamma \sum_{j=1}^{P} \left(x_i, x_{i'}\right)^2\right) \tag{2-25}$$

En comparación con muchos métodos estadísticos tradicionales, las máquinas de soporte vectorial tienen un mejor rendimiento al abordar muchos problemas de clasificación litológica, en gran medida gracias a la optimización de los parámetros asociados por medio de la validación cruzada (James et al., 2013). Algunas de las ventajas de usar SVM es que proporciona flexibilidad para elegir el valor umbral para el hiperplano, por medio de la implementación de funciones de Kernel se ocupa de la transformación no lineal, elimina los problemas de sobreajuste (Borra et al., 2019) y se escala relativamente bien a datos de alta dimensión(Kamusoko, 2019).

- **Aprendizaje No supervisado para clasificación de imágenes satelitales:**

  El aprendizaje no supervisado es implementado cuando no se tiene información previa, acerca de las clases de un conjunto de datos, situación muy común, dado que etiquetar la totalidad de los elementos, en especial en datos de teledetección, suele ser muy costoso en términos de tiempo, o sencillamente imposible en terrenos inaccesibles (Gupta, 2018).

  El principio más usado para la clasificación no supervisada de datos de teledetección es el de agrupamiento (Clustering). Los algoritmos en este caso, al no contar con información previa de clases, realizan el agrupamiento de las características espectrales de manera natural, basándose solamente en los valores de reflectancia (digital number – DN) o intensidad de brillo, teniendo en cuenta la proximidad o las similitudes en el espacio espectral de características (Borra et al., 2019), este es un proceso iterativo en el que cada píxel posterior debe compararse con todos los grupos anteriores, lo que implica que su rendimiento depende del número de grupos solicitados, así como del tamaño de los datos; generalmente son procesos costosos en términos tecnológicos y de tiempo. El reto consiste en definir el número de clases que se deben generar; puesto que el algoritmo generara tantos grupos de clases como se le asigne, por tanto, se requiere de experiencia y conocimiento

del área de estudio por parte del analista.

Generalmente las técnicas de aprendizaje no supervisado son usadas como etapa preliminar y exploratoria de los datos, para identificar datos atípicos, así como para la reducción de dimensionalidad o de ruido que puede ser provocado por redundancia entre las variables, entre otras aplicaciones (Dangeti, 2017).

El algoritmo de clasificación no supervisada más usado en teledetección es el de K-medias *(K-means)*, éste divide el conjunto de datos en K grupos distintos; el primer paso es especificar el número de grupos K que se desean formar, posteriormente el algoritmo asignará cada observación a un grupo teniendo en cuenta la distancia mínima al centroide de este y teniendo en cuenta dos condiciones clave, la primera es que cada observación debe pertenecer al menos a un K grupo y la segunda que los grupos no se superpongan, es decir que ninguna observación pertenezca a más de un grupo, esto se logra garantizando la mínima variabilidad dentro de cada grupo y la máxima entre grupos (James et al., 2013).

2.3.5. Evaluación de desempeño de la clasificación de datos de teledetección:

Evaluar la precisión y exactitud es un paso muy importante, pero frecuentemente subestimado en el proceso de clasificación de imágenes satelitales para cartografiar la cobertura y uso del terreno (Thenkabail, 2016). Se considera que la tarea de clasificación ha terminado cuando se evalúa la precisión de los clasificadores, esto es, comparar a través de parámetros cuantitativos el nivel de concordancia entre las etiquetas asignadas y la información de referencia. Un análisis cualitativo es recomendado como complemento para evaluar la importancia de las variables predictoras en el proceso de clasificación (Borra et al., 2019).
En teledetección suelen usarse múltiples métodos para evaluar el rendimiento de los modelos predictivos, unos de los más usados y confiable es la validación cruzada iterativa *(K-fold)*; éste método busca estimar cómo se comportará el modelo con datos no vistos o que no estuvieron en el proceso de entrenamiento y así evitar el sobre ajuste que es muy común cuando se validan los algoritmos con conjuntos de entrenamiento y prueba divididos aleatoriamente, lo que no garantiza que la prueba se realice con datos conocidos por los clasificadores. El método de validación cruzada divide el conjunto de entrenamiento en K particiones y realiza K experimentos, en donde se entrenan los algorimos con *K-1* particiones y se reserva una partición para validar; de manera iterativa turna las particiones para entrenamiento y validación, hasta haber pasado por todos los grupos, garantizando que la validación siempre ocurre con datos que el algotimo no ha visto; entonces se obtienen K resultados de precisión, los cuales se promedian y se obtiene la precisión general del modelo (Golblatt et al., 2016).

Otro método para evaluar cuantitativamente la calidad o precisión de la clasificación es el análisis de la matriz de error de clase o matriz de confusión, la cual compara el mapa final y las etiquetas de referencia (Congalton y Green, 2009).

La matriz es un arreglo de dimensiones $n\ x\ n$ clases, en el que se reporta el número de falsos positivos y negativos, así como verdaderos positivos y negativos.

Tabla 2-4: Matriz de confusión

		Valores Actuales	
		Positivo	Negativo
Valores predichos	Positivo	**TP** (true positive)	**FP** (False positive)
	Negativo	**FN** (False negative)	**TP** (True Negative)

Fuente: Elaboracion Propia

En la tabla 2-4 se observa que las columnas representan las clases asignadas (valores actuales) y las filas las clases reales (predicción)(Chan y Bai, 2018).

TP: Clasificado como positivo y la clase actual es positiva
FN: Clasificado como negativo y la clase actual es positiva
TN: Clasificado como negativo y la clase actual es negativa

Las métricas que se pueden medir a través de la matriz de confusión son:

- Precisión general: es un valor que representa el porcentaje de pixeles clasificados correcta- mente (Borra et al., 2019), y se calcula como el número total de píxeles identificados correctamente sobre el número total de pixeles de la muestra.
- Error de omisión: refleja el error de clasificación, resultado de clasificar pixeles de clase verdadera como pixeles de clase falsa (Borra et al., 2019).
- Error de Comisión: refleja el error de clasificación, resultado de clasificar pixeles de clase falsa como pixeles de clase verdadera (Borra et al., 2019).
- Precisión del productor: refleja la proporción de pixeles clasificados correctamente en una clase, respecto al total de pixeles de esa clase en los datos de referencia, equivale a 1-error de omisión, (Borra et al., 2019).
- Precisión del usuario: refleja la proporción de pixeles clasificados correctamente en una clase, respecto al total de pixeles categorizados en esa clase por el clasificador, equivale a 1-error de Comisión (Borra et al., 2019).

- Coeficiente Kappa (k): es una medida del rendimiento del clasificador derivada de la matriz de error, sin tener en cuenta el sesgo resultante entre la salida del clasificador y los datos de referencia (Richards, 2013).

- El coeficiente toma valores entre 0 y 1, donde 0 significa que no hay acuerdo en la predicción, y a medida que tiende a 1, más cercana esta la predicción del valor de referencia (Chan y Bai, 2018).

3 Objetivos

3.1. Objetivo general

Realizar clasificación litológica en afloramientos de roca mediante técnicas de aprendizaje supervisado y no supervisado usando datos de teledetección (imágenes satelitales).

3.2. Objetivos específicos

- Aplicar algoritmos de clasificación de aprendizaje supervisado y no supervisado, para encontrar combinaciones espectrales optimas de imágenes satelitales, con el fin de discriminar litologías en el área de estudio.
- Seleccionar el mejor modelo de clasificación, mediante análisis de técnicas evaluación de desempeño de los algoritmos empleados.
- Extrapolar los resultados de la clasificación litológica a áreas de contexto geológico similares, en las cuales exista dificultad para realizar geología de campo lo que genera incertidumbre acerca de la validez o calidad de la cartografía existente.

4 Metodologia

4.1. Generación de base de Datos:

4.1.1. Área de estudio:

Para realizar el presente estudio se selecciona un área de geología conocida, es decir, se cuenta con información de referencia, también conocida como *ground thruth*, y su respectivo mapa geológico; la zona seleccionada es árida y se caracteriza por tener un clima desértico y por la ausencia de vegetación, lo que permite minimizar los efectos distorsionadores característicos de las imágenes satelitales como son, efectos atmosféricos, presencia de nubes y/o contaminación, entre otros.

El área de interés *Area Of Interest - AOI* de este trabajo se encuentra localizada en la cuenca de Ouarzazate, en Marruecos, en el límite septentrional de desierto del Sahara y cubre un área de 382 Km^2.Se trata de una zona de geología variada y estudiada ampliamente (Tesón, 2009,), en la que la subjetividad del mapa geológico puede ser considerada muy baja siendo por tanto ideal para ser utilizada como información de referencia para validar los algoritmos de clasificación supervisada.En la figura 4-1, se detalla el área usada para este trabajo sobre el mapa de Marruecos; dicha sección es visualizada a través de la combinación de color verdadero (RGB) de una imagen del satélite Sentinel 2A MSI y cortada posteriormente, definiendo así, el área de interés que será usado en las secciones siguientes.

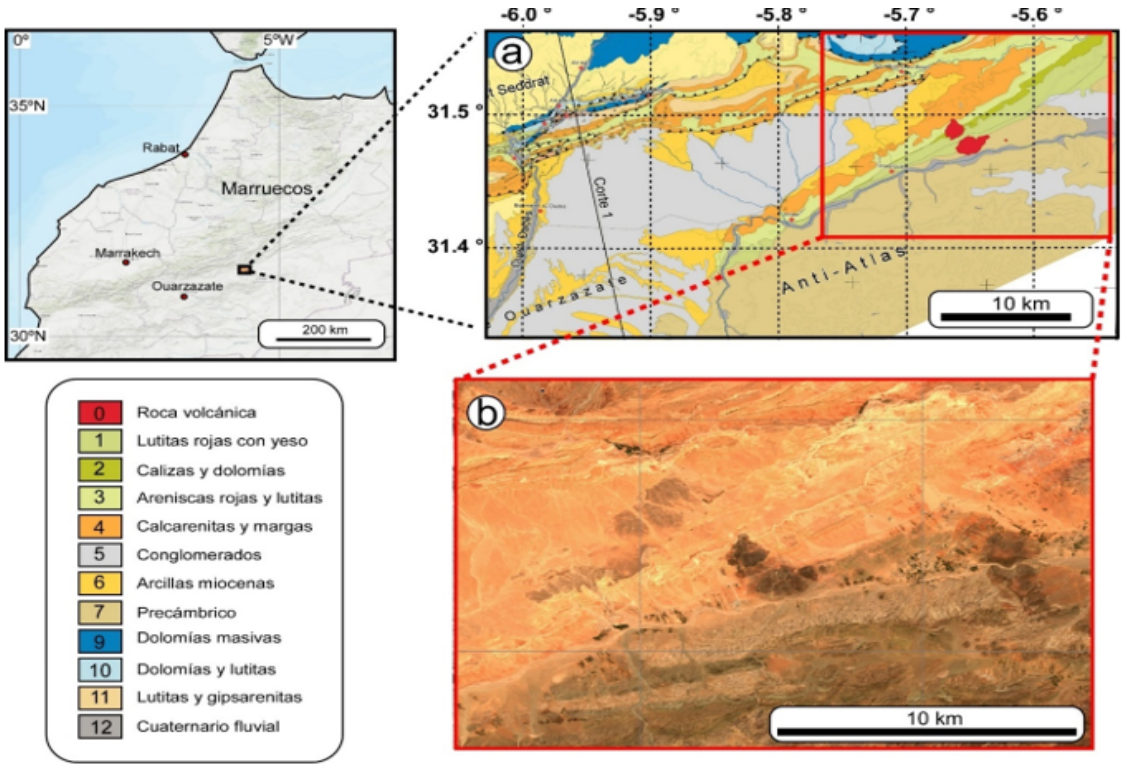

Imagen 4-1: . Área de interés a). Mapa geológico (grounf thruth) b. Imagen Sentinel 2A MSI sobre el AOI, el corte obedece a la zona elegida para clasificación litológica. Visualización de las andas RGB o true color..

4.1.2. 2 Selección de Datos:

Para llevar a cabo el objeto de estudio, se eligió el satélite SENTINEL-2 el cual es una misión europea de imágenes multiespectrales de alta resolución y amplia gama, lanzado en 2015. La especificación completa de la misión de los satélites gemelos que vuelan en la misma órbita, pero con una fase de 180° está diseñada para ofrecer una alta frecuencia de revisita de cinco días en el ecuador. Este satélite lleva a bordo un instrumento óptico el MSI (MultiSpectral Instrument) que mide la radiancia reflejada de la Tierra en 13 bandas espectrales que van desde VNIR hasta SWIR de las cuales cuatro bandas tienen 10 m, seis bandas tienen 20 m y tres bandas a 60 m de resolución espacial, el ancho de la franja orbital es de 290 km. Ver tabla 4-1. Este satélite se utiliza para respaldar una variedad de servicios y aplicaciones ofrecidos por Copernicus, incluida la gestión de tierras, agricultura, silvicultura, control de desastres, operaciones de ayuda humanitaria, mapeo de riesgos y cuestiones de seguridad.
Los parámetros que se usaron para elegir la fuente de los datos se describen a continuación:

- **Accesibilidad:** Para este estudio se opta por usar imágenes de libre acceso, pues en la actualidad existen múltiples fuentes abiertas al público, desde donde se puede

acceder de manera ágil y sencilla a grandes catálogos de datos con excelente calidad provenientes de misiones gubernamentales y comerciales.

- **Antecedentes de estudio respecto a la aplicación:** Recientes estudios han arrojado resultados satisfactorios realizando clasificación litológica con imágenes del Satélite Sentinel 2 MSI, en comparación con imágenes ASTER y Landsat; Ge, W., Cheng Q., Tang, Y., Jing, L. y Gao Ch., 2018 en Mongolia, China y Bentahar, I. y Raji, M., 2020 en el Atlas Central de Marruecos cuando realizaron clasificación litológica encontraron que la precisión de la clasificación supervisada mejoraba en alrededor de un 5 % cuando se trabajaba con imágenes Sentinel, respecto a Landsat y Aster.

- **Resolución espectral y espacial:** Las Imágenes del sensor MSI de Sentinel tienen mejor resolución espacial en las bandas de onda corta SWIR1 y SWIR2 y de infrarrojo cercano, respecto a los satélites Landsat y Aster, ver tabla 2-1. Así mismo las imágenes provenientes del sensor MSI tienen mejor resolución radiométrica (12bits, escala de grises 0-4095) ver tabla 4-1, respecto a las provenientes de Landsat y Aster (8 bits, escala de grises 0-255), a excepción de la banda térmica de Aster que tiene (12 bits, escala de grises 0-4095). Una mejor resolución espacial y radiométrica produce mejores métricas de clasificación porque aumenta la precisión de las observaciones muy conveniente para entrenar los algoritmos de clasificación.

Tabla 4-1: Características de las imágenes de Sentinel 2 MSI

BAND		**WAVELENGTH (min-max in micrometers)**	**SPACIAL RESOLUTION (meters)**
Band 1	Coastal Aerosol	0.421 - 0.457	60
Band 2	Blue	0.439 - 0.535	10
Band 3	Green	0.537 - 0.582	10
Band 4	Red	0.646 - 0.685	10
Band 5	Vegetation red edg	0.694 - 0.714	20
Band 6	Vegetation red edg	0.731 - 0.749	20
Band 7	Vegetation red edg	0.768 - 0.796	20
Band 8	NIR (near infared)	0.767 - 0.908	10
Band 8a	NIR (near infared)	0.848 - 0.881	20
Band 9	Narrow NIR	0.931 - 0.958	60
Band 10	Cirrus	1.338 - 1.414	60
Band 11	SWIR (Short wave infared)	1.539 - 1.681	20
Band 12	SWIR (Short wave infared	2.072 - 2.312	20

Fuente: Elaboracion Propia

4.1.3. Ground thruth:

es la información de referencia usada para generar las clases litológicas para el proceso de clasificación. Esta información ha sido recolectada en campo por profesionales de geología y plasmado en el mapa geológico que se muestra en la imagen 4-1, en donde se aprecia el mapa de la zona seccionado por el AOI.

4.1.4. Herramientas computacionales:

Para el desarrollo de este trabajo de investigación se usa Google Earth Engine, que es una plataforma interactiva de cómputo en la nube que cuenta con un catálogo a escala de petabytes de imágenes de satélite y de datos geoespaciales disponibles para descarga, visualización y procesamiento; se caracteriza por tener una infraestructura dedicada al cómputo de alto desempeño lo que convierte en una herramienta muy potente gracias a su API (Application Programming Interface) disponible en los lenguajes Python y JavaScript; cuenta con un editor de código en la web mediante el cual se puede acceder a imágenes de diferentes satélites y procesarlas mediante complejos algoritmos para el análisis geoespacial, en muy corto tiempo y con alta precisión. Además, es una plataforma de libre acceso que tiene como principal objetivo brindar una herramienta poderosa para al análisis e investigación de la superficie de la tierra para fines científicos primordialmente.

4.2. Preprocesamiento:

El preprocesamiento es una parte primordial en la clasificación litológica porque en este paso se genera la geodatabase, cuya calidad afecta decisivamente la precisión de la clasificación; a continuación, se mencionan las pautas que se siguen en el preprocesamiento el cual se realiza desde el paso cero en la plataforma GEE:

- Selección y descarga en el editor de código de GEE de la colección de imágenes contenidas en el sensor elegido según el paso 4.1.2.
- **Enmascaramiento de nubes:** en este paso se calcula la probabilidad de nube por píxel en la colección de imágenes originales, proporcionando un método flexible para enmascarar con precisión píxeles nublados, mediante una función generada en GEE junto con la banda de calidad QA60 de Sentinel-2. Se genera una base de datos libre de nubes y preparada para ejecutar procedimientos de clasificación.
- **Filtrado por metadatos:** A este punto se tiene una colección inmensa de imágenes tomadas desde que empezó a operar el sensor alrededor de la tierra; es preciso disminuir ese tamaño y seleccionar las que interesan para el objeto de estudio. El primer filtro es el rango de fecha sobre el que se quiere trabajar, este dependerá del interés de estudio, en este caso se busca una época seca, lo más libre de nubes y vapor de agua en la

atmósfera. El segundo filtro es por geometría es donde se acota el área de interés, y el tercero se procede a definir el porcentaje de nubes que se va a permitir en las imágenes.

- **Correcciones propias del sensor o del medio:** En este paso se efectúan las correcciones necesarias para eliminar defectos o ruido que afecten el procesamiento. Para este caso particular, trabajar en GEE tiene un aventaja ya que la colección de imágenes descargadas al inicio están ortorrectificadas (corrección de reflectancia) y corregidas atmosféricamente

- **Interpretación por medio de composición de color:** La visualización es importante para interpretar la geodatabase, para esto se utilizan técnicas de composición de color verdadero y falso color, en donde se combinan las diferentes bandas espectrales y se visualizan poder apreciar el AOI desde diferentes perspectivas.

4.3. Procesamiento:

La primera parte del procesamiento es generar las variables predictoras que permitan la extracción de características y el reconocimiento de patrones que van a aportar información específica para el posterior entrenamiento de los algoritmos de clasificación. A continuación, se mencionan las técnicas usadas para resaltar información espectral y topográfica para generar nuevas variables:

4.3.1. Extracción de características:

1. Análisis de índices espectrales y relación de bandas (Band ratio): se aplican técnicas transformación de las imágenes para generar características que servirán como insumo en el entrenamiento de los algoritmos de clasificación. Estas técnicas generan nuevas variables (bandas) de entrada que aportan información especifica y detallada en forma de atributos para facilitar la discriminación de las unidades litológicas durante la clasificación. Como se puede apreciar en la imagen 4-2, las unidades litológicas tienen una mayor reflectancia en el rango de 1600 nm del espectro o lo que es lo mismo en la B11 (SWIR), lo que indica que las variables predictoras generadas a partir de operaciones aritméticas de banda, alrededor de este rango del espectro, proporcionan información importante en el modelamiento y clasificación. A continuación de proponen los siguientes índices para generar atributos que resalten minerales propios de la litología estudiada en las diferentes bandas (ver imagen 4-3):

    - **Índice de minerales arcillosos** – *CMR (Clay Mineral Ratio)*: el CMR es un índice geológico para identificar arcilla y alunita mediante la razón entre bandas infrarrojas de onda corta SWIR 1 y SWIR 2, muy útil a la hora de realizar mapeo

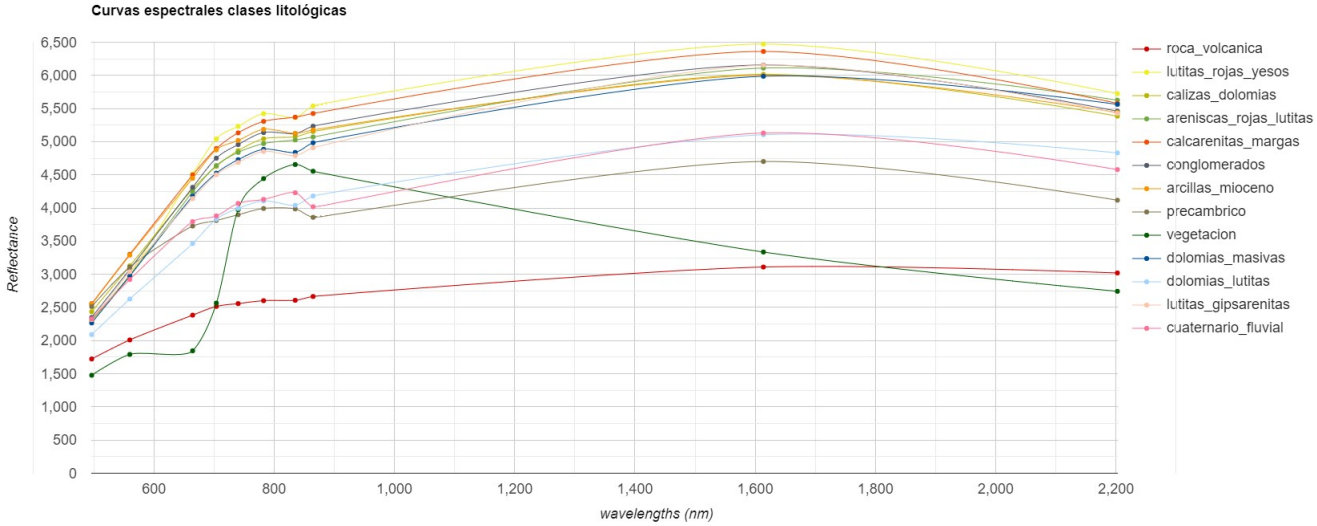

Imagen 4-2: Curvas espectrales de las unidades litológicas.

litológico (Cardile et al., 2024).

$$CMR = \frac{B11\ (SWIR\ 11)}{B12\ (SWIR\ 12)} \tag{4-1}$$

- **Índice de óxido de hierro -** *IOR (Iron Oxide Ratio)*: el IOR es un índice geológico que permite identificar rocas alteradas hidrotermalmente por ejemplo de volcanes, que contienen sulfuros que contienen hierro y que se han oxidado; se calcula mediante la razón de la banda del rojo y el azul (Cardile et al., 2024).

$$IOR \frac{B4\ (Red)}{B2\ (Blue)} \tag{4-2}$$

- **Caliza:** es una relación de bandas que consigue resaltar la presencia de diferentes tipos de caliza, siendo esta una roca sedimentaria, que gracias al ground thruth se sabe que esta presente en el área de estudio. Se calcula mediante la operación aritmética ente las bandas B4, B6, B7 y B8, (Tangestani y Shayeganpour, 2020) así:

$$caliza = \frac{(B4\ (Red) + B6\ (Red\ Edge\ 2))}{(B7\ (Red\ Edge\ 3) + B8\ (Nir)} \tag{4-3}$$

- **NDVI:** el índice de vegetación se calcula para evidenciar la escasa vegetación presente en el área de interés y se calcula como sigue:

$$NDVI = \frac{(B8 - B4)}{(B8 + B4)} \tag{4-4}$$

2. DEM-Modelo digital de terreno: en los modelos de clasificación se incluye información de elevación y pendiente *(slope* del terreno que aporta información valiosa a la discriminación ltológica, dado que la zona estudiada tiene una topografía accidentada (Serbouti et al., 2022. Para este fin, se usa la base NASADEM *(Digital Elevation Model from NASA)* del catálogo de datos de GEE, que consisite en un reprocesamiento de datos STRM *(Shuttle Radar Topography Mission)* con el análisis se incluye información de elevación del terreno que aporta información valiosa a la con precisión mejorada al incorporar datos auxiliares de los conjuntos de datos ASTER (NASA, 2020).

3. PCA - Análisis de componentes principales: Las transformaciones de imagen basadas en píxel como el análisis de componentes principales, son una buena alternativa para reducir la dimensionalidad y para generar variables con nuevos atributos (componentes principales) que se incluyen en el modelo de clasificación, porque revelan información espectral que no es evidente en la imagen original.

4. TEXTURA- Transformación de imagen basada en vecindad: En el mapeo litológico la información de textura es clave en el proceso de extracción de características, para este estudio se aplica el método de Matriz de co-ocurrencia de escala de grises, GLCM *(Gray Level co-ocurrence Matrix)*. Con la matriz de co-ocurrencia calculada, se generan dos índices de textura claves para el mapeo litológico, la entropía y el contraste. Se generan tantos índices de textura como bandas espectrales tenga la geodatabase.

5. Clasificación no supervisada, algoritmo k-medias: por la naturaleza de la clasificación no supervisada que consiste en el auto aprendizaje organizado, en donde gracias a un algoritmo de agrupamiento (K-means) los pixeles son segmentados de acuerdo a su similitud espectral, suele ser muy útil para determinar una descomposición de clases espectrales de los datos y en este estudio es usado como una variable adicional en los modelos de clasificación supervisada (Richards, 2013).

4.3.2. Apilamiento de variables predictoras y normalización de geodatabase:

- Apilamiento de variables: Se procede a generar varios conjuntos de datos que serán tenidos en cuenta para los modelos de clasificación supervisada.

- Normalización de las características: Las variables generadas previamente, se deben normalizar antes de generar los modelos de clasificación, los algoritmos de machine learning funcionan mejor cuando todas las características tienen el mismo rango (Cardile et al., 2024). Se aplica una función para normalizar la imagen con todas las nuevas variables generadas, en donde los valores de pixeles deben estar entre 0 y 1.

Tabla 4-2: Modelos propuestos para aplicar a los clasificadores.

	VARIABLES PREDICTORAS
Modelo 0	Bandas espectrales propias de la imagen: B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12.
Modelo 1	Bandas + Modelo digital de terreno. B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12, DEM.
Modelo 2	Bandas + Cluster + Modelo digital de terreno. B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12, K-means, DEM.
Modelo 3	Bandas + Cluster + Modelo digital de terreno + Índices B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12, K-means, DEM, cmr, ior, caliza.
Modelo 4	Bandas + Cluster + Modelo digital de terreno + Índices + PCA B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12, K-means, DEM, cmr, ior, caliza, pc1, pc2, pc3, pc4, pc5, pc6, pc7, 8, pc9, pc10
Modelo 5	Bandas + Cluster + Modelo digital de terreno + Índices + PCA + Textura B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12, K-means, DEM, cmr, ior, caliza, pc1, pc2, pc3, pc4, pc5, pc6, pc7, 8, pc9, pc10, E2, E3, E4, E5, E6, E7, E8, E11, E12, C2, C3, C4, C5, C6, C7, C8, C11, C12

4.3.3. Clasificación supervisada: La clasificación supervisada se lleva a cabo mediante 3 etapas primordiales, las cuales se especifican a continuación:

- Selección de datos de Entrenamiento: En la práctica, el proceso de entrenamiento consiste en usar información de referencia en este caso el mapa geológico, para construir las muestras de entrenamiento, que generaran un conjunto de patrones espectrales para cada categoría; estos patrones son los que servirán para construir la regla de clasificación que, con posterioridad, se utilizará para asignar al resto de píxeles de la imagen. Para esto se tienen en cuenta las siguientes consideraciones:

  1. Definir las clases: en este caso se especifican las unidades litológicas que se quieren clasificar. Se incluye la vegetación como una clase, porque a pesar de estar trabajando sobre una zona desértica, hay algo de vegetación que se debe tener en cuenta para evitar errores en la clasificación, ver tabla 4-3.

  2. Generar las muestras de entrenamiento: estas pueden ser puntos o polígonos; en este caso se tomaron polígonos dada la extensión del área de estudio ($382Km^2$), ver Figura(4-3). Estos se trazan uno a uno sobre el mapa geológico previamente importado al Code Editor de GEE. Los polígonos son trazados sobre zonas homogéneas evitando muestrear sobre bordes de clase.

  3. Cantidad: se procura mantener el equilibrio de clases, tomando polígonos que cubran aleatoriamente alrededor del 2 % del área de cada clase y así evitar el desbalanceo o el sobre ajuste en la clasificación, ver tabla 4-3, en donde se especifica el número de polígonos y área por clase y área de entrenamiento.

4. Cobertura: Los polígonos se distribuyen alrededor de toda la imagen, evitando dejar zonas sin etiquetar, garantizando la uniformidad y la cobertura del muestreo.

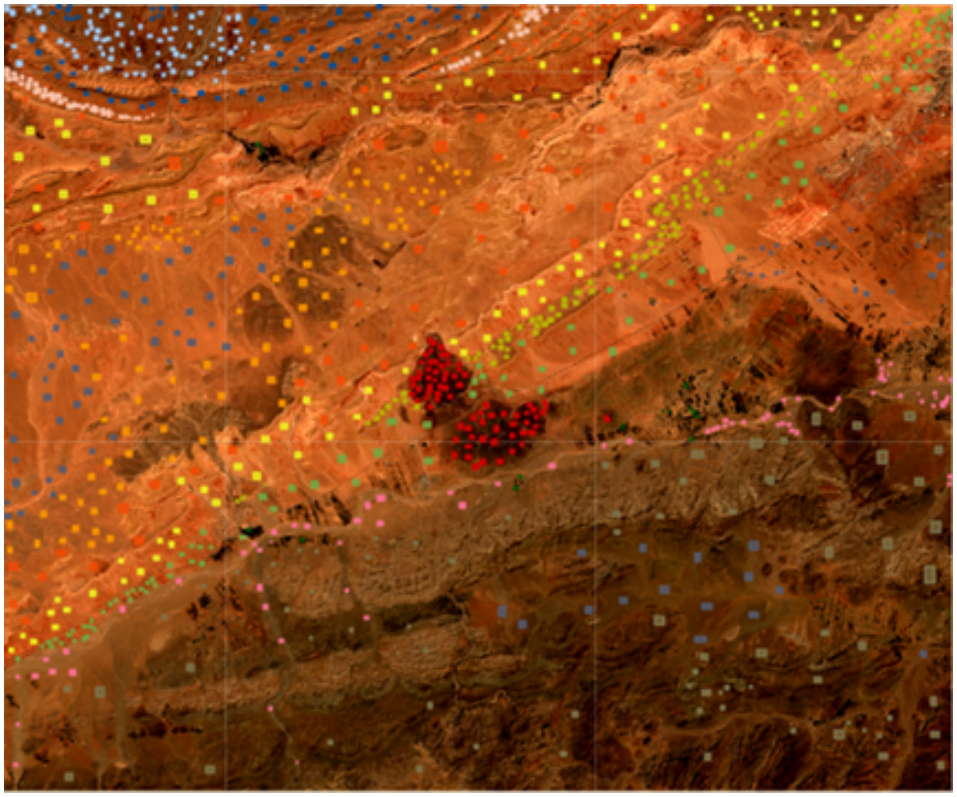

Imagen 4-3: Imagen Color verdadero contrastada con los polígonos de entrenamiento.

Tabla 4-3: Clases litológicas usadas para entrenar los algoritmos

Color	Clase	Unidades litológicas	Escala de tiempo	Poligonos	Area por clase km2	Proporcion
	0	roca_volcánica	Cuaternario	21	4.1	2.26%
	1	Lutitas rojas con yesos	Cretaceo tardio	81	31	2.08%
	2	calizas_dolomias	Cretaceo tardio	54	9.8	1.86%
	3	areniscas_rojas_lutitas	cretaceo temprano	100	41	1.20%
	4	calcarenitas_margas	Paleoceno y eoceno	99	50	1.91%
	5	conglomerados	Cuaternario	100	48.3	1.51%
	6	arcillas_mioceno	Mioceno	58	24.1	1.96%
	7	precámbrico	Precámbico	82	132	1.28%
	8	vegetación	N/A	60	10	0.20%
	9	dolomias_masivas	Jurásico tempranp	70	12.7	1.74%
	10	dolomias_lutitas	Jurásico medio	67	8.3	1.92%
	11	lutitas_gipsarenitas	Mioceno tardio	15	2	1.60%
	12	cuaternario fluvial	cuaternario	80	17.6	1.07%

- Clasificación: El proceso de clasificación inicia con la generación del conjunto de entrenamiento a partir de la geodatabase previamente preprocesada, la coleccion de características de las classes (que es el compendio de los polígonos de las clases) y las bandas y/o variables que serán usadas para la predicción, de acuerdo con cada modelo diseñado (ver tabla 4-2). Con los datos de entrenamiento se procede a entrenar cada algoritmo (CART, RF, MD, SPV, y NB), en este caso se realiza dentro de un proceso de validación cruzada, en donde el conjunto de entrenamiento se divide en 5 grupos diferentes e iterativamente se entrena con 4 grupos y se deja uno para validación, asi sucesivamente hasta haber completado los 5 experimentos. De la validación cruzada se obtienen 5 valores de precisión los cuales se promedian y se logra sacar un valor de precisión general del clasificador, finalmente se procede a generar la matriz de confusión y visualizar la imagen clasificada.

- Evaluación de precisión: La precisión se evalúa de manera cuantitativa mediante el análisis de la precisión general promedio obtenida a partir de la validación cruzada y de la matriz de confusión analizando la precisión del productor o error de comisión y la precisión del consumidor o error de omisión, así mismo se realiza una validación cualitativa por medio del análisis de importancia de las variables o el peso que toma cada una dentro de cada modelo de clasificación.

5 Resultados Técnicas de clasificación basadas en Machine Learning:

5.1. Geodatabase

En el editor de código de GEE, se descarga la colección de datos "Sentinel-2 MSI: MultiSpectral Instrument, Level-2A". El producto Level-2A proporciona reflectancia de superficie ortorrectificada y corregida atmosféricamente, con registro multiespectral de subpíxeles. El producto descargado incluye un mapa de clasificación de escenas (nube, sombras de nubes, vegetación, suelos/desiertos, agua, nieve, etc.). Se usa el parámetro de nubosidad (Banda QA60) para hacer enmascaramiento de nubes.

5.2. Preprocesamiento:

Sobre la colección de imágenes descargada se procede a aplicar los filtros para reducir el tamaño de la geodabase.

5.2.1. Flitro de metadatos:

- Filtro por geometría: Se define el polígono de AOI y se utiliza para la generación de la geodatabase definitiva
- Filtro de fecha: Se elige el periodo sobre el cual se quiere trabajar. En este caso se trabaja con imágenes tomadas entre el 01 de enero al 30 de junio del año 2023.
- Enmascaramiento de nubes: se trae la función de enmascaramiento de nubes y se elige el porcentaje de nubosidad permitido para los datos, en este caso se trabaja con el 10 % de nubosidad.

5.2.2. Interpretación:

Se genera la visualización de la imagen mediante técnica de composición de color.

- True Color composite o composición de color verdadero - RGB: es una técnica de visualización de imágenes mediante la combinación de bandas espectrales a través del

código de color. El color verdadero es conseguido combinando las bandas RGB (Rojo, verde y azul), proporcionando una imagen muy parecida a la real, percibida según el espectro visible, ver imagen 5-1.

- False color composite o composición de false color: Se consigue combinando las bandas RGB con las bandas de infrarrojo cercano y de onda corta, obteniendo imágenes más contrastadas que las de color verdadero, por tanto, requiere una profunda interpretación, dado que los colores apreciados distan de tener sentido real. Este tipo de visualización es muy útil para detectar cambios mineralógicos, o para hacer seguimiento a la salud de la vegetación, entre otras aplicaciones, ver Figura 5-1.

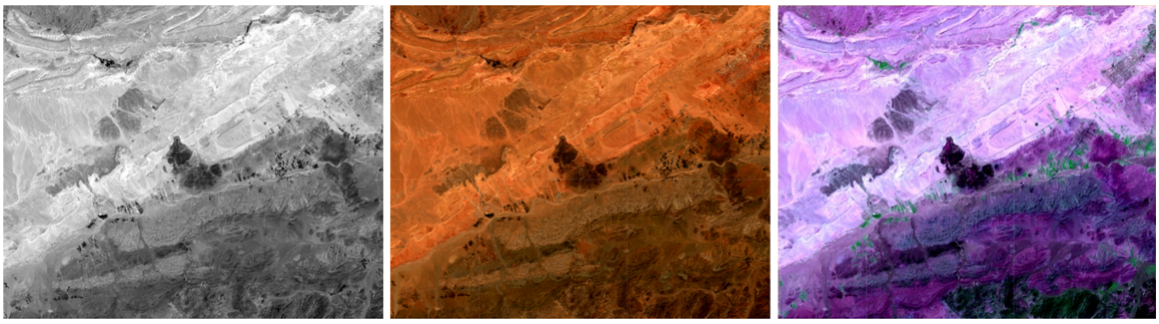

Imagen 5-1: Izquierda: banda B4(rojo). Centro: imagen color verdadero RGB combinación de bandas B4, B3, B2. Derecha: imagen falso color combinación de bandas B12, B8, B11.

5.3. Procesamiento

5.3.1. Análisis de índices espectrales:

A continuación, se muestran las imágenes de los índices generados que van a formar parte de las variables predictoras en los diferentes modelos. En la imagen 5-2 se aprecian los índices de óxidos de hierro, arcillas y calizas, se observa que, a más intensidad de color, mas presencia de esos minerales y/o rocas sedimentarias. Se calculó el NDVI para evidenciar la escasa presencia de vegetación, pero que aun así se tiene en cuenta como una clase en los modelos. Del modelo digital de terreno se extraen 2 imágenes, una que corresponde a la pendiente y otra a la elevación, ver imagen 5-2.

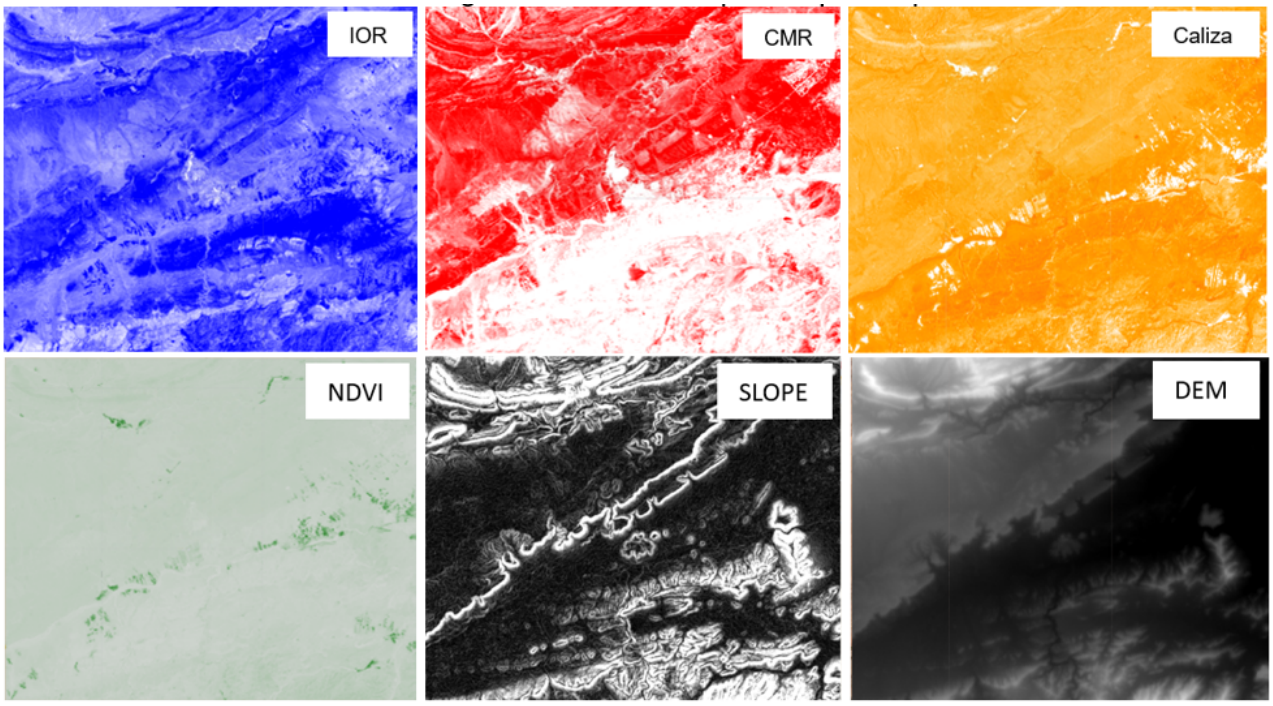

Imagen 5-2: Índices espectrales y modelo digital de terreno.

5.3.2. Transformación de imagen:

- **Análisis de componentes principales – PCA:** se realiza el análisis de componentes principales sobre las 10 bandas espectrales de la imagen y los resultados se reflejan en la tabla 5-1, en donde se aprecian la matriz de valores propios y la varianza de cada componente. En la imagen 5-3 se se aprecian 9 de los 10 compontes generados. Se observa que el componente 1 expresa el 96 % de la variabilidad de la imagen original y representa el albedo o porcentaje de la radiación reflejada por la superficie, las superficies claras tienen valores de albedo superiores a las oscuras. Los componentes 2, 4 y 5 y tienen la principal carga en las bandas 11 y 12, lo que permite concluir que representan las litologías ricas en arcillas y calizas de acuerdo con el índice CMR calculado previamente. Los componentes 8 y 9 tienen la mayor carga en las bandas 2 y 4 que indica presencia de óxidos de hierro, importantes también para este estudio. Con base en este análisis se propone usar todos los componentes principales en los modelos que incluyen análisis de componentes principales y evaluar su impacto en la precisión de la clasificación.

Tabla 5-1: Resultado del Análisis de Componentes Principales.

		Vectores Propios									
Bandas	(nm)	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
B2	496.6	-0.0859	-0.1463	0.4664	0.3032	0.3472	-0.0937	0.2195	-0.6438	0.1945	-0.1787
B3	560.0	-0.1401	-0.1536	0.5360	0.3643	0.2709	0.1570	-0.1090	0.5710	-0.2650	0.1748
B4	664.5	-0.2764	-0.0721	0.4290	-0.3681	-0.2701	-0.3164	-0.2482	0.1976	0.5654	-0.0735
B5	703.9	-0.3115	-0.1042	0.2876	-0.3947	-0.3442	-0.0335	0.2595	-0.1837	-0.6545	-0.0674
B6	740.2	-0.3227	-0.2492	-0.0698	-0.1239	-0.0507	0.5633	0.0570	-0.1972	0.2488	0.6241
B7	782.5	-0.3382	-0.2730	-0.2071	0.0004	0.0503	0.4602	0.0539	0.1677	0.1110	-0.7145
B8	835.1	-0.3488	-0.2842	-0.2634	0.0708	0.1837	-0.2725	-0.7051	-0.2343	-0.2479	0.0408
B8A	864.8	-0.3567	-0.2737	-0.3317	0.1443	0.1113	-0.5102	0.5515	0.2405	0.0829	0.1598
B11	1613.7	-0.4362	0.5052	-0.0275	0.5623	-0.4777	0.0269	-0.0425	-0.0753	0.0312	-0.0107
B12	2202.4	-0.3766	0.6256	-0.0228	-0.3558	0.5789	0.0312	0.0407	0.0360	-0.0174	0.0154
Varianza		96.47	2.05	1.04	0.23	0.13	0.03	0.02	0.01	0.01	0.01
Valor Propio		0.054383	0.001157	0.000585	0.000130	0.000071	0.000020	0.000012	0.000007	0.000004	0.000003

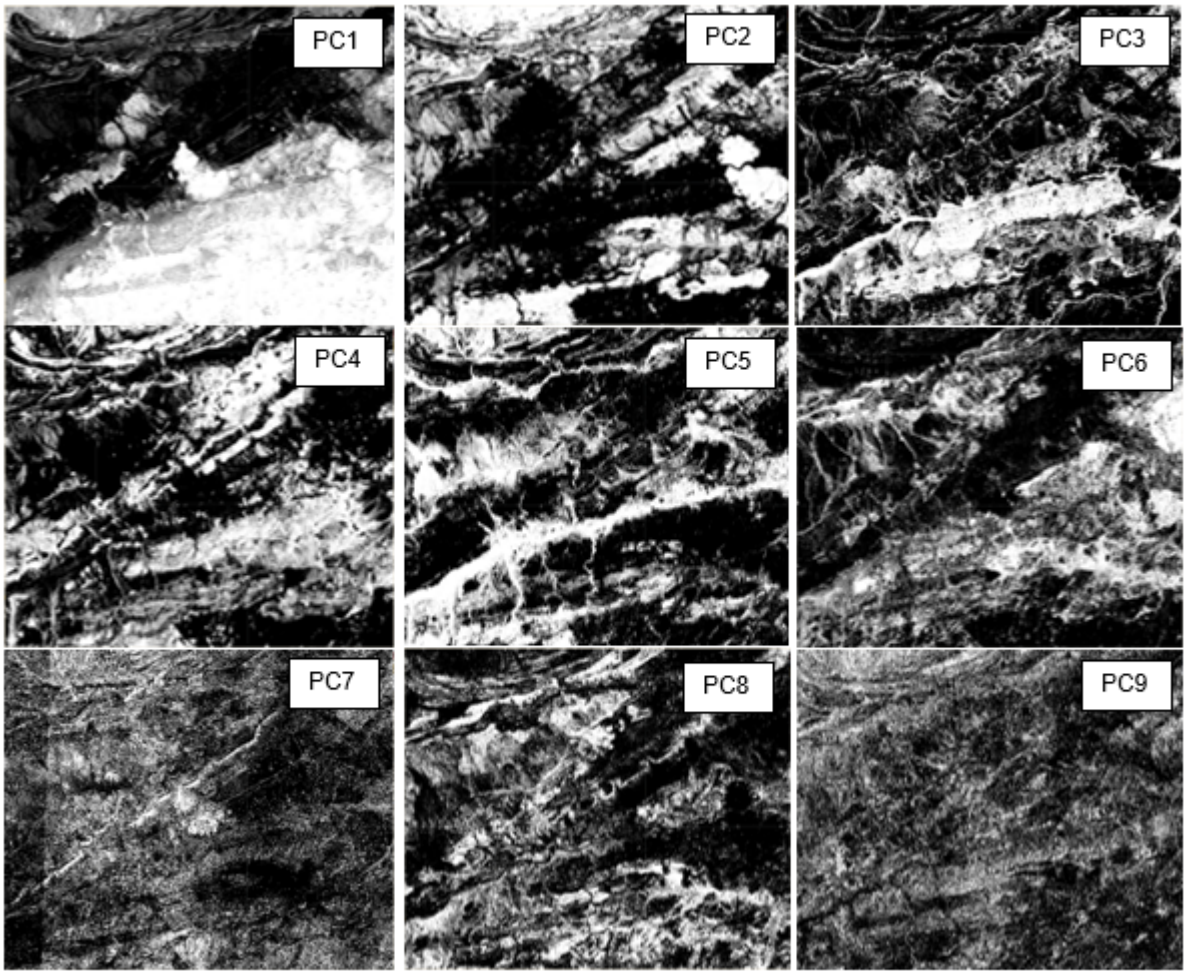

Imagen 5-3: Componentes principales del PCA

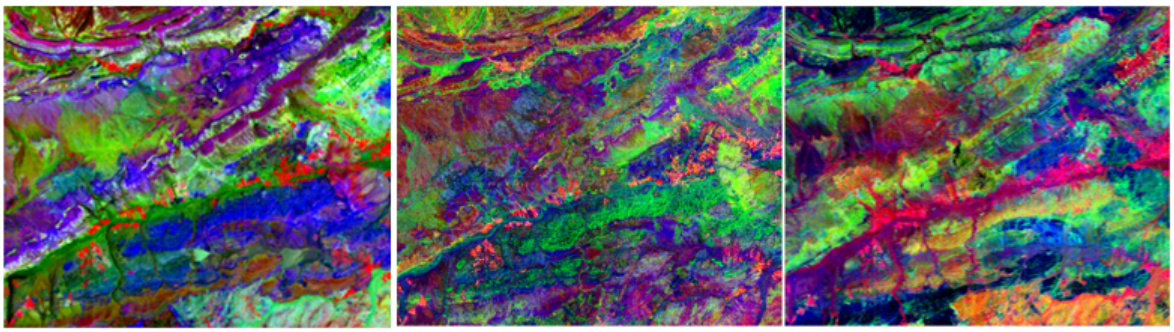

Imagen 5-4: Composición de color de PC. Izquierda PC3, PC4, PC5. Centro PC3, PC8, PC9. Derecha PC2, PC5, PC7.

A modo de interpretación, se muestra la composición de color de algunos componentes principales en donde por ejemplo de resalta muy bien la vegetación a la izquierda de la imagen5-4.

- **Análisis de textura GLCM:**

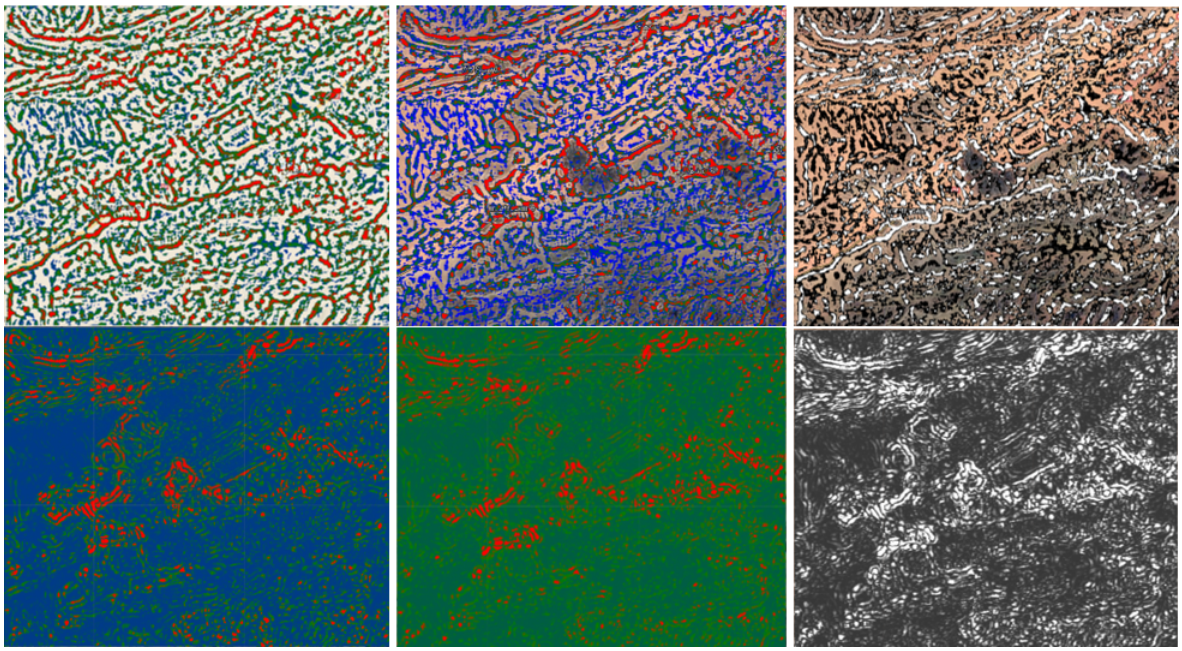

Imagen 5-5: Entropia B2, B8, color composite RGB (arriba) y contraste B2, B8 y RGB (abajo).

1. **Entropía y contraste:** Se generan los índices de entropía y contraste, uno por cada banda, (10 en total), para ser usadas en el modelo 5 de los clasificadores para evaluar el impacto las dos métricas de textura en el desempeño de la clasificación. Las métricas de textura complementan la información proporcionada los

parámetros del modelo digital de terreno. En la imagen 5-5 se aprecian las bandas B2, B8 y color composite de entropía (arriba) y contraste (abajo).

5.3.3. K-means:

Se genera el modelo de clasificación No supervisado (imagen 5-6) con el algoritmo K-means para 13 grupos (son 13 clases conocidas). Esta información es introducida como una variable en los modelos de clasificación superisada propuestos para realizar el mapeo litológico.

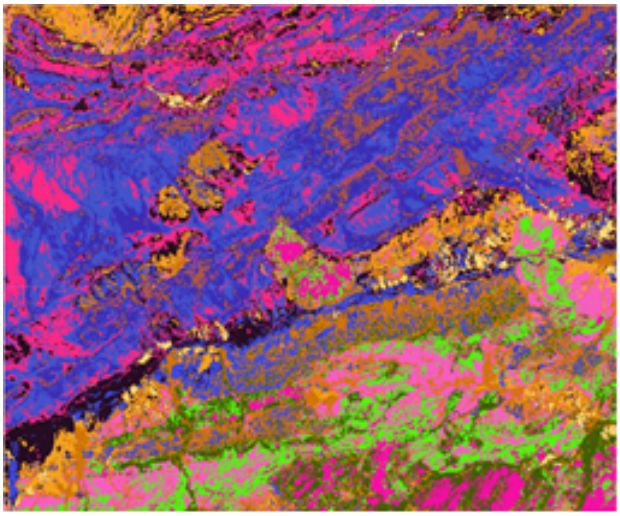

Imagen 5-6: Algoritmo no supervisado k-means.

5.3.4. Clasificación supervisada:

Uno a uno se van generando los modelos de clasificación (ver tabla 4-2), adicionando en cada uno las variables predictoras propuestas previamente normalizadas y así evaluar la importancia de estas en del desempeño de cada modelo. A continuación, se detallan los resultados de cada algoritmo junto con los modelos de variables predictoras propuesto.

1. **Classification and Regression Trees:** En la figura (5-7) se aprecian las imágenes resultantes del algoritmo CART en cada modelo, se puede ver que, a medida que se agregan variables predictoras, mejora visiblemente la predicción, siendo el mejor modelo el que incluye las variables predictoras de índices espectrales, topográficos, PCA y textura juntos, con una precisión del 86 %, aunque se observa que hay clases que fueron clasificadas correctamente y otras presentan errores de predicción. En la tabla 10 se consignan los valores de precisión del usuario y del productor por clase con detalle.

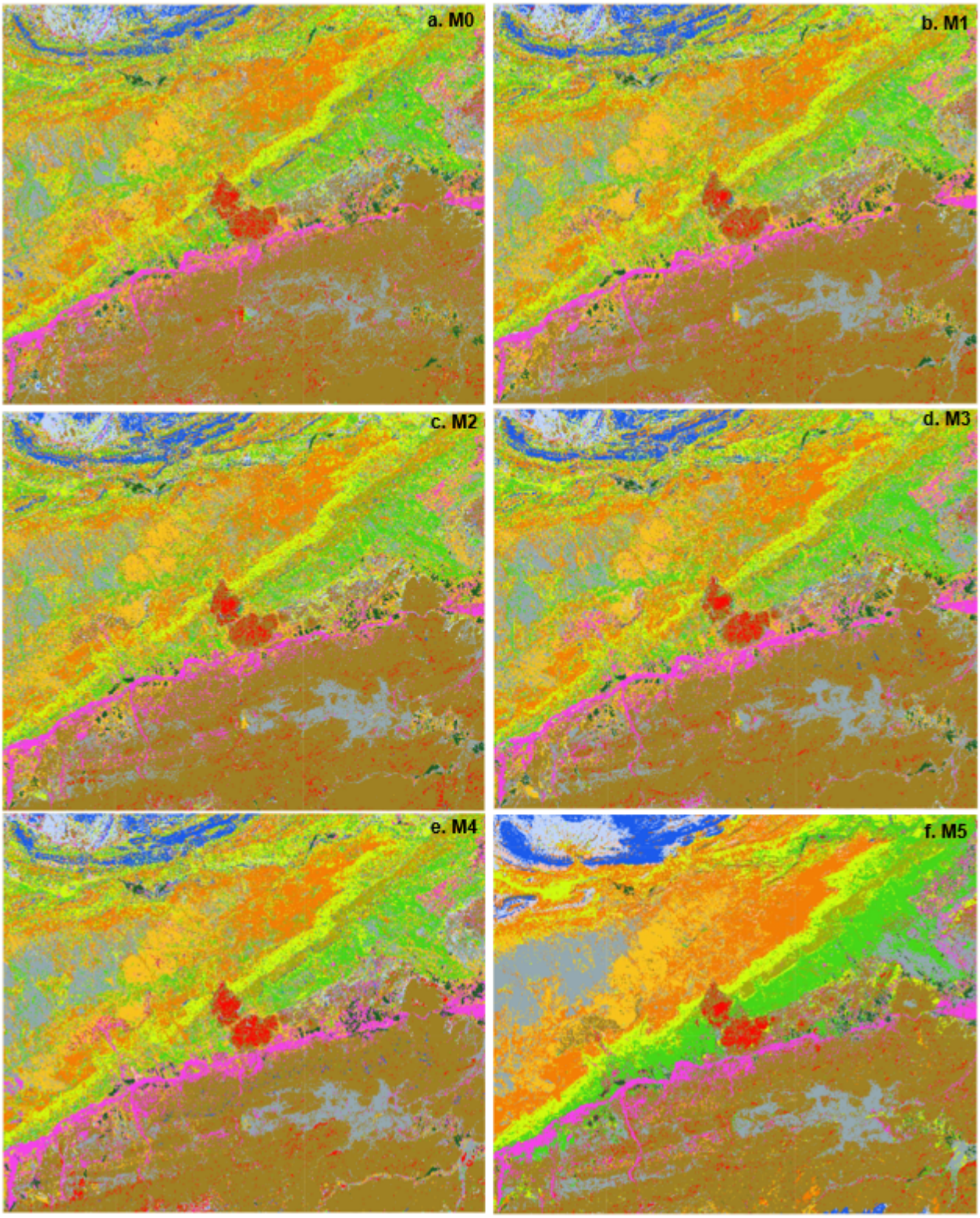

Imagen 5-7: Imágenes obtenidas a través del algoritmo CART. a) M0 (73 %). b) M1 (75 %), c) M2 (76 %), d) M3 (78 %), e) M4 (85 %), f) M5 (86 %).

2. **Random Forest:** En la imagen (5-8) se muestran las imágenes generadas por el algoritmo Random Forest para los modelos propuestos. Este algoritmo presenta mejores métricas que el CART y al igual que este, su desempeño mejora a medida que se entrena con variables predictoras topográficas y de textura.

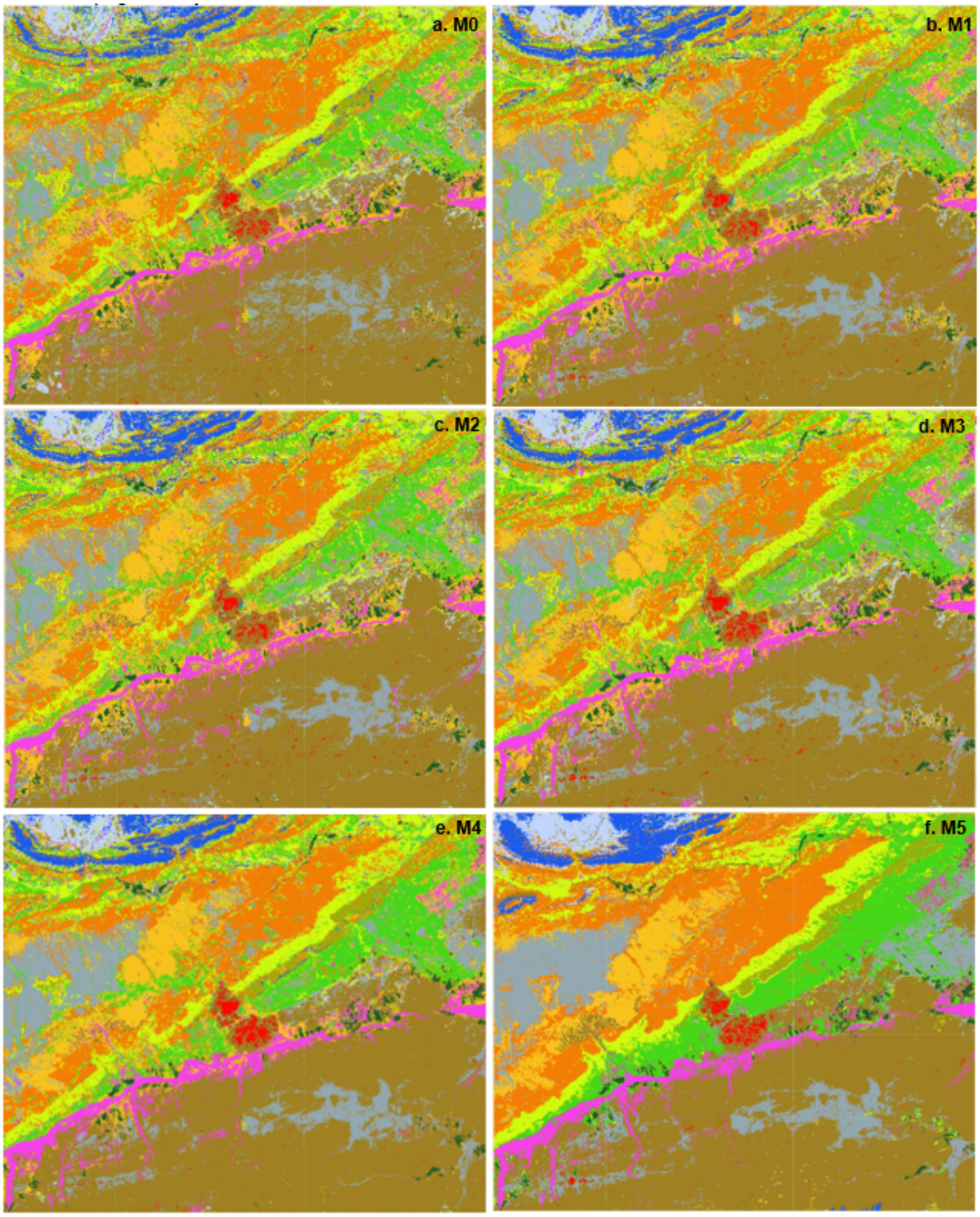

Imagen 5-8: Imágenes obtenidas a traves del algoritmo Random Forest. a) M0 (80 %), b) M1 (84 %), c) M2 (85 %), d) M3 (88 %), e) M4 (93 %), f) M5 (95 %).

3. **Minimum Distance:** los resultados con este algoritmo presentan unas métricas aceptables y aunque el desempeño mejora a medida que se agregan variables, no es el esperado, presentando varios errores de precisión por clase.

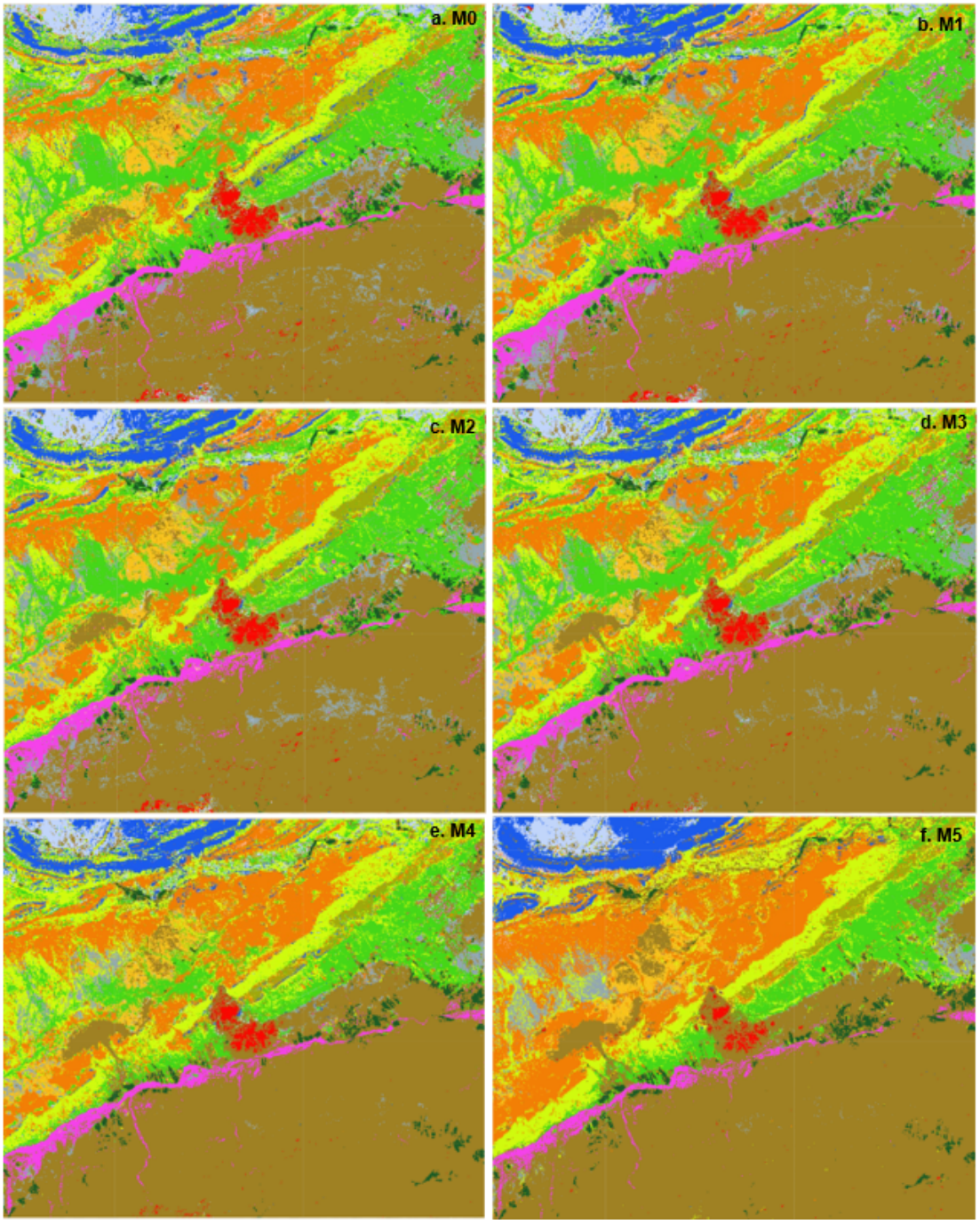

Imagen 5-9: Imágenes obtenidas a través del agoritmo Minimum Distance. a) M0 (6 %) b) M1 (65 %), c) M2 (67 %), d) M3 (68 %), e) M4 (70 %), f) M5 (71 %).

4. **SVM:** Las métricas de los modelos generados por este algoritmo son aceptables y si bien, mejoran a medida que se incorporan información del análisis de componentes principales y de textura, no son tan favorables como los anteriores.

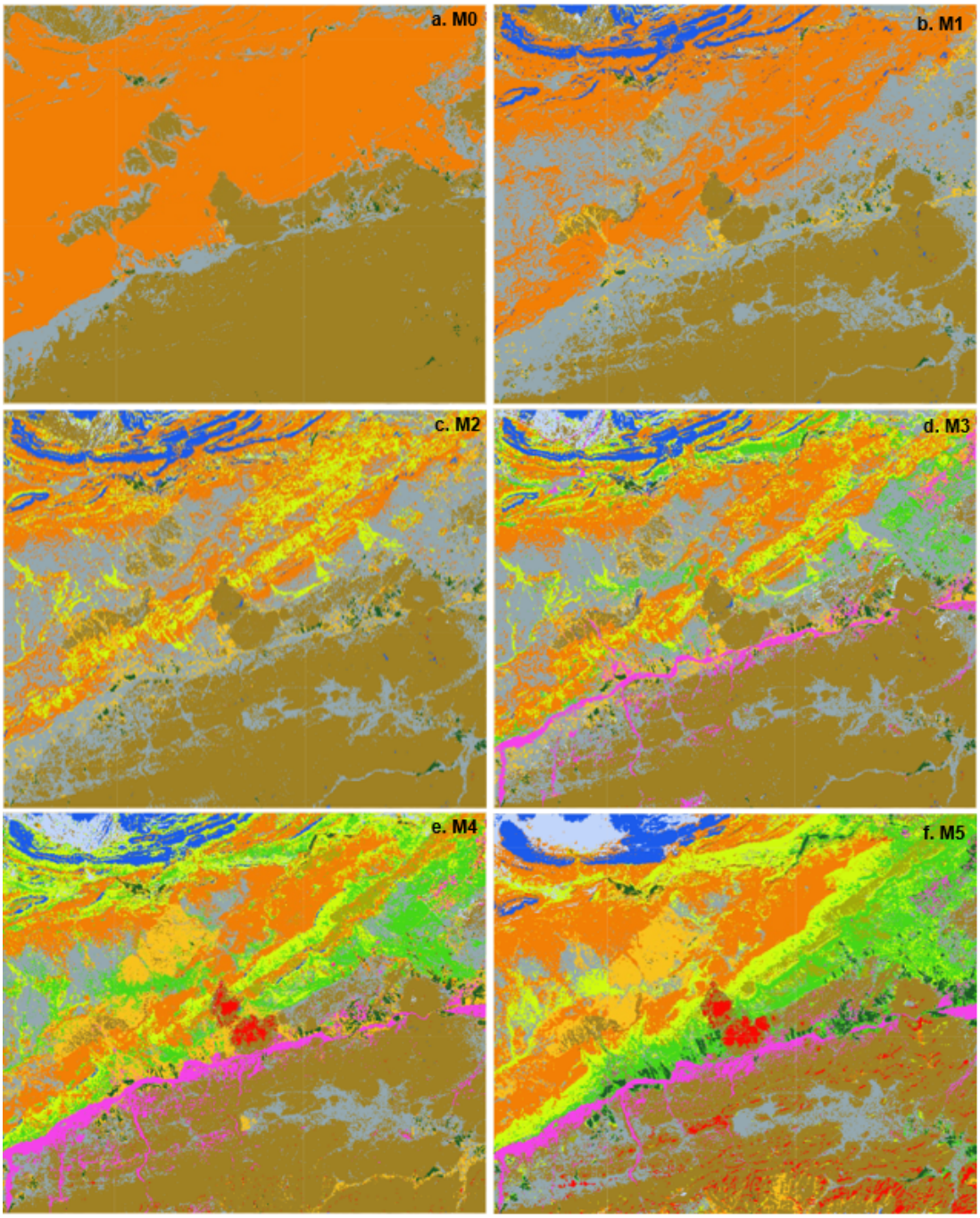

Imagen 5-10: Imágenes obtenidas por medio del algotimo SVM a) M0 (45 %), b) M1 (50 %), c) M2 (52 %), d) M3 (56 %), e) M4 (67 %), f) M5 (68 %).

5. **Naïve Bayes:** los resultados obtenidos a través de este algoritmo de enfoque paramétrico no son favorables y no se consigue que mejoren por medio de la adición de mas variables a los modelos, por tanto, descarta para realizar clasificación litológica.

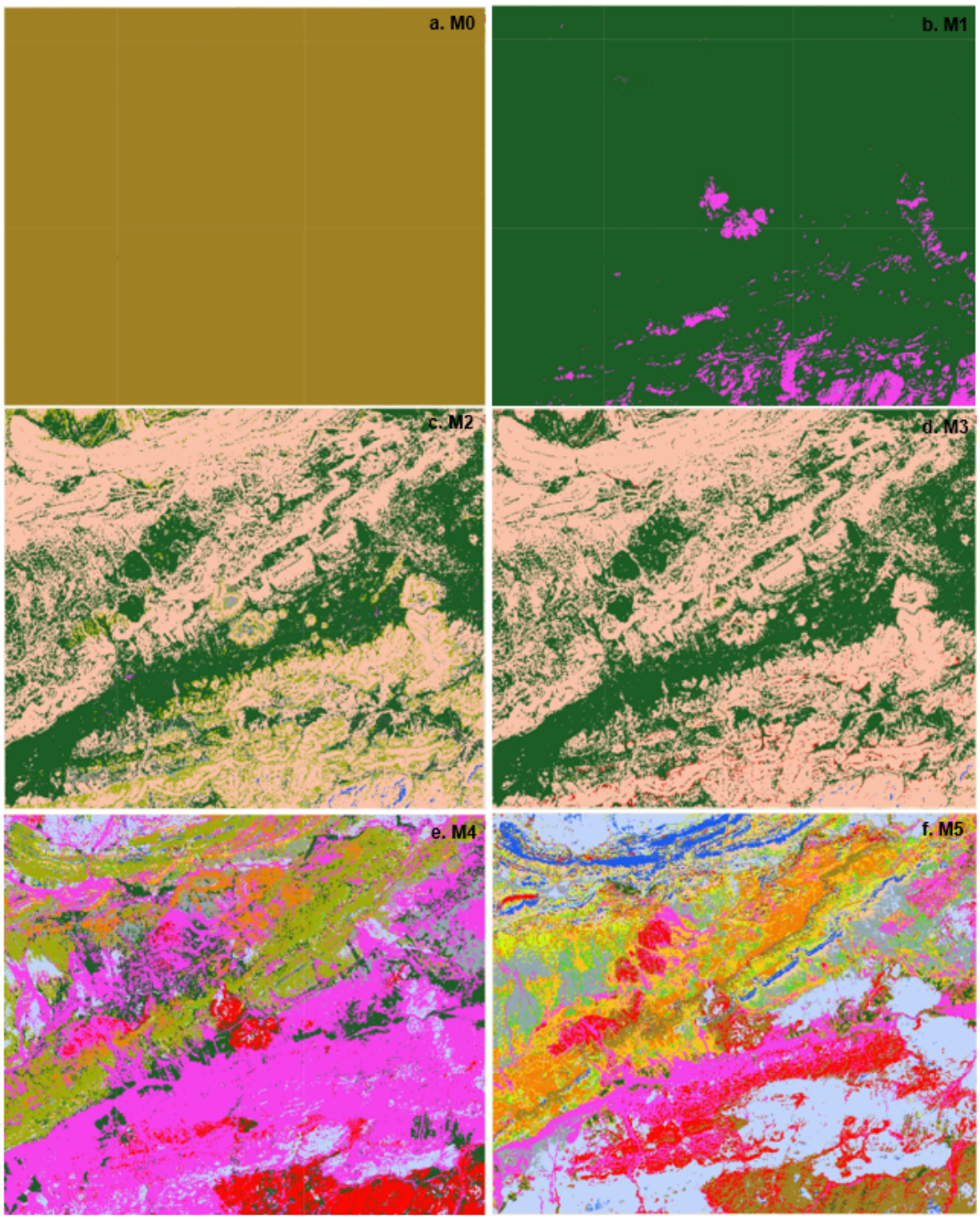

Imagen 5-11: Imágenes obtenidas por medio del algotimo Naïve Bayes. a) M0 (8 %), b) M1 (2 %), c) M2 (2 %), d) M3 (4 %), e. M4 (13 %), f) M5 (14 %).

5.3.5. Evaluación de desempeño de la clasificación:

La evaluación de desempeño de los modelos de clasificación supervisada se realizó por medio de la técnica de validación cruzada con 5 iteraciones, en donde se calculó la precisión general de las 5 particiones independientes y luego se promediaron los resultados.

Tabla 5-2: Precisión general de los modelos de clasificación.

K- Fold (K = 5) Validación cruzada					
Precisión General - OA %					
Modelos MLA´s	RANDOM FOREST	CART	MINIMUM DISTANCE	SUPPORT VECTOR MACHINE	NAIVE BAYES
Modelo 0: Bandas espectrales	80.1	72.6	63.5	45.3	8.1
Modelo 1: Bandas + slope	84.2	75.5	64.9	50.2	2.1
Modelo 2: Bandas + kmeans + DEM	84.9	76.1	67.1	51.9	2.2
Modelo 3: Bandas + kmeans + Dem + Índices	87.9	78.2	67.9	56.3	4.3
Modelo 4: Bandas + kmeans + Dem + Índices + PCA	92.8	84.6	70.1	67.5	13.3
Modelo 5: Bandas + kmeans + Dem + Índices + PCA + textura	95.4	86.3	71.4	68.1	14.2

Tabla 5-3: Precisión del productor y del consumidos del modelo 5 de Random Forest.

OA %	**RANDOM FOREST**		**CART**		**MINIMUM DISTANCE**		**SUPPORT VECTOR MACHINE**		**NAIVE BAYES**	
	93		**86**		**71**		**68**		**14**	
	PA %	**CA %**	**PA %**	**CA %**	**PA %**	**CA %**	**PA %**	**CA %**	**PA %**	**CA %**
Roca_volcánica	80.0	97.6	86.0	74.1	46.0	92.0	23.6	14.6	23.6	5.8
Lutitas rojas con yesos	97.0	97.8	94.1	90.6	76.0	57.4	10.8	52.7	10.8	12.1
Calizas_dolomias	95.0	99.1	87.1	94.4	66.4	95.1	0.8	78.3	0.8	17.1
Areniscas_rojas_lutitas	98.0	96.4	90.9	93.1	66.8	90.0	5.2	80.6	5.2	47.7
Calcarenitas_margas	98.0	93.8	94.3	93.3	82.9	62.0	33.3	61.9	33.3	28.6
Conglomerados	95.0	97	92.5	93.1	18.8	82.6	33.5	59.8	33.5	20.2
Arcillas_mioceno	95.0	95.3	89.6	89.6	26.2	73.7	23.4	61.0	23.4	35.5
Precámbrico	99.0	98.2	97.9	97.6	98.7	74.9	35.6	91.9	35.6	74.3
Vegetación	92.0	100	66.7	100.0	100.0	100.0	88.2	100.0	88.2	55.6
Dolomias_masivas	96.0	93.8	93.7	96.4	97.2	70.1	31.9	89.9	31.9	30.6
Dolomias_lutitas	86.0	96.7	90.4	94.9	76.0	86.8	76.3	73.6	76.3	25.5
Lutitas_gipsarenitas	62.0	100	76.2	84.2	42.9	90.0	6.7	0.0	6.7	3.7
Cuaternario fluvial	96.0	98.1	93.6	94.5	71.8	90.8	79.8	90.9	79.8	49.7

OA: Overall Accuracy PA: Producer accuracy CA: Consumer accuracy

Nota: Precisión del productor: Proporción de pixeles clasificados correctamente en

una clase, respecto al total de pixeles de esa clase 1- error de omisión. Precisión del consumidor: proporción de pixeles clasificados correctamente en una clase, respecto al total de pixeles categorizados en esa clase por el clasificador, equivale a 1-error de Comisión.

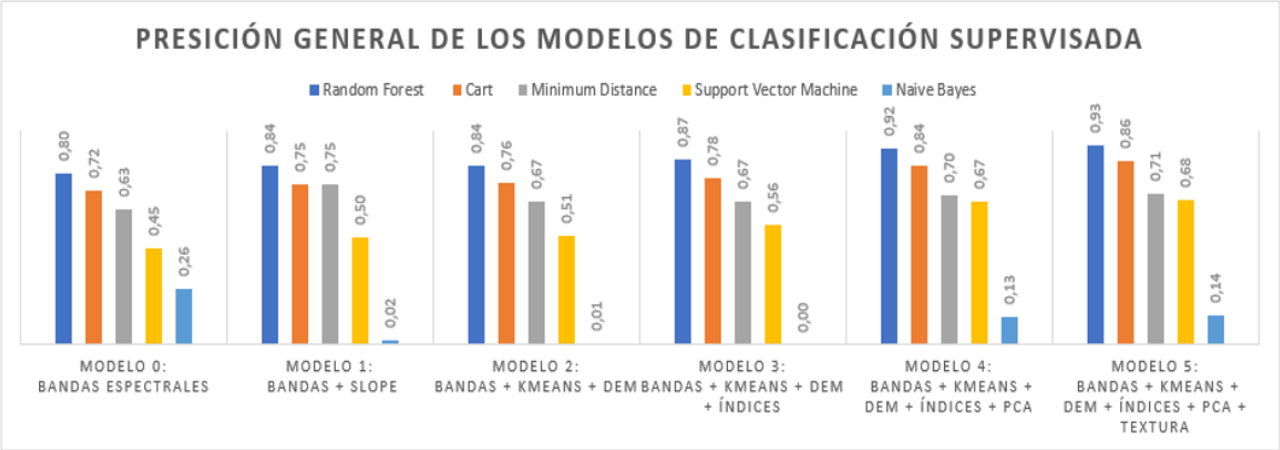

Imagen 5-12: Precisión general modelos de clasificación supervisada.

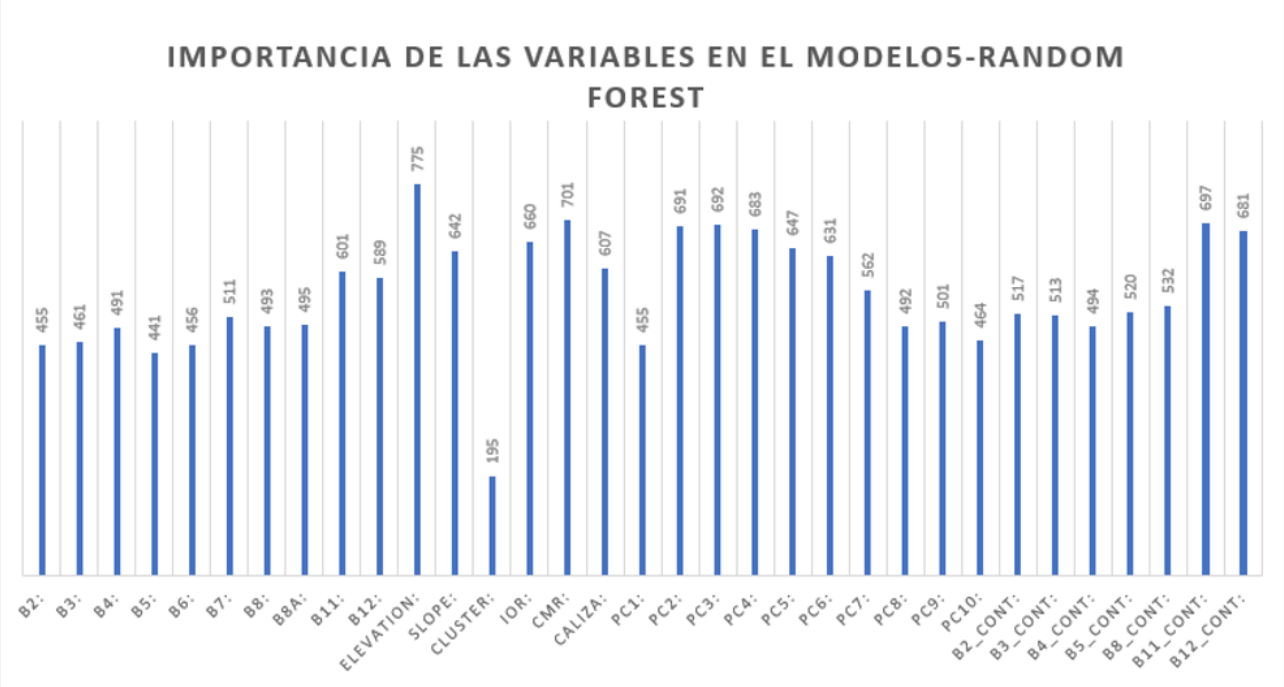

Imagen 5-13: Importancia de cada variable predictora en el modelo 5 con Random Forest.

De la tabla (5-2) y (5-3) y del gráfico (5-12) y (5-13) en donde se condesan los resultados de la validación cruzada de los modelos de clasificación supervisada se sacan las

siguientes apreciaciones:

a) El algotitmo Random Forest reproduce las mejores métricas de desempeño (precisión general) respecto a los otros clasificadores, seguido de algorimo CART y éste a su vez del clasificador Minimun Distance, que tiene similar desempeño que el algoritmo Support Vector Machine.

b) El algoritmo Naive Bayes produjo la precisión más baja en todos los modelos propuestos y esto se debe al enfoque paramétrico de este clasificador, en donde se asume que las variables son independientes entre ellas; algo que es difícil de encontrar en datos de teledetección, en donde normalmente se tiene alta correlación entre bandas espectrales de la imagen original.

c) El mejor modelo es el generado por el algoritmo Random Forest ejecutado con todas las variables predictoras con un 95 % de precisión general. En la imagen 5-14 se presentan los resultados del mejor modelo, contrastado con el mapa geológico, en donde se puede apreciar la similitud de la imagen clasificada y el mapa esperado.

d) Se evidencia que la adición de información adicional como índices espectrales, modelo digital de terreno (pendiente y elevación) y textura, en forma de variables predictoras en los modelos de clasificación, produce un mejor desempeño de los clasificadores, esto gracias a que estas variables ayudan a extraer características propias de la imagen tomada por el satélite y destacan la naturaleza topográfica del terreno estudiado.

e) La inclusión de los componentes principales en la clasificación mejora significativamente la precisión de la predicción, (del 88 % al 93 % con el clasificador Random Forest) esto ocurre porque el principio del análisis de componentes principales es generar nuevas variables (componentes) no correlacionados entre sí, en donde cada uno resalta una información diferente y especifica a partir de los datos de origen y adicionalmente permite disminuir la dimensión de los datos gracias a que los primeros componentes contiene la mayoría de la variabilidad de las bandas de la imagen original. En la tabla 5-13 se puede apreciar la importancia de las variables en del algoritmo Random Forest con el modelo 5, allí se aprecia que las variables mas relevantes en el modelamiento son el modelo digital de terreno, los primeros componentes principales y las bandas B11 y B12 y la textura (contraste) derivada de ellas, resultado esperado dado que en las curvas espectrales de la imagen 4-2, se observó que las litologías objetivo del estudio tienen mayor reflectancia en la longitud de onda de estas bandas.

f) La variable predictora generada a partir del algoritmo de clasificación no supervisada K-means, denominada “cluster” en la imagen 5-6, tiene un peso poco significativo en la clasificación con el algoritmo Random Forest y todas las variables

propuestas. Se concluye que la presencia o ausencia de esta característica no impacta directamente la precisión de la clasificación, por tanto, se puede prescindir de ella.

g) En la tabla 5-3 se detalla la precisión de la clasificación de los pixeles por cada clase tomada del modelo con todas las variables predictoras, respecto a los cinco algorirmos. Del algoritmo Random Forest se puede decir que tiene una muy buena precisión de consumidor o bajo error de omisión, dado que la proporción de pixeles clasificados correctamente por clase es alta (95 %) en todas las clases, excepto en las clases dolomias_masivas y calcarenitas_margas que fue de 93.8 %. En el clasificador CART se observa que la precisión del productor o la proporción de pixeles clasificados correctamente respecto al total de pixeles de esa clase de vegetación es relativamente baja (67 %) junto con la clase lutitas_gipserenitas (76 %), esto se debe probablemente a que son las clases con menos área y por tanto menos muestras por área de clase tienen y al muestrear puede haberse tocado alguna frontera de clase, sin embargo, la precisión del consumidor buena para estas dos clases, 100 % para vegetación y 84 % para las lutitas_gipsarenitas.

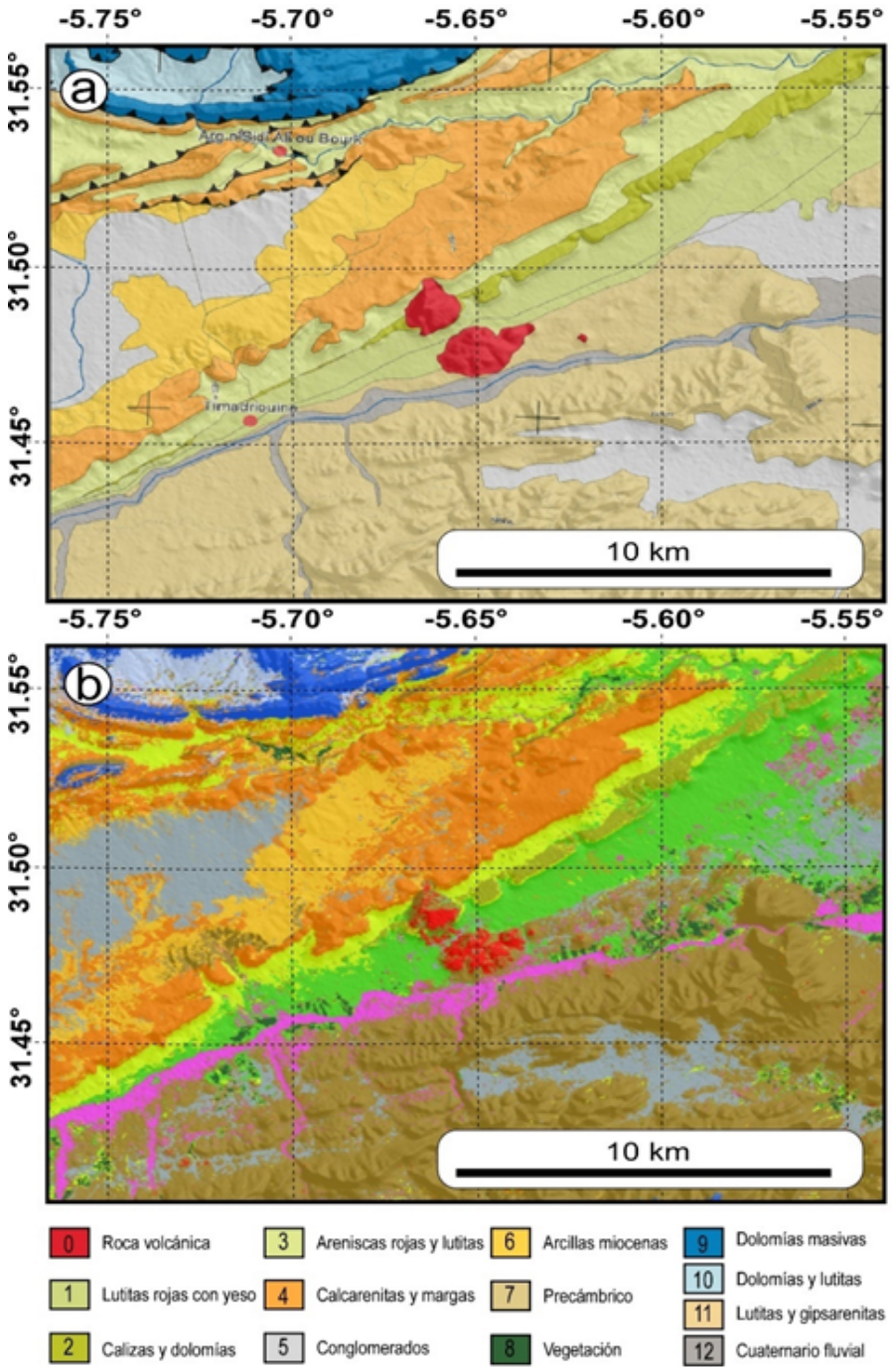

Imagen 5-14: a).Mapa geológico de la zona estudiada. b.) Imagen clasificada por Random Forest M5.

6. **Extrapolación de resultados a zona de similares características:** Habiendo elegido al algoritmo Random Forest junto con todas las variables predictoras excepto la variable del algoritmo no supervisado K-means, pues se definió en la discusión del apartado de resultados que no tiene un peso significativo. A continuación, se muestra

el resultado de la clasificación sobre el área extendida y desconocida, pero de similares características litológicas.

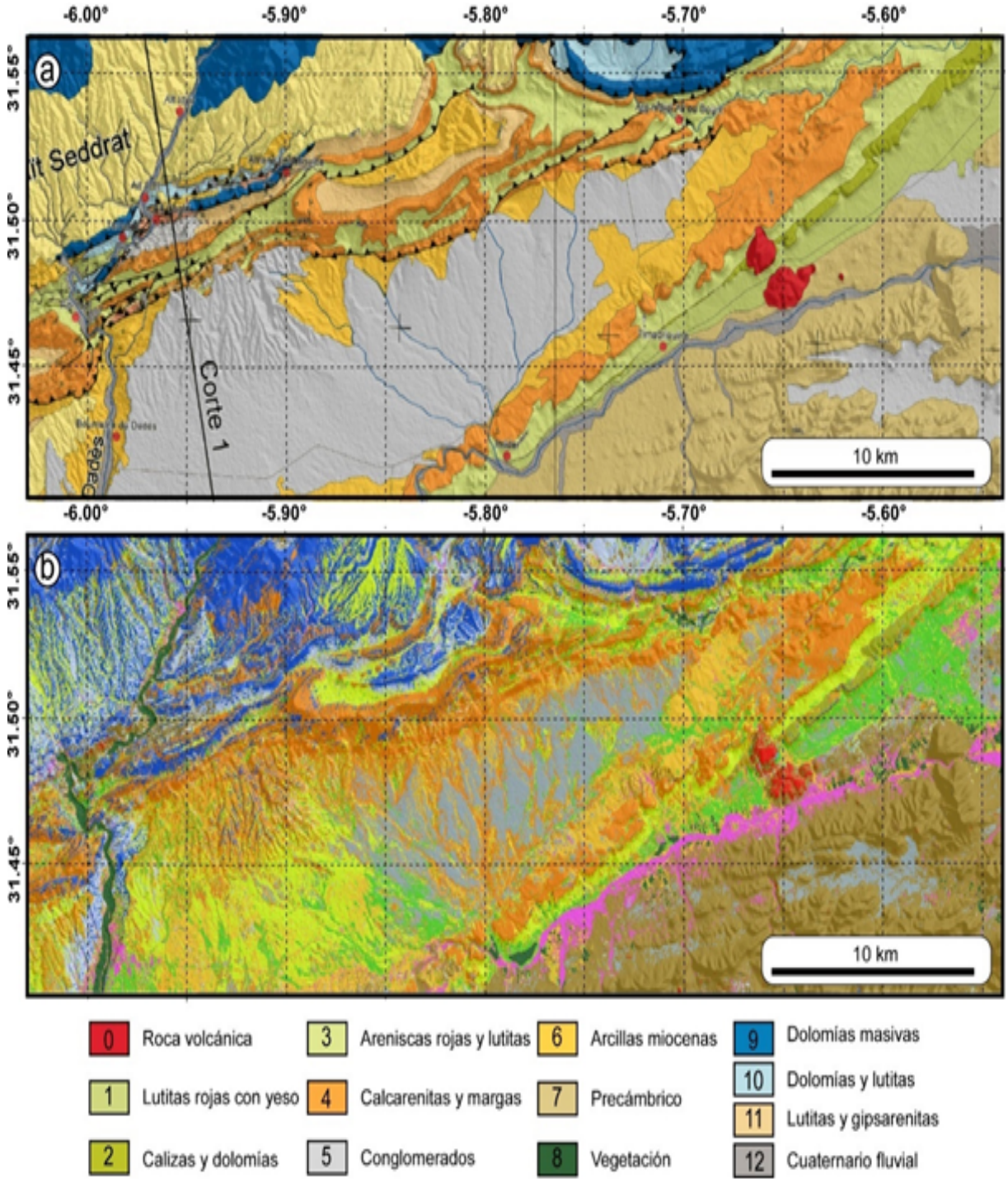

Imagen 5-15: a.) Mapa geológico de la zona extendida b.) Imagen clasificada Modelo 5 Random Forest 95 %.

En la imagen 5-15 se presenta la imagen clasificada con el 95 % de precisión general promedio, sobre el área extendida con al algoritmo Random Forest y las variables de índices espectrales, terreno, PCA y textura, contrastado con el mapa geológico, contrastada con el mapa geológico, desde el cual se pueden distinguir algunas diferencias de clasificación por clases, a pesar de haber obtenido una precisión considerablemente buena. De la imagen obtenida se pueden hacer las siguientes observaciones:

a) En el área extendida hay una clase nueva (color amarillo claro, arriba a la izquierda) denominada conglomerados del Mioceno, la cual es asignada por el algoritmo a la clase Dolomias masivas, puesto que a pesar de desconocer la unidad litológica nueva la relaciona con una clase conocida, y esto ocurre poque en efecto existe una similitud mineralógica entre estas dos clases, las dos son ricas en carbonatos, eso explica porque fue clasificada en esa categoría.

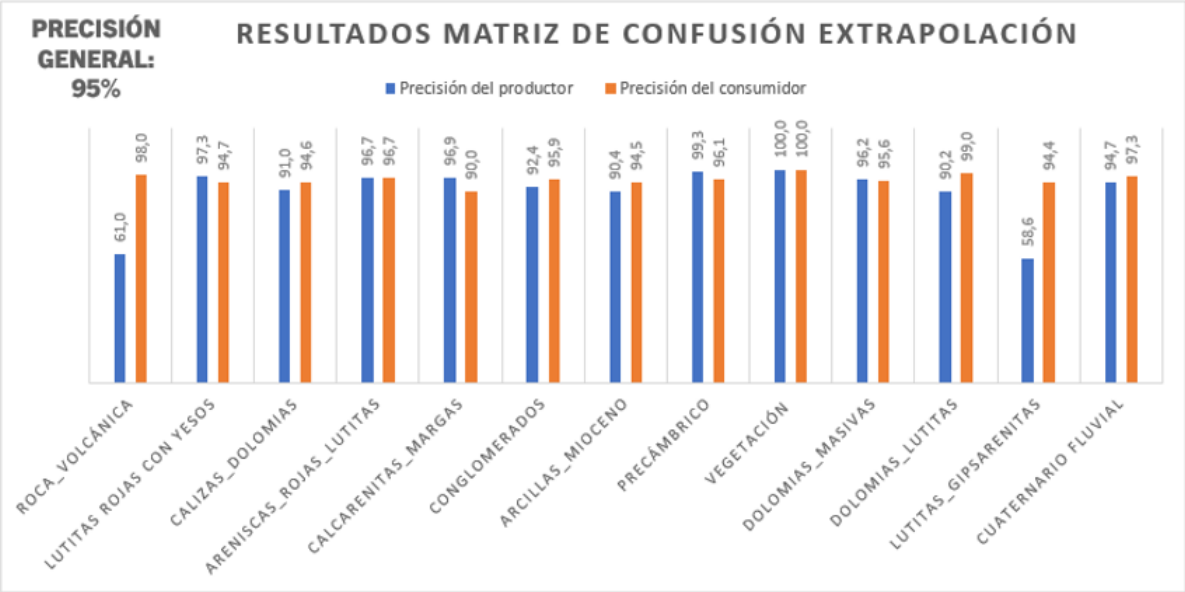

Imagen 5-16: Importancia de cada variable predictora en el modelo 5 con Random Forest en el área extendida.

b) La proporción de pixeles clasificados correctamente respecto al total de pixeles en las clases de roca volcánica y lutitas gipsarenitas disminuyó considerablemente (ver imagen 5-15), esto se debe probablemente a que el área de estas clases es pequeña respecto a las otras y por ellos se tomaron pocas muestras o se muestreó sobre alguna frontera de clase; en el mapa extendido el algoritmo no logra identificar muy bien estas dos litologías por la poca representatividad de las muestras tomadas al principio.

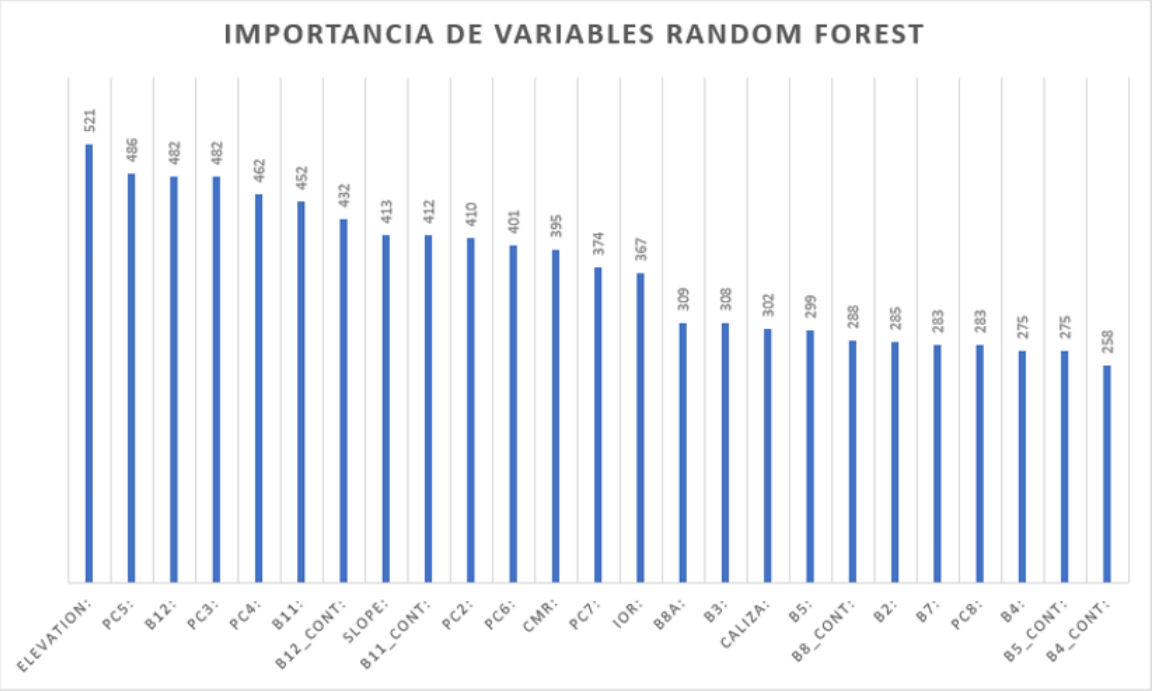

Imagen 5-17: Importancia de las variables predictoras dentro del modelo de Random Forest.

c) En el gráfico 5-16 se puede apreciar el peso de cada variable sobre el modelo de clasificación sen el área extendida, como este de esperarse las variables topográficas, las relacionadas con las bandas B11 y B12, los primeros componentes principales tienen mas peso en la predicción, esto debido a la naturaleza del terreno y las características espectrales de las unidades litológicas.

6 Conclusiones:

6.1. Modelamiento Clasificación Supervisada

- A raíz de los resultados obtenidos en este estudio, se puede afirmar que realizar clasificación litológica por medio de algoritmos de clasificación supervisada es una alternativa interesante como herramienta para hacer cartografía geológica porque se consiguen resultados concluyentes y satisfactorios; durante el proceso es importante contar con un especialista en el campo de la geología que contribuya al planteamiento de los modelos de acuerdo a las características del terreno y la litología, fortalezca la interpretación de los resultados y pueda proponer ajustes asertivos.

- Es clave resaltar que los resultados fueron satisfactorios por dos factores importantes; el primero fue el modelamiento de los datos a través de diferentes algoritmos de clasificación supervisada buscando el que entendiera y aprendiera mejor de los datos y el segundo, la adición de variables predictoras de diferentes naturaleza que consiguieron resaltar las caracerísticas de las unidades litológicas y robustecer la predicción.

- Todos los procesos llevados a cabo en este estudio se realizaron en la plataforma geomática con procesamiento en la nube Google Earth Engine - GEE con excelentes resultados, en especial en los tiempos de procesamiento que fueron increíblemente cortos aún en el modelamiento de algoritmos con múltiples variables, así como en la variabilidad y complejidad de algoritmos y funciones que permitieron ejecutar a cabalidad los objetivos propuestos en este estudio, sumado a la disponibildiad del catalogo de datos del Satelite Sentinel-2A MSI y el Modelo Digital de Terreno.

6.2. Extrapolación a un área extendida:

- Extender un modelo de clasificación a un área desconocida pero de características litológicas similiares a la entrenada, brinda buenos resultados al generar una herramienta preliminar para realizar cartografía geológica. Esto supone un ahorro considerable en tiempo y esfuerzo durante el proceso cartográfico.

6.3. Códigos trabajados en el estudio:

Los scripts generados en este trabajo se relacionan a continuación:

- https://code.earthengine.google.com/31cdb93a6bd5f60b8873a30af874ad6e
- https://code.earthengine.google.com/89e584a7c948c3e91cd62a8fccc08992
- https://code.earthengine.google.com/8680435461e2bbf5f13d03a81cc4ed46

Bibliografía

Borra, S., Thanki, R., & Dey, N. (2019). *Satellite Image Analysis: Clustering and Classification*. Springer Nature Singapore Pte Ltd.

Breiman, L. (2001). *Random Forest* [Tesis doctoral]. Tesis doctoral.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (2017). *Classification And Regression Trees*. Chapman Hall/CRC.

Cao, R., Chen, J., Chen, Y., Zhu, X., & Chen, M. (2020). Thick cloud removal in Landsat images based on autoregression of Landsat time-series data. *Remote Sensing of Enviroment*, *249*, 112001.

Cardile, F., Crowly, M., Saah, D., & Clinton, N. (Eds.). (2024). *Cloud-Based Remote Sensing with Google Earth Engine. Fundamentals and Applications*. Springer.

Chan, B., & Bai, K. (2018). *Multisensor Data Fusion and Machine Learning for Environmental Remote Sensing*. CRC Press.

Congalton, R., & Green, K. (2009). *Assessing the Accuracy of Remotely Sensed Data*. CRC Press.

Cracknell, M., & Reading, A. (2013). Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers and Geosciences*, *63*, 22, 23.

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

Estornell, J., Martí-Gavilá, J., Sebatiá, M., & Mengual, J. (2013). Principal Component analysis applied to remote sensing. *Modelling in science Education and Learningl*, *6(2)*, No. 7.

Ge, Q., W.and Cheng, Tang, Y., Jing, L., & Gao, C. (2018). Lithological Classification Using Sentinel-2A Data in the Shibanjing Ophiolite Complex in Inner Mongolia, China. *Remote Sensing*, *10*, 638.

Golblatt, R., You, W., Hanson, G., & Khandelwal, A. (2016). Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine. *Remote Sensing*, *8*, 634-662.

Gupta, R. (2018). *Remote Sensing Geology*. Springer.

Haralick, R., Shanmugan, K., & Dinstain, I. (1973). Textural Features of Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-3 No.6*, 610-621.

Hargitai, H. (Ed.). (2019). *Planetary Cartography and GIS*. Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. Springer.

Kamusoko, C. (2019). *Remote Sensing (Image Classification in R)*. Springer.

Lavender, S., & A., L. (2023). *Practical Handbook of Remote Sensing*. CRC Press.

Mahood, A., Maxwell, B., Spiers, A., Koontz, M., Ilangakoon, N., Solvik, K., Quarderer, N., Mcglinchy, J., Scholl, V., St Denis, L., Nagy, C., Braswell, A., Rossi, M., Herwehe, L., Wasser, L., Cattau, M., Iglesias, V., Yao, F., Leyk, S., & Balch, J. (2023). Ten simple rules for working with high resolution remote sensing data. *Peer Community Journal*, *3, article e4*.

Marghany, M. (Ed.). (2022). *Remote sensing and image processing in mineralogy*. CRC Press.

Moser, G., & Zerubia, J. (Eds.). (2018). *Mathematical Models for Remote Sensing Image Processing*. Springer.

NASA, J. (2020). *NASADEM Merged DEM Global 1 arc second V001 [Data set]. NASA EOSDIS Land Processes DAAC*. GEE.

Ninomiya, Y. (2004). Lithologic mapping with multispectral ASTER TIR and SWIR data. *The International Society for Optical Engineering*, *5234*, 180-190.

Pokhariya, H., Singh, D., & Prakash, R. (2023). Evaluation of different machine learning algorithms for LULC classification in heterogeneous Landscape by using Remote sensing and GIS techniques. *Engineering Research Express*, *DOI 10.1088/2631-8695/acfa64*.

Prost, G. (2014). *Remote Sensing for geoscientists, image analysis and integration)*. CRS Press.

Quirós, E. (2009). *Clasificación de imágenes multiespectrales ASTER mediante funciones adaptativas* [Tesis doctoral]. Tesis doctoral.

Richards, J. (2013). *Remote Sensing Digital Image Analysis*. Springer.

Serbouti, I., Raji, M., Hakdaoui, M., El Kamel, F., Pradhan, B., Gite, S., Alamri, A., Maulud, K., & Dikshit, A. (2022). Improved Lithological Map of Large Complex Semi-arid Regions Using Spectral and Textural Datasets within Google Earth Engine and Fused Machine Learning Multi-Classifiers. *Remote Sensing*, *14*, 5498.

Shirmard, H., Farahbakhsh, E., Muller, D., & Chandra, R. (2022). A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sensing of Enviroment*, *268*, 112750.

Tangestani, M., & Shayeganpour, S. (2020). Mapping a Lithologically complex terrain using Sentinel-2A data: a case study of Suriyan area, Southwestern Iran). *International Journal of Remote Sensing*, *41-9*, 3558-3574.

Tesón, E. (2009). *Estructura y cronología de la deformación en el borde Sur del Alto Atlas de Marruecos a partir del registro tectono-sedimentario de la cuenta de ante país de Ouarzazate* [Tesis doctoral]. Tesis doctoral.

Thenkabail, P. (Ed.). (2016). *Remotely sensed data characterization, Classification, and accuracies*. CRC Press.

Tso, B., & Mather, P. (2009). *Classification methods for remotely sensed data*. CRC Press.