Universidad de Valladolid

ESCUELA DE INGENIERÍA INFORMÁTICA

GRADO EN INGENIERÍA INFORMÁTICA
Mención en Computación

Desarrollo de modelos para la clasificación de granates en el marco de la exploración espectroscópica in situ de Marte

---

Alumno: Vega López, Paula Xiaozhen

Tutor: López Reyes, Guillermo Eduardo
Julve González, Sofía

Agradecimientos

Quiero agradecerles a mis padres, por haber estado siempre apoyándome sin importar las circunstancias. A todas las figuras y tutores que he tenido, por haberme enseñado a ser como soy. Y a las matemáticas, que a pesar de nuestra relación unilateral, no se han apartado de mi lado.

Resumen

El proyecto realizado consiste en la creación de un clasificador automático de granates mediante Python con el objetivo de categorizar espectros Raman en los seis subgrupos principales de este mineral.

En este documento se describe el proceso llevado a cabo para preprocesar los datos, entrenar y seleccionar los clasificadores, y la construcción del modelo final, junto con la discusión de los resultados obtenidos.

Abstract

This project involves the creation of an automatic garnet classifier using Python with the aim of categorizing Raman spectra into the six main subgroups of this mineral.

This document describes the process undertaken to preprocess the data, train and select the classifiers, and construct the final model, along with a discussion of the obtained results.

Índice general

Lista de Figuras

Lista de Tablas

Capítulo 1

Introducción

1.1. Contexto

Hoy en día, la exploración espacial en Marte es cada vez más importante a la hora de buscar vida en otros planetas. Organizaciones como la NASA, la ESA, o empresas privadas han centrado su mirada en el estudio de este planeta y de sus minerales para resolver la pregunta de si alguna vez existió vida en Marte.

Dos importantes misiones espaciales, *Mars2020* y *ExoMars*, cuentan con la colaboración del grupo de investigación ERICA de la Universidad de Valladolid. Su participación consiste en el desarrollo de dos dispositivos que utilizan técnicas de espectroscopia a bordo de dos misiones espaciales a Marte. El primero es un instrumento de Espectroscopia Laser Raman (RLS), una de las tres herramientas del ALD (*Analytical Laboratory Drawer*) del Rover Rosalind Franklin para la identificación y caracterización de minerales y marcadores biológicos. Y el segundo es SuperCam, un conjunto de herramientas de detección remota del Rover Perseverance, lanzado por la NASA en la misión de Marte de 2020 y actualmente en servicio, para examinar rocas y suelos con una cámara láser y varios espectrómetros con el objetivo de encontrar signos de posibles condiciones habitables y evidencia de agua y vida microscópica en el pasado en Marte. [1, 2, 3]

La espectroscopia Raman es un tipo de espectroscopia vibracional frecuentemente utilizado en el campo de la química, se puede obtener información estructural acerca de los modos vibracionales normales, sus frecuencias y los niveles de energía de las moléculas que componen una muestra. Esta técnica resulta muy útil en el estudio de minerales debido a su análisis rápido que no requiere la destrucción de la muestra. Se basa en la dispersión inelástica de fotones conocida como dispersión Raman. Utilizando una fuente de luz monocromática, típicamente un láser dentro del espectro visible, cercano al infrarrojo o ultravioleta, se ilumina la muestra a analizar. La luz del láser interactúa con las moléculas excitándolas, lo que modifica la energía de los fotones enviados. Este desplazamiento de la energía es el que proporciona información sobre los modos vibracionales de los átomos que componen la muestra.

La radiación electromagnética resultante se recoge a través de unas lentes y filtros, proporcionando un espectro. Cada mineral, definido por su composición química, presenta un patrón espectro identificable llamado huella dactilar estructural por las que las moléculas pueden ser identificadas.[4, 5, 6, 7]

1.2. Motivación

El estudio de las fases minerales que presentan ciertas soluciones sólidas como son los granates se puede utilizar para extraer información precisa acerca de su composición química. Utilizando algoritmos multivariantes y modelos de clasificación, se podrá obtener esta información elemental de minerales más complejos y se podrá caracterizarlos de manera más precisa para lograr entender la composición de Marte y su historia geológica. [8, 9]

A largo plazo, se espera que la clasificación geoquímica de los minerales de espectros Raman pueda tener un importante impacto en el campo de la exploración planetaria.

1.3. Objetivos

El objetivo principal de este proyecto es implementar un modelo que permita clasificar con facilidad los espectros de granates obtenidos en Marte utilizando herramientas de espectroscopia Raman. Una tasa de éxito aceptable estaría alrededor del 80 % sin sobreajuste. Esto permitiría a los investigadores automatizar un trabajo que normalmente se realiza de forma manual y conlleva un coste temporal alto.

1.4. Estructura de la memoria

Este documento se estructura de la siguiente forma:

Capítulo 2 Materiales y métodos Describe las muestras geológicas utilizadas (granates), así como la obtención y procesamiento de sus espectros y la metodología empleada para la resolución del problema.

Capítulo 3 Optimización de modelos y resultados Describe los modelos utilizados y los resultados obtenidos junto con una discusión de estos últimos.

Capítulo 4 Conclusiones Describe las conclusiones a las que se ha llegado y las posibles líneas de trabajo futuras.

Bibliografía Listado de las referencias utilizadas.

Anexo A Resumen de enlaces adicionales: Incluye enlaces de interés sobre el proyecto, como el repositorio de código.

Capítulo 2

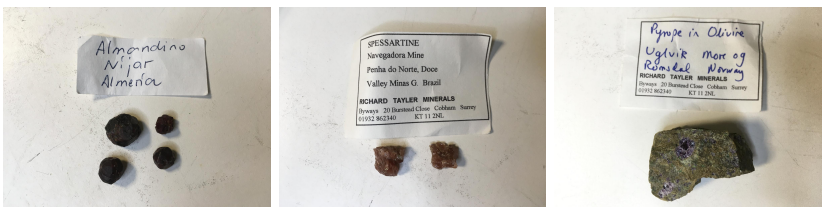
Materiales y métodos

2.1. Muestras geológicas: granates

Los granates son un grupo de minerales perteneciente a la categoría de los neosilicatos, compuestos por átomos de silicio y oxígeno, cuya fórmula química genérica es $X_3Y_2(SiO_4)_3$. Los elementos en la posición X representan los cationes divalentes, mientras que los de la posición Y representan los trivalentes. Estos pueden variar para formar diferentes soluciones sólidas. El rango del elemento en X es Ca, Mg, Fe, Mn y el del elemento Y es Al, Fe, Mn, Cr, Ti, V [10].

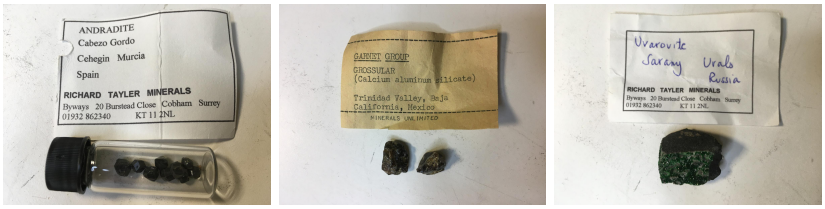
El grupo de los granates está dividido en dos series que contienen a su vez los seis subgrupos más comunes: piralspitas, cuando Y = Al y la X varía entre Mg, Fe^{2+}, Mn, y ugranditas, cuando X = Ca y la Y varía entre Cr, Al, Fe^{3+}. Dentro de las piralspitas se encuentran los almandinos ($Fe_3Al_2(SiO_4)_3$), los piropos ($Mg_3Al_2(SiO_4)_3$) y las espesartinas ($Mn_3Al_2(SiO_4)_3$). Y dentro de las uvarovitas, las andraditas ($Ca_3Fe_2(SiO_4)_3$), las grosulares ($Ca_3Al_2(SiO_4)_3$) y las uvarovitas ($Ca_3Cr_2(SiO_4)_3$).

En las figuras 2.1 y 2.2 se pueden observar algunos de estos tipos de minerales que han sido utilizados en el trabajo.



(a) Almandino (b) Espesartina (c) Piropo

Figura 2.1: Piralspitas



(a) Andradita (b) Grosular (c) Uvarovita

Figura 2.2: Ugranditas

Los granates que se han utilizado para este trabajo son 32 muestras diferentes pertenecientes a las seis subcategorías del mineral. Estas muestras estaban incrustadas dentro de un sustrato más grande por lo que la difícil extracción del mineral para su procesamiento ha hecho que el número de espectros obtenidos de cada muestra no haya podido ser el mismo. Debido a la escasez de muestras en el laboratorio, también se han utilizado 47 muestras tomadas de RRUFF, una base de datos abierta de espectroscopia Raman, junto a sus espectros.

El listado de las 79 muestras utilizadas (laboratorio + RRUFF) es el siguiente. Esta lista contiene los minerales resultantes tras el proceso de descarte de muestras y espectros no válidos. En la figura 2.3, se puede ver gráficamente las muestras totales, a la izquierda las piralspitas y a la derecha las ugranditas (siguiendo el mismo orden que en la lista).

- Piralspitas
  - Almandinos: 15 (7 + 8)
  - Espesartinas: 7 (4 + 3)
  - Piropos: 11 (3 + 8)
- Ugranditas
  - Andraditas: 19 (9 + 10)
  - Grosularias: 22 (6 + 16)
  - Uvarovitas: 5 (3 + 2)

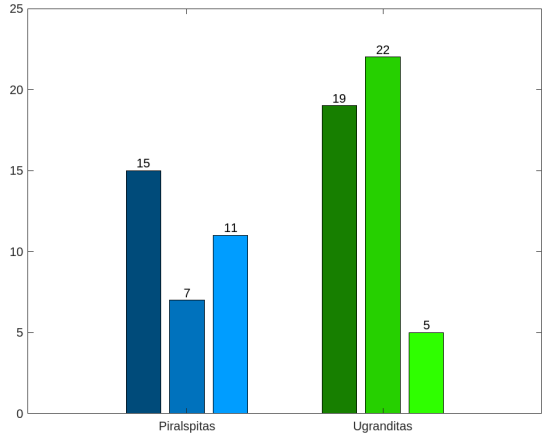
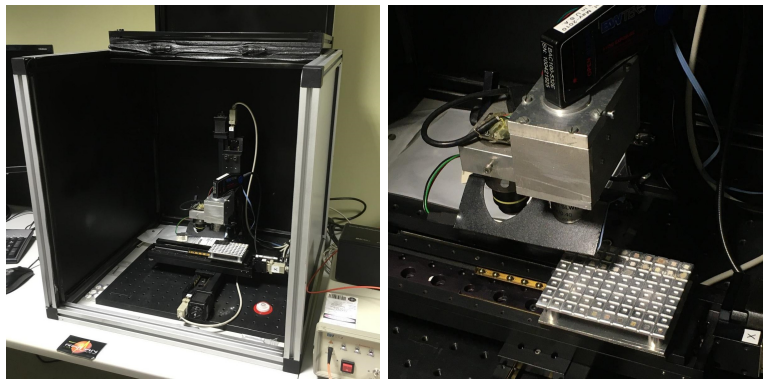


Figura 2.3: Muestras totales

2.2. Conjunto de datos

2.2.1. Espectros Raman

Los espectros utilizados habían sido obtenidos anteriormente con la herramienta láser Elforlight G4-PSU, utilizando el cabezal Raman Horiba-JY Superhead con láser de 532 nm y el microscopio Zeiss Axiotech 30. El espectrómetro y CCD utilizados son del equipo Horiba-JY Induram. Estos espectros se han tomado utilizando un objetivo 50x con longitud focal larga y una potencia de láser sobre la muestra por defecto de 17 mW. Mediante esta herramienta se habían tomado 3 espectros de cada muestra disponible. Para tomar estas medidas no fue necesario que la muestra estuviese pulverizada. Aun así, se necesitaba un tiempo demasiado largo para poder obtener cada espectro, por lo que para solucionar este inconveniente se recurrió al Simulador del RLS de la Figura 2.4 (se trata de una caja hermética que contiene el instrumento Raman junto con una tablilla donde se depositan las muestras a analizar), que representa al Raman RLS de ExoMars. En total se han obtenido 15 espectros por muestra, de las que estaban en polvo, combinando 12 espectros del simulador y 3 del Induram; mientras que solo se poseen los 3 espectros del Induram para aquellas muestras que no estaban pulverizadas.



(a) Parte exterior (b) Parte interior

Figura 2.4: Simulador

Una vez obtenidos, se ha implementando un filtro tipo *XOR* para eliminar la luz blanca de la lámpara de calibrado KOSI HCA de los espectros. Después se ha realizado la corrección de la línea de base y por último, los espectros han sido normalizados para que el promedio de su intensidad sea la unidad. En total, se han utilizado 372 espectros, de los cuales 325 se han obtenido a través de los equipos del laboratorio y 47 a través de RRUFF.

En la Figura 2.5 se muestra el formato de un espectro, en este caso de un almandino. Está compuesto de una serie de puntos (x,y) que representan la intensidad normalizada y el desplazamiento Raman respectivamente. Este último representa la diferencia entre la luz láser incidente y la luz dispersada, y depende de la frecuencia de vibración de los átomos de la molécula y del cambio en la composición del material estudiado. Normalmente, esta frecuencia se mide en números de onda. En el espectro, los datos presentan una serie de

picos que forman un patrón identificativo del tipo de enlaces moleculares que componene la muestra. A estos picos también se les llama bandas. El estudio de estas bandas permite la identificación del tipo mineralógico. [11, 12, 13]

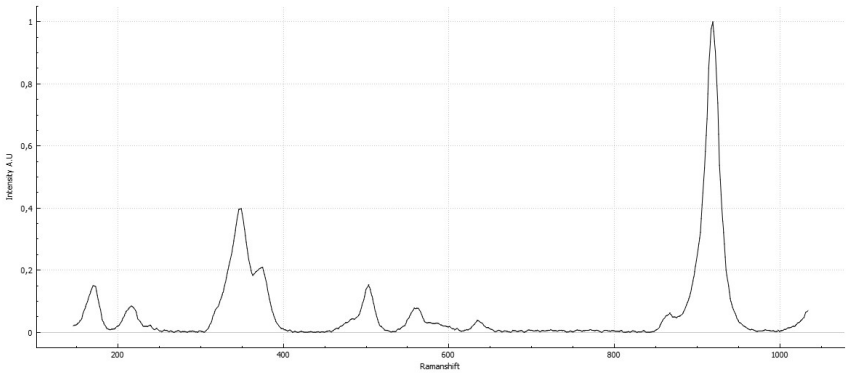


Figura 2.5: Espectrograma de un almandino

Matemáticamente, las bandas siguen un perfil de Voigt, una función que es una convolución de una distribución Cauchy-Lorentz y una distribución Gausiana, que sigue la siguiente fórmula [14]:

$$V(x;\sigma,\gamma) = \int_{-\infty}^{\infty} G(x';\sigma) \cdot L(x-x';\gamma) \cdot dx \tag{2.1}$$

Donde el perfil Gausiano es:

$$G(x;\sigma) \equiv \frac{e^{-x^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} \quad \text{con} \quad \sigma = \frac{f}{(2\sqrt{2\ln(2)}} \tag{2.2}$$

Y el perfil Lorentziano es:

$$L(x;\gamma) \equiv \frac{\gamma}{\pi(x^2+\gamma^2)} \quad \text{con} \quad \gamma = \frac{f}{2} \tag{2.3}$$

Esta convolución (2.1) se puede simplificar utilizando perfiles Pseudo-Voigt, una combinación lineal de las distribuciones con factor lorentziano-gausiano η:

$$V_p(x,f) = \eta \cdot L(x,f) + (1-\eta) \cdot G(x,f) \quad \text{con} \quad 0 < \eta < 1 \tag{2.4}$$

Utilizando un ajuste de bandas mediante la ecuación (2.5), se pueden comparar los parámetros de los picos de cada espectro. La función para calcular el perfil de Voigt tiene como entrada los datos en x, los parámetros de intensidad del pico i, anchura del pico en el punto medio f, centro del pico c y factor lorentziano λ.

$$V_p(x; i, f, c, \lambda) = i \cdot ((1 - \lambda) * G(x; c, f) + \lambda * L(x; c, f)) \tag{2.5}$$

2.2.2. Los datos

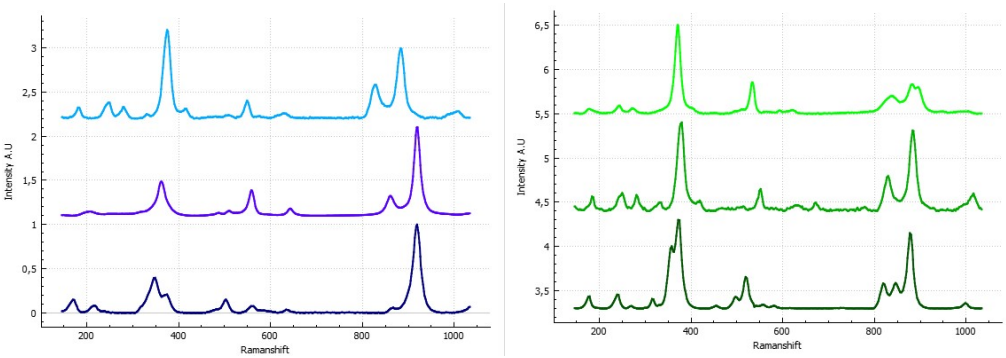
Antes de discutir los modelos, se deben comentar ciertas propiedades de los datos que se han utilizado y que han condicionado la posterior creación de los clasificadores. En primer lugar, el número de muestras disponibles para entrenar y validar el clasificador es bastante limitado, 79 muestras, donde la mayoría no son completamente puras (el mineral a estudiar de la muestra está presente en pequeñas concentraciones que, al extraerlo para su análisis, puede presentar trazas de la piedra donde se encontraba). A la cantidad de muestras también hay que añadir que no se tiene el mismo número de muestras para cada grupo y subgrupo de los granates, lo que crea clases desbalanceadas con pocas instancias en unas y muchas en otras. Por último, estas muestras han sido tomadas en un ambiente que simula las condiciones reales en las que se analizan las rocas en Marte, por lo que los datos contienen un ruido sistemático en los atributos. Como el clasificador se utilizará en un entorno con ruido, es preferible dejarlo y solo reducir aquel que también se vaya a reducir en el preprocesamiento de los espectros.

Para paliar el número limitado de muestras, se han ampliado las instancias añadiendo las muestras de la base de datos RRUFF.[15] Estos datos, están tomados por diferentes instrumentos con otras calibraciones, por lo que sirven para añadir más variedad a los datos y evitar el sobreajuste de los modelos. También se ha tomado más de un espectro por muestra para que el clasificador se pueda entrenar con más instancias. Y para resolver el problema de las clases desbalanceadas, se ha utilizado una separación guiada de los conjuntos de entrenamiento, prueba y validación.

Una vez recopilados todos los datos a usar en el trabajo, se ha procedido a la fase de preprocesamiento de los datos, donde se han llevado a cabo varias transformaciones para acabar con un formato estándar: descartar aquellas muestras o espectros que no son válidos, cortar los espectros para que estén situados en el mismo intervalo de frecuencias, normalizar los espectros por intensidad de sus picos, obtener los parámetros de las bandas de cada espectro y pasar los datos a formato .csv.

Las imágenes presentadas en la figura 2.6 muestran un ejemplo de espectro de cada clase (siguiendo el mismo código de colores que la figura 2.3). Como se puede ver, la cantidad y posición de las bandas en cada tipo de granate difiere, pero se pueden encontrar varias similitudes entre aquellos espectros pertenecientes a cada uno de los dos grupos.

En primera instancia se realizó una selección de parámetros de las bandas espectrales como datos de entrada para los clasificadores. Para ello, se ha utilizado la aplicación de SpectPro. Las bandas seleccionadas de cada espectro se escogieron manualmente, buscando la mayor cantidad de picos comunes entre muestras que indiquen los enlaces moleculares presentes y ausentes de las mismas. Con esta selección de parámetros, se probaron algunos clasificadores utilizando Weka [16], para comprobar si los atributos elegidos eran los adecuados, debido a la difícil comparación entre bandas existentes en solo ciertos tipos de muestra. Tras validar con diferentes métodos, incluyendo árboles y modelos basados en distancias como C4.5, Naive



(a) Espectros de espesartina, piropo y almandino, de arriba a abajo

(b) Espectros de uvarovita, grosular y andradita, de arriba a abajo

Figura 2.6: Ejemplos de espectros

Bayes, K-Vecinos, etc. se llegó a la conclusión de que eran insuficientes (estos resultados no se mostrarán debido a que se descartó esta selección de atributos). Después del análisis de los resultados anteriores, se modificaron los datos intentando usar toda la información disponible de los espectros, es decir, todos los puntos (x,y) de intensidad y frecuencia. Para obtener un menor número de facetas, los datos se redujeron mediante Análisis de Componentes Principales (PCA). [17]

Al final, para preprocesar los espectros, se utilizó un script escrito en Matlab [18] que recogía todos los datos y realizaba PCA de las intensidades de cada espectro. El número de componentes resultantes elegidas han sido 10, con una varianza del 95 % de los datos.

En la siguiente tabla (2.1) se muestra la estructura final de los datos con 16 atributos, dos de ellos para representar la clase. Los atributos que se le darán a los clasificadores serán los 10 componentes principales (PCs) y el grupo o subgrupo correspondiente como clase.

Nombre	**Tipo**	**Descripción**
Tipo	Nominal	Nombre del tipo de muestra
O	Nominal	Origen de la muestra (simulador o RRUFF)
N	Número	Número de muestra
E	Número	Número de espectro de la muestra
1-10	Número	Componentes principales (10 atributos)
Subgrupo	Nominal	Subgrupo al que pertenece la muestra (6 clases)
Grupo	Nominal	Grupo al que pertenece la muestra (2 clases)

Tabla 2.1: Atributos

2.3. El clasificador

El modelo elegido es el modelo mixto que consta de tres clasificadores con seis clases finales: almandinos, piropos, espesartinas, andraditas, grosulares y uvarovitas. Cada clasificador se entrena de forma independiente a los demás, pero todos son validados con el mismo conjunto de datos. En la figura 2.7 se puede ver el esquema del modelo planteado.

El primer clasificador recibe las PCs y separa las muestras en dos grupos: piralspitas y ugranditas. Una vez separadas las muestras, se pasan a los dos clasificadores siguientes los componentes y el subgrupo correspondiente para que las categoricen entre las subclases de piralspitas y ugranditas respectivamente. Para calcular la tasa de error, se cuentan solo las muestras mal clasificadas al final, dentro de las seis clases.

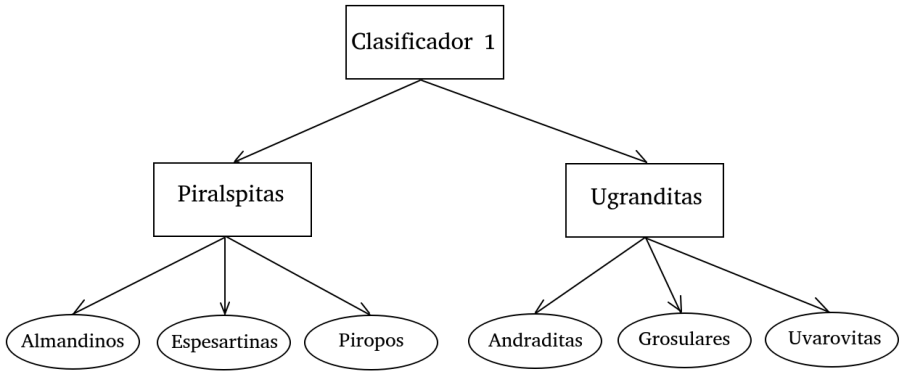


Figura 2.7: Esquema del modelo mixto

A la hora de elegir qué tipo de modelo de clasificación automática usar, se ha optado por realizar pruebas entre estos cinco: K-Medias, árboles de clasificación C4.5, K-Vecinos, Naive Bayes y Perceptrón Multicapa (MLP). Esta elección ha sido determinada por la cantidad y formato de las instancias ya que, con un número limitado de instancias para trabajar, es preferible la creación de un modelo clasificatorio más simple. [19, 20, 21, 22]

Los datos se dividen en tres partes: entrenamiento, prueba y validación. Para entrenar y probar los modelos, se utiliza la técnica de validación cruzada de 3 particiones con los datos de entrenamiento y prueba. Y para comprobar la tasa de aciertos del modelo final, se usa el conjunto de validación con instancias totalmente nuevas para el clasificador.

La separación del conjunto de validación se ha tenido que hacer de manera semi-manual debido a que los espectros de la misma muestra deben utilizarse todos juntos, al no haber mucha variación entre estos y ser casi iguales. La parte manual de la partición ha consistido en la separación de un grupo que contenga el mismo número o similar de muestras de cada grupo, subgrupo y origen para una validación completa. El resto de instancias: los conjuntos de entrenamiento y prueba, se han dividido automáticamente utilizando la validación cruzada estratificada con 3 particiones de Sklearn, StratifiedKFold.

La técnica de validación cruzada de k particiones, también llamada *k-fold* se utiliza para la validación de modelos cuando no existe un conjunto de datos de prueba suficientemente grande. En ella, los datos se dividen en k particiones, en este caso 3, y se realizan tantas iteraciones como particiones. Uno de los subconjuntos se utiliza como datos de prueba y el resto (k-1 conjuntos) como datos de entrenamiento. En cada iteración, el conjunto de prueba y entrenamiento varía. Una vez realizada la media aritmética de los resultados de cada iteración, se obtiene una precisión final única. De esta manera es posible evaluar la bondad del clasificador con pocas instancias, sin caer en el sobreajuste. Aunque en la práctica lo más común es utilizar la validación cruzada de 10 iteraciones, la cantidad de instancias de uvarovitas ha hecho que el número de estas solo pueda ser tres.

Después de entrenar y calcular la precisión de los modelos, se comparan cuáles funcionan mejor en qué parte. En función de estos resultados, se crean los posibles modelos finales.

2.3.1. Modelos

Se han elegido cinco modelos diferentes de clasificación supervisada: árbol de decisión, K-Medias, K-Vecinos, Naive Bayes y Perceptrón Multicapa. Los modelos se han implementado a través de Python utilizando la biblioteca para aprendizaje automático de *Scikit-learn* [23]. En los siguientes apartados se explicará brevemente en qué consisten [24, 25, 26] y el motivo de su elección.

C4.5

Dentro de los árboles de decisión, se ha utilizado el algoritmo C4.5, basado en ID3, que utiliza el concepto de entropía de la información para clasificar los datos. Este modelo es uno de los más simples, por lo que en general, es más robusto frente a nuevos datos.

El árbol está formado por nodos, que representan atributos y hojas, que representan las clases finales. Por cada nodo de decisión, se elije el atributo que mejor divida las muestras en subconjuntos que contengan datos de una clase u otra, utilizando como criterio la mayor ganancia de información que resulte al dividir las muestras de esa manera. Este proceso se repite recursivamente hasta que todas las instancias del subconjunto pertenezcan a la misma clase; cuando eso sucede, se crea un nodo hoja que representa esa clase. También puede ocurrir que no se consiga ninguna ganancia de información, en este caso, C4.5 crea un nodo de decisión en un nivel superior del árbol con el valor de la clase.

Algorithm 1 C4.5

1. Comprobar los **casos base**.
2. Por cada atributo ***a*** encontrar la **ganancia de información normalizada** de la división de ***a***.
3. Siendo ***a_best*** el atributo con la **ganancia de información normalizada más alta**.
4. Crear **nodo de decisión** de la división de ***a_best***
5. Repetir en las sublistas obtenidas por división de ***a_best*** y agregar estos nodos como hijos de **nodo**.

K-Medias

K-Medias es un método de agrupamiento muy utilizado en minería de datos. Se basa en particionar el conjunto de observaciones en k grupos donde cada instancia pertenece al grupo cuyo valor medio es más cercano, también llamado centroide. Este método, basado en distancias, utiliza la distancia euclídea al cuadrado entre las muestras, para formar las agrupaciones. Normalmente este problema es computacionalmente costoso, considerado un tipo NP-difícil. Aun así, al estar basado en distancias, es posible que funcione bastante bien por el tipo de atributos numéricos que se utilizan en este problema. Para el primer clasificador k será igual a 2, debido que se dividen las muestras en dos grupos, y en los dos clasificadores siguientes, k será igual a 3, por las tres subclases de cada mineral.

Algorithm 2 K-medias

1. Elegir el número de *clusters* ***k***.
2. Elegir de manera aleatoria los centroides ***c_k***.
3. Hasta llegar al número de **iteraciones máximo** o **converger**:
   a) Por cada punto ***x_i***, calcular la **distancia a los centroides** y asignarlo al **más cercano**.
   b) Por cada cluster ***k_i***, **actualizar el centroide** por la media de todos los puntos asignados al mismo.

K-Vecinos

Este método, también llamado k vecinos más próximos, es un modelo de clasificación supervisada no paramétrico, es decir, no asume una distribución subyacente de los datos. K-Vecinos estima el valor de una función de probabilidad a posteriori de que una muestra pertenezca a una clase u otra, según la clase mayoritaria de las k muestras más cercanas que

la rodean. K-vecinos también se basa en la distancia euclídea para calcular la proximidad entre instancias, por lo que puede funcionar bastante bien al agrupar y diferenciar entre las dos grandes clases. En los clasificadores, el número de vecinos k será igual a 1.

Algorithm 3 K-vecinos

1. Elegir el número de vecinos más próximos ***k***.
2. Por cada punto ***x_i***:
   a) Calcular la **distancia euclídea** entre ***x_i*** y **todos los puntos**.
   b) Ordenar la lista de distancias de **menor a mayor**.
   c) Elegir los ***k*** *primeros* puntos.
   d) Asignar la clase según la **clase mayoritaria del grupo** de puntos.

Naive Bayes

Naive Bayes o clasificador bayesiano ingenuo es un tipo de clasificador probabilístico basado en el teorema de Bayes, que vincula la probabilidad de un evento A, dado otro evento B, con la probabilidad de B, dado A. Este clasificador asume la independencia de las variables predictoras. La estimación de los parámetros utiliza el criterio de máxima verosimilitud. Naive Bayes normalmente no requiere de una gran cantidad de datos para el entrenamiento del modelo y se utiliza con atributos categóricos. Debido a que los datos de los espectros son numéricos, se ha utilizado la versión Gaussiana de Naive Bayes para este modelo.

Algorithm 4 Naive Bayes

1. Calcular la **media** y la **desviación estándar** de cada atributo predictor por cada clase.
2. Por cada **atributo**:
   a) Calcular la **probabilidad** de ***f_i*** utilizando la ecuación de densidad de gauss en cada clase.
3. Calcular la **estimación para cada clase**.
4. Asignar a cada instancia la clase de la **estimación más probable**.

Perceptrón Multicapa

El perceptrón multicapa (MLP) es una red neuronal artificial formada por varias capas de perceptrones, que se utiliza a menudo para resolver problemas no linealmente separables.

La red neuronal diferencia entre tres tipos de capa: la capa de entrada, que introduce los datos en la red, la capa de salida, que devuelve el valor de la clase resultante y la capa oculta, que pueden ser varias y está formada por neuronas que procesan la información. Este algoritmo utiliza la retropropagación del error junto con descenso del gradiente para optimizar los parámetros de la función de activación. Para los clasificadores, se utilizará un tamaño de la capa oculta de 10 neuronas.

Algorithm 5 Perceptrón Multicapa

1. **Inicialización** de los pesos $\boldsymbol{w}$.
2. Por cada par de entradas-salidas:
   a) Calcular la **salida** de la muestra $\boldsymbol{x}$.
   b) **Comparar** con la salida correcta y **calcular el error**.
   c) Calcular las **derivadas parciales** del error y **ajustar los pesos** de cada neurona para reducirlo.

Capítulo 3

Optimización de modelos, resultados y discusión

El objetivo principal de este trabajo es poder realizar una clasificación automática de granates según sus espectros. Como se ha explicado en anteriores secciones, los granates se dividen en dos clases principales y tres subclases por cada una de ellas. El planteamiento principal consta de un modelo mixto con un primer clasificador que distinga entre piralspitas y ugranditas, seguido de otros dos clasificadores que separen las muestras catalogadas entre las subclases restantes.

3.1. Primer clasificador

Los resultados de cada modelo probado para el primer clasificador se han recogido en la tabla 3.1. Para este clasificador, al diferenciar entre dos clases que poseen características bastante distintivas entre sí, se esperaban tasas de acierto mayores que en los dos siguientes.

Fold	**K-Medias**	**C4.5**	**K-Vecinos**	**NB**	**MLP**
1	0.74	0.88	0.84	0.67	0.72
2	0.83	0.62	0.69	0.55	0.64
3	0.70	0.61	0.65	0.65	0.70
Media	0.76	0.70	0.73	0.62	0.69
Desvest	0.07	0.15	0.10	0.06	0.04

Tabla 3.1: Porcentaje de aciertos del primer clasificador

Comparando la tasa de aciertos media de cada modelo y ordenando de mejor a peor los clasificadores tenemos: K-Medias, K-Vecinos, C4.5, MLP, Naive Bayes. Para crear el clasificador final se probará con los dos mejores, en este caso, K-Medias y K-Vecinos.

Pero primero se debe comprobar si realmente existen diferencias a la hora de utilizar un modelo u otro. Esto se puede verificar con el test de Student remuestreado pareado para validación cruzada, que permite saber si existe una diferencia significativa entre los modelos para este problema. El test determina con un nivel de confianza α, habitualmente 5 %, utilizando el valor z de las tablas del test de Student de $(k-1)$ grados de libertad para α ($\alpha/2$ si se utilizan las tablas de una cola), si existe una diferencia significativa entre los clasificadores. Tras calcular el estadístico t cuya fórmula es la 3.1 donde S_d^2 es la varianza muestral de la diferencia entre resultados de cada modelo ($d = \overline{x} - \overline{y}$) y el número de folds es k, se comprueba si $t \geq z$. Si lo es, la prueba indica que la diferencia entre los clasificadores es significativa.

$$t = \frac{\overline{d}}{\sqrt{\frac{S_d^2}{k}}} \tag{3.1}$$

A la hora de realizar la prueba, se ha elegido un $\alpha = 10\,\%$, siendo $z = 2.92$, y $k = 3$ ya que se han utilizado tres particiones. Tras realizar los cálculos necesarios, se obtiene el estadístico $t = 3.4183$ y comparándolo con z, t es mayor, por lo que la diferencia es significativa. Se utilizarán ambos clasificadores como posible primer clasificador del modelo final.

3.2. Clasificador de piralspitas

Los resultados del segundo clasificador de la tabla 3.2 muestran los porcentajes de aciertos de los modelos. Comparando sus tasas y ordenándolos de mejor a peor quedan de la siguiente manera: K-Vecinos, MLP, C4.5, Naive Bayes, K-Medias. En este caso, los dos mejores clasificadores son K-Vecinos y MLP. Los aciertos para este clasificador varían más que los aciertos del primero, destacando una peor precisión en el algoritmo de K-Medias.

Como en el anterior apartado, se deben comparar los modelos para saber si son significativamente diferentes. Utilizando la ecuación 3.1 para calcular el estadístico entre estos dos clasificadores se obtiene un $t = 1.6352$. Al ser menor que el valor de referencia z, la diferencia entre los modelos no es significativa, es decir, para construir el clasificador final no habría una gran diferencia al usar un tipo u otro. Por lo que para el modelo final, se utilizará K-Vecinos como clasificador de piralspitas.

Fold	**K-Medias**	**C4.5**	**K-Vecinos**	**NB**	**MLP**
1	0.72	0.83	0.86	0.63	0.91
2	0.56	0.78	0.94	0.86	0.83
3	0.44	0.67	0.56	0.56	0.56
Media	0.57	0.76	0.79	0.68	0.77
Desvest	0.14	0.08	0.20	0.16	0.19

Tabla 3.2: Porcentaje de aciertos del clasificador de piralspitas

3.3. Clasificador de ugranditas

Los resultados del clasificador de ugranditas se pueden ver en la tabla 3.3. En este caso, ordenados de mayor porcentaje de aciertos a menor: MLP, K-Medias, Naive Bayes, K-Vecinos, C4.5. Los dos mejores son MLP y K-Medias. Como en los dos clasificadores anteriores, se utiliza el test de Student para comprobar si son diferenciables. Calculando el estadístico con la ecuación 3.1, $t = 1.4607$ siendo también menor que z por lo que no hay una diferencia significativa entre ambos. Por consiguiente, para el modelo final se utilizará MLP.

Fold	**K-Medias**	**C4.5**	**K-Vecinos**	**NB**	**MLP**
1	0.23	0.58	0.92	1	0.96
2	0.82	0.23	0.22	0.23	0.64
3	0.64	0.86	1	0.96	0.93
Media	0.56	0.55	0.72	0.64	0.84
Desvest	0.30	0.32	0.43	0.43	0.18

Tabla 3.3: Porcentaje de aciertos del clasificador de ugranditas

3.4. Modelo final

Una vez elegidos los tipos de clasificadores a probar para el modelo final, se crean y se entrenan utilizando todos los datos excepto los de validación. Tras los resultados anteriores, se han hecho dos versiones del modelo final. La primera está formada por un primer clasificador de K-Medias seguido por K-Vecinos para el grupo de las piralspitas y MLP para el grupo de las ugranditas. Y la segunda versión solo modifica el primer clasificador, intercambiándolo por K-Vecinos. En las figuras 3.1 y 3.2 se puede ver gráficamente el esquema de los modelos.

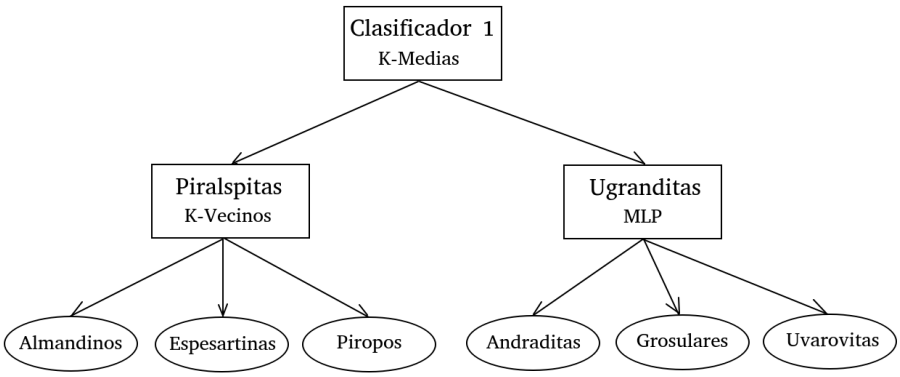


Figura 3.1: Modelo 1

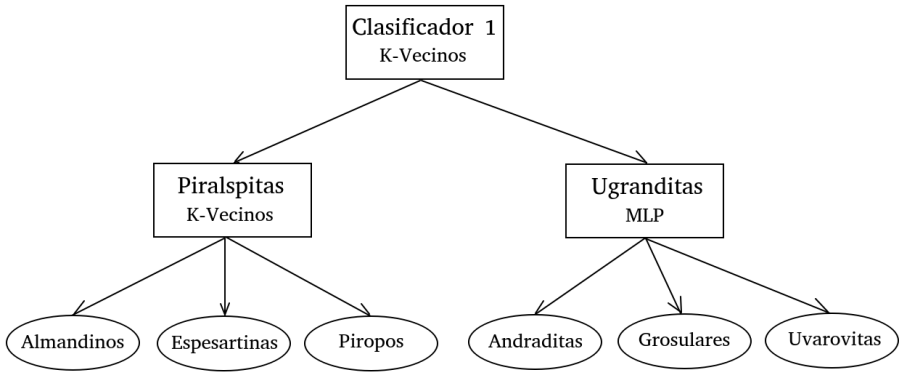


Figura 3.2: Modelo 2

Una vez entrenados los clasificadores, se utiliza el conjunto de validación para calcular sus aciertos. Los datos pasan por el primer clasificador y son separados entre los dos grupos, ugranditas y piralspitas. Después, se cuentan las instancias erróneamente clasificadas y se dan como entrada a los siguientes clasificadores solo aquellas que están clasificadas correctamente. Una vez que los clasificadores de piralspitas y ugranditas devuelven la salida, se hace el conteo de las instancias mal clasificadas, se suman a las de la primera fase y se calcula la tasa de aciertos total.

En la tabla 3.4 aparecen los resultados de los dos modelos finales. Los clasificadores tienen una tasa de aciertos bastante similar y cerca del 85 %, el mejor de ambos es la segunda versión del modelo mixto con K-Vecinos como primer clasificador y clasificador del grupo de piralspitas y MLP como clasificador de ugranditas.

	Primero	**Piralspitas**	**Ugranditas**	**Aciertos**
Modelo 1	K-Medias	K-Vecinos	MLP	83 %
Modelo 2	K-Vecinos	K-Vecinos	MLP	87 %

Tabla 3.4: Modelos finales

3.5. Discusión

Ahora se va a proceder a discutir los resultados obtenidos de los clasificadores anteriores, empezando por los resultados de las tablas 3.1, 3.2 y 3.3. En el primer clasificador, los resultados son bastante homogéneos, la media de todos los métodos está cerca del 70 % de aciertos. Se puede ver en la tabla 3.1 que Naive Bayes clasifica peor mientras que los dos mejores son K-Medias y K-vecinos, es posible que esto se deba al hecho de que estos dos últimos están basados en distancias, mientras que Bayes es probabilístico. Pasando al clasificador de piralspitas, en la tabla 3.2 existe una media global de 71 % ligeramente superior a la del primer clasificador. Aun así, los resultados varían más, destacando K-Medias, cuyo porcentaje de aciertos medio no llega al 60 %, probablemente debido a no converger en el

mínimo global en las tres iteraciones que se han realizado. Por último, en el clasificador de ugranditas se obtiene una media global de 67 %. En esta tabla hay mucha variación entre las tasas de cada método según la iteración. Como se ha mostrado antes en la tabla 3.3, todos los modelos excepto MLP tienen tasas de error que no superan el 50 % en alguna iteración. Esto puede ser a causa del ruido dentro de los espectros y de su similitud dentro del grupo de las ugranditas (ver figura 2.6b).

Dentro del modelo final, el porcentaje de aciertos que han tenido las dos versiones es bastante bueno, aproximándose al 85 % y la diferencia de aciertos no es demasiado grande, solo un 4.35 %, por lo que sería posible utilizar cualquiera de los dos modelos con resultados aceptables. Los modelos 1 y 2 solo se diferencian en el tipo del primer clasificador, siendo en el número uno K-Medias y en el dos, K-Vecinos. Ambos clasificadores están basados en distancias, pero varían en la forma del agrupamiento de las muestras. Es posible que K-Vecinos funcione mejor que K-Medias porque este último haya convergido en un mínimo local, llegando a un resultado no óptimo. Cabe destacar que en la comparación de ambos métodos al utilizar el test de Student, si se reduce el α a la mitad, el test indica que no hay una diferencia significativa entre ambos, por lo que apoya la teoría del mínimo local.

Dentro de los clasificadores en la segunda fase, piralspitas y ugranditas, las tablas 3.2 y 3.3 muestran un buen rendimiento de K-Vecinos y MLP en ambas clases aunque también destaca el árbol de clasificación en las piralspitas. Observando las muestras que han clasificado erróneamente cada modelo, se puede ver que fallan en casi las mismas instancias. En la tabla 3.5 se han listado los espectros donde cada modelo no acierta a la hora de clasificar.

	Modelo 1				**Modelo 2**			
Clasificador	**Tipo**	**O**	**Nº**	**E**	**Tipo**	**O**	**Nº**	**E**
Primero	Andradite	S	16	1	Andradite	S	16	1
	Andradite	S	16	2	Andradite	S	16	2
	Andradite	R	7	1	Andradite	R	7	1
	Uvarovite	R	4	1	Almandine	R	14	1
	Almandine	S	3	1	Pyrope	R	8	1
	Almandine	S	3	2				
	Almandine	S	3	3				
	Almandine	S	3	4				
	Pyrope	R	8	1				
Piralspitas	Almandine	S	3	8	Almandine	S	3	8
	Almandine	S	3	14	Almandine	S	3	14
Ugranditas	Grossular	S	3	13	Grossular	S	3	13
					Uvarovite	R	4	1

Tabla 3.5: Espectros mal clasificados

Al comprobar los espectros mal clasificados y compararlos con otros espectros del mismo grupo, se puede intentar ver las diferencias y teorizar sobre los motivos por los que han sido clasificadas erróneamente. En la figura 3.3 se pueden ver los espectros de una andradita en negro y la andradita nº16 (espectros 1 y 2) en verde. Se puede apreciar claramente que en los espectros verdes existe una mayor cantidad de ruido que en el negro, es posible que esa sea la razón de su clasificación errónea.

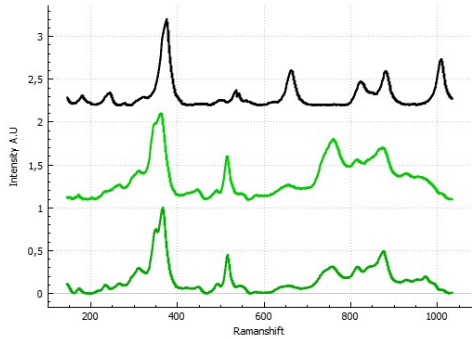
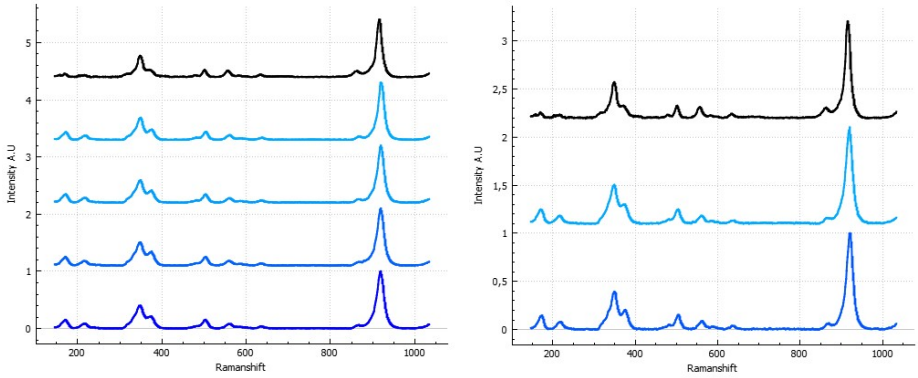


Figura 3.3: Comparación de espectros: andradita nº16

Otra de las muestras mal clasificadas es el almandino nº3 (espectros 1, 2, 3, 4, 8 y 14). En la figura 3.4 se compara esta muestra con otro espectro de almandino, pero no hay una gran diferencia apreciable entre estos; la única distinción es la intensidad de las dos primeras bandas situadas cerca de los 200 nm.



(a) Espectros 1, 2, 3 y 4 de almandino nº3 (b) Espectros 8 y 14 de almandino nº3

Figura 3.4: Comparación de espectros: almandino nº3

Es posible que la clasificación errónea de las muestras presentadas en la tabla 3.5 sea debido a la cantidad de ruido dentro de los espectros, que se ha trasladado posteriormente a las componentes principales utilizadas como atributos de clasificación al realizar PCA. Otra opción es el tipo de muestra utilizada, no todas son completamente puras y puede haber mezclas entre subgrupos de la misma clase, por ejemplo, una muestra con un 60 % de andradita y 40 % de grosular o una andradita con algún porcentaje menor de otro mineral distinto a los granates (ya que algunas muestras están incrustadas en otros minerales, como se puede ver en las figuras 2.1 y 2.2). Estos tipos de muestra presentaría características mixtas de varias subclases con picos de intensidad y posición variadas, por lo que daría problemas a la hora de su clasificación. Aun así, el clasificador funciona con una tasa de éxito aceptable y, si solo se cuentan las muestras (no los espectros) de la tabla 3.5, el número es casi idéntico en ambas versiones.

Capítulo 4

Conclusiones

4.1. Conclusiones

El estudio de la vida pasada en otros planetas presenta un reto tanto técnico como logístico a la hora de recoger y analizar datos, así como para encontrar las trazas de vida primigenia que pudiera haber habitado en otros cuerpos del sistema solar. En este sentido, el estudio de la mineralogía del lugar es clave ya que permite comprender y establecer las condiciones ambientales (presencia de agua, temperatura, atmósfera, etc.) bajo las que se formaron los minerales encontrados, de forma que se pueda establecer la habitabilidad de una determinada región, en este caso, de Marte. Un caso de estudio relevante son los granates, una solución sólida en tres fases donde hay un gran interés en ser capaces de cuantificar/clasificar las muestras según sus diferentes tipos. La espectroscopia Raman permite obtener un espectro vibracional cuyas características vienen dadas por la estructura molecular del material. Es por ello que gracias a la utilización de técnicas de espectroscopia Raman se ha pretendido analizar la composición de los granates con el objetivo de proporcionar información valiosa para el estudio de la potencialidad de la existencia de vida en Marte. Para ello, en este trabajo se estudia la factibilidad de realizar modelos para la clasificación de muestras minerales de granates basados en técnicas de espectroscopia Raman.

La manera de abordar este trabajo ha consistido en la realización de un preprocesamiento de los datos donde se han transformado las intensidades de los espectros mediante PCA, obteniendo 10 facetas con las que poder caracterizar cada tipo de granate. Después, se ha procedido a elegir el método de clasificación, comparando entre cinco métodos frecuentemente utilizados en técnicas de aprendizaje automático: K-Medias, K-Vecinos, árboles de decisión, perceptrón multicapa y Naive Bayes. Tras obtener los resultados de las pruebas, se han construido dos modelos mixtos finales, el primero formado por K-Medias para separar entre piralspitas y ugranditas, K-Vecinos para clasificar los subgrupos de piralspitas y MLP para clasificar los de ugranditas; y el segundo modelo final, formado por K-Vecinos en su primer clasificador e idéntico al anterior en los dos siguientes clasificadores. A través de este proyecto, se ha logrado construir dos clasificadores automáticos de espectros con un 82 % y 86 % de

aciertos del conjunto de datos disponibles.

4.2. Líneas de trabajo futuras

Aunque el modelo final consigue un porcentaje de aciertos bastante aceptable; debido a la cantidad limitada de datos, es posible que con nuevas instancias su rendimiento no sea igual de eficiente. Por ello se propone la creación de un modelo más robusto frente a datos con ruido y tomados con instrumentos diferentes (utilizando también calibraciones distintas), basado en atributos más paramétricos de los espectros como son la intensidad y posición relativa de los picos entre sí y su anchura en el punto medio. Para ello se debería realizar un estudio más intensivo sobre la correspondencia de las bandas de los espectros y buscar maneras de hacerlas comparables. También se sugiere utilizar técnicas de *oversampling* para paliar la falta de muestras en algunas clases, como son las uvarovitas debido a su escasez y rareza natural.[27, 28]

Bibliografía

[1] D. Bolle, "Perseverance Science Instruments - NASA science." https://science.nasa.gov/mission/mars-2020-perseverance/science-instruments/, 2024. Accessed: 2024-05-14.

[2] ERICA, "Erica homepage." http://erica.uva.es/, 2024. Accessed: 2024-05-10.

[3] ESA, "ESA - ExoMars Factsheet." https://www.esa.int/Science_Exploration/Human_and_Robotic_Exploration/Exploration/ExoMars/ExoMars_Factsheet, 2022. Accessed: 2024-05-20.

[4] G. S. Senesi, "Laser-Induced Breakdown Spectroscopy (LIBS) applied to terrestrial and extraterrestrial analogue geomaterials with emphasis to minerals and rocks," *Earth-Science Reviews*, vol. 139, pp. 231–267, 2014.

[5] M. Veneranda, J. A. Manrique, A. Sanz-Arranz, S. J. Gonzalez, C. P. Garcia, E. P. Sanchez, M. Konstantinidis, E. Charro, J. M. Lopez, M. A. Gonzalez, F. Rull, and G. Lopez-Reyes, "Application of chemometrics on raman spectra from mars: Recent advances and future perspectives," *Journal of Chemometrics*, vol. 39, 07 2022.

[6] S. P. Mulvaney and C. D. Keating, "Raman spectroscopy," *Analytical Chemistry*, vol. 72, no. 12, pp. 145–158, 2000.

[7] J. A. Jaszczak, "Word to the wise: Raman spectroscopy in the identification and study of minerals," *Rocks & Minerals*, vol. 88, no. 2, pp. 184–189, 2013.

[8] J. A. Manrique-Martinez, G. Lopez-Reyes, A. Alvarez-Perez, T. Bozic, M. Veneranda, A. Sanz-Arranz, J. Saiz, J. Medina-Garcia, and F. Rull-Perez, "Evaluation of multivariate analyses and data fusion between raman and laser-induced breakdown spectroscopy in binary mixtures and its potential for solar system exploration," *Journal of Raman Spectroscopy*, vol. 51, 01 2020.

[9] G. Lopez-Reyes, P. Sobron, C. Lefebvre, and F. Rull, "Multivariate analysis of raman spectra for the identification of sulfates: Implications for exomars," *American Mineralogist*, vol. 99, p. 1570–1579, 2014.

[10] Wikipedia contributors, "Garnet — Wikipedia, the free encyclopedia," 2024. [Online; accessed 1-May-2024].

[11] M. Fu, J. Dai, and L. Zhao, "A study on the raman spectral characteristics of garnet from the jiama copper polymetallic deposit in tibet," *Minerals*, vol. 12, p. 1578, 12 2022.

[12] P. Gillet, G. Fiquetw, J. M. Malézieux, and C. A. Geiger, "High-pressure and high-temperature raman spectroscopy of end-member garnets: pyrope, grossular and andradite," *European Journal of Mineralogy*, vol. 4, pp. 651–664, 08 1992.

[13] P. Mingsheng, H.-k. Mao, L. Dien, and E. Chao, "Raman spectroscopy of garnet-group minerals," *Chinese Journal of Geochemistry*, vol. 13, pp. 176–183, 04 1994.

[14] Wikipedia contributors, "Voigt profile — Wikipedia, the free encyclopedia," 2024. [Online; accessed 1-May-2024].

[15] RRUFF, "Database of Raman spectroscopy." `https://rruff.info/`. Accessed: 2024-04-27.

[16] Weka, "Weka 3 - Data Mining with Open Source Machine Learning Software." `https://ml.cms.waikato.ac.nz/weka/index.html`. Accessed: 2024-06-17.

[17] Y. Bi, Y. Zhang, J. Yan, Z. Wu, and Y. Li, "Classification and discrimination of minerals using laser induced breakdown spectroscopy and raman spectroscopy*," *Plasma Science and Technology*, vol. 17, p. 923, nov 2015.

[18] MathWorks, "MATLAB - El lenguaje del cálculo técnico." `https://es.mathworks.com/products/matlab.html`. Accessed: 2024-06-17.

[19] J. Schönig, H. von Eynatten, and R. e. a. Tolosana-Delgado, "Garnet major-element composition as an indicator of host-rock type: a machine learning approach using the random forest classifier," *Contrib Mineral Petrol*, vol. 176, 11 2021.

[20] S. T. Ishikawa and V. C. Gulick, "An automated mineral classifier using raman spectra," *Computers & Geosciences*, vol. 54, pp. 259–268, 2013.

[21] C. Carey, T. Boucher, S. Mahadevan, P. Bartholomew, and M. Dyar, "Machine learning tools formineral recognition and classification from raman spectroscopy," *Journal of Raman Spectroscopy*, vol. 46, 08 2015.

[22] Y. Qi, D. Hu, Y. Jiang, Z. Wu, M. Zheng, E. X. Chen, Y. Liang, M. A. Sadi, K. Zhang, and Y. P. Chen, "Recent progresses in machine learning assisted raman spectroscopy," *Advanced Optical Materials*, vol. 11, no. 14, p. 2203104, 2023.

[23] Scikit-learn, "scikit-learn: Machine Learning in Python." `https://scikit-learn.org/stable/`. Accessed: 2024-06-17.

[24] T. M. Mitchell, *Machine learning*, vol. 1. McGraw-hill New York, 1997.

[25] F. Berzal, *Redes Neuronales and Deep Learning*. Fernando Berzal, 2018.

[26] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems, Amsterdam: Morgan Kaufmann, 3 ed., 2011.

[27] H. Song, H. Leng, Z. Hou, R. Gao, C. Chen, C. Meng, J. Sun, C. Li, and B. Ma, "Grouped-sampling technique to deal with unbalance in raman spectral data modeling," *Photodiagnosis and Photodynamic Therapy*, vol. 40, p. 103059, 2022.

[28] C. Chen, X. Wu, E. Zuo, C. Chen, X. Lv, and L. Wu, "R-gdorus technology: Effectively solving the raman spectral data imbalance in medical diagnosis," *Chemometrics and Intelligent Laboratory Systems*, vol. 235, p. 104762, 2023.

Apéndice A

Anexos

Los enlaces útiles de interés en este Trabajo Fin de Grado son:

- Repositorio del código: `https://gitlab.inf.uva.es/pauvega/garnet-classification`.