



**viu**

**Universidad  
Internacional  
de Valencia**

# **Predicción de la magnitud de la banda prohibida (bandgap) en compuestos inorgánicos mediante técnicas de Machine Learning**

Titulación:  
Máster en Big Data y  
Data Science

Curso Académico  
2023-2024

Alumno: Sandoval Brito  
Rodrigo Eduardo  
DNI: 1724006810

Director: Jose Carlos  
González, PhD.

Convocatoria:

Primera

# Índice general

<b>Índice de figuras</b>	<b>2</b>
<b>Índice de cuadros</b>	<b>2</b>
<b>1 Introducción</b>	<b>1</b>
1.1 Motivación . . . . .	1
1.2 Objetivos . . . . .	3
1.2.1 Objetivos generales . . . . .	3
1.2.2 Objetivos específicos . . . . .	3
<b>2 Estado del Arte</b>	<b>5</b>
2.1 Marco teórico . . . . .	5
2.1.1 Informática de Materiales . . . . .	5
2.1.2 Bandgap y Estructura Electrónica de los Materiales . . . . .	6
2.1.3 Técnicas de Machine Learning para la Predicción de Propiedades de Materiales . . . . .	7
2.1.4 Aplicaciones Prácticas y Futuras del Bandgap . . . . .	9
2.2 Trabajos relacionados . . . . .	11
2.2.1 Multi-fidelity machine learning models for accurate bandgap predictions of solid. (Pilania et al. 2017) . . . . .	11
2.2.2 Predicting the bandgaps of Inorganic Solids by Machine Learning. (Zhuo et al. 2018) . . . . .	12
2.2.3 Machine-Learning-Assisted Accurate bandgap Predictions of Functionalized MXene. (Rajan et al. 2018) . . . . .	12
2.2.4 Machine learning the bandgap properties of kesterite I <sub>2</sub> -II-IV-V <sub>4</sub> quaternary compounds for photovoltaics applications. (Weston et al. 2018) . . . . .	13
2.2.5 Bandgap Prediction for Large Organic Crystal Structures with Machine Learning. (Olsthoorn et al. 2019) . . . . .	13
2.2.6 Atom table convolutional neural networks for an accurate prediction of compounds properties. (Zeng et al. 2019) . . . . .	14
2.2.7 Extracting Knowledge from DFT: Experimental bandgap Predictions Through Ensemble Learning. (Kauwe et al. 2020) . . . . .	15
2.2.8 Learning properties of ordered and disordered materials from multi-fidelity data. (Chen et al. 2021) . . . . .	15
<b>3 Metodología y Desarrollo</b>	<b>17</b>
3.1 Selección de datos . . . . .	18
3.1.1 Herramientas de trabajo . . . . .	18
3.1.2 Extracción de datos de la API de Materials Project . . . . .	19
3.1.3 Características utilizadas . . . . .	20
3.2 Pre-procesamiento de datos . . . . .	23
3.2.1 Exploración inicial de los datos . . . . .	24
3.2.2 Limpieza de los datos . . . . .	26

3.2.3	Análisis de correlaciones y relevancia de las características . . . . .	26
3.2.4	Visualización de características relevantes . . . . .	31
3.3	Transformación de datos . . . . .	40
3.3.1	Aprendizaje no supervisado para selección de características . . . . .	40
3.3.2	Generación de características derivadas . . . . .	44
3.4	Minería de datos . . . . .	45
3.4.1	Aprendizaje supervisado en modelo de regresión . . . . .	45
3.4.2	Aprendizaje supervisado en modelo de clasificación . . . . .	52
3.5	Evaluación de resultados . . . . .	55
3.5.1	Modelo final . . . . .	55
3.5.2	Comparativa de resultados con otros trabajos relacionados . . . . .	57
3.5.3	Despliegue del modelo final con Amazon Web Services (AWS)	61
<b>4</b>	<b>Discusión y Conclusiones</b>	<b>67</b>
4.1	Conclusiones . . . . .	67
4.2	Trabajo futuro y limitaciones . . . . .	70

# Índice de figuras

<u>1.1</u>	<u>Bandgap en materiales conductores, semiconductores y aislantes</u>	2
<u>3.1</u>	<u>Correlaciones entre las variables</u>	28
<u>3.2</u>	<u>Importancia de las características</u>	31
<u>3.3</u>	<u>Distribución de los materiales del dataset según la simetría cristalina</u>	32
<u>3.4</u>	<u>Bandgap y densidad de los materiales según la simetría cristalina</u>	33
<u>3.5</u>	<u>Relación entre la densidad, el volumen y la simetría cristalina</u>	35
<u>3.6</u>	<u>Comparación de propiedades promedio según la simetría cristalina</u>	37
<u>3.7</u>	<u>Relación entre la densidad y energía por átomo según su estabilidad</u>	38
<u>3.8</u>	<u>Histograma del bandgap</u>	39
<u>3.9</u>	<u>Clustering de las características</u>	42
<u>3.10</u>	<u>Comparación de MAE promedio vs el Bandgap</u>	50
<u>3.11</u>	<u>Predicciones vs valores reales del bandgap</u>	52
<u>3.12</u>	<u>Matriz de confusión del RFC</u>	54
<u>3.13</u>	<u>Curva ROC del Modelo RFC</u>	55
<u>3.14</u>	<u>Comparación de predicciones y valores reales del modelo final</u>	56
<u>3.15</u>	<u>Interfaz web del usuario para la predicción del bandgap alojada en AWS</u>	62
<u>3.16</u>	<u>Resultado del bandgap recibido en email del usuario</u>	65

# Índice de cuadros

3.1	<u>Tipos de datos de las variables</u>	25
3.2	<u>Variables y abreviaciones</u>	27
3.3	<u>Comparación de resultados en tarea de regresión</u>	57
3.4	<u>Comparación de resultados en tarea de clasificación</u>	60

## Resumen

El presente TFM aborda la predicción de la banda prohibida (bandgap) en diversos materiales mediante técnicas avanzadas de Machine Learning. Utilizando un extenso dataset con cerca de 30,000 compuestos inorgánicos, obtenido a través de la API de Materials Project, se desarrollaron modelos robustos capaces de predecir con precisión el bandgap, una propiedad fundamental que determina las características electrónicas y ópticas de los materiales.

La metodología empleada siguió rigurosamente el proceso KDD (Knowledge Discovery in Databases), abarcando desde la extracción y preprocesamiento de datos hasta la aplicación de técnicas de Machine Learning y la comparación entre los distintos modelos evaluados. Se realizaron también análisis exploratorios exhaustivos, estudiando correlaciones entre variables, evaluando la importancia de características y generando visualizaciones reveladoras, que aportaron una comprensión más detallada de los factores que influyen en el bandgap.

Mediante la combinación estratégica de un modelo de clasificación Random Forest y un modelo de regresión Gradient Boosting en un ensamble final, se logró un rendimiento altamente competitivo, superando en algunos casos incluso el estado del arte en este campo de estudio. Consiguiendo así un modelo de clasificación que alcanzó una precisión superior al 92 % en la distinción entre compuestos metálicos y no metálicos, mientras que el modelo de regresión obtuvo un error medio absoluto de 0.3 eV y un coeficiente de determinación de 0.89, para la predicción del valor del bandgap.

Finalmente en este trabajo, se implementó exitosamente el despliegue del modelo final en una aplicación web alojada en AWS, proporcionando así una herramienta accesible y eficiente para la comunidad científica. De esta manera los usuarios pueden ingresar las características de un compuesto y recibir una predicción precisa del bandgap en cuestión de minutos. De este modo, mediante la combinación de inteligencia artificial y ciencia de materiales hemos logrado desarrollar aplicaciones prácticas que poseen un gran potencial para acelerar el descubrimiento de nuevos materiales con propiedades a medida.

**Palabras clave:** Bandgap, Machine Learning, Materials Informatics, AWS

# 1. Introducción

## 1.1. Motivación

El presente trabajo se centra en el campo emergente de la informática de materiales, también conocido como Materials Informatics. El cual es un campo interdisciplinario que combina principios de la ciencia de materiales, la informática y la ciencia de datos con el propósito de impulsar el descubrimiento y diseño de nuevos materiales de manera más eficiente.

La iniciativa del Material Genome Initiative (MGI), lanzada en 2011 por el gobierno de los Estados Unidos, ha dado un fuerte impulso a este esfuerzo por acelerar el descubrimiento, desarrollo y despliegue de nuevos materiales. El MGI ha facilitado la colaboración y el acceso a recursos, marcando una era en la que predecir rápidamente las propiedades de materiales aún no sintetizados se ha vuelto esencial.

El progreso tecnológico depende en gran medida de nuestra capacidad para descubrir y diseñar nuevos materiales con propiedades funcionales específicas. A pesar de los avances en recursos computacionales y algoritmos, la complejidad combinatoria de este campo exige enfoques innovadores. Lo cual solo es posible debido a la acumulación de datos científicos y la disponibilidad de bases de datos abiertas, tales como The Materials Project, un proyecto colaborativo entre el Laboratorio Lawrence Berkeley y el Laboratorio del MIT, que ofrece a la comunidad científica una base de datos open-source con propiedades de diversos materiales.

De esta manera en lugar de depender exclusivamente de la experimentación tradicional, el ensayo y el error, la Informática de Materiales utiliza técnicas computacionales y algoritmos para recolectar, analizar y modelar extensos conjuntos de datos. Estos datos abarcan una amplia gama de propiedades de los materiales, tales como sus estructuras cristalinas, composiciones químicas y diversas características físicas. Al hacerlo, se busca maximizar la eficiencia en la identificación y desarrollo de materiales con propiedades altamente específicas o aplicaciones innovadoras, reduciendo significativamente el tiempo y los recursos necesarios para su descubrimiento.

Existen numerosas propiedades de un material, tales como la conductividad eléctrica o el espectro de absorción óptica, que están fundamentalmente determinadas por su estructura electrónica, la cual se origina en la naturaleza cuántica de los electrones subyacentes. La capacidad de descubrir o incluso diseñar materiales con propiedades electrónicas específicas es crucial para mantener el avance tecnológico actual.

Dentro de las múltiples propiedades que pueden caracterizar a un material, el bandgap o banda prohibida emerge como una de las más relevantes, esta propiedad refleja la diferencia energética entre la banda de valencia, donde se encuentran los electrones más energéticamente estables, y la banda de conducción, el nivel energético más bajo no ocupado. Lo cual convierte al bandgap en una de las propiedades más influyentes de un material diseñado para el mundo altamente electrónico y moderno en el que vivimos hoy en día.

El bandgap es crucial para una amplia variedad de aplicaciones, incluyendo, pero no limitándose a, la generación de energía fotovoltaica, dispositivos electrónicos y procesos de catálisis. El valor del bandgap no solo permite determinar si un material es o no un metal, sino que también ofrece la posibilidad de diseñar y descubrir materiales con propiedades electrónicas deseadas de manera más directa y económica. Esto representa una alternativa atractiva frente a los enfoques tradicionales, que a menudo requieren de experimentación física costosa o cálculos computacionales de gran intensidad.

Además el valor del bandgap determina si los materiales se comportan como aislantes, semiconductores o conductores. Un material con un bandgap amplio es típicamente un aislante, impidiendo el libre movimiento de electrones y, por ende, la conducción eléctrica. En contraste, un bandgap estrecho indica un semiconductor, permitiendo la movilidad electrónica bajo ciertas condiciones y posibilitando una conductividad controlada. Los materiales conductores o metales, por otro lado, presentan un bandgap nulo, facilitando una alta movilidad electrónica y una eficiente conductividad eléctrica, tal como se observa en la Figura 1.1.

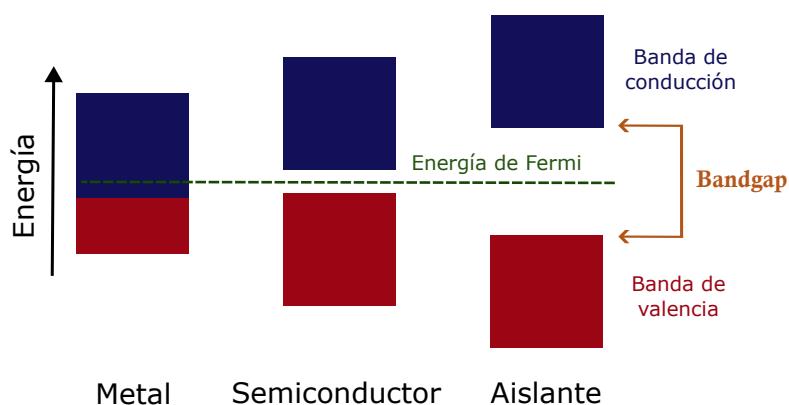


Figura 1.1: Bandgap en materiales conductores, semiconductores y aislantes

La predicción del bandgap por medio de técnicas de Machine Learning posibilita la identificación y combinación de materiales con bandgaps variados para crear estructuras compuestas optimizadas para aplicaciones específicas, tanto en la ciencia

como en la industria. Este conocimiento abre la puerta al desarrollo de dispositivos electrónicos más eficientes, materiales avanzados para la captura y almacenamiento de energía y catalizadores innovadores para reacciones químicas. Además la combinación de materiales con el bandgap adecuado junto con propiedades elásticas suaves abre un abanico de posibilidades hacia el desarrollo de dispositivos electrónicos flexibles novedosos. Todas estas múltiples aplicaciones marcan un camino prometedor hacia nuevos horizontes de innovación en diversos sectores de la ciencia de materiales y la ingeniería.

En aplicaciones como LEDs y células solares, el tamaño del bandgap también juega un rol crucial. Estas dos potenciales aplicaciones son especialmente prometedoras, dada la capacidad de la energía solar para contribuir significativamente más allá del 4 por ciento del consumo energético mundial actual.

La importancia del bandgap en el avance de tecnologías electrónicas es indiscutible, subrayando su papel esencial en la innovación de materiales. Esta propiedad no solo es clave en el diseño y desarrollo de los dispositivos electrónicos que definen nuestra era, sino que también impulsa la búsqueda de materiales superiores. La capacidad de predecir el bandgap, permite a los investigadores descubrir materiales con propiedades optimizadas, promoviendo así el desarrollo de tecnologías más eficientes, duraderas y sostenibles. De esta manera el enfoque innovador del campo de la Informática de Materiales nos brinda oportunidades sin precedentes para afrontar los desafíos presentes y prever las demandas tecnológicas futuras.

## 1.2. Objetivos

### Objetivos generales

- Aplicar técnicas de Machine Learning para construir un modelo de regresión que realice predicciones del valor del bandgap.
- Aplicar técnicas de Machine Learning para construir un modelo de clasificación que separe los compuestos metálicos (bandgap igual a cero) de los no metálicos (bandgap distinto de cero).

### Objetivos específicos

- Realizar un análisis exploratorio detallado de las relaciones entre las características de los materiales y el bandgap.
- Aplicar aprendizaje no supervisado para identificar subconjuntos de características altamente correlacionadas o redundantes, que permitan una reducción efectiva de la dimensionalidad del dataset y una representación más compacta de las propiedades clave.

- Implementar estrategias de ingeniería de características, escogiendo las más relevantes, combinando varias de ellas o extrayendo nuevas características con la finalidad de optimizar el dataset para facilitar el aprendizaje del modelo.
- Diseñar, entrenar y optimizar modelos de regresión y clasificación mediante aprendizaje supervisado que utilicen el dataset preprocesado para predecir la magnitud de la banda prohibida (bandgap) y evaluar el desempeño de los distintos modelos utilizados.
- Desarrollar una interfaz web alojada en AWS (Amazon Web Services), con un front-end intuitivo para ingresar datos de nuevos compuestos inorgánicos, un back-end robusto que procese la información mediante algoritmos de aprendizaje automático, y un módulo de notificación por correo electrónico que envíe los resultados de la predicción del bandgap al usuario en pocos minutos.

## 2. Estado del Arte

### 2.1. Marco teórico

#### Informática de Materiales

La informática de materiales (Materials Informatics) es un campo emergente que combina los principios de la ciencia de materiales con las herramientas y técnicas de la ciencia de datos. Esta disciplina ha surgido como respuesta a la creciente necesidad de acelerar el descubrimiento, diseño y optimización de nuevos materiales. En un mundo donde la tecnología avanza rápidamente, la demanda de materiales con propiedades específicas y alto rendimiento es cada vez mayor. Sin embargo, el proceso tradicional de desarrollo de materiales es lento, costoso y a menudo se basa en un enfoque de prueba y error.

La informática de materiales busca abordar estos desafíos mediante la integración de métodos computacionales, aprendizaje automático y minería de datos con la ciencia de materiales tradicional. Al aprovechar la gran cantidad de datos generados por experimentos y simulaciones, la informática de materiales permite a los investigadores identificar patrones, correlaciones y relaciones estructura-propiedad que pueden ser difíciles de descubrir utilizando enfoques convencionales. Esto conduce a un proceso más eficiente y dirigido para el descubrimiento y diseño de materiales.

Un aspecto clave de la informática de materiales es el desarrollo de bases de datos completas y bien estructuradas que contienen información sobre las propiedades, estructuras y métodos de síntesis de una amplia gama de materiales. Estas bases de datos, como Materials Project, AFLOW y OQMD, no solo sirven como repositorios de conocimiento, sino que también permiten la aplicación de técnicas de aprendizaje automático para extraer información valiosa y hacer predicciones sobre nuevos materiales.

El aprendizaje automático, una subcategoría de la inteligencia artificial, juega un papel fundamental en la informática de materiales. Los algoritmos de aprendizaje automático, como las redes neuronales, las máquinas de vectores de soporte y los árboles de decisión, pueden aprender de los datos existentes y hacer predicciones precisas sobre las propiedades de los materiales. Estas técnicas permiten a los investigadores explorar vastos espacios de diseño de materiales de manera eficiente, identificar candidatos prometedores y guiar los experimentos hacia direcciones más fructíferas.

Además del aprendizaje automático, la informática de materiales se beneficia y complementa otros métodos computacionalmente más costosos, como los cálculos

de estructura electrónica basados en la teoría del funcional de la densidad (DFT). Estos cálculos proporcionan información valiosa sobre las propiedades electrónicas, ópticas y magnéticas de los materiales, que pueden utilizarse para alimentar los modelos de aprendizaje automático y mejorar su capacidad predictiva.

La combinación de la ciencia de datos con la ciencia de materiales en la informática de materiales está revolucionando la forma en que se descubren y diseñan los materiales. Al aprovechar el poder de los datos y la computación, los investigadores pueden acelerar significativamente el proceso de desarrollo de materiales, reducir los costos y descubrir materiales con propiedades sin precedentes. Esta sinergia entre dos disciplinas aparentemente distintas está abriendo nuevas oportunidades y allanando el camino para un futuro más brillante y sostenible.

## Bandgap y Estructura Electrónica de los Materiales

El bandgap, también conocido como banda prohibida o brecha de banda, es una propiedad fundamental de los materiales que desempeña un papel crucial en la determinación de sus propiedades electrónicas y ópticas. En términos simples, el bandgap se refiere a la diferencia de energía entre la parte superior de la banda de valencia y la parte inferior de la banda de conducción en la estructura electrónica de un material. Esta estructura electrónica surge de la teoría de bandas, que describe el comportamiento de los electrones en un sólido.

Dado que el bandgap representa una diferencia de energía, se mide en Electronvoltios (eV). El electronvoltio es una unidad de energía que se utiliza comúnmente en física atómica, molecular y de partículas. Se define como la energía cinética adquirida por un electrón cuando es acelerado por una diferencia de potencial eléctrico de un voltio en el vacío.

En un átomo aislado, los electrones ocupan niveles de energía discretos. Sin embargo, cuando los átomos se unen para formar un sólido, sus orbitales atómicos se superponen y forman bandas continuas de energía. Estas bandas representan los rangos de energía permitidos para los electrones en el sólido. La banda de valencia contiene los electrones de valencia que participan en los enlaces químicos, mientras que la banda de conducción contiene los electrones que están débilmente unidos y pueden moverse libremente por el material, contribuyendo así a la conductividad eléctrica.

El bandgap es la diferencia de energía entre estas dos bandas. Su tamaño y naturaleza determinan si un material es un conductor, un semiconductor o un aislante. En los conductores, como los metales, la banda de valencia y la banda de conducción se superponen, lo que resulta en un bandgap nulo. Esto permite que los electrones se muevan fácilmente entre las bandas, lo que explica la alta conductividad eléctrica de los metales.

Por otro lado, los semiconductores y los aislantes tienen un bandgap distinto de cero. En los semiconductores, el bandgap es relativamente pequeño, típicamente entre 0.5 y 3 eV. A temperatura ambiente, algunos electrones pueden ser excitados térmicamente desde la banda de valencia hasta la banda de conducción, creando portadores de carga libres que contribuyen a la conductividad eléctrica. Esta conductividad puede ser controlada mediante el dopaje, que implica la introducción de impurezas en el material para modificar su estructura electrónica.

En los aislantes, el bandgap es mucho mayor, generalmente superior a 4 eV. Esta gran diferencia de energía hace que sea muy difícil para los electrones saltar a la banda de conducción, lo que resulta en una conductividad eléctrica muy baja o nula. Los materiales aislantes encuentran aplicaciones donde se requiere una alta resistencia eléctrica, como en los dieléctricos y en los materiales de aislamiento.

El bandgap también desempeña un papel esencial en las propiedades ópticas de los materiales. Cuando un fotón con una energía igual o superior al bandgap es absorbido por un material, puede excitar un electrón desde la banda de valencia hasta la banda de conducción. Este es el principio básico detrás de los dispositivos optoelectrónicos, como las células solares y los fotodetectores. La energía del bandgap determina el rango de longitudes de onda de la luz que puede ser absorbida o emitida por un material.

La comprensión y el control del bandgap son fundamentales para el diseño y la optimización de materiales funcionales. Mediante la ingeniería de la estructura electrónica, los científicos pueden crear materiales con propiedades a medida para aplicaciones específicas. Por ejemplo, los semiconductores con bandgaps ajustados se utilizan en células solares para maximizar la absorción de luz y la eficiencia de conversión de energía.

En resumen, el bandgap es una característica intrínseca de la estructura electrónica de un material que gobierna sus propiedades electrónicas y ópticas. Su origen reside en la teoría de bandas y su magnitud determina si un material es un conductor, un semiconductor o un aislante. El control del bandgap a través de la ingeniería de materiales es esencial para el desarrollo de tecnologías avanzadas en áreas como la electrónica, la optoelectrónica y la energía renovable.

## Técnicas de Machine Learning para la Predicción de Propiedades de Materiales

El machine learning (ML) se ha convertido en una herramienta poderosa en la informática de materiales para predecir las propiedades de los materiales. Las técnicas de ML permiten a los investigadores aprovechar las grandes cantidades de datos generados por experimentos y simulaciones para descubrir patrones y relaciones

complejas que pueden ser difíciles de identificar utilizando métodos tradicionales. Entre las diversas técnicas de ML, los algoritmos de Random Forest (RF) y Gradient Boosting (GB) han demostrado ser particularmente efectivos para la predicción de propiedades de materiales.

Random Forest es un algoritmo de aprendizaje supervisado que pertenece a la familia de los métodos de ensemble learning. Se basa en la construcción de múltiples árboles de decisión independientes que se entrenan en diferentes subconjuntos de los datos de entrenamiento. Cada árbol de decisión se construye utilizando un subconjunto aleatorio de las características (features) y un subconjunto aleatorio de las muestras de entrenamiento. Durante la predicción, cada árbol individual proporciona una predicción y la predicción final se obtiene mediante un voto mayoritario o un promedio de las predicciones de todos los árboles.

La fortaleza de Random Forest radica en su capacidad para manejar datos de alta dimensionalidad y evitar el sobreajuste (overfitting). Al promediar las predicciones de múltiples árboles, Random Forest reduce la varianza y mejora la generalización del modelo. Además, Random Forest puede manejar características ruidosas o irrelevantes, ya que cada árbol se entrena en un subconjunto diferente de características. Esta propiedad es particularmente útil en la informática de materiales, donde el número de características puede ser grande y no todas ellas pueden ser relevantes para la propiedad objetivo.

Por otro lado, Gradient Boosting es otro algoritmo de ensemble learning que combina múltiples modelos débiles para crear un modelo robusto. A diferencia de Random Forest, donde los árboles se construyen de forma independiente, en Gradient Boosting los árboles se construyen secuencialmente, donde cada árbol subsiguiente se entrena para corregir los errores cometidos por los árboles anteriores. El proceso de entrenamiento implica la minimización de una función de pérdida, como el error cuadrático medio, utilizando el descenso del gradiente.

Una de las ventajas de Gradient Boosting es su capacidad para capturar relaciones no lineales complejas entre las características y la propiedad objetivo. Al combinar múltiples modelos débiles, Gradient Boosting puede aproximar funciones complejas con alta precisión. Además, Gradient Boosting permite la optimización de diferentes funciones de pérdida, lo que lo hace adaptable a diversos tipos de problemas, como regresión, clasificación y ranking.

Tanto Random Forest como Gradient Boosting han demostrado un excelente rendimiento en la predicción de diversas propiedades de materiales, como la energía de formación, la conductividad eléctrica, la dureza y la estabilidad térmica. Por ejemplo, en un estudio realizado por Ward et al. [6], se utilizó el algoritmo de Random Forest como parte de un marco de trabajo general para predecir propiedades de materiales. Demostrando la capacidad de Random Forest para manejar datos de al-

ta dimensionalidad y hacer predicciones precisas en problemas complejos de ciencia de materiales.

En otro estudio, Kauwea et al. [2] utilizaron algoritmos como Gradient Boosting y Random Forest para construir ensambles que permitan la predicción de propiedades de materiales. Este trabajo destaca la importancia de combinar diferentes modelos y fuentes de datos para mejorar la precisión de las predicciones. El enfoque de ensamble permite aprovechar las fortalezas de cada algoritmo individual. Por ejemplo, Gradient Boosting es conocido por su capacidad para capturar relaciones no lineales complejas, mientras que Random Forest es robusto frente al sobreajuste y puede manejar características irrelevantes.

Además de su capacidad predictiva, tanto Random Forest como Gradient Boosting ofrecen la posibilidad de interpretar la importancia de las características utilizadas en el modelo. Esto es crucial en la informática de materiales, ya que no solo se busca hacer predicciones precisas, sino también obtener información sobre los factores subyacentes que influyen en las propiedades de los materiales. Al analizar la importancia de las características, los investigadores pueden obtener una comprensión más profunda de las relaciones estructura-propiedad y guiar el diseño de nuevos materiales.

En resumen, las técnicas de Machine Learning, como Random Forest y Gradient Boosting, han demostrado ser herramientas poderosas para la predicción de propiedades de materiales. Su capacidad para manejar datos de alta dimensionalidad, capturar relaciones complejas y proporcionar interpretabilidad las convierte en valiosas herramientas en la informática de materiales. A medida que la cantidad de datos disponibles siga creciendo, se espera que estas técnicas desempeñen un papel cada vez más importante en el descubrimiento y diseño de nuevos materiales con propiedades optimizadas.

## Aplicaciones Prácticas y Futuras del Bandgap

La predicción del bandgap de los materiales tiene numerosas aplicaciones prácticas en diversos campos tecnológicos, especialmente en la optoelectrónica y la energía renovable. Dos de las aplicaciones más destacadas son los diodos emisores de luz (LEDs) y las células solares, donde el tamaño del bandgap juega un papel fundamental en el rendimiento y la eficiencia de estos dispositivos.

En el caso de los LEDs, la predicción precisa del bandgap es esencial para el diseño y la optimización de estos dispositivos. Los LEDs son ampliamente utilizados en iluminación, pantallas y señalización debido a su alta eficiencia energética, larga vida útil y capacidad para emitir luz de colores específicos. El color de la luz emitida por un LED está determinado por la energía del bandgap del material semiconductor utilizado. Cuanto mayor sea el bandgap, mayor será la energía de los fotones emi-

tidos y, por lo tanto, la luz tendrá una longitud de onda más corta, acercándose al extremo azul del espectro visible. Por otro lado, un bandgap más pequeño da lugar a la emisión de fotones de menor energía y luz de mayor longitud de onda, hacia el extremo rojo del espectro.

Mediante la predicción del bandgap, los investigadores pueden identificar y diseñar materiales semiconductores con bandgaps específicos para obtener LEDs que emitan luz de colores deseados. Esto permite la creación de LEDs altamente eficientes y personalizados para diversas aplicaciones, como la iluminación de estado sólido, las pantallas de dispositivos electrónicos y la señalización óptica. Además, la capacidad de ajustar el bandgap también permite el desarrollo de LEDs que emiten luz en el rango ultravioleta o infrarrojo, ampliando aún más las aplicaciones potenciales de estos dispositivos.

Por otro lado, en las células solares, la predicción del bandgap es crucial para optimizar la eficiencia de conversión de energía. Las células solares convierten la energía de la luz solar en electricidad mediante el efecto fotovoltaico. Cuando un fotón con una energía igual o mayor que el bandgap del material es absorbido, se genera un par electrón-hueco, que posteriormente se separa y se recolecta para generar una corriente eléctrica. El bandgap óptimo para una célula solar debe permitir la absorción de una amplia gama de longitudes de onda de la luz solar, maximizando así la cantidad de energía capturada y convertida en electricidad.

Si el bandgap de una célula solar es demasiado grande, no podrá absorber eficientemente los fotones de baja energía (luz de mayor longitud de onda), lo que limitará su capacidad para aprovechar una parte significativa del espectro solar. Por otro lado, si el bandgap es demasiado pequeño, aunque la célula solar podrá absorber una mayor parte del espectro, la energía excedente de los fotones de alta energía se perderá en forma de calor en lugar de convertirse en electricidad útil. Por lo tanto, existe un bandgap óptimo que maximiza la eficiencia de conversión de energía de una célula solar.

Mediante la predicción del bandgap, los investigadores pueden identificar materiales semiconductores con bandgaps ideales para aplicaciones fotovoltaicas. Esto permite el diseño de células solares de alta eficiencia que puedan aprovechar al máximo el espectro solar y convertir una mayor proporción de energía lumínica en electricidad. Además, la predicción del bandgap también es relevante para el desarrollo de células solares de múltiples uniones, donde se apilan materiales con diferentes bandgaps para absorber eficientemente diferentes partes del espectro solar, lo que resulta en eficiencias de conversión aún mayores.

Más allá de los LEDs y las células solares, la predicción del bandgap tiene aplicaciones en otros campos, como la fotocatálisis, donde los materiales con bandgaps adecuados pueden utilizarse para catalizar reacciones químicas impulsadas por la

luz, como la descomposición del agua para producir hidrógeno como combustible limpio. También es relevante en el desarrollo de detectores de luz, dispositivos espintrónicos y computación cuántica, donde el control preciso del bandgap es esencial para el funcionamiento óptimo de estos sistemas.

Mirando hacia el futuro, se espera que la predicción del bandgap desempeñe un papel cada vez más importante en el descubrimiento y diseño de nuevos materiales funcionales. A medida que se generan más datos experimentales y computacionales, y se desarrollan algoritmos de aprendizaje automático más avanzados, será posible predecir con mayor precisión el bandgap de una amplia gama de materiales, incluso antes de su síntesis. Esto acelerará enormemente el proceso de descubrimiento de materiales y permitirá el diseño eficaz de materiales con propiedades optoelectrónicas y fotovoltaicas optimizadas.

En resumen, la predicción del bandgap tiene aplicaciones prácticas significativas en tecnologías como los LEDs y las células solares, donde el tamaño del bandgap determina el color de la luz emitida y la eficiencia de conversión de energía, respectivamente. A medida que avanzamos hacia un futuro más sostenible y energéticamente eficiente, la capacidad de predecir y controlar el bandgap de los materiales será cada vez más valiosa, impulsando la innovación en la iluminación, la energía renovable y otras áreas tecnológicas clave.

## 2.2. Trabajos relacionados

### **Multi-fidelity machine learning models for accurate bandgap predictions of solid. (Pilania et al. 2017)**

En este trabajo [4] los autores abordan la predicción del bandgap utilizando modelos de aprendizaje automático de múltiples fidelidades. Ya que reconocen que aunque sería ideal entrenar modelos con datos experimentales o cálculos de alta precisión, obtener suficientes datos de alta calidad es extremadamente difícil debido a los altos costos computacionales y experimentales. Para superar esta limitación, proponen un enfoque de aprendizaje automático de múltiples fidelidades que combina datos de baja fidelidad (pero abundantes) con datos de alta fidelidad (pero escasos).

Los autores emplean procesos gaussianos de múltiples fidelidades basados en la regresión de co-kriging (Cok) para mejorar significativamente la precisión de las predicciones del bandgap en comparación con los modelos de fidelidad única. La regresión de co-kriging es una técnica geoestadística que combina datos de diferentes fidelidades para mejorar la precisión de las predicciones. Modela la relación entre los datos mediante una función de covarianza que captura la similitud espacial y la correlación entre las fuentes. Luego, interpola y predice valores en ubicaciones no muestreadas, considerando datos de baja y alta fidelidad, lo que permite obte-

ner predicciones más precisas y confiables, especialmente cuando los datos de alta fidelidad son limitados.

### **Predicting the bandgaps of Inorganic Solids by Machine Learning. (Zhao et al. 2018)**

En este trabajo [9] los autores proponen un enfoque que consiste en desarrollar un modelo capaz de estimar los valores experimentales de bandgap sin depender de cálculos computacionalmente costosos basados en la teoría del funcional de la densidad (DFT). Este enfoque se basa en un modelo de aprendizaje supervisado que aborda tanto tareas de regresión como de clasificación utilizando especialmente algoritmos de Gradient Boosting, Random Forest y vectores de soporte SVR.

El modelo SVR se entrena utilizando el conjunto de características seleccionadas y los bandgaps experimentales correspondientes. SVR es un algoritmo de aprendizaje supervisado que busca encontrar un hiperplano que maximice el margen entre las clases y minimice el error de predicción. Utiliza una función kernel para transformar los datos de entrada en un espacio de características de mayor dimensión, donde se puede encontrar un hiperplano de separación óptimo.

También se sugiere una corrección adicional que tiene en cuenta la contribución de la diferencia de electronegatividad general, lo que proporciona una explicación del efecto de curvatura (bowing) observado en la curva de bandgap frente a la composición. En términos simples, el efecto de curvatura describe cómo la banda prohibida de un material compuesto puede desviarse de manera significativa de una línea recta cuando se combina con otros materiales. Esto significa que la relación entre la composición del material y su banda prohibida no es lineal, como podría esperarse intuitivamente, sino que muestra una curvatura en la gráfica de la banda prohibida en función de la composición.

### **Machine-Learning-Assisted Accurate bandgap Predictions of Functionalized MXene. (Rajan et al. 2018)**

En este artículo [5] los autores presentan un enfoque de aprendizaje automático para predecir el bandgap de los MXenes funcionalizados. Los MXenes son una clase de materiales bidimensionales con propiedades electrónicas y ópticas fascinantes, y la funcionalización de su superficie ofrece una vía prometedora para ajustar su bandgap. Sin embargo, la exploración exhaustiva de los MXenes funcionalizados mediante cálculos utilizando primeros principios es computacionalmente costosa debido al gran espacio de posibles estructuras. Para abordar este desafío, los autores exploran varios algoritmos como el Proceso Gaussiano (GPR) con el kernel Matérn 5/2 para regresión y el Bagging para clasificación.

El GPR es un modelo no paramétrico basado en distribuciones de probabilidad, donde se asume que el valor de predicción es el resultado de un proceso gaussiano afectado por un ruido aditivo independiente. Dado un conjunto de datos de entrenamiento y un nuevo punto de entrada, el GPR puede predecir la distribución de probabilidad de la variable objetivo. El kernel Matérn 5/2 es una función de covarianza que mide la similitud entre dos puntos de entrada, teniendo en cuenta la distancia entre ellos.

Por otro lado el bagging, o Bootstrap Aggregating, es una técnica de conjunto que busca mejorar la precisión de la clasificación al combinar las predicciones de múltiples subclásificadores. Bajo este enfoque, se generan múltiples conjuntos de datos de entrenamiento mediante remuestreo bootstrap y se entrena un clasificador base en cada uno. Luego, las predicciones de cada clasificador base se promedian o votan para producir una predicción final. Este enfoque ayuda a reducir la varianza y mejorar la generalización del modelo, lo que puede conducir a un mejor rendimiento predictivo en comparación con un único clasificador.

### **Machine learning the bandgap properties of kesterite I2-II-IV-V4 quaternary compounds for photovoltaics applications. (Weston et al. 2018)**

Este trabajo [7] explora la aplicación de técnicas de aprendizaje automático para predecir las propiedades del bandgap en compuestos cuaternarios tipo kesterita I2-II-IV-V4. Estos materiales son de gran interés para aplicaciones fotovoltaicas debido a su alta eficiencia y bajo costo. Sin embargo, la búsqueda experimental de nuevas composiciones con las propiedades deseadas es un proceso costoso y que consume mucho tiempo.

Los autores proponen un enfoque computacional para acelerar el descubrimiento de nuevos compuestos kesterita prometedores. Utilizando una base de datos de materiales conocidos, se entrena distintos modelos de aprendizaje automático para predecir el bandgap a partir de las características composicionales y estructurales de los compuestos. El modelo final demuestra una alta capacidad para identificar nuevos materiales con propiedades óptimas para celdas solares. Este trabajo sienta las bases para el diseño eficaz de compuestos cuaternarios tipo kesterita mediante técnicas de machine learning, abriendo nuevas posibilidades para el desarrollo de tecnologías fotovoltaicas eficientes y sostenibles.

### **Bandgap Prediction for Large Organic Crystal Structures with Machine Learning. (Olsthoorn et al. 2019)**

En este trabajo [3] los autores presentan un nuevo conjunto de datos llamado OMDB-GAP1, que contiene valores de bandgap calculados mediante la teoría del funcional de la densidad (DFT) para 12,500 estructuras cristalinas orgánicas. Este

conjunto de datos es único debido a su consistencia y al gran tamaño de las estructuras cristalinas, con un promedio de 82 átomos por celda unitaria. Los autores utilizan este conjunto de datos para evaluar el rendimiento de dos métodos de aprendizaje automático: la regresión de ridge kernel (KRR) con el kernel SOAP (Smooth Overlap of Atomic Positions) y una red neuronal profunda llamada SchNet, comúnmente utilizada en el modelado de sistemas químicos y de materiales.

La regresión de ridge (KRR) con el kernel SOAP transforma las coordenadas atómicas en una representación invariante a la rotación y traslación, capturando la distribución de los vecinos de cada átomo. La similitud entre estructuras se calcula como el producto escalar de sus descriptores SOAP. KRR encuentra una relación lineal entre estos descriptores y la propiedad de interés, minimizando el error cuadrático medio regularizado. Al igual que la regresión SVR, el modelo KRR utiliza funciones kernel para transformar los datos a un espacio de mayor dimensionalidad, sin embargo, a diferencia de KRR que se enfoca solamente en minimizar el error cuadrático medio regularizado, SVR busca un equilibrio entre la minimización del error y la maximización del margen, lo que suele permitirle tener una mayor capacidad de generalización.

Además, en este estudio los autores utilizan el kernel SOAP aprendido para medir la similitud estructural entre los materiales y visualizar la relación entre las estructuras cristalinas y sus bandgaps utilizando el algoritmo de reducción de dimensionalidad t-SNE. Esta visualización proporciona una forma intuitiva de navegar por el paisaje estructura-propiedad de los cristales orgánicos e identificar materiales con una funcionalidad específica. Los autores también analizan los embeddings elementales aprendidos por el modelo SchNet y demuestran que el modelo ha aprendido similitudes entre ciertos tipos de átomos de manera análoga a la tabla periódica.

### **Atom table convolutional neural networks for an accurate prediction of compounds properties. (Zeng et al. 2019)**

En este artículo [8] los autores desarrollan un framework de aprendizaje automático basado en una red neuronal convolucional llamada ATCNN (Atom Table Convolutional Neural Networks) que utiliza exclusivamente información composicional de un material para aprender directamente sus propiedades experimentales. El modelo ATCNN se aplica para predecir propiedades como el bandgap para distintos compuestos.

Los autores demuestran que la precisión del modelo ATCNN supera los cálculos de la teoría del funcional de la densidad (DFT) para la predicción de las propiedades predichas. Además, los autores utilizan el análisis de componentes principales (PCA) para visualizar las características aprendidas de los elementos del grupo principal y descubren que el modelo ATCNN efectivamente aprende las propiedades de los elementos y reproduce las tendencias químicas que reflejan la física subyacente

de las propiedades predichas.

Este estudio presenta un marco general de aprendizaje automático que aprende características directamente de la información composicional y logra una alta precisión en la predicción de diversas propiedades de los compuestos, siendo este artículo de Nature uno de los principales exponentes del enorme potencial de los enfoques de aprendizaje automático para acelerar el descubrimiento y diseño de nuevos materiales.

### **Extracting Knowledge from DFT: Experimental bandgap Predictions Through Ensemble Learning. (Kauwe et al. 2020)**

En este artículo [2] los autores presentan un enfoque novedoso para mejorar la predicción del bandgap en materiales mediante el aprovechamiento de datos de diferentes fuentes usando técnicas de aprendizaje automático ensemble. El objetivo es combinar de manera eficaz datos experimentales limitados pero de alta calidad con grandes cantidades de datos computacionales de menor precisión obtenidos mediante DFT.

El modelo ensemble desarrollado logra reducir significativamente el error en la predicción del bandgap en comparación con modelos entrenados usando solo datos experimentales. Esto se logra entrenando un modelo neuronal con los datos DFT para generar predicciones que luego se combinan con modelos entrenados en datos experimentales mediante un meta-learner. Este enfoque permite extraer conocimiento útil de los extensos datos DFT sin introducir el sesgo inherente en los cálculos DFT. Los resultados demuestran el potencial de integrar inteligentemente múltiples fuentes de datos para mejorar las predicciones de propiedades de materiales y acelerar el descubrimiento de nuevos compuestos prometedores.

### **Learning properties of ordered and disordered materials from multi-fidelity data. (Chen et al. 2021)**

En este artículo de Nature [1] los autores desarrollan una novedosa arquitectura de redes neuronales de grafos multi-fidelidad como un enfoque universal para lograr predicciones precisas de propiedades de materiales con conjuntos de datos pequeños. La idea central es combinar una gran cantidad de datos computacionales de baja fidelidad pero de bajo costo (por ejemplo, cálculos DFT con funcionales PBE) con conjuntos más pequeños de datos de alta fidelidad (por ejemplo, cálculos DFT con funcionales híbridos o datos experimentales) para entrenar modelos predictivos superiores.

Este trabajo demuestra que la inclusión de datos de los bandgaps calculados con PBE mejora significativamente la capacidad de los modelos para extraer características estructurales clave de los materiales, llevando a una reducción significativa en

el error absoluto medio (MAE) de las predicciones de los bandgaps experimentales. Además, muestran que los embeddings elementales aprendidos por la red proveen una forma natural de representar el desorden en materiales, lo cual aborda una limitación importante de los métodos computacionales actuales. En general, el enfoque de redes de grafos multi-fidelidad permite aprovechar eficientemente datos abundantes de baja fidelidad para mejorar la precisión en la predicción de propiedades obtenidas por métodos experimentales o computacionales de alto costo.

### 3. Metodología y Desarrollo

El presente trabajo se desarrollará siguiendo la metodología KDD (Knowledge Discovery in Databases), cuyo enfoque consiste en extraer conocimiento útil a partir de grandes conjuntos de datos. La metodología KDD consta de las siguientes etapas:

1. Selección de datos
2. Pre-procesamiento de datos
3. Transformación de datos
4. Minería de datos
5. Evaluación de resultados

En la etapa de **selección de datos**, se utilizará la API de Materials Project para extraer un dataset con propiedades de alrededor de 30 mil compuestos inorgánicos. Esto permitirá contar con una amplia base de datos para el posterior análisis.

Durante el **pre-procesamiento de datos**, se realizará una exploración inicial de los datos para entender su estructura y características. Luego se procederá a limpiar el dataset, con el fin de evitar registros o campos con valores nulos. También se generarán visualizaciones para apreciar relaciones entre las características más relevantes del dataset.

La **transformación de los datos** involucrará técnicas de aprendizaje no supervisado para realizar clustering y selección de características. Además, se generarán características derivadas a partir de las originales para enriquecer la representación de los compuestos.

En la fase de **minería de datos**, se aplicarán algoritmos de aprendizaje supervisado tanto para enfoques de regresión como de clasificación. Esto permitirá combinar ambos acercamientos en un modelo final que predice el valor numérico del bandgap y a su vez clasifica los materiales entre metales y no metales.

Finalmente, en la **evaluación** se analizarán los resultados de los modelos desarrollados. Se comparará su desempeño con trabajos previos relacionados, demostrando que el modelo final desarrollado presenta un mejor desempeño con respecto a varios trabajos relacionados. Como paso adicional, se implementará el despliegue del modelo final en una página web, donde el usuario podrá ingresar las características de un compuesto y recibir vía email la predicción del bandgap correspondiente.

Por lo tanto, el presente trabajo aplica rigurosamente la metodología KDD, desde la obtención de datos hasta el despliegue de un producto final. Lo cual permite descubrir conocimiento valioso sobre la relación entre el bandgap y otras características de los compuestos inorgánicos. A continuación se abordará el desarrollo realizado en cada una de las etapas de la metodología KDD.

### 3.1. Selección de datos

En esta sección se profundizará en temas como las herramientas utilizadas en el desarrollo de este proyecto, el proceso de selección de datos para nuestro modelo de predicción del bandgap y la relevancia de las características que compondrán nuestro dataset.

#### Herramientas de trabajo

En el desarrollo de este trabajo se han utilizado diversas herramientas entre las que se destacan varias librerías de Python y servicios de AWS. A continuación, se detallan las principales herramientas empleadas:

- **Pymatgen:** Es una biblioteca de Python ampliamente utilizada en el campo de la ciencia de materiales. En este proyecto, se ha utilizado la clase MPRester de Pymatgen para acceder a la base de datos de materiales del Materials Project. Mediante el uso de la API de Materials Project, se ha obtenido información sobre los compuestos inorgánicos de interés, incluyendo sus propiedades estructurales y electrónicas.
- **Scikit-learn:** Es una biblioteca de aprendizaje automático de código abierto para Python. Se ha utilizado extensivamente en este proyecto para diversas tareas, como el entrenamiento de modelos, la división de datos en conjuntos de entrenamiento y prueba (train test split), el escalado de características (MinMaxScaler), la evaluación de modelos (r2 score, mean absolute error) y la optimización de hiperparámetros (RandomizedSearchCV, GridSearchCV).
- **NumPy:** Es una biblioteca fundamental para el cálculo científico en Python. Se ha utilizado para diversas operaciones matemáticas y manipulación de datos, como la creación de arrays, el cálculo de promedios y la realización de operaciones elemento a elemento.
- **AWS (Amazon Web Services):** Para el despliegue del modelo, se ha utilizado la plataforma de servicios en la nube de Amazon. Específicamente, se han empleado los servicios de API Gateway, S3 y Lambda. API Gateway se ha utilizado para crear una API que permita interactuar con el modelo, S3 para almacenar el dataset de entrenamiento y alojar el front-end, y Lambda para ejecutar el modelo de predicción en respuesta a las solicitudes de la API.

La combinación de estas herramientas ha permitido llevar a cabo un flujo de trabajo completo, desde la obtención de los datos hasta la puesta en marcha del modelo para su uso en la predicción del bandgap de compuestos inorgánicos.

## Extracción de datos de la API de Materials Project

Para el proceso de recopilación de información sobre compuestos inorgánicos se ha utilizado la API proporcionada por el Materials Project. A continuación, se detalla los pasos más importantes realizados para la obtención del dataset utilizado:

1. Instalación de la biblioteca Pymatgen: Se ha utilizado sistema de gestión de paquetes pip para instalar la biblioteca Pymatgen, que proporciona una interfaz para interactuar con la API de Materials Project.
2. Autenticación en la API: Se ha utilizado una clave de API personal para autenticarse y acceder a los servicios de la API de Materials Project.
3. Obtención de campos disponibles: Utilizando la clase MPRester de Pymatgen, se ha obtenido la lista de campos disponibles (`list_of_available_fields`) que se pueden consultar según su índice relacionado (`material_id`).
4. Extracción de datos para distintos rangos del índice relacionado: Se ha implementado un bucle para realizar consultas a la API de Materials Project para distintos rangos del `material_id`. Es decir para cada iteración, se obtiene información sobre un material específico utilizando su `material_id` correspondiente.
5. Extracción de la información: Para cada material consultado exitosamente, se han extraído diversas propiedades y características relevantes, como la fórmula química, el bandgap, el número de sitios atómicos, el volumen, la densidad, la simetría cristalina, la energía por átomo, la estabilidad, las propiedades magnéticas, entre otras. Estos datos se han almacenado en un diccionario de python para su posterior uso.
6. Construcción del dataset: Los datos extraídos de cada material se han agregado como una nueva fila en un DataFrame de Pandas. Esto permite ir construyendo un conjunto de datos completo a medida que se recorren los índices relacionados con cada material.
7. Manejo de errores y compuestos faltantes: Durante el proceso de extracción, se ha encontrado que existen rangos completos de `material_id` para los cuales no hay datos disponibles en la API. Para manejar estos casos, se ha utilizado un bloque de manejo de errores que permite continuar con la ejecución del código incluso si no se encuentran datos para un `material_id` específico.

Es importante destacar que el proceso de recopilación de datos ha sido laborioso debido a la naturaleza dispersa de los datos en la API de Materials Project. Ha sido

necesario realizar un barrido exhaustivo de miles o incluso cientos de miles de índices `material_id` para obtener un conjunto de datos completo de cerca de 30,000 compuestos inorgánicos. Esta extracción en su totalidad ha llevado varios días de ejecución.

Además, cabe mencionar que la API entrega los datos de cada compuesto en formato de documento (tipo JSON), lo que requiere un procesamiento adicional para extraer las características relevantes y convertirlas en un formato tabular adecuado para su posterior análisis. Se han seleccionado principalmente características de tipo numérico y se han descartado características con entradas mayoritariamente nulas para garantizar la calidad de los datos.

Por lo tanto, el proceso de extracción de datos de la API de Materials Project ha implicado un esfuerzo significativo debido a la naturaleza dispersa de los datos y la necesidad de procesar grandes cantidades de información. Sin embargo, el resultado final es un conjunto de datos valioso que contiene información detallada sobre las propiedades y características de una amplia gama de compuestos inorgánicos, lo que sienta las bases para las futuras etapas de nuestro proyecto.

## Características utilizadas

El dataset obtenido del Material Project consta de la variable de predicción `bandgap` junto con 19 variables predictoras, a estas características las podemos dividir en dos categorías principales: composicionales y no composicionales. A continuación, se proporciona una descripción física de cada característica.

### Características composicionales

Las características composicionales de un material se refieren a la información relacionada con su composición química, es decir, los elementos químicos presentes y sus proporciones relativas. Estas características describen la identidad y la cantidad de los componentes básicos que forman el material. En este trabajo utilizaremos las siguientes características composicionales:

- **Fórmula química (formula\_pretty):** Representa la composición elemental del material, indicando los elementos presentes y su proporción estequiométrica. La fórmula química es fundamental para comprender la estructura y las propiedades del material. Además, como veremos más adelante este atributo nos será de utilidad para enriquecer nuestro dataset generando nuevas características a partir de la fórmula química.
- **Número de elementos (nelements):** Indica la cantidad de elementos químicos distintos presentes en el material. Este parámetro proporciona información sobre la complejidad composicional del material.

- **Número de sitios atómicos (nsites)**: Representa el número total de sitios atómicos en la estructura cristalina del material. Este valor está relacionado con el tamaño de la celda unitaria y la complejidad estructural del material.

### Características no composicionales

Las características no composicionales de un material abarcan una amplia gama de propiedades y atributos que van más allá de su composición química. Estas características describen aspectos estructurales, físicos, termodinámicos y electrónicos del material. En este trabajo utilizaremos las siguientes características no composicionales:

- **Volumen**: Representa el volumen de la celda unitaria del material en unidades de Angstroms cúbicos ( $\text{\AA}^3$ ). El volumen está directamente relacionado con el tamaño y la densidad del material.
- **Densidad**: Indica la densidad del material en unidades de gramos por centímetro cúbico ( $\text{g/cm}^3$ ). La densidad es una propiedad física fundamental que depende de la composición y la estructura cristalina del material.
- **Densidad atómica**: Representa la densidad del material en términos del número de átomos por unidad de volumen. Esta medida proporciona información sobre el empaquetamiento atómico en la estructura cristalina.
- **Simetría cristalina**: Describe la simetría de la estructura cristalina del material, como cúbica, tetragonal, hexagonal, etc. La simetría cristalina está estrechamente relacionada con las propiedades físicas y electrónicas del material.
- **Número de simetría**: Es un valor numérico que cuantifica la simetría cristalina del material. Un número de simetría más alto indica una mayor simetría en la estructura cristalina.
- **Lados de la celda unitaria (sides\_abc)**: Representa las longitudes de los lados de la celda unitaria del material en las direcciones a, b y c. Estos valores proporcionan información sobre las dimensiones y la forma de la celda unitaria.
- **Ángulos de la celda unitaria (angles\_abc)**: Indica los ángulos entre los lados de la celda unitaria del material, denotados como  $\alpha$ ,  $\beta$  y  $\gamma$ . Estos ángulos definen la geometría de la celda unitaria y están relacionados con la simetría cristalina.
- **Energía por átomo no corregida**: Representa la energía total por átomo del material calculada sin correcciones adicionales. Esta energía se obtiene directamente de los cálculos de primeros principios.
- **Energía por átomo**: Indica la energía total por átomo del material después de aplicar correcciones y ajustes necesarios. Esta energía proporciona una estimación más precisa de la estabilidad energética del material.

- **Energía de formación por átomo:** Representa la energía requerida para formar el material a partir de sus elementos constituyentes en condiciones estándar. Una energía de formación negativa indica que el material es estable termodinámicamente.
- **Energía sobre el casco convexo (energy\_above\_hull):** Indica la diferencia de energía entre el material y el estado de menor energía en el diagrama de fases correspondiente. Una energía sobre el casco convexo de cero indica que el material es estable, mientras que valores positivos sugieren metaestabilidad.
- **Estabilidad (is\_stable):** Es una variable binaria que indica si el material es termodinámicamente estable (valor de 1) o no (valor de 0) según los cálculos realizados.
- **Energía de Fermi (efermi):** Representa la energía del nivel de Fermi del material, que es un parámetro fundamental en la descripción de las propiedades electrónicas y la estructura de bandas del material, ya que se encuentra en medio de la banda de valencia y la banda de conducción, tal como lo muestra la Figura 1.1.
- **Magnetismo (is\_magnetic):** Es una variable binaria que indica si el material exhibe propiedades magnéticas (valor de 1) o no (valor de 0) según los cálculos realizados.
- **Magnetización total:** Representa la magnetización total del material, que es una medida de la densidad de momento magnético neto en el material.
- **Número de sitios magnéticos:** Indica la cantidad de sitios atómicos en el material que contribuyen al comportamiento magnético observado.

La predicción del bandgap utilizando tanto características composicionales como no composicionales ofrece varios beneficios en comparación con enfoques que solo consideran la composición química. A continuación, se destacan algunos de ellos:

1. **Información estructural:** Al incluir características no composicionales como la simetría cristalina, los parámetros de la celda unitaria y la densidad, se captura información valiosa sobre la estructura del material. Estas características estructurales pueden tener un impacto significativo en las propiedades electrónicas y el bandgap del material.
2. **Propiedades físicas:** Características como el volumen, la densidad y la energía por átomo proporcionan información sobre las propiedades físicas del material. Estas propiedades están estrechamente relacionadas con la estructura electrónica y pueden influir en el valor del bandgap.
3. **Estabilidad termodinámica:** La inclusión de características como la energía de formación y la energía sobre el casco convexo permite evaluar la estabilidad termodinámica del material. Los materiales estables tienden a exhibir propiedades electrónicas y bandgaps más confiables y reproducibles.

4. **Propiedades magnéticas:** Considerar características como el magnetismo y la magnetización total puede ser relevante para ciertos materiales en los que las interacciones magnéticas influyen en la estructura electrónica y el bandgap.
5. **Generalización y transferibilidad:** Al utilizar un conjunto más amplio de características, incluyendo tanto composicionales como no composicionales, se puede desarrollar un modelo de predicción más robusto y generalizable. Esto permite aplicar el modelo a una gama más amplia de materiales, incluso cuando la información composicional es limitada.

A pesar de la importancia de las características composicionales, nuestro enfoque de utilizar tanto características composicionales como no composicionales presenta varias fortalezas. En primer lugar, la composición química por sí sola no siempre captura completamente la complejidad y la diversidad de los materiales. Materiales con la misma composición pueden exhibir diferentes estructuras cristalinas, propiedades físicas y comportamientos electrónicos. Al incluir características no composicionales, podemos capturar estas sutilezas y mejorar la precisión de nuestras predicciones.

Además, la consideración de características no composicionales nos permite investigar relaciones y patrones más allá de la composición química. Debido a que la relación entre las diversas características de un material y su bandgap no es trivial ni directa. Este enfoque puede permitir la exploración de correlaciones entre la estructura cristalina, las propiedades físicas y el bandgap, lo cual nos puede proporcionar una comprensión más profunda de los factores subyacentes que influyen en las propiedades electrónicas de los materiales.

De modo que, la predicción del bandgap utilizando características composicionales y no composicionales obtenidas de la base de datos del Materials Project nos brinda una descripción más completa y detallada de los materiales. Este enfoque nos permite capturar la complejidad y la diversidad de los materiales, mejorar la precisión de nuestras predicciones y obtener una comprensión más profunda de los factores que influyen en las propiedades electrónicas. A pesar de la importancia de la composición química, nuestro enfoque integral nos brinda una herramienta poderosa para el descubrimiento y diseño de nuevos materiales con propiedades electrónicas deseadas.

### 3.2. Pre-procesamiento de datos

En esta sección se abordará el pre-procesamiento de los datos, incluyendo una exploración inicial para comprender su estructura y la importancia de sus características, la limpieza del dataset y la generación de visualizaciones que nos permitan apreciar las relaciones entre las características más relevantes del dataset.

## Exploración inicial de los datos

Al realizar una exploración inicial de nuestro dataset observamos que la mayoría de las variables son numéricas, salvo las siguientes excepciones:

- `formula_pretty` (String) : Debido a ser la formula química del material compuesta por letras y números.
- `crystal_symmetry` (String) : Debido a que hace referencia al tipo de simetría que posee la red cristalina del compuesto, por ejemplo: cubica, hexagonal, etc. La simetría cristalina nos brinda información sobre la forma en la que se ordenan los átomos en el espacio.
- `is_magnetic` (Bool) : Debido a que toma el valor de verdadero si el material es magnético y falso en el caso contrario.
- `is_stable` (Bool) : Debido a que toma el valor de verdadero si el material es estable y falso en el caso contrario.

En la Tabla 3.1 podemos observar en más detalle los tipos de datos para cada una de las variables que componen nuestro dataset. Con la finalidad de realizar un análisis numérico sobre las características convertiremos las características de tipo booleano a tipo entero, tomando el valor 1 para verdadero y 0 para falso. En el caso de las variables tipo string realizaremos una conversión más minuciosa utilizando principios físicos.

Por ejemplo, para la variable `crystal_symmetry` que representa el tipo de simetría que posee la red cristalina del compuesto, le daremos un valor numérico según el grado de simetría de la estructura, ordenando los 7 diferentes tipos, de mayor a menor simetría, o de manera equivalente, de menor a mayor complejidad:

- **Cubic** (1): Mayor simetría, más simple. Todos los lados y ángulos son iguales, formando un cubo perfecto.
- **Tetragonal** (2): Simetría alta, ligeramente más compleja que la cúbica. Similar al cubo, pero con un eje más largo o más corto que los otros dos.
- **Hexagonal** (3): Simetría alta, comparable a la tetragonal pero con un sistema de ejes diferente. Presenta una estructura en forma de prisma hexagonal.
- **Trigonal** (4): Similar a la hexagonal en términos de simetría, pero con diferencias en la disposición atómica. Puede considerarse como una variante de la estructura hexagonal.
- **Orthorhombic** (5): Menor simetría que las anteriores, con lados de diferentes longitudes y ángulos rectos. Forma de prisma rectangular con tres lados desiguales.

Cuadro 3.1: Tipos de datos de las variables

Variables y tipos de dato	
Variable	Tipo de Dato
band_gap	float64
formula_pretty	string
nsites	int64
nelements	int64
volume	float64
density	float64
density_atomic	float64
crystal_symmetry	string
symmetry_number	int64
sides_abc	float64
angles_abc	float64
uncorrected_energy_per_atom	float64
energy_per_atom	float64
formation_energy_per_atom	float64
energy_above_hull	float64
is_stable	bool
equilibrium_reaction_energy_per_atom	float64
efermi	float64
is_magnetic	bool
total_magnetization	float64
num_magnetic_sites	float64

- **Monoclinic** (6): Aún menor simetría, con ángulos no ortogonales. Presenta una forma de paralelepípedo inclinado, con un ángulo diferente de 90 grados.
- **Triclinic** (7): La menor simetría, más compleja en términos de simetría y disposición atómica. No presenta ningún ángulo recto y todos los lados son de diferente longitud.

Para la variable `formula_pretty` que nos proporciona información sobre los tipos de elementos que conforman un compuesto, el principal dato numérico que podemos extraer será su peso molecular total. El peso molecular es la suma de los pesos atómicos de todos los átomos que componen una molécula. Para calcular el peso molecular, es necesario analizar el string de la fórmula química, reconocer sus elementos constituyentes y sumar los pesos de cada una de sus partes.

Para realizar este procedimiento de manera más directa utilizaremos la librería de python llamada `periodictable` junto con las expresiones regulares de reconocimiento de caracteres para la identificación de los elementos contenidos en la formula química del compuesto. De esta manera podemos reemplazar la variable tipo string

formula\_pretty por la variable numérica molecular\_weight.

## Limpieza de los datos

Como se mencionó anteriormente, durante el proceso de extracción de datos de la API de Materials Project, se ha realizado un esfuerzo significativo para obtener un conjunto de datos completo y de calidad. Dado que la API entrega los datos de cada compuesto en formato de documento (tipo JSON), es necesario aplicar un procesamiento adicional para extraer las características relevantes y convertirlas en un formato tabular adecuado para su posterior análisis. Durante este proceso, se han seleccionado principalmente características de tipo numérico y se han descartado aquellas con entradas mayoritariamente nulas, como por ejemplo el campo equilibrium\_reaction\_energy\_per\_atom, para garantizar la calidad de los datos.

Una vez extraídos los datos, se ha llevado a cabo un proceso de limpieza y preparación del dataset. Se han aplicado las funciones `dropna()` y `drop_duplicates()` para eliminar filas con valores nulos y registros duplicados, respectivamente. Se ha optado también por no utilizar ninguna técnica de eliminación de outliers o imputación de valores faltantes (filling) con el fin de mantener la integridad y calidad de los datos.

Como resultado, se ha obtenido un dataset rico en características, con 20 campos completos (sin valores nulos) y un total de 29,222 instancias de distintos compuestos inorgánicos. Este dataset constituye una base sólida para las posteriores etapas del proyecto, ya que contiene información valiosa y detallada sobre las propiedades y características de una amplia gama de materiales.

Después de la respectiva limpieza de los datos observamos que algunas variables de nuestro dataset poseen nombres muy extensos que son difícilmente legibles en un gráfico, por lo cual generaremos una abreviación compuesta por 3 letras (inicio, medio y final) para cada una de las características, obteniendo la lista de abreviaciones mostradas en la Tabla 3.2.

## Análisis de correlaciones y relevancia de las características

Considerando las abreviaciones establecidas podemos continuar con nuestra exploración generando un gráfico de correlaciones, como se muestra en la Figura 3.1. Donde se observa que la variable del band\_gap (bdp) tiene una mayor correlación con las variables efermi (eei) y formation\_energy\_per\_atom (fem).

La energía de Fermi (efermi) es el nivel más alto que pueden ocupar los electrones en un material a una temperatura de cero absoluto (0 Kelvin). Este nivel actúa como una línea divisoria entre la banda de valencia y la banda de conducción.

Cuadro 3.2: **Variables y abreviaciones**

<b>Variables y abreviaciones</b>	
<b>Variable</b>	<b>Abreviación</b>
band_gap	bdp
nsites	nis
nelements	nms
volume	vle
density	dsy
density_atomic	dyc
crystal_symmetry	c_y
symmetry_number	syr
sides_abc	ssc
angles_abc	aec
uncorrected_energy_per_atom	unm
energy_per_atom	epm
formation_energy_per_atom	fem
energy_above_hull	ebl
is_stable	ite
efermi	eei
is_magnetic	igc
total_magnetization	tnn
num_magnetic_sites	nes
molecular_weight	mat

Donde todos los niveles de energía por debajo de la energía de Fermi estarán completamente llenos de electrones, mientras que todos los niveles por encima estarán completamente vacíos, tal como lo vimos en la Figura 1.1. La fuerte correlación negativa entre la energía de Fermi y el bandgap puede entenderse considerando que a medida que la energía de Fermi aumenta, se llenan más estados electrónicos en la banda de valencia, lo que reduce la diferencia de energía entre el estado ocupado más alto y el estado desocupado más bajo. Esta reducción en la diferencia de energía se traduce en un bandgap más estrecho.

Por otro lado, la energía de formación por átomo representa la energía requerida para formar el material a partir de sus constituyentes elementales. Una mayor energía de formación por átomo indica una mayor estabilidad termodinámica y enlaces químicos más fuertes en el material. La fuerte correlación negativa entre el bandgap y la energía de formación por átomo puede entenderse considerando que una energía de formación más negativa indica una mayor estabilidad termodinámica y enlaces químicos más fuertes en el material. Estos enlaces fuertes y la estabilidad estructural resultante conducen a una mayor separación energética entre los orbitales enlazantes y antienlazantes, lo que se traduce en un bandgap más amplio.

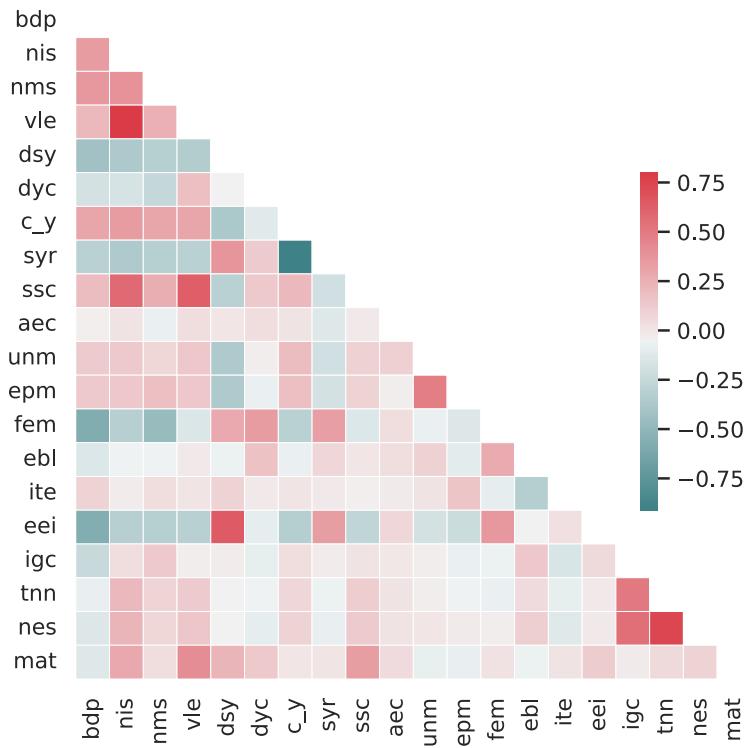


Figura 3.1: Correlaciones entre las variables

Además, una estructura cristalina más estable y ordenada puede resultar en una menor densidad de estados electrónicos cerca del nivel de Fermi, contribuyendo así a un bandgap más grande.

Del mismo modo la Figura 3.1 nos muestra correlaciones significativas entre la variable `num_magnetic_sites` (`nes`) con la variable `total_magnetization` (`tnn`), la variable `symmetry_number` (`syr`) con la variable `crystal_symmetry` (`c_y`), la variable `efermi` (`eei`) con la variable `density` (`dsy`) y la variable `nsites` (`nis`) con la variable `volume` (`vle`).

La fuerte correlación positiva entre el número de sitios magnéticos (`nes`) y la magnetización total (`tnn`) se explica por la relación directa entre la cantidad de átomos magnéticos en un material y su respuesta magnética global.

La fuerte correlación negativa entre el número de simetría (`syr`) y la simetría del cristal (`c_y`) puede entenderse considerando la relación inversa entre la complejidad de la estructura cristalina y el grado de simetría. El número de simetría es una medida cuantitativa del número de operaciones de simetría que dejan invariante la

estructura cristalina, mientras que la simetría del cristal es una descripción cualitativa de la disposición espacial de los átomos en la red. A medida que la estructura cristalina se vuelve más compleja y de menor simetría, el número de operaciones de simetría permitidas disminuye. Por ejemplo, los sistemas cristalinos de alta simetría, como el cúbico y el hexagonal, tienen un mayor número de operaciones de simetría en comparación con los sistemas de menor simetría, como el triclínico. Por lo tanto, la correlación negativa entre el número de simetría y la simetría del cristal refleja la relación fundamental entre la complejidad estructural y las restricciones geométricas impuestas por la simetría cristalina del material.

La fuerte correlación positiva entre la energía de Fermi (eei) y la densidad (dsy) puede explicarse por la relación intrínseca entre la densidad de estados electrónicos y el nivel de Fermi en un material. A medida que aumenta la densidad del material, hay más átomos por unidad de volumen, lo que resulta en una mayor superposición de los orbitales atómicos. Esto conduce a una mayor densidad de estados electrónicos cerca del nivel de Fermi. Como consecuencia, la energía de Fermi aumenta para acomodar los electrones adicionales en los estados disponibles. Por lo tanto, la fuerte correlación positiva entre la energía de Fermi y la densidad se deriva de la relación fundamental entre la estructura electrónica y la densidad del material, donde una mayor densidad conduce a una mayor ocupación de los estados electrónicos y, por lo tanto, a una energía de Fermi más alta.

La fuerte correlación positiva entre el número de sitios atómicos (nis) y el volumen (vle) puede entenderse considerando la relación directa entre la cantidad de átomos y el espacio que ocupan en la estructura cristalina. El número de sitios atómicos se refiere a la cantidad total de posiciones atómicas en la celda unitaria del cristal, mientras que el volumen representa el espacio tridimensional ocupado por la celda unitaria. A medida que aumenta el número de sitios atómicos, se requiere un mayor volumen para acomodar los átomos adicionales en la estructura cristalina. Esto se debe a que cada átomo ocupa un espacio finito y contribuye al tamaño total de la celda unitaria. Además, la presencia de más átomos en la estructura puede conducir a una expansión del volumen debido a las interacciones interatómicas y los efectos de empaquetamiento. Por lo tanto, la fuerte correlación positiva entre el número de sitios atómicos y el volumen se deriva de la relación intrínseca entre la composición atómica y las dimensiones espaciales de la celda unitaria en los materiales cristalinos.

A pesar de que el análisis de las correlaciones entre las variables nos brinda cierta información sobre la relevancia o impacto del resto de características sobre el bandgap, podemos expandir nuestro análisis sobre la importancia de las características utilizando un modelo de regresión basado en Random Forest.

Random Forest es un algoritmo de aprendizaje automático que combina múltiples árboles de decisión para realizar predicciones robustas y precisas. Una de las

ventajas de este algoritmo es su capacidad para evaluar la importancia de las características utilizadas en el modelo.

Para obtener los mejores resultados, se ha realizado una búsqueda aleatoria de hiperparámetros para encontrar la configuración óptima del modelo Random Forest. Los hiperparámetros seleccionados incluyen:

- **bootstrap:** False, indica que no se utiliza el remuestreo bootstrap durante el entrenamiento.
- **max\_depth:** 30, establece la profundidad máxima de cada árbol de decisión.
- **max\_features:** 'sqrt', especifica que se utilizará la raíz cuadrada del número total de características al considerar la división en cada nodo.
- **min\_samples\_leaf:** 1, define el número mínimo de muestras requeridas para estar en un nodo hoja.
- **min\_samples\_split:** 2, indica el número mínimo de muestras requeridas para dividir un nodo interno.
- **n\_estimators:** 700, establece el número de árboles en el bosque aleatorio.

Una vez definidos los mejores parámetros, se ha inicializado el modelo dividiendo los datos de forma aleatoria entre training y test en una proporción de 80/20. Este proceso permite al modelo aprender patrones y relaciones entre las características y el bandgap.

Después de entrenar el modelo, se ha obtenido la importancia de cada característica utilizando el atributo `feature_importances_` del modelo. Este atributo proporciona una puntuación de importancia para cada característica, donde un valor más alto indica una mayor contribución en la predicción del bandgap. De esta manera podemos generar una puntuación de la importancia de las características, tal como se muestra en la Figura 3.2.

De esta manera se puede observar que el gráfico de la importancia de las características nos confirma ciertas conclusiones obtenidas del análisis de correlaciones, dado que justamente las variables de la energía de Fermi `efermi` (`eei`) y la energía de formación por átomo `formation_energy_per_atom` (`fem`) eran las más fuertemente correlacionadas con el bandgap y también dichas variables son identificadas por el modelo Random Forest como las más relevantes (y por mucho).

Observamos también que existen características muy poco relevantes como por ejemplo la información sobre la estabilidad del material `is_stable` (`ite`) ya que posiblemente el aporte de esta característica ya haya sido considerado en la variable `formation_energy_per_atom`, la cual como se mencionó anteriormente se encuentra estrechamente relacionada con la estabilidad del material. Este análisis de la

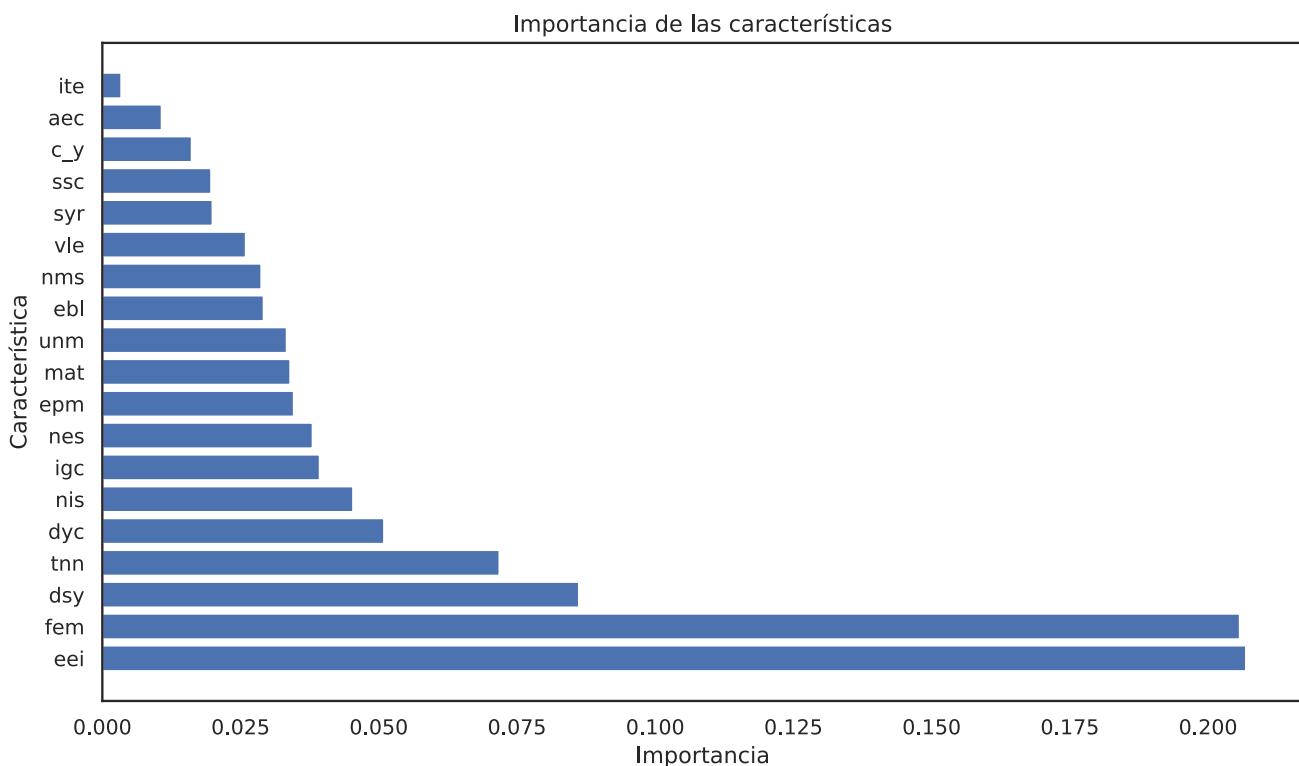


Figura 3.2: Importancia de las características.

importancia de características es fundamental para llevar a cabo la selección y la ingeniería de características (feature engineering), como se discutirá más adelante.

## Visualización de características relevantes

Con la finalidad de realizar una exploración más exhaustiva de nuestro dataset hemos generado una serie de gráficas que nos permitirán apreciar de forma visual relaciones físicas existentes entre ciertas características relevantes de nuestros datos.

Nuestro dataset contiene los 7 tipos de simetrías de redes cristalinas existentes, las cuales representan todas las posibles configuraciones de redes cristalinas en la naturaleza. La simetría de un material depende de varios factores, como la composición química, los enlaces atómicos, las condiciones de formación y las interacciones entre átomos o moléculas. A continuación se presentan las simetrías de varios materiales conocidos:

- **Cúbica:** diamante, oro, plata, cloruro de sodio (sal común)
- **Tetragonal:** rutilo ( $\text{TiO}_2$ ), circonia ( $\text{ZrO}_2$ )

- **Hexagonal:** grafito, cuarzo, hielo
- **Ortorrómbica:** olivino, topacio, azufre
- **Monoclínica:** yeso, mica, augita
- **Triclinica:** plagioclasa, caolinita, turmalina

En la Figura 3.3 podemos observar la diversidad de las simetrías cristalinas presentes en nuestro dataset, donde se aprecia que los materiales con mayor simetría como la red cúbica, tetragonal o hexagonal, existen en mayor abundancia que los materiales con menor simetría como la triclinica o la monoclínica. Esta distribución aparentemente favorable hacia los materiales con mayor simetría en nuestro dataset de compuestos inorgánicos puede ser originada por varias razones:

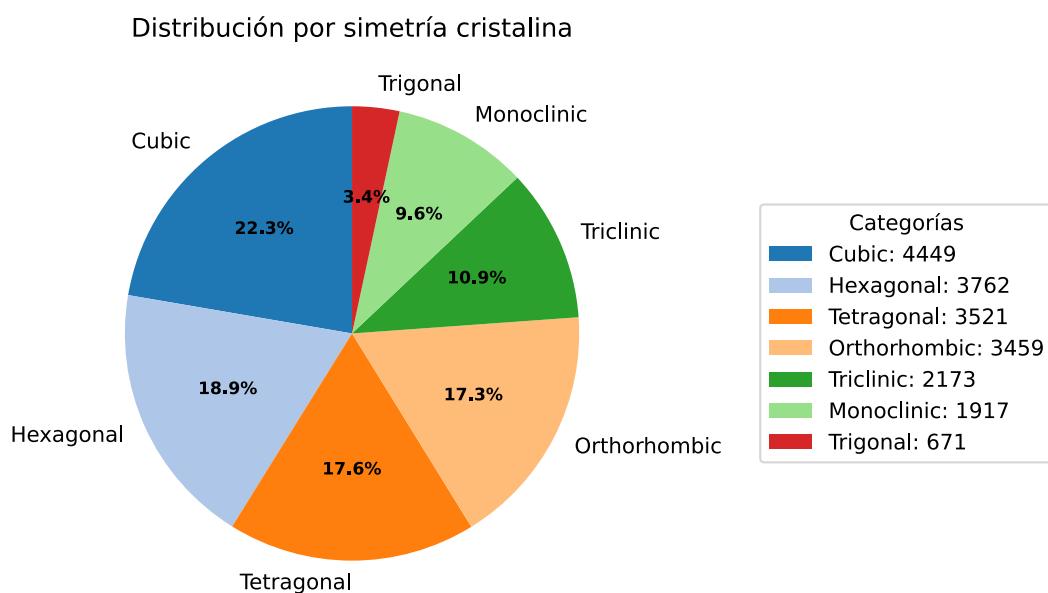


Figura 3.3: Distribución de los materiales del dataset según la simetría cristalina

1. **Estabilidad termodinámica:** Las estructuras cristalinas con mayor simetría suelen ser más estables termodinámicamente debido a su menor energía libre. Esto se debe a que las estructuras simétricas permiten un empaquetamiento más eficiente de los átomos o iones, lo que resulta en interacciones más fuertes y una menor energía potencial.
2. **Enlace químico:** La naturaleza de los enlaces químicos en los compuestos inorgánicos favorece la formación de estructuras simétricas. Los enlaces iónicos y covalentes tienden a formar geometrías regulares y simétricas para minimizar la energía del sistema y maximizar la estabilidad.

3. **Tamaño y carga de los iones:** Los compuestos inorgánicos a menudo están compuestos por iones con tamaños y cargas similares, lo que facilita la formación de estructuras cristalinas simétricas. Los iones de tamaño y carga similar pueden empaquetarse de manera más eficiente en arreglos simétricos.
4. **Condiciones de formación:** Las condiciones bajo las cuales se forman los compuestos inorgánicos, como la temperatura y la presión, pueden favorecer la formación de estructuras cristalinas con mayor simetría. Las altas temperaturas y presiones pueden promover la formación de fases más simétricas y estables.
5. **Simplicidad y frecuencia:** Las estructuras cristalinas con mayor simetría, como las cúbicas y tetragonales, son relativamente simples y se encuentran con mayor frecuencia en la naturaleza. Esto se debe en parte a su estabilidad y a las condiciones favorables para su formación.

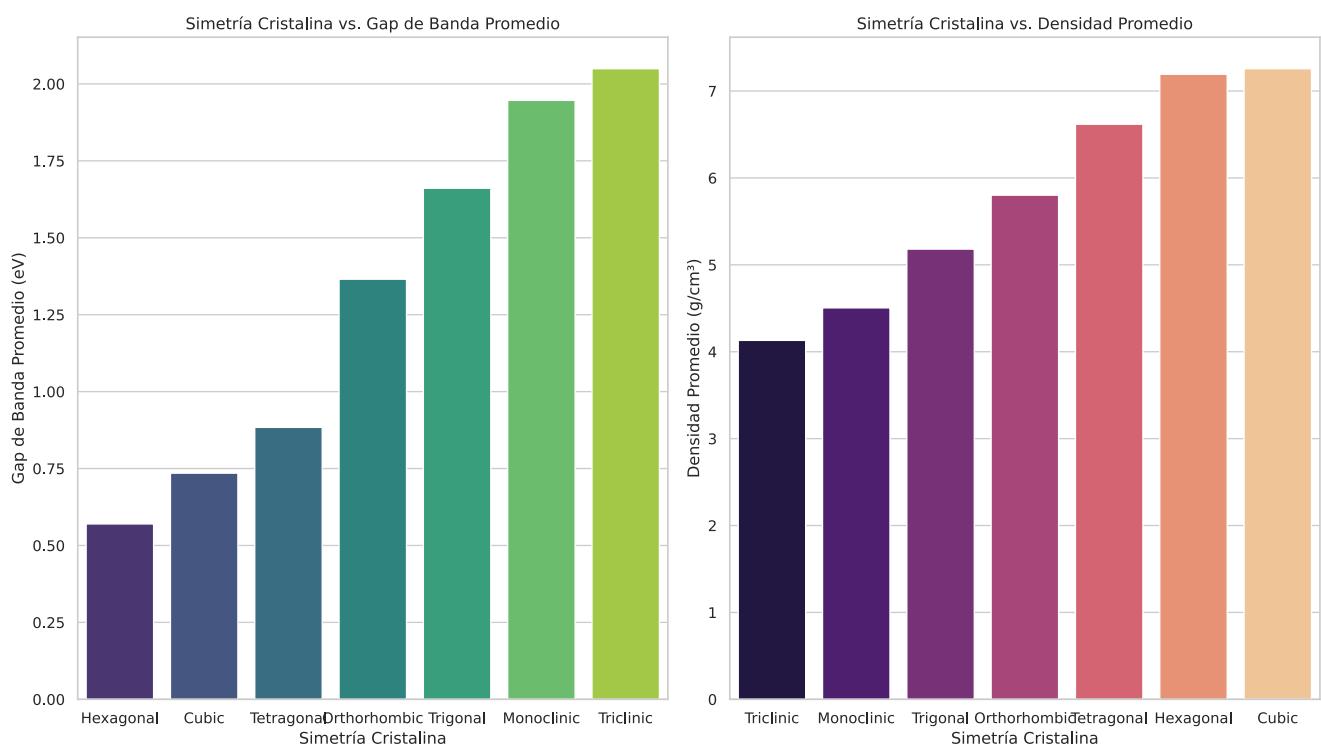


Figura 3.4: Bandgap y densidad de los materiales según la simetría cristalina

Ahora si analizamos el bandgap promedio de los materiales según la simetría cristalina, tal como se observa en la Figura 3.4, podemos apreciar también que materiales con menor simetría generalmente presentan mayor bandgap que materiales más simétricos. Entre las razones por las cuales se puede producir esta relación inversamente proporcional entre el bandgap y la simetría cristalina tenemos las siguientes:

1. **Reducción del solapamiento de orbitales:** En estructuras menos simétricas, el solapamiento de los orbitales atómicos puede ser menos eficiente, lo que resulta en una menor deslocalización de los electrones. Esto puede conducir a una mayor separación entre la banda de valencia y la banda de conducción, aumentando así el bandgap.
2. **Distorsión de la red cristalina:** La menor simetría puede provocar distorsiones en la red cristalina, lo que puede alterar la periodicidad del potencial cristalino. Estas distorsiones pueden modificar la estructura de bandas electrónicas, pudiendo incrementar la separación entre las bandas y, por lo tanto, aumentar el bandgap.
3. **Menor degeneración de estados electrónicos:** En estructuras más simétricas, es común encontrar una mayor degeneración de los estados electrónicos debido a la presencia de elementos de simetría. La degeneración se refiere a la existencia de múltiples estados electrónicos con la misma energía. Por el contrario, en estructuras menos simétricas, la degeneración puede ser menor, lo que puede contribuir a un aumento del bandgap.
4. **Localización de estados electrónicos:** La menor simetría puede favorecer la localización de los estados electrónicos, lo que significa que los electrones están más confinados espacialmente. Esta localización puede dificultar la transición de los electrones desde la banda de valencia a la banda de conducción, resultando en un mayor bandgap.
5. **Efecto de la anisotropía:** Los materiales con menor simetría a menudo exhiben propiedades anisotrópicas, lo que significa que sus propiedades varían según la dirección. Esta anisotropía puede influir en la estructura de bandas electrónicas y, en algunos casos, puede contribuir a un aumento del bandgap.

También podemos apreciar que la estructura hexagonal es la que presenta el menor bandgap promedio, propiedad crucial para determinar la conducción eléctrica de un material, esta característica nos muestra el potencial de la estructura hexagonal, la cual otorga propiedades únicas a los materiales que la poseen, como ocurre en el caso del grafeno, el cual se compone de varias capas de átomos de carbono posicionados en cada vértice de una red hexagonal plana.

Del mismo modo como observamos también en la Figura 3.4 si analizamos la densidad promedio de los materiales según su simetría cristalina podemos apreciar que materiales con mayor simetría generalmente presentan mayor densidad que materiales menos simétricos. Entre las razones por las cuales se puede producir esta relación directamente proporcional entre la densidad y la simetría cristalina tenemos las siguientes:

1. **Empaquetamiento más eficiente:** Las estructuras cristalinas con mayor simetría tienden a permitir un empaquetamiento más eficiente de los átomos o

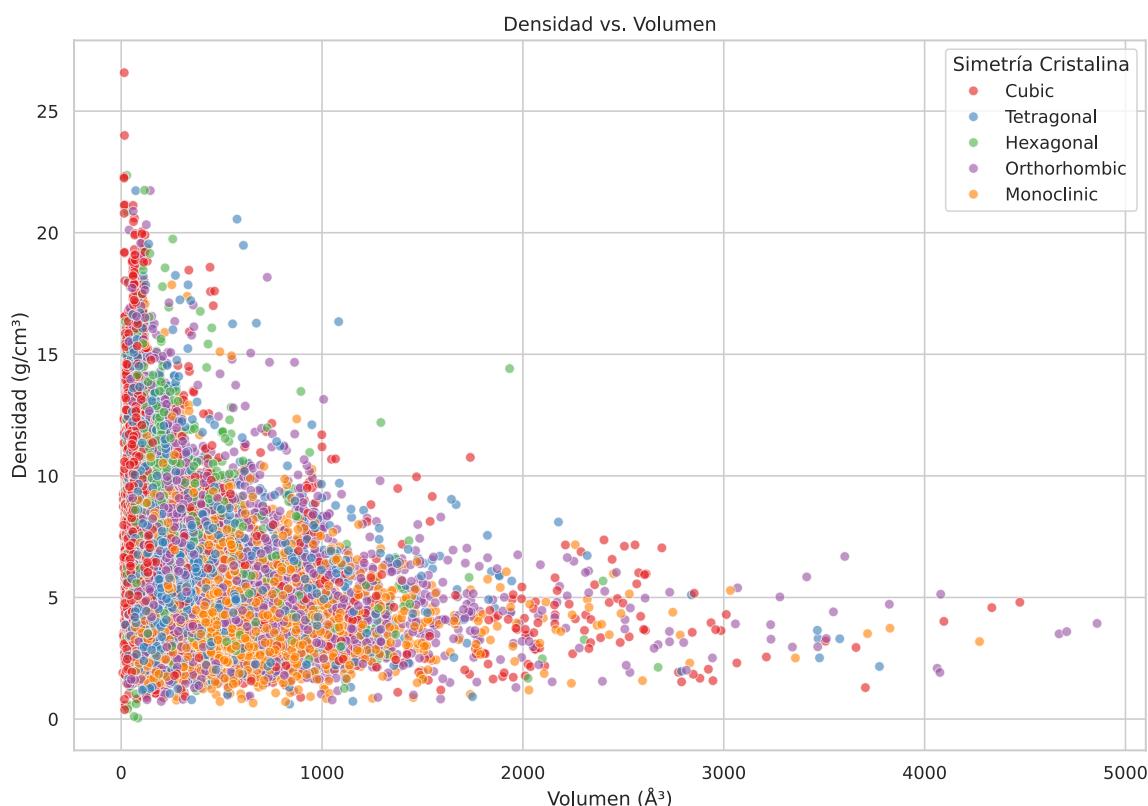


Figura 3.5: Relación entre la densidad, el volumen y la simetría cristalina

moléculas. Esto se debe a que las unidades estructurales pueden ocupar el espacio de manera más óptima, minimizando los espacios vacíos. Un empaquetamiento más eficiente conduce a una mayor densidad del material.

- Distancias interatómicas más cortas:** En estructuras más simétricas, las distancias entre los átomos o moléculas adyacentes tienden a ser más uniformes y potencialmente más cortas. Esto se debe a que la simetría permite una distribución más regular de las unidades estructurales. Distancias interatómicas más cortas implican una mayor proximidad entre los átomos, lo que puede resultar en una mayor densidad.
- Fuerzas de atracción más fuertes:** La simetría cristalina puede influir en las fuerzas de atracción entre los átomos o moléculas. En estructuras más simétricas, las fuerzas de atracción pueden ser más uniformes y potencialmente más fuertes debido a la distribución regular de las cargas electrónicas. Fuerzas de atracción más fuertes pueden conducir a una mayor cohesión y, por lo tanto, a una mayor densidad.
- Menor presencia de defectos:** Los materiales con mayor simetría cristalina tienden a tener una menor presencia de defectos estructurales, como vacan-

tes o dislocaciones. Estos defectos pueden introducir espacios vacíos en la estructura, reduciendo así la densidad. En materiales más simétricos, la probabilidad de formación de defectos puede ser menor, lo que contribuye a una mayor densidad.

5. **Influencia de la composición química:** La simetría cristalina y la densidad también pueden estar influenciadas por la composición química del material. Algunos elementos o combinaciones de elementos pueden favorecer estructuras más simétricas y, al mismo tiempo, contribuir a una mayor densidad debido a sus propiedades intrínsecas, como el tamaño atómico y la masa atómica.

Para analizar a mayor detalle la relación entre la densidad de los materiales con su simetría cristalina podemos generar un gráfico de dispersión entre los principales tipos de simetrías cristalinas en relación con la densidad y el volumen del material, tal como se muestra en la Figura 3.5.

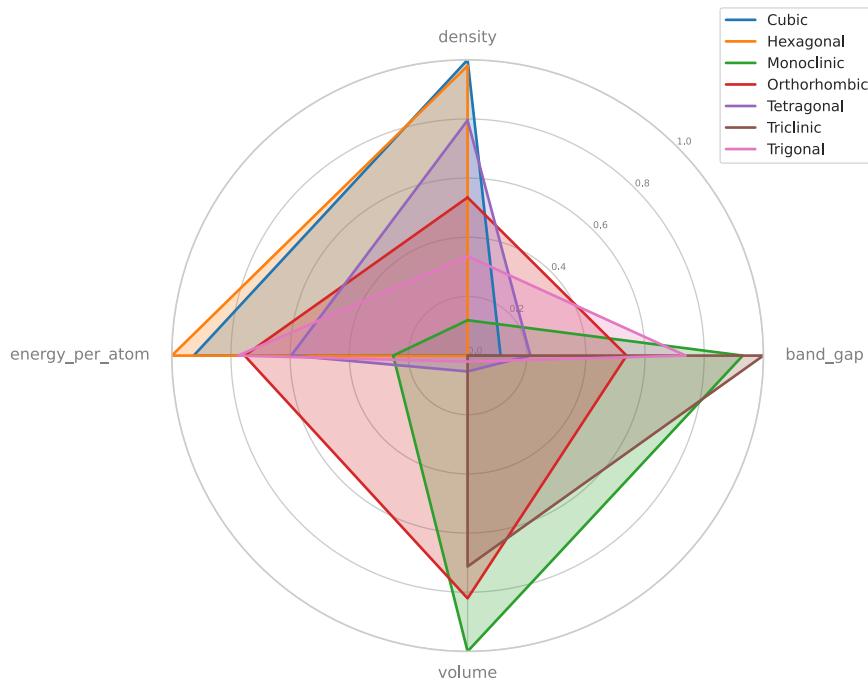
Esta gráfica nos muestra una clara tendencia en la relación entre la simetría cristalina y la densidad de los materiales. Los puntos rojos, que representan los materiales con redes cúbicas, y los puntos verdes, correspondientes a las estructuras hexagonales, se agrupan predominantemente en la región izquierda del gráfico. Este comportamiento se debe a que estas simetrías son las más compactas y eficientes en términos de empaquetamiento atómico, lo que resulta en volúmenes reducidos y, por consiguiente, en densidades más altas.

En contraste podemos observar que los materiales con simetría monoclínica, representados por los puntos naranjas, exhiben una distribución más amplia y dispersa en el gráfico. Esta distribución sugiere que los materiales monoclínicos tienen una mayor variabilidad en sus volúmenes y densidades, lo que se puede atribuir a su menor grado de simetría y empaquetamiento atómico menos eficiente en comparación con las redes cúbicas y hexagonales. Como resultado, los materiales monoclínicos tienden a ocupar volúmenes más grandes y presentan densidades más bajas y dispersas.

Para complementar el análisis de la relación entre las diversas propiedades del material con su tipo de simetría cristalina podemos realizar una comparación de los valores promedio de las propiedades del bandgap, densidad, volumen y energía por átomo para cada simetría cristalina, tal como se observa en la Figura 3.5, la cual nos muestra un claro contraste entre las simetrías cúbica y hexagonal (estructuras más simétricas y compactas), que se encuentran prácticamente en el cuadrante opuesto a las simetrías monoclínica y triclínica (estructuras menos simétricas y compactas).

Este contraste puede explicarse por las diferencias en el empaquetamiento atómico y la eficiencia de llenado del espacio en estas estructuras. Las simetrías cúbica y hexagonal, al ser más compactas y tener un empaquetamiento más eficiente, exhiben una mayor densidad en comparación con las simetrías monoclínica y triclínica.

Comparación de propiedades promedio por simetría cristalina

Figura 3.6: Comparación de propiedades promedio según la simetría cristalina

Esto se debe a que en las estructuras cúbica y hexagonal, los átomos están dispuestos de manera más regular y cercana entre sí, aprovechando mejor el espacio disponible.

Además, se puede observar que las estructuras cúbica y hexagonal tienden a tener un menor volumen promedio en comparación con las estructuras monoclínica y triclínica. Esto es consistente con el hecho de que las simetrías cúbica y hexagonal son más compactas, lo que resulta en un menor volumen ocupado por cada celda unitaria.

En cuanto a la energía por átomo, se observa que las estructuras cúbica y hexagonal tienden a tener valores más bajos en comparación con las estructuras monoclinica y triclinica. Lo cual es consistente con el hecho de que las simetrías cúbica y hexagonal son energéticamente más estables, lo cual puede atribuirse a su mayor eficiencia de empaquetamiento y a las interacciones más favorables entre los átomos en estas configuraciones.

También cabe mencionar que durante el análisis de las relaciones más relevantes entre las distintas propiedades presentes en nuestro dataset, se encontró una visualización particularmente atractiva e informativa al analizar la relación entre la energía por átomo y la densidad de diferentes compuestos, con respecto a su es-

tabilidad, tal como se observa en la Figura 3.7. El gráfico de dispersión presenta una clara distinción entre los compuestos estables (puntos naranjas) y los inestables (puntos azules).

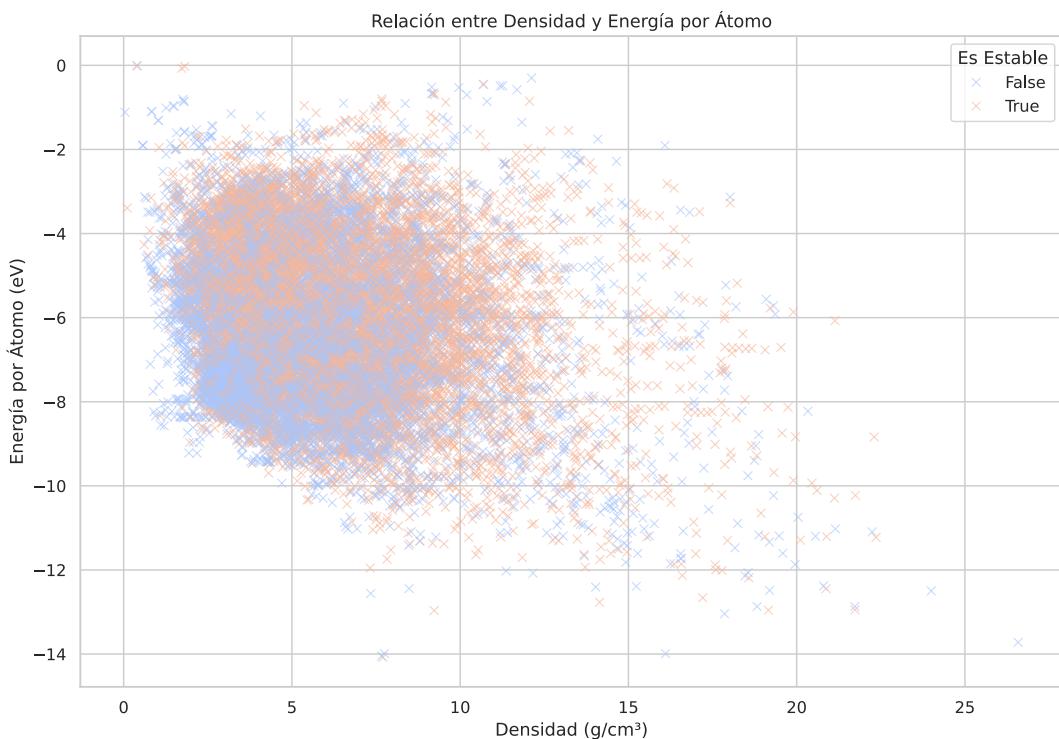


Figura 3.7: Relación entre la densidad y energía por átomo según su estabilidad

Una observación destacada es que la mayoría de los compuestos inestables se concentran en la región inferior izquierda del gráfico, correspondiente a bajas densidades y bajas energías por átomo. Esta distribución no uniforme sugiere que la combinación de estos dos factores tiene una influencia significativa en la estabilidad de los materiales.

Es decir, la combinación de estos dos factores aumenta la probabilidad de inestabilidad en un material. Desde un punto de vista físico, una baja energía por átomo indica enlaces más débiles entre los átomos constituyentes, lo que significa que se requiere menos energía para alterar o romper estos enlaces, conduciendo a una potencial inestabilidad estructural.

De la misma manera, una baja densidad implica un menor empaquetamiento atómico y, por lo tanto, distancias interatómicas mayores. Esto debilita las interacciones entre los átomos, ya que la fuerza de atracción disminuye con la distancia. Como resultado, los materiales con baja densidad son más propensos a sufrir deformacio-

nes, transformaciones de fase o reacciones químicas indeseadas, contribuyendo a su inestabilidad. De esta manera podemos concluir que la combinación de enlaces débiles y un empaquetamiento poco compacto crea condiciones favorables para la inestabilidad de un material.

Finalmente para comprender de mejor manera la distribución de los valores del bandgap para los distintos intervalos del dataset, construiremos un histograma, tal como observa en la Figura 3.8.

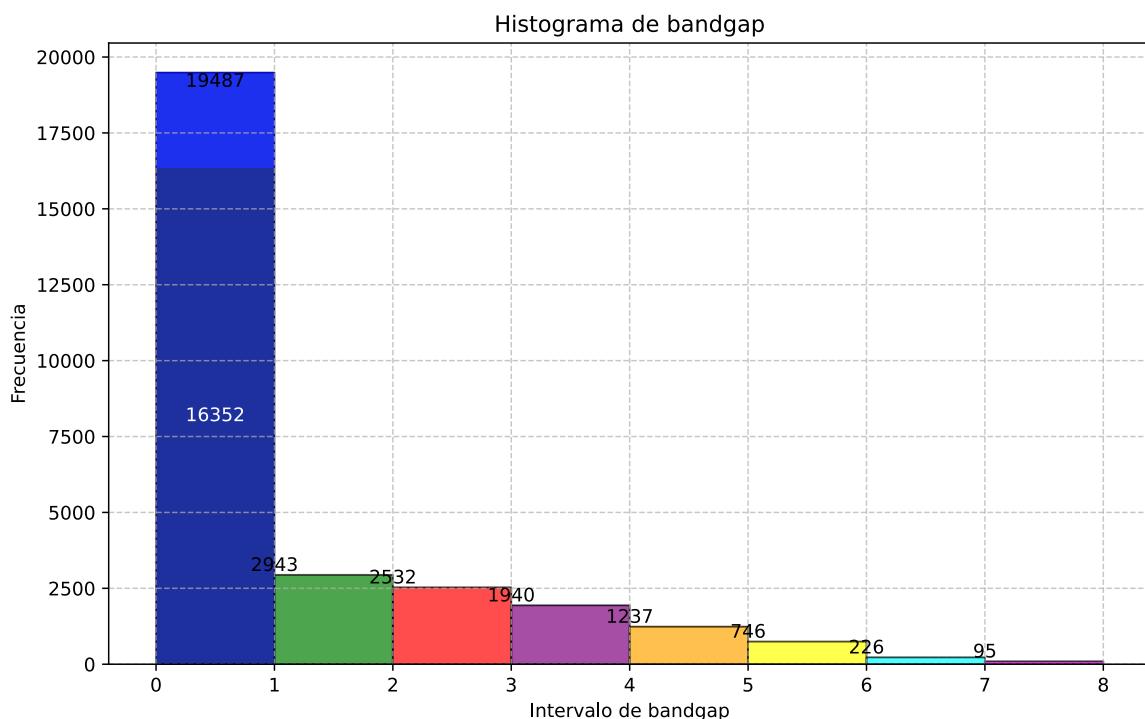


Figura 3.8: Histograma del bandgap

De esta figura podemos observar que de los 29222 compuestos que contiene nuestro dataset, casi la mitad de ellos (16352) poseen bandgap igual a cero y el resto con bandgap distinto de cero. En el intervalo de 0 a 1 eV observamos que existen 3135 compuestos, en el intervalo de 1 a 2 existen 2943 compuestos, encontramos 2532 compuestos para el intervalo de 2 a 3 y así sucesivamente observamos que existe una disminución del número de compuestos entre mayor sea el valor del bandgap. Por lo tanto debido a la naturaleza de nuestro dataset, también esperaríamos que los modelos aplicados decrezcan en precisión mientras mayor sea el valor del bandgap, ya que existirán menos datos para su entrenamiento.

Es importante notar que la distribución presente en nuestro dataset es un reflejo de la naturaleza, donde de manera general los compuestos con menor bandgap suelen poseer mayor estabilidad química que los compuestos con mayor bandgap,

y por lo tanto se encuentran presentes en mayor cantidad, ya que los materiales con bandgaps muy grandes a menudo requieren enlaces químicos muy fuertes y específicos, lo que limita su formación y estabilidad.

### 3.3. Transformación de datos

En esta sección se tratará la transformación de los datos, donde se aplicarán técnicas de aprendizaje no supervisado para realizar clustering y selección de características, además de generar características derivadas a partir de las originales con el objetivo de enriquecer la representación de los compuestos y mejorar el rendimiento de los modelos de aprendizaje automático.

#### Aprendizaje no supervisado para selección de características

Ahora que ya poseemos mayor información acerca de las correlaciones de las características y su importancia con respecto al bandgap, complementaremos nuestro análisis con la implementación de aprendizaje no supervisado para utilizarlo posteriormente en la ingeniería de características.

Para lo cual se ha llevado a cabo un proceso de clustering de características con el propósito de agrupar aquellas que presentan similitudes entre sí. Este enfoque tiene como objetivo identificar patrones subyacentes en los datos y reducir la dimensionalidad del conjunto de características, lo que puede facilitar la interpretación de los resultados y mejorar el rendimiento de los modelos de aprendizaje automático.

Antes de aplicar el clustering, se ha realizado una transformación de los nombres de las características para hacerlos más compactos y manejables. Para ello, se ha utilizado las abreviaturas de tres letras para cada característica. Estas abreviaturas se crean tomando la primera letra, la letra del medio y la última letra de cada nombre, tal como se mencionó en la Tabla 3.2.

Una vez transformados los nombres de las características, se ha realizado la transposición del dataset, lo que significa que las características se convierten en instancias y viceversa. Esto se lleva a cabo con el objetivo de agrupar características similares, utilizando el método de enlace de Ward y la distancia euclídea como medida de similitud. El método de enlace de Ward es un algoritmo jerárquico aglomerativo que busca minimizar la varianza total dentro de los clústeres. Comienza considerando cada característica como un clúster individual y, en cada iteración, fusiona los dos clústeres más similares según la medida de similitud elegida. Este proceso se repite hasta que se alcanza el número deseado de clústeres o hasta que se cumple algún criterio de parada.

Para determinar el número óptimo de clústeres, se ha utilizado el coeficiente de silueta como medida de evaluación. El coeficiente de silueta es una métrica que

cuantifica la calidad de la agrupación, considerando tanto la cohesión de los elementos dentro de un mismo clúster como la separación entre diferentes clústeres. Se ha evaluado el coeficiente de silueta para diferentes valores de corte, es decir, para distintos números de clústeres. El valor de corte que maximiza el coeficiente de silueta se considera como el número óptimo de clústeres. Este enfoque permite obtener una agrupación de características que capture de manera adecuada la estructura subyacente de los datos.

Después de realizar el clustering, se ha aplicado la técnica de Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de las características y facilitar su visualización en un espacio bidimensional. PCA es una técnica de reducción de dimensionalidad que busca proyectar los datos en un nuevo espacio de menor dimensión, manteniendo la mayor cantidad de información posible. En este caso, se han seleccionado las dos primeras componentes principales para representar las características en un plano.

Los resultados del clustering se han visualizado mediante un scatter plot, donde cada punto representa una característica y los colores nos permiten apreciar la formación de clusters. Esta representación gráfica permite apreciar la distribución de las características en el espacio bidimensional y observar la separación entre los diferentes clústeres. Además, se ha incluido una leyenda que muestra las abreviaturas de las características agrupadas en cada clúster, lo que facilita la interpretación de los resultados.

La visualización de los resultados del clustering es una herramienta valiosa para comprender la estructura de los datos y detectar posibles patrones o relaciones entre las características. Puede ayudar a identificar características redundantes o altamente correlacionadas, así como a descubrir subgrupos de características que comparten propiedades similares. Esta información puede ser útil para seleccionar un subconjunto representativo de características y reducir la dimensionalidad del conjunto de datos, lo que a su vez puede mejorar la eficiencia y el rendimiento de los modelos de aprendizaje automático.

De esta manera obtenemos finalmente el clustering de las 20 características de nuestro dataset, tal como se muestra en la Figura 3.9, de donde podemos extraer principalmente 4 clusters o agrupaciones de las características:

1. Primer Cluster: Extremo izquierdo de la figura (puntos verdes y azules claros)
  - tnn: total\_magnetization
  - nes: num\_magnetic\_sites
  - ebl: energy\_above\_hull
  - vle: volume
  - dyc: density\_atomic

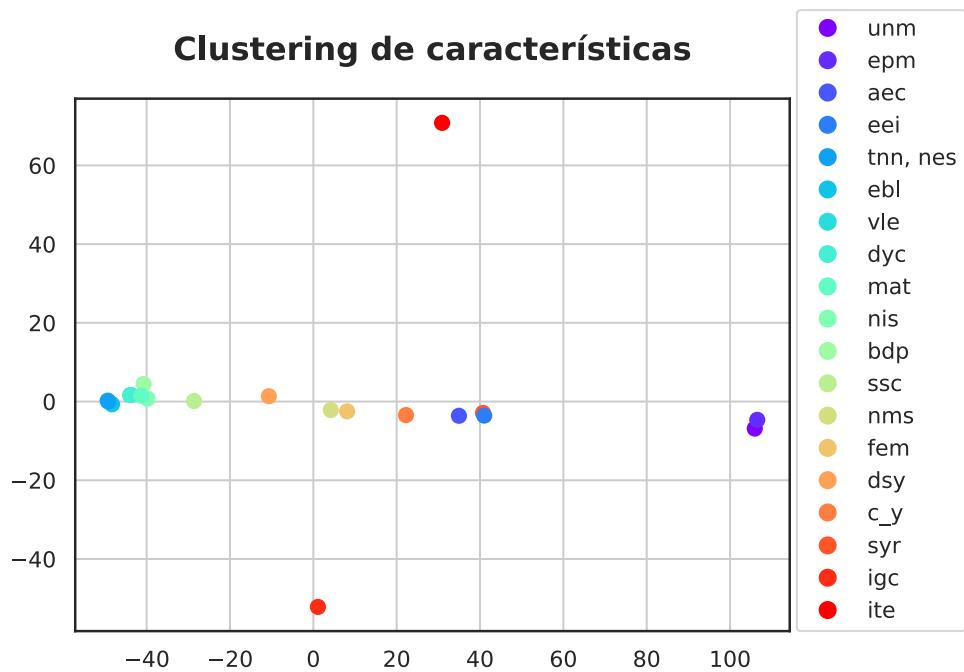


Figura 3.9: Clustering de las características

- mat: molecular\_weight
  - nis: nsites
  - bdp: band\_gap
2. Segundo Cluster: Centro-izquierda (puntos verde oscuro y naranja)
- nms: nelements
  - fem: formation\_energy\_per\_atom
3. Tercer Cluster: Centro-derecha (puntos azules oscuros y rojo claro)
- aec: angles\_abc
  - eei: efermi
  - syr: symmetry\_number
4. Cuarto cluster: Extremo derecho (puntos violeta)
- unm: uncorrected\_energy\_per\_atom
  - epm: energy\_per\_atom

Además observamos que existen características un poco aisladas del resto de clusters como ocurre con:

- dsy: density
- c\_y: crystal\_symmetry
- ssc: sides\_abc

Y un par de características muy separadas del resto, como son:

- igc: is\_magnetic
- ite: is\_stable

Esta información obtenida del clustering de características la podemos combinar con lo aprendido del dataset a partir de sus correlaciones e importancia de sus características. De este análisis podemos realizar una serie de experimentos que nos permita identificar cual es la mejor selección o combinación de características que ayude a mejorar el rendimiento de los modelos de aprendizaje automático implementados.

Entre algunos de los experimentos de feature engineering realizados tenemos:

- Eliminación de la característica de estabilidad (ite) debido a ser la característica con menor importancia (Figura 3.2), además de ser una de las características más alejadas de otros clusters (Figura 3.9).
- Combinación de las características del número de sitios (nis) y el volumen (vle) debido a que poseen una fuerte correlación entre ellas (Figura 3.1), además de ambas pertenecer al primer cluster (Figura 3.9).
- Combinación de las características del número de sitios magnéticos (nes) y la magnetización total (tnn) debido a que poseen una fuerte correlación entre ellas (Figura 3.1), además de ambas pertenecer al primer cluster (Figura 3.9).
- Combinación de las características de la energía Hull (ebl) y la densidad atómica (dyc) debido a que poseen una correlación significativa entre ellas (Figura 3.1), además de ambas pertenecer al primer cluster (Figura 3.9).
- Combinación de las características de energía de formación por átomo (fem) y número de elementos (nms) debido a que poseen una correlación significativa entre ellas (Figura 3.1), además de ambas pertenecer al segundo cluster (Figura 3.9).
- Combinación de las características energía por átomo no corregida (unm) y la energía por átomo (epm) debido a que poseen una correlación significativa entre ellas (Figura 3.1), además de ambas pertenecer al cuarto cluster (Figura 3.9).

Para combinar las características primero aplicamos la técnica de normalización MinMaxScaler para asegurar que todas las características estén en una escala comparable. Luego, creamos las nuevas características combinadas calculando el promedio de las características respectivas.

Curiosamente de los múltiples experimentos llevados a cabo, en los que se exploraron diversos escenarios y combinaciones de características, incluyendo las previamente mencionadas, los resultados obtenidos no mostraron una mejora significativa en el rendimiento de los modelos de predicción. De hecho, el desempeño de los modelos entrenados con el conjunto completo de características disponibles en nuestro dataset fue siempre superior al obtenido mediante las diferentes estrategias de selección y combinación de características implementadas. Este hallazgo sugiere que, para este caso en particular, la utilización de todas las características disponibles puede ser la opción más adecuada para obtener un rendimiento óptimo en la tarea de predicción deseada.

Este fenómeno se puede atribuir, fundamentalmente, a que después de las etapas de extracción y pre-procesamiento se ha conseguido obtener un dataset con un número relativamente reducido de características, pero todas ellas con un alto contenido informativo y valor numérico. Cada una de estas características, en mayor o menor grado, contribuye de manera significativa al aprendizaje de los modelos de predicción, proporcionando información relevante y complementaria. Esta riqueza en la calidad de los datos ha permitido que, incluso con un conjunto limitado de variables, se pueda obtener un rendimiento satisfactorio en las tareas de aprendizaje automático, sin la necesidad de recurrir a técnicas adicionales de selección y combinación de características.

## Generación de características derivadas

Tal como observamos en la anterior sección, el procedimiento de selección y combinación de características utilizado hasta el momento no consiguió una mejora en el rendimiento de los modelos de predicción empleados.

Debido a lo cual se realizaron diversos experimentos adicionales de feature engineering, de los cuales un solo procedimiento se destacó del resto, al conseguir una mejora del rendimiento de los modelos en alrededor de un 10 %.

Este procedimiento consiste en extraer mayor cantidad de información de la fórmula química del compuesto (`formula_pretty`), de manera similar al procedimiento realizado para obtener la característica del peso molecular (`molecular_weight`).

Nuevamente con la ayuda de la biblioteca de python `periodictable` y las expresiones regulares (`re`) se identificaron todos los elementos únicos de todos los compuestos del dataset y se crearon dos nuevas columnas para cada uno de ellos:

una columna llamada <nombre del elemento>\_count que indica la cantidad de átomos de ese elemento en cada compuesto, y otra columna llamada <nombre del elemento>\_mass\_ratio que representa la proporción de masa de ese elemento en cada compuesto, calculada como la masa atómica del elemento por la cantidad de átomos dividido para el peso molecular del compuesto.

De esta manera nuestro dataset pasó de tener 20 columnas a más de 200. Lo cual sin aumentar significativamente el peso del dataset, consiguió ser el único método de ingeniería de características que mejoró de manera considerable el aprendizaje de los modelos empleados.

## 3.4. Minería de datos

En esta sección se abordará la minería de datos, donde se aplicarán diversos algoritmos de aprendizaje supervisado tanto para enfoques de regresión como de clasificación, con el propósito de desarrollar modelos capaces de predecir el valor numérico del bandgap y clasificar los materiales entre metales y no metales, evaluando su desempeño y comparándolos entre sí.

### Aprendizaje supervisado en modelo de regresión

Una vez que hemos obtenido un conjunto de datos limpio, preprocesado y enriquecido con características adicionales, estamos en condiciones de aplicarlo en modelos de aprendizaje supervisado. Para ello, hemos seleccionado los algoritmos más ampliamente utilizados en la literatura científica para la predicción de propiedades de materiales, junto con otros métodos prometedores.

Después de un exhaustivo proceso de evaluación y comparación de diversos algoritmos y combinaciones, a continuación, se detallan los modelos y ensambles de modelos que exhibieron el mejor desempeño en la tarea de predicción del bandgap:

#### 1. Modelos

##### ■ Gradient Boosting Regression (GBR)

El algoritmo de Gradient Boosting Regression (GBR) es un método de ensemble basado en árboles de decisión que combina múltiples modelos débiles para crear un modelo predictivo robusto. El proceso de entrenamiento del GBR se realiza de manera iterativa, donde en cada iteración se ajusta un nuevo árbol de decisión a los residuos (errores) del modelo anterior. Los árboles se construyen de manera secuencial, y cada uno se enfoca en corregir los errores cometidos por los árboles anteriores. La predicción final se obtiene mediante la suma ponderada de las predicciones de todos los árboles. El GBR utiliza un algoritmo de optimización basado en gradientes para minimizar una función de pérdida, como el

error cuadrático medio. Además, emplea técnicas de regularización, como la reducción de la tasa de aprendizaje y la limitación de la profundidad de los árboles, para evitar el sobreajuste y mejorar la capacidad de generalización del modelo.

#### ■ **Random Forest Regression (RFR)**

El algoritmo de Random Forest Regression (RFR) es un método de ensemble basado en árboles de decisión que combina múltiples árboles de regresión para obtener una predicción más precisa y robusta. A diferencia del GBR, donde los árboles se construyen secuencialmente, en el RFR los árboles se construyen de manera independiente y paralela. Cada árbol se entrena utilizando una muestra aleatoria con reemplazo (bootstrap) de los datos de entrenamiento, y en cada división del árbol se considera solo un subconjunto aleatorio de las características. Esta aleatoriedad introduce diversidad en el ensemble y reduce la correlación entre los árboles, lo que mejora la capacidad de generalización del modelo. La predicción final se obtiene promediando las predicciones de todos los árboles. El RFR es robusto frente al ruido y los valores atípicos, y puede manejar características de alta dimensión y relaciones no lineales entre las variables.

#### ■ **Kernel Ridge Regression (KRR)**

El algoritmo de Kernel Ridge Regression (KRR) es una extensión del método de regresión lineal que utiliza técnicas de kernel para modelar relaciones no lineales entre las variables de entrada y la variable objetivo. En KRR, los datos de entrada se transforman a un espacio de características de alta dimensión mediante una función kernel, como el kernel gaussiano o el kernel polinomial. Luego, se aplica una regresión lineal en este espacio de características transformado para encontrar la mejor solución que minimice el error cuadrático medio regularizado. La regularización se introduce a través de un parámetro de complejidad que controla el equilibrio entre el ajuste a los datos de entrenamiento y la complejidad del modelo. KRR es especialmente útil cuando las relaciones entre las variables son no lineales y cuando se trabaja con conjuntos de datos de alta dimensión.

#### ■ **Support Vector Regression (SVR)**

El algoritmo de Support Vector Regression (SVR) es una extensión de las máquinas de vectores de soporte (SVM) para problemas de regresión. El objetivo del SVR es encontrar una función que se ajuste a los datos de entrenamiento con un margen de tolerancia especificado, al tiempo que se mantiene lo más plana posible. El SVR mapea los datos de entrada a un espacio de características de alta dimensión utilizando una función kernel, como la de base radial (RBF). Luego, se busca un hiperplano óptimo en este espacio de características que minimice el error de predicción, teniendo en cuenta una función de pérdida que penaliza los errores fuera

del margen de tolerancia. Los puntos de datos que se encuentran dentro del margen de tolerancia no contribuyen al error, mientras que los puntos fuera del margen son penalizados. El SVR utiliza solo un subconjunto de los datos de entrenamiento, llamados vectores de soporte, para construir el modelo, lo que lo hace eficiente en términos computacionales. En la implementación de este modelo se utilizaron los hiperparámetros optimizados del trabajo de Zhuo et al [9].

#### ■ Adaboost Regressor

A diferencia de Adaboost, que se enfoca en la clasificación, la versión de regresión de Adaboost ajusta los pesos de las instancias basándose en el error de regresión en lugar de la clasificación correcta o incorrecta. En cada iteración, se entrena un modelo de regresión débil (como un árbol de decisión de profundidad limitada) en el conjunto de datos ponderado. Los pesos de las instancias se actualizan de acuerdo con el error de regresión, dando mayor peso a las instancias con mayor error. Este proceso se repite durante varias iteraciones, y en cada iteración se entrena un nuevo modelo de regresión débil en el conjunto de datos ponderado. La predicción final de AdaBoostRegressor se obtiene mediante una combinación ponderada de las predicciones de los modelos de regresión débiles, donde los modelos con menor error reciben un mayor peso en la decisión final. AdaBoostRegressor es capaz de capturar relaciones no lineales entre las características y la variable objetivo, y puede mejorar el rendimiento en comparación con un solo modelo de regresión.

#### ■ K-Nearest Neighbors (KNN)

El algoritmo K-Nearest Neighbors (KNN) es un método de aprendizaje supervisado utilizado tanto para clasificación como para regresión. En KNN, la predicción para una nueva instancia se basa en las k instancias más cercanas en el conjunto de entrenamiento. La cercanía se mide utilizando una métrica de distancia, como la distancia euclídea. Para la clasificación, la clase más frecuente entre los k vecinos más cercanos se asigna como la clase predicha para la nueva instancia. En el caso de la regresión, la predicción se obtiene promediando los valores de los k vecinos más cercanos. El valor de k es un hiperparámetro que se debe ajustar y determina el número de vecinos a considerar. KNN es un algoritmo sencillo pero efectivo, especialmente cuando las clases o los valores de regresión están bien separados en el espacio de características.

#### ■ Least Absolute Shrinkage and Selection Operators (Lasso)

El algoritmo Lasso es un método de regresión lineal regularizada que busca encontrar los coeficientes óptimos de un modelo lineal minimizando la suma de los errores cuadráticos y aplicando una penalización basada en la norma L1 de los coeficientes. La regularización Lasso introduce

un término de penalización que favorece soluciones dispersas, es decir, modelos con muchos coeficientes igual a cero. Esto permite realizar una selección automática de características, ya que los coeficientes no importantes se reducen a cero durante el proceso de optimización. La fuerza de la regularización se controla mediante un hiperparámetro llamado alfa, que determina el equilibrio entre el ajuste a los datos y la complejidad del modelo. Lasso es útil cuando se tiene un gran número de características y se desea identificar las más relevantes para la predicción.

- **Tikhonov Regularization (Ridge)**

El algoritmo Ridge es otro método de regresión lineal regularizada similar a Lasso, pero que utiliza la norma L2 de los coeficientes para la penalización. A diferencia de Lasso, Ridge no realiza una selección automática de características, sino que reduce la magnitud de los coeficientes sin llevarlos a cero. La regularización Ridge introduce un término de penalización que favorece soluciones con coeficientes pequeños y distribuidos de manera más uniforme. Esto ayuda a controlar la complejidad del modelo y evitar el sobreajuste. Al igual que en Lasso, la fuerza de la regularización se controla mediante el hiperparámetro alfa. Ridge es útil cuando se tienen características correlacionadas y se desea mantener todos los coeficientes en el modelo, pero reduciendo su impacto.

- **ElasticNet**

ElasticNet es una combinación de los algoritmos Lasso y Ridge, que busca aprovechar las fortalezas de ambos métodos de regularización. ElasticNet introduce dos términos de penalización: uno basado en la norma L1 (como en Lasso) y otro basado en la norma L2 (como en Ridge). La combinación de estas penalizaciones permite realizar una selección de características y, al mismo tiempo, manejar la multicolinealidad en los datos. El hiperparámetro alfa controla la fuerza total de la regularización, mientras que el hiperparámetro `L1_ratio` determina la proporción de la penalización L1 y L2. Cuando `L1_ratio` es igual a 1, ElasticNet se reduce a Lasso, y cuando `L1_ratio` es igual a 0, se convierte en Ridge. Valores intermedios de `L1_ratio` permiten ajustar el equilibrio entre la dispersión de los coeficientes y la regularización de la magnitud. ElasticNet es útil cuando se tienen características correlacionadas y se desea realizar una selección de características mientras se controla la complejidad del modelo.

## 2. Ensamble de modelos

- **Voting**

El algoritmo de Voting es un método de ensemble que combina las predicciones de múltiples modelos base para obtener una predicción final.

En el Voting promedio, las predicciones numéricas de los modelos base se promedian para obtener la predicción final. El objetivo del Voting es aprovechar la diversidad de los modelos base y reducir el riesgo de seleccionar un modelo individual subóptimo. Al combinar las predicciones de diferentes modelos, el Voting puede mejorar la robustez y la precisión de las predicciones finales.

#### ■ Stacking

El algoritmo de Stacking, también conocido como apilamiento o blending, es otra técnica de ensemble que combina múltiples modelos base para obtener una predicción mejorada. A diferencia del Voting, que combina directamente las predicciones de los modelos base, el Stacking utiliza un modelo adicional llamado metamodelo para aprender cómo combinar de manera óptima las predicciones de los modelos base. El proceso de Stacking consta de dos etapas: en la primera etapa, se entrena los modelos base utilizando los datos de entrenamiento y se obtienen sus predicciones.

Luego, estas predicciones se utilizan como características de entrada para entrenar el metamodelo en la segunda etapa. Es decir, el metamodelo recibe como entrada las predicciones generadas por los modelos base en lugar de los datos originales. Cada modelo base se considera como una columna adicional en el conjunto de datos de entrada del metamodelo. Por ejemplo, si se tienen tres modelos base, el metamodelo recibirá tres columnas correspondientes a las predicciones de cada modelo base. De esta manera, el metamodelo aprende a combinar las predicciones de los modelos base de manera óptima, asignando pesos o importancia a cada una de ellas.

El metamodelo aprende a ponderar y combinar las predicciones de los modelos base de manera óptima. Una vez entrenado, el metamodelo se utiliza para hacer predicciones finales en nuevos datos. El Stacking permite aprovechar las fortalezas de diferentes modelos base y puede mejorar significativamente el rendimiento en comparación con el uso de un solo modelo.

En la Figura 3.10 se puede apreciar el rendimiento de los modelos y ensambles antes mencionados. Donde el algoritmo Gradient Boosting (GBR) y Random Forest (RFR) fueron los que presentaron el mejor comportamiento en la tarea de predicción del bandgap.

En la imagen se puede apreciar que el algoritmo de regresión del Random Forest (color verde) aunque presenta mejor comportamiento para valores bajos del bandgap después su error MAE se disparó para valores más altos en comparación con el

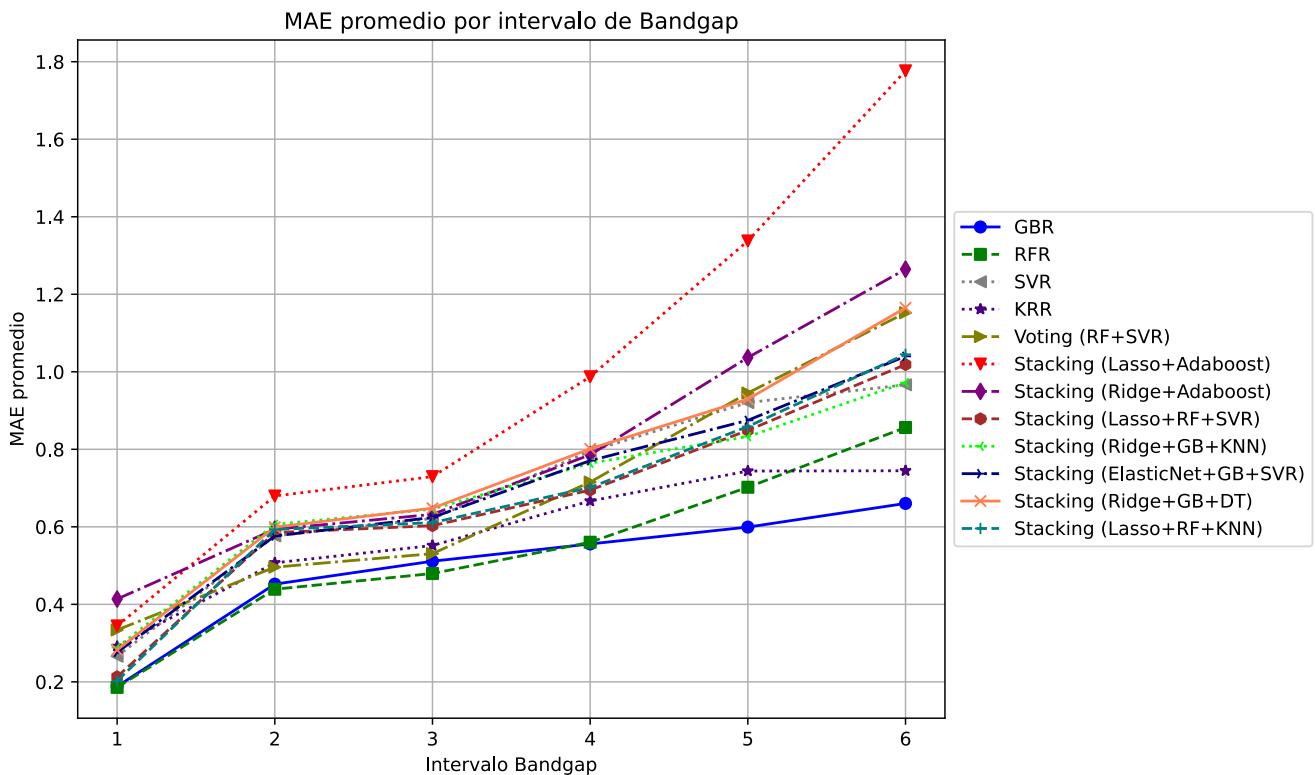


Figura 3.10: Comparación de MAE promedio vs el Bandgap

algoritmo de regresión del Gradient Boosting (color azul). Debido a esto, el algoritmo GBR ha sido el que ha mostrado el mejor comportamiento y el menor MAE total entre todos los modelos y ensambles evaluados.

Además comprobamos que de manera general los modelos aumentan su error, o disminuyen su precisión entre mayor sea el valor del bandgap, tal como lo observamos en la Figura 3.8, esto es debido al limitado número de compuestos contenidos en nuestro dataset con valores altos de bandgap. Este mismo comportamiento también es mencionado por Zhuo et al. [9], quienes reportan una disminución en la precisión de su modelo para compuestos con bandgaps ultra amplios.

Para el entrenamiento y evaluación de estos modelos se realizó una división de los datos en un ratio 80/20 entre training y test, junto con la normalización de las características utilizando MinMaxScaler ajustadas para los datos de entrenamiento.

Para encontrar los mejores hiperparámetros del modelo Gradient Boosting se utilizó el método RandomSearch para una búsqueda más amplia y aleatoria, seguido por el método GridSearch para una búsqueda más focalizada y específica. A continuación se detallan los mejores hiperparámetros encontrados para el modelo GBR:

- `learning_rate`: 0.05. Este hiperparámetro controla la tasa de aprendizaje del modelo, es decir, la magnitud de la contribución de cada árbol en cada iteración del algoritmo. Un valor de 0.05 indica que cada árbol contribuirá con un 5% a la predicción final.
- `max_depth`: 10. Este hiperparámetro define la profundidad máxima de cada árbol de decisión en el modelo. Un valor de 10 significa que cada árbol puede tener hasta 10 niveles de profundidad.
- `max_features`: 'sqrt'. Este hiperparámetro determina el número de características a considerar al buscar la mejor división en cada nodo. El valor 'sqrt' indica que se considerará la raíz cuadrada del número total de características.
- `min_samples_leaf`: 4. Este hiperparámetro especifica el número mínimo de muestras requeridas para estar en un nodo hoja. Un valor de 4 significa que cada nodo hoja debe contener al menos 4 muestras.
- `min_samples_split`: 10. Este hiperparámetro define el número mínimo de muestras requeridas para dividir un nodo interno. Un valor de 10 indica que se necesitan al menos 10 muestras para considerar una división.
- `n_estimators`: 620. Este hiperparámetro establece el número de árboles de decisión a construir en el modelo. Un valor de 620 significa que se crearán 620 árboles.
- `subsample`: 0.99. Este hiperparámetro determina la fracción de muestras a utilizar para cada árbol. Un valor de 0.99 indica que se utilizará el 99% de las muestras para construir cada árbol.

Para visualizar de mejor manera la calidad de las predicciones obtenidas en nuestro modelo GBR compararemos los valores predichos del bandgap con sus valores reales, tal como se muestra en la Figura 3.11. Donde se observa que la gráfica de valores predichos vs. valores reales exhibe una notable concentración de puntos a lo largo de la diagonal del plano, lo cual evidencia una alta concordancia entre las predicciones del modelo y los valores reales del Bandgap. Con ciertas predicciones atípicas principalmente cuando el bandgap es cero, lo cual será solventando en gran medida implementando adicionalmente un modelo de clasificación que separe las clases de compuestos metálicos (bandgap igual a cero) de los no metálicos (bandgap distinto de cero).

De esta manera observamos que la combinación de estos hiperparámetros ha demostrado ser altamente efectiva, permitiendo obtener un modelo de Gradient Boosting con un rendimiento incluso superior a varios estudios previos en la literatura científica relacionada con la predicción del Bandgap. Este notable desempeño será evidenciado y discutido en detalle en las secciones posteriores, donde se realizará una comparativa con los resultados obtenidos por otros investigadores en este campo de estudio.

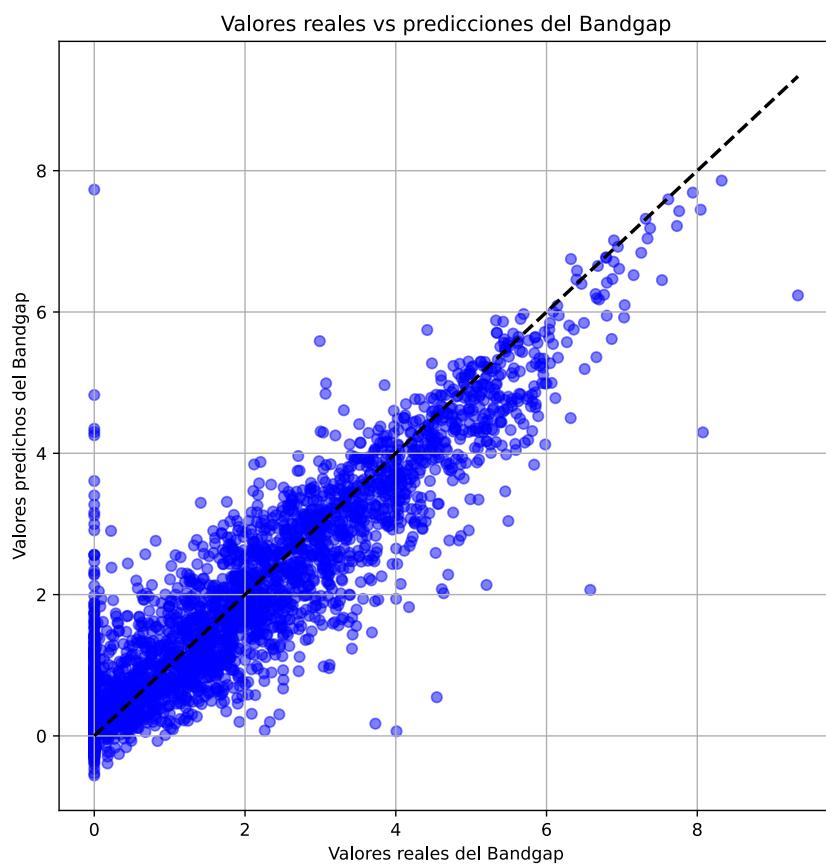


Figura 3.11: Predicciones vs valores reales del bandgap

## Aprendizaje supervisado en modelo de clasificación

Como lo mencionamos en la anterior sección, la Figura 3.11 muestra mayor cantidad de predicciones atípicas para los valores del bandgap igual a cero. Motivo por el cual surge la necesidad de encontrar un mecanismo de separación de los compuestos metálicos (con bandgap igual a cero) y los no metálicos (con bandgap distinto de cero).

Lo cual nos ha llevado a una exploración y evaluación de varios algoritmos de clasificación, entre los cuales constan las versiones de clasificación de los modelos mencionados en la anterior sección. Inclusive la Figura 3.10 nos brinda cierta pista sobre el modelo que mejor comportamiento podrá presentar en la clasificación, como observamos en la imagen el algoritmo de regresión Random Forest demostró un rendimiento superior al resto de modelos para valores bajos del bandgap.

Y dado que nuestro objetivo es establecer una frontera de clasificación en el valor más bajo posible del bandgap para distinguir entre las clases con valores cero y distintos de cero, el algoritmo Random Forest para clasificación (RFC) resultó ser

la opción más adecuada. Este algoritmo logró alcanzar la precisión más alta entre todos los modelos evaluados, lo que lo convirtió en la elección más conveniente para abordar este problema de clasificación específico.

Además es importante notar que tal como lo observamos en la Figura 3.8, nuestro dataset original se encuentra bastante balanceado con casi la mitad de los compuestos con bandgap igual a cero y el resto con valores distintos de cero. Durante el desarrollo del modelo de clasificación, también se experimentó con técnicas de balanceo de clases como DownSampling y UpSampling. Sin embargo, se obtuvieron los mejores resultados al preservar todas las instancias del dataset original sin aplicar remuestreo.

Para el entrenamiento de los modelos de clasificación el primer paso fue convertir la variable continua del bandgap en una variable discreta, es decir se transformó su tipo de dato de decimal a booleano, lo cual se consigue reemplazando los valores distintos de cero del bandgap por la unidad.

El resto del tratamiento de los datos fue similar al realizado para el entrenamiento de los modelos de regresión, es decir, división de los datos entre training y test, normalización de los mismos y aplicación de RandomSearch y GridSearch para la optimización de los hiperparámetros. A continuación se detallan los mejores hiperparámetros encontrados para el modelo RFC:

- n\_estimators: 681
- min\_samples\_split: 4
- min\_samples\_leaf: 1
- max\_features: log2
- max\_depth: 37
- bootstrap: False

Como se observa en la matriz de confusión de la Figura 3.12 el modelo desarrollado demuestra una notable capacidad para distinguir con precisión entre compuestos inorgánicos metálicos y no metálicos. La matriz muestra que el modelo logra clasificar correctamente 3056 compuestos no metálicos (clase 0) y 2363 compuestos metálicos (clase 1). Estos valores en la diagonal principal de la matriz indican un alto número de clasificaciones acertadas para ambas clases.

Por otro lado, se observa que el modelo comete un número relativamente bajo de errores de clasificación. Solo 190 compuestos no metálicos son clasificados erróneamente como metálicos (falsos positivos), mientras que 236 compuestos metálicos son clasificados incorrectamente como no metálicos (falsos negativos). Estos

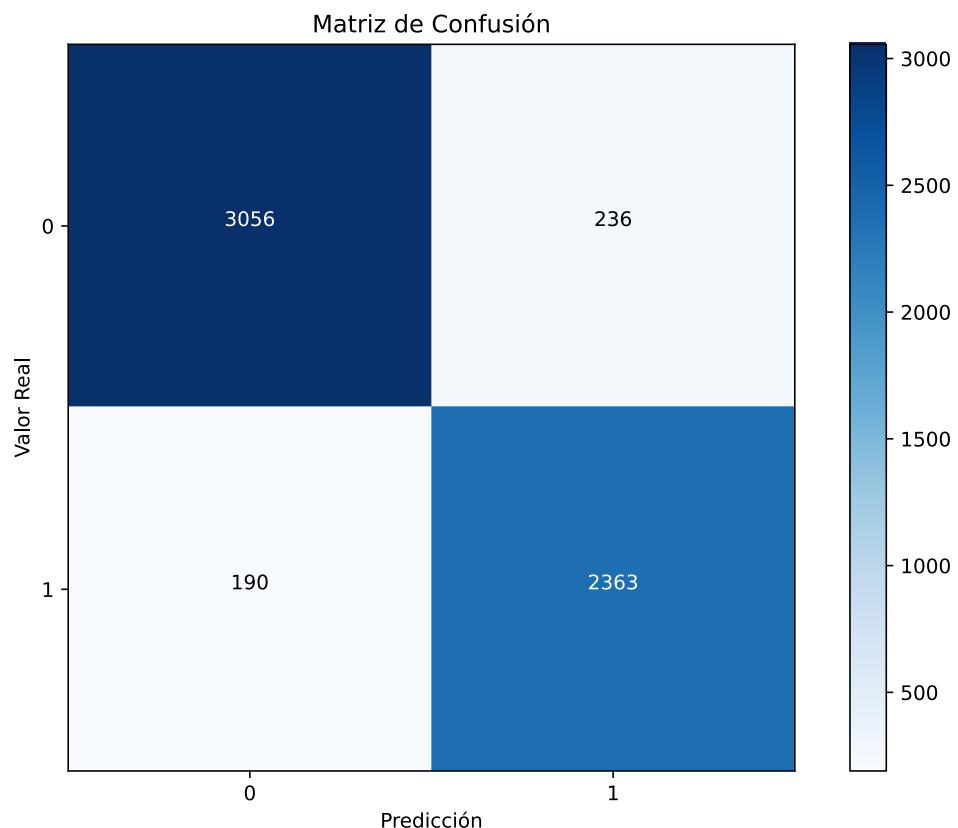


Figura 3.12: [Matriz de confusión del RFC](#)

valores fuera de la diagonal principal son significativamente menores en comparación con las clasificaciones correctas, lo cual resalta la precisión del modelo.

También considerando que los falsos positivos y los falsos negativos además de ser valores bajos son casi iguales en orden de magnitud, significa que el modelo es altamente efectivo para discriminar entre las dos clases. Lo cual se evidencia con un área debajo de la curva ROC cercana al 100 %, tal como se observa en la Figura 3.13.

De esta manera hemos conseguido desarrollar un modelo que presenta un rendimiento incluso superior a varios estudios previos en la literatura científica relacionada con la clasificación de materiales metálicos y no metálicos basados en su bandgap. Lo cual será discutido en detalle en las secciones posteriores, donde se realizará una comparativa con los resultados obtenidos por otros investigadores en este campo de estudio.

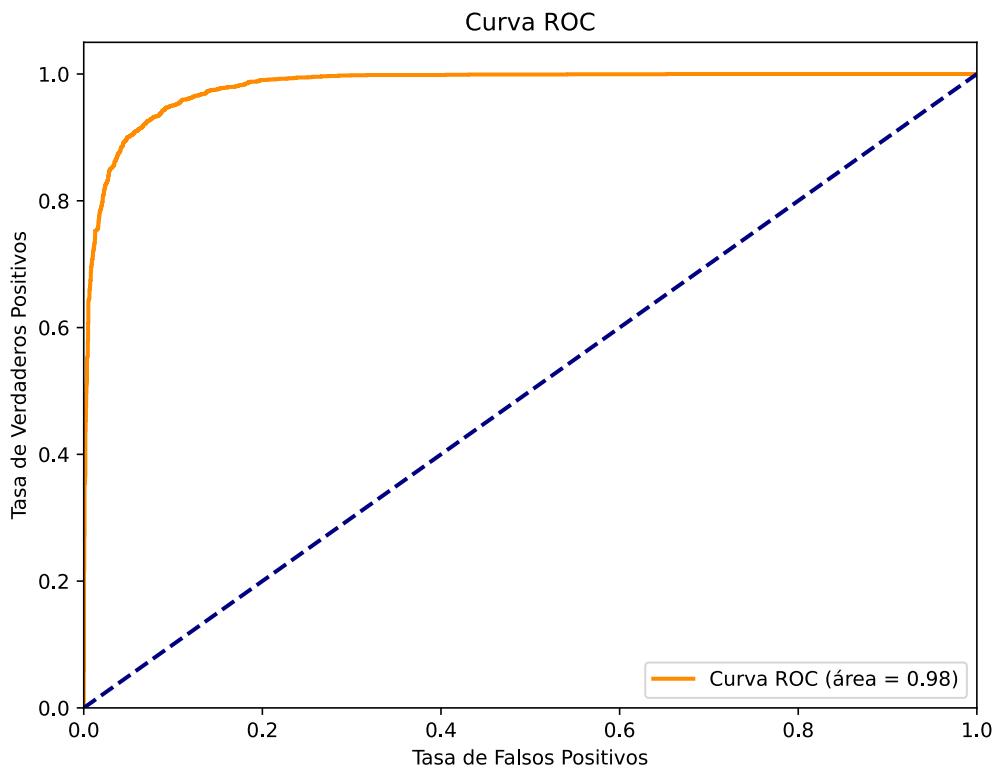


Figura 3.13: [Curva ROC del Modelo RFC](#)

### 3.5. Evaluación de resultados

En esta sección se llevará a cabo la evaluación de los resultados obtenidos, analizando el rendimiento de los modelos desarrollados, comparando su desempeño con trabajos relacionados en la literatura científica referente a la predicción del bandgap, y describiendo el proceso de despliegue de este modelo en una página web alojada en AWS, donde los usuarios podrán ingresar las características de un compuesto y recibir la predicción del bandgap correspondiente vía email.

#### Modelo final

Para aprovechar al máximo las fortalezas de los modelos desarrollados, tanto de nuestro modelo de clasificación RFC como del modelo de regresión GBR, construiremos un modelo final por medio de un ensamble ad-hoc que consistirá en primero pasar los datos a través del modelo de clasificación, seguido por el modelo de regresión.

Considerando que convenientemente el modelo de clasificación funciona ligeramente mejor para predecir la clase 0, que la clase 1, tal como lo vimos en la Figura 3.12. Al ingresar nuevos compuestos al modelo final tendremos dos posibles esce-

narios:

1. El modelo RFC clasifica al nuevo compuesto ingresado en la clase 0 (bandgap igual cero).
2. El modelo RFC clasifica al nuevo compuesto ingresado en la clase 1 (bandgap distinto de cero), por lo que se aplica secuencialmente el modelo GBR para obtener una predicción concreta del bandgap.

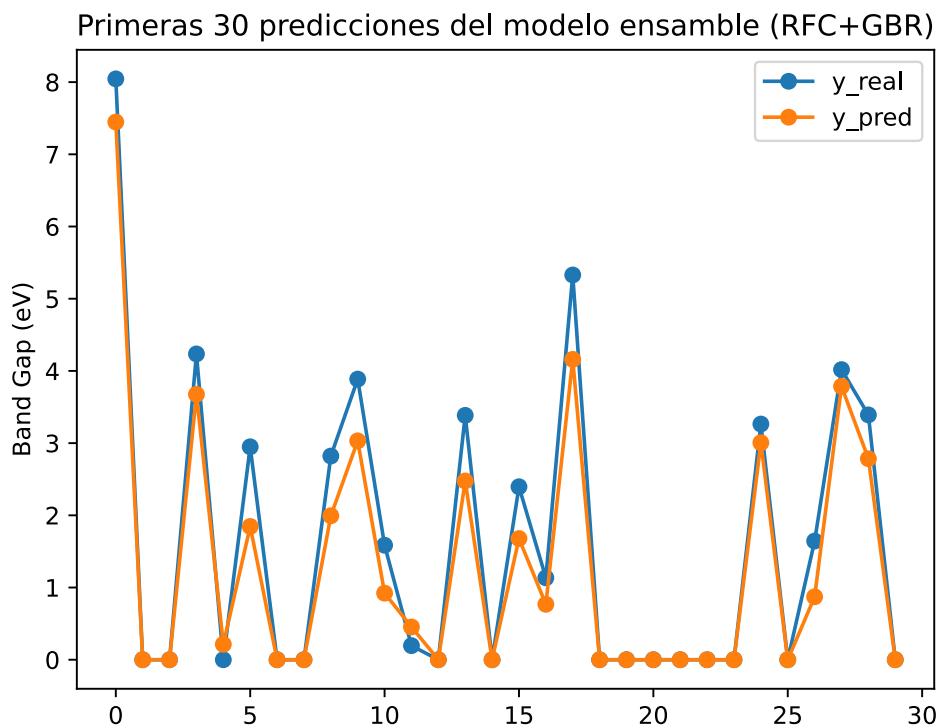


Figura 3.14: Comparación de predicciones y valores reales del modelo final

De esta manera obtenemos un modelo final que consigue un coeficiente de determinación  $R^2$  cercano a 0.9 lo cual significa que aproximadamente el 90 % de la variación en la variable dependiente puede ser explicada por las variables independientes del modelo.

Además este modelo final del ensamble (RFC+GBR) presenta un error MAE de 0.25 eV, lo cual es relativamente bajo considerando que nuestro dataset posee valores de bandgap de hasta 9.33 eV y un valor medio de 2.38 eV. En la Figura 3.14 se puede apreciar la alta concordancia entre las predicciones realizadas y los valores reales en el conjunto de testing.

## Comparativa de resultados con otros trabajos relacionados

Para la tarea de regresión el presente TFM se basó en un dataset de 29222 compuestos inorgánicos y empleó el algoritmo Gradient Boosting Regression (GBR) junto con técnicas de optimización de hiperparámetros, como RandomSearch y Grid-Search. Este enfoque resultó en un MAE de 0.30 eV y un RMSE de 0.52 eV, lo cual representa un desempeño competitivo en relación al estado del arte de la literatura científica mencionada en la sección 2.2, tal como lo observamos en la Tabla 3.3.

Cuadro 3.3: Comparación de resultados en tarea de regresión

Comparación de Resultados en Modelo de Regresión					
Trabajo	MAE	RMSE	R-squared	Algoritmo utilizado	Dataset
Rajan et al. (2018)	0.11 eV	0.14 eV	0.83	GPR (Matérn 5/2)	23 870
Weston et al. (2018)	Sin dato	0.28 eV	0.95	SVR (RBF)	184
Zhuo et al. (2018)	Sin dato	0.45 eV	0.90	SVR (RBF)	3896
TFM	0.302 eV	0.52 eV	0.89	GBR	29222
Zeng et al. (2019)	0.307 eV	Sin dato	0.94	Red Neuronal ATCNN	3896
Kauwe et al. (2020)	Sin dato	0.54 eV	0.883	SVR+RFR+GBR+ANN	47051
Chen et al. (2021)	0.37 eV	Sin dato	Sin dato	Red Neuronal MFGN	52348
Olsthoorn et al. (2019)	0.38 eV	Sin dato	Sin dato	KRR(SOAP)+SchNet	12500
Pilania et al. (2017)	Sin dato	Sin dato	Sin dato	GPR(CoK)	640

A continuación se realizará una comparación con estos trabajos relacionados en la tarea de regresión:

- En el trabajo de Rajan et al. (2018) [5] reportaron el MAE más bajo encontrado hasta la fecha en la literatura científica, con un valor de 0.11 eV. Este trabajo abordó la predicción del bandgap de los materiales bidimensionales MXenes mediante el modelo de Proceso Gaussiano (GPR), con un RMSE de 0.14 eV y un  $R^2$  de 0.83 en su dataset de 23,870 compuestos. Donde se implementó Lasso (Least Absolute Shrinkage and Selection Operator) para reducir las características de 47 a 15. Luego extendieron a millón de características compuestas no lineales y aplicaron Lasso nuevamente para reducir a 8 características finales. Para la construcción del regresor gaussiano GPR se utilizó el kernel Matérn 5/2 con la optimización de sus dos hiperparámetros: la longitud de escala que controla la suavidad de la función en el espacio de entrada y la varianza de las muestras aleatorias extraídas del proceso gaussiano. Dicha optimización se realizó por medio de la maximización de la verosimilitud.
- En el trabajo de Weston et al. (2018) [7] lograron predecir la magnitud del bandgap fundamental en compuestos de kesterita I2-II-IV-V4 con un RMSE de 0.283 eV utilizando regresión de vectores de soporte (SVR) con un kernel de sesgo radial (RBF). Donde se realizó una cuidadosa ingeniería de características, utilizando solo tres descriptores por elemento: la electronegatividad, el

radio iónico y la fila en la tabla periódica. Además, los hiperparámetros de la SVR (parámetro de costo C y coeficiente gamma del kernel) fueron optimizados mediante búsqueda en cuadrícula (grid search) y validación cruzada de 10 pliegues.

- En el trabajo de Zhuo et al.(2018) [9] desarrollaron un modelo de aprendizaje automático basado en la regresión de vector de soporte (SVR) con kernel RBF para predecir el bandgap de sólidos inorgánicos. El modelo alcanzó una notable precisión, con un error cuadrático medio (RMSE) de 0.45 eV y un coeficiente de determinación ( $R^2$ ) de 0.90. Esto se logró gracias a un conjunto de entrenamiento compuesto exclusivamente por bandgaps experimentales de 3896 compuestos y al uso de descriptores basados únicamente en la composición. En este trabajo evitaron emplear bandgaps calculados por DFT, evitando así la subestimación sistemática inherente a los funcionales estándar. Además, aplicaron una selección de características mediante un algoritmo genético con el método de mínimos cuadrados parciales, aunque finalmente incluyeron todas las características. La optimización de hiperparámetros se realizó mediante validación cruzada de 10 pliegues
- En el trabajo de Zeng et al.(2019) [8] propusieron un enfoque novedoso llamado Redes Neuronales Convolucionales sobre Tablas Atómicas (ATCNN) para predecir propiedades de materiales directamente a partir de la composición. La arquitectura de la red consiste en 2 capas convolucionales, 1 max pooling y 2 capas densas. Para la predicción del bandgap, su modelo alcanzó un error absoluto medio (MAE) de 0.307 eV, superando a los cálculos DFT estándar (MAE  $\approx$  0,6 eV), aunque ligeramente superior al MAE obtenido en el presente TFM. La clave de este trabajo fue tratar los compuestos como tablas atómicas y utilizar redes neuronales convolucionales para aprender automáticamente las características relevantes a partir de la composición, sin requerir un conjunto de descriptores predefinidos. Los hiperparámetros de la red (kernels por capa convolucional, tamaño de kernel, número de capas densas y neuronas por capa convolucional) también fueron optimizados mediante búsqueda en cuadrícula y validación cruzada.
- En el trabajo de Kauwe et al. (2020) [2] obtuvieron el mejor rendimiento en la predicción del bandgap al construir un ensamble compuesto por 4 modelos: regresión de vectores de soporte (SVR), regresión de bosques aleatorios (RFR), regresión de Gradient Boosting (GBR) entrenados con datos experimentales, y una red neuronal artificial (ANN) entrenada con datos de bandgap calculados mediante DFT. Este ensamble de modelos consiguió un error RMSE de 0.54 eV, ligeramente superior al valor obtenido en el presente TFM. La construcción del ensamble consta de dos niveles. En el primer nivel, se entranan cuatro modelos base diferentes (SVR, RFR, GBR en datos experimentales y una red neuronal en datos DFT) para generar predicciones individuales del bandgap. Luego, en el segundo nivel, se utiliza un meta-aprendiz que toma las

predicciones de los cuatro modelos del primer nivel como características de entrada y aprende a combinarlas óptimamente para producir la predicción final de bandgap experimental. La arquitectura de la red neuronal consta de 3 capas densas con ReLU y dropout. Para la optimización de hiperparámetros de la red neuronal (cantidad de nodos y tasa de dropout) empleó un enfoque combinado de algoritmo genético y pruebas manuales (trial and error) para obtener los resultados óptimos.

- En el trabajo de Chen et al. (2021) [1] se propone un enfoque con redes neuronales de multi-fidelidad basadas en grafos (multi-fidelity graph networks o red MDGN) para predecir propiedades de materiales, tales como el bandgap, utilizando datos de diferentes niveles de cálculo y experimentales. La clave fue incorporar una gran cantidad de datos de baja fidelidad (bandgaps calculados con el funcional PBE) junto con datos de alta fidelidad (funcionales híbridos o experimentos) en un marco de aprendizaje profundo basado en grafos. Esto permitió aprender representaciones latentes estructurales efectivas que condujeron a mejoras en las predicciones de alta fidelidad, consiguiendo así un MAE de 0.37 eV.
- En el trabajo de Olsthoorn et al. (2019) [3] se evalúan dos modelos de aprendizaje automático, regresión de kernel ridge (KRR) con el kernel SOAP (Smooth Overlap of Atomic Positions) y la red neuronal profunda SchNet, para predecir el bandgap en cristales orgánicos complejos. Utilizaron un nuevo conjunto de datos, OMDB-GAP1, que contiene 12,500 estructuras con un promedio de 82 átomos por celda unidad, lo que representa un desafío significativo en comparación con los conjuntos de datos existentes. Después de optimizar los hiperparámetros, el modelo SOAP alcanzó un error absoluto medio (MAE) de 0.430 eV, mientras que SchNet logró un MAE ligeramente mejor de 0.415 eV. Combinando las predicciones de ambos modelos mediante un ensamble por promediado, lograron reducir aún más el MAE a 0.388 eV.
- Finalmente, en el trabajo de Pilania et al. (2017) [4] se propone un enfoque interesante de aprendizaje automático multi-fidelidad (multi-fidelity machine learning) para predecir el bandgap de materiales sólidos. La clave fue combinar cálculos cuánticos de diferentes niveles de fidelidad utilizando un modelo de regresión gaussiana (GPR) con un kernel co-kriging (Cok). Específicamente, utilizaron cálculos de bandgap de bajo costo con el funcional PBE (baja fidelidad) junto con un subconjunto de cálculos más precisos y costosos con el funcional HSE06 (alta fidelidad). Se utilizó un dataset de 640 compuestos de haluro de perovskita, 599 compuestos de baja fidelidad calculados con PBE y 250 de alta fidelidad calculados con HSE06. El marco de co-kriging permite aprovechar la información de los datos de baja fidelidad para realizar predicciones precisas a nivel de alta fidelidad. Optimizaron los hiperparámetros del modelo co-kriging, como las longitudes de escala del kernel gaussiano, mediante estimación de máxima verosimilitud (MLE). Este estudio muestra que el

rendimiento del modelo depende del ratio de los datos de baja y alta fidelidad, pero no incluye ninguna métrica específica con la cual se pueda comparar los resultados de nuestro TFM.

Por otro lado para la tarea de clasificación el presente TFM utilizó el mismo dataset de 29222 compuestos inorgánicos y empleó el algoritmo de clasificación de Random Forest (RFC) junto con técnicas de optimización de hiperparámetros, como RandomSearch y GridSearch. Este enfoque resultó en una precisión superior al 92 % y con un área bajo la curva ROC del 98 %, lo cual representa un desempeño competitivo en relación al estado del arte de la literatura científica mencionada en la sección 2.2, tal como lo observamos en la Tabla 3.4.

Cuadro 3.4: **Comparación de resultados en tarea de clasificación**

Comparación de Resultados en Modelo de Clasificación				
Trabajo	Precisión	ROC AUC	Algoritmo utilizado	Dataset
Rajan et al. (2018)	0.94	0.99	Bagging	7200
TFM	0.927	0.98	RFC	29222
Zhuo et al. (2018)	0.92	0.97	SVC(RBF)	4916
Zeng et al. (2019)	0.91	0.97	Red Neuronal ATCNN	4916
Weston et al. (2018)	0.89	Sin dato	Logistic Regression	184

A continuación se realizará una comparación con estos trabajos relacionados en la tarea de clasificación:

- En el trabajo de Rajan et al. (2018) [5] desarrollaron un modelo de clasificación para separar los MXenes semiconductores de los metálicos con una precisión del 94 %. Para construir este modelo, utilizaron un conjunto de datos de 7200 MXenes seleccionados aleatoriamente, que fueron optimizados y clasificados mediante cálculos basados en la teoría del funcional de la densidad (DFT). Este modelo utilizó árboles de decisión bootstrap-agregados (bagged). Bagging es un método de ensamble que mejora la precisión de la clasificación al combinar las predicciones de varios subclásificadores.
- En el trabajo de Zhuo et al. (2018) [9] reportaron una precisión de 0.92 usando RFC, ligeramente inferior a la precisión presentada en nuestro TFM. En este estudio de clasificación, emplearon Support Vector Classification (SVC) con un kernel de función de base radial (RBF). Haciendo uso de un dataset con 2458 compuestos metálicos y 2458 no metálicos. Utilizaron un esquema de validación cruzada de 10 iteraciones y optimizaron los hiperparámetros clave del SVC (la constante de costo y el parámetro libre).
- En el trabajo de Zeng et al. (2019) [8] utilizaron también la red neuronal convolucional de tablas atómicas (ATCNN) para la tarea de clasificación. Con la diferencia que ahora esta red fue construida con una sola capa convolucional,

una de max pooling y 2 densas. También se utilizó la misma cantidad de datos que en el modelo SVC del estudio de Zhuo et al., el cual contenía 2458 aislantes y 2458 metales únicos. Alcanzando así una precisión de 0.91, aunque bastante alta, levemente inferior a la alcanzada en nuestro TFM.

- En el trabajo de Weston et al. (2018) [7] desarrollaron un clasificador binario utilizando regresión logística. Inicialmente, el modelo se entrenó usando un espacio de características simple de 12 dimensiones, logrando una precisión del 73 %. Para mejorar las predicciones, se realizó ingeniería de características, creando combinaciones polinómicas de las características originales. Usando combinaciones polinómicas de segundo orden, la precisión del clasificador aumentó al 83 %. Sin embargo, el uso de características polinómicas de tercer orden condujo a una reducción en la precisión al 81 %, lo que sugiere un sobreajuste. Para abordar el problema del sobreajuste y mantener la ventaja de conservar algunos términos de orden superior, se utilizó el método de selección de características con extracción recursiva. De esta manera, el espacio de características polinómicas de alta dimensión se reduce a un pequeño subconjunto de características que tienen el mayor peso en la determinación del resultado. Utilizando la extracción de características y ajustando el clasificador con características polinómicas de tercer orden, la puntuación de precisión aumentó al 89 %, con un número óptimo de 30 características.

A pesar de que hemos considerado los trabajos relacionados más relevantes y recientes en la literatura científica relacionada con la predicción del bandgap por medio de técnicas de Machine learning. Es importante tener en cuenta que la comparación directa de los modelos puede ser limitada debido a las diferencias en los conjuntos de datos y los enfoques específicos utilizados en cada estudio.

Considerando esto, de la revisión anterior observamos que existen tres trabajos relacionados que han mostrado mejores resultados que los obtenidos en el TFM, sin embargo, estos trabajos tienen la particularidad de ser útiles solo para un tipo muy específico de materiales o haber sido entrenado con un conjunto de datos muy pequeño en comparación con el utilizado en nuestro TFM, lo cual podría ser un limitante en la verdadera capacidad de generalización de estos modelos.

Por otro lado, nuestro trabajo exhibe mejores resultados a los alcanzados por 4 estudios recientes en cuanto a la tarea de regresión y también supera a 3 trabajos destacados en la tarea de clasificación, demostrando así el notable desempeño de los modelos desarrollados en el presente TFM.

## Despliegue del modelo final con Amazon Web Services (AWS)

Para el despliegue de nuestro modelo final creamos una solución alojada en AWS donde el usuario recibirá la predicción del bandgap un par de minutos después de haber ingresado los datos del nuevo compuesto a predecir, junto con el correo

electrónico donde desea recibir la información, tal como se observa en la 3.15. A continuación, se detallan los pasos seguidos para el despliegue del modelo final.

← → G bandgap-prediction.s3.amazonaws.com/index.html

### Predictor del Bandgap

Formula Química: KVH4O7

Numero de Sitios: 26

Numero de Elementos: 4

Volumen: 334.6445738712368

Densidad: 2.045052741615885

Densidad Atómica: 12.870945148893725

Simetría Cristalina: Triclinic

Numero de Simetría: 1

Lados ABC: 6.573337181775936

Ángulos ABC: 106.61764612415584

Energía no corregida por atomo: -5.520940623846154

Energía por atomo: 6.021632931538462

Energía de formación por atomo: -1.524572346519231

Ehull: 0.0

Es Estable:

Efermi: 1.81443274

Es Magnético:

Magnetización Total: 0.0006752

Numero de Sitios Magnéticos: 0.0

Correo Electrónico: rodraxsan@hotmail.com

En unos minutos recibirá la predicción del bandgap en su correo electrónico: rodraxsan@hotmail.com.

Figura 3.15: Interfaz web del usuario para la predicción del bandgap alojada en AWS

### Carga del Dataset en AWS S3

El primer paso consistió en cargar el dataset original en un bucket de Amazon S3. Este dataset posee cerca de 30 mil instancias con los 20 campos originales, como el bandgap, la fórmula química, el número de sitios, el número de elementos, el volumen, la densidad, la simetría cristalina, entre otros. También es necesario configurar los permisos adecuados en el bucket para garantizar un acceso seguro y controlado a los datos.

## Creación de la API REST con AWS API Gateway

Para permitir la interacción entre la página web y el backend, se creó una API REST utilizando AWS API Gateway. Se configuraron las rutas necesarias y se definieron los métodos permitidos, como GET, POST y OPTIONS. Además, se establecieron los permisos de autorización mediante el uso de IAM Roles, que permiten controlar el acceso a los recursos de AWS.

Se configuraron también los CORS (Cross-Origin Resource Sharing) para permitir la comunicación entre el origen donde estará alojada la página web y la API. Esto es fundamental para evitar problemas de seguridad y garantizar que solo se permitan las interacciones autorizadas.

También se definieron los Stages donde se desplegará la API, lo que permite tener diferentes entornos (desarrollo, pruebas, producción) y gestionar las versiones de la API de manera eficiente.

Para facilitar el monitoreo y la resolución de problemas, se conectó la API con CloudWatch, un servicio de AWS que permite recopilar y analizar los registros de logs de la API en tiempo real.

Finalmente, se configuró la integración de la API con la función Lambda, que se encargará de procesar las solicitudes y generar las respuestas apropiadas.

## Entrenamiento del modelo final por medio de AWS Lambda

Se creó una función Lambda utilizando Python para realizar el entrenamiento del modelo final y enviar la predicción del bandgap al correo electrónico del usuario. Sin embargo, esta tarea presentó varios desafíos iniciales, como por ejemplo la falta de bibliotecas necesarias para el funcionamiento del modelo, tales como scikit-learn, periodictable, numpy y la biblioteca smtplib para enviar correos electrónicos con Python.

Para solucionar este problema, se descargaron las bibliotecas requeridas desde el índice de paquetes de Python (pypi.org) en formato .whl. Luego, se agruparon en un archivo comprimido y se cargaron como una capa (layer) en la función Lambda. Esto permitió que la función tuviera acceso a todas las dependencias necesarias.

Una vez configurada la capa con los paquetes requeridos, se utilizó la biblioteca boto3 de Python para programar la función Lambda. Se creó un cliente S3 para acceder al dataset cargado en el bucket de S3. Al leer el archivo CSV, se cargaron los datos y se aplicaron las transformaciones necesarias para el pre-procesamiento tales como la extracción de características derivadas de la fórmula química, la conversión de variables a tipo numérico, y demás transformaciones mencionadas en la

sección 3.2.

Además, la función Lambda permite cargar el cuerpo del evento, que contiene los datos enviados por el usuario a través de la interfaz web. Estos datos vienen en formato JSON, por lo que se realiza una conversión a un diccionario de Python para poder acceder a los valores de cada característica ingresada por el usuario.

Luego, se normaliza el dataset utilizando MinMaxScaler, ajustándolo a los datos de entrenamiento, que en este caso corresponderán al dataset completo, dado que el rendimiento del modelo ya ha sido evaluado previamente en conjuntos de entrenamiento y prueba. Los datos del nuevo compuesto ingresado por el usuario también serán normalizados utilizando los parámetros de ajuste obtenidos de los datos de entrenamiento del dataset completo, garantizando así la consistencia en la escala de las características.

Una vez entrenado el modelo de clasificación, se le pasan los datos del nuevo compuesto ingresado por el usuario, y el modelo clasifica si pertenece a la clase 0 (bandgap igual a cero) o a la clase 1 (bandgap distinto de cero). En función de la predicción del modelo de clasificación, se realiza una lógica condicional:

- Si la predicción es de la clase 0, la función Lambda envía la predicción de bandgap igual a cero al correo electrónico del usuario.
- Si la predicción es de la clase 1, se entrena el modelo de regresión Gradient Boosting Regressor (GBR) y se envía al correo electrónico del usuario una predicción del bandgap como un número real con 16 decimales.

### Interfaz del usuario por medio de Frontend en AWS

Ahora que ya tenemos construido el backend. El último paso será definir el frontend, que también estará alojado en AWS. Para lo cual se creó un código HTML que genera una página web sencilla y funcional donde el usuario puede ingresar los datos del nuevo compuesto a predecir. Para lo cual se creó un formulario con los nombres de todas las características necesarias para la predicción y se envían todos los datos como una solicitud POST a la API REST en formato JSON.

El archivo index.html con el código de la interfaz web del usuario se ubicó en otro bucket de S3 configurado como público. Para este bucket se realizaron configuraciones especiales, como las políticas del bucket, para permitir la comunicación entre los componentes interactuantes. También se configuró CORS para definir los métodos permitidos (GET, POST, OPTIONS, etc) y se activó la propiedad Static Website Hosting en el bucket, lo que permite alojar directamente una página web con un dominio de AWS desde un bucket de S3.

Para realizar pruebas de todas las conexiones entre el frontend y el backend, se utilizó AWS Cloud9, un entorno de desarrollo integrado (IDE) basado en la nube. Después de una exhaustiva serie de ensayos y meticulosas pruebas utilizando los registros de logs generados por la API y la función Lambda en CloudWatch, se logró una conexión exitosa entre el frontend y el backend. Lo cual permite al usuario ingresar los datos de un nuevo compuesto junto con su correo electrónico y, en menos de 2 minutos, recibir una predicción precisa del bandgap correspondiente, tal como se observa en la Figura 3.16.

### Prediccion BangGap

 correo.████████.autoenviado@outlook.com  
21:30

Para: rodraxsan@hotmail.com

El valor del BandGap predicho por el modelo es: 3.0383186226263494

Figura 3.16: Resultado del bandgap recibido en email del usuario

Es importante también notar la precisión de nuestro modelo final, donde para este ejemplo en particular los datos ingresados en la interfaz web, hacen referencia a un compuesto semiconductor nunca antes visto por el modelo y que según la información descargada de la API del Materials Project posee un valor real del bandgap de 2.84 eV, bastante cercano al valor 3.03 eV que recibe el usuario en su correo electrónico.

En el caso se tenga interés en evaluar el modelo implementado en AWS, simplemente es necesario acceder a la siguiente página web:

<https://bandgap-prediction.s3.amazonaws.com/index.html>

e ingresar los datos del compuesto de interés, utilizando el tipo de dato correcto en cada caso, tal como se muestra en la Figura 3.15. Considerando por ejemplo que el campo de formula química deberá ser un string al igual que la simetría cristalina que solo podrá contener uno de los 7 tipos mencionados en la Sección 3.2.1 o que los datos booleanos de estabilidad y magnetismo se encuentran representados por una casilla donde para True se la tendrá que marcar y para False dejarla desmarcada.

Este proceso demuestra la capacidad de AWS para integrar diferentes servicios y crear soluciones escalables y eficientes para el despliegue de modelos de Machine Learning. Mostrando así la capacidad de la combinación de almacenamiento en la

nube, API REST, funciones serverless y alojamiento web que permite crear aplicaciones robustas y accesibles para los usuarios finales.

De esta manera comprobamos que a través de una cuidadosa configuración y pruebas exhaustivas, se ha logrado una solución funcional donde el usuario puede ingresar los datos de un nuevo compuesto en una página web y recibir una predicción precisa del bandgap en su correo electrónico en cuestión de minutos.

## 4. Discusión y Conclusiones

### 4.1. Conclusiones

Este trabajo de fin de máster ha logrado con gran éxito el desarrollo de modelos robustos de machine learning para la predicción precisa del bandgap en compuestos inorgánicos, utilizando un extenso dataset de cerca de 30 mil materiales obtenido a través de la API de Materials Project. La calidad y amplitud de este dataset ha sido fundamental para el entrenamiento de modelos generalizables.

Se ha demostrado contundentemente la eficacia de combinar técnicas de aprendizaje supervisado, específicamente un modelo de clasificación Random Forest y un modelo de regresión Gradient Boosting, en un ensamble final que supera el rendimiento reportado en varios estudios recientes del estado del arte en este campo. Esta estrategia de combinar la fortaleza de múltiples enfoques ha resultado ser un diferenciador clave.

Además se ha demostrado que debido a la naturaleza de nuestro dataset la precisión del modelo final decrece a medida que el bandgap crece, tal como lo reportan también otros investigadores [9]. Por lo tanto, podemos concluir que nuestro modelo genera predicciones con alta confiabilidad para materiales metálicos o conductores, confiabilidad moderada para materiales semiconductores y baja confiabilidad para materiales aislantes.

La adopción de una rigurosa metodología siguiendo el proceso KDD (Knowledge Discovery in Databases), que abarca desde la extracción y preprocesamiento de los datos hasta técnicas avanzadas de ingeniería de características y evaluación de los modelos utilizados, ha sido fundamental para obtener modelos predictivos de alto desempeño. Este enfoque sistemático ha permitido un aprovechamiento óptimo de la información contenida en el dataset.

Los análisis exploratorios detallados realizados, que incluyen el estudio de correlaciones entre variables, la evaluación de la importancia de características y la generación de visualizaciones reveladoras, han permitido una comprensión profunda de los factores que influyen en el bandgap. Este valioso conocimiento no solo ha facilitado el desarrollo de modelos más robustos, sino que también ha brindado importantes insights físicos sobre los materiales estudiados.

La implementación de técnicas avanzadas de feature engineering, como la extracción de información adicional de la fórmula química de los compuestos, ha demostrado ser un aspecto clave para mejorar significativamente el rendimiento de los modelos. Este hallazgo resalta la importancia de incorporar conocimiento experto

del dominio en el proceso de modelado.

El despliegue exitoso del modelo final en una página web alojada en AWS, con una arquitectura robusta que integra múltiples servicios en la nube como S3, API Gateway y Lambda, demuestra la viabilidad de ofrecer la predicción del bandgap como un servicio accesible y eficiente para la comunidad científica. Esta implementación abre la puerta a la democratización de herramientas avanzadas de descubrimiento de materiales.

Los resultados obtenidos no solo tienen implicaciones prácticas para acelerar el desarrollo de nuevos materiales con propiedades electrónicas deseadas, sino que también demuestran el inmenso potencial de la sinergia entre la ciencia de datos y la física de materiales.

Con respecto a los objetivos planteados inicialmente en la sección 1.2.1 podemos comprobar que todos han sido abordados en el presente trabajo:

- Objetivo general 1: Este objetivo se cumplió exitosamente en el trabajo. Se aplicaron diversas técnicas de aprendizaje automático, como Gradient Boosting Regression (GBR), Random Forest Regression (RFR), Kernel Ridge Regression (KRR), Support Vector Regression (SVR), entre otras, para desarrollar modelos de regresión capaces de predecir el valor del bandgap. El modelo GBR demostró el mejor desempeño, logrando un error medio absoluto (MAE) de 0.3 eV y un coeficiente de determinación ( $R^2$ ) de 0.89, superando incluso varios estudios recientes en la literatura científica relacionada con la predicción del bandgap.
- Objetivo general 2: Este objetivo también se alcanzó de manera satisfactoria, debido a que se empleó el algoritmo de clasificación Random Forest (RFC) junto con técnicas de optimización de hiperparámetros, como RandomSearch y GridSearch, para desarrollar un modelo capaz de distinguir entre compuestos metálicos y no metálicos basándose en su bandgap. El modelo RFC logró una precisión superior al 92 % y un área bajo la curva ROC del 98 %, demostrando un rendimiento competitivo en relación al estado del arte en la literatura científica para esta tarea de clasificación.
- Objetivo específico 1: Este objetivo se cumplió ampliamente en el trabajo, debido a que se llevó a cabo un análisis exploratorio exhaustivo de las relaciones entre diversas características de los materiales, como la simetría cristalina, la densidad, el volumen y la energía por átomo, y su influencia en el bandgap. Se generaron visualizaciones reveladoras, como gráficos de correlación y dispersión, que permitieron una comprensión profunda de los factores que afectan el bandgap. De esta manera, el análisis realizado brindó conocimientos físicos valiosos sobre los materiales estudiados y las relaciones más relevantes entre sus características.

- Objetivo específico 2: Este objetivo se abordó en el trabajo, aunque con resultados mixtos. Se aplicaron técnicas de aprendizaje no supervisado, como el clustering de características utilizando el método de enlace de Ward y la distancia euclídea, para identificar subconjuntos de características similares. Sin embargo, los experimentos realizados con la selección y combinación de características basadas en estos clusters no mejoraron significativamente el rendimiento de los modelos de predicción. Se concluyó que el dataset ya contaba con un número relativamente reducido de características, todas ellas con alto contenido informativo, lo que limitó el impacto de la reducción de dimensionalidad en este caso particular.
- Objetivo específico 3: Este objetivo se cumplió de manera destacada en el trabajo, ya que se implementaron diversas estrategias de ingeniería de características, como la extracción de información adicional de la fórmula química de los compuestos, generando nuevas columnas para cada elemento presente. Esta técnica demostró ser altamente efectiva, mejorando el rendimiento de los modelos en alrededor de un 10 %. Además, se llevaron a cabo experimentos combinando características fuertemente correlacionadas, aunque estos no resultaron en mejoras significativas. La selección de características basada en su importancia, determinada por el algoritmo Random Forest, también fue una estrategia valiosa para identificar las variables más influyentes en la predicción del bandgap.
- Objetivo específico 4: Este objetivo se cumplió de manera sobresaliente, debido a que se diseñaron, entrenaron y optimizaron múltiples modelos de regresión y clasificación utilizando técnicas de aprendizaje supervisado. Para la tarea de regresión, el modelo Gradient Boosting Regression (GBR) demostró el mejor desempeño, superando varios estudios recientes en la literatura científica. Para la clasificación, el modelo Random Forest Classifier (RFC) alcanzó una precisión notable, comparable con el estado del arte. La optimización de hiperparámetros mediante técnicas como RandomSearch y GridSearch fue fundamental para obtener el máximo rendimiento de los modelos. Además, se evaluó exhaustivamente el desempeño de los distintos enfoques utilizados, lo que permitió una comparación rigurosa con otros trabajos relacionados.
- Objetivo específico 5: Este objetivo se cumplió de manera exitosa y destacada, por medio del desarrollo de una interfaz web intuitiva y funcional, alojada en AWS, que permite a los usuarios ingresar los datos de un nuevo compuesto inorgánico y recibir la predicción del bandgap en su correo electrónico en cuestión de minutos. La arquitectura del sistema, que integra diversos servicios de AWS como S3, API Gateway y Lambda, demostró ser robusta y eficiente. El back-end procesa la información ingresada por el usuario mediante los algoritmos de aprendizaje automático desarrollados, mientras que el módulo de notificación por correo electrónico garantiza una comunicación rápida y efectiva de los resultados. Este despliegue del modelo final como un servicio accesible

y fácil de usar representa un logro significativo en términos de transferencia tecnológica y democratización de herramientas avanzadas para la predicción de propiedades de materiales.

## 4.2. Trabajo futuro y limitaciones

A pesar del excepcional rendimiento demostrado por los modelos desarrollados, es importante reconocer que su capacidad de generalización a materiales significativamente diferentes a los presentes en el dataset de entrenamiento podría ser limitada. Para abordar esta limitación, sería valioso ampliar la diversidad del dataset incorporando nuevas clases de compuestos, como materiales 2D, estructuras orgánicas complejas o materiales amorfos.

Explorar arquitecturas más complejas de deep learning, como redes neuronales de grafos o convolucionales, podría permitir capturar relaciones no lineales más sutiles entre las características de los materiales y el bandgap, y así mejorar aún más la precisión de las predicciones, especialmente para valores extremos del bandgap. Estos enfoques podrían ser particularmente beneficiosos para materiales con estructuras electrónicas altamente correlacionadas.

Investigar técnicas avanzadas de transfer learning para aprovechar el conocimiento de modelos pre-entrenados en datasets masivos como AFLOW o OQMD, podría reducir la necesidad de grandes cantidades de datos etiquetados y acelerar el desarrollo de modelos especializados para clases específicas de materiales. Esta estrategia podría ser especialmente útil para estudiar materiales menos comunes o con propiedades exóticas.

Extender la metodología presentada para predecir no solo el bandgap sino múltiples propiedades optoelectrónicas simultáneamente, mediante enfoques de aprendizaje multi-tarea, brindaría una herramienta aún más potente para el diseño integral de materiales. Esto permitiría una optimización holística de las propiedades de los materiales para aplicaciones específicas.

Además en un trabajo futuro, se podría realizar una optimización más exhaustiva de los hiperparámetros de los modelos actuales, por medio de RandomSearch y GridSearch, para potencialmente incrementar aún más la precisión del modelo final. A su vez también se podrían integrar los modelos predictivos desarrollados con técnicas avanzadas de optimización y búsqueda, como algoritmos genéticos o de optimización Bayesiana, que permitan incluso implementar un marco de diseño inverso que pueda sugerir composiciones óptimas de materiales dado un bandgap deseado. Este enfoque de diseño guiado por IA presenta un gran potencial para el aceleramiento del descubrimiento de nuevos materiales con propiedades a medida.

Si bien el despliegue actual del modelo en una aplicación web es funcional y

eficiente, futuras versiones podrían incorporar visualizaciones interactivas de los resultados que permitan a los usuarios explorar de manera intuitiva las relaciones entre las características de los materiales y el bandgap predicho. Además, la implementación de una función de predicción en lotes podría acelerar la evaluación de grandes grupos de compuestos candidatos.

Aunque este trabajo se ha centrado en la predicción del bandgap, la metodología desarrollada podría extenderse a otras propiedades críticas de los materiales, como la conductividad térmica, la resistencia mecánica o la estabilidad química. Sin embargo, esto requeriría la recopilación y curación de datasets de alta calidad para estas propiedades, lo cual puede ser un desafío dada la escasez de datos experimentales en algunos casos.

La integración de los modelos de machine learning con simulaciones físicas de primeros principios, como los cálculos de DFT, podría permitir un enfoque híbrido que combine la eficiencia de los modelos de IA con la precisión de los métodos ab initio. Además se podría incorporar en un trabajo futuro, la incorporación de distintos datasets, cada uno con distinto nivel de fiabilidad, tal como lo proponen Pilania et al. [4].

Finalmente, será crucial desarrollar métricas y protocolos rigurosos para validar y garantizar la confiabilidad de los modelos de IA en el contexto de la ciencia de materiales. Esto implicará no solo la evaluación exhaustiva del rendimiento predictivo, sino también la interpretabilidad de los modelos y su capacidad para capturar relaciones físicamente significativas. Para lo cual será crucial la estrecha colaboración entre expertos en ciencia de datos y científicos de materiales.

# Bibliografía

- [1] Chi Chen, Yunxing Zuo, Weike Ye, Xiangguo Li, and Shyue Ping Ong. Learning properties of ordered and disordered materials from multi-fidelity data. *Nature Computational Science*, 1(1):46–53, 2021.
- [2] Steven K Kauwe, Taylor Welker, and Taylor D Sparks. Extracting knowledge from dft: experimental band gap predictions through ensemble learning. *Integrating materials and manufacturing innovation*, 9(3):213–220, 2020.
- [3] Bart Olsthoorn, R Matthias Geilhufe, Stanislav S Borysov, and Alexander V Balatsky. Band gap prediction for large organic crystal structures with machine learning. *Advanced Quantum Technologies*, 2(7-8):1900023, 2019.
- [4] Ghanshyam Pilania, James E Gubernatis, and Turab Lookman. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science*, 129:156–163, 2017.
- [5] Arunkumar Chitteth Rajan, Avanish Mishra, Swanti Satsangi, Rishabh Vaish, Hiroshi Mizuseki, Kwang-Ryeol Lee, and Abhishek K Singh. Machine-learning-assisted accurate band gap predictions of functionalized mxene. *Chemistry of Materials*, 30(12):4031–4038, 2018.
- [6] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.
- [7] L Weston and C Stampfl. Machine learning the band gap properties of kesterite i 2- ii- iv- v 4 quaternary compounds for photovoltaics applications. *Physical Review Materials*, 2(8):085407, 2018.
- [8] Shuming Zeng, Yinchang Zhao, Geng Li, Ruirui Wang, Xinming Wang, and Jun Ni. Atom table convolutional neural networks for an accurate prediction of compounds properties. *NPJ Computational Materials*, 5(1):84, 2019.
- [9] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters*, 9(7):1668–1673, 2018.