Universidad Abierta y a Distancia de México

UnADM

División de Ciencias Exactas, Ingenierías y Tecnología

Ingeniería en Desarrollo de Software

Asignatura:

Fundamentos de Investigación

DS-DFIN-2002-B2-017

Unidad 5: Análisis de datos y el informe de resultados

Actividad 1. Análisis de datos

Alumno:

Angel Fernando Cisneros Gaytán

ES202113379

Docente:

Carmita Castillo Sastré

Noviembre, 2020

Introducción

Machine Learning (ML) es una disciplina científica del ámbito de la inteligencia artificial que desarrolla sistemas basados en algoritmos que aprenden automáticamente a tráves del análisis de datos. Actualmente el campo de ML está tomando cada vez mayor relevancia debido al aumento exponencial de información accesible al investigador generada mediante experimentos o cálculos computacionales. Desde hace algunos años existen bases de datos que contienen información sobre algunas de las propiedades de materiales obtenidas principalmente por cálculos cuánticos. Por ejemplo, *Aflow* cuenta en este momento con más de 3 millones de materiales catalogados mientras que *Materials Project* cuenta con casi 2 millones de materiales. Ante esta mina de datos cabe preguntarse cómo analizar toda esta información o parte de ella. Una de las ramas de ML, conocida como ML No Supervisado, desarrolla algoritmos para la visualización de datos.

1. Realiza una lista de todos los datos cuantitativos que se pueden recolectar en el tema de investigación que propones.

La correcta selección de características o propiedades resulta ser un paso crucial al desarrollar algortimos eficientes de ML con un alto poder predictivo. En cuanto a la aplicación de ML en el área de ciencia de materiales, los siguientes datos han demostrado ser determinantes en el diseño de nuevos materiales **(Wang, 2019)**.

- Energía: Calculada mediante programas de estructura electrónica.
- Energía por átomo: Energía normalizada por átomo en la celda unitaria.
- Volumen: Volumen final del material.
- Densidad electrónica: Disposición de los electrones en el espacio tridimensional.
- Elementos: Arreglo de los elementos en el material.
- Estructural cristalina: Disposición geométrica de los átomos dentro de la celda unitaria.
- Cargas atómicas: Carga parcial de cada átomo.
- Densidad de Estados, Estructura de bandas y *band gap*: Propiedades electrónicas de los materiales.
- Momento magnético.
- Propiedades termodinámicas y elásticas.

2. Revisa bibliografía o páginas oficiales que sean fuentes fiables de donde puedas obtener la información que necesitas para la recolección de datos necesarios de acuerdo con tu tema.

La información necesaria para desarrolar el presente proyecto de investigación puede ser obtenida de las siguientes bases de datos:

- Materials Project
- AFLOW (Automatic-FLOW for Materials Discovery)
- Crystallography Open Database

3. Selecciona la información necesaria con base en el área de tu investigación

De acuerdo a la planeación desarrollada y entregada en una de las actividades anteriores, se propone desarrollar un modelo de Machine Learning de la categoría del Aprendizaje No Supervisado para diseñar nuevos materiales empleados como cátodos en baterias de ion litio. El entrenamiento del modelo se realizará utilizando datos sobre las propiedades magnéticas, termodinámicas y especialmente sobre las electrónicas, tales como la densidad electrónica y las cargas atómicas.

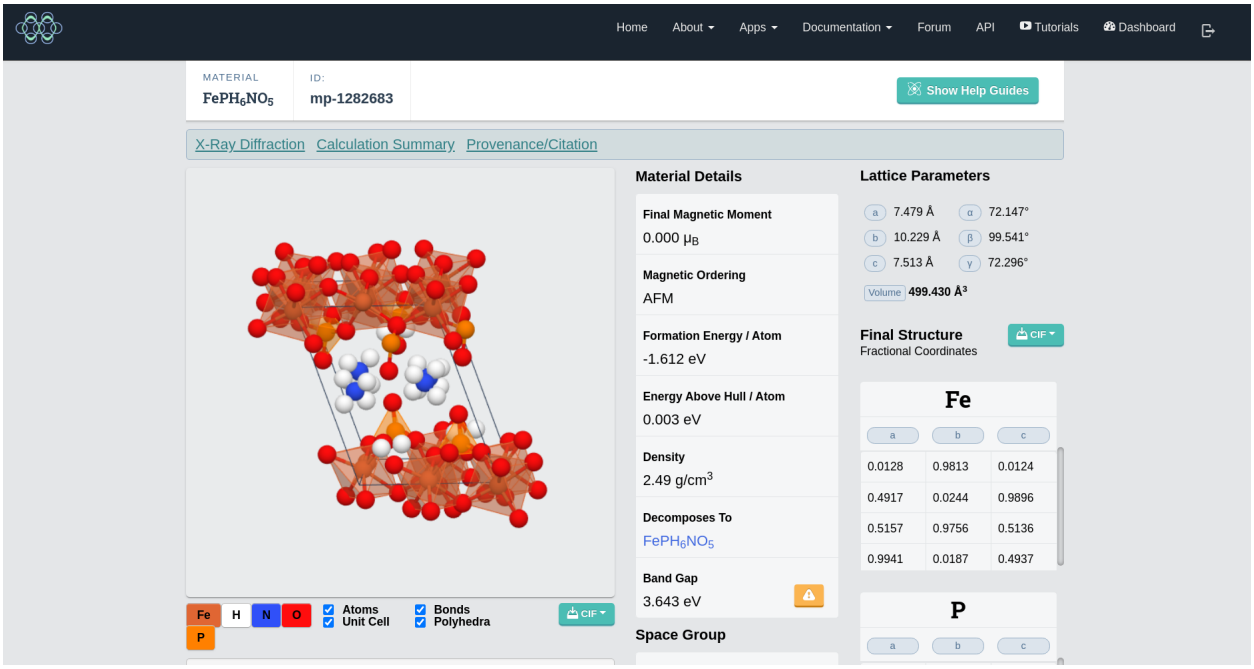


Figure 1: Celda unitaria del material $FePH_6NO_5$ y algunas de sus propiedades obtenidas por cálculos cuánticos y publicados en la base de datos Materials Project.

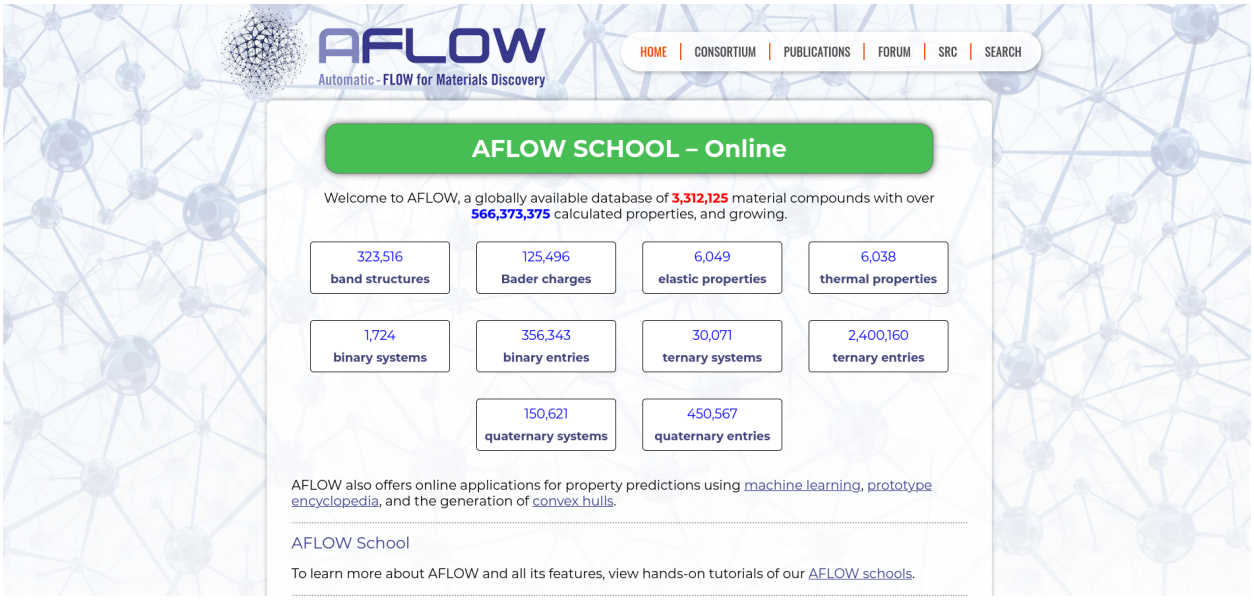
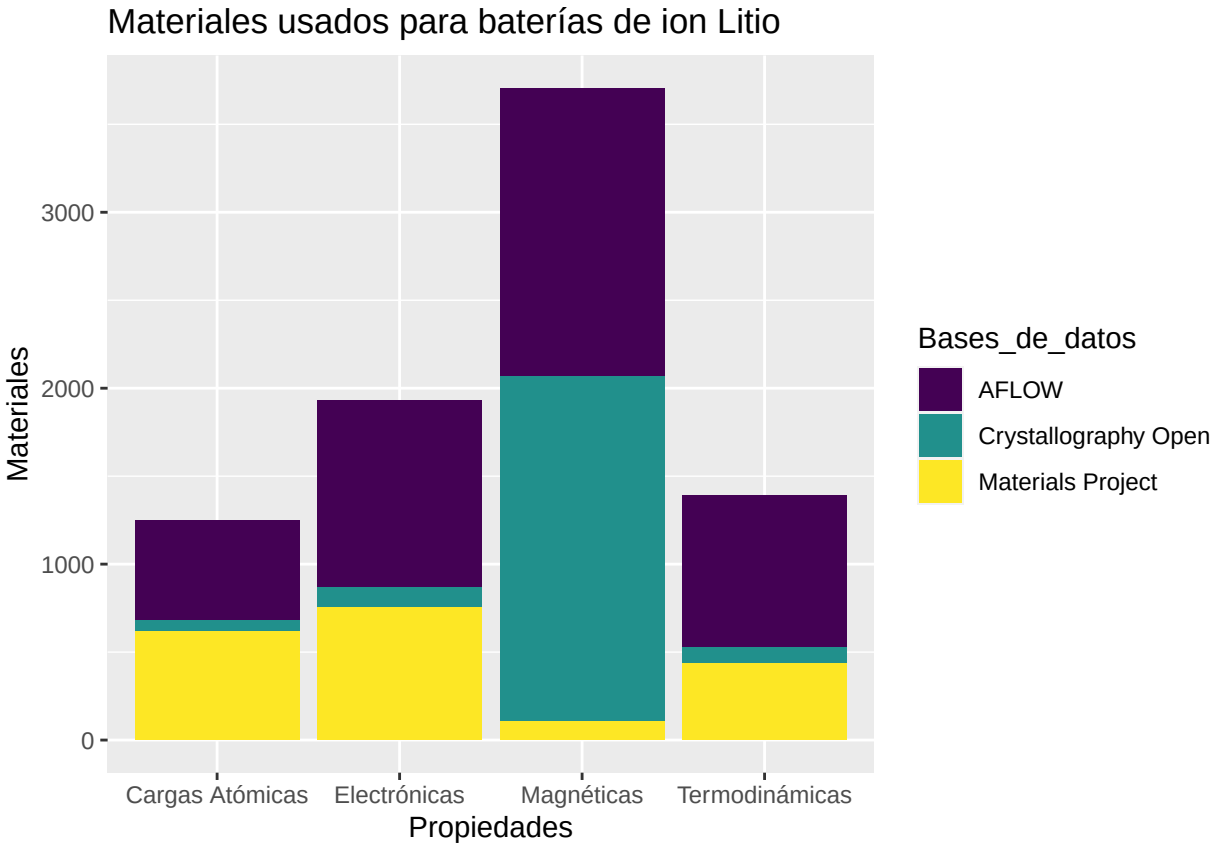


Figure 2: AFLOW es otra de las bases de datos más populares y grandes con mas de 500 millones de propiedades calculadas.

4. Analiza la información más importante sobre los objetivos que planteaste en la unidad 2, y la lista que realizaste al inicio de la actividad, después:

a) Selecciona al menos dos series de datos de los indicadores que consideres y elabora cuando menos dos gráficas, una de barras y una de pastel, de cada serie de datos con base en tu información.

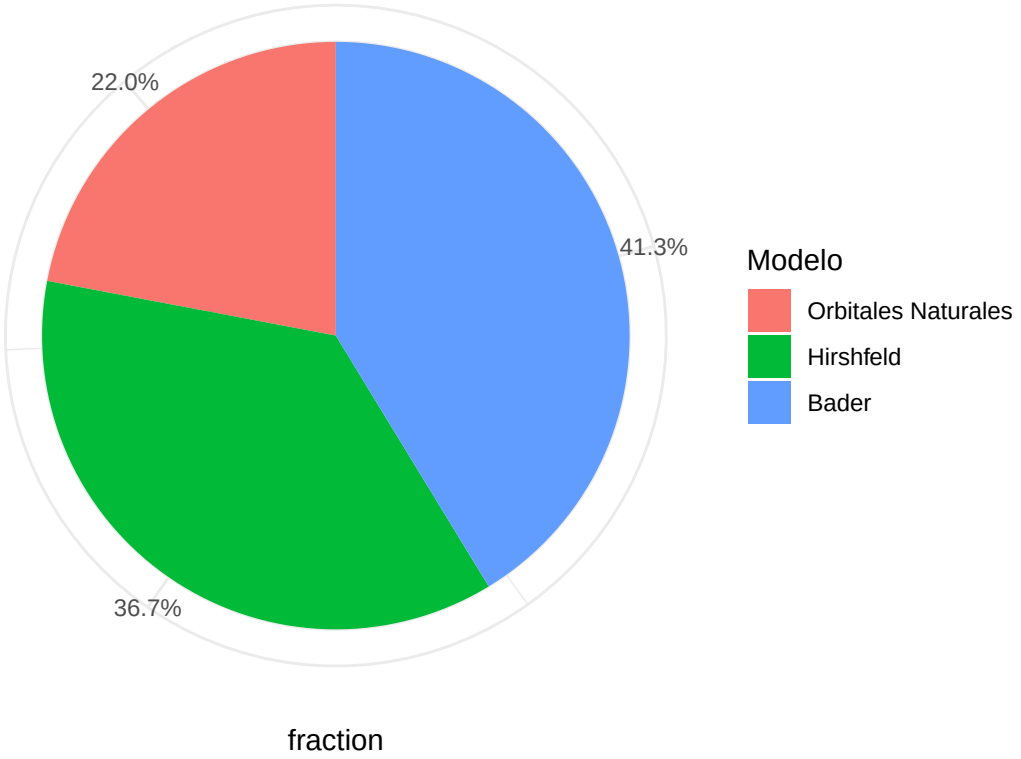
```
## Loading required package: viridisLite
```




```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:viridis':
##
##     viridis_pal
```

Modelo o teoría cuántica utilizada para calcular las cargas atómicas.



b) Redacta el análisis de los datos obtenidos (reflexión), por cada gráfica, sobre el proceso de graficado y sus resultados obtenidos.

En cuanto al proceso de graficado, dependerá totalmente de la herramienta utilizada, las cuales pueden ser clasificadas en dos grandes grupos: aquellos programas basados en una interfaz gráfica tales como Excel y OriginLab y los programas que utilizan la linea de comandos como Gnuplot, matplotlib de python o **ggplot** de R, el cual fue utilizado en esta actividad. La ventaja de utilizar programas como ggplot es que permite la automatización de tareas y el manejo de grandes cantidades de datos. Además, y es una mera opinión personal, la estética de las gráficas supera por mucho las relizadas por Excel. La desventaja es necesidad de tener que aprender una cantidad considerable de comandos para poder graficar lo que uno realmente desea. Y respecto a los resultados, la gráfica de barras muestra como la información sobre las propiedades químicas de los materiales para baterías de ion litio varía en las tres bases de datos consultadas. Por lo tanto, para el entrenamiento del modelo de ML será necesario obtener información de las tres bases. De la gráfica de pastel se observa que las cargas atómicas son más comúnmente calculadas usando la teoría de Bader ya que es menos costosa computacionalmente, lo cual se refiere a que el tiempo de cómputo es menor a comparación de los otros dos modelos.

Conclusiones

Actualmente existen bases de datos que almacenan enormes cantidades de propiedades electrónicas, termodinámicas, elásticas, magnéticas, entres otras calculadas por medio de teorías cuánticas a la espera de ser analizadas y así obtner un conocimiento más profundo sobre el comportamiento de los materiales en diferentes escenarios de aplicación. Es aqui donde las técnicas de Machine Learning basadas en probabilidad, algebra lineal se unen la teoría de bases de datos y la programación para analizar datos y convertirlos en conocimiento.

Referencias

[1] Wang, M., Wang, T., Cai P. & Chen, X. (2019). Nanomaterials discovery and design trough machine learning Small Method. 1900025: 1-7.

[2] Stefano, C. (2012). AFLOW: An automatic framework for high-throughput materials discovery. Computational Materials Science, 58 (2012) 218-226.

[3] Ceder, G. (2013). The Materials Project: A materials genome approach to accelerating materials innovation APL Materials, 1(1), 011002.

[4] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

[5] R Core Team, (2020). R: A Language and Environment for Statistical Computing, R Foundation for Statistical, Vienna, Austria, https://www.R-project.org/.