Reporte

November 5, 2023

1 Minería de datos en ciencias de materiales

- Fabián Castro Contreras
- Sebastian Monteiro
- Simón Campos Rojas
- Dylan Riquelme

- Benjamín Mancilla

1.1 Introducción: Problema y Motivación

La ciencia de materiales es un campo en constante evolución que desempeña un papel fundamental en la creación de tecnologías innovadoras y en la mejora de los materiales existentes para una amplia gama de aplicaciones. En este contexto, la minería de datos emerge como una poderosa herramienta que nos permite desentrañar patrones, descubrir relaciones y extraer información valiosa a partir de vastos conjuntos de datos relacionados con materiales.

Existen poderosas bases de datos que han utilizado algoritmos complejos y supercomputadoras para predecir propiedades de materiales. Una base de datos famosa y ampliamente utilizada debido a la calidad de sus datos es *Materials Project*, la cual presenta información de más de 150,000 materiales. Por lo tanto, encontrar patrones en las propiedades desemboca en un área para la minería de datos.

1.2 Materials Project

Materials Project cuenta con una API especialmente diseñada para el acceso a esta base de datos no SQL escrita en MongoDB. Dado la vasta cantidad de datos, se presentará el código utilizado para extraer los datos, pero **NO se sugiere ejecutar este código** ya que tarda bastante en descargar todo.

La API utilizada se llama *mp_api* y puede ser instalada usando *pip*.

La base de datos contiene varias clases, las cuales hay que descargar una por una con las propiedades de interés.

2 Exploración de Datos

2.1 1.- Consiguiendo los datos:

En primera instancia, se obtiene una llave mediante una cuenta creada en la pagina de *Materials Project* para acceder a la base de datos de la misma, luego se descargan los datos y se almacenan

en archivos .csv. Después se crea un repositorio en *Github* con todos los datos, los requerimientos para el proyecto (librerías) y los archivos *.ipynb* para el futuro código. Para revisar los atributos extraídos para cada clase o dataset, ir a la tabla anexada.

2.2 2.- Limpieza de datos

A priori no se requiere una limpieza de datos tan exhaustiva para los datos, dado que *Materials Project* tiene los datasets muy completos. Sin embargo, se tuvo que eliminar una columna repetida llamada *"Unnamed: 0"* porque correspondían a los ID´s repetidos y se dejaron todas las columnas solicitadas en el query inicial.

Luego se procede a unir los dataframes mediante un merge en base a la ID de cada objeto, para facilitar la exploración de datos y la interpretación de estos, resultando en un dataframe con 7290 filas y 22 columnas. A continuación se muestran los primeros 5 objetos del dataframe:

Descripción de los datos luego del merge

Notar que cada columna es un atributo físico de los materiales. Al hacer merge, algunos materiales ya no aparecen en la lista, se redujo la cantidad de filas de ~150.000 a ~7.000. Esto se produjo porque, para el dataset de materiales dieléctricos, no están los cálculos de los atributos para todos los materiales.

Este análisis se realiza para poder hacer una mirada superficial a las matrices de covarianza, correlación y estadísticas varias. Más adelante es posible que se trabaje con los dataframes por separado debido a la variedad de estructura entre los distintos materiales, o también de alguna forma establecer una estructura universal para poder realizar el merge, pero no perdiendo una cantidad de datos tan grandes.

2.3 3.- Estadísticas de los datos

A continuación se muestra la descripción de los datos.

Estadísticas de los datos

Notar que los atributos **e_total, e_ionic, e_electronic y n** tienen valores muy extremos, por esto eso que poseen una desviación estándar tan alta. Además, la diferencia entre los valores mínimos y máximos de estos atributos es muy grande, lo que reafirma la dispersión de los datos. Esto se puede ver en los histogramas de cada atributo:

Histogramas de cada atributo

Como se puede observar, la gran mayoría de estos histogramas siguen un patrón de distribución χ^2, no obstante, los histogramas de energía de fermi (efermi) y de energía por átomo (energy_per_atom) parecen acercarse más a una distribución normal. Hay que notar que para los histogramas de **e_total, e_ionic, e_electronic y n**, se eliminaron los valores extremos para poder apreciar mejor la distribución de los datos.

Se presenta la matriz de correlación de los datos:

Matriz de correlación

Esta matriz desvela distintas características de los datos: - Entre la energía de reacción en equilibrio y la entalpía de descomposición hay una gran correlación de 0.93. - Entre la energía de fermi y la densidad hay una correlación interesante de 0.53 y posiblemente útil para la caracterización de los

materiales. - Entre la energía total y n hay correlación de 0.4 y entre e_total y e_electronic hay una correlación de 0.44. - Entre e_total y e_ionic hay una correlación de 0.93. - Entre el band_gap y la energía de formación por átomo, la densidad y la energía de fermi existe una correlación inversa de -0.47, -0.39 y -0.53 respectivamente.

Sobre las demás relaciones, hay muy poco más que se pueda rescatar, dada la baja relación entre los atributos.

A continuación se presenta la matriz de covarianza de los datos:

Matriz de covarianza

Desafortunadamente, esta matriz no nos entrega mucha información, ya que no nos deja ver la variabilidad de los datos.

3 Inquietudes a responder

- ¿Existen grupos de materiales con características similares?
- ¿Podemos encontrar materiales extraños, por ejemplo, material magnético que no sea metálico?
- ¿Podemos identificar una característica que influya en lo bueno que puede ser un conductor?
- ¿Existen grupos de materiales con características similares?

Al no tener labels, es directo que se va a usar clustering. Se probará con DBSCAN, con K-means y jerárquico.

Para el caso de k-means, se utilizará el método del codo.

DBSCAN no debería ser útil, debido a que aglomera usando las densidades, cosa que , por intuición fisica, no debería influir en el tipo de material, pero de todas maneras se realizará este análisis.

El clustering jerárquico también nos podría entregar un resultado interesante, ya que podrian haber subconjuntos de materiales.

Luego, una pregunta interesante es ver que característica es mas decididora para determinar la pertenencia de un punto a un clustering. Por ejemplo, si ploteamos los puntos con colores que dependen del valor de un atributo, y este color es muy intenso dentro de un cluster en particular, no sería descabellado pensar que esta característica influye bastante en este material y sus características.

- ¿Podemos identificar una característica que influya en lo bueno que puede ser un conductor?

Primero, vamos a filtrar a todos los conductores usando el atributo band_gap $= 0$, esto pues se sabe que todos los conductores tienen este valor. Los semiconductores tienen un bandgap > 0 y los aislantes tienen band gap mucho mas grande. La siguiente imagen ejemplifica un material con band gap $= 0$.

Notar como en el gráfico de la derecha, que muestra la densidad de estados, no hay saltos en la función de color rojo.

Gráfico de bandgap (1)

El siguiente material tiene un band gap distinto de cero, es decir, los electrones necesitan un "pequeño empujón" para pasar de un estado ligado al estado libre de un conductor. Este "pequeño

empujón" se puede conseguir con una diferencia de potencial o con el efecto túnel. Aquí la función roja no es continua en todo el espacio, específicamente alrededor de 0.

Gráfico de bandgap (2)

Luego, debido a que no tenemos labels, clasificar no tendría sentido. Por esta razón, se realizará un clustering.

Primero, se probará con k-means . Para inspeccionar la cantidad idónea de clusters, primero se realizará la técnica del codo.

Con la cantidad de clusters obtenida, se realizará k-means.

También se probará con DBSCAN. La intuición física nos dice que no servirá mucho, pues la densidad de puntos no debería ser un factor de clasificación.

El cluster que nos interesa es el jerárquico, pues queremos identificar que tan buenos son los conductores. Queremos identificar alguna característica que tenga un gradiente que coincida con la forma en la que se agrupan los clusters. Una forma de visualizar esto sería algo así:

Gráfico de ejemplo de cluster

Notar como los puntos se vuelven más rojos a medida que nos acercamos al centro, una especie de gradiente hacia el centro nos podría indicar que esta característica influye directamente en lo bueno (o malo) que es un conductor. Se utilizará PCA para dimensionar a un gráfico.

Posteriormente se realizará una validación de los métodos de clustering e intentaremos identificar alguna característica o estructura interesante.

4 Anexos

4.1 Tabla de atributos

Clase	Atributos
properties_mat (MaterialsDoc)	material_id, composition, volume, density, density_atomic
properties_thermo (ThermoDoc)	material_id, energy_per_atom, formation_energy_per_atom, equilibrium_reaction_energy_per_atom, decomposition_enthalpy
properties_electro (ElectroDoc)	material_id, band_gap, efermi, is_metal, is_stable
properties_magnetic (MagneticDoc)	material_id, is_magnetic, exchange_symmetry, num_magnetic_sites
properties_dielectric (DielectricDoc)	material_id, e_total, e_ionic, e_electronic, n
properties_oxidation_states (Oxidation-StateDoc)	material_id, possible_species, possible_valences, average_oxidation_states

Atributos extraídos según la información brindada en la documentación de Materials Project.

4.2 Resultado del merge de los dataframes

[ ]:
```
result.head()
```

```
  material_id    e_total    e_ionic  e_electronic         n  band_gap  \
0    mp-28944  21.521069  13.942520      7.578549  2.752916    1.5135
1    mp-28096   3.218928   0.598031      2.620897  1.618918    2.8804
2   mp-863678  10.737604   6.585117      4.152487  2.037765    1.6514
3    mp-10461  13.624477   9.998830      3.625647  1.904113    2.9095
4     mp-8756   8.849831   4.815258      4.034573  2.008625    2.5447

     efermi  is_metal  is_magnetic  exchange_symmetry  …      volume  \
0  1.973873     False        False                186  …  208.023425
1 -1.097983     False        False                 43  …  465.547991
2  2.395559     False        False                146  …  261.201622
3  2.551343     False        False                167  …  331.052135
4  0.345471     False        False                129  …  146.040316

    density  density_atomic                possible_species  \
0  5.939485       34.670571         ['Bi3+', 'Cl-', 'Te2-']
1  1.926612       29.096749                   ['S+', 'Cl-']
2  5.454615       13.060081   ['O2-', 'K+', 'Sb5+', 'Zn2+']
3  5.052121       15.047824  ['O2-', 'Sb5+', 'Na+', 'Sr2+']
4  2.842588       24.340053           ['Li+', 'Se2-', 'K+']

                                  possible_valences  \
0                 [3.0, 3.0, -2.0, -2.0, -1.0, -1.0]
1  [1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, -1.0,…
2  [1.0, 2.0, 2.0, 2.0, 2.0, 5.0, 5.0, 5.0, -2.0,…
3  [1.0, 1.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 5.0, …
4                   [1.0, 1.0, 1.0, 1.0, -2.0, -2.0]

                     average_oxidation_states energy_per_atom  \
0          {'Bi': 3.0, 'Te': -2.0, 'Cl': -1.0}       -3.902939
1                        {'S': 1.0, 'Cl': -1.0}       -3.466514
2   {'K': 1.0, 'Zn': 2.0, 'Sb': 5.0, 'O': -2.0}       -5.816968
3  {'Na': 1.0, 'Sr': 2.0, 'Sb': 5.0, 'O': -2.0}       -6.392569
4            {'K': 1.0, 'Li': 1.0, 'Se': -2.0}       -3.542389

  formation_energy_per_atom  equilibrium_reaction_energy_per_atom  \
0                 -0.958829                             -0.051867
1                 -0.474021                             -0.075092
2                 -1.922136                             -0.022649
3                 -2.738011                             -0.088787
4                 -1.370645                             -0.034785

   decomposition_enthalpy
0               -0.051867
```

```
1               -0.075092
2               -0.022649
3               -0.086070
4               -0.034785

[5 rows x 22 columns]
```

4.3 Estadisticas de los datos

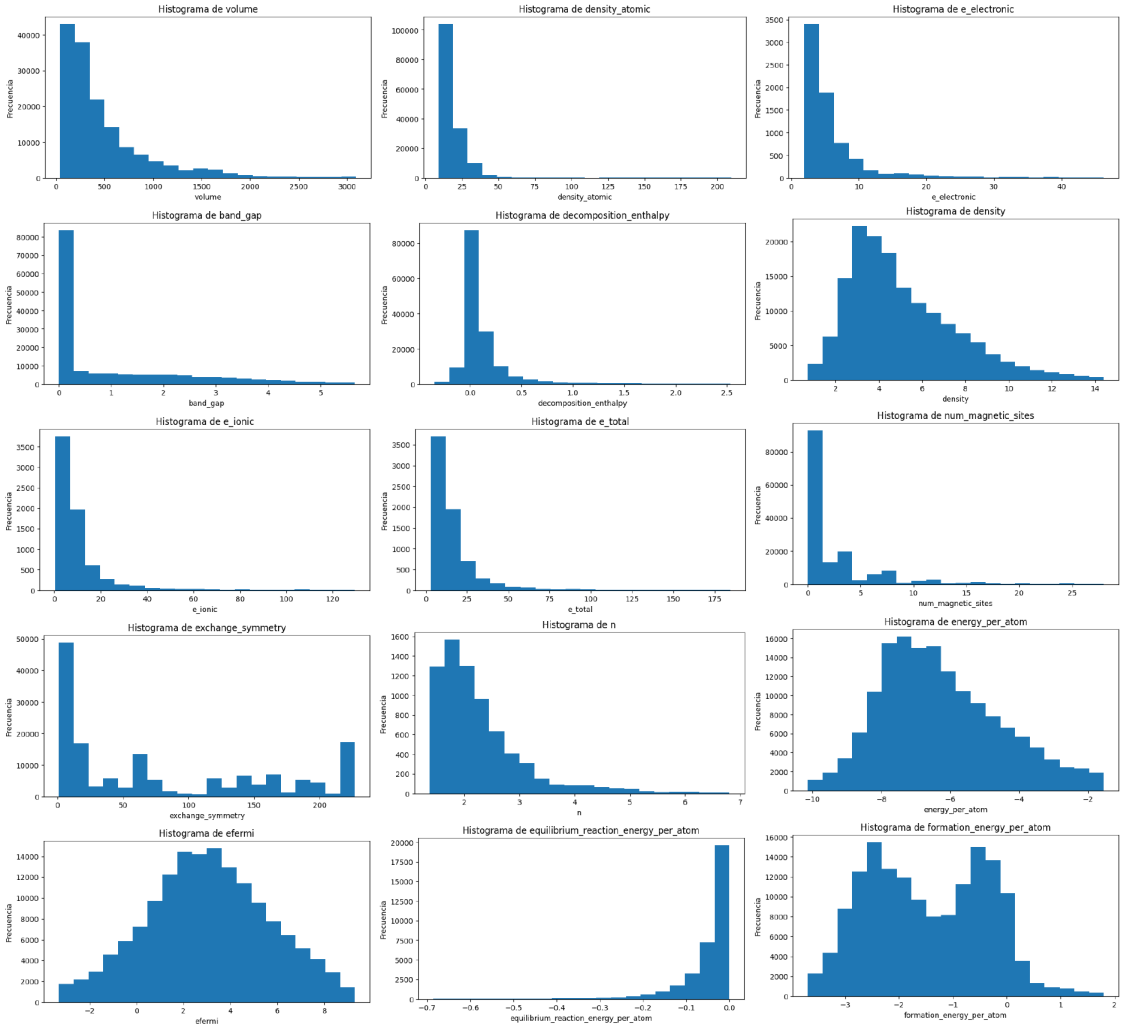
[ ]:
```
numeric_columns.describe()
```

```
            e_total        e_ionic  e_electronic            n     band_gap  \
count   7290.000000    7290.000000   7290.000000  7290.000000  7290.000000
mean      50.920612      32.398492     18.522120     2.436020     2.336182
std     1655.107848    1487.919679    625.658741     3.549444     1.697134
min        1.155248       0.000000    -64.837332     0.000000     0.000000
25%        7.931442       4.000849      2.940330     1.714739     1.003525
50%       11.662250       6.472804      4.264421     2.065047     1.991500
75%       19.016789      11.364285      6.558800     2.561015     3.418875
max   126575.316823  126567.273642  46857.910510   216.466881    12.139100

            efermi  exchange_symmetry  num_magnetic_sites       volume  \
count  7290.000000        7290.000000         7290.000000  7290.000000
mean      1.724809          94.329081            0.560494   289.837992
std       2.388488          76.760838            1.642730   221.638154
min      -7.804680           1.000000            0.000000    11.286588
25%       0.215491          15.000000            0.000000   143.530177
50%       1.749932          72.000000            0.000000   233.974644
75%       3.227443         164.000000            0.000000   367.666530
max      11.149808         230.000000           24.000000  3998.471538

           density  density_atomic  energy_per_atom  \
count  7290.000000     7290.000000      7290.000000
mean      4.402732       18.878844        -5.782469
std       1.862653        8.587621         1.714053
min       0.023670        5.643294       -11.047931
25%       3.084127       12.598164        -7.131346
50%       4.129389       15.950793        -5.739387
75%       5.351808       23.677849        -4.422131
max      17.732855      132.548261        -0.219326

       formation_energy_per_atom  equilibrium_reaction_energy_per_atom  \
count                7290.000000                           4647.000000
mean                   -1.733503                             -0.125225
std                     0.982938                              0.348267
min                    -4.491319                             -4.427975
25%                    -2.500231                             -0.100782
50%                    -1.686184                             -0.041113
```

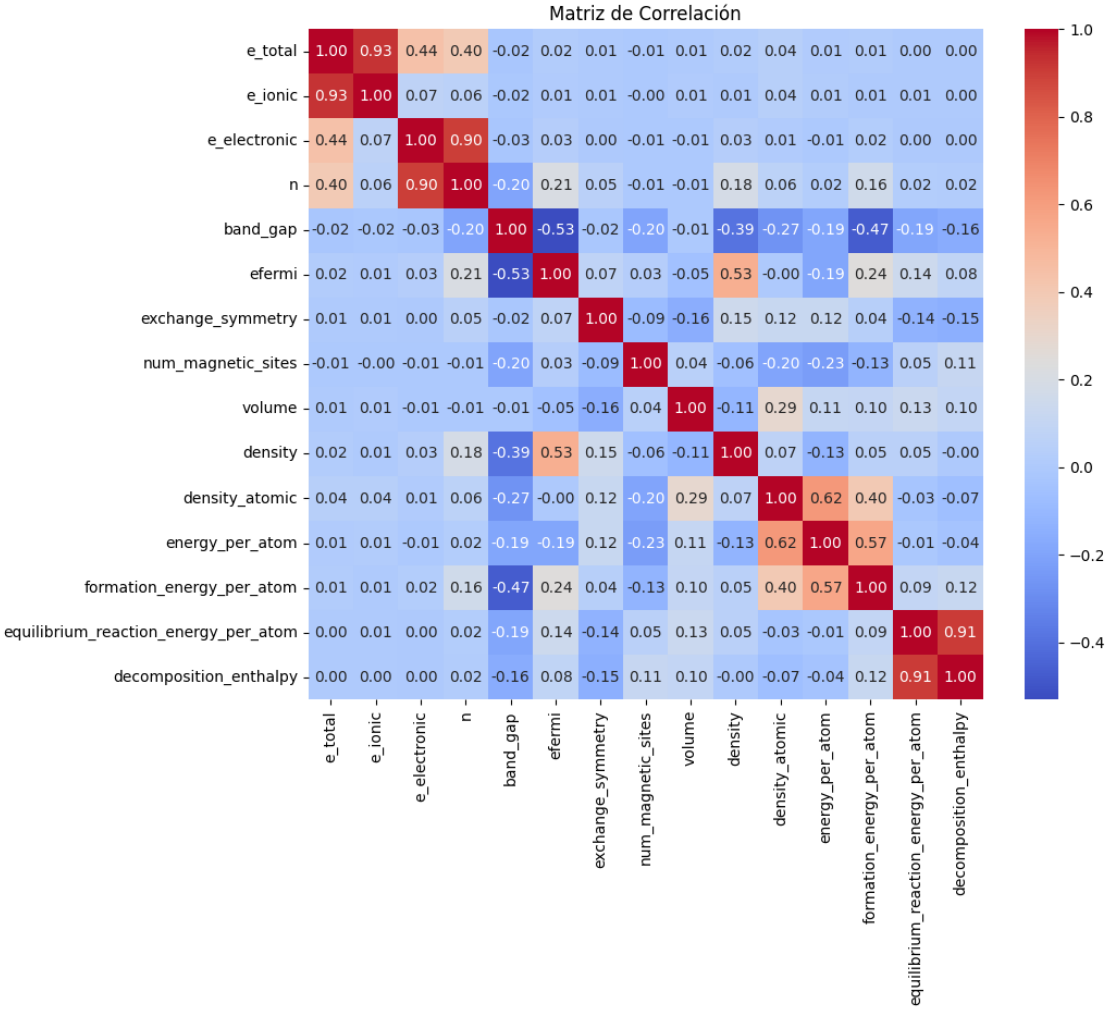
75%	-0.925363	-0.016220
max	5.212425	0.000000

	decomposition_enthalpy
count	7290.000000
mean	-0.075678
std	0.316180
min	-4.427975
25%	-0.078266
50%	-0.023512
75%	0.000381
max	5.212425

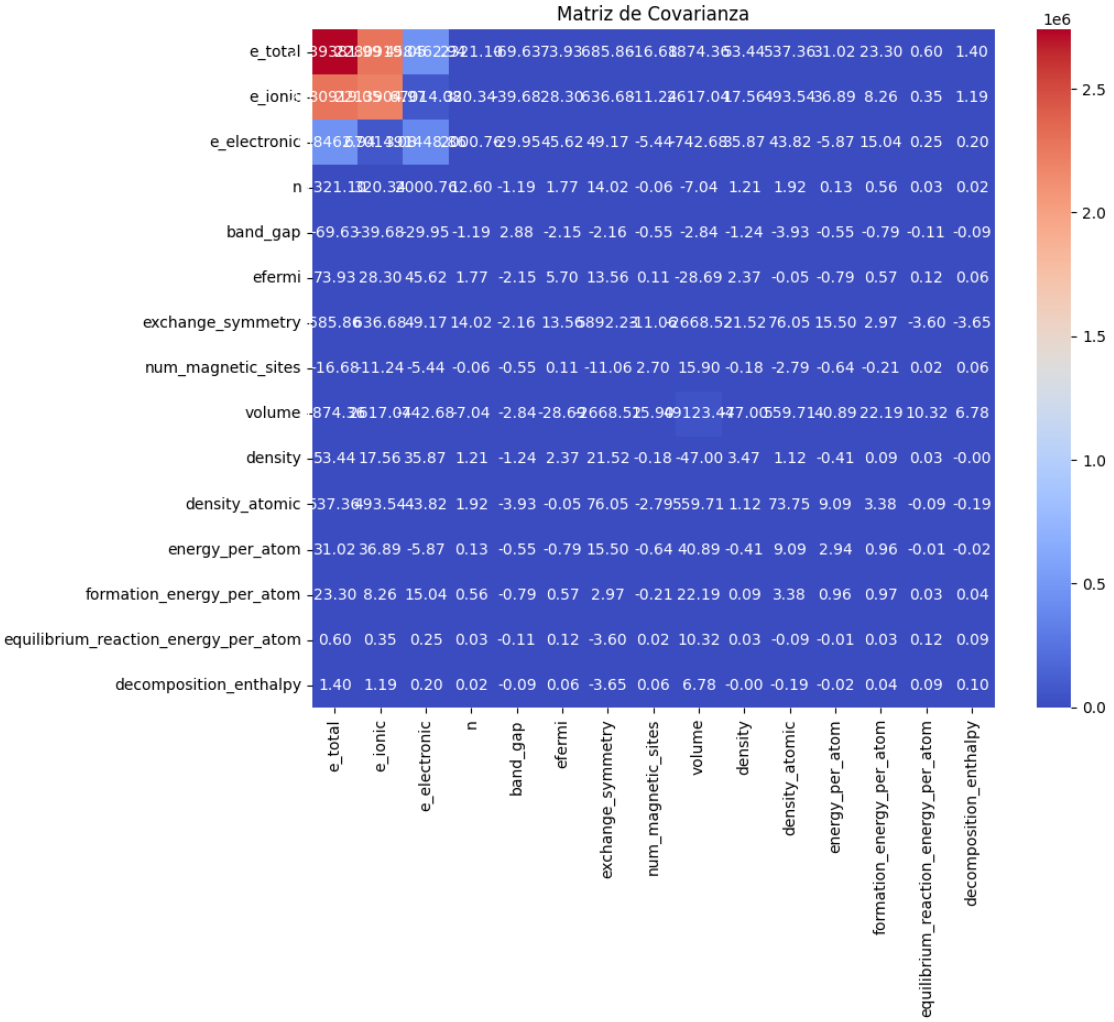
4.4 Histogramas de cada atributo



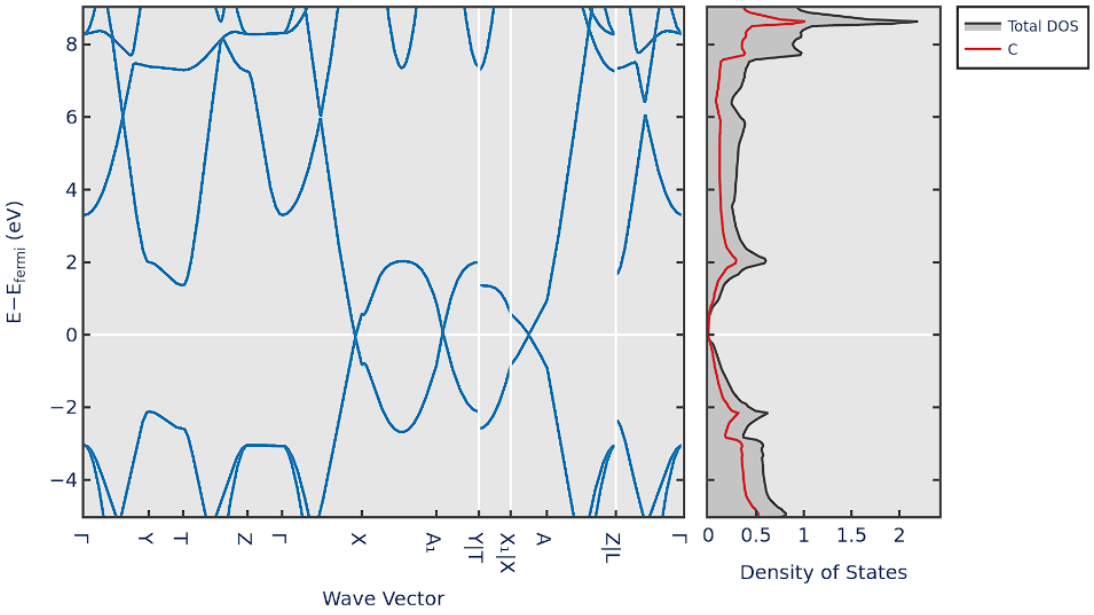
4.5 Matriz de correlacion de los datos



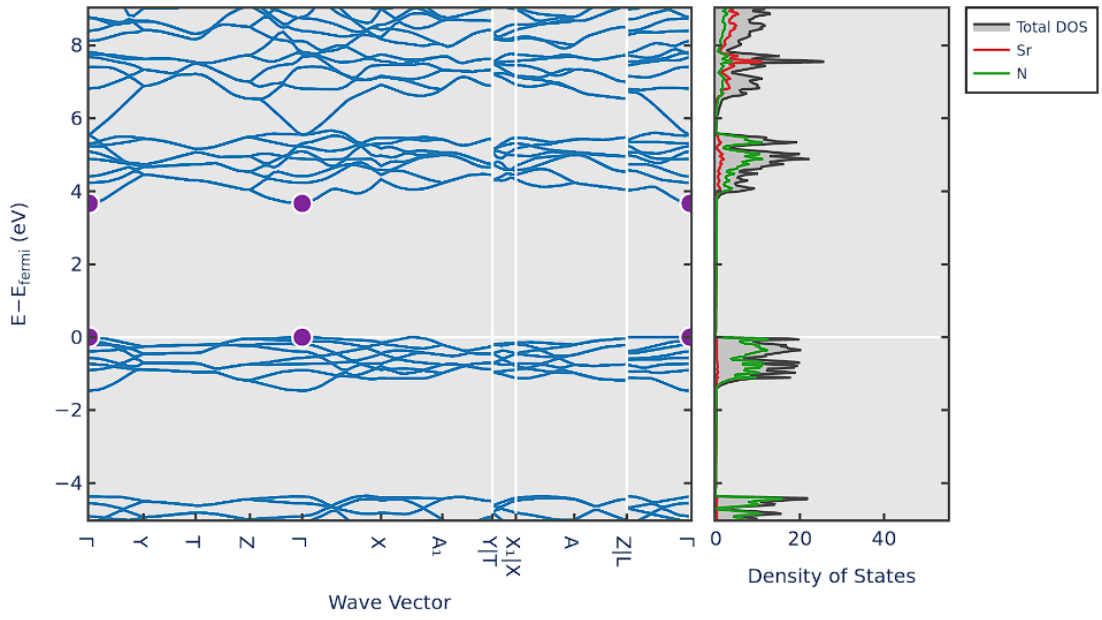
4.6 Matriz de covarianza de los datos



4.7 Bandgap (1)

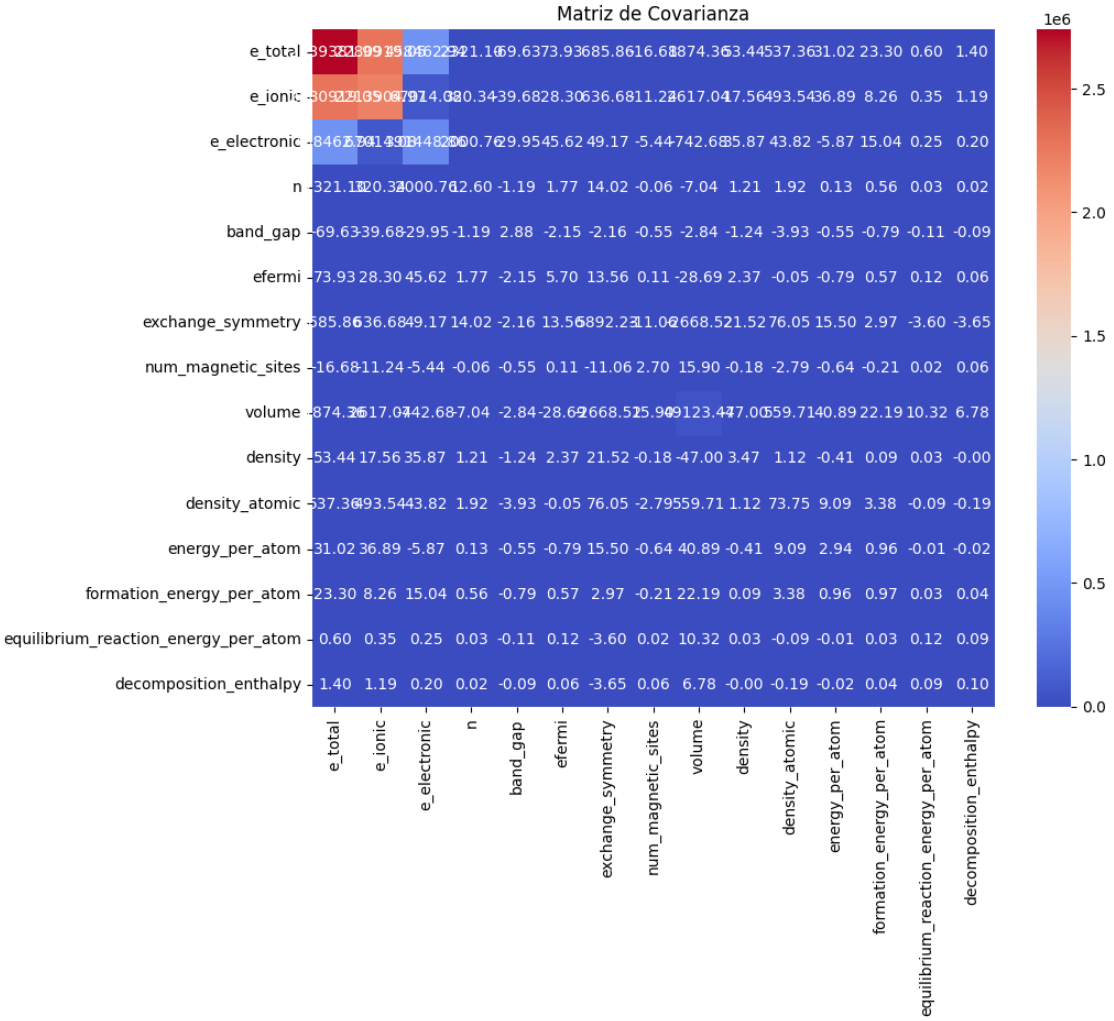


4.8 Bandgap (2)



4.9 Ejemplo Cluster

A continuación se presenta la matriz de covarianza de los datos:

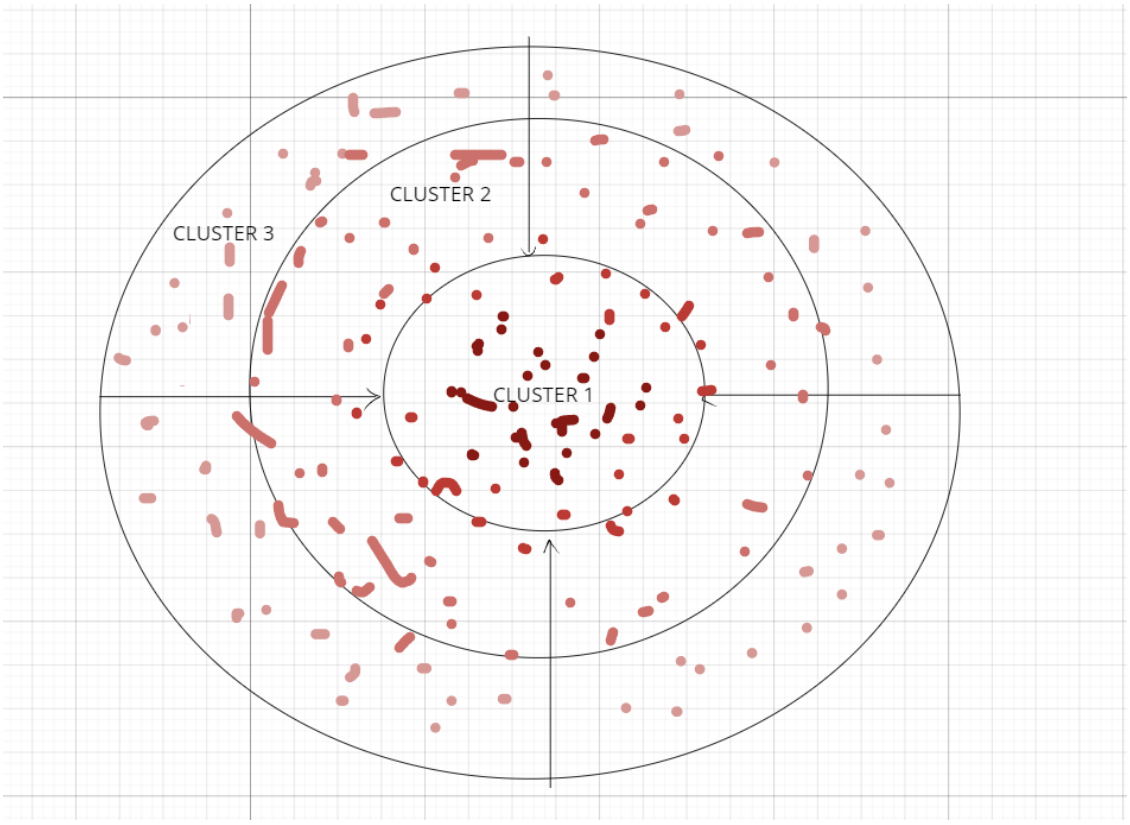


Desafortunadamente, esta matriz no nos entrega mucha información, ya que no nos deja ver la variabilidad de los datos.

5 Inquietudes a responder

- ¿Existen grupos de materiales con características similares?
- ¿Podemos identificar una caracteristica que influya en lo bueno que puede ser un conductor?
- ¿Existen grupos de materiales con características similares?

Al no tener labels, es directo que se va a usar clustering. Se probará con DBSCAN, con K-means y jerárquico.



5.1 Repositorio del proyecto

https://github.com/Fabian-Castro-C/Miner-a-de-datos

5.2 Contribución de miembros del equipo

Miembro	Tarea
Simón Campos	Metodología de investigación
Fabián Castro	Implementación de la API y clustering
Benjamín Mancilla	Limpieza de datos y maestría en git
Sebastián Monteiro	Formulación de preguntas y procedimiento experimental
Dylan Riquelme	Redacción y organización del informe