UNIVERSIDAD
DE ANTIOQUIA

1 8 0 3

PROYECTO ENTREGA 2:
Predicción de propiedades moleculares

Facultad de Ingeniería
Universidad de Antioquia
Abril 2023

Santiago Salazar Correa
Cc. 1000570053

Julián Andrés Rosero
Cc. 1004214604

Linda Vanessa Ramos Pabón
Cc.1085341028

Introducción

La constante escalar de acoplamiento es un parámetro importante en varias áreas de la ciencia y la ingeniería, incluyendo la física de partículas, la física cuántica, la teoría de campos y la teoría de la relatividad, esta hace una descripción cuantitativa de las interacciones entre átomos de una molécula, así como realizar predicciones y cálculos precisos para la corroboración teorías físicas.

En el presente proyecto se pretende realizar la predicción de la constante escalar de acoplamiento correspondiente a la interacción entre dos átomos en una molécula determinada. Para dicho propósito se hará uso de los datos entregados en la competencia de kaggle "*Predicting Molecular Properties*" (https://www.kaggle.com/c/champs-scalar-coupling). A partir de los datasets entregados se extraerá y analizará las distintas características sobre las moléculas y sus respectivos átomos con el objetivo de encontrar una relación entre las mismas, el cual permita dar con un modelo que logre capturar adecuadamente la naturaleza de los datos y la forma en que estos se distribuyen.

Descripción de los datasets

Inicialmente se encontró con la siguiente información en el zip o dataset '*champs-scalar-coupling*' cuando se realizó una extracción de este:

- **Momento dipolar *(dipole_moments.csv)***
- Contribución al escalar coupling *(scalar_coupling_contributions.csv)*
- Cargas mulliken *(mulliken_charges.csv)*
- **Energía potencial *(potential_energy.csv)***
- Tensores de blindaje magnético *(magnetic_shielding_tensors.csv)*
- **Datos de entrenamiento *(Train.csv)***
- **Datos de testeo *(Test.csv)***
- **Estructuras *(structures.csv)***

Se puede observar que el dataset contiene información y mediciones importantes que se le realizaron a diferentes moléculas presentes en los data frame como puede ser su momento dipolar, energía potencial entre otros. Cuando se realizó el análisis de cada uno de los data frames existentes se encontró que:

1. Train: Contiene información como el nombre de la molécula, el tipo de interacción que tienen estas, es decir interacción átomo-átomo, estás presentan todas las posibles combinaciones de iteraciones entre los átomos que se están estudiando en las diferentes moléculas, los índices de los átomos y el valor de la constante de acoplamiento, siendo este el valor relevante para el estudio, como se puede observar en la *figura 1*.

	id	molecule_name	atom_index_0	atom_index_1	type	scalar_coupling_constant
0	0	dsgdb9nsd_000001	1	0	1JHC	84.80760
1	1	dsgdb9nsd_000001	1	2	2JHH	-11.25700
2	2	dsgdb9nsd_000001	1	3	2JHH	-11.25480

Figura 1. Dataframe de Train

2. Test: Contiene la misma información que *Train*, pero con la diferencia de que no está considerada la constante coupling, puesto que, está es quien estamos buscando.
3. Structure: Posee información como nombre de molécula, el tipo de átomo presente en la molécula y sus coordenadas respectivamente, como lo indica la *figura 2*.

	molecule_name	atom_index	atom	x	y	z
0	dsgdb9nsd_000001	0	C	-0.012698	1.085804	0.008001
1	dsgdb9nsd_000001	1	H	0.002150	-0.006031	0.001976
2	dsgdb9nsd_000001	2	H	1.011731	1.463751	0.000277
3	dsgdb9nsd_000001	3	H	-0.540815	1.447527	-0.876644
4	dsgdb9nsd_000001	4	H	-0.523814	1.437933	0.906397

Figura 2. Dataframe de Structure

4. Dipole Moments: Posee información de nombre de molécula e información del vector de momento dipolar de la molécula, descompuesto en coordenadas cartesianas (x,y,z) para las respectivas interacciones átomo-átomo. Este momento dipolar es importante porque habla del comportamiento químico y físico que puede tener la interacción de átomos en una molécula, haciendo propiedades como su solubilidad.

	molecule_name	X	Y	Z
0	dsgdb9nsd_000001	0.0000	0.0000	0.0000
1	dsgdb9nsd_000002	-0.0002	0.0000	1.6256
2	dsgdb9nsd_000003	0.0000	0.0000	-1.8511
3	dsgdb9nsd_000004	0.0000	0.0000	0.0000

Figura 3. Dataframe dipole moment

5. Potential Energy: Este data frame contiene información acerca del nombre de la molécula, la energía potencial de la molécula.
6. Mulliken Charges: Posee información como el valor de mulliken, el cual entrega información de la distribución de la carga electrónica en una molécula.
7. Magnetic Shielding Tensors: Aquí se tiene información como la interacción de átomo-átomo en diferentes coordenadas.

Iteraciones de desarrollo

1. Visualización de los datos:

Se comenzó por hacer una análisis de los datos que se tenían, el cual constó de una visualización de los data frames con el objetivo de verificar con qué tipo de datos se contaba, las columnas, filas y la forma en la cual se organizaba la información, posteriormente mediante el uso de gráficos se visualiza la información de los datos para conocer la distribución de los datos que se consideraron pertinentes para el proyecto,.

El tipo de gráfico empleado para la visualización inicial fue un histograma, puesto que, él mismo permite conocer la distribución de los datos y la cantidades de los mismos, el histograma fue empleado para visualizar la constante escalar de acoplamiento, energía potencial, los valores de momento dipolar el cual consta de tres coordenadas cada coordenada con una columna propia en el dataset.

2. Selección y limpieza de los datos a usar:

A partir de toda la información entregada en los múltiples datasets, se optó por hacer uso de el indice atomico, el tipo de interacción, momento dipolar y energía potencial como información para realizar la predicción de la constante de acoplamiento escalar, se decidió hacer uso de dichas características debido a que de esta forma se obtiene información relevante sobre cada molécula y las interacciones entre cada par de átomos en la misma, dado que, el objetivo es predecir la constante de acoplamiento que existe entre cada par de átomos presentes en determinada molécula. Se descartó hacer uso de las cargas mulkien y tensores de blindaje magnético porque las filas entregadas no corresponden con las filas que se tienen en los anteriores dataframes. Además, con las características anteriormente descritas se espera tener suficiente información para llevar a cabo la predicción sin complejizar demasiado dicho propósito.

Una vez se tuvo claro qué valores se iban a usar se procedió a unir toda esta información en un mismo dataframe haciendo uso de la librería pandas. Luego se procedió a hacer las siguientes modificaciones en los datos:

- Debido a que los datos correspondientes al tipo de interacción son presentados con una nomenclatura que describe la interacción entre los tipos de átomos como una variable categórica, se procede a separar dicha información en columnas que denotan el tipo de interacción con un 1 si se trata de dicha interacción o un 0 si no se trata de ella.
- Se realiza el mismo procedimiento para el índice químico tanto para el *index_atom_0* como para el *índex_atóm_1*.
- Se procede a convertir el 5% de los datos de las tres columnas correspondientes al momento dipolar en datos faltantes tipo NAN.

Las modificaciones anteriormente expuestas se realizaron con el principal objetivo de cumplir con las condiciones establecidas en el proyecto de contar con al menos 30 columnas y un 5% de datos faltantes en al menos tres columnas.

Posteriormente se procedió a reemplazar los datos faltantes en las tres columnas correspondientes al momento dipolar por el promedio de la columna y a eliminar las dos columnas correspondientes a *index_atom_0* e *index_atom_1,* puesto que, ya están siendo descritas por las nuevas columnas creadas.

Para realizar los modelos se decidió extraer del data frame train y el creado por el equipo, con el propósito de sacar unos datos de entrenamiento y otros de testeo, posteriormente a esto se realizó un modelo de machine learning mediante el método de regresión lineal. Lo que se obtuvo fue errores mínimos. Esta prueba se realizó solo utilizando el archivo de train.csv preprocesado, también se evaluaron pruebas de valor absoluto medio, así como error cuadrático medio, Los errores fueron aproximadamente 0 y los score fueron aproximadamente el 100% como se ve en la *figura 4.*

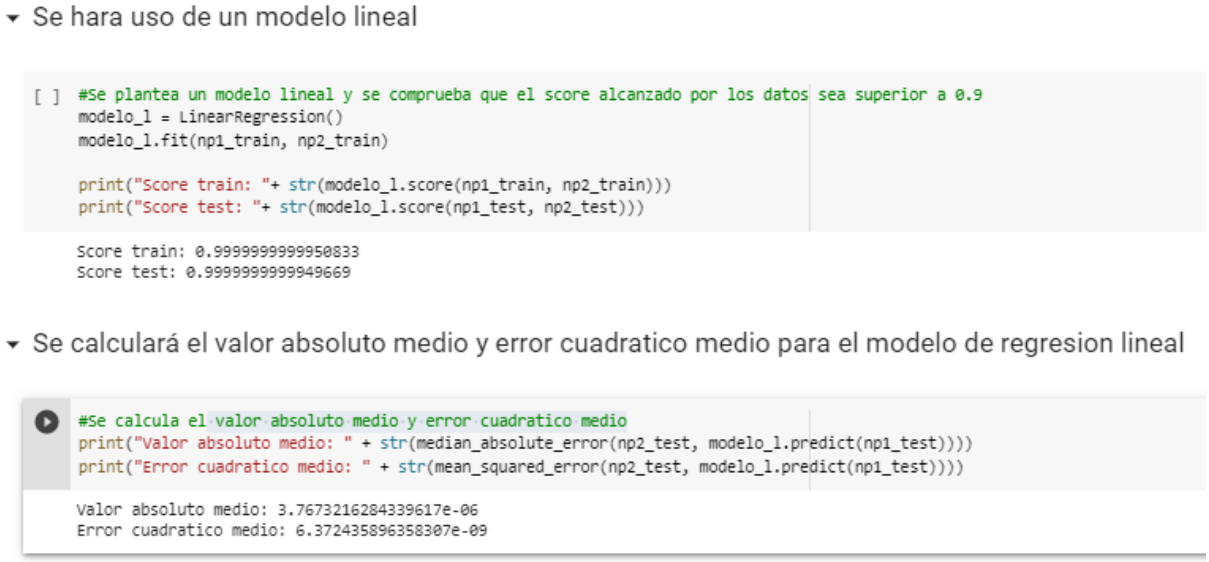

Figura 4. Modelo de regresión lineal usando machine learning

Retos y consideraciones de despliegue

Inicialmente se consideró realizar el proyecto haciendo uso de todas las filas entregadas en el dataset el cual contaba con más de 4 millones de las mismas, sin embargo trabajar con dicha cantidad de datos se tornó en algo sumamente complicado debido a limitaciones de RAM en el entorno (Google Colab); constantemente se reiniciaba el proyecto y los cambios no eran guardados, además de que los tiempos de ejecución de ciertas tareas de limpieza y organización de los datos se tornaba en una tarea de horas complejizando mucho el proyecto.

Debido a las complicaciones anteriormente expuestas se decidió usar una muestra de todos los datos entregados, se procuró entonces tomar alrededor de 500 mil filas teniendo en cuenta que la última molécula en el dataset se encontrara completa y con la información de todos sus átomos, por lo que, finalmente se usaron exactamente 500023 filas del data frames y con las mismas se continuará el proyecto de predicción.

Consideramos que la cantidad de datos usados para el procesamiento de la información es adecuada, puesto que, se está cumpliendo con el requisito de tener al menos 5 mil filas además de que se está respetando el que todas las moléculas empleadas en el dataset tengan la información sobre sus átomos completa.

Una vez entendida, limpia y organizada la información recolectada de los data frames, se procederá a hacer un análisis de la misma haciendo uso de modelos que se adapten y logren

capturar el comportamiento y distribución de los datos empleados, como se mencionó anteriormente se comenzó planteando un modelo lineal, pero se buscará hallar modelos como SVM para realizar una comparación y mostrar el modelo que mejor se ajusta a los datos. Por otro lado, también realizar pruebas de valor absoluto medio, así como error cuadrático medio entre otros.

Conclusiones

1. El progreso alcanzado es bueno, puesto que generalmente una de las partes más complicadas de este tipo de proyectos es la realización de la limpieza y el entendimiento de los datos, así como comprender de forma asertiva el cómo se proseguirá con el análisis para el proyecto final.
2. Hasta el momento los resultados obtenidos mediante el modelo lineal han sido buenos, puesto que, se está obteniendo un alto coeficiente, sin embargo es necesario evaluar los resultados con la métrica establecida en la competencia de kaggle.
3. Es importante tener en cuenta las limitaciones que pueden llegar a tener las herramientas que se empleen para el proyecto en este caso puntual debido a limitaciones de RAM de google colab el servidor se reiniciaba constantemente, puesto que se excede la capacidad lo cual hacía perder el progreso que se tenía con respecto a la limpieza de los datos.