ENTREGA 2 PROYECTO INTELIGENCIA ARTIFICIAL

Integrante:
Juan José Ramúrez Cuervo 1000869826

Tema del proyecto:
Predicción de propiedades moleculares

El presente proyecto es acerca de la predicción de propiedades moleculas de un conjunto de datos obtenidos en kaggle. El objetivo de este proyecto es predecir mediante machine learning la interacción magnética entre dos átomos de cada una de las moléculas que se encuentran en la base de datos.

Lo primero que se realizó fue la carga de la base de datos completa a Google Colab, ara trabajar allí con ella. Luego, se realizó un preprocesamiento de los datos incluidos en el archivo train.csv al adjuntarle diferentes variables más que se encontraban en otros archivos csv de forma que pudieran aportar información. También se pasó la variable categórica que se refiere al tipo de molécula a variable numérica mediante el método onehot

Base de datos original:

	id	molecule_name	atom_index_0	atom_index_1	type	scalar_coupling_constant
0	0	dsgdb9nsd_000001	1	0	1JHC	84.807600
1	1	dsgdb9nsd_000001	1	2	2JHH	-11.257000
2	2	dsgdb9nsd_000001	1	3	2JHH	-11.254800
3	3	dsgdb9nsd_000001	1	4	2JHH	-11.254300
4	4	dsgdb9nsd_000001	2	0	1JHC	84.807400

Base de datos después del preprocesamiento inicial:

id	atom_index_0	atom_index_1	scalar_coupling_constant	dipole_moment_X	dipole_moment_Y	dipole_moment_Z	potential_energy	molecule_name
0	1	0	84.807600	0.0000	0.0000	0.0	-40.523680	dsgdb9nsd_000001
1	1	2	-11.257000	0.0000	0.0000	0.0	-40.523680	dsgdb9nsd_000001
2	1	3	-11.254800	0.0000	0.0000	0.0	-40.523680	dsgdb9nsd_000001
3	1	4	-11.254300	0.0000	0.0000	0.0	-40.523680	dsgdb9nsd_000001
4	2	0	84.807400	0.0000	0.0000	0.0	-40.523680	dsgdb9nsd_000001

Base de datos tras adjuntar la variable categórica a numérica:

	id	molecule_name	atom_index_0	atom_index_1	scalar_coupling_constant	type_1JHC	type_1JHN	type_2JHC	type_2JHH	type_2JHN	type_3JHC	type_3JHH
0	0	dsgdb9nsd_000001	1	0	84.807600	1	0	0	0	0	0	0
1	1	dsgdb9nsd_000001	1	2	-11.257000	0	0	0	1	0	0	0
2	2	dsgdb9nsd_000001	1	3	-11.254800	0	0	0	1	0	0	0
3	3	dsgdb9nsd_000001	1	4	-11.254300	0	0	0	1	0	0	0
4	4	dsgdb9nsd_000001	2	0	84.807400	1	0	0	0	0	0	0

Posterior a esto se realizó un modelo de machine learning mediante el método de regresión lineal. Los resultados obtenidos fueron de errores mínimos. Esta prueba se realizó solo utilizando el archivo de train.csv preprocesado

```
[17] lr = LinearRegression()
     lr.fit(Xtr, ytr)
     lr.score(Xtr, ytr), lr.score(Xts, yts)

     (0.9999999999944946, 0.999999999994517)
```

```
[18] r2_score(yts, lr.predict(Xts))

     0.999999999994517
```

```
[19] median_absolute_error(yts, lr.predict(Xts))

     3.218204964738902e-06
```

```
[20] mean_squared_error(yts, lr.predict(Xts))

     6.6879108703320546e-09
```

Los errores fueron aproximadamente 0 y los score fueron aproximadamente el 100%.