UNIVERSIDAD
DE ANTIOQUIA
1 8 0 3

PROYECTO ENTREGA 1:
Predicción de propiedades moleculares

Facultad de Ingeniería
Universidad de Antioquia
2023

Santiago Salazar Correa
Cc. 1000570053

Julián Andrés Rosero
Cc. 1004214604

Linda Vanessa Ramos Pabón
Cc.1085341028

1. Descripción del problema predictivo a resolver.

Se busca predecir la constante de acoplamiento escalar, la cual describe cómo son los acoplamientos a través de los enlaces, mostrando así cómo se relaciona la interacción magnética entre dos átomos en una molécula. Por otro lado, esta competencia no solo buscará predecir los pares de átomos en cada molécula, sino que también predecir los pares que se encuentran en los archivos, es decir, si por ejemplo tenemos una molécula cuya composición hay flúor (F), pues este no predecirá la constante escalar para ningún par que contenga F.

Dicha interacción, permite maximizar aplicaciones como la Resonancia Magnética Nuclear (RMN), permitiendo obtener resultados más claros acerca de composición molecular de tejidos, estructura y organización de proteínas, permitiendo detectar anomalías en pacientes, para el caso de aplicaciones médicas y biomédicas.
Actualmente, la identificación de dichas constantes es un proceso sumamente costoso, tanto en tiempo como en recursos, puesto que, requiere semanas de trabajo utilizando métodos cuánticos para estimar la constante de acoplamiento para una sola molécula tridimensional.

2. Dataset que va a utilizar.

En Kaggle se encontró la competencia “Predicting Molecular Properties”, la cual presenta un dataset en el cual se cuenta con 131 mil archivos y 47 columnas. Dicho dataset en estas tablas contiene información como nombre de la molécula, los índices de acoplamiento que se buscan predecir, número de átomos, entre otros.

Enlace de la competencia de Kaggle:
https://www.kaggle.com/competitions/champs-scalar-coupling/overview)

3. Métricas de desempeño requeridas (de machine learning y de negocio).

El desempeño del modelo será medido con base en el logaritmo del promedio del error absoluto, calculado para cada tipo de acoplamiento escalar y luego promediado entre tipos, de modo que una disminución del 1 % en MAE para un tipo proporciona la misma mejora en la puntuación que una disminución del 1 % para otro tipo.

$$score = \frac{1}{T}\sum_{t=1}^{T}\log\left(\frac{1}{n_t}\sum_{i=1}^{n_t}|y_i - \hat{y}_i|\right)$$

- T es el número de tipos de acoplamiento escalar
- nt es el número de observaciones de tipo
- yi es la constante de acoplamiento escalar real para la observación

- ŷi es la constante de acoplamiento escalar predicha para la observación

4. Un primer criterio sobre cuál sería el desempeño deseable en producción.

La implementación de este modelo predictivo permite comprender aún más la interacción molecular que a su vez ofrece mayor información acerca de la actividad celular de los tejidos analizados, además de permitir diseñar moléculas que optimicen dicha actividad celular. No obstante, debido a que el modelo aún está en proceso de investigación, no es posible determinar con certeza el desempeño ofrecido en el mismo.

El criterio a tomar en cuenta para decir que el modelo es viable, es tener un score aproximado de -3.23968 o aún más negativo, dado que, el score perfecto que se puede lograr para la predicción es de -20.7232. El score de -3.23968 fue el valor alcanzado por el equipo ganador de la competencia. El objetivo de la predicción es principalmente investigativo, por lo que, esencialmente alcanzar dicho valor o uno más cercano al score perfecto sería de gran relevancia, en virtud de que, se realizaría un aporte al modelo ganador del concurso, el cual en teoría, es el mejor entre todos actualmente.