INTRODUCCIÓN

El objetivo de este reto es crear modelos de aprendizaje automático que utilicen datos de emisiones de código abierto (procedentes de observaciones por satélite Sentinel-5P) para predecir las emisiones de carbono. Como métrica de machine learning vamos a usar el Root Mean Squared Error (RMSE) que es el que define la propia competencia. Como métrica de negocio se podría estimar los niveles de emisiones de carbono en África, incluso en lugares donde no es posible el seguimiento sobre el terreno, ya que es la misión de las organizaciones gubernamentales velar porque estos incrementos no se produzcan de una manera descontrolada.

Se hace necesaria una exploración de las medidas recopiladas y aportadas para esta competencia para lograr una mejor predicción y de alguna manera filtrar aquellos datos que por su naturaleza intrínseca no posean una estrecha relación con la variable objetivo. Esto lo podemos hacer ya que el dataset tiene bastantes variables que pueden ser utilizadas con los diferentes modelos predictivos de machine learning, y aunque desconocemos las distribuciones reales de las variables y el grado de correlación, el objetivo es implementar las mejores técnicas de diseño de algoritmos que mejoren el índice de predicción.

Si las partes por millón de CO2 que se encuentran en la atmosfera no aumentan o disminuyen en más del 10%, el proyecto es sostenible debido a que contribuiría a la mejora del cambio climático.

Exploración descriptiva del dataset

La mayoría de los datos suministrados en el dataset hacen alusión a los niveles de algunas sustancias especificas liberadas a la atmosfera en diferentes regiones geográficas, inicialmente no se observa una relación directa de estas mediciones con la cantidad de emisión de CO2 en la población objetivo, pero con transformaciones se puede lograr encontrar alguna relación que nos permita mejorar la predicción.

Algunas variables presentan relaciones interesantes, pero al no ser con la variable objetivo se puede descartar su aporte a la solución buscada. Otras variables presentan distribuciones muy complejas de abordar en este proyecto e igualmente no se observa correlación con la variable objetivo.

Iteraciones de desarrollo

Se realizaron dos iteraciones con diferentes modelos y parámetros, pero en ninguna de las dos se logró conseguir un ajuste aceptable en cuanto a los valores predictivos de la regresión, se podrían realizar mas adelante realizar otro tipo de filtrado de datos o conseguir otros datos más relacionados.

También se logra un aprendizaje en la elección de las variables, ya que el tener una gran cantidad de variables no garantiza mejores resultados, porque depende de demasiados factores y de entender la base científica de lo que estamos procesando.

Retos y consideraciones de despliegue

Para el despliegue es muy importante tener en cuenta que el procesamiento de los equipos donde este se realice podría afectar considerablemente la ejecución de los algoritmos incluso causar congelamientos y detención imprevista en la ejecución.

También se requiere una integración con los repositorios para poder administrar el esquema de repositorios con el cual se está trabajando.

Conclusiones

El abordar temas de ciencia de datos siempre va a ser algo interesante ya que partimos de la experiencia o acontecimientos estudiados en el pasado para mejorar los resultados que obtengamos en el futuro.

Igualmente en esta área se manejan temas de gran complejidad para lo cual es importante adquirir una muy buena experiencia con temas estadísticos y de manipulación de datos principalmente con lenguajes y APIs diseñadas para tal fin.

Es importante entender que no siempre se consiguen los resultados esperados y menos al primer intento.