Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Puebla

Inteligencia Artificial para la ciencia de datos TC3007C

Reto
Momento de Retroalimentación: Reto Datos

Fernando Jiménez Pereyra	A01734609
Daniel Flores Rodríguez	A01734184
Alejandro López Hernández	A01733984
Daniel Munive Meneses	A01734205

26 de octubre de 2022

Reto

La empresa Naatik desea desarrollar una demo para análisis de churn con el fin de conseguir futuros clientes.

La demo debe ser capaz de recibir archivos de tipos csv con diferentes sets de datos que pudieran tener variables diferentes, y procesarlos para generar perfiles de clientes, y determinar su churn. Entre las limitaciones planteadas por el socio formador están:

- Funciona completamente de manera local.
- Funciona sobre Windows.
- El sistema no debe poder correr en la plataforma de demostración, una computadora con i7 de 11 generación y 16 gigabytes de ram.

El desarrollo del reto se llevará a cabo en tres etapas:

1. Se generará un sistema capaz de procesar el set de datos actual, y un modelo de predicción de churn y de clustering.
2. Se ajustará el sistema para ser capaz de procesar archivos con diferentes sets de datos.
3. Se considera que al finalizar la etapa anterior ya se tiene un producto que cumple con las necesidades del socio formador. Sin embargo durante esta etapa se terminará de refinar el proyecto para garantizar la máxima calidad del proyecto.

Herramientas y tecnologías

Debido a las limitaciones planteadas por el socio formador, se usará el framework pandas para hacer el procesamiento de los datos. Ya que las limitaciones de hardware y software impiden la implementación de algo más avanzado como un cluster de spark.

Almacenamiento de los datos

El alcance del reto no plantea que el sistema sea capaz de acceder a bases de datos o sistemas de archivos externos, limitándose a acceder únicamente a archivos locales. A su vez la limitación sobre el sistema operativo sobre el cual debe trabajar hace que el sistema deba ser capaz de trabajar correctamente sobre el sistema de archivos NTFS.

Con el fin de que el sistema sea lo más eficiente posible en cuanto a tiempo de operación, a medida que el sistema vaya procesando un set de datos irá guardando archivos intermedios. Una vez que el sistema haya procesado los datos, y generado los grupos junto con sus análisis de churn, el sistema generará un reporte sobre el set de datos, el cual se guardará de forma local.

Big Data en el reto

Consideramos que si se podría dar un enfoque a Big Data, ya que como es necesario que el sistema pueda trabajar con diferentes sets de datos con diversidad

de variables, será necesario que se implemente un proceso de ETL que se pueda ajustar a los diferentes sets de datos de forma automática, por lo que se estaría cumpliendo el criterio de veracidad.

Además como el tamaño de los sets de datos que el sistema va a manejar dependerá directamente de las acciones del usuario final, cabiendo la posibilidad de que lleve el sistema al límite para los recursos planteados, se podría decir que cumple el criterio de volumen de forma parcial, ya que no se tiene ninguna certeza de que se cumpla este escenario.