TECNOLÓGICO DE MONTERREY®

Inteligencia artificial avanzada para la ciencia de datos

Reto

Etapa 4. Reporte de modelación

Diego Arturo Padilla Domínguez - A01552594

Keyuan Zhao - A01366831

Carolina Herrera Martínez - A01411547

Cutberto Arizabalo Nava - A01411431

Jose Pablo Cobos Austria - A01274631

Campus Querétaro

04 de noviembre de 2022

Etapa 4: “Modeling”

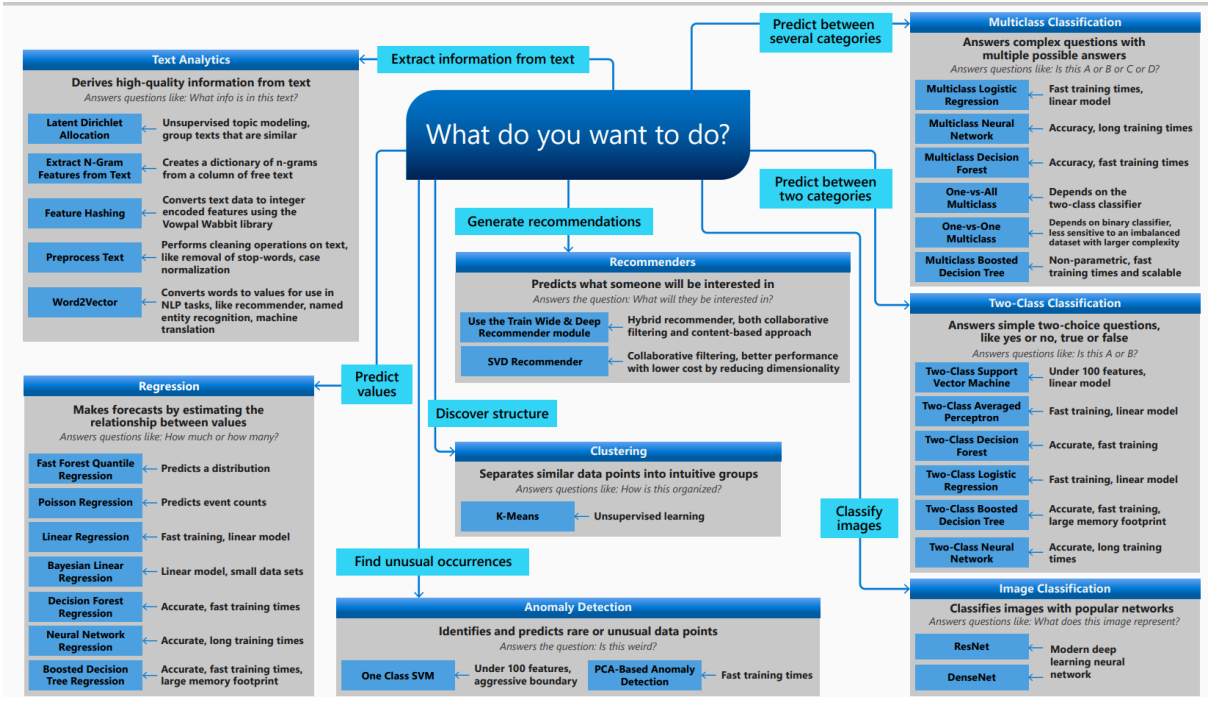
Para ejecutar esta etapa es esencial enfocarnos en el objetivo de la minería de datos, el cual es el siguiente:

“Obtener datos representativos sobre las rutas Origen-Destino mediante la limpieza, transformación y análisis del set de datos inicial. Utilizar dichos datos para entrenar un modelo que sea capaz de predecir la cantidad de viajes a una comuna en base a las características de dicha comuna. “

Por ende, resulta natural que una vez obtenida la matriz de viajes procedamos a buscar cuál es el mejor modelo para dicha predicción.

En el área de ML existen 3 grandes ramas con un gran abanico de modelos que podemos utilizar. Por lo tanto, es importante saber distinguir qué es lo que queremos hacer, para poder elegir el modelo adecuado.

Como base para esta decisión, usaremos la guía de ML de Microsoft Azure, la cual se muestra a continuación:



En nuestro caso, queremos predecir valores (cantidad de visitas a una comuna) en base a los atributos de la comuna, por lo que necesitamos usar una Regresión.

Vemos que existen distintos tipos de regresión, cada tipo tiene distintas ventajas y requisitos para su uso.

Como el modelo más sencillo de regresión es el de regresión lineal, lo utilizaremos como modelo base/benchmark, y este será nuestro punto de comparación con el resto de modelos.
Además, el modelo de regresión lineal nos brinda la ventaja de que podemos analizar las propiedades estadísticas para obtener insights como la significancia de cada variable independiente.

Modelo 1. Regresión Lineal.
Framework: Sci-Kit Learn

Descripción de datos:
Este modelo recibe como variables independientes lo siguiente:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas

Y como variable dependiente:

- Número de viajes hacia la comuna

Resultados:
Este modelo base tuvo un Mean Absolute Error de 2835.56. Dicho valor será utilizado como benchmark para comparación con los siguientes modelos.

Modelo 2. Regresión Lineal.
Framework: Sci-kit Learn

Descripción de datos:
Para este modelo regresamos a la parte de preparación de datos, ya que nos percatamos de que estábamos incluyenddo muy pocas variables en el modelo, por lo que el error era elevado.
La nueva variable a incluir en el modelo es la hora del día en la que se hizo el viaje, por lo que tomamos el valor de la hora (de 0 a 23) y generamos variables “dummies” en base a ellas. Esto nos permite tener un modelo que considere la hora del día para hacer la predicción.

Este modelo recibe como variables independientes lo siguiente:

- Número de escuelas
- Número de hospitales
- Número de iglesias
- Número de zonas típicas
- 23 variables dummies para las horas del día

Y como variable dependiente:

- Número de viajes hacia la comuna

Resultados:
Este modelo base tuvo un Mean Absolute Error de 127.96. Obtuvimos una mejora considerable con respecto a nuestro modelo benchmark, sin la necesidad de utilizar un tipo modelo más complejo.