Equipo Jerry

Reporte Etapa : Modeling

Objetivo:

Predecir la cantidad de viajes que realizan las comunas a las universidades en determinados horarios.

Selección de modelo

Tipos de modelo a emplear

De acuerdo a la problemática a resolver, se establecen los siguientes posibles modelos:

- Modelo de Regresión Lineal: Este tipo de modelo describe la relación entre una variable dependiente y una o más variables independientes, para nuestro caso este modelo resulta de utilidad dado que buscamos hacer una predicción de una cantidad, determinando la relación entre las variables dependiente con las independientes. Como ventajas tenemos que es un modelo fácil de implementar. interpretar y entrenar, y tiene un desempeño bueno hablando de datos linealmente separados. Con este modelo para verificar su calidad y validez, primeramente se dividirán los datos en conjuntos de validación, entrenamiento y pruebas, y se utilizarían métodos como cross-validation, y por medio de porcentajes de error, y variables como R^2 y coeficientes. El modelo será construido con el conjunto de entrenamiento y su calidad estimada con los conjuntos de validación y testeo.
- Random Forest: Este es un modelo que consiste en varios árboles de decisión, el cual puede ser implementado en problemas de regresión y clasificación. Es un posible modelo ya que puede llegar a mejorar la precisión de las predicciones, reduce problemas de overfitting y de varianza. Para la evaluación del desempeño de este modelo, de igual forma se dividirían los datos en conjuntos de prueba, entrenamiento y validación y es posible utilizar técnicas como matrices de confusión, y el resultado de variables como el accuracy_score. El modelo será construido con el conjunto de entrenamiento y su calidad estimada con los conjuntos de validación y testeo.
- RNN (Recurrent Neuronal Network): Este es un tipo de red neuronal que puede minimizar errores de predicción al ajustar variables del modelo, como lo son los pesos. Este tipo de red resultaría útil ya que se busca encontrar una cantidad por medio de datos previos, el cual es uno de sus propósitos principales. Para la evaluación del modelo se utilizarían variables como porcentaje de error o bien accuracy_score.

Diseño de las pruebas

- Revisar los diseños de pruebas para cada objetivo de minería de datos por separado

  Dentro de los objetivos de minería de datos que se establecieron en la primera etapa, se habla de dos puntos:

  1. Identificar la comuna de origen que registre los mayores viajes a cada una de las universidades, para el entrenamiento del modelo, se va a tomar en cuenta la cantidad de viajes que se hace de cada comuna hacia las universidades.
  2. Determinar los rangos de horarios donde se hacen más viajes hacia las universidades, para esto, se va a tomar en cuenta la cantidad de viajes que se hace a cada hora en el transcurso de un día cualquiera.

- Decidir sobre los pasos necesarios
  1. Partiendo de la matriz de viajes, podemos agrupar la cantidad de viajes de viajes por hora y por comuna de origen.
  2. Después realizar la división de los datos para entrenamiento, validación y prueba
  3. Creación y preparación del modelo
  4. Entrenamiento
  5. Validación del modelo y evaluación de este.