UNIVERSIDAD MAYOR DE SAN SIMÓN

POSGRADO FACULTAD DE CIENCIAS Y TECNOLOGÍA

MACHINE LEARNING

USA AIRPORTS

Presentado por:

Alvarez Ortiz Kevin Franco
Mamani Chambi Luis Eduardo

Docente:

MsC. Lesly Zerna Orellana

COCHABAMBA - BOLIVIA

II – 2019

INTRODUCCION

El nuevo y buen procesamiento de datos, HASTA ahora desconocido, puede utilizarse para guiar a los seres humanos en sus acciones. – Peter Naur (1974)

Este proyecto busca aplicar técnicas de machine learning a un set de datos que corresponde a los aeropuertos de Estados Unidos, este conjunto de datos es un récord de más de 3.5 millones de vuelos nacionales de EEUU desde 1990 hasta 2009. Se ha tomado del sitio web de OpenFlights que tiene una enorme base de datos de diferentes medios de viaje en todo el mundo.

OBJETIVO

Aplicar técnicas de Machine Learning para obtener un pronostico de la cantidad de pasajeros que realizan viajes en los EEUU.

DESARROLLO

A continuación detallamos los pasos que se siguieron para completar la practica final del módulo.

- Se escogio un dataset, tomando en cuenta el tipo de modelo de machine learning a aplicar.
- Se realizo un análisis exploratorio de datos para definir el tamaño del dataset, los tipos de datos de las columnas, la cantidad de valores nulos y tener información mas precisa sobre el dataset.
- Se realizo la limpieza de datos, convirtiendo valores al tipo de dato correspondiente, eliminando filas que carecían de cierta información relevante y creando nuevas columnas.
- Se realizo la visualización de información que era interesante y/o relevante, para poder entender un poco más el dataset. Además de la correlación entre variables.
- Se normalizaron los valores mas significativos, se evaluo la curtosis y asimetría de diferentes campos, sobre los cuales se aplicó métodos matemáticos para tener datos más concisos.

- Se aplicaron dos modelos de ML para ver cuál era más preciso para predecir la cantidad de pasajeros. Se hizo uso de la librería scikit-learn.
- Se aplicó más métodos de visualización sobre el resultado final para obtener conocimiento sobre el modelo creado.

El set de datos usado contiene las siguientes columnas después de haberse realizado la limpieza de datos:

- **Origin_airport**: código de aeropuerto de tres letras del aeropuerto de origen
- **Destination_airport**: código de aeropuerto de tres letras del aeropuerto de destino
- **Origin_city:** nombre de la ciudad de origen
- **Destination_city:** nombre de la ciudad de destino
- **Origin_state**: acrónimo del estado de origen
- **Destination_state:** acrónimo del estado de destino
- **Passengers**: Número de pasajeros transportados desde el origen hasta el destino.
- **Seats**: número de asientos disponibles en vuelos desde el origen hasta el destino
- **Flights**: número de vuelos entre el origen y el destino (registros múltiples durante un mes, muchos con vuelos> 1)
- **Distance**: Distancia (a la milla más cercana) volada entre el origen y el destino
- **Fly_date**: la fecha (aaaamm) del vuelo
- **Origin_population**: población de la ciudad de origen según lo informado por el censo de EE. UU.
- **Destination_population**: población de la ciudad de destino según lo informado por el censo de EE. UU.
- **Quantity_seats**: cantidad de asientos

BIBLIOGRAFÍA

https://www.kaggle.com/flashgordon/usa-airport-dataset