Challenge-Data-Scientist-Rolf-Traeger

Work about predicting the probability of delay of the flights that land or take off from the airport of Santiago de Chile (SCL). `delay_15` reflect if the flight was a delay or not.

Índice o Proceso de trabajo

Definición del problema

El problema consiste en predecir la probabilidad de retraso de los vuelos(representado por la columna `delay_15`) que aterrizan o despegan del aeropuerto de Santiago de Chile (SCL). Para eso tendrás un dataset con datos públicos y reales donde cada fila corresponde a un vuelo que aterrizó o despegó de SCL durante el 2017.

Para cada vuelo se encuentra disponible la siguiente información: Columns in `data\dataset_SCL.csv`:

```
"Fecha-I" : "Scheduled date and time of the flight"
"Vlo-I" : "Scheduled flight number"
"Ori-I" : "Programmed origin city code"
"Des-I" : "Programmed destination city code"
"Emp-I" : "Scheduled flight airline code"
"Fecha-O" : "Date and time of flight operation"
"Vlo-O" : "Flight operation number of the flight"
"Ori-O" : "Operation origin city code"
"Des-O" : "Operation destination city code"
"Emp-O" : "Airline code of the operated flight"
"DIA" : "Day of the month of flight operation"
"MES" : "Number of the month of operation of the flight"
"AÑO" : "Year of flight operation"
"DIANOM" : "Day of the week of flight operation"
"TIPOVUELO" : "Type of flight, I =International, N =National"
"OPERA" : "Name of the airline that operates"
"SIGLAORI" : "Name city of origin"
"SIGLADES" : "Destination city name"
```

Columns to make:

```
"high_season" : 'Binary variable that reflects if the flight was on a high demand
date or not'
"period_day" : 'Reflects the situation on the day the flight lands'
"min_diff" : 'Flight delay in minutes'
"delay_15" : 'Binary variable that reflects if the flight was a delay or not'
```

Librearias utilizadas

Comandos de instalación:

```
conda install plotnine
conda install pandas
conda install scikit-learn
conda install -c anaconda git
```

Data de estudio

Toda la data a utilizar se encuentra en la carpeta `...\data\`

Columnas a utilizar

Tomando en cuenta que existen varias columnas que reflejan lo mismo o informaicón que representa el futuro como `Fecha-O` o el atributo construido `min_diff`, se debe reducir las columnas que ensucien el proceso y que no influyan sobre `delay_15`.