National College of Ireland

Configuration Manual

MSc Research Project
Data Analytics

Nikhil Salvi
Student ID: X20179529

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Nikhil Salvi
Student ID:	X20179529
Programme:	MSc. Data Analytics **Year:** 2022-23
Module:	Research Project
Supervisor	**:** Dr. Christian Horn
Submission Due Date:	15th December 2022
Project Title:	Critical analysis on flight cancellations and predictive analysis on flight delays using automated machine learning
Word Count:	711 **Page Count:** 4

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Nikhil Salvi

Date: 15th December 2022

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Nikhil Salvi

X20179529

1 Introduction

This document represents the instructions to reproduce the classification of customer reviews for business development and to predict customer satisfaction. The steps and requirements for reproducing the machine learning models are as follows.

2 System Configuration

Hardware and software setup for the research work is explained below with respective diagrams.

2.1 Hardware configuration

For hardware configuration, MacBook air has been used. The specifications are, apple M1 chip, 8 core CPU, (4 performance and 4 efficiency), memory 8 GB with 512 GB storage.

Figure 1: Hardware configuration

2.2 Software configuration

For software configuration two software are used, like jupyter notebook, and a hosted jupyter notebook of google collaboratory. Figure 2 shows the version of jupyter notebook that has been used with the help of Anaconda Navigator.

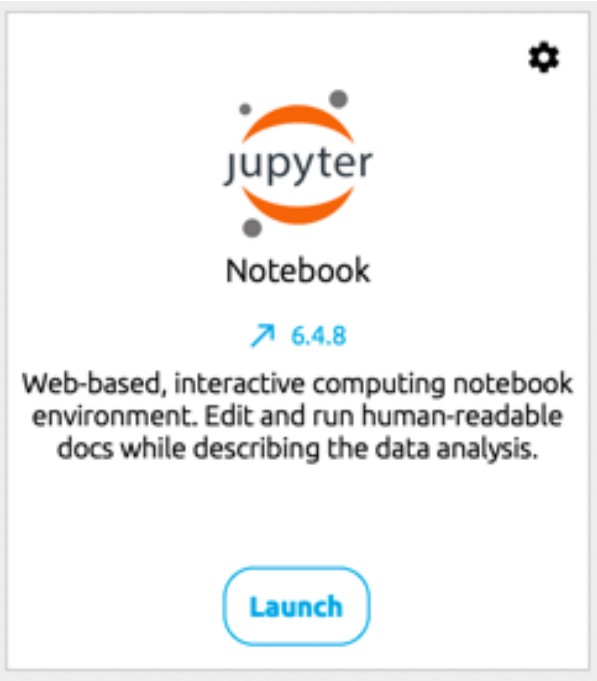

Figure 2: Jupyter notebook

The initial process of creating a sample data from population data is completed using jupyter notebook. Google colaboratory is used further to build analytical and predictive models. Figure 3 shows the python version used for this project, 3.8.16.

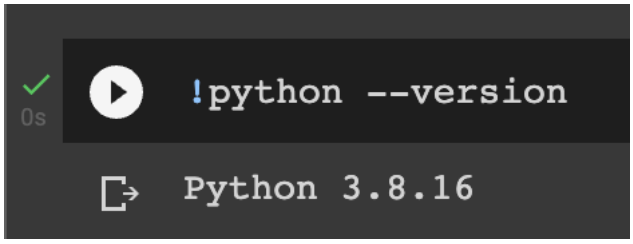

Figure 3: Python version

3 Implementation

3.1 Data Source

The data source is acquired from an open source, Kaggle.com. The link is below:
https://www.kaggle.com/datasets/divyansh22/flight-delay-prediction
The data source has 3 datasets. One dataset contains information of all the airlines, another data set has the information of all the airports, and the last data is the main data of all the flights.

3.2 Feature Engineering

For feature engineering, pandas, and numpy, modules are used. In the process of EDA, data is visualised using seaborn, matplot and GeoPandas libraries. Below figure 4 shows all the libraries used in this project.

```
import geopandas as gpd
import pandas as pd
import matplotlib.pyplot as plt
from shapely import geometry
from matplotlib.text import FontProperties
import seaborn as sns
import numpy as np
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from imblearn import over_sampling, under_sampling
from imblearn.over_sampling import RandomOverSampler
from pycaret.classification import *
```

Figure 4: Required python libraries and modules.

The dataset “airports.csv” contains the latitudes and longitudes of all the airports. Using pandas, the “airports.csv” dataset and “flights.csv” datasets are merged (left join). This dataset is the master dataset, which has been used for the entire project.

4 Visualisation

Using Geopandas module and the data of longitudes and latitudes of the airports, the locations of the airports are plotted on the map of North America continent. The figure 5 below shows the plot.

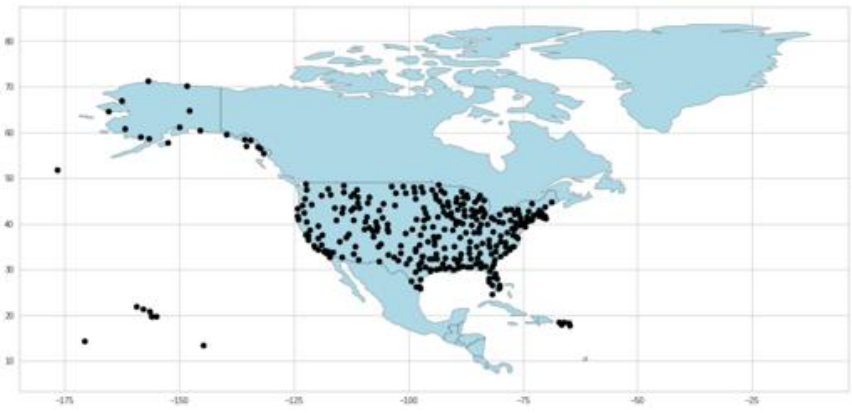

Figure 5: GeoPandas plot of locations of all the airports in dataset

To transform the data, lable encoding is used. The figure 6 shows the transformed data.

	index	MONTH	DAY	DAY_OF_WEEK	AIRLINE	TAIL_NUMBER	TAXI_OUT	WHEELS_OFF	AIR_TIME	DISTANCE	WHEELS_ON	TAXI_IN	ARRIVAL_DELAY
0	0	6	29	3	13	3848	8.0	700.0	45.0	308	745.0	4.0	-21.0
1	1	10	27	0	13	3834	15.0	648.0	132.0	1164	1100.0	5.0	-20.0
2	2	1	22	0	9	1609	15.0	1118.0	87.0	541	1145.0	8.0	-2.0
3	3	1	21	6	0	1866	41.0	2126.0	99.0	802	2305.0	9.0	77.0
4	4	7	18	2	3	331	43.0	1824.0	251.0	1947	1935.0	11.0	36.0
5	5	10	28	1	13	2195	7.0	1649.0	77.0	404	1806.0	4.0	0.0
6	6	6	19	0	0	3488	16.0	1413.0	77.0	500	1530.0	10.0	0.0
7	7	3	22	3	13	3846	11.0	1555.0	226.0	1488	1641.0	8.0	14.0
8	8	5	9	2	9	1669	21.0	1142.0	53.0	331	1235.0	4.0	-4.0
9	9	10	25	5	3	4307	12.0	1835.0	169.0	1306	2024.0	3.0	-18.0
10	10	6	28	2	7	2681	17.0	1047.0	26.0	134	1113.0	3.0	-11.0
11	11	3	23	4	4	3165	17.0	1644.0	26.0	143	1710.0	5.0	44.0
12	12	0	27	2	9	4046	21.0	832.0	75.0	522	947.0	4.0	-17.0
13	13	10	26	6	7	2589	14.0	2308.0	69.0	396	17.0	3.0	68.0
14	14	6	17	5	0	1847	23.0	815.0	50.0	258	905.0	18.0	13.0

Figure 6: Lable encoding

5 Model building and Evaluation

For model building, automated machine learning PyCaret is used. This model will train the data using many classification algorithms, and based on accuracy, it will identify the best algorithm. The below figure 7 is the final score table of the model.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lr	Logistic Regression	0.9608	0.9862	0.8526	0.9239	0.8867	0.8631	0.8642	25.596
svm	SVM - Linear Kernel	0.9596	0.0000	0.8467	0.9228	0.8826	0.8583	0.8597	6.011
dt	Decision Tree Classifier	0.9040	0.8388	0.7372	0.7312	0.7341	0.6755	0.6756	383.528
knn	K Neighbors Classifier	0.8946	0.8482	0.4657	0.8999	0.6136	0.5596	0.6005	875.313
ridge	Ridge Classifier	0.8925	0.0000	0.4083	0.9853	0.5772	0.5275	0.5951	8.103
nb	Naive Bayes	0.4878	0.5445	0.5926	0.1955	0.2940	0.0322	0.0443	3.044

Figure 7: Evaluation