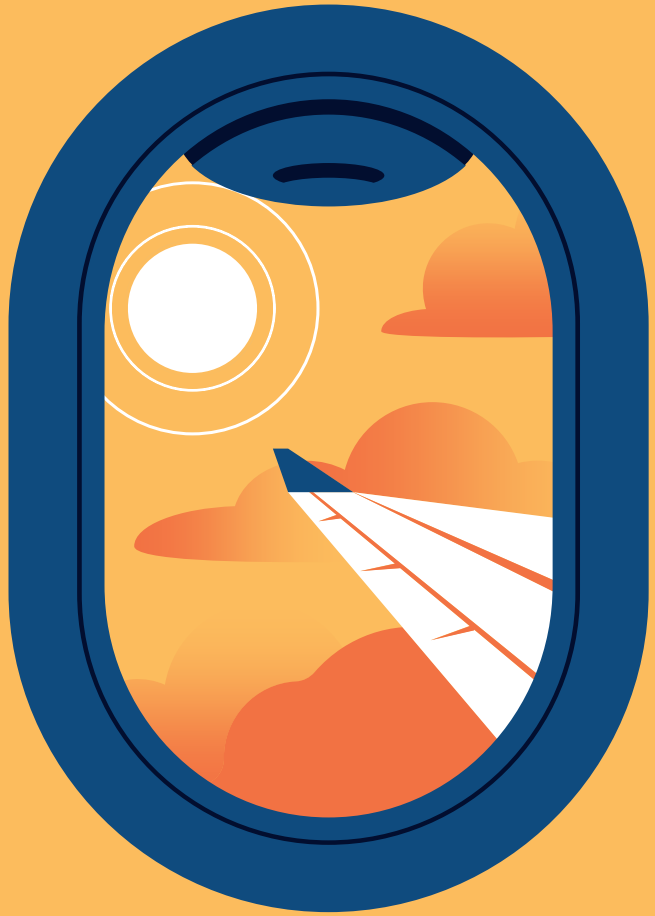
T5 - Data Science BootCamp Project

Airline Passenger Satisfaction

Presented by: Mohammed Ahmed Faraj

INTRODUCTION

This project was implemented in the Data Science BootCamp at SDAIA Academy.
This project aims to satisfy travelers with their travels. The data used in this project is provided by Kaggle, and the data is categorized using the Sentiment Density Analysis tool, with a trained random set from the sklearn library and an accuracy of 97%. The flow is used to create an interactive dashboard to visualize and communicate the end results.



Ambition

Riyadh season

This project is an example of my future work, which is to study the happiness of visitors in the Riyadh season, and it helps in the quality of life program.

Suggest

I will develop the work in the future using new algorithms such as: SVM or FM to study users behavior and find out a suggestion according to the interest of each visitor.

Project Goal

The goal of this project was to use binary classification models to predict the type of cover type in different wilderness areas with good accuracy.

Data Exploration

Datasets

Datasets contain 103904 instances

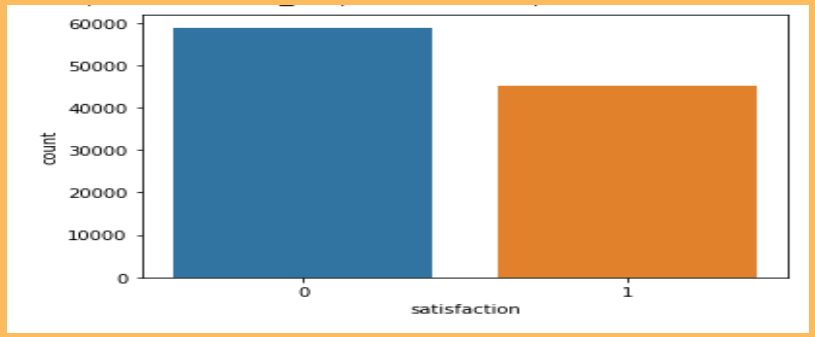
Datasets contain 25 Attribute

Datasets has 3 classes for Satisfaction

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location
0	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3	1
1	1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	3	3
2	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2	2
3	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	5	5
4	4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	3	3

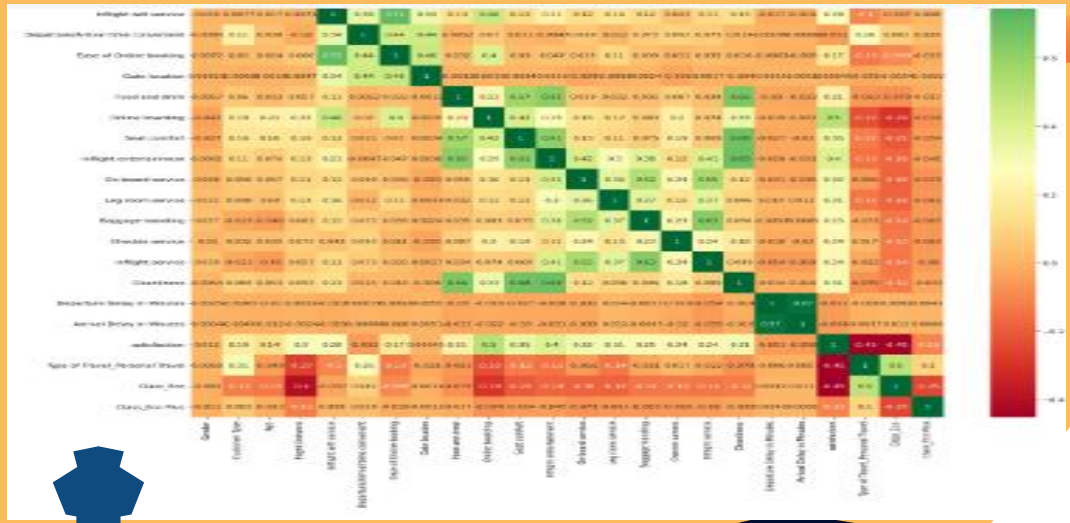
Data Exploration

- For each class label, i display the code of the class and the name of that class.
- Summaries the class distribution using a suitable graph
- Display a statistical summary for all the attributes



Data Processing

1. Check if null value exists
2. Check duplicated data
3. Set outlayers (IQR)
4. (pearson correlation) feature is most important for our dataset (for training)



Model

train_accuracy 0.991802
validation_accuracy 0.86947
test_accuracy 0.8723

Decision Tree Classifier

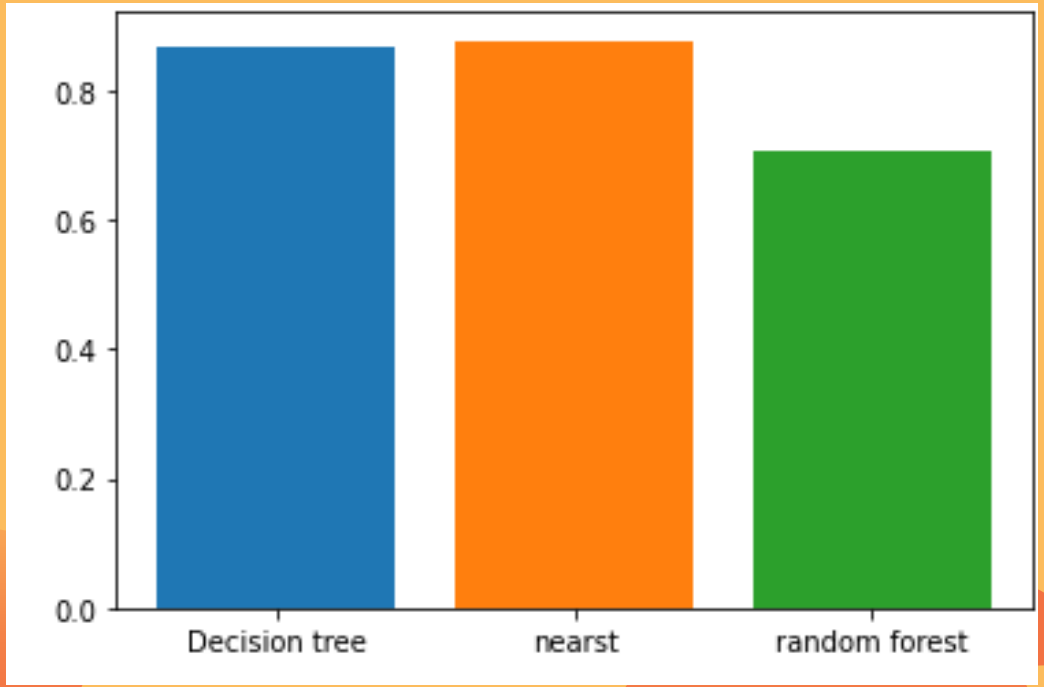
train_accuracy 0.938189
validation_accuracy 0.881621
test_accuracy 0.882300807

Nearest Neighbour

train_accuracy 0.700813
validation_accuracy 0.697912
test_accuracy 0.700042

Random Forest Classifier

Results
Algorithm Comparison



Thank you ☺

Mohammed Faraj