CLASSIFICATION OF AIRLINE PASSENGER REVIEWS

CS 508 – Project Report

"We've been rolling out a suite of customer self-service tools, which integrate technology in all aspects of travel and improve the customer experience as we work to reduce our costs."
-Robin Hayes, CEO of JetBlue Airways.

"Our focus will be on continuing to improve customer service and expanding United's network to offer customers more choice."

-Scott Kirby, President of United Airlines.

Sri Krishna Sreedhar | Anik Alvi | Rahul Chowdary Garigipati

1. Introduction:

Customer satisfaction is always top of mind for airlines. Unhappy or disengaged customers naturally mean fewer passengers and less revenue. It's important that customers have an excellent experience every time they travel. On-time flights, good in-flight entertainment, more snacks, and more legroom might be the obvious contributors to a good experience and more loyalty. This project deals with the Airline Passenger reviews data. Whenever a passenger takes a particular flight, the airline company prefers to take the review of inflight services from the passengers, which helps the company to improve its services. If the reviews provided by the passengers are bad, then there is a high chance of losing their customers if they don`t act upon their bad services. The airline company uses the data from the passengers and finds which service is better among all the services and might want to advertise about it. In this way, the reviews from the passengers help the companies to maintain a good and respectable name and stay in the growing competition.

2. Problem Definition:

Once, they have collected the reviews through various kinds of feedback forms, if the airline companies manually check each and every review by the customers, it takes them a lot of time and manpower and also it might not be the best way to check and analyse the data. So, in-order to reduce the effort and manpower and to analyse the data in the best way possible, they might want to find a better approach in classifying the collected reviews into positive and negative ones. So, that they can respond quickly based on the number of positive and negative reviews.

3. Motivation:

From the data mining techniques that were taught to us in this class, we thought classification models can be solution for this problem and is the best approach that can classify millions of reviews in minutes of time. Also, if there are text data reviews, sentiment analysis is the approach that can classify the data into positive and negative. These machine learning models won't take much time to provide the results to the company and with high accuracy rates.

4. Solution:

Reviews can be collected in multiple ways by Airline companies from the passengers. Few companies will build an application with various services mentioned that are provided during travel and will ask the passenger to rate each of their services. In this case, huge amounts of statistical data will be collected by the company. The other case is where they give a textbox in their application and ask passengers to give them feedback based on their experience. Sometimes the passenger itself might tweet about the experience after travelling. In either of the case, the review is in a text format and will have to deal with lots of text data.

Thus, the below is the explanation of how we dealt with both kinds of data and our observations:

4.1. For Statistical Data:

4.1.1 Data Collection:

We downloaded the statistical dataset from the Kaggle Website and the link is provided below:
https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction?select=test.csv

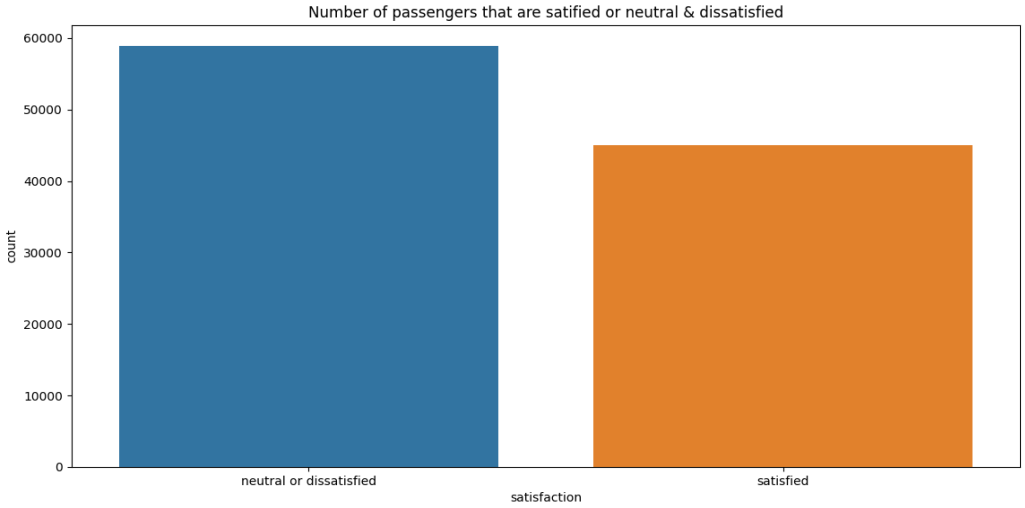
There are two datasets available (Train.CSV and test.CSV) train.CSV to train the classifying model and test.CSV to test the model. The dataset has 25 different columns including the class label that must be predicted. Training dataset has 1,03,905 rows of data and the testing dataset has 25,977 rows of data.

4.1.2 Data Pre-processing:

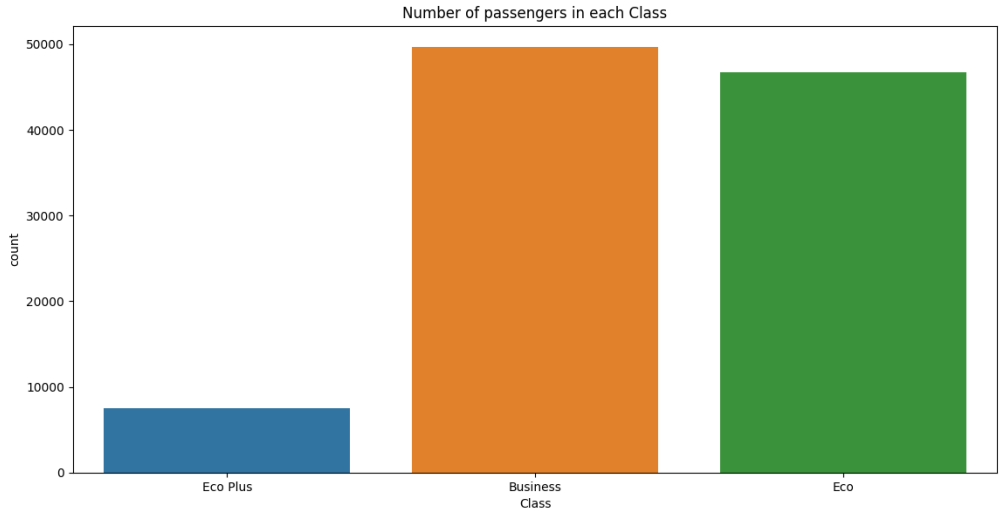
One of the most important tasks in every machine learning task is the data pre-processing section. In our project, few attributes had missing values. To resolve this problem, we replaced these missing values by the average value of that attribute since the real value is not known. Normalization was also required since many of the attributes had different values and needs to be scaled to a particular range. For this, we used Min Max Scaler which transform features by scaling each feature to a given range. e.g., between zero and one. Moreover, attribute ID of customer was also removed from the dataset since it is not related to the classification problem we are trying to solve.

4.1.3. Data Analysis:

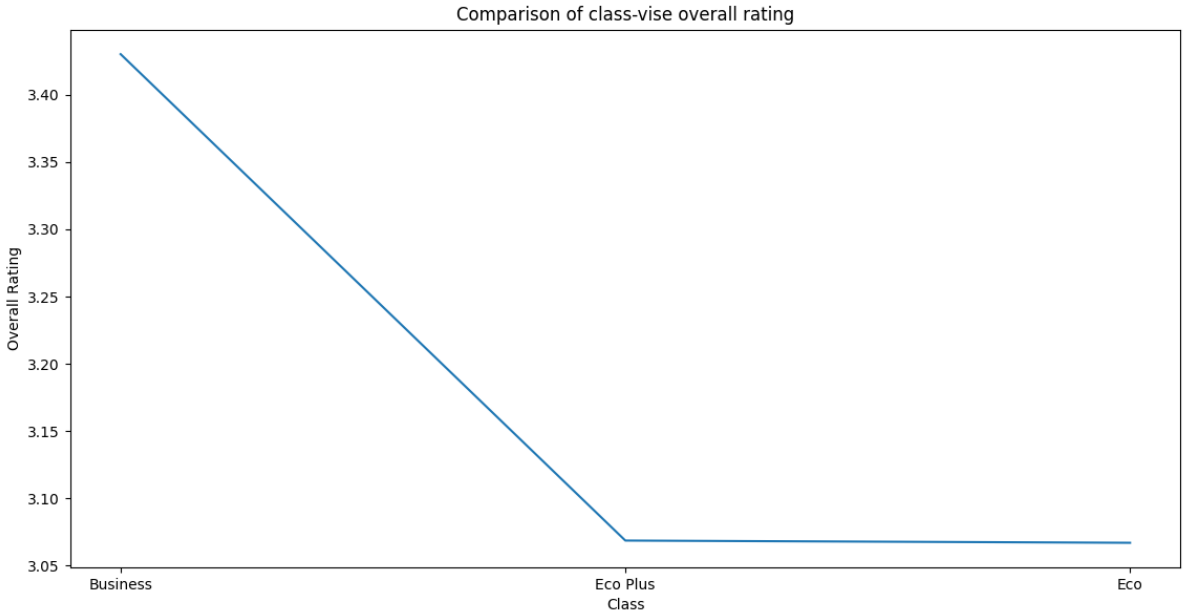
Using the data, the analysis has been done to find out the relation between various columns and compared how the columns are related.
The below is the plot to show the count of passengers that are classified into 'satisfied' and 'neutral or dissatisfied'.



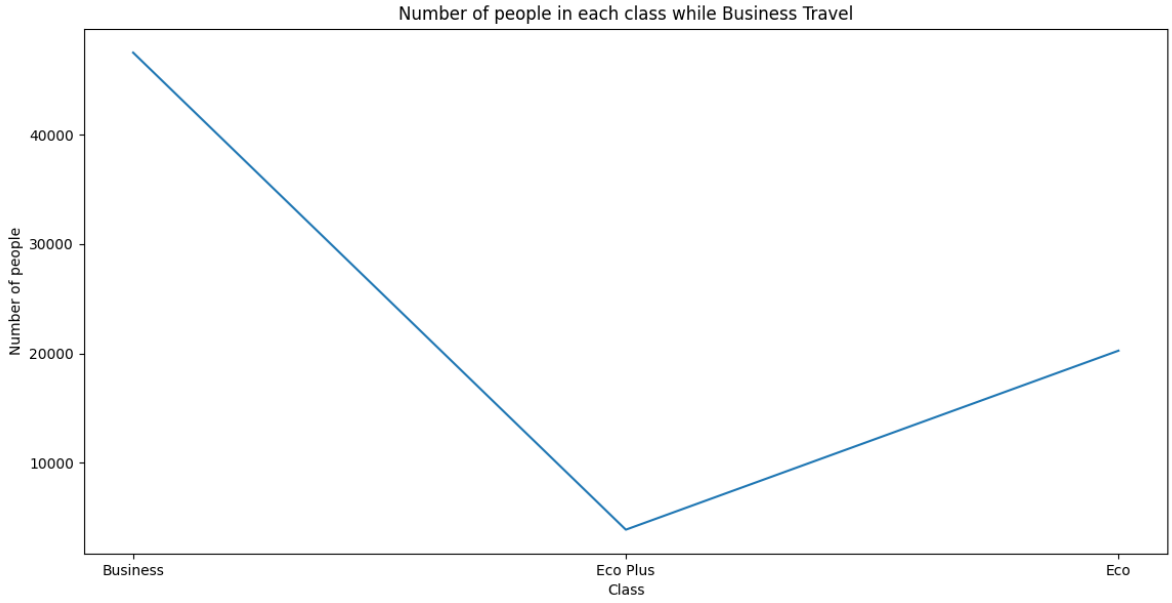
The below is the plot to show the count of passengers that preferred to travel in three available classes of the flight that are Business, Eco Plus and Economy.

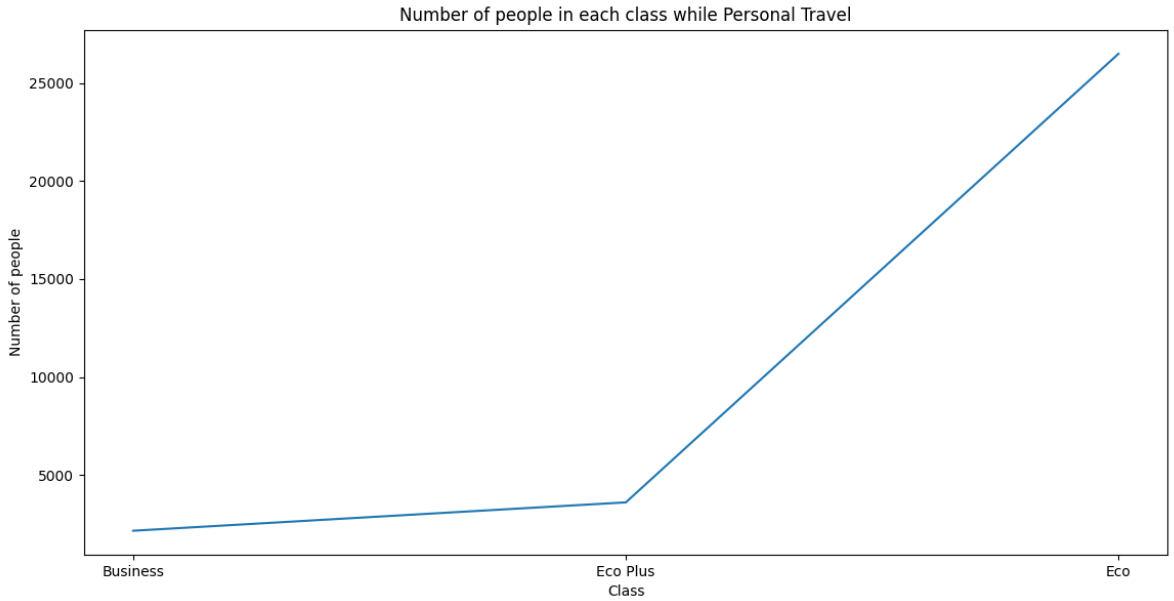


The below graph is to show the overall rating given by each passenger from three different classes. It's through taking the mean of all services ratings and again calculating the mean of all means for each class and seems like business class have provided better services when compared among all three classes.



The next two plots are to show how many passengers preferred to travel in what classes based on the purpose of their travel. From the below plots we can tell that, while travelling on business purpose, most of the passengers preferred to travel in business class and while travelling on person reasons, most of them chose to travel in Economy class.



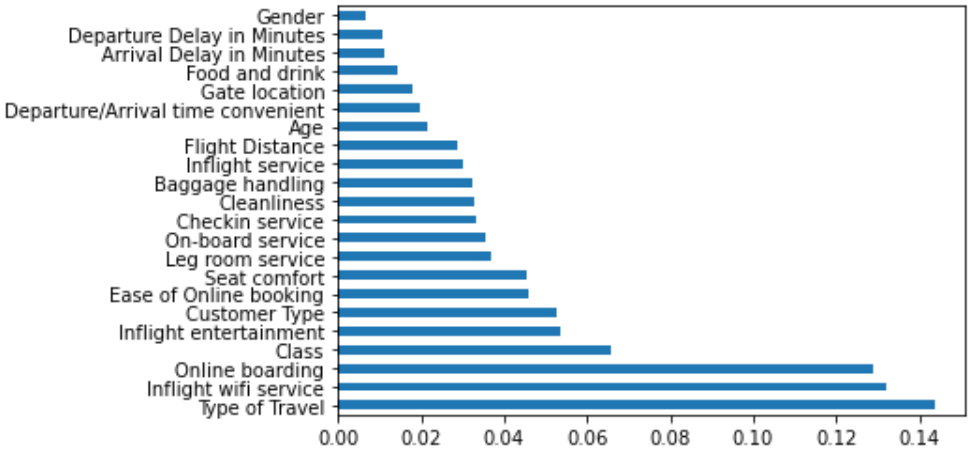


4.1.4 Feature Analysis:

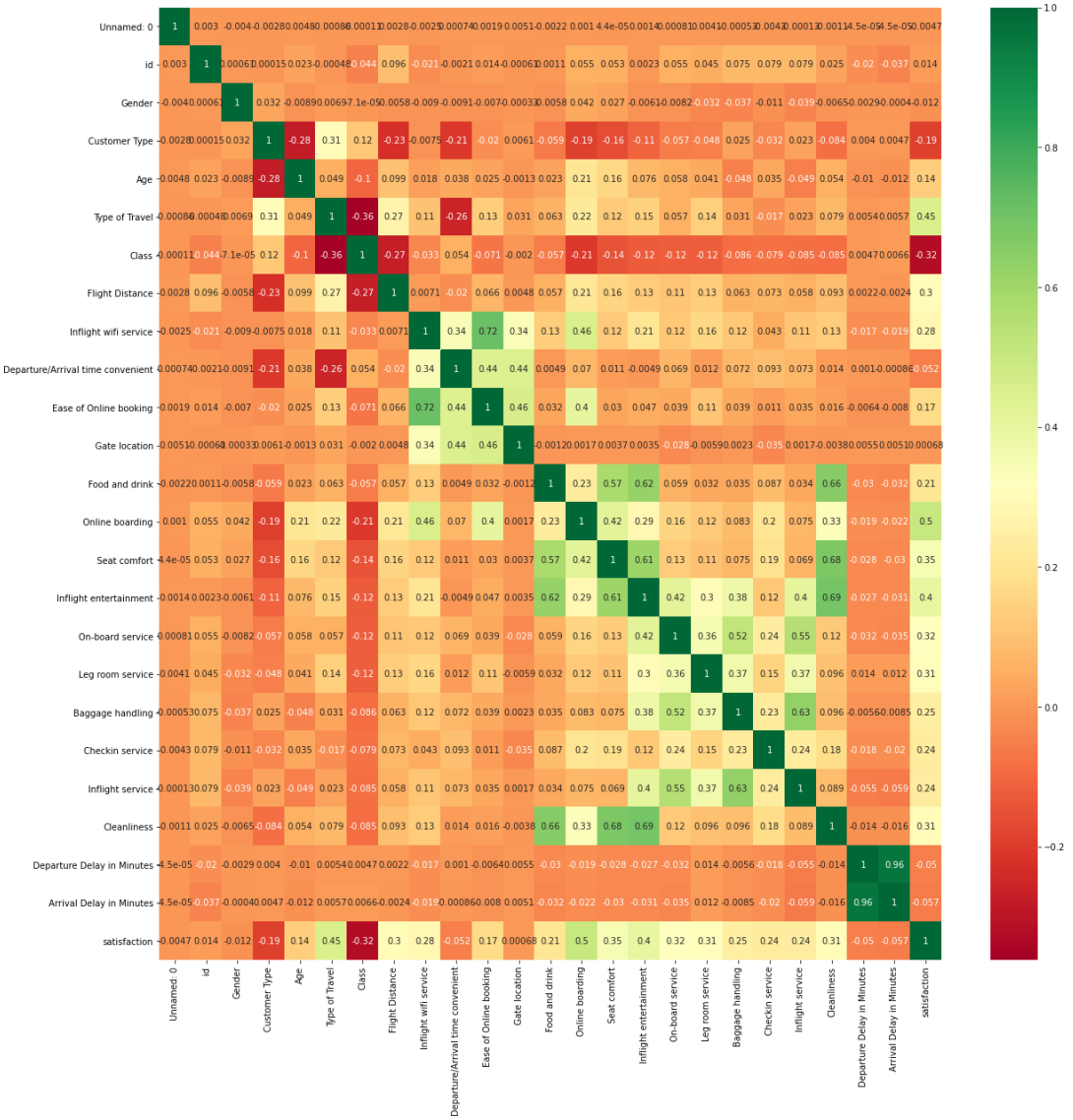
Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Few irrelevant features in your data can decrease the accuracy of the models. Benefits of performing feature analysis and picking the most important attributes in the dataset are it reduces overfitting, reduces training time and improves accuracy.

There are three different feature selection techniques implemented to select 12 most influencing features among the 22 columns. First technique of them is the **Univariate selection**, in this technique, scikit-learn library provides the **SelectKBest** class that can be used with a suite of different statistical tests to select a specific number of features that shows huge influence on the model when compared to other features.

In the next technique **Feature Importance**, an inbuilt class that comes with Tree Based Classifiers, we will be using Extra Tree Classifier for extracting the required number of influencing features. It gives you a score for each feature of your data, the higher the score more important or relevant is the feature.



The third technique is **Correlation Matrix with Heatmap,** Correlation states how the features are related to each other.it can be positive or negative. Heatmap makes it easy to identify which features are most related to the target variable.



From all the three feature analysis techniques, the top 12 influencing features has been selected and created a dataset. The important features are Customer Type, Type of Travel, Class, Inflight Wi-Fi Service, Ease of Online booking, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg-room service, Baggage Handling, Cleanliness.

4.1.5 Classification Models:

Decision Tree: Decision tree builds classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The result is a tree with decision nodes and leaf nodes.

KNN Classifier: K-Nearest Neighbour is a supervised machine learning model. This algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

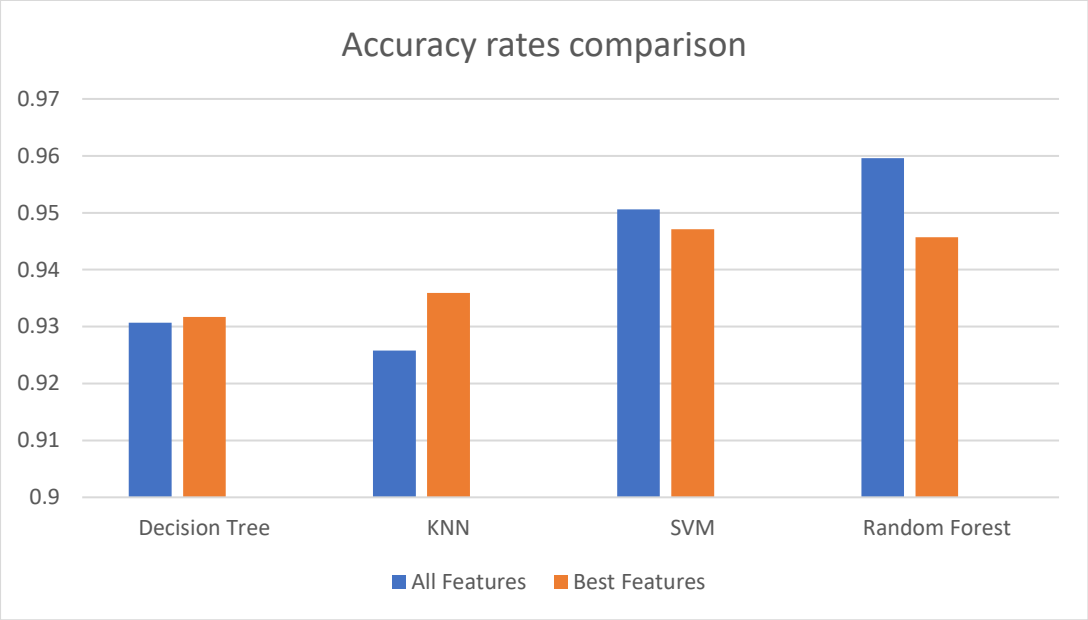
SVM Classifier: The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. It classifies the data by finding the maximum distance between two class labels.

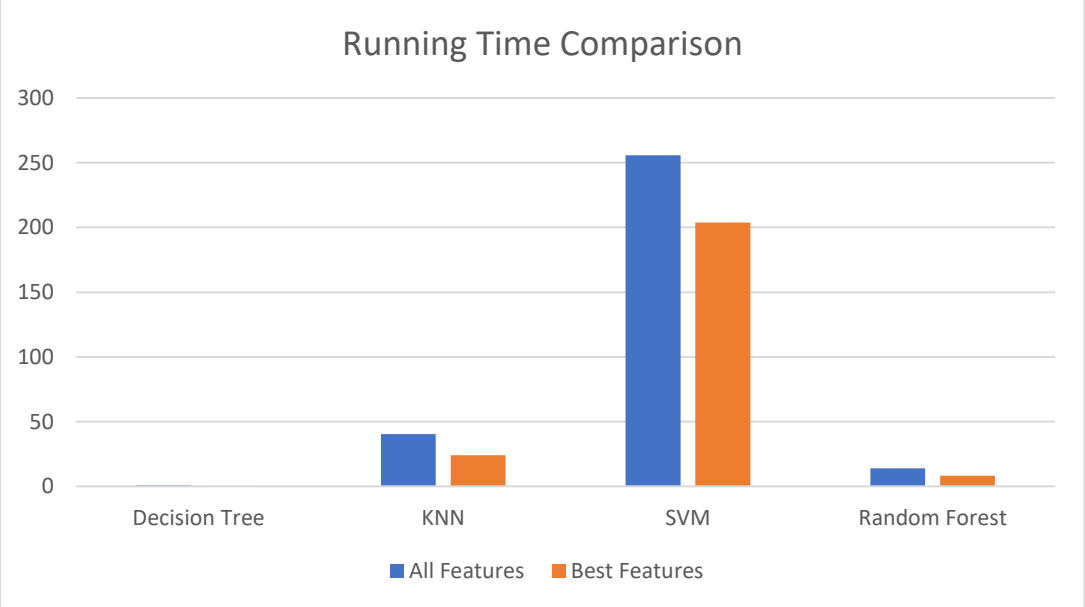
The next classifier is something not taught in this course, but we implemented to try a model of our own:

Random Forest: This algorithm is a combination of decision tress working as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. Trees protect each other from their individual errors and that makes this model more powerful. Random forest takes advantage by allowing each individual tree to randomly sample from the dataset with replacement, resulting in different trees. This process is known as bagging and it results in prediction by committee is more accurate than that of any individual tree.

We ran the whole dataset and also the dataset with best features derived from feature selection and trained the model using both datasets and tested the performance.

4.1.6 Results:





4.1.7 Observations:

The accuracy rates when compared between models Random Forest classifier has performed the best of all and when compared between the datasets, best features dataset derived from the feature selection process and whole dataset, best features performance was not bad when compared to all features as best features accuracy was more than all features accuracy in two classifiers and almost equal in the other two models.

Comparing running times of all 4 algorithms, SVM model took the highest time and Decision tree took the least and was almost negligible when compared to others. When compared between datasets, best features dataset took less time to run for all 4 classifier models.

4.2 For Text Data:

4.2.1 Data Collection:

For text data, we have done **web scraping** and collected tweets from twitter targeting the airline reviews search word and collected 2000 tweets. The process for web scraping from twitter is as below:

- ➔ Though there are many ways to scrap data from websites, the most efficient way to collect data from twitter would be through using twitter API library **Tweepy**.
- ➔ For using the Twitter API, you need to have a developer access Twitter account. Request for the same it might take 1–2 days to get an approval after answering several questions about your project to twitter team. Once, you're done with the set up create an app, in it, you will get Keys and tokens, which will help us retrieve data from Twitter. They act as login credentials.
- ➔ Next step is to extract tweets by giving a search word and it'll collect data based on word that was mentioned. Basically, it will extract tweets that contain the word which is valid for our project.

4.2.2 Data pre-processing:

Once the data is collected, they need to be pre-processed as the tweets contains retweets, account mentions, hashtags and hyperlinks. To remove this data, we have matched the pattern using regular expressions and cleaned it from that sentence. Then the cleaned data is added to a data frame.

4.2.3 Classification of Data:

For classifying the text data, **Sentiment Analysis** technique is used. TextBlob is the library which is used for **Natural Language Processing**. From textblob, the polarity and subjectivity can be calculated for each sentence.

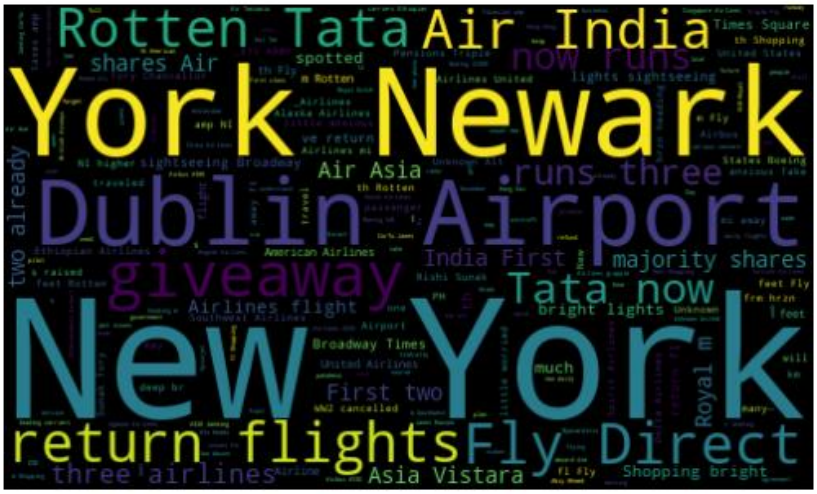
Polarity: It is a score, that says if a sentence is a positive or negative sentence. It is calculated by finding the polarity score of each word from sentiment dictionary and then calculating the average of all words in that sentence. Polarity scores could range between -1 to 1.

Subjectivity is about how deep meaning the sentence carries about the search word and how personal an opinion is. If the subjectivity score is less, that means that sentence is meaningless.

A statement is classified into a positive review or a negative review based on the polarity score. If polarity score is greater than 0, it means it is a positive review and if the score is below zero, the statement is a negative review.

4.2.4 Results:

WordCloud is a library that's used for generating a wordcloud image in which we can see the most frequent words in our collected data. That means the words shown in that image or the most targeted words in the tweets. So, if they are positive words, the number of positive tweets could be more and vice versa.



For the search word 'airline Passengers', if the total tweets collected are 2000, the number of tweets classified into each class are:

Class	Number of tweets
Positive	1074
Negative	166
Neutral	760

Running times:

Task	Time taken
Collecting Data	197.256 seconds
Cleaning Data	0.008 seconds
Classifying Data	1.443 seconds

5. References:

https://blogs.perficient.com/2018/05/14/customer-satisfaction-in-the-airline-industry/

https://towardsdatascience.com/extracting-data-from-twitter-using-python-5ab67bff553a

https://medium.com/red-buffer/sentiment-analysis-let-textblob-do-all-the-work-9927d803d137

https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e

https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

https://towardsdatascience.com/understanding-random-forest-58381e0602d2