Airline Passenger Satisfaction

A Project Submitted in Partial Fulfillment of the Requirements of Passing for the Course of

CSI 382 - Data Mining and Knowledge Discovery

By

Sayeda Sanzida Ferdous Ruhi

ID: CSE 072 08182

Department of Computer Science and Engineering

STAMFORD UNIVERSITY BANGLADESH

12th December, 2024

Abstract

The problem domain is about Airline Passenger Satisfaction. This domain has 24 attributes which has 129880 instances each. The last attribute of this domain is the target attribute. And the rest attributes are connected to the target. But the values of these attributes need to customize. So I applied some normalization techniques to normalize the values. And also dropped some attributes that are irrelevant. After applying techniques I found accuracy for each method and found some errors. So I think there is a great scope to remove those errors.

Acknowledgement

I would like to extend my appreciation to the following individuals, institutions, and organizations for their valuable contributions to the Airline Passenger Satisfaction Dataset.

And also to our respected course teacher, Adnan Ferdous Ashrafi, Senior Lecturer Department of CSE, Stamford University Bangladesh.

Table of Contents

1 Introduction

Airline passenger satisfaction is a critical aspect of the aviation industry that encompasses the overall contentment and fulfillment of passengers with the services provided by an airline. It is a multifaceted concept that goes beyond merely reaching the destination safely. Passenger satisfaction is influenced by various factors throughout the entire travel experience, from booking a ticket to disembarking at the final destination. Airlines recognize the importance of prioritizing passenger satisfaction not only as a means to retain customers but also to enhance their reputation and competitiveness in a highly dynamic market.

In this domain, we are trying to figure out satisfaction level of airline passengers through various factors. By doing this, airline agencies can figure out what features best satisfied their customers and what features take away their customers.

Though data is available as many people travel through airlines, but it is quite hard to implement its implications as satisfying customers depends on various attributes and while implementing, airline agencies need to consider all of them.

2 Dataset Description

Creating a dataset for airline passenger satisfaction involves collecting and organizing relevant information that reflects the various aspects of the passenger experience. The dataset aims to analyze factors influencing passenger satisfaction with airline services, facilitating insights into areas of improvement for airlines and the aviation industry as a whole.

2.1 Dataset Source

This dataset contains an airline passenger satisfaction survey. The source link: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

2.2 Dataset Attributes

The dataset contains 24 attributes. They are:

- Gender: Gender of the passengers (Female, Male)
- Customer Type: The customer type (Loyal customer, disloyal customer)
- Age: The actual age of the passengers
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)

- Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)

- Flight distance: The flight distance of this journey
- Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
- Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
- Ease of Online booking: Satisfaction level of online booking
- Gate location: Satisfaction level of Gate location
- Food and drink: Satisfaction level of Food and drink
- Online boarding: Satisfaction level of online boarding
- Seat comfort: Satisfaction level of Seat comfort
- Inflight entertainment: Satisfaction level of inflight entertainment
- On-board service: Satisfaction level of On-board service
- Leg room service: Satisfaction level of Leg room service
- Baggage handling: Satisfaction level of baggage handling
- Check-in service: Satisfaction level of Check-in service
- Inflight service: Satisfaction level of inflight service
- Cleanliness: Satisfaction level of Cleanliness
- Departure Delay in Minutes: Minutes delayed when departure
- Arrival Delay in Minutes: Minutes delayed when Arrival
- Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

2.3 Chapter Summary

The data set was modified from this dataset by John D. It has been cleaned up for the purposes of classification.

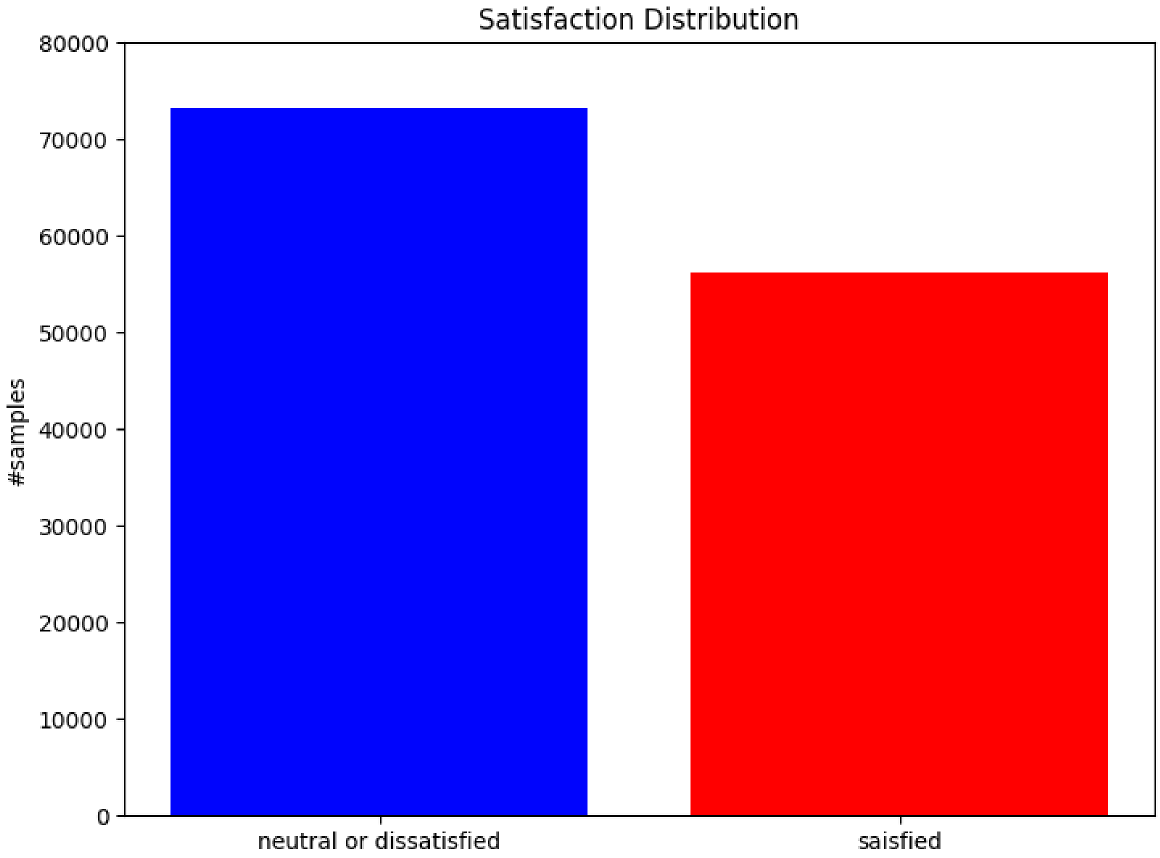
3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical phase in the data analysis process that involves visually and statistically summarizing, interpreting, and understanding the main characteristics of a dataset. The primary goal of EDA is to gain insights into the underlying patterns, relationships, and trends within the data before formal modeling or hypothesis testing takes place. By exploring the data in a systematic way, analysts and data scientists can make informed decisions about subsequent steps in the analysis and better understand the story the data is telling.

3.1 Histogram

In order to understand shape of a dataset, we plot histograms. Histograms shows count of different categories in an attribute.

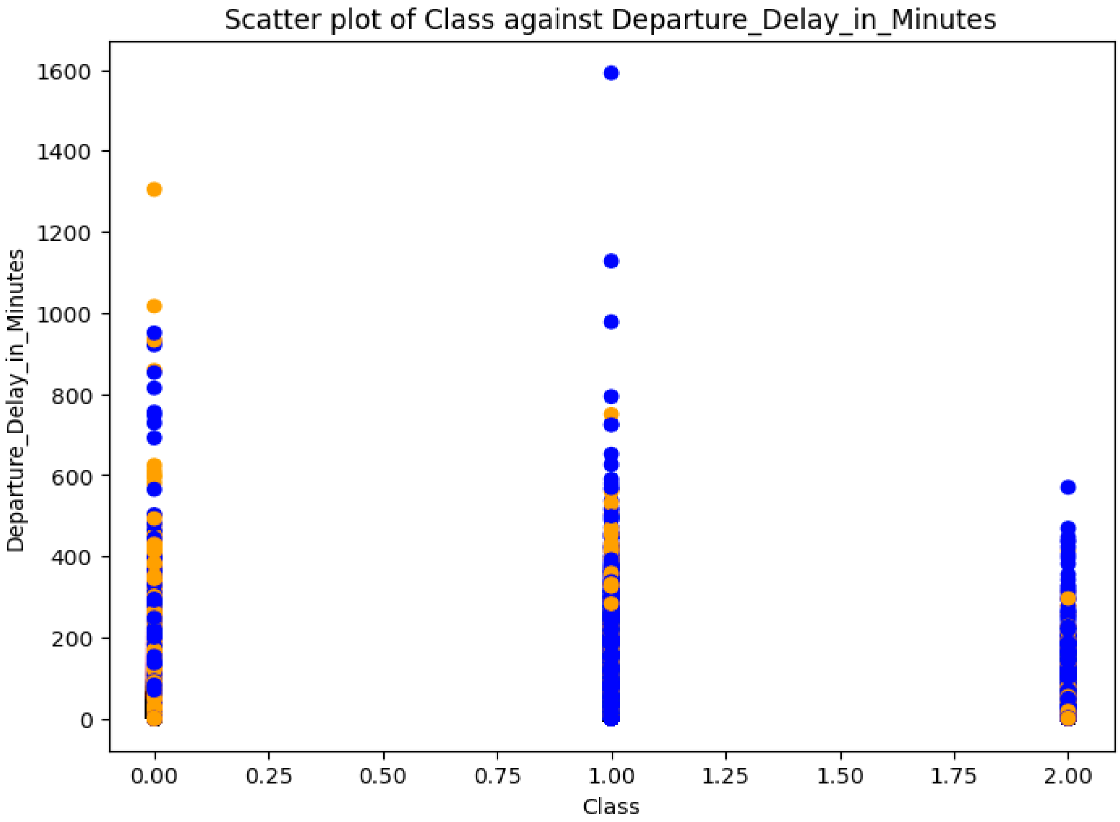
The histogram given below shows total number of satisfied and neutral or dissatisfied passengers. And we can see that there are more neutral or dissatisfied passengers than satisfied passengers.

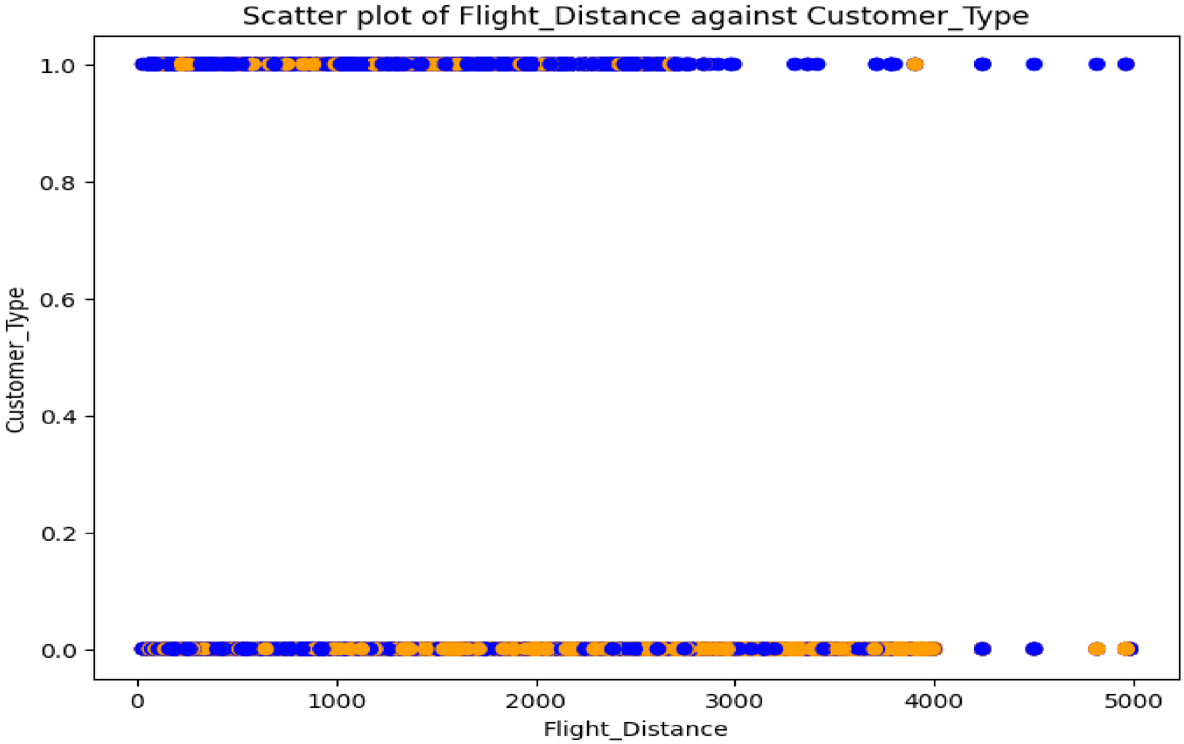


3.2 Scatterplots

We plot scatterplots to visualize relationships between variables. It also shows if there is any outliers in the attribute. Some scatterplots showing relationships and outliers are given below:

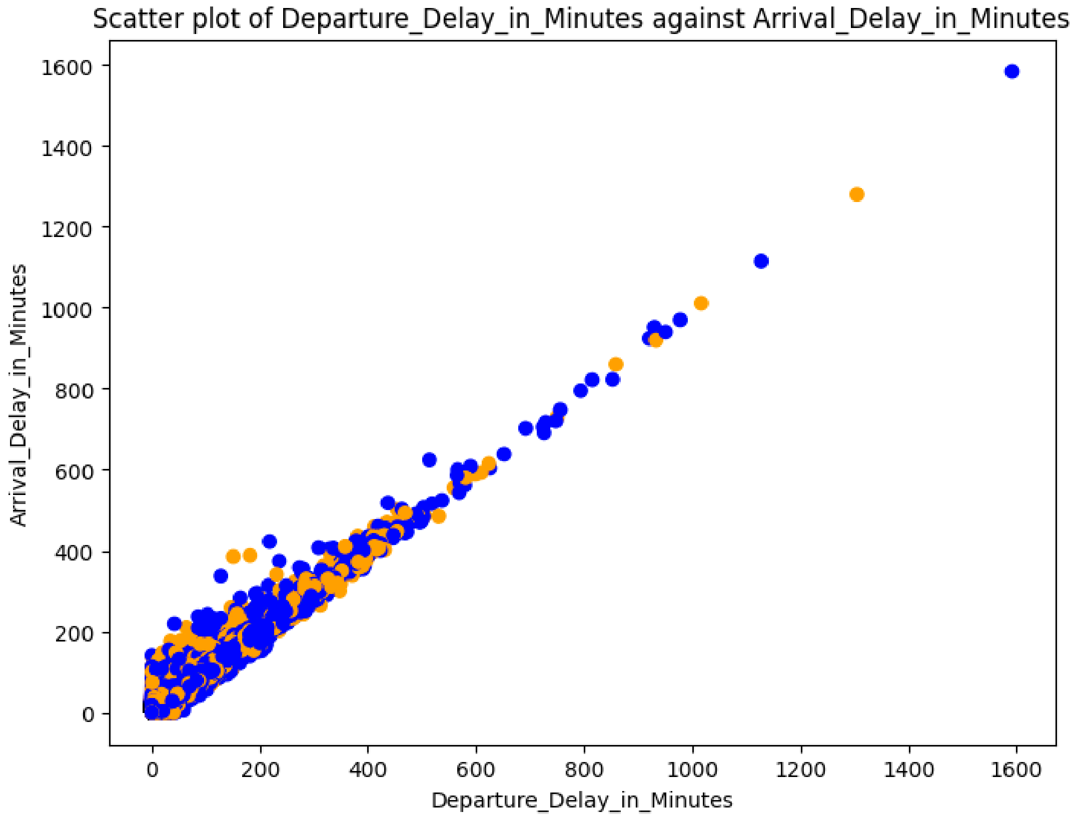
- This scatterplots shows relationship between Departure Delay in Minutes and Class. And also shows some outliers. Here in this scatterplot, we find 5 outliers.


Scatter plot of Class against Departure_Delay_in_Minutes
Departure_Delay_in_Minutes
Class
0
200
400
600
800
1000
1200
1400
1600
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

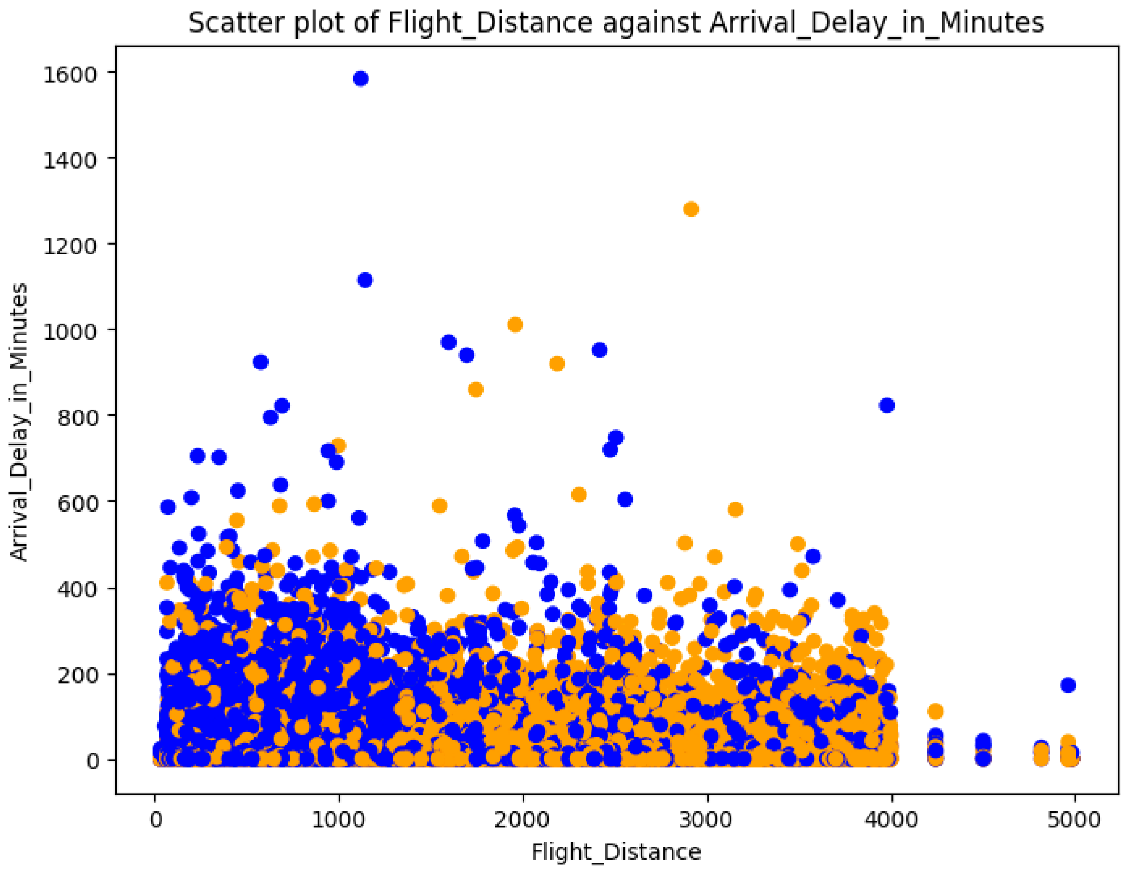

Scatter plot of Flight_Distance against Customer_Type
Customer_Type
Flight_Distance
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
5000

- This scatterplots shows relationship between Customer Type and Flight Distance. And also shows some outliers. Here in this scatterplot, we find 9 outliers.

- This scatterplots shows relationship between Departure Delay in Minutes and Arrival Delay in Minutes. And also shows some outliers. Here in this scatterplot, we find 3 outliers.

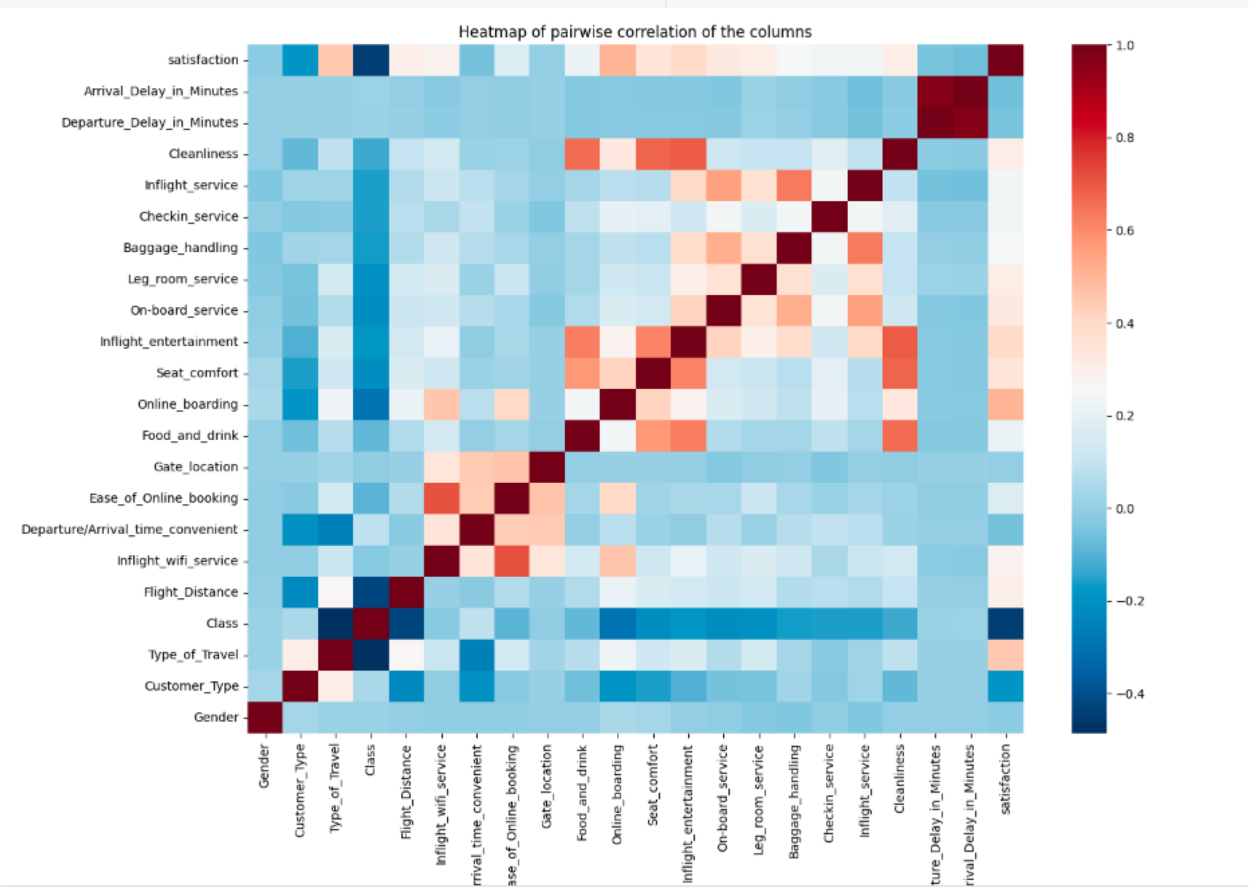


- This scatterplots shows relationship between Flight Distance and Arrival Delay in Minutes. And also shows some outliers. Here in this scatterplot, we find many outliers but most recognizable are 3 outliers.



3.3 Co-related Variables

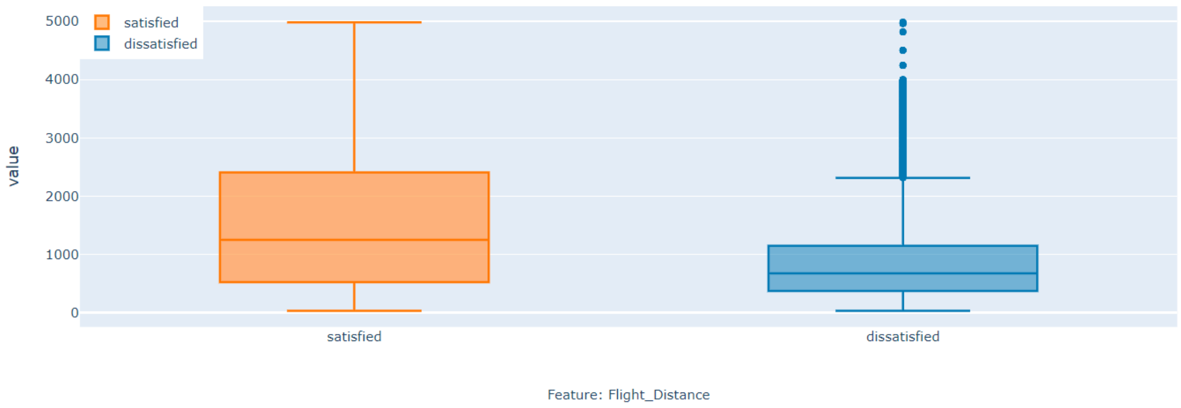
Correlated variables refer to two or more variables that have a statistical relationship with each other. The nature of this relationship can be positive, negative, or neutral, and it is typically measured by correlation coefficients. Correlation is a statistical technique that quantifies the strength and direction of the linear relationship between two continuous variables. Here we provide a correlation matrix to show relationships between variables. And after finding close relation we eliminate one attribute from all attributes.



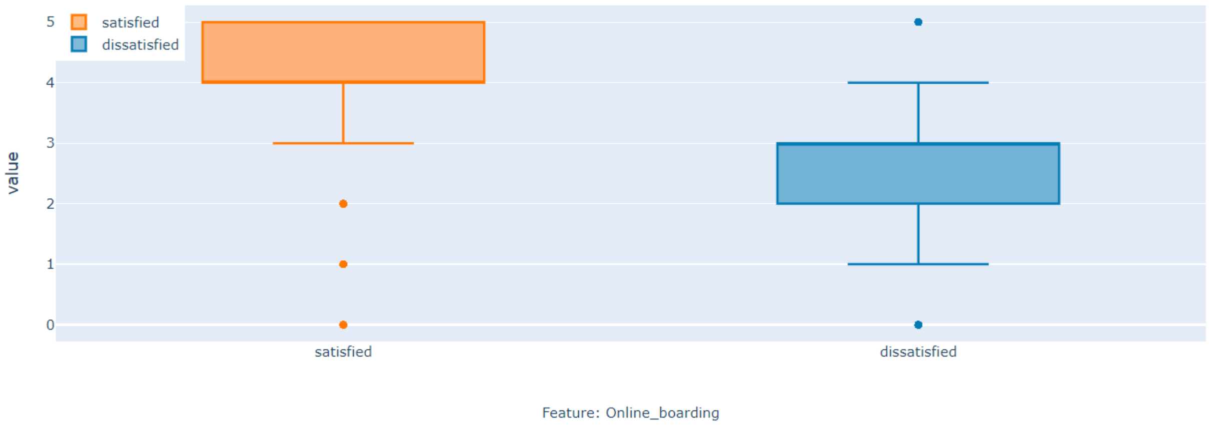
3.4 Box-plot

A box plot, also known as a box-and-whisker plot, is a graphical representation that displays the distribution and summary statistics of a set of data. It provides a visual summary of the central tendency, spread, and skewness of the data. Some box-plots are given below showing min, max and median range:

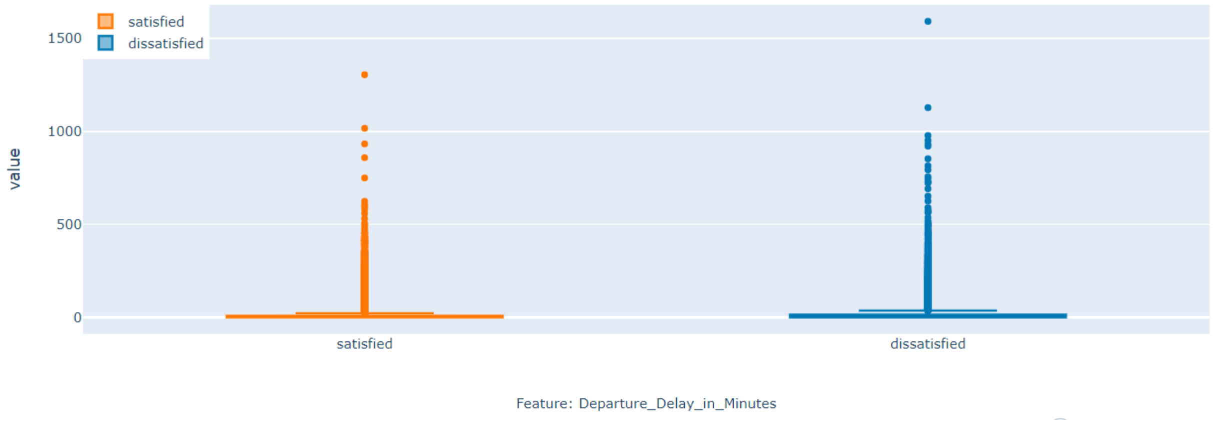
- In this box-plot, we can see that in both cases median is close to Q1. For satisfied, there is no outliers. And for dissatisfied, there are many outliers.



- ➢ In this box-plot, we can see that in satisfied case median is close to Q1 and in dissatisfied case median is close to Q3. For satisfied case, there are 3 outliers. And for dissatisfied case, there are 2 outliers.



- ➢ In this box-plot, we can see that in satisfied and dissatisfied case Q1, median and Q3 are all equal. For satisfied and dissatisfied cases, there are many outliers.



3.5 Chapter Summary

In this chapter, we analyze some diagrams to find patterns and to find outliers. We analyze some histograms, scatterplots, correlation matrix and box-plots to find patterns and outliers.

4 Statistical Inference

Statistical inference is the process of drawing conclusions or making predictions about a population based on a sample of data from that population. It involves using statistical methods and models to make inferences about the characteristics of a larger group or population from which the sample was drawn. Statistical inference is crucial in scientific research, data analysis, and decision-making.

4.1 Statistical Measures

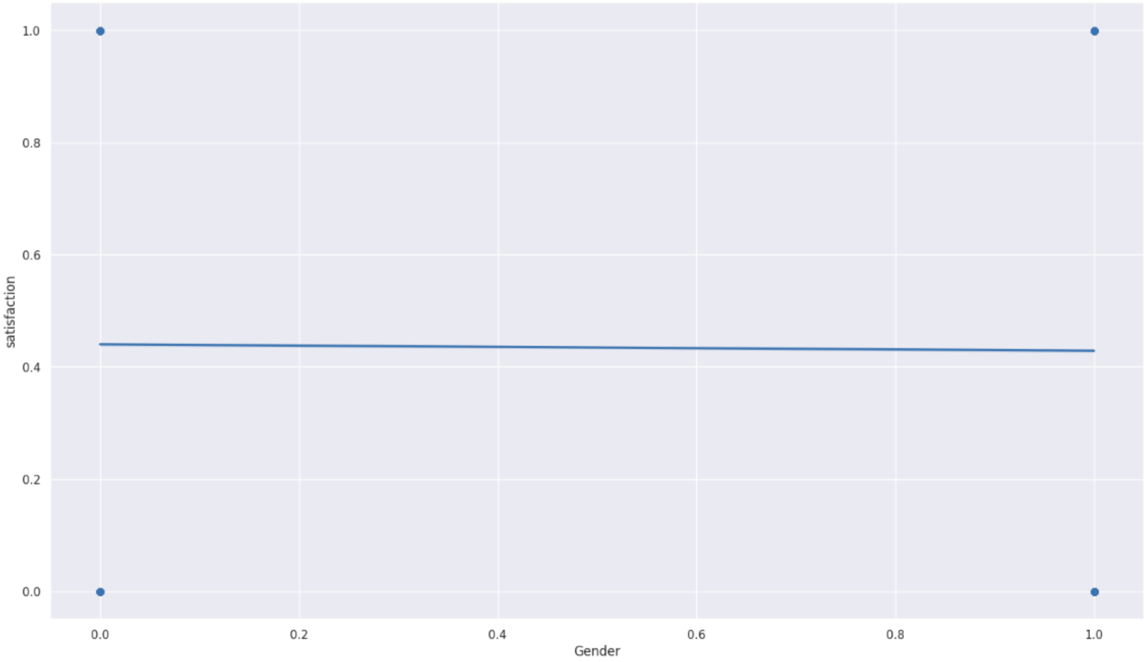
Statistical measures are quantitative values calculated from a set of data to provide insights into various aspects of its distribution, central tendency, dispersion, and shape. These measures help summarize and describe the characteristics of a dataset.

4.2 Confidence Measures

Confidence measures, particularly confidence intervals, are statistical tools used to estimate the range within which a population parameter is likely to fall with a certain level of confidence. Confidence intervals provide a measure of the uncertainty associated with estimating a population parameter based on sample data. The most common confidence level is 95%, but other levels, such as 90% or 99%, can also be used. For this dataset, the confidence measure we use is 95%.

4.3 Regression

Regression analysis is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. The goal of regression analysis is to understand how changes in the independent variables are associated with changes in the dependent variable. It is commonly used for prediction, understanding the strength and direction of relationships, and identifying the key factors that influence an outcome. Here I am attaching a regression plot below showing regression line between satisfaction and gender:



4.4 Chapter Summary

Regression analysis is widely used in various fields, including economics, finance, biology, and social sciences, for understanding and predicting relationships between variables. In this chapter, we plot regression graph to show regression and line and provide confidence measure for the dataset.

5 Model Building

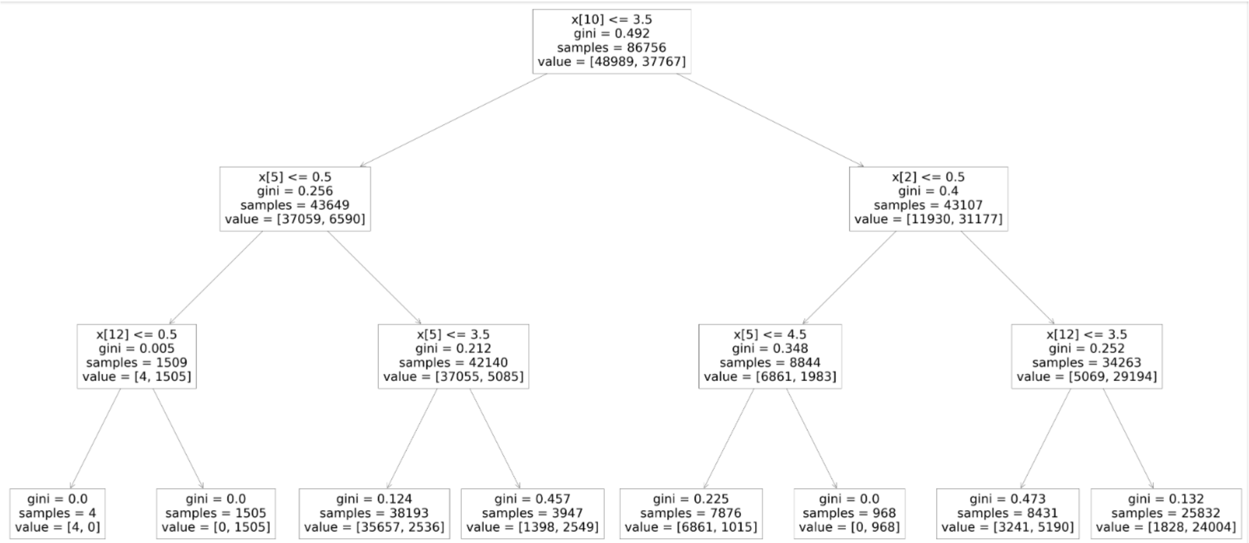
Building a model involves creating a mathematical representation of a system or process based on available data to make predictions, gain insights, or understand relationships between variables. The process of model building varies depending on the type of model (e.g., statistical, machine learning) and the specific objectives of the analysis.

5.1 Decision Tree

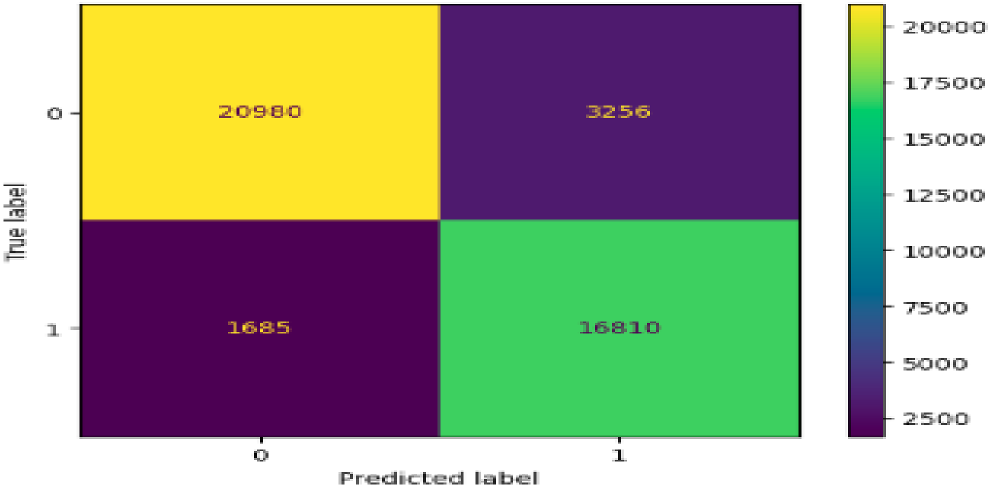
Decision tree is a popular machine learning algorithm used for both classification and regression tasks. It models decisions based on a tree-like structure, where each internal node represents a decision based on the value of a particular feature, each branch represents an outcome of that decision, and each leaf node represents the final predicted output.

5.1.1 CART Algorithm

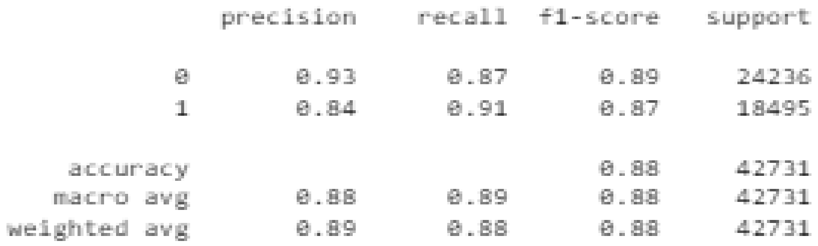
CART, which stands for Classification and Regression Trees, is a specific type of decision tree algorithm that can be used for both classification and regression tasks. The CART algorithm was introduced by Leo Breiman and is widely used for its simplicity and effectiveness. Here I am attaching the decision tree below using CART algorithm for max depth 3:



And the confusion matrix for the dataset using CART algorithm is given below:



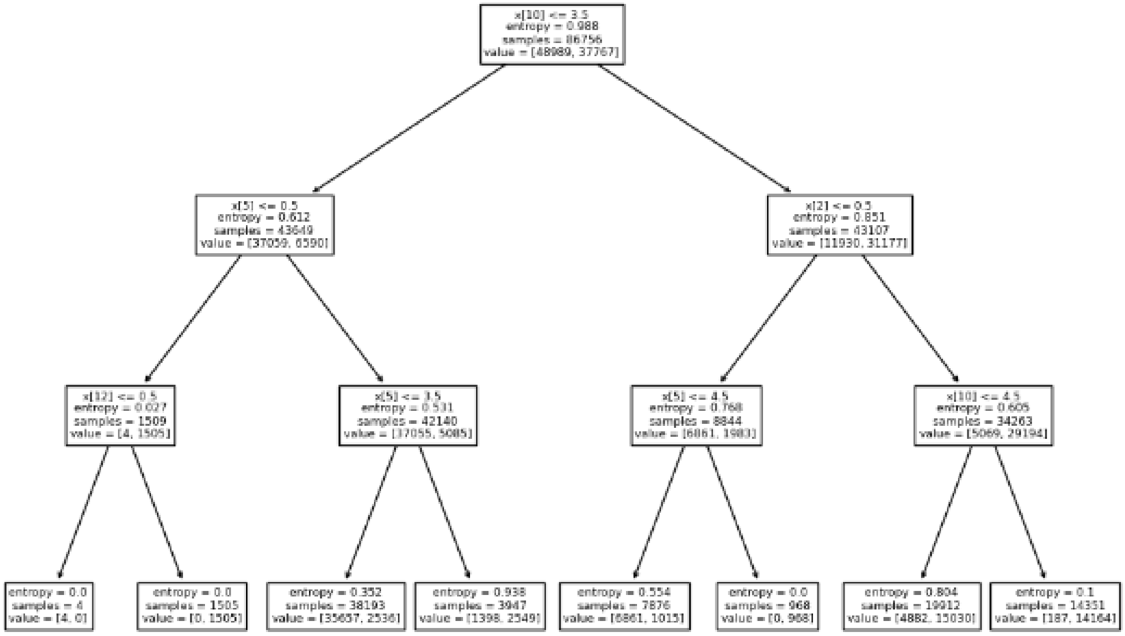
And the classification report for this dataset using CART algorithm is given below:



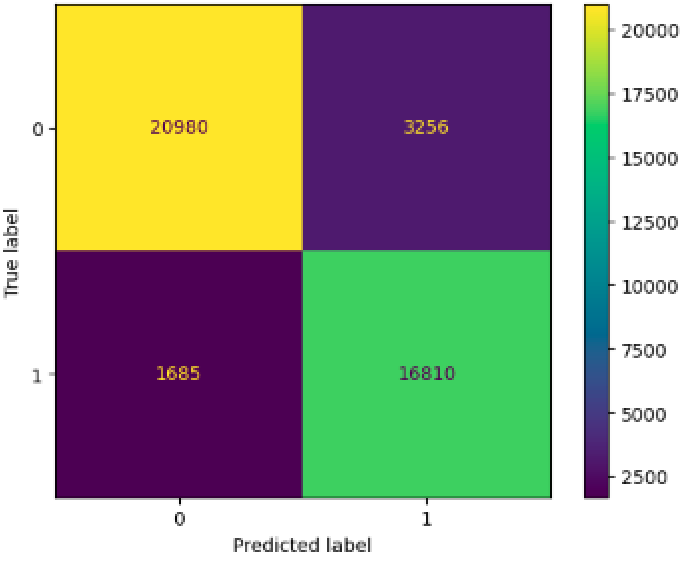
```
              precision    recall  f1-score   support

           0       0.93      0.87      0.89     24236
           1       0.84      0.91      0.87     18495

    accuracy                           0.88     42731
   macro avg       0.88      0.89      0.88     42731
weighted avg       0.89      0.88      0.88     42731
```

5.1.2 C4.5 Algorithm

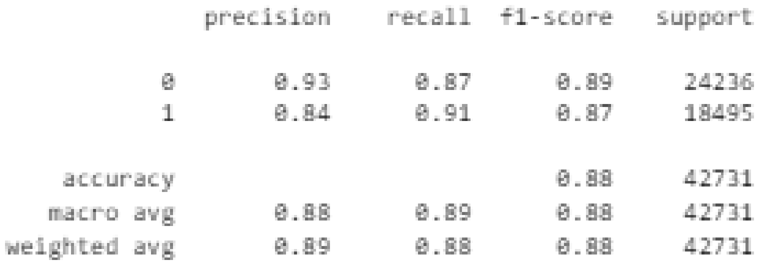
C4.5 is a classic machine learning algorithm used for generating decision trees. It was developed by Ross Quinlan and is an extension of his earlier ID3 (Iterative Dichotomiser 3) algorithm. C4.5 is primarily used for classification tasks but can be adapted for regression as well. It's notable for its ability to handle both discrete and continuous attributes, as well as its approach to handling missing values. Here I am attaching the decision tree below using C4.5 algorithm for max depth 3:



And the confusion matrix for the dataset using C4.5 algorithm is given below:



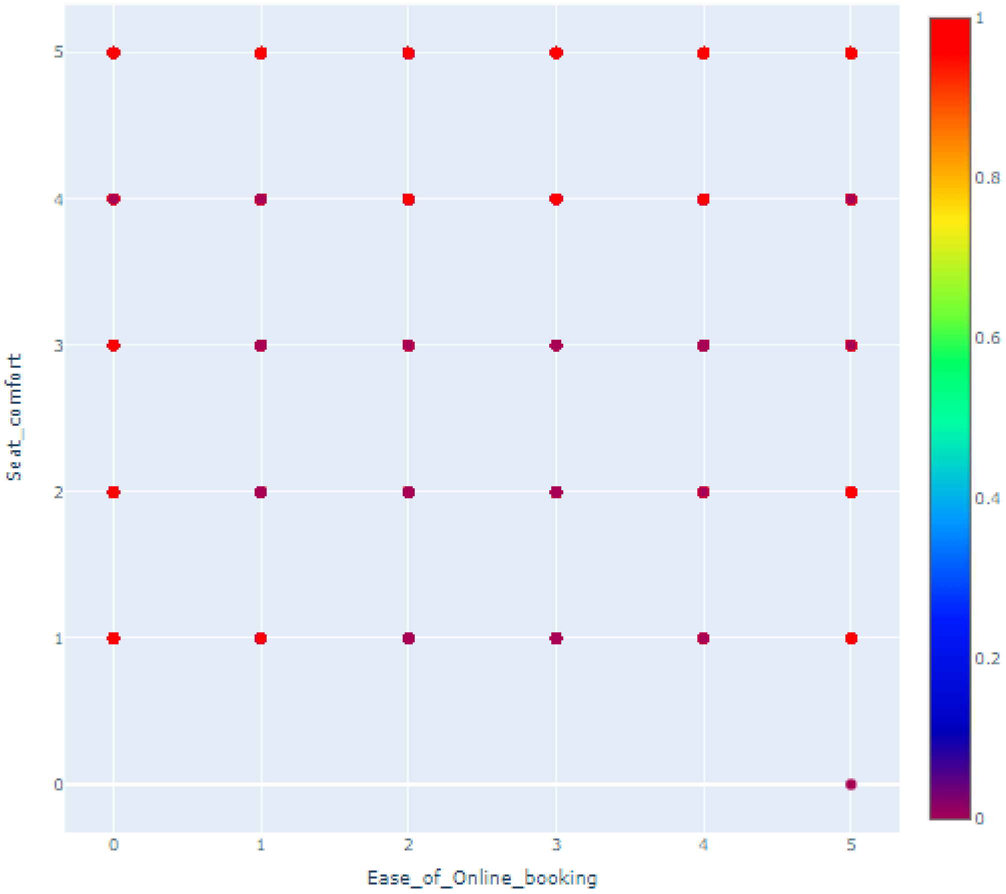
And the classification report for this dataset using C4.5 algorithm is given below:



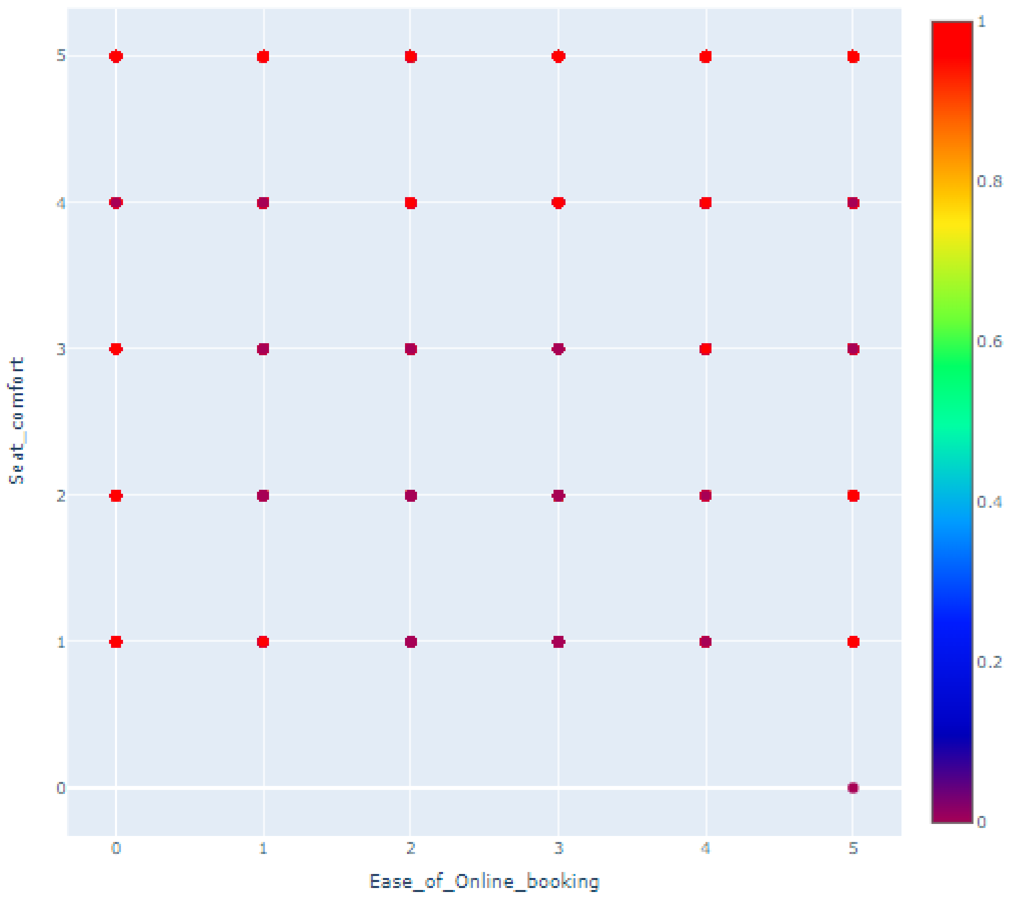
```
              precision    recall  f1-score   support

           0       0.93      0.87      0.89     24236
           1       0.84      0.91      0.87     18495

    accuracy                           0.88     42731
   macro avg       0.88      0.89      0.88     42731
weighted avg       0.89      0.88      0.88     42731
```

5.2 K-Nearest Neighbor

The k-Nearest Neighbors (k-NN) algorithm is a simple and widely used machine learning algorithm for both classification and regression tasks. It is a type of instance-based learning, where the model doesn't explicitly learn from the training data but instead memorizes it. The prediction is based on the majority class (for classification) or the average value (for regression) of the k-nearest neighbors in the feature space. Here I am attaching two scatterplots below showing nearest neighbor values:



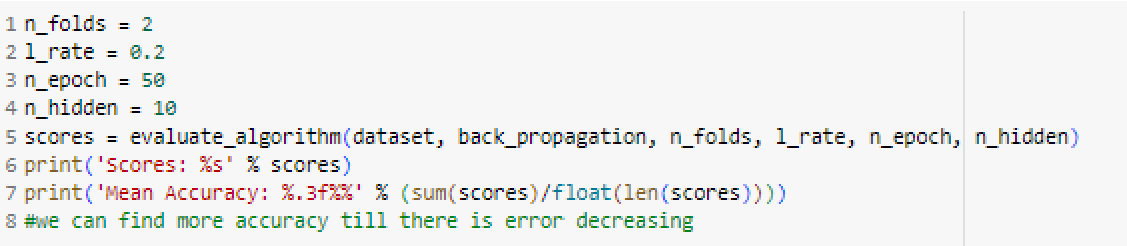
This scatterplot shows points before appending dataframe.



This scatterplot show points after appending dataframe.

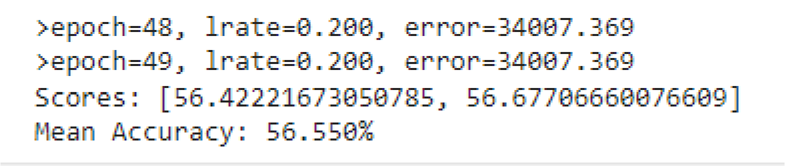
5.3 Neural Network

A neural network is a computational model inspired by the way biological neural networks in the human brain function. It's a key component of machine learning and artificial intelligence. Neural networks are used for a wide range of tasks, including pattern recognition, classification, regression, image and speech recognition, and more.



```python
n_folds = 2
l_rate = 0.2
n_epoch = 50
n_hidden = 10
scores = evaluate_algorithm(dataset, back_propagation, n_folds, l_rate, n_epoch, n_hidden)
print('Scores: %s' % scores)
print('Mean Accuracy: %.3f%%' % (sum(scores)/float(len(scores))))
#we can find more accuracy till there is error decreasing
```

After running the neural network, we find that:



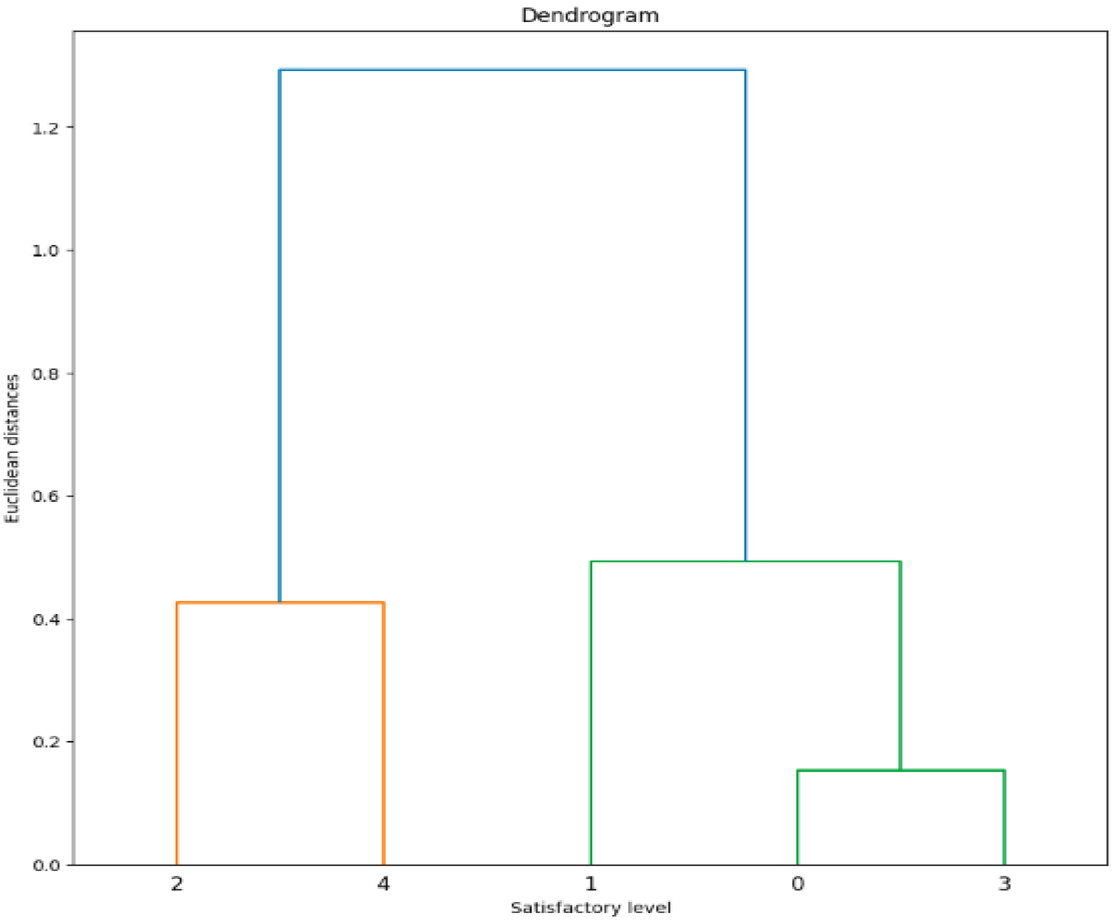
```
>epoch=48, lrate=0.200, error=34007.369
>epoch=49, lrate=0.200, error=34007.369
Scores: [56.42221673050785, 56.67706660076609]
Mean Accuracy: 56.550%
```

5.4 Hierarchical & K-Means Clustering

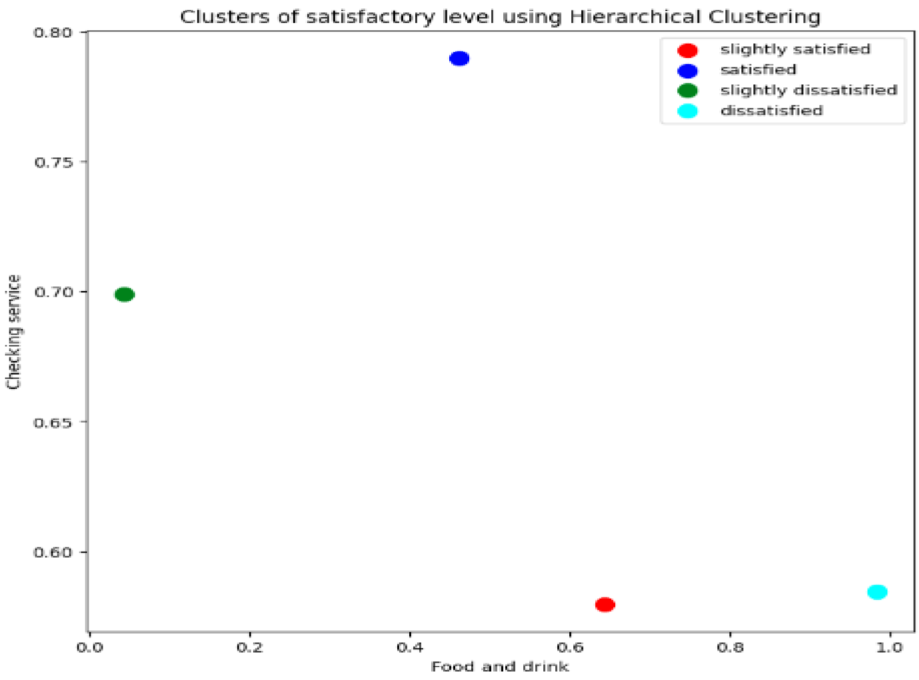
Hierarchical clustering and K-means clustering are two popular techniques in unsupervised machine learning used for grouping similar data points into clusters. Each method has its own characteristics and use cases.

5.4.1 Hierarchical Clustering

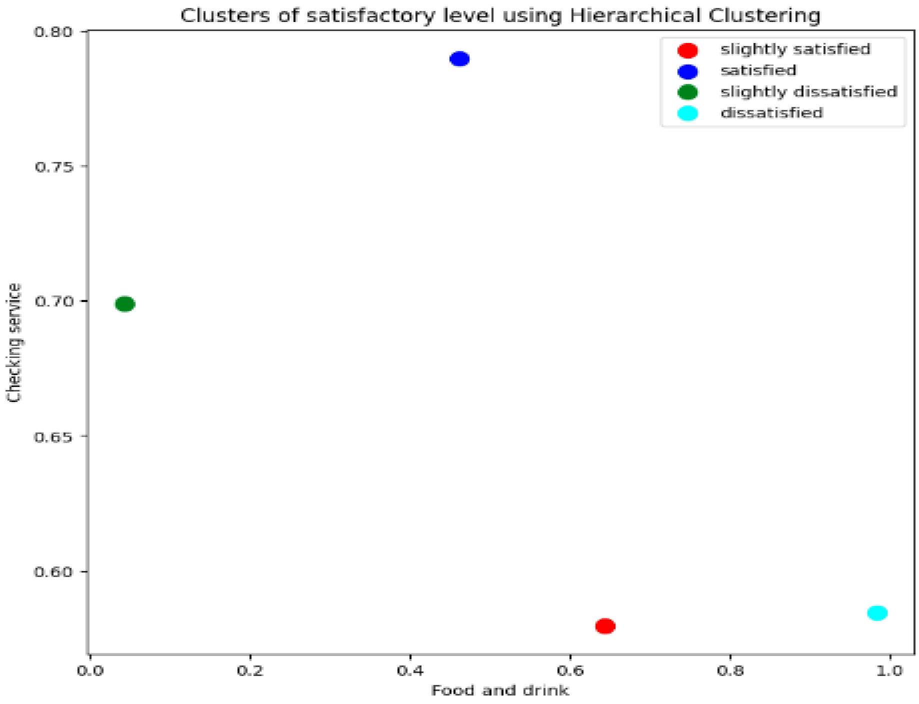
Hierarchical clustering is a method that creates a hierarchy of clusters. It can be visualized as a tree-like structure, known as a dendrogram. The dendogram for the dataset is given below:



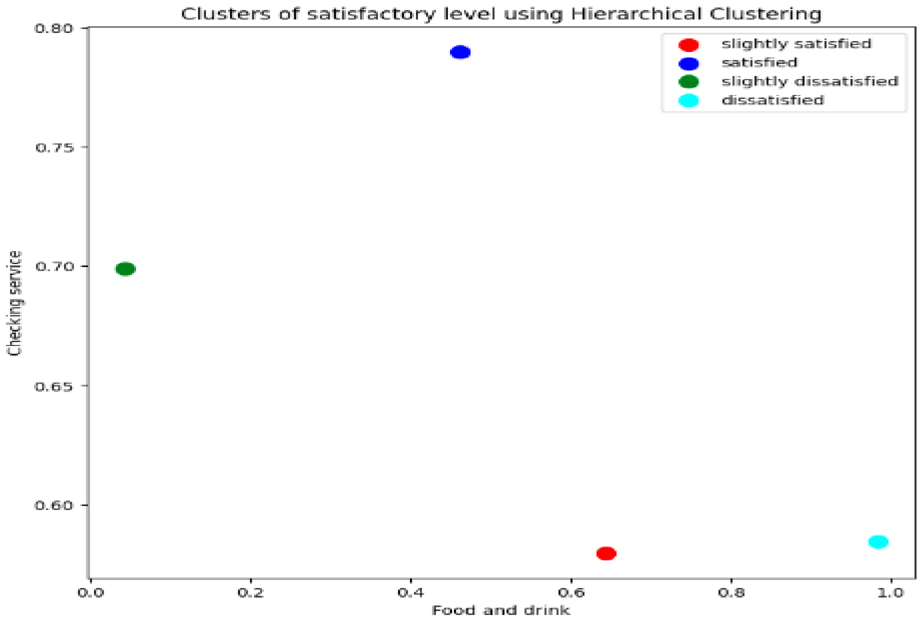
Here I am attaching some figures showing Hierarchial clustering:



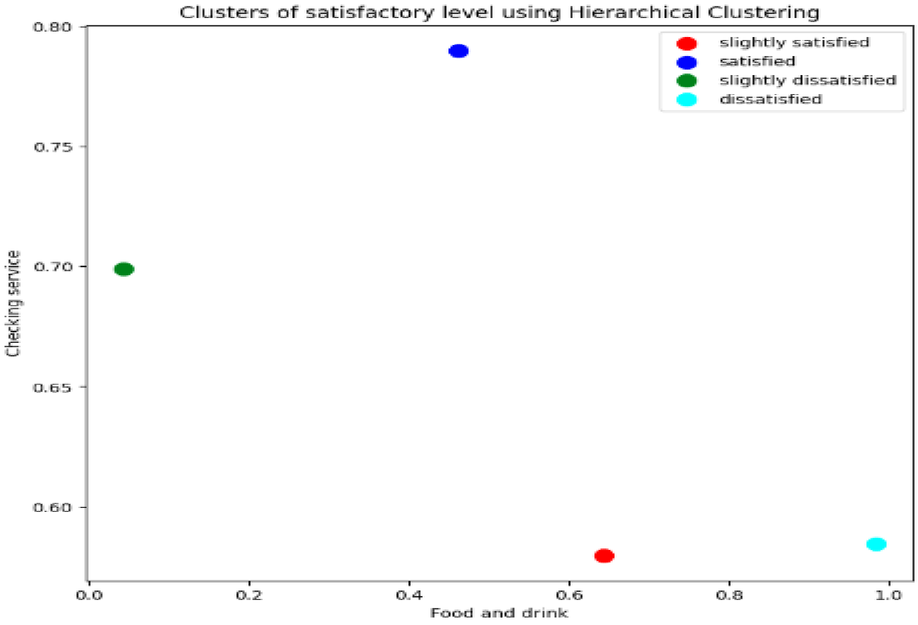
This figure shows clustering using single linkage.



This figure shows clustering using complete linkage.



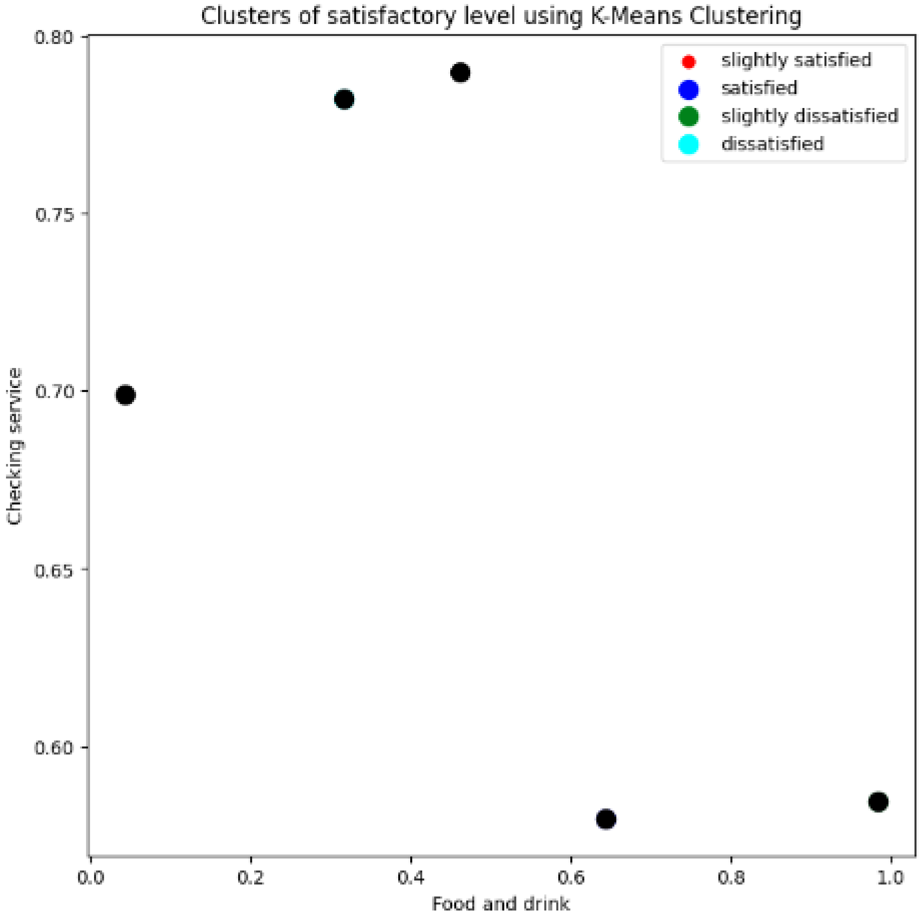
This figure shows clustering using average linkage.



This figure shows clustering using ward linkage.

5.4.2 K-Means Clustering

K-means clustering is a partitioning method that divides a dataset into a predefined number of clusters (K). The goal is to assign data points to clusters in a way that minimizes the sum of squared distances from each point to the center of its assigned cluster (centroid). Here I am attaching a figure below showing the centroid of the clusters:



5.5 Chapter Summary

This chapter states different type of model. This chapter includes models like Decision Tree, K-Nearest Neighbor, Neural Network, and Hierarchical & K-Means Clustering.

6 Model Evaluation

Model evaluation is a critical step in the machine learning workflow to assess the performance and generalization ability of a trained model. The choice of evaluation metrics depends on the type of task (classification, regression, clustering) and the specific goals of the analysis.

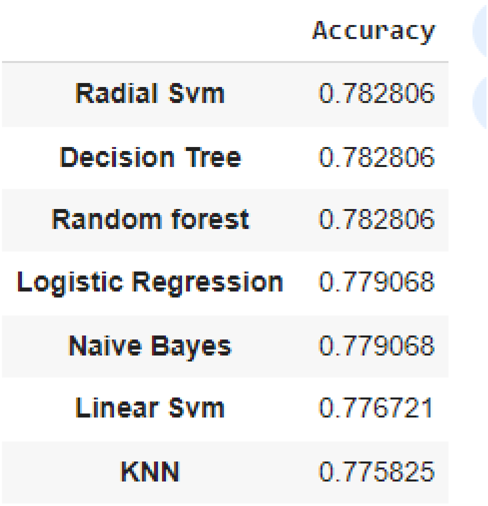
6.1 Evaluation Metrics

I have used several types of evaluation metrics in this dataset. The metrics are:

- Radial Svm
- Decision Tree
- Random forest
- Logistic Regression
- Naive Bayes
- Linear Svm
- KNN

6.2 Accuracy

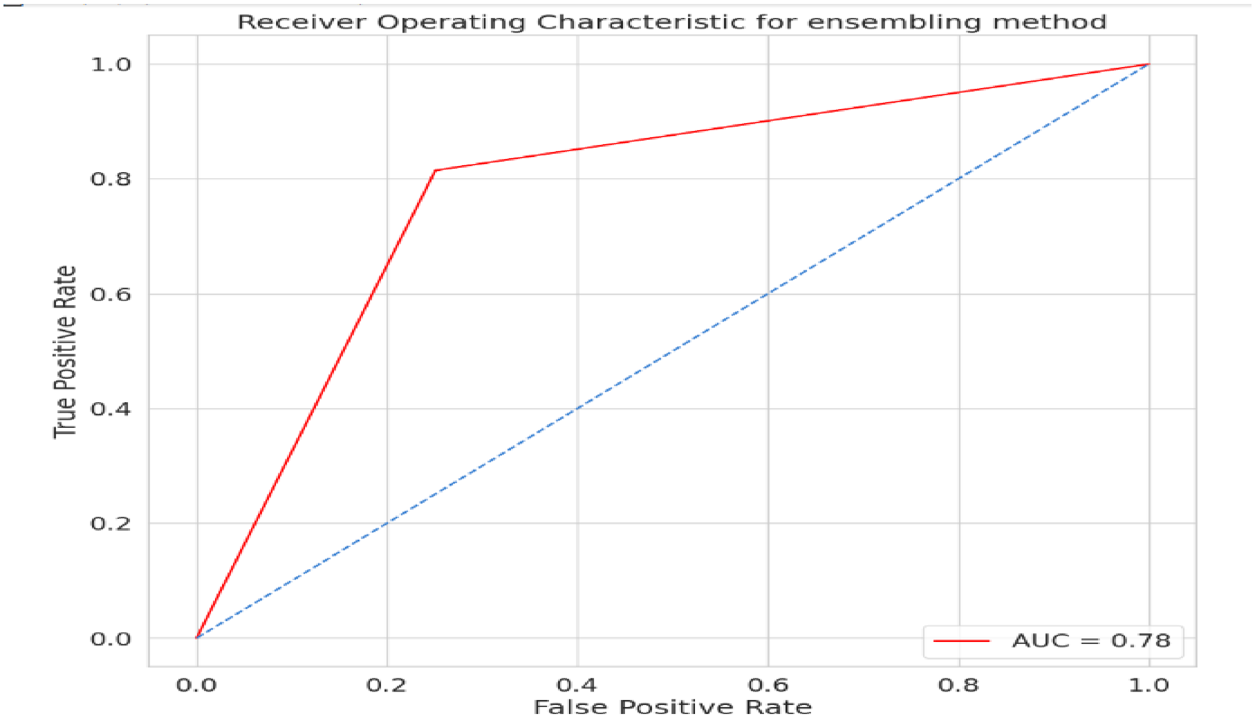
Accuracy is the measure of the overall correctness of the model. The accuracy for different models over this dataset are given below:



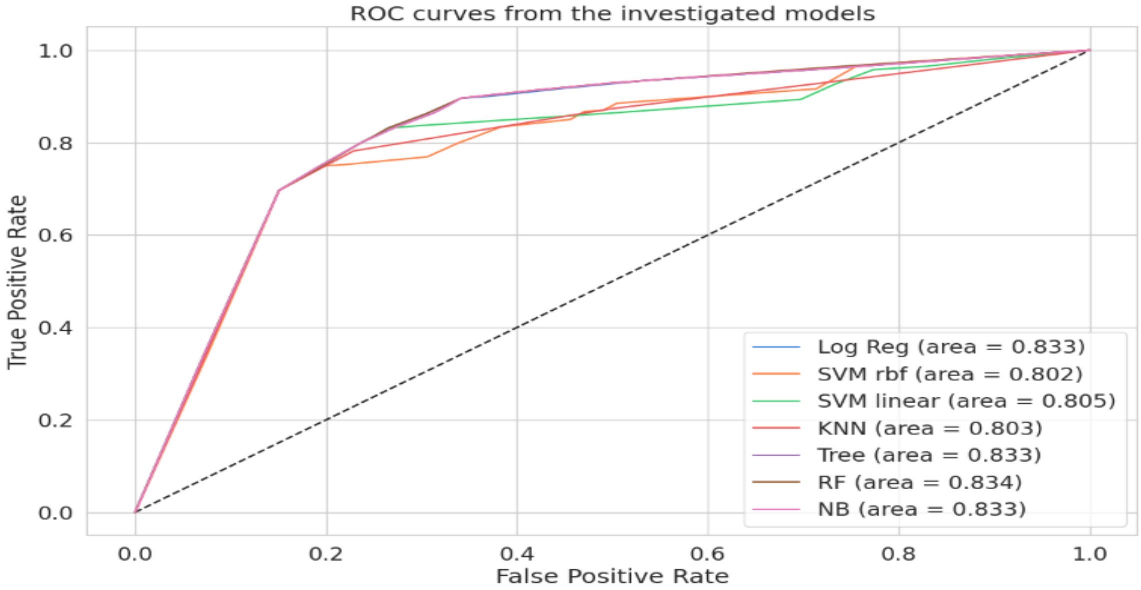
	Accuracy
Radial Svm	0.782806
Decision Tree	0.782806
Random forest	0.782806
Logistic Regression	0.779068
Naive Bayes	0.779068
Linear Svm	0.776721
KNN	0.775825

6.3 ROC Curve

ROC Curve measures the area under the Receiver Operating Characteristic (ROC) curve. Here I am attaching two figures showing two figures. One of them is for only ensembling method and the other one is for the other models.



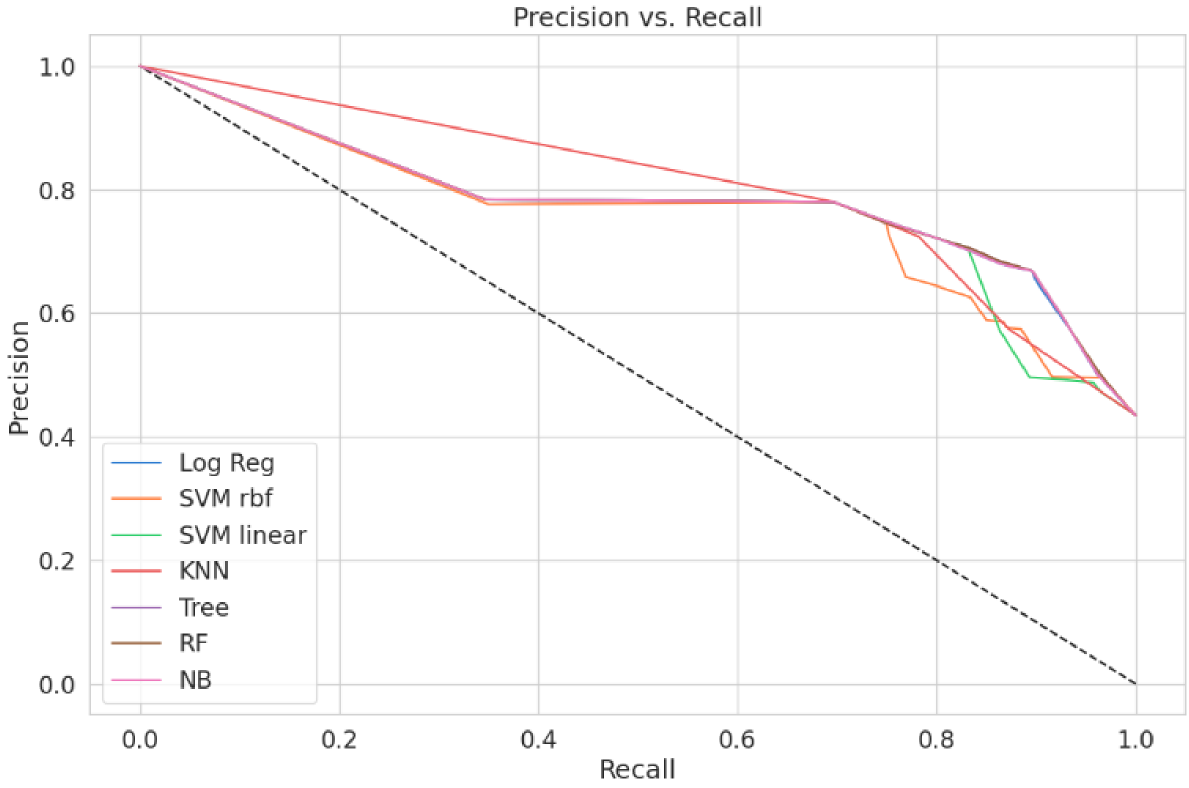
This figure is only for ensembling model.



This figure is for other models.

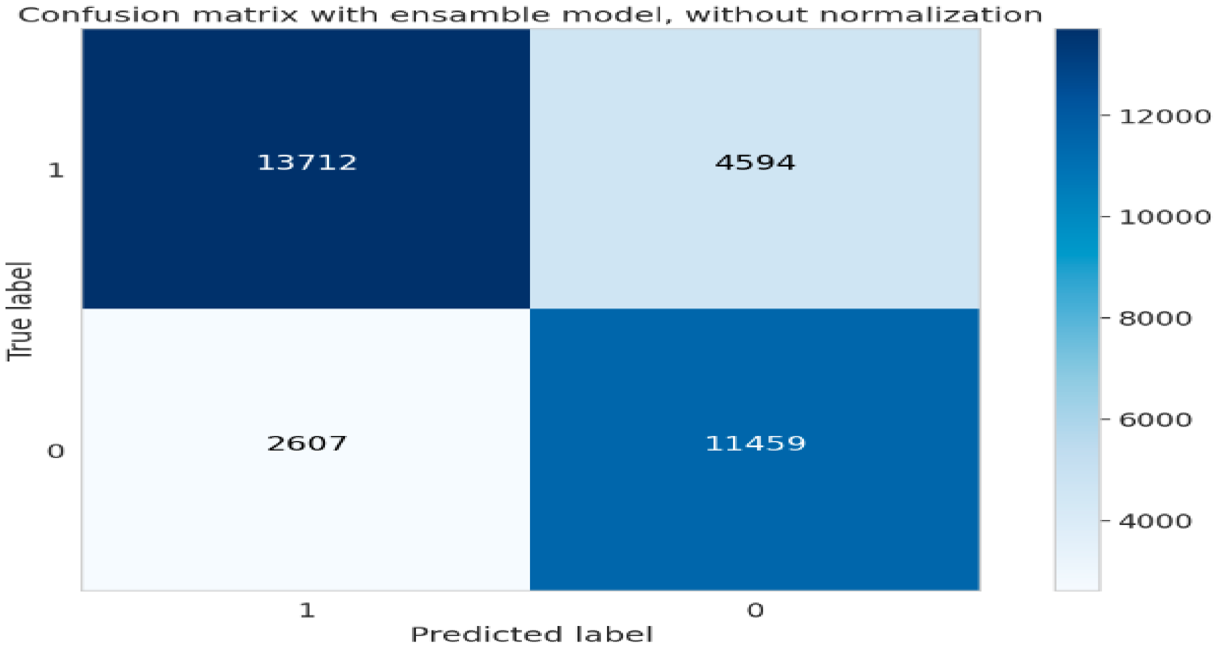
6.4 Precision Recall Curve

The Precision-Recall (PR) curve is a graphical representation of the trade-off between precision and recall for different thresholds in a binary classification model. It is particularly useful when dealing with imbalanced datasets, where one class is much more frequent than the other. The Precision Recall Curve for all models is given below:

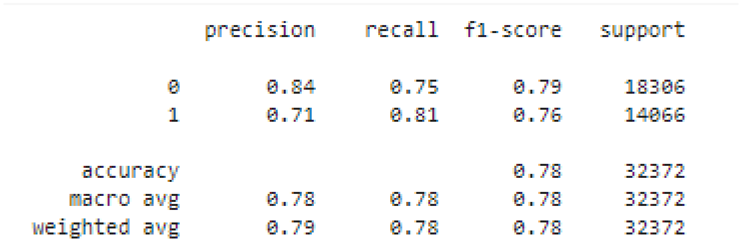


6.5 Confusion Matrix

A confusion matrix is a table used in machine learning to evaluate the performance of a classification model. It summarizes the results of the model's predictions on a classification problem, showing the counts of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. The confusion matrix is a useful tool to assess the performance of a model and understand its behavior. The confusion matrix for this dataset using ensemble model is given below:



The classification report for this model is given below:



```
              precision    recall  f1-score   support

           0       0.84      0.75      0.79     18306
           1       0.71      0.81      0.76     14066

    accuracy                           0.78     32372
   macro avg       0.78      0.78      0.78     32372
weighted avg       0.79      0.78      0.78     32372
```

6.6 Chapter Summary

In this chapter, we can see some evaluation metrics that are used to evaluate some models on our dataset. To evaluate our models, we used ROC Curve, Precision Recall Curve, Confusion Matrix and so on.

7 Conclusion

In conclusion, our analysis of the airline passenger satisfaction dataset has provided valuable insights into the factors influencing passenger contentment and areas where improvements can be made. The overall satisfaction level, indicated by [56.42221673050785, 56.67706660076609] and mean accuracy [56.550%], serves as a foundation for understanding passenger sentiments.

7.1 Limitations

The limitations of this dataset is there are too many attributes that holds great impact on the accuracy level of models. Also there is no binary attributes that we can use to apply association rules. So association rules cannot be applied in this dataset.

7.2 Future Works

As the accuracy level is not so high and error rate is too much high, so there is a great scope to work to increase accuracy and model efficiency.

References

1) https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction