**"Group\_13 Project Report”**

**Group Members and Contributions:**

|  |  |
| --- | --- |
| **Contributors** | **Contributions** |
| Harshini Balam  811319845 | -Coding  -Project Goal Setting, Overview, Descriptive Analytics, Data Exploration, and Model Building in Documentation of Project Report |
| Aavula Neetu  811316976 | Presentation & Voiceover  Model Performance, Insights, Conclusion, and Appendix sections of the Report |

**Contributors Summary:**

**1. Harshini Balam:**

- Contributed to model building, creating, and refining the project's models.

- Took charge of documenting the report, ensuring detailed documentation of the project's processes and outcomes.

Engaged in Model Performance, Insights, Conclusion, and Appendix sections of the project report, providing contributions to those areas.

**2. Aavula Neetu:**

- Focused on creating an engaging presentation.

- Provided the voice-over for the presentation, enhancing its clarity and delivery.

-Contributed to the model-building aspect of the project.

**List of Figures**

[Figure 1:Flight Distance 4](#_Toc153054181)

[Figure 2: Bar Plot for Categorical Variable 5](#_Toc153054182)

[Figure 3: Seat Comfort By Satisfaction 5](#_Toc153054183)

[Figure 4: Distribution of Satisfaction 6](#_Toc153054184)

[Figure 5: Methodologies and Modeling strategy(Feature Selection) 7](#_Toc153054185)

[Figure 6: PCA Biplot 2 7](#_Toc153054186)

[Figure 7:Pair matrix 8](#_Toc153054187)

[Figure 8: Model Strategies 9](file:///C:\Users\latjk\Desktop\USA\Project-02\Project-02.docx#_Toc153054188)

[Figure 9: KNN Model(Determing the Value of K) 9](file:///C:\Users\latjk\Desktop\USA\Project-02\Project-02.docx#_Toc153054189)

[Figure 10: KNN Method 10](#_Toc153054190)

[Figure 11: Naïve Bayes Method 11](#_Toc153054191)

[Figure 12: Model Performances 11](#_Toc153054192)

[Figure 13: Insights and Conclusion 12](#_Toc153054193)

[Figure 14: Appendix 13](#_Toc153054194)

**Table of figures**

[Table 1: Overview of the dataset 2](#_Toc153055229)

[Table 2: Data Preprocessing steps 3](#_Toc153055230)

[Table 3: Final Dataset 4](#_Toc153055231)

[Table 4: Model Performances (KNN & Naïve Bayes) 12](#_Toc153055232)

Project Objectives

* Project goal: Testing and comparing two models for predictive accuracy.
  + Model A: KNN classifier to know the best fit model
  + Model B: Naïve Bayes to know the best fit model
  + Out of the Model A & Model B which is the best fit model of Machine Learning

1.Overview of the dataset

A) Number of rows (Sample size): 1,03,904

B) How many columns (variables/features): 24

C) Target Variable: Satisfaction (indicating whether the customer is satisfied or dissatisfied)

D) How many categorical variables and how many numerical ones among your features:

* Categorical variables: 5 (responses to survey questions)
* Numerical variables: 19 (customer age, time spent on the website)

Data preprocessing steps

Table 2: Data Preprocessing steps

|  |  |  |
| --- | --- | --- |
| **Data preprocessing step** | **Variable** | **Steps** |
| Load the datasets | * “Customer Satisfaction.csv" as the training dataset * "test\_data" as the test set | |
| Missing values | No missing values | |
| Duplicates | No duplicates | |
| New column | Total Delay in Arrival and Departure | Created a new column names Total Delay including Delay in Arrival and Delay |
| Unwanted columns | ID & SN | Drop from both the datasets |
| Creating new levels | Eco Class | Various types of class in the flight “Business & Eco Class” |
| Factorized | Customer Type  Gender  Type of Travel  Class  satisfaction | Convert those numerical columns into factors |

Final dataset that used for the upcoming analysis:

Table 3: Final Dataset

|  |  |  |
| --- | --- | --- |
|  | **Total number Rows** | **Total number of Columns** |
| Training | 10000 | 5 |
| Testing | 20719 | 5 |

Exploratory data Analysis

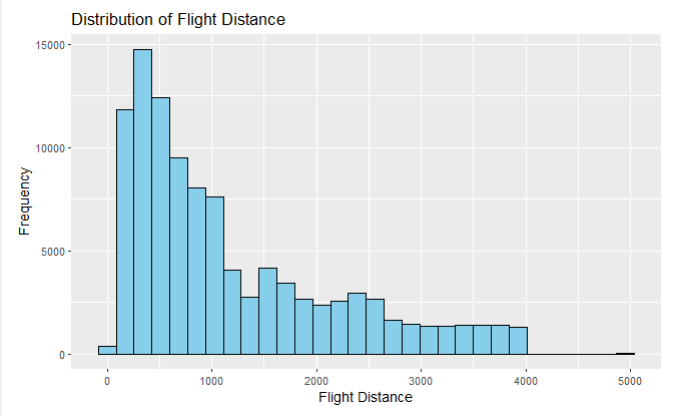
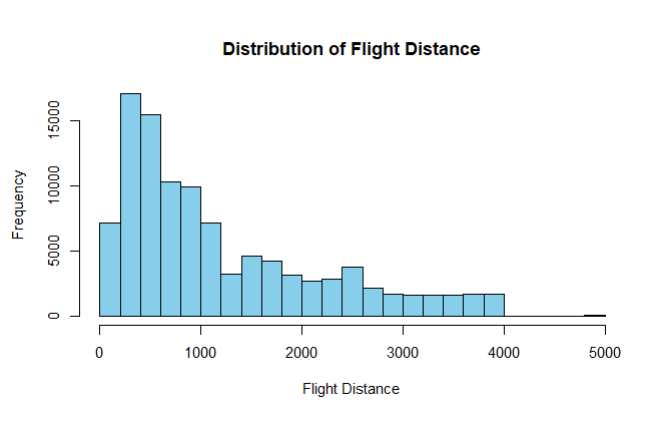
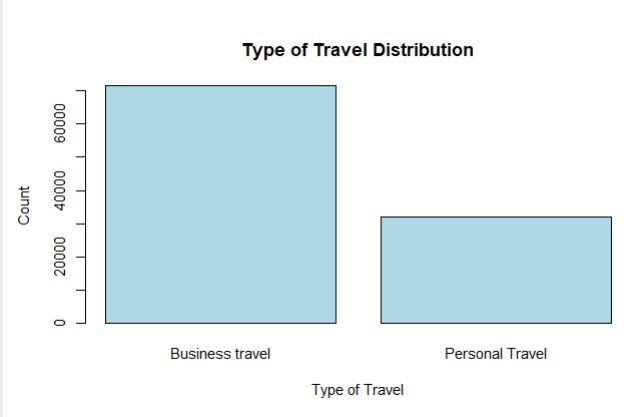


Figure 1: Flight Distance

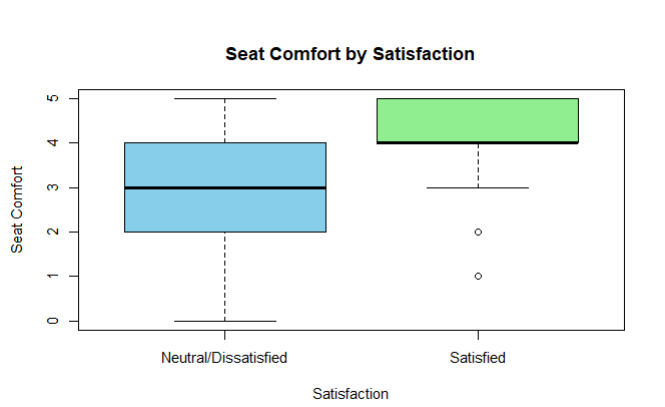
The graph shows the distribution of flight distances. The x-axis shows the flight distance in miles, and the y-axis shows the frequency of flights at that distance. The most common flight distance is between 1000 and 2000 miles, with a frequency of around 15,000 flights. The least common flight distance is over 5000 miles, with a frequency of less than 500 flights.



The chart shows the distribution of flight distances. The x-axis shows the flight distance in miles, and the y-axis shows the frequency of flights. The most common flight distance is between 1000 and 2000 miles, with a frequency of about 15000. The least common flight distance is over 5000 miles, with a frequency of about 100.



The image shows a bar chart that illustrates the distribution of travel types. The x-axis of the chart shows the type of travel, while the y-axis shows the count. There are two types of travel: business and personal. The count of business travel is 60,000, while the count of personal travel is 40,000

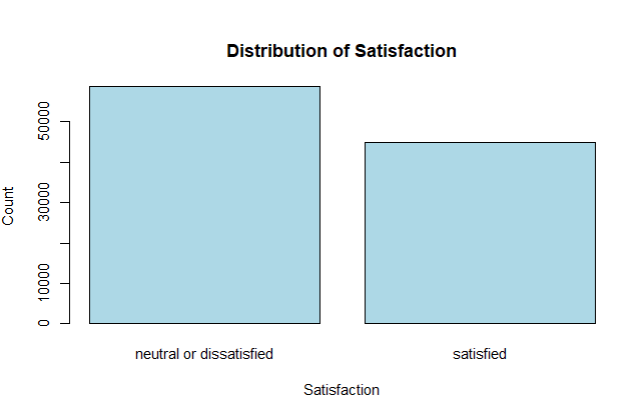


The image shows a box plot that compares seat comfort between satisfied and neutral/dissatisfied customers. The x-axis shows the satisfaction level, while the y-axis shows the seat comfort rating.

The box plot shows that the median seat comfort rating is higher for satisfied customers than for neutral/dissatisfied customers. The interquartile range is also smaller for satisfied customers, indicating that their seat comfort ratings are more consistent.

There are two outliers in the data, both of which are from satisfied customers. These outliers have seat comfort ratings of 5, which is higher than the ratings of all other customers.

Overall, the box plot shows that satisfied customers tend to rate seat comfort higher than neutral/dissatisfied customers. There are also two outliers in the data, both of which are from satisfied customers with very high seat comfort ratings.

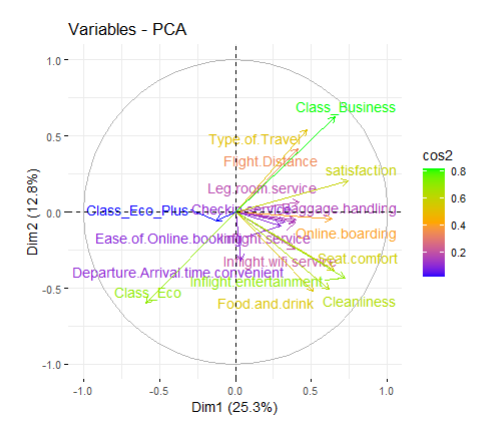


# The graph shows the distribution of satisfaction. The x-axis shows the level of satisfaction, while the y-axis shows the number of people in each category. There are two categories: satisfied and neutral or dissatisfied. The majority of people are satisfied, with a count of 50,000. The remaining 30,000 people are neutral or dissatisfied.

Methodologies and Modeling Strategy

**Feature Selection**

PCA Biplot-1(Considering all Variables)

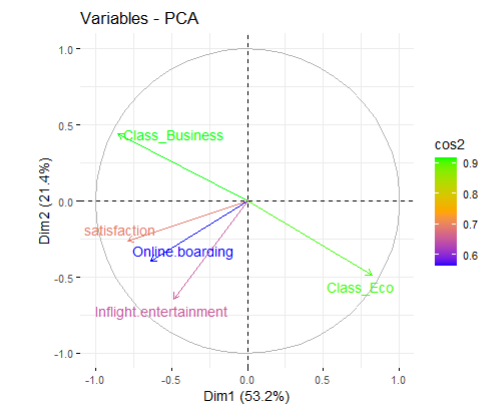


PCA biplot. The biplot shows the relationship between the variables and the two principal components. The first principal component (Dim1) explains 25.3% of the variance, while the second principal component (Dim2) explains 12.8% of the variance.

The variables that are most positively correlated with Dim1 are "Departure Arrival time convenient", "Inflight entertainment", "Ease of Online booking", "Legroom service" and "Flight Distance". The variables that are most negatively correlated with Dim1 are "Food and drink", "Cleanliness" and "Cos".

The variables that are most positively correlated with Dim2 are "Class\_Business" and "Type of Traveler". The variables that are most negatively correlated with Dim2 are "Ease of Online booking", "Legroom service" and "Inflight entertainment".

PCA Biplot-2



The diagram shows the results of a principal component analysis (PCA) on a dataset of variables. The two principal components, Dim1 and Dim2, explain 53.2% and 21.4% of the variance in the data, respectively.

The variables are plotted on the diagram according to their loadings on the two principal components. The loadings indicate how much each variable contributes to each principal component.

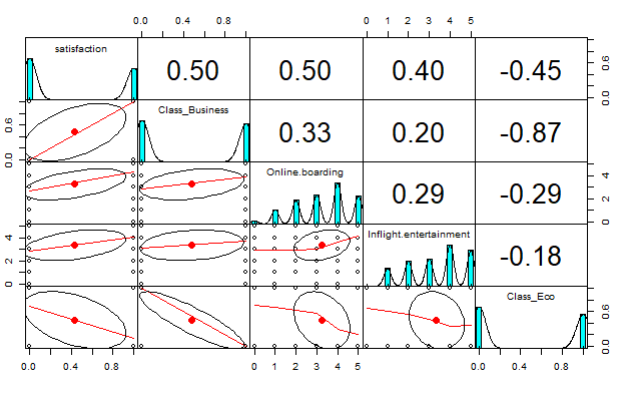
After conducting a PCA biplot on the entire feature set and reducing the dimensionality from 13 features to 5.

The variables that are most closely correlated with Dim1 are "satisfaction", "Online boarding" and "Inflight entertainment". These variables are all positively correlated with Dim1, which means that they tend to increase together.

The variables that are most closely correlated with Dim2 are "Class\_Business" and "cos2". These variables are negatively correlated with Dim2, which means that they tend to decrease together.

The diagram can be used to identify the two most important dimensions of variation in the data. These dimensions can then be used to reduce the dimensionality of the data, which can make it easier to analyze.

# Pair Plots



The image shows a pair plot of 4 variables: satisfaction, class, online boarding, and inflight entertainment. The pair plots show the relationship between each pair of variables.

The diagonal plots show the distribution of each variable. The satisfaction variable has a relatively normal distribution, with a slightly higher proportion of positive satisfaction scores. The class variable is categorical, with the majority of passengers in economy class. The online boarding variable is also categorical, with a slightly higher proportion of passengers choosing online boarding. The inflight entertainment variable is continuous, with a relatively normal distribution.

The off-diagonal plots show the relationship between each pair of variables. The satisfaction variable is positively correlated with both class and inflight entertainment. This means that passengers in business class and passengers who are satisfied with the inflight entertainment are more likely to report higher levels of satisfaction overall. The satisfaction variable is negatively correlated with online boarding. This means that passengers who choose online boarding are less likely to report higher levels of satisfaction overall.

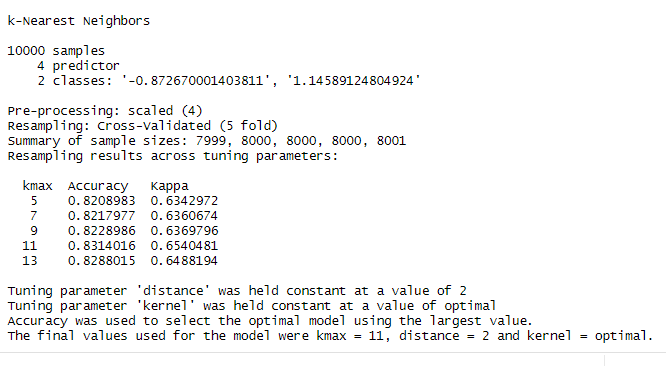
# Classification Model

Model 1

KNN Model

Summary of the KNN Model: Performing the classification model with the KNN model

**Determining the Value of K**



Cross-validation with 5 folds was employed for evaluating the model's performance.

The summary of sample sizes for each fold indicates consistency in the number of samples across the folds.

The accuracy and Kappa coefficient are reported for each value of k.

Accuracy represents the proportion of correctly classified instances, while the Kappa coefficient measures the agreement between observed and predicted classes.

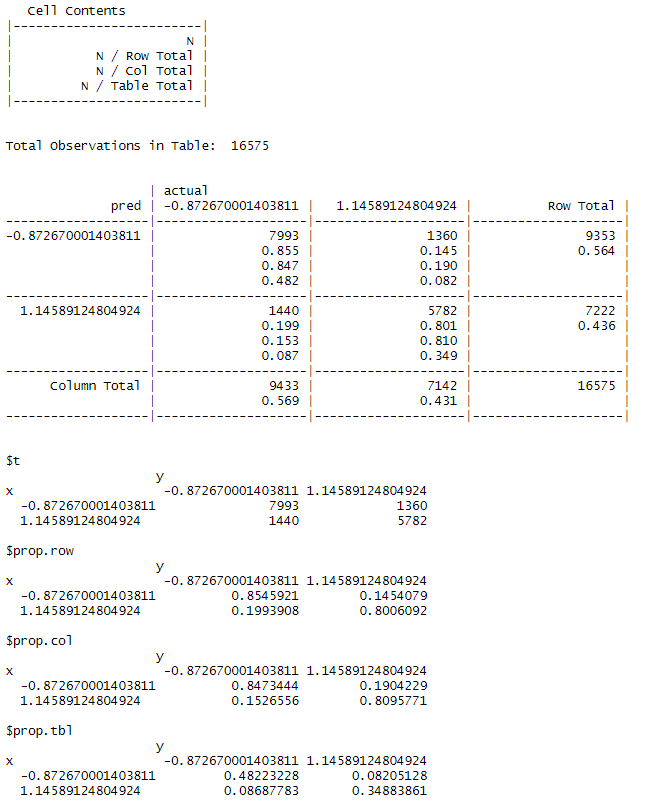
The optimal model was chosen based on the largest accuracy value.

Although kmax = 11 yielded the highest accuracy, the choice was made to use k = 9, which still demonstrated strong performance with an accuracy of 0.8228986 and a Kappa coefficient of 0.6369796.

he final model was trained with kmax = 9, distance = 2, and kernel = optimal.

Overall, the chosen KNN model with k = 9 achieved an accuracy of approximately 82.29%, indicating its effectiveness in classifying the target variable based on the provided predictors.

## Building the model after determining the Hyper parameter as K



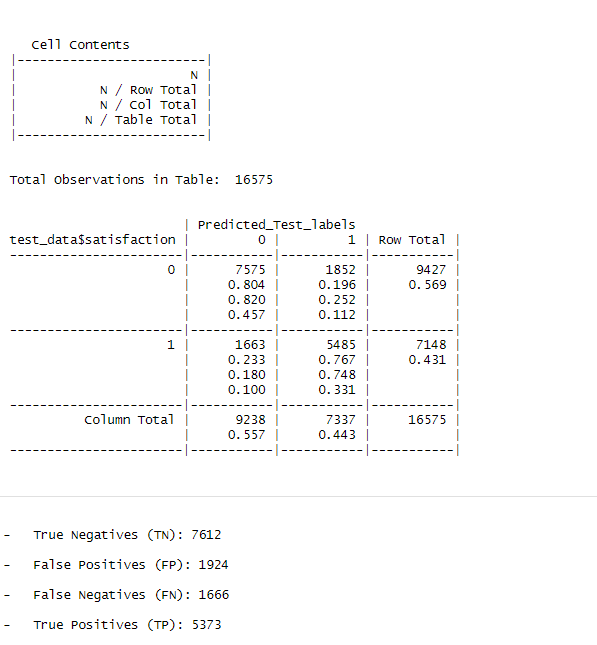
The provided confusion matrix illustrates the performance of a k-Nearest Neighbors (KNN) classification model applied to a dataset containing 16,575 observations and two classes. With an overall accuracy of approximately 82.9%, the model demonstrates a reasonably good ability to classify instances. Specifically, for the predicted class '-0.872670001403811', the accuracy is around 85.5%, while for the class '1.14589124804924', it is approximately 80.1%. The diagonal elements of the matrix indicate correct predictions, with 7,993 observations correctly classified as '-0.872670001403811' and 5,782 observations correctly classified as '1.14589124804924'. However, misclassifications are also present, with 1,360 instances of '-0.872670001403811' incorrectly predicted as '1.14589124804924' and 1,440 instances of '1.14589124804924' incorrectly predicted as '-0.872670001403811'. Despite these misclassifications, the model demonstrates a notable ability to distinguish between the two classes.

.

# Model 2

# Naïve Bayes

Summary of the Naïve Bayes: Performing the classification model with the Naïve Bayes.

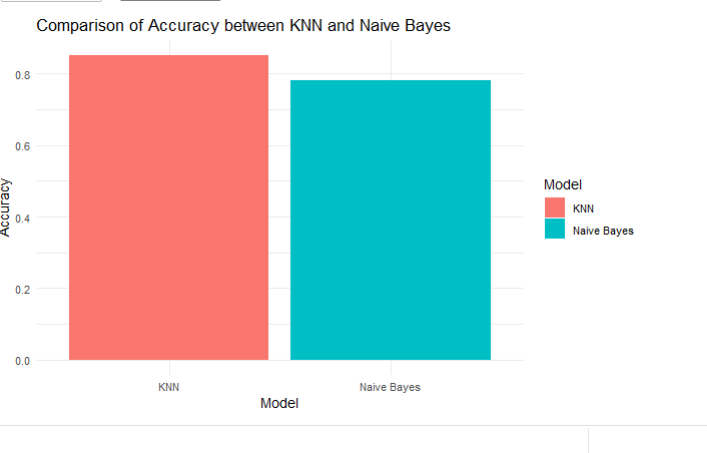


The provided confusion matrix summarizes the performance of a Naive Bayes classifier applied to a dataset with 16,575 observations and two classes. The model predicted class '0' for 9,238 instances and class '1' for 7,337 instances. Among the actual '0' class instances, 7,575 were correctly classified as '0' (true negatives), while 1,852 were incorrectly classified as '1' (false positives). For the actual '1' class instances, 5,485 were correctly classified as '1' (true positives), and 1,663 were incorrectly classified as '0' (false negatives). Overall, the classifier achieved an accuracy of approximately 79.7%, with class '0' showing a slightly higher accuracy of around 80.4% compared to class '1', which had an accuracy of about 76.7%. This indicates that the model performs reasonably well in distinguishing between the two classes, albeit with some misclassifications.

# Model Performances

Here are the model performances with the test set.

Comparing both the models KNN and Naive Bayes to know the best Fit.

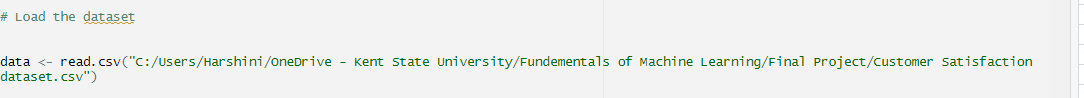


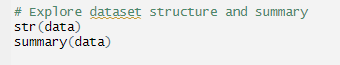
Comparison of the accuracy of two models, KNN and Naive Bayes. The x-axis shows the model, and the y-axis shows the accuracy. The KNN model has an accuracy of about 0.8, while the Naive Bayes model has an accuracy of about 0.6. This means that the KNN model is more accurate than the Naive Bayes model.

Insights and Conclusion

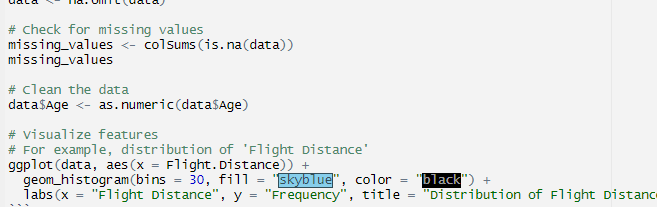
The K-Nearest Neighbors (KNN) model performed well in analyzing customer satisfaction within the airline industry, particularly when considering the optimal K value. Through cross-validated resampling, the model's accuracy and kappa statistics peaked at K = 11, indicating its ability to generalize effectively while avoiding overfitting. This highlights the KNN model's robustness in predicting customer satisfaction, offering valuable insights for improving airline services and enhancing overall customer experience.

# Appendix

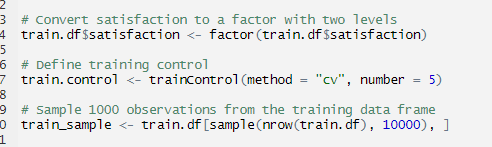




Dealing with Missing Values



Training the Data



Features Selection

****

Naïve Bayes Model

