CLASSIFICATION MODELS BASED ON AIRLINE SATISFACTION RATING

KEERTHANA SENTHILKUMAR
T00711122

Abstract

The focus of this paper is to explore the potential of various machine learning techniques, such as Support Vector Machines (SVM), Decision Trees, Naive Bayes, K-Nearest Neighbors (KNN), and Random Forests, to uncover valuable safety insights within airline data. The analysis involves dividing the dataset into training and testing subsets, assessing the models' accuracy, and examining any potential overfitting concerns. Prior to evaluating results, outliers present in continuous variables are addressed. Classification models take customer satisfaction ratings as their target variable. The confusion matrix serves as a key tool for evaluating each model's performance. By comparing accuracy metrics from the training and testing datasets, the best performing classification model can be identified. The results of this study aim to enhance industry operations by highlighting improvement opportunities and informing data-driven decisions that advance overall air travel efficiency and safety.

Keywords
Accuracy, satisfaction rating, classification models, overfit

I. INTRODUCTION

The aviation industry has always been a critical component of the global economy, connecting people and businesses across the world. However, the Covid-19 pandemic has disrupted the industry, leading to a significant decline in air travel and causing financial losses for airlines worldwide. Considering these challenges, this research aims to explore how machine learning techniques can be utilized to analyze and improve the efficiency of airline operations. The goal of this research will investigate the use of popular classification algorithms, including K-Nearest Neighbours, Support Vector Machines, Decision Trees, Random Forest Classifier, and Naive Bayes Classifier, to analyze airline data. By evaluating and comparing the performance of these algorithms, to identify the most effective model and provide insights into how it can be leveraged to enhance the safety and efficiency of airline travel.

The GitHub link for the R code:
https://github.com/KEERTHANA-SENTHILKUMAR/DASC-5420-AIRLINE-FINAL-PROJECT

II. DATASETS

A. Description of airline dataset:

The airline dataset is a comprehensive source of data containing various aspects of customer satisfaction in the aviation industry. With 9,342 observations of 23 variables, it offers a wide range of data points that can be used to train classification models such as SVM, KNN, Random Forest, Naive Bayes, and Decision Trees. The target variable is the satisfaction rating, which is classified into five categories, including very satisfied, satisfied, neutral, dissatisfied, and very dissatisfied. Overall, the airline dataset is a useful resource for studying and predicting customer satisfaction in the aviation industry [1].

B. Exploring dataset:

The descriptive analysis on the airline dataset is to better understand the data by using the function "summary" and "structure" especially on the continuous variables. The mean age of the customers is 40.47 years and the flight distance range from 67 miles to 3997 miles, with a mean of 1082 miles. The average delay in departure time is 15.6 minutes, while the average delay in arrival time is 18.5 minutes. Overall, the customers seem to be satisfied with the airline services, as most of the satisfaction ratings are above 3 on a scale of 1 to 5. checking any missing or null values in the dataset to avoid any errors in the further modelling performance.

C. Exploratory data analysis:

Identifying and removing outliers in continuous variables such as flight distance, arrival delay, and departure delay is necessary because outliers can lead to overfitting or underfitting of the models, which can ultimately affect the accuracy of the predictions. Thus, removing outliers can help to improve the accuracy and performance of classification models.

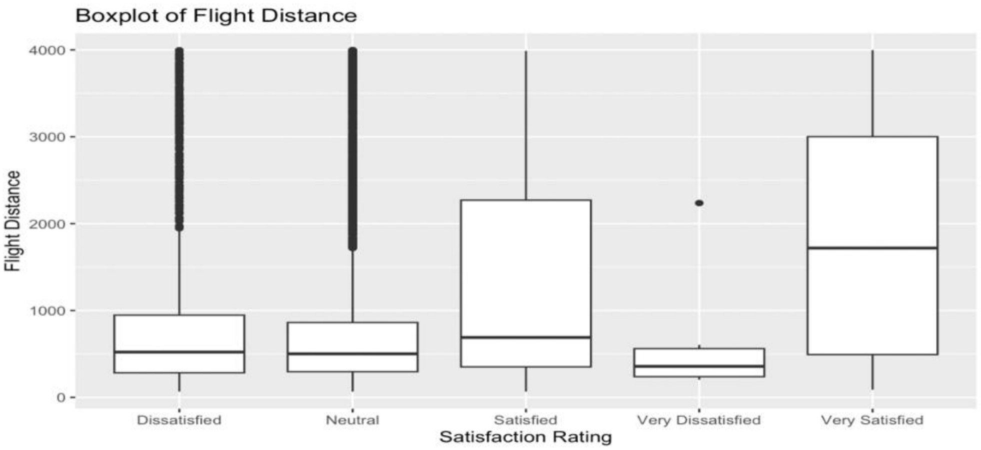

Figure 1: Boxplot for continuous variable "flight distance" against the target variable

Using “satisfaction rating” as the target variable for all five classification models in an airline dataset can help airlines gain insights into customer needs and preferences and improve their services.

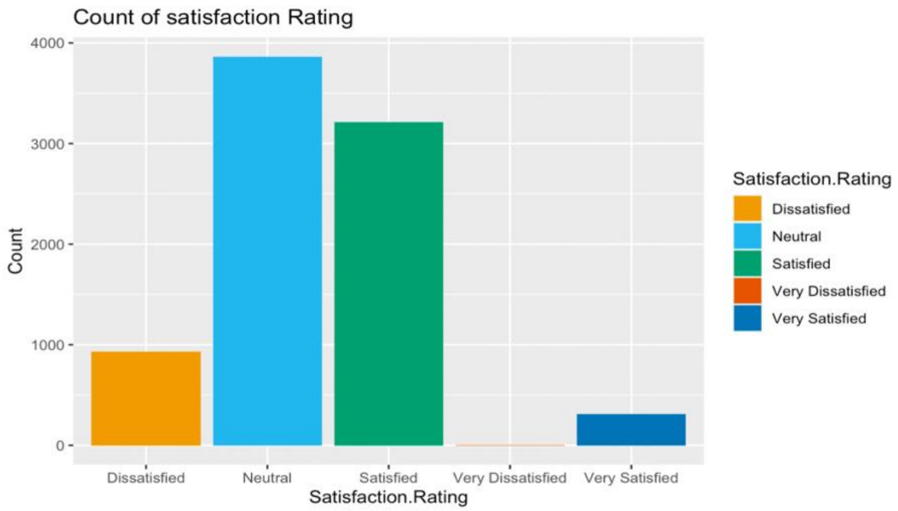

Figure 2: Bar plot for the target variable “satisfaction rating”

III. METHODOLOGY

The airline dataset is divided into a 60% training set and a 40% testing dataset for all five classification models.

A. Support Vector Machine:

The Support Vector Machine (SVM) classification model predicts customer satisfaction. Using a linear kernel, the SVM model trains on the training set and predicts on the test set. Model accuracy is determined by comparing predicted and actual values in the test set. A confusion matrix plot visualizes performance across various classes. Additionally, to assess overfitting, SVM model accuracy is calculated on the training set.

B. Decision tree:

- A decision tree is a graphical representation of a decision-making model in the form of a tree-like structure [2]. The decision tree model is trained on the train data set, and the accuracy is evaluated on the test data set. The accuracy of the decision tree model on the test data set and train dataset were calculated.

- The maximum depth of the tree was taken to be five for this model. Figure 3 illustrates the resulting decision tree obtained after fitting the data.

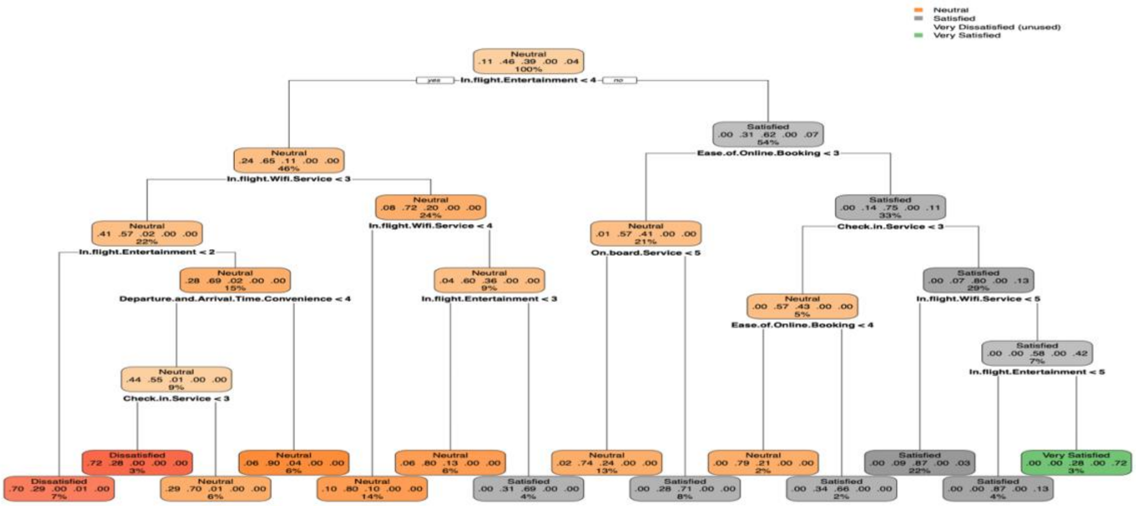
🟧 Neutral
▨ Satisfied
Very Dissatisfied (unused)
🟩 Very Satisfied
Neutral .11 .46 .39 .00 .04 100%
yes
In.flight.Entertainment < 4
no
Satisfied .00 .31 .62 .00 .07 54%
Ease.of.Online.Booking < 3
Neutral .24 .65 .11 .00 .00 46%
In.flight.Wifi.Service < 3
Neutral .08 .72 .20 .00 .00 24%
In.flight.Wifi.Service < 4
Satisfied .00 .14 .75 .00 .11 33%
Check.in.Service < 3
Neutral .41 .57 .02 .00 .00 22%
In.flight.Entertainment < 2
Neutral .01 .57 .41 .00 .00 21%
On.board.Service < 5
Neutral .28 .69 .02 .00 .00 15%
Departure.and.Arrival.Time.Convenience < 4
Neutral .04 .60 .36 .00 .00 9%
In.flight.Entertainment < 3
Satisfied .00 .07 .80 .00 .13 29%
In.flight.Wifi.Service < 5
Neutral .00 .57 .43 .00 .00 5%
Ease.of.Online.Booking < 4
Satisfied .00 .00 .58 .00 .42 7%
In.flight.Entertainment < 5
Neutral .44 .55 .01 .00 .00 9%
Check.in.Service < 3
Dissatisfied .72 .28 .00 .00 .00 3%
Neutral .06 .90 .04 .00 .00 6%
Neutral .06 .80 .13 .00 .00 6%
Neutral .02 .74 .24 .00 .00 13%
Neutral .00 .79 .21 .00 .00 2%
Satisfied .00 .09 .87 .00 .03 22%
Very Satisfied .00 .00 .28 .00 .72 3%
Dissatisfied .70 .29 .00 .01 .00 7%
Neutral .29 .70 .01 .00 .00 6%
Neutral .10 .80 .10 .00 .00 14%
Satisfied .00 .31 .69 .00 .00 4%
Satisfied .00 .28 .71 .00 .00 8%
Satisfied .00 .34 .66 .00 .00 2%
Satisfied .00 .00 .87 .00 .13 4%

Figure 3: Decision tree model on test dataset

C. *Random Forest:*

- The algorithm creates many decision trees (100), each using a random subset of the available data and features. It then combines the outputs of these trees to make a final prediction. This helps to reduce overfitting and improve the accuracy of the model. The accuracy metrics is used to predict the target variable “satisfaction rating” because it has more than two categories than using f1, recall as metrics to evaluate the model’s performance.

- 100 trees were built before taking the averages of predictions while fitting the random forest classifier.

D. *Naive bayes:*

The Naive Bayes function is used to train the model. The function takes the response variable and the predictors as input and returns a trained Naive Bayes model. The “predict” function is then used to make predictions on the test dataset. Finally, the accuracy of the model on the training dataset is calculated and the same procedure for the test dataset.

E. *K Nearest Neighbor:*

KNN (K-Nearest Neighbor) is a classification algorithm that works based on the distance between the observations [3]. The KNN model is trained and make the prediction on test dataset and the accuracy based on training and testing dataset were calculated to check the model’s performance.

IV. RESULT

A. The confusion matrix for all five-classification model were plotted.

The confusion matrix analyzes the performance metric “accuracy” of each classification model by providing information on the number of true positives, true negatives, false positives, and false negatives [4]. A true positive occurs when the model correctly predicts that a customer is satisfied, a true negative occurs when the model correctly predicts that a customer is dissatisfied, a false positive occurs when the model predicts that a customer is satisfied but they are dissatisfied, and a false negative occurs when the model predicts that a customer is dissatisfied but they are satisfied.

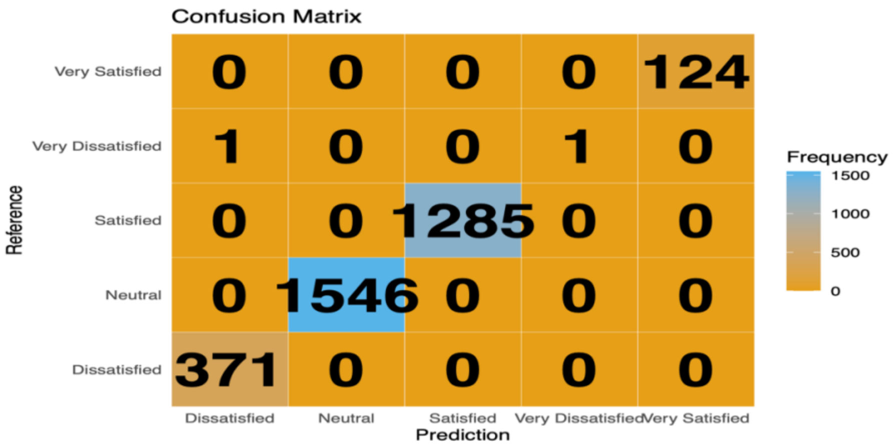

Figure 4: Confusion matrix for SVM model

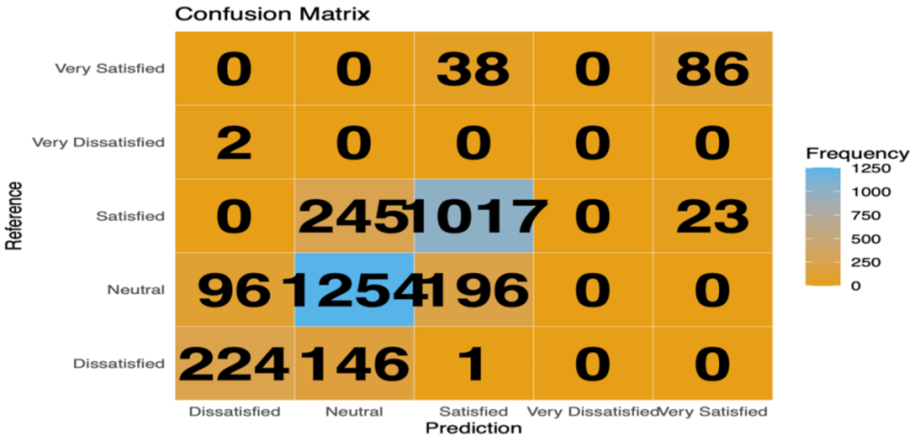

Figure 5: Confusion matrix for Decision tree

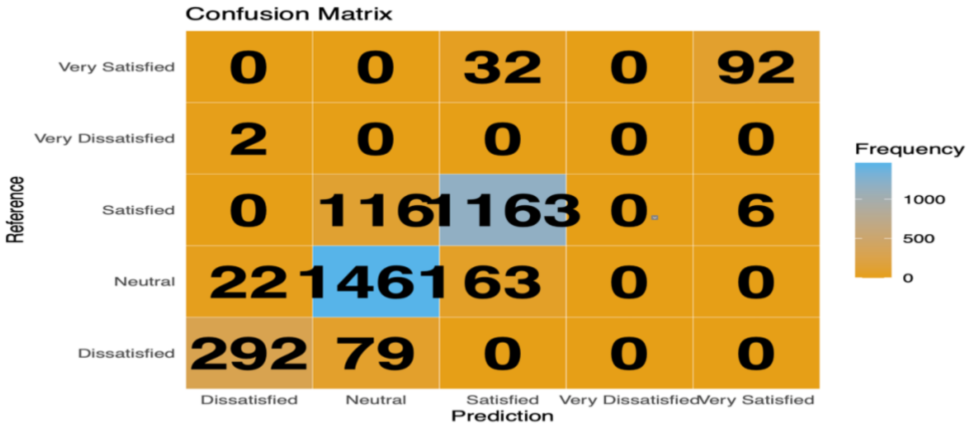

Figure 6: Confusion matrix for Random Forest model

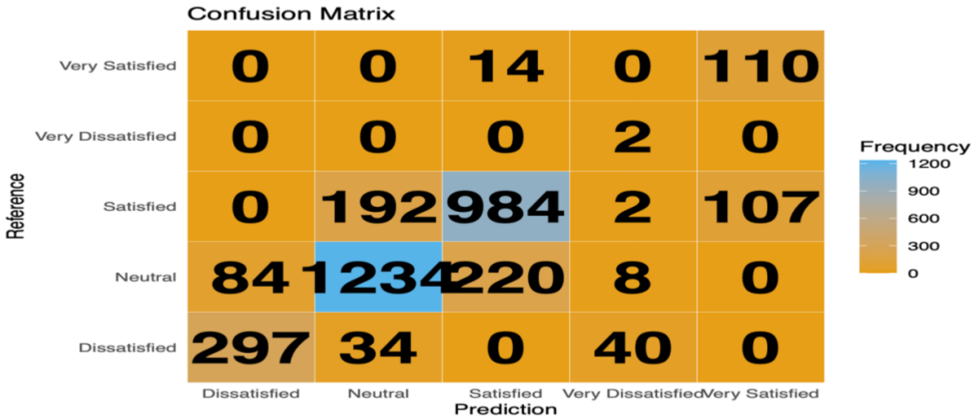

Figure 7: Confusion matrix for Naive bayes model

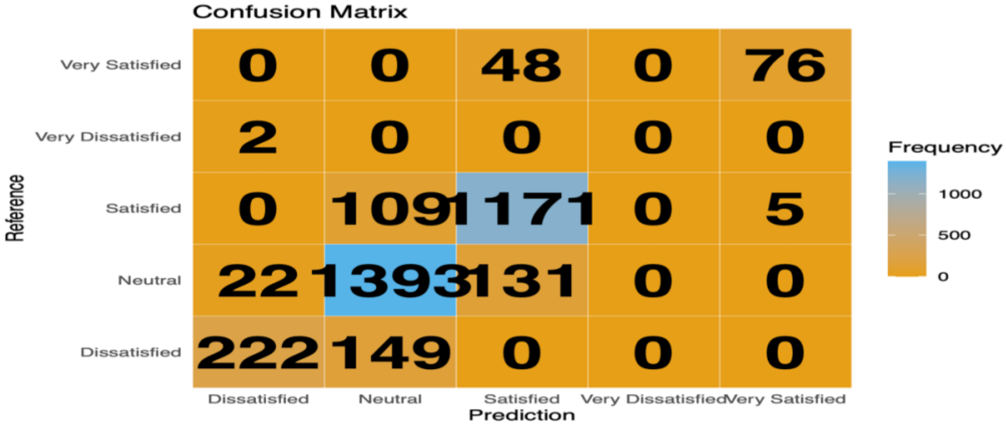

Figure 8: Confusion matrix for KNN

B. Comparison of five classification models based on testing and training accuracy.

This table shows the accuracy of the testing and training dataset by five classification methods.

ACCURACY	**TESTING_ACCURACY**	**TRAINING_ACCURACY**
Support vector machine	99.97	99.98
Decision tree	77.55	78.44
Random forest	90.38	99.96
Naive bayes	78.94	79.28
K nearest neighbor	86.0	89.35

Table 1: Testing and Training accuracy

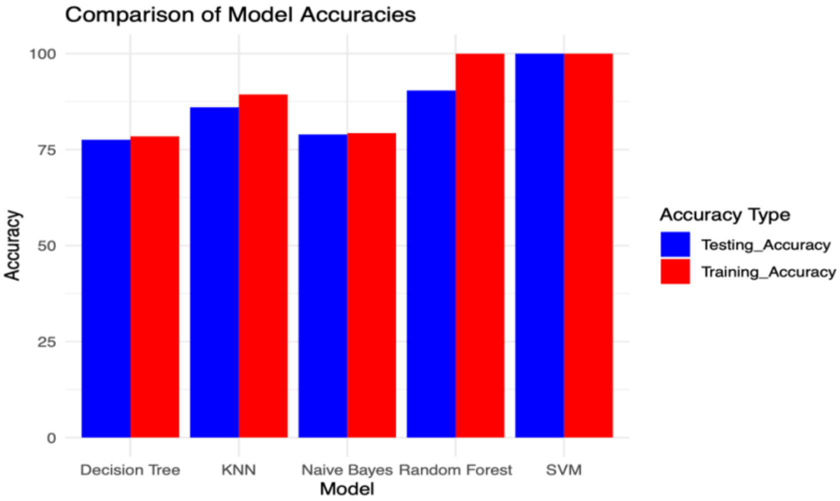

Figure 9: Bar chart for the accuracy comparison

- Figure 9 shows the testing accuracy and training accuracy as blue and red respectively.
- Overfitting happens when the model fits well in training data, and not much effective to unseen test data [5].
- This will be accepted that there is a difference in accuracy between training data and test data.
- And that difference must be larger (Train set accuracy>> Test set accuracy)
- In SVM classification model, testing and training set having lesser difference in their accuracy. Since the accuracy is **99.97%** and **99.98%** for testing and training dataset respectively, the model’s performance is good and there is no overfitting of model.
- In decision tree, the difference of accuracy is lesser and comparatively poor performance of model since the accuracy is around **77.55%** and **78.44%** for testing and training data respectively from Table 1.

- Though the random forest model having bigger difference between the accuracy, we conclude this model is not overfitting and having good model performance since the accuracy is around **90.38%** and **99.96%** for test and train respectively from Table 1.
- The naive bayes model's performance is poor when compared with SVM and random forest and the accuracy difference is small, so there is no overfitting of model.
- KNN model is slightly performing well than decision tree and naive bayes and this model also not overfits.

V. CONCLUSION

In conclusion, the SVM and random forest models performed well with high accuracy on both training and testing datasets, indicating they are not overfitting, and can be used for airline prediction tasks such as flight delays or cancellations. The KNN models also performed decently with a lesser difference in accuracy between training and testing data, while the naive bayes model and decision tree showed poor performance. These models can be further improved by incorporating new data and variables for further studies in airline satisfaction rating.

VI. REFERENCES

[1] Karthika Muruganandam. Airline Passenger Satisfaction (10K Dataset).2022. Retrieved from https://www.kaggle.com/datasets/karthikamuruganandam/dataset-airline-v4/download?datasetVersionNumber=1.

[2] Arun Saini, Dothang Truong, Jing Yu Pan. Airline efficiency and environmental impacts-Data envelopment analysis. International Journal of Transportation Science and Technology. 2022.ISSN 2046-0430.https://doi.org/10.1016/j.ijtst.2022.02.005.

[3] Zhang S, Li X, Zong M, Zhu X, Wang R. Efficient KNN Classification with Different Numbers of Nearest Neighbors. IEEE Trans Neural Netw Learn Syst. 2018. doi: 10.1109/TNNLS.2017.2673241.

[4] Ruuska S, Hämäläinen W, Kajava S, Mughal M, Matilainen P, Mononen J. Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle. Behav Processes. 2018. doi: 10.1016/j.beproc.2018.01.004.

[5] Subramanian J, Simon R. Overfitting in prediction models - is it a problem only in high dimensions? Contemp Clin Trials. 2013. doi: 10.1016/j.cct.2013.06.011.

[6] Huang Y, Fields KG, Ma Y. A Tutorial on Generative Adversarial Networks with Application to Classification of Imbalanced Data. Stat Anal Data Min. 2022. doi: 10.1002/sam.11570.