e-ISSN 2598-9596

COMPARISON OF BAGGING, BOOSTING, AND STACKING ENSEMBLE MODELS FOR AIRLINE CUSTOMER SATISFACTION ANALYSIS

[PERBANDINGAN MODEL ENSEMBLE BAGGING, BOOSTING, DAN STACKING UNTUK KEPUASAN PELANGGAN MASKAPAI PENERBANGAN]

Melvin Lee[1]*
[1]Department of Computer Science, Universitas Pelita Harapan,
Jl. Imam Bonjol No.6 Lantai 5 - 7, Petisah Tengah, Medan, Indonesia
*Corresponding author: 01671220001@student.uph.edu

ABSTRACT

By the end of COVID-19 pandemic and subsequent lockdowns last year, air travel has soared high, with an increase of 30.1% compared to last year according to one report. The rise of number of passengers means a good opportunity for the airline carriers to recoup losses due to lockdowns, and competition becomes heated as rival carriers try to lure new and old customers into their services. To remain competitive, more and more companies are turning towards machine learning to analyze large amounts of data to gain an edge towards their competitors, with ensemble learning being one of the many methods employed for the analysis work. In this study, Decision Tree, Random Forest, Boosting, and Stacking methods will be chosen for comparative study, which will be supplied with Airline Satisfaction dataset which is cleaned of null values and changing data types, for the study itself and then compared with each other using confusion matrix, precision-recall-f1-scoreaccuracy metrics, ROC curve, and feature importances. The results have shown that while the three chosen classifiers are almost similar in their overall success rate, with Bagging method reaching 96.117%, Boosting with a rate of 96.037%, and stacking with a rate of 96.264%, overall Stacking has the highest rate among all. These results show the almost negligible differences on all three main ensemble learning methods in terms of efficacy. Additional studies with larger datasets, and more varieties of ensemble learning methods can improve the overall judgement of the results.

Keywords: airline satisfaction; bagging; boosting; ensemble learning; stacking

ABSTRAK

Dengan berakhirnya pandemi COVID-19 dan lockdown yang terjadi tahun lalu, perjalanan udara melonjak tinggi, dengan peningkatan sebesar 30,1% dibandingkan tahun lalu menurut sebuah laporan. Peningkatan jumlah penumpang berarti peluang bagus bagi maskapai penerbangan untuk menutup kerugian akibat lockdown, dan persaingan menjadi memanas ketika maskapai pesaing mencoba memikat pelanggan baru dan lama untuk menggunakan layanan mereka. Agar tetap kompetitif, semakin banyak perusahaan yang beralih ke pembelajaran mesin untuk menganalisis data dalam jumlah besar guna mendapatkan keunggulan dibandingkan pesaing mereka, dengan pembelajaran ansambel menjadi salah satu dari banyak metode yang digunakan untuk pekerjaan analisis. Dalam studi ini, metode Decision Tree, Random Forest, Boosting, dan Stacking akan dipilih untuk studi komparatif, yang akan dilengkapi dengan dataset Kepuasan Maskapai yang dibersihkan dari *nilai null dan tipe data yang berubah, untuk studi itu sendiri dan kemudian dibandingkan dengan masing-masing metode. lainnya menggunakan matriks konfusi, metrik akurasi skor recall-f1, kurva ROC, dan kepentingan fitur. Hasilnya menunjukkan bahwa meskipun ketiga pengklasifikasi yang dipilih memiliki tingkat keberhasilan keseluruhan yang hampir serupa, dengan metode Bagging mencapai 96,117%, Boosting dengan tingkat 96,037%, dan penumpukan dengan tingkat 96,264%, secara keseluruhan Penumpukan memiliki tingkat tertinggi di antara pengklasifikasi lainnya. semua. Hasil ini menunjukkan perbedaan yang hampir dapat diabaikan pada ketiga metode pembelajaran ansambel utama dalam hal kemanjuran. Studi tambahan dengan kumpulan data yang lebih besar, dan lebih banyak variasi metode pembelajaran ansambel dapat meningkatkan penilaian hasil secara keseluruhan.*

Kata kunci: *airline satisfaction; bagging; boosting; ensemble learning; stacking*

INTRODUCTION

As of 2023, total passengers boarding airlines has increased by 30.1% compared to last year, showing strong recovery from COVID-19 pandemic and will continue to see a strong growth trend in the future (Airlines IATA, 2023), with another source predicting that global passenger traffic will fully recover by 2024 and may reach 9.4 billion passengers (Figure 1).

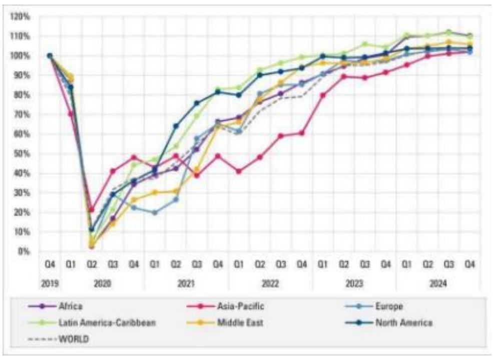


Figure 1. The passenger traffic by each region from 2019 to 2024 prediction. Source: (Airports Council International, 2023).

With the resurgence of airline traffic post-pandemic, airline industry will need to recover the losses sustained during the lockdown era (Nair, 2023) and the competition to obtain as many passengers as possible will be challenging as industries are struggling to survive 1 and recover (Bouwer *et al.*, 2021). To stay afloat and competitive, airliners must attract potential customers to them while building customer loyalty and recommendation, and one of the best ways to do so would be to increase customer satisfaction (Dong *et al.*, 2021).

With the complexity of identifying and analyzing the overall customer satisfaction, airline industries have turned to machine learning, specifically ensemble learning for making complex calculations and reporting on customer satisfaction analysis. Machine learning in its basic definition, describes the ability of a system to learn from given data related to analytics and solving given problems, which works by slowly learning meaningful patterns and relationships between pieces of data through examples and observations

(Janiesch *et al.*, 2021). Combining multiple machine learning algorithms will combine the output methods to perform more complex calculations for better results which is called Ensemble Learning (Zhou, 2009).

Past studies on analyses of a multi-dimensional problem have seen higher prediction accuracy using ensemble learning than single-based machine learning techniques (Akano & James, 2022), with many literature reviews on various ensemble learning techniques (Dong *et al.*, 2020). However, despite the strengths, ensemble learning process still have its weaknesses to be aware of, which despite several strategies and techniques still have limitations in terms of generalization, training difficulties, and more (Tasci *et al.*, 2021). Thus, this thesis which is titled "Comparison of Several Ensemble Models for Airline Customer Satisfaction", will firstly explore what factors will increase airline 2 satisfaction from past studies, and then applying them into the ensemble learning study and comparison. The study will then take an airline dataset which will then be pre-processed and cleaned before using it for both model training and testing, and then building each ensemble learning methods for testing to obtain performance results, which consists of ensemble learning performance metrics, an ROC and performance curve, and a confusion matrix, with which the results are then compared side-by-side with each ensemble learning methods to determine which among them have the best performance.

RESEARCH AND METHODOLOGY

Problem Limitations

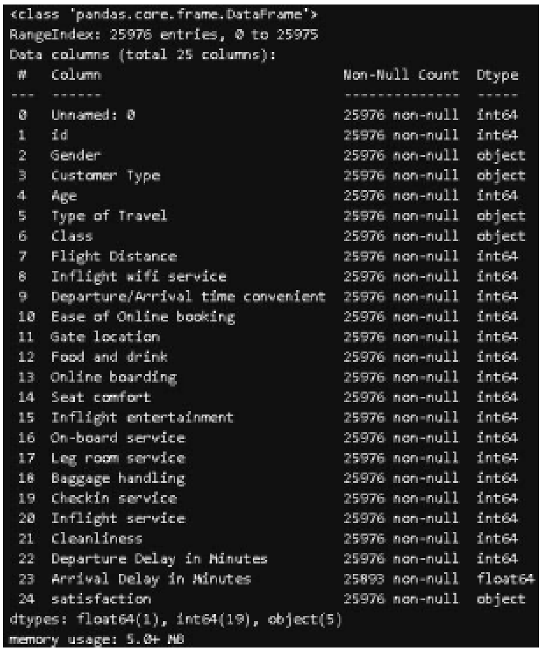
With the problems for this study identified, the next step will be to determine what focus will this study be, thus the limitations in this thesis are thus:

1. The number of Ensemble learning techniques to be used for this study.
2. The factors of airline satisfaction will be the values and parameters in a dataset which will be used for the ensemble learning study.
3. The expected results will be the accuracy and performance scores in the form of numeric values, which are supplied with charts supporting it.
4. The ensemble learning methods will be built and tested on Jupyter Notebook, using Python programming language, with NumPy, Seaborn, and Pandas plugin to facilitate data gathering, and model building.
5. The survey responses which will be used for the data collection will have data values in either Boolean or numeric values with no open-ended questions.

Research Methodology

Dataset

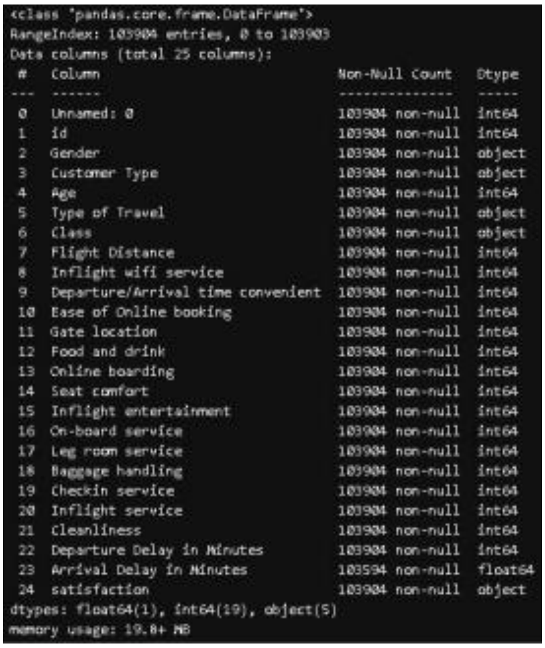
For this study, the dataset “Airline Passenger Satisfaction” and containing US passenger details, modified to be more cleaned-up from a previous dataset by the author TJ Klien, is obtained from an open-source dataset website Kaggle, with their rating for each of the airline’s aspects, which will be called “training” dataset, which features 25 columns, and numbering with 103.904 records.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25976 entries, 0 to 25975
Data columns (total 25 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Unnamed: 0                         25976 non-null  int64
 1   id                                 25976 non-null  int64
 2   Gender                             25976 non-null  object
 3   Customer Type                      25976 non-null  object
 4   Age                                25976 non-null  int64
 5   Type of Travel                     25976 non-null  object
 6   Class                              25976 non-null  object
 7   Flight Distance                    25976 non-null  int64
 8   Inflight wifi service              25976 non-null  int64
 9   Departure/Arrival time convenient  25976 non-null  int64
 10  Ease of Online booking             25976 non-null  int64
 11  Gate location                      25976 non-null  int64
 12  Food and drink                     25976 non-null  int64
 13  Online boarding                    25976 non-null  int64
 14  Seat comfort                       25976 non-null  int64
 15  Inflight entertainment             25976 non-null  int64
 16  On-board service                   25976 non-null  int64
 17  Leg room service                   25976 non-null  int64
 18  Baggage handling                   25976 non-null  int64
 19  Checkin service                    25976 non-null  int64
 20  Inflight service                   25976 non-null  int64
 21  Cleanliness                        25976 non-null  int64
 22  Departure Delay in Minutes         25976 non-null  int64
 23  Arrival Delay in Minutes           25893 non-null  float64
 24  satisfaction                       25976 non-null  object
dtypes: float64(1), int64(19), object(5)
memory usage: 5.0+ MB
```

Figure 2. Dataset attributes of the dataset "test".

The second dataset that will be used for this study is the “test” dataset (Figure 2), which contains similar columns and data types as the “training” dataset (Figure 3), has several 25.977 records present inside. The “test” dataset will be used for the actual evaluation of the ensemble learning methods, while the “training” dataset will be used for the model training for the ensemble learning methods.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103904 entries, 0 to 103903
Data columns (total 25 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   Unnamed: 0                         103904 non-null  int64
 1   id                                 103904 non-null  int64
 2   Gender                             103904 non-null  object
 3   Customer Type                      103904 non-null  object
 4   Age                                103904 non-null  int64
 5   Type of Travel                     103904 non-null  object
 6   Class                              103904 non-null  object
 7   Flight Distance                    103904 non-null  int64
 8   Inflight wifi service              103904 non-null  int64
 9   Departure/Arrival time convenient  103904 non-null  int64
 10  Ease of Online booking             103904 non-null  int64
 11  Gate location                      103904 non-null  int64
 12  Food and drink                     103904 non-null  int64
 13  Online boarding                    103904 non-null  int64
 14  Seat comfort                       103904 non-null  int64
 15  Inflight entertainment             103904 non-null  int64
 16  On-board service                   103904 non-null  int64
 17  Leg room service                   103904 non-null  int64
 18  Baggage handling                   103904 non-null  int64
 19  Checkin service                    103904 non-null  int64
 20  Inflight service                   103904 non-null  int64
 21  Cleanliness                        103904 non-null  int64
 22  Departure Delay in Minutes         103904 non-null  int64
 23  Arrival Delay in Minutes           103594 non-null  float64
 24  satisfaction                       103904 non-null  object
dtypes: float64(1), int64(19), object(5)
memory usage: 19.8+ MB
```

Figure 3. Dataset for the "training" dataset

There are some unneeded columns which will be removed in the pre-processing stage later, thus the description of each column for the “training” and “test” dataset which will be used for this study is as follows:

1. Gender: contains a binary data type between male and female.
2. Customer Type: contains binary data between loyal and disloyal customers.
3. Age: contains numerical values stating the actual age of a passenger.
4. Type of Travel: contains string values of the purpose of the flight of the passenger.
5. Class: contains string values of the type of flight taken, which is either business, economy, or other.

6. Flight Distance: contains integer values showing the flight distance of the travelling to their destination.
7. Departure Delay in Minutes: measures the delay on flight departure the passenger must tolerate.
8. Arrival Delay in Minutes: measures the delay on flight arrival the passenger must tolerate.

The following column within the dataset contains numerical values ranging from 1 to 5 on the level of the satisfaction with 1 being the lowest and 5 being the highest, and values of 0 means the passenger does not rate it:

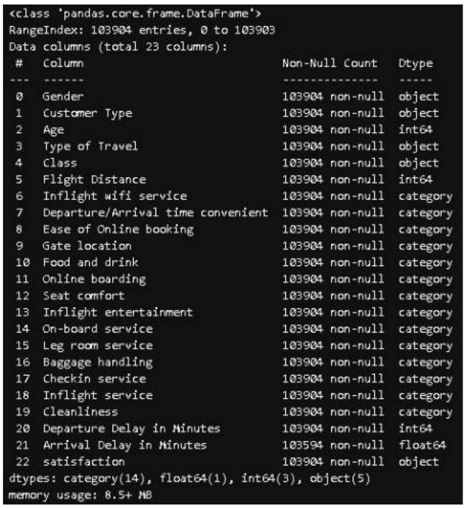
1. Inflight Wi-Fi Service: ratings of the satisfaction with Wi-Fi service onboard.
2. Departure/Arrival Time Convenient: ratings of the satisfaction with the departure/arrival time of a flight.
3. Ease of Online Booking: ratings of satisfaction on how easy it is to book a flight online.
4. Gate Location: ratings of the satisfaction of the location of the boarding gates of a flight.
5. Food and Drink: ratings of the satisfaction of the food and drinks provided on the flight.
6. Online Boarding: ratings of the satisfaction of the online boarding check in of the flight.
7. Seat Comfort: ratings of the satisfaction of how comfortable the seats in the flight are.
8. Inflight Entertainment: ratings of satisfaction on the entertainment options within the flight.
9. On-board Service: ratings of satisfaction on the services provided within the flight.
10. Leg Room Service ratings of satisfaction on the leg room services provided within the flight.
11. Baggage Handling: ratings of satisfaction on the handling of passengers' baggage by the flight.
12. Check-in Service: ratings of satisfaction on the check-in services provided by the flight.
13. Inflight Service: ratings of satisfaction on the other services provided by the flight during the flight.
14. Cleanliness: ratings of satisfaction on the overall cleanliness of the flight.

The last column consists of string values containing the overall rating of the flight by the passengers, which is divided into three types of responses: Satisfied, Neutral, and Dissatisfied.

Data Pre-Processing

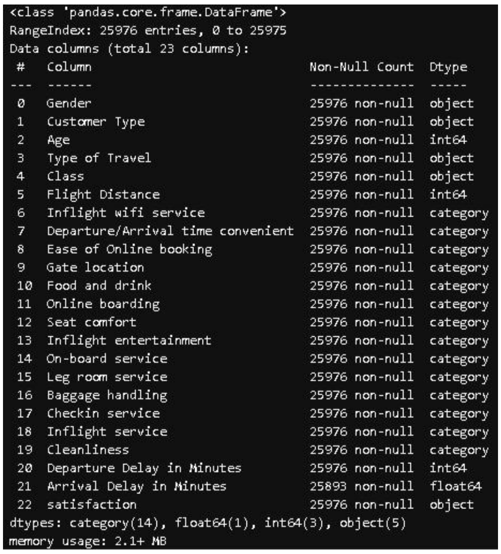
While both "training" and "test" datasets have been cleaned and pre-processed to be more concise, the dataset still contains null values, unnecessary columns and data, and improper data types which may affect the study itself, which is why this subsection will focus on further pre-pro. Figure 4 shows the dataset “training” after the data types have been fixed and Figure 5 shows the dataset “test” after the data types have been fixed. In both Figure 4 and Figure 5, the “unnamed column” together with “id” column as seen on Figure 1 has been removed, and most datatypes from number 6 to 19 have been changed into “category” data type to better fit for inputting on ensemble learning.

The next problem will be to solve the issue of missing values present on both datasets. Figure 6 shows the list of missing values on “training” dataset and Figure 7 is the “test” dataset showing the number of missing values present in the dataset. In Figure 6 and Figure 7, the number of missing values in Arrival Delay in Minutes column.



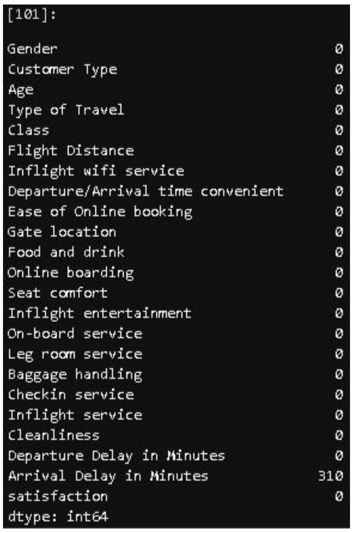
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 103904 entries, 0 to 103903
Data columns (total 23 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   Gender                             103904 non-null  object
 1   Customer Type                      103904 non-null  object
 2   Age                                103904 non-null  int64
 3   Type of Travel                     103904 non-null  object
 4   Class                              103904 non-null  object
 5   Flight Distance                    103904 non-null  int64
 6   Inflight wifi service              103904 non-null  category
 7   Departure/Arrival time convenient  103904 non-null  category
 8   Ease of Online booking             103904 non-null  category
 9   Gate location                      103904 non-null  category
 10  Food and drink                     103904 non-null  category
 11  Online boarding                    103904 non-null  category
 12  Seat comfort                       103904 non-null  category
 13  Inflight entertainment             103904 non-null  category
 14  On-board service                   103904 non-null  category
 15  Leg room service                   103904 non-null  category
 16  Baggage handling                   103904 non-null  category
 17  Checkin service                    103904 non-null  category
 18  Inflight service                   103904 non-null  category
 19  Cleanliness                        103904 non-null  category
 20  Departure Delay in Minutes         103904 non-null  int64
 21  Arrival Delay in Minutes           103594 non-null  float64
 22  satisfaction                       103904 non-null  object
dtypes: category(14), float64(1), int64(3), object(5)
memory usage: 8.5+ MB
```

Figure 4. The data attributes of the dataset "training" after the fix.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25976 entries, 0 to 25975
Data columns (total 23 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Gender                             25976 non-null  object
 1   Customer Type                      25976 non-null  object
 2   Age                                25976 non-null  int64
 3   Type of Travel                     25976 non-null  object
 4   Class                              25976 non-null  object
 5   Flight Distance                    25976 non-null  int64
 6   Inflight wifi service              25976 non-null  category
 7   Departure/Arrival time convenient  25976 non-null  category
 8   Ease of Online booking             25976 non-null  category
 9   Gate location                      25976 non-null  category
 10  Food and drink                     25976 non-null  category
 11  Online boarding                    25976 non-null  category
 12  Seat comfort                       25976 non-null  category
 13  Inflight entertainment             25976 non-null  category
 14  On-board service                   25976 non-null  category
 15  Leg room service                   25976 non-null  category
 16  Baggage handling                   25976 non-null  category
 17  Checkin service                    25976 non-null  category
 18  Inflight service                   25976 non-null  category
 19  Cleanliness                        25976 non-null  category
 20  Departure Delay in Minutes         25976 non-null  int64
 21  Arrival Delay in Minutes           25893 non-null  float64
 22  satisfaction                       25976 non-null  object
dtypes: category(14), float64(1), int64(3), object(5)
memory usage: 2.1+ MB
```

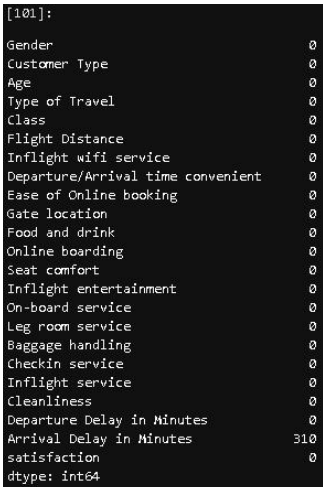
Figure 5. Dataset attributes of the dataset "test" after the fix.



```
[101]:

Gender                                 0
Customer Type                          0
Age                                    0
Type of Travel                         0
Class                                  0
Flight Distance                        0
Inflight wifi service                  0
Departure/Arrival time convenient      0
Ease of Online booking                 0
Gate location                          0
Food and drink                         0
Online boarding                        0
Seat comfort                           0
Inflight entertainment                 0
On-board service                       0
Leg room service                       0
Baggage handling                       0
Checkin service                        0
Inflight service                       0
Cleanliness                            0
Departure Delay in Minutes             0
Arrival Delay in Minutes             310
satisfaction                           0
dtype: int64
```

Figure 6. The list of total missing values on each column of the "training" dataset.

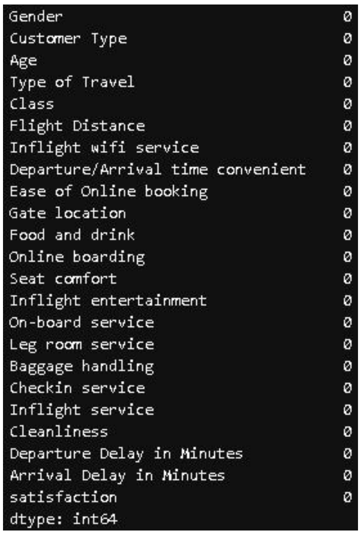


```
[101]:

Gender                                 0
Customer Type                          0
Age                                    0
Type of Travel                         0
Class                                  0
Flight Distance                        0
Inflight wifi service                  0
Departure/Arrival time convenient      0
Ease of Online booking                 0
Gate location                          0
Food and drink                         0
Online boarding                        0
Seat comfort                           0
Inflight entertainment                 0
On-board service                       0
Leg room service                       0
Baggage handling                       0
Checkin service                        0
Inflight service                       0
Cleanliness                            0
Departure Delay in Minutes             0
Arrival Delay in Minutes             310
satisfaction                           0
dtype: int64
```

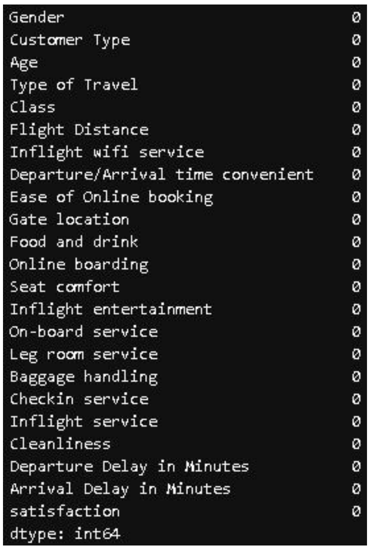
Figure 7. The list of total missing values on each column of the “test" dataset.

Figure 8 is the “training” dataset after the missing values have been fixed and Figure 9 is the “test” dataset after the missing values have been fixed.



```
Gender                                0
Customer Type                         0
Age                                   0
Type of Travel                        0
Class                                 0
Flight Distance                       0
Inflight wifi service                 0
Departure/Arrival time convenient     0
Ease of Online booking                0
Gate location                         0
Food and drink                        0
Online boarding                       0
Seat comfort                          0
Inflight entertainment                0
On-board service                      0
Leg room service                      0
Baggage handling                      0
Checkin service                       0
Inflight service                      0
Cleanliness                           0
Departure Delay in Minutes            0
Arrival Delay in Minutes              0
satisfaction                          0
dtype: int64
```

Figure 8. The list of total missing values on the "training" dataset after the fix.



```
Gender                                0
Customer Type                         0
Age                                   0
Type of Travel                        0
Class                                 0
Flight Distance                       0
Inflight wifi service                 0
Departure/Arrival time convenient     0
Ease of Online booking                0
Gate location                         0
Food and drink                        0
Online boarding                       0
Seat comfort                          0
Inflight entertainment                0
On-board service                      0
Leg room service                      0
Baggage handling                      0
Checkin service                       0
Inflight service                      0
Cleanliness                           0
Departure Delay in Minutes            0
Arrival Delay in Minutes              0
satisfaction                          0
dtype: int64
```

Figure 9. The list of total missing values on the "test" dataset after the fix.

Exploratory Data Analysis

To better understand the characteristics and gaining insight on the dataset used for the training, an exploratory data analysis will be conducted to learn more about the dataset itself before running

ensemble learning methods for testing, and to be able to predict the results more carefully. Since this study is all about airline satisfaction, the overall values of satisfied/dissatisfied or neutral within the dataset can be seen in Figure 10.

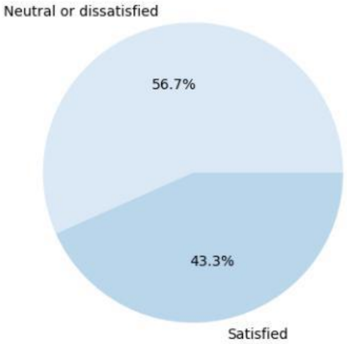


Figure 10. Pie chart showing the overall number of neutral or dissatisfied, and satisfied passengers within the dataset, overall “training” dataset is almost near balanced among both level of satisfaction.

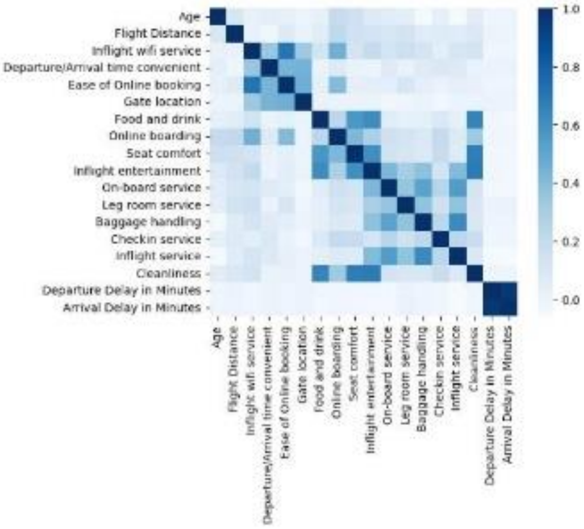


Figure 11. Correlation heatmap showing the relationship of each column with one another in ”training” dataset. Note the strong correlation between the column departure delay and arrival dela

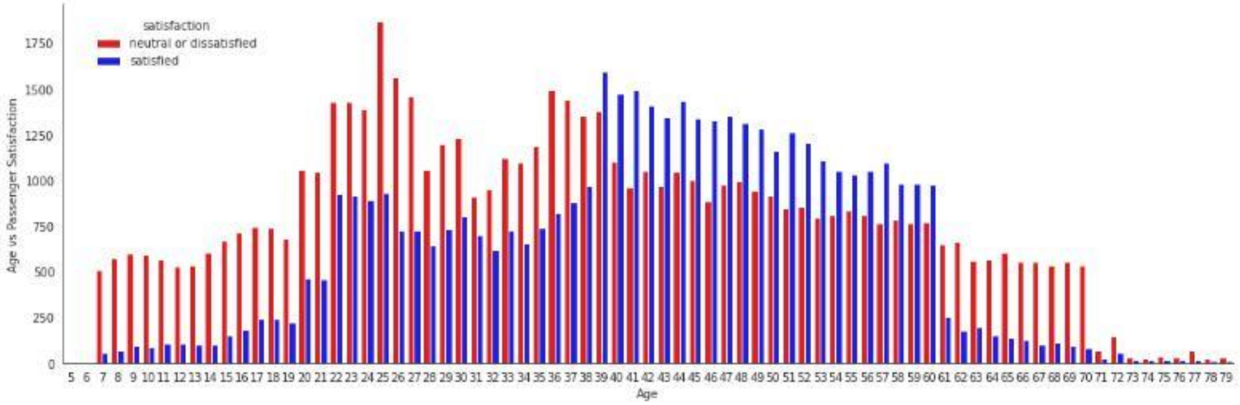


Figure 12. The correlation heatmap showing the relations between quantitative values within the dataset. Note the strong correlation between departure delay in minutes with the arrival delay in minutes.

Figure 11 shows the heatmap cluster between data columns within the “training” dataset. Note that “test” dataset, has similar columns and features as shown in Figure 3. Interestingly, Figure 12 shows the differences of satisfaction levels of ages from the youngest (5) to the oldest (79) within the dataset, and it shows that people aging from 39 - 60 usually have higher satisfaction rates than the ages ranging from 7 - 38, and 61 - 79.

Although gender can affect the ratings of certain aspects of an airline, Figure 13 shows that in overall satisfaction levels, there is only a slight difference, 32 where women tend to be a little bit more dissatisfied than the men, while satisfied levels remain the same for both genders.

The bias of customers based on the loyalty type, and data exploration has shown a stark contrast between loyal and disloyal customer types, with loyal customer rating more frequently than disloyal customers, and the level of satisfaction is the lowest on disloyal customer types (Figure 14).

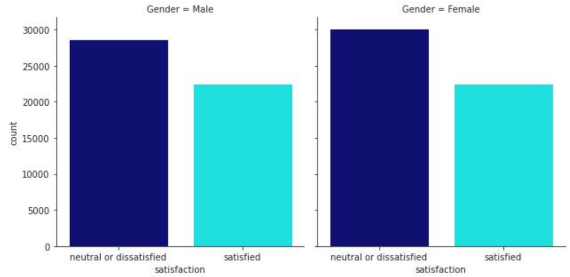


Figure 13. Bar chart showing the differences of satisfaction levels between the two genders.

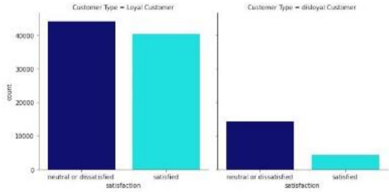


Figure 14. Bar chart showing the total number of satisfied and neutral or dissatisfied passengers, split between loyal and disloyal types of customer.

Figure 15 shows an interesting insight that the longer the flight distance, the higher the level of satisfaction of a passenger regarding the inflight entertainment and leg room service on average, showing that inflight entertainment and leg room service can be important factor in affecting airline satisfaction, with the caveat that the longer the distance, the more important it will be.

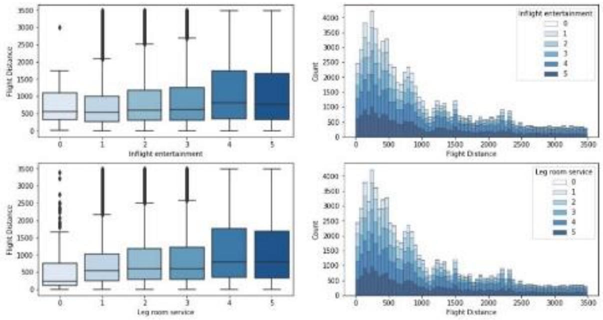
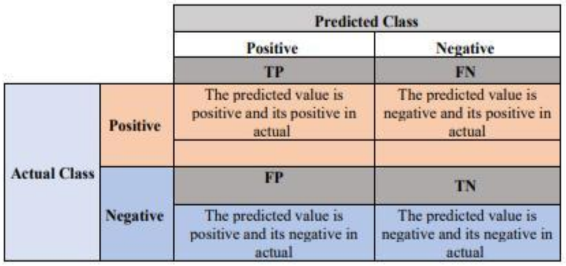


Figure 15. Box chart and histogram plot showing the relations of flight distance and in-flight entertainment with satisfaction levels.

Evaluation Method

To determine the performance of each ensemble learning models, there will be four metrics to be used for this study to measure its efficacy: Precision which is the accuracy of the model to predict positive labels from the given data, Recall which calculates how much actual positive data can be obtained by the model with the true positive data labels, F1 Score which is a calculation with weighting from the precision results, and Accuracy which measures how many times can the model classify data correctly. All of these four metrics will be laid out on tables comparing each methods to each other, supplied with confusion matrix showing the predicted values on four dimensions, which are: True Positive which means the model accurately predicts a positive data sample, False Negative where the model incorrectly predicts a negative data sample incorrectly, False Positive where the model incorrectly predicts a positive data sample, and True Negative where the model accurately predicts a negative data sample. Figure 16 shows how a confusion matrix would look like.



		Predicted Class	
		Positive	Negative
		TP	FN
Actual Class	Positive	The predicted value is positive and its positive in actual	The predicted value is negative and its positive in actual
	Negative	FP	TN
		The predicted value is positive and its negative in actual	The predicted value is negative and its negative in actual

Figure 16. Confusion matrix and what each value signifies.

For Bagging ensemble learning, the method used will be the Random Forest classifier, and the Decision Tree classifier methods, while the Boosting Method uses the XGBoost method which is one of the most used sub-methods in Boosting, and for the Stacking method, it will employ the standard Blending method which is widely used.

ROC, or Receiver Operating Characteristic will also be used for comparison and evaluation between the ensemble learning models, where the ROC shows the test accuracy where the closer the graph is to the top and left-hand border the more accurate the test is, vice versa. The test accuracy is also shown in the area under the curve where the greater the area under the curve, the more accurate the test is. Figure 17 shows what is an ROC curve.

Feature importance will also be used for the evaluation, where it lists all the available data columns used in machine learning and weighted with scores, which the higher the score, the more that data column will have a larger effect on the model that is being used for the prediction. Figure 18 shows the example of a feature importance.

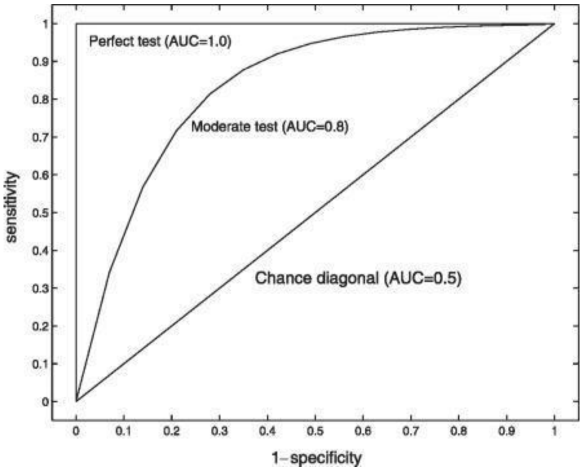
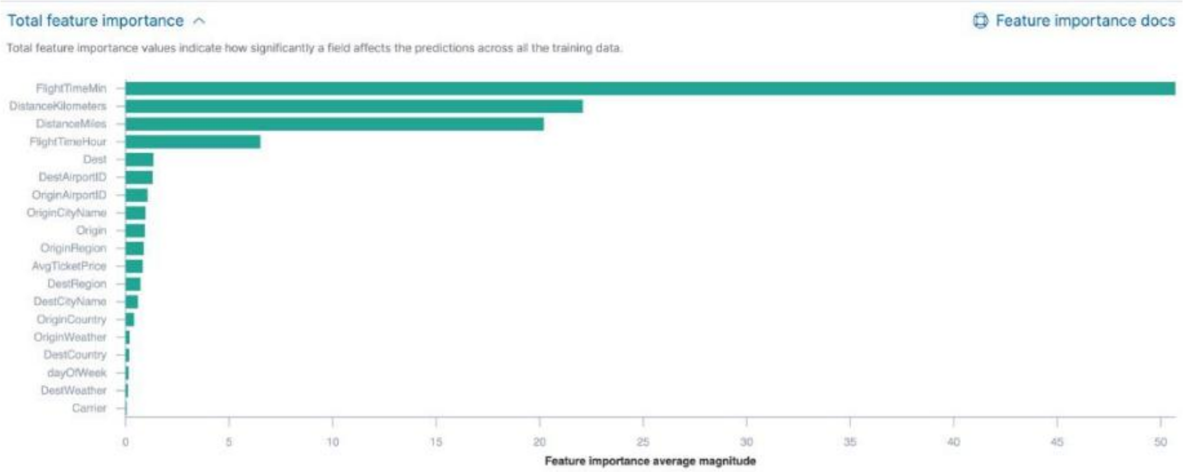


Figure 17. ROC curve, where scores above 0.5 and higher are accurate, while scores under 0.5 are less accurate. Values larger than 0.5 also indicate that model has an ability to discriminate


Total feature importance

Total feature importance values indicate how significantly a field affects the predictions across all the training data.

Figure 18. Bar chart of Feature Importance. The higher the score is, the more it will affect the overall model scoring

Calculation Methods

This following section will detail the calculation methods of all Ensemble learning methods and algorithms that will be used for the analysis, and the discussion for this study. The sampling done within this study will be stratified random sampling, grouped according to the age, time taken for the flight to arrive which in each row flight distance data is divided by 20 plus the arrival in minutes, and overtake which in each row, arrival delay in minutes is subtracted with the departure delay in minutes. The calculation present in this section will be divided into two parts: the calculation used to measure the results that will be used in the evaluation, and the calculation that will be used for the ensemble learning method algorithms itself.

To gauge the efficacy of the ensemble learning model, firstly the formula to calculate precision is as follows:

$$Precision = \frac{TP}{TP + FP}$$

Moving on, the equation to calculate the recall score is thus:

$$Recall = \frac{TP}{TP + FN}$$

These two results will then be used for the F1 Score calculation formula, which is:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

In addition to these three equations, the next and the last important equation is the accuracy of the ensemble learning model, which is:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Next up are the formulas of each ensemble learning models from chapter 2 which will be used for this study. The first formula to look up to will be the for Bagging:

$$f^{(x)} = \frac{1}{B}\sum_{B=1}^{B} f_{b(x)}$$

where $fb(x)$ represents the weak learners present in the machine learning model, B 1 generates the bootstrapping sets. Next is the equation for boosting method, which is:

$$f(x) = \sum_{t} a_t h_t(x)$$

where $ht(x)$ is created from several weak classifiers through training data and model building of it, which creates a second model that attempts to correct the errors, which is at. Finally, the formula for Stacking method is:

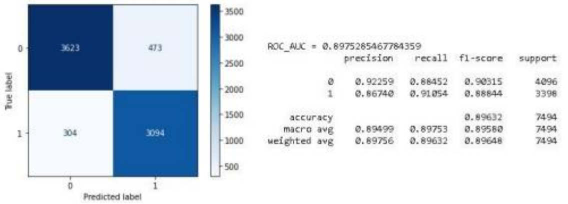
$$f_s(x) = \sum_{i=1}^{n} a_i f_i(x)$$

where $fi(x)$ is the output of the Nth base model, with N denoting the length of the dataset, ai denotes the weight of the Nth base model of the input X.

RESULTS AND DISCUSSION

Results

Figure 19, Figure 20, and Figure 21 are the test results on the Decision Tree ensemble learning method. Figure 22, Figure 23, and Figure 24 are the test results on the Random Forest ensemble learning method. Figure 25, Figure 26, and Figure 27 are the test results on the Boosting ensemble learning method. Figure 28, Figure 29, and Figure 30 are the test results on the Stacking ensemble learning method.



```
ROC_AUC = 0.8975285467784359
              precision    recall  f1-score   support

           0    0.92259   0.88452   0.90315      4096
           1    0.86740   0.91054   0.88844      3398

    accuracy                        0.89632      7494
   macro avg    0.89499   0.89753   0.89580      7494
weighted avg    0.89756   0.89632   0.89648      7494
```

Figure 19. Confusion matrix and precision, recall, f1-score, and support scores of Decision Tree ensemble learning method.

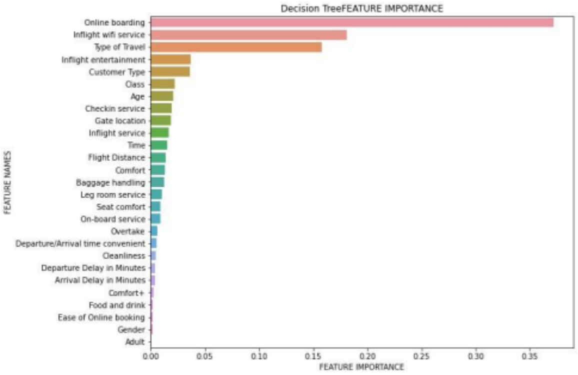


Figure 20. Feature importance for Decision Tree ensemble learning method.

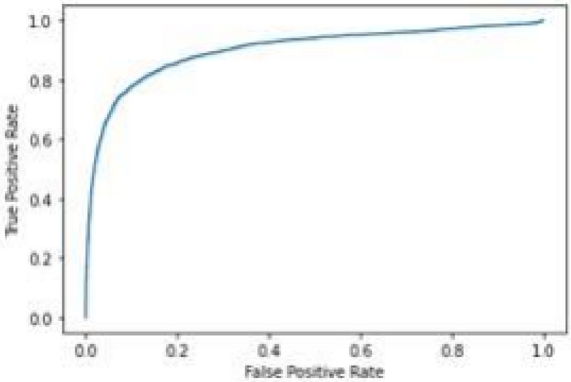
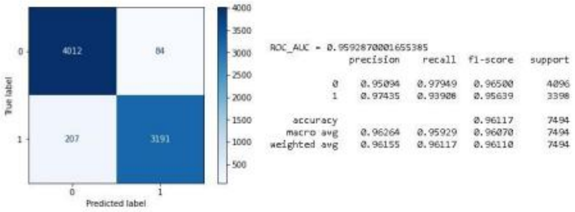


Figure 21. ROC curve for Decision Tree ensemble learning method.



```
ROC_AUC = 0.9592870001655385
              precision    recall  f1-score   support

           0    0.95094   0.97949   0.96500      4096
           1    0.97435   0.93908   0.95639      3398

    accuracy                        0.96117      7494
   macro avg    0.96264   0.95929   0.96070      7494
weighted avg    0.96155   0.96117   0.96110      7494
```

Figure 22. Confusion matrix and precision, recall, f1-score, and support scores for the Random Forest ensemble learning method.

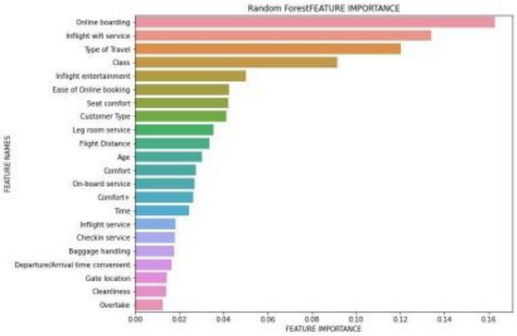


Figure 23. Features importance of Random Forest ensemble learning method.

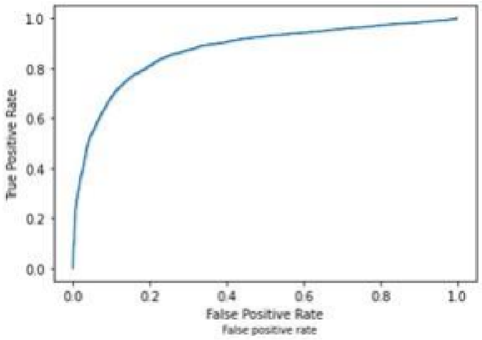
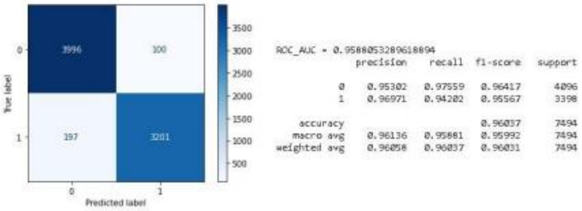


Figure 24. ROC curve of Random Forest ensemble learning method.



```
ROC_AUC = 0.9588053289618894
              precision    recall  f1-score   support

           0    0.95302   0.97559   0.96417      4096
           1    0.96971   0.94202   0.95567      3398

    accuracy                        0.96037      7494
   macro avg    0.96136   0.95881   0.95992      7494
weighted avg    0.96058   0.96037   0.96031      7494
```

Figure 25. Confusion matrix and precision, recall, f1-score, and support scores for the Boosting ensemble learning method.

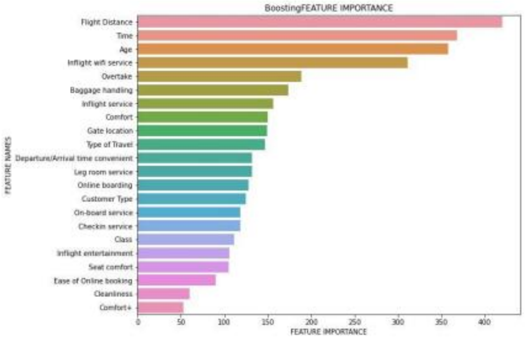


Figure 26. Features importance chart of the boosting ensemble learning method.

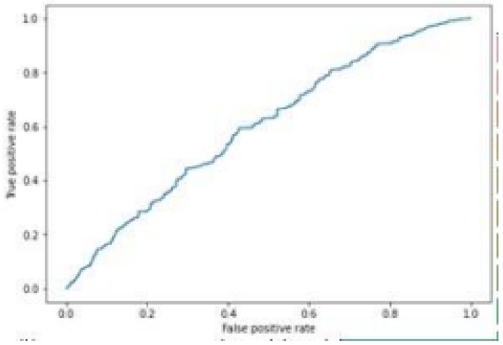
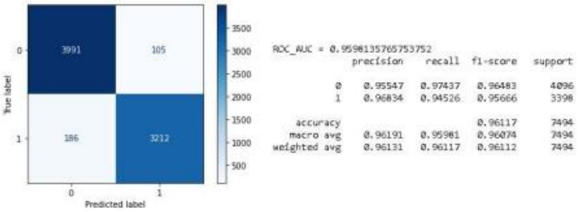


Figure 27. ROC curve of the Boosting ensemble learning method.



```
ROC_AUC = 0.9598135765753752
              precision    recall  f1-score   support

           0    0.95547   0.97437   0.96483      4096
           1    0.96834   0.94526   0.95666      3398

    accuracy                        0.96117      7494
   macro avg    0.96191   0.95981   0.96074      7494
weighted avg    0.96131   0.96117   0.96112      7494
```

Figure 28. Confusion matrix and precision, recall, f1-score, and support scores for Stacking ensemble learning method.

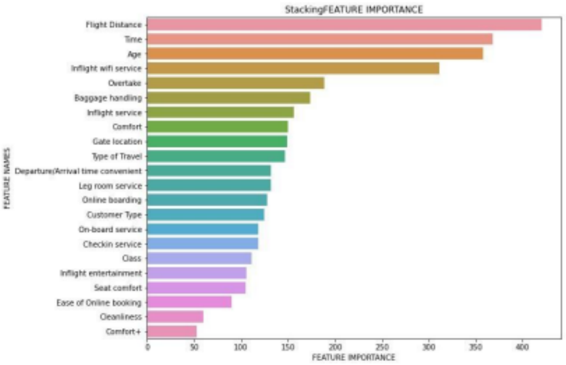


Figure 29. Features importance chart of the Stacking ensemble learning method.

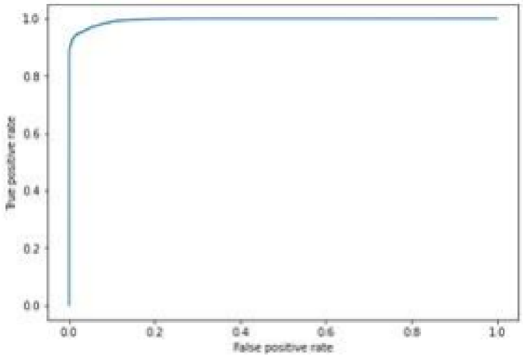


Figure 30. ROC curve of Stacking ensemble learning method.

Discussion

As shown on the charts and scores on all figures, the four methods, Random Forest, Decision Tree, Boosting, and Stacking, interestingly have almost reached parity with each other: Random Forest Bagging ensemble method has a success rate of 96.117%, Boosting has a success rate of 96.037%, and stacking has a success rate of 96.264%. However, when it comes to Decision Tree, it performs the worst among all, with a success rate of 89.63%, showing the inherent advantages of the Random Forest method in accuracy improvement over Decision Tree as discussed in chapter 2, with better recall score as shown in the confusion matrix, and less prone to false predictions. Interestingly, in terms of feature importance, the Decision Tree and Random Forest ensemble learning method, both belonging to Bagging method family on ensemble learning, has Online Boarding as the highest level of importance in affecting the overall study, while Flight Distance seems to affect the overall score the most on Stacking and Boosting method. The ROC curves have shown that overall, XGBoost has the worst graph shape of all, while Stacking has the best graph shape.

CONCLUSION

The ensemble learning methods used for this study, which are Random Forest and Decision Tree for Bagging methods, Boosting, and Stacking, have shown different results with each other. Random Forest together with Boosting and Stacking have a success rate of 96.117%, 96.037%, and 96.264% respectively. The Decision Tree method seemed to perform the most poorly with a success rate of 89.63%. Among all the ensemble learning methods, Stacking has the highest overall success rate, showing that ensemble learning can identify the aspects of satisfaction of each passenger towards the flight with many factors of satisfaction within the given dataset. Meanwhile, the worst performing ensemble learning method is the Decision Tree. This study hopefully acts as a guidance towards the ensemble learning enthusiasts of all levels of experience and providing some knowledge and insight towards those interested in gauging airline satisfaction. However, in the future research on ensemble learning, there will be plans to expand to more ensemble learning methods, with larger datasets for a better analysis in the efficacy of ensemble learning methods available.

REFERENCES

Airlines IATA. (2023). *Passenger demand posts solid growth*. Retrieved November 10, 2023 from https://airlines.iata.org/2023/11/10/passenger-demand-posts-solid-growth

Airports Council International. (2023). *The trusted source for air travel demand updates*. Retrieved September 27, 2023 from https://aci.aero/2023/09/27/global-passenger-traffic-expected-torecover-by-2024-and-reach-9-4-billion-passengers/

Akano, T. T., & James, C. C. (2022). An assessment of ensemble learning approaches and single-based machine learning algorithms for the characterization of undersaturated oil viscosity. *Beni-Suef University Journal of Basic and Applied Sciences*, *11*, 149. https://doi.org/10.1186/s43088-022-00327-8

Bouwer, J., Saxon, S., & Wittkamp, N. (2021). *Back to the future? Airline sector poised for change post COVID-19*. Retrieved October 4, 2023 from https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/ourinsights/back-to-the-future-airline-sector-poised-for-change-post-covid-19

Dong, Y., Liang, J., Zhao, Z., & Ding, D. (2021). Research on the relationship between customer satisfaction and compensation plan in U.S Airline industry. *Advances in Economics, Business and Management Research*, 506-510.

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, *14*, 241–258. https://doi.org/10.1007/s11704-019-8208-z

Janiesch, C., Zschech, P. & Heinrich, K. (2021). Machine learning and deep learning. *Electron Markets*, *31*, 685–695. https://doi.org/10.1007/s12525-021-00475-2

Nair, A. (2023). *Corona virus lockdown – A dramatic impact on the aviation industry*. Retrieved October 4, 2023 from https://straitsresearch.com/article/corona-virus-lockdown-a-dramatic-impact-on-the-aviation-industry

Tasci, E., Uluturk, C., & Ugur, A. (2021). A voting-based ensemble deep learning method focusing on image augmentation and preprocessing variations for tuberculosis detection. *Neural Computing and Applications*, *33*, 15541–15555. https://doi.org/10.1007/s00521-021-06177-2

Zhou, Z. H. (2009). Ensemble learning. In: Li, S. Z., & Jain, A. (Eds.) *Encyclopedia of biometrics* (pp. 270-273). Springer. https://doi.org/10.1007/978-0-387-73003-5_293