



IPSA

INSTITUT POLYTECHNIQUE DES SCIENCES
AVANCÉES

BE Ma412 - Air passenger satisfaction

JUNE 17TH 2024

MAGDELEINE DYLAN

Table of contents

1	Data Analysis and Understanding	1
2	Dimensionality Reduction and Exploration of the data with Principal Component Analysis (PCA)	6
3	Evaluation of Multiple Training Models	8
3.1	Logistic Regression	8
3.2	Support Vector Machine (SVM)	10
3.3	Random Forests	12
3.4	K-Neighbor	14
4	Selection of the Best Model	16
4.1	Selection Criteria	16
5	Conclusion	20

Introduction

Our project takes place during the global health crisis, in the aerospace sector. Our goal is to understand, assess, and improve passenger satisfaction to enhance this industry. Our project focuses on the analysis of data from a survey on air passenger satisfaction presented in an Excel spreadsheet. The objective is to develop a predictive model capable of determining the level of passenger satisfaction.

Our project will be divided into several phases. Initially, we will analyze our data and select an optimal machine learning model. To do this, we will use various machine learning techniques, including logistic regression, support vector machines (SVM), decision trees, and random forests. We will also explore dimensionality reduction methods such as Principal Component Analysis (PCA) and clustering techniques like K-Means. Finally, we will discuss the potential impact of this model on decisions and on improving customer satisfaction.

1 Data Analysis and Understanding

In the first step, we will analyse our data. We'll see by ourself how differents parameters impact the passagers satisfaction. To do it we will plot many graphics and analyse them. to analyse we will take the data *train.csv*.

First, for a better understanding of the data, we will see the proportion of flight distance and age.

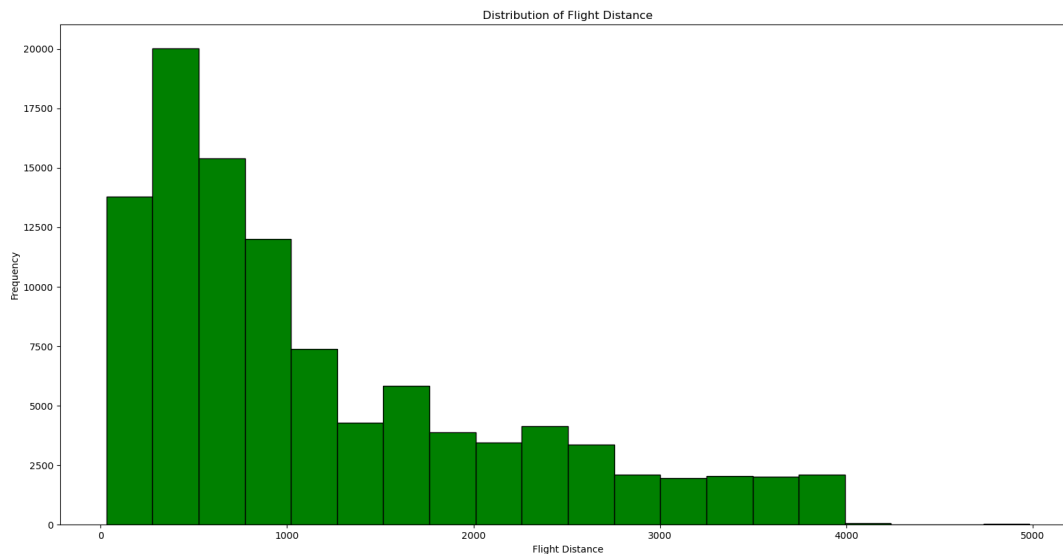


FIGURE 1 – fr quency of flight distance

We observe that in our data, we have more small flight distance than long distance. We have to take it in consideration, it may influence the passenger satisfaction.

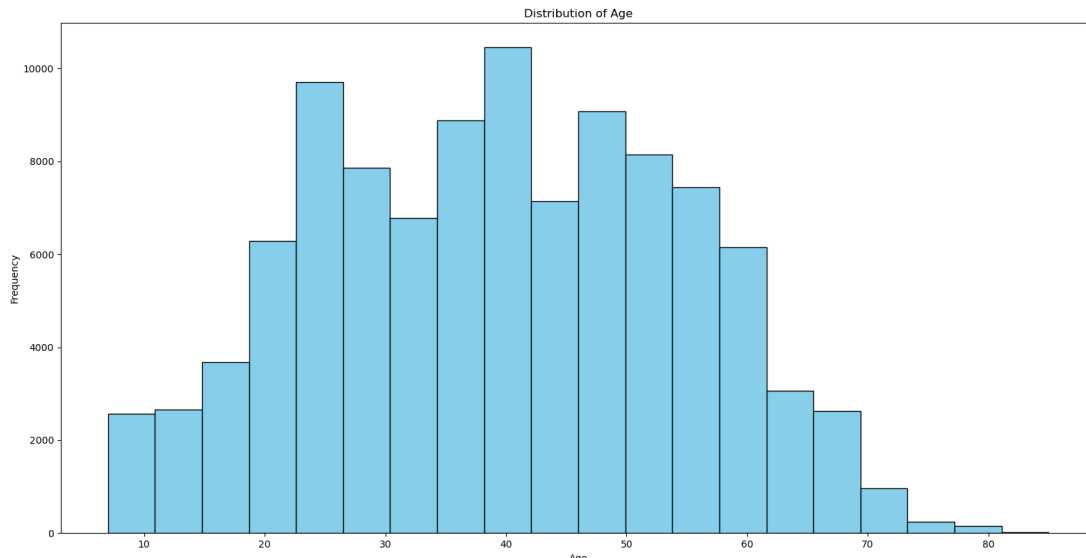


FIGURE 2 – frequency of age

We also have to know our client. So it is interesting to see the distribution of age of the passenger. We observe that the distribution of age is approximately like a Gaussian. There are not a lot of people traveling under 18 years old and over 65 years old. The age does not influence the passenger experience so every opinion counts, whatever the age.

Now we will try to explain which parameter influence the most the passenger satisfaction by plotting some graph. First, we could plot the passenger satisfaction in function of the traveling class.

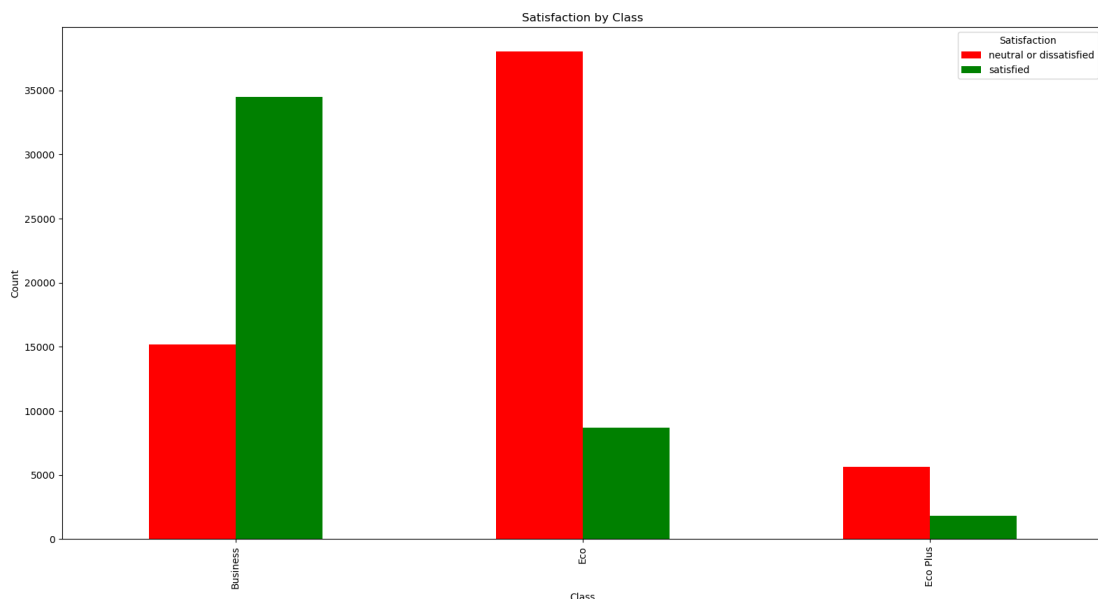


FIGURE 3 – Passenger satisfaction in function of the traveling class

Here on the graph, the green represent passengers satisfied and in red passengers who are neutral or dissatisfied. We can see that there are a lot of passengers who are neutral or dissatisfied when they travel in class Eco. Conversely, we observe that in class Business, passengers are most likely satisfied by there travel. In the class Eco-Plus we wee that the proportion is a little bit equivalent. It is explain by the fact that in Eco-Plus, the passenger does not have a lot of expectations.

Now we can compute the graph of the Passenger satisfaction in function of the type of travel.

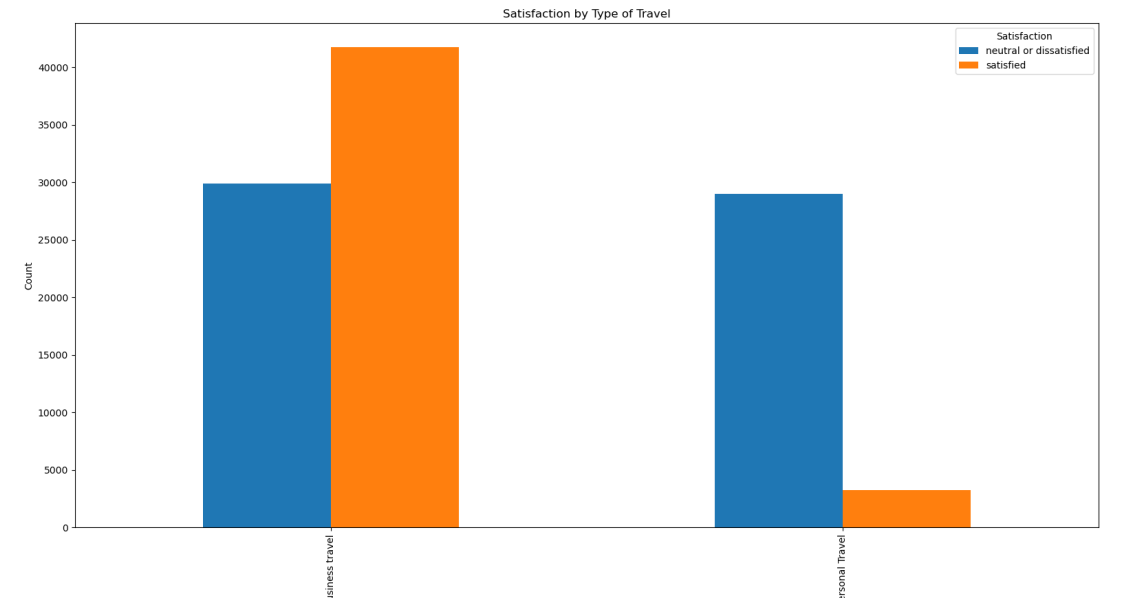


FIGURE 4 – Passenger satisfaction in function of the type of travel

We observe that when the travel is for business the opinions are balanced between satisfied and neutral/dissatisfied with a little advantage for the satisfied. For personal travel the passengers are most likely neutral or dissatisfied. To understand that, we can observe the proportion of type of travel (business or personal travel) by the Class choosen for these travels.

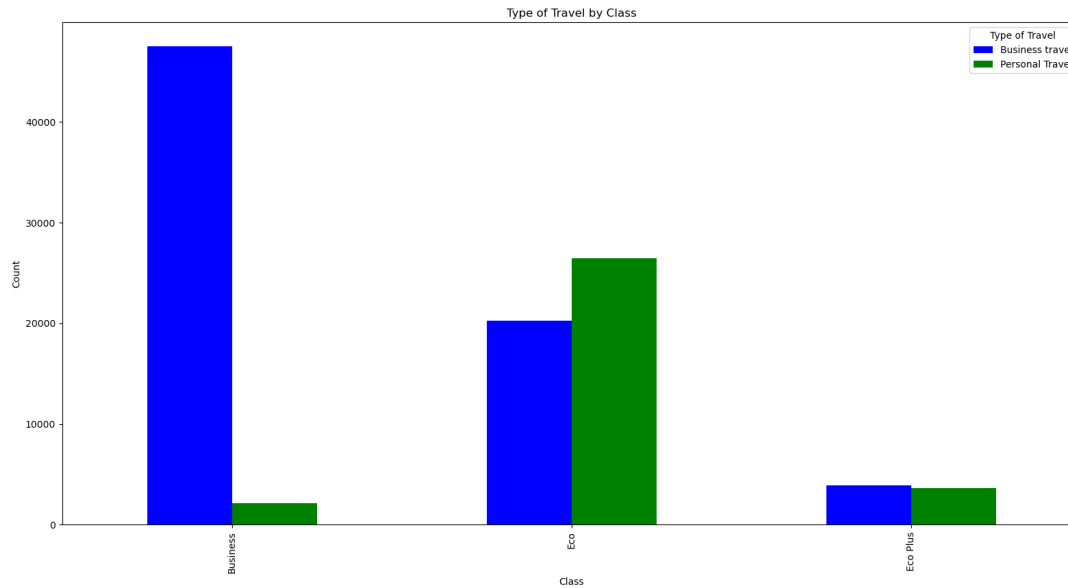


FIGURE 5 – Class in function of the type of travel

This graph shows us that for business travel, the clients choose most likely business class. In eco class the type of travel is balanced between personal and business travel. This is the same for Eco-Plus class. We saw previously that the clients traveling in business class and for business were satisfied. These data are correlated with the last graph we saw that when the clients travel for business, they choose business class.

Now we can compute two last graphs to see who (age) is traveling in which class and how they are satisfied.

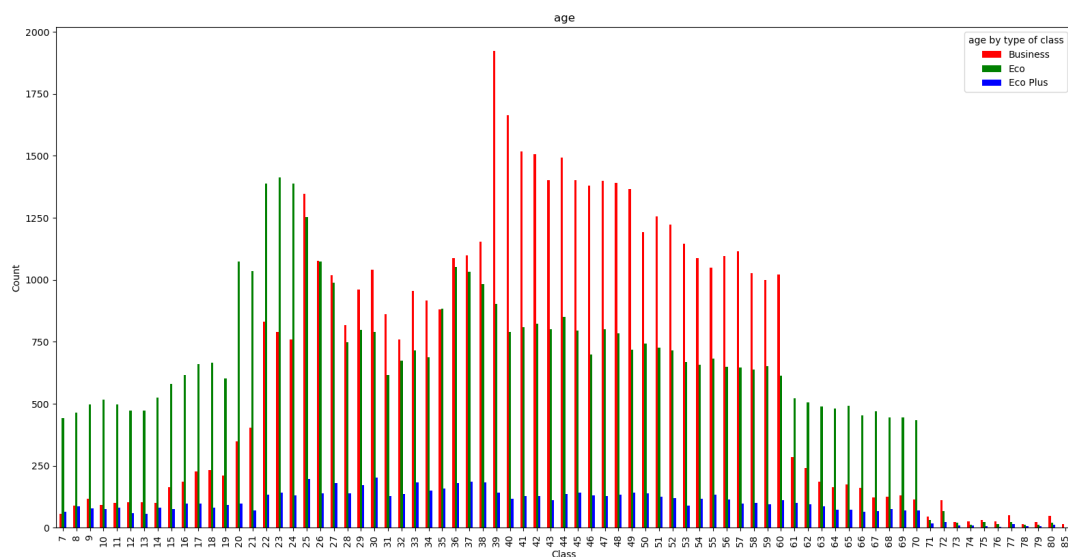


FIGURE 6 – Class in function of the age

On this first graph, we can see that between 26 and 60 years old passenger travel most likely in business. This is explain by the fact that people are working or have a good salary so they can travel in buisness class. We can observe that, for the Eco class, we have a pick in 20-26 years old group. Maybe because they have not a lot of money to spend in business travel. We see also that the Eco-Plus is approximately constent for all age.

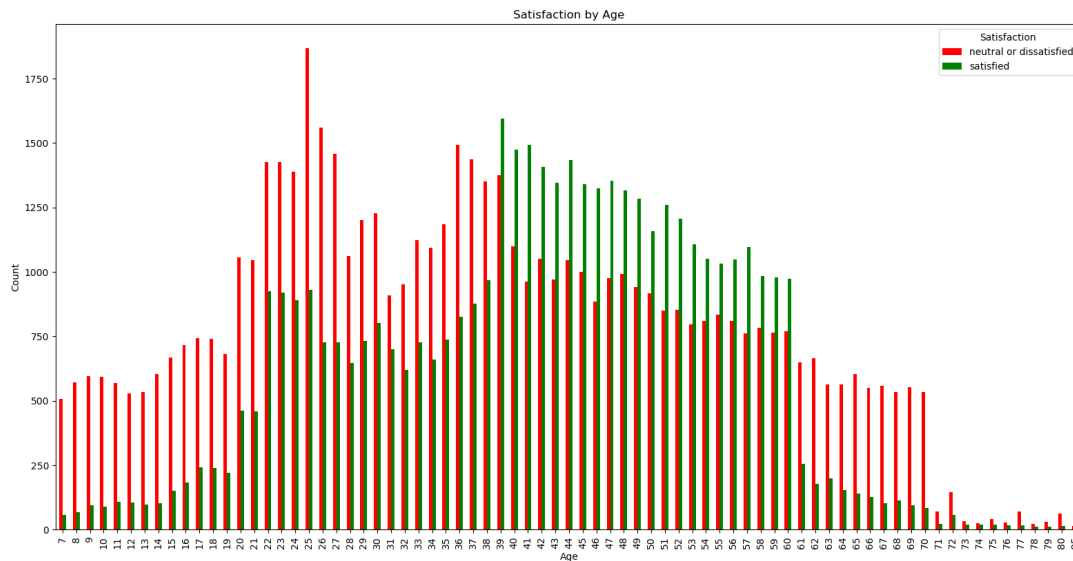


FIGURE 7 – satisfaction in function of age

This last graph show that the passengers are satisfied between 40 and 60 years old. These data is correlated with the precedent graph because we saw that this age group travel the most in business class. Moreover we also saw that passenger traveling in business class are most satisfied by there travel.

However we can see some data that not influence the satisfaction like the ID number or the gender. I have remove the column ID from the data for a better analysis. Also for the comming parts I've changed the value in some dataframe columns :

1. "Gender" : 1 for male and 0 for female.
2. "Satisfaction" : 1 for satisfied and 0 for neutral or dissatisfied.
3. "Type of travel" : 1 for business Travel and 0 for Personal Travel.
4. "Class" : 2 for business, 1 for Eco and 0 for Eco plus.
5. "Customer Type" : 1 for loyal customers and 0 for disloyal customer.

Now every columns of our data frame are number so they are more easier to manipulate to implement our learning models.

2 Dimensionality Reduction and Exploration of the data with Principal Component Analysis (PCA)

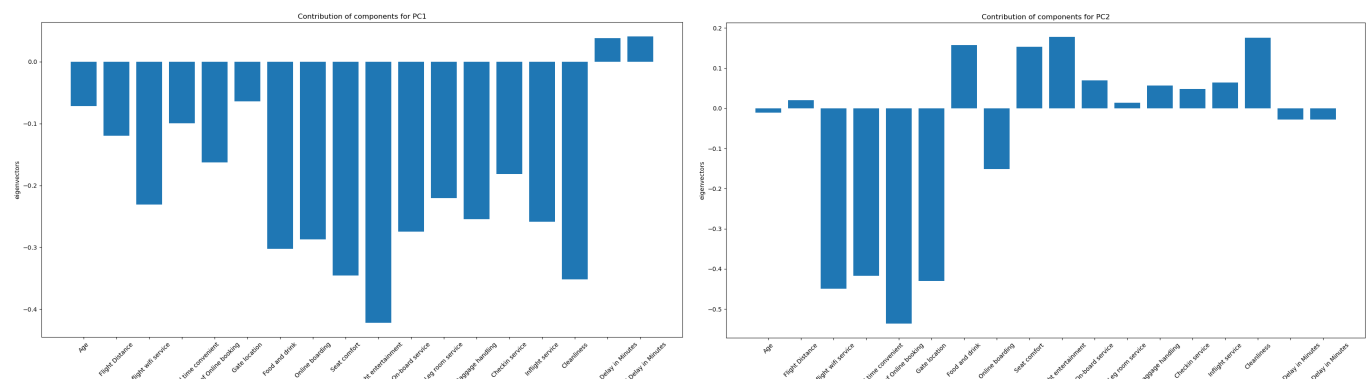
In this part, we will see two methods to reduce our data and find the components with more height. To do that we will apply the principal component analysis.

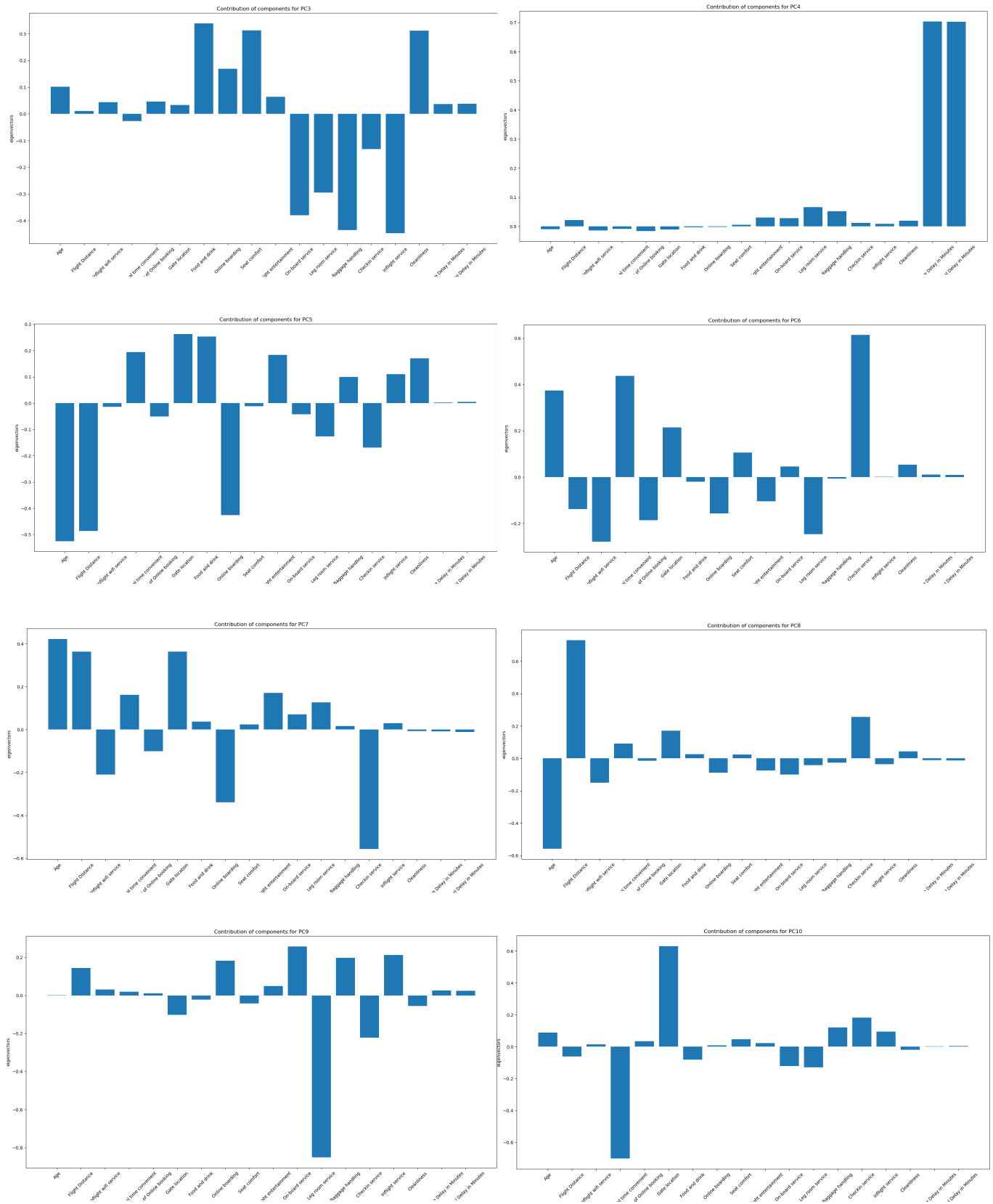
The principal component analysis is a technique that can be used to simplify a dataset. It is a linear transformation that chooses a new coordinate system for the data set such that greatest variance by any projection of the dataset comes to lie on the first axis (then called the first principal component), the second greatest variance on the second axis, and so on.

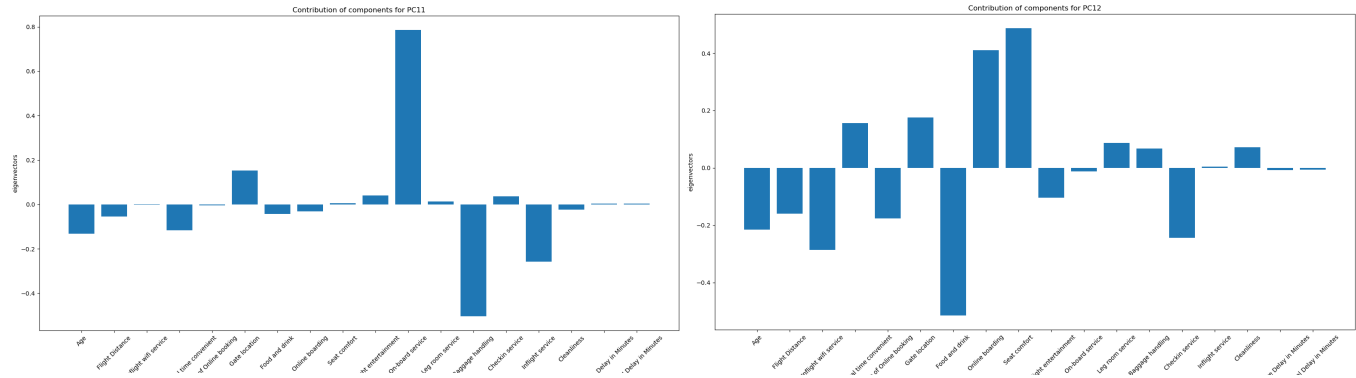
This technique is made of many steps :

1. **Step 1 : Standardization** (range of the continuous initial variables so that each one of them contributes equally to the analysis)
2. **Step 2 : Covariance matrix computation** (see if there is any relationship between variables)
3. **Step 3 : Compute the eigenvectors and eigenvalues of the covariance matrix** to identify the principal components (concepts that we need to compute from the covariance matrix)
4. **Step 4 : Analyse the height of component** (look wich component has the most height)

First, we have to know how many principal components we have to choose to have the better result. So I've made a code who choose only the explained variance over 90%. After that we know that the better number is 12. So we can compute the PCA, and extract the eigenvectors. With that we can see the different heights of all the component of our data. Here is the twelve graphs of each principal component.







Every bars (in the order left to right) represent : Age, Flight Distance, Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness, Departure Delay in Minutes, Arrival Delay in Minutes. The principal components don't correspond directly to specific variables in our original dataset. Each principal component is a linear combination of the original variables. They are calculated in such a way as to capture the maximum variance in the data, starting with the direction in which the data varies the most (PC1), then the second greatest variance (PC2), and so on. So to see the height of components we can observe the pick (negative or positive) of the bars.

3 Evaluation of Multiple Training Models

In this part we will explore some training models and see the different results. We'll see three models seen in course : Logistic Regression ; Support Vector Machine and Random Forests. Additionally, we'll test a model not seen in class : K-neighbor.

To do all these tests, we have to transform all the string variables into int variables. Next, we set X and Y variables for train data and test data. basically the Y represent the satisfaction column, and the X represent the rest of the dataframe. After this step, it's possible that some variables are lacking so we fill the dataframe. To finish we scale our variables X and Y for train and test. Now we can compute all the models.

3.1 Logistic Regression

Logistic regression is a supervised learning method used for binary classification. It models the probability of data belonging to a particular class (label 1) as a function of the values of its characteristics (features). The logistic function, or sigmoid, transforms a linear combination of features into a probability between 0 and 1. If this probability exceeds a certain threshold (generally 0.5), the observation is classified in the positive class (1) ; otherwise, it is classified in the negative class (0).

We can compute the confusion matrix for the Logistic Regression model.

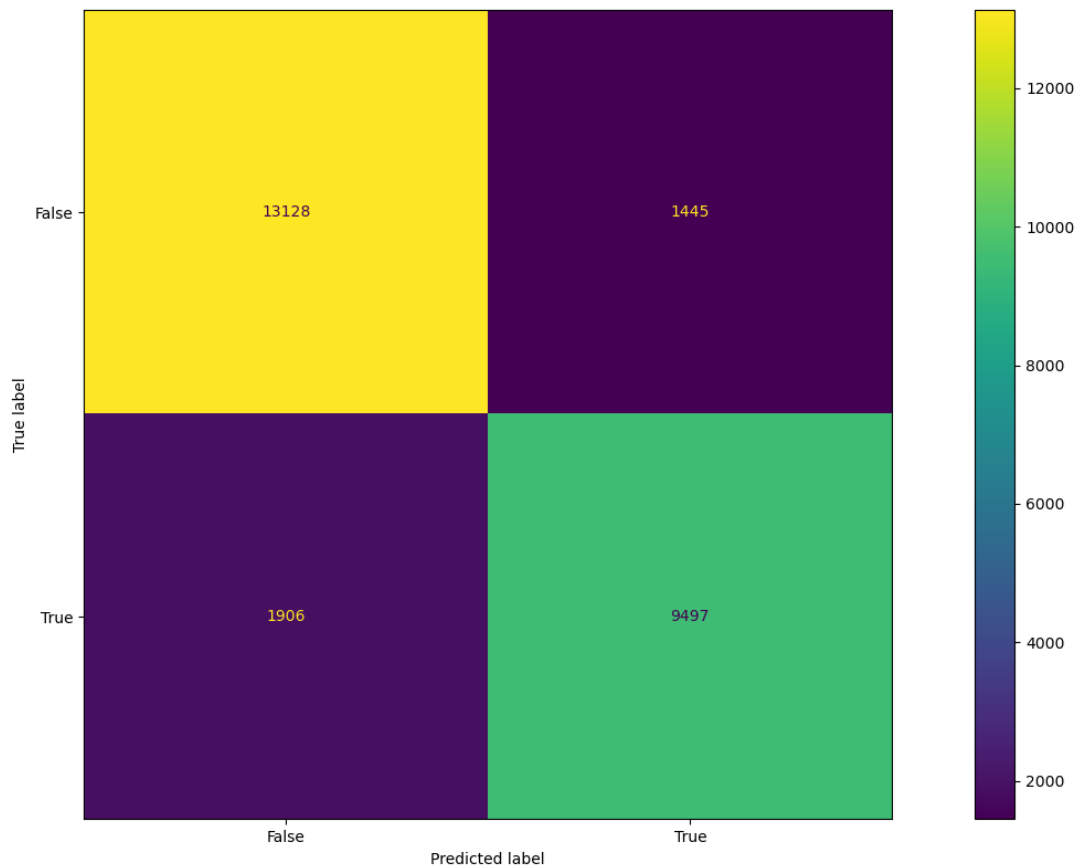


FIGURE 8 – Confusion matrix for Logistic Regression

True Negatives (False, False) : 13128 The model correctly predicted 13128 instances as being of class "False".

False Positives (False, True) : 1445

The model incorrectly predicted 1445 instances as being of class "True" when they were actually "False".

False Negatives (True, False) : 1906

The model incorrectly predicted 1906 instances as being of class "False" when they were "True".

True Positives : 9497

The model correctly predicted 9497 instances as being of the "True" class.

Now we can see the performance of the model.

```

Model: Logistic regression
Classification Report:
support
          0          0.87          0.90          0.89          14573
          1          0.87          0.83          0.85          11403

accuracy
macro avg          0.87          0.87          0.87          25976
weighted avg       0.87          0.87          0.87          25976

```

FIGURE 9 – Performance of the Logistic Regression model

Here is the table corresponding to the picture :

Classe	Précision	Rappel	F1-score	Support
0	0.87	0.90	0.89	14573
1	0.87	0.83	0.85	11403
Total	0.87	0.87	0.87	25976

We can see that Logistic regression model have 87% precision in total to predict who will be satisfied (1) and neutral or dissatisfied (0).

3.2 Support Vector Machine (SVM)

SVM are a supervised learning method used for classification and regression. In classification, an SVM seeks to find the optimal hyperplane that separates the data into two distinct classes. This hyperplane maximises the margin, i.e. the distance between the data points closest to the hyperplane (the support vectors) of the two classes. SVM can also use kernels to transform non-linear data into a higher-dimensional feature space where a linear hyperplane can be used for separation.

Here is the confusion matrix for Support Vector Machine :

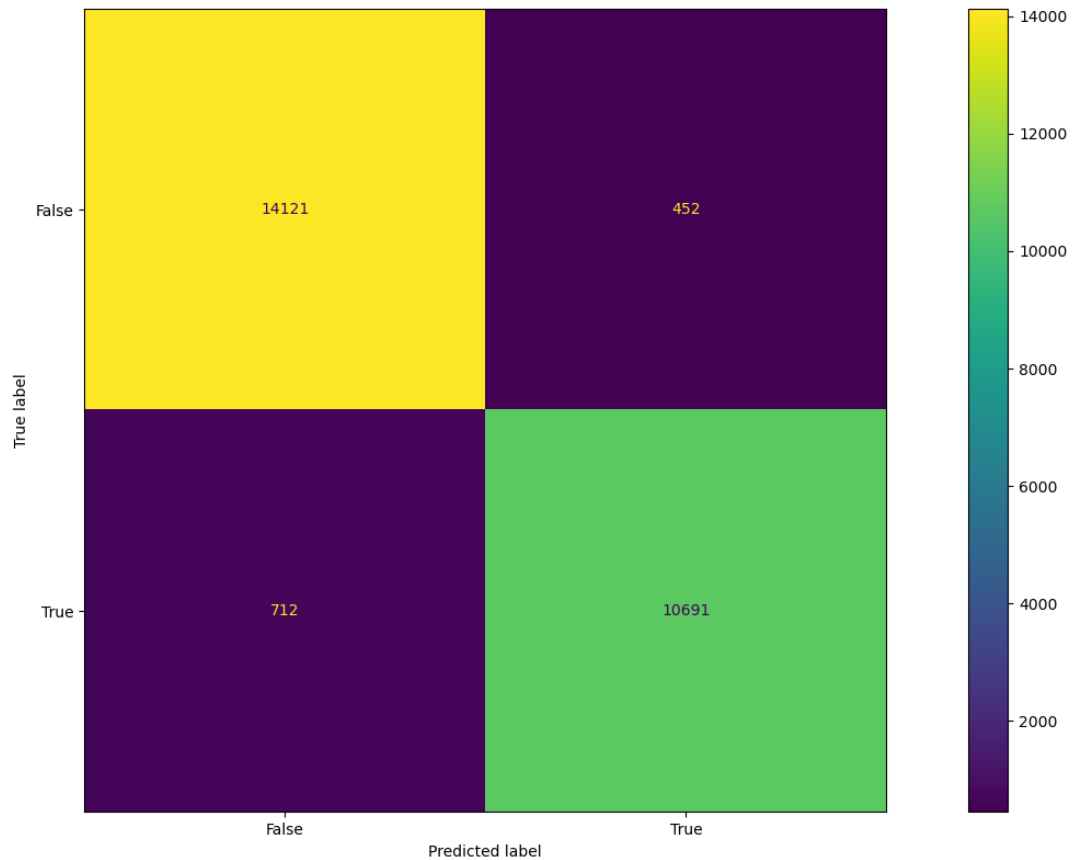


FIGURE 10 – Confusion matrix for SVM

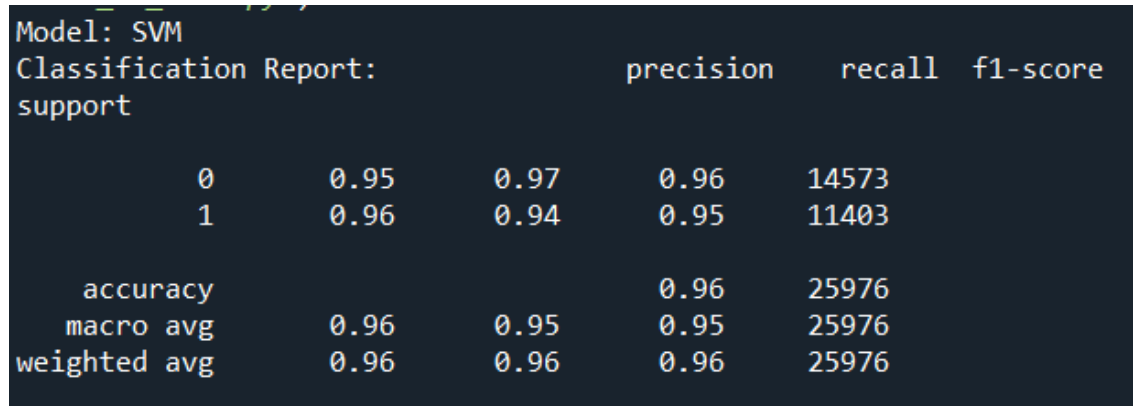
True Negatives (False, False) : 14121 The model correctly predicted 14121 instances as being of class "False".

False Positives (False, True) : 452
The model incorrectly predicted 452 instances as being of class "True" when they were actually "False".

False Negatives (True, False) : 712
The model incorrectly predicted 712 instances as being of class "False" when they were "True".

True Positives : 10691
The model correctly predicted 10691 instances as being of the "True" class.

Now we can see the performance of the model.



```

Model: SVM
Classification Report:
              precision    recall  f1-score   support

      0       0.95      0.97      0.96     14573
      1       0.96      0.94      0.95     11403

   accuracy       0.96
  macro avg       0.96
weighted avg       0.96
  
```

FIGURE 11 – Performance of the SVM model

Here is the table corresponding to the picture :

Classe	Précision	Rappel	F1-score	Support
0	0.95	0.97	0.96	14573
1	0.96	0.94	0.95	11403
Total	0.96	0.96	0.96	25976

We can see that Logistic regression model have 96% precision in total to predict who will be satisfied (1) and neutral or dissatisfied (0). It's better than logistic regression.

3.3 Random Forests

Random Forest is a supervised learning technique used for classification and regression. It works by constructing a multitude of decision trees during training and then outputting the majority class (classification) or the average of the predictions (regression) from these trees. Each tree is constructed from a random sample of the data with replacement (bagging), and each decision within the tree is based on a random subset of the features. This approach reduces overlearning and improves model accuracy.

Here is the confusion matrix for Random forests model :

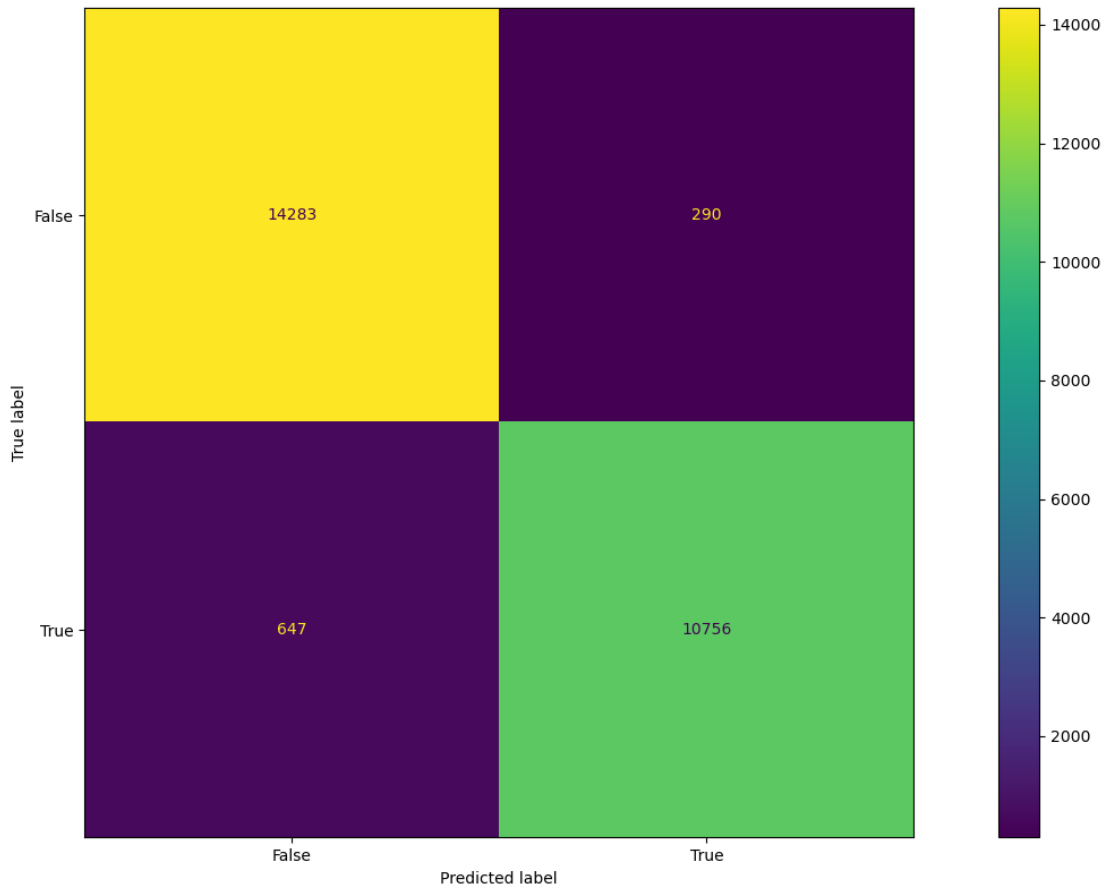


FIGURE 12 – Confusion matrix for Random Forests

True Negatives (False, False) : 14283 The model correctly predicted 14283 instances as being of class "False".

False Positives (False, True) : 290

The model incorrectly predicted 290 instances as being of class "True" when they were actually "False".

False Negatives (True, False) : 647

The model incorrectly predicted 647 instances as being of class "False" when they were "True".

True Positives : 10756

The model correctly predicted 10756 instances as being of the "True" class.

Now we can see the performance of the model.

	precision	recall	f1-score
0	0.96	0.98	0.97
1	0.97	0.94	0.96
accuracy	0.96		25976
macro avg	0.97	0.96	0.96
weighted avg	0.96	0.96	0.96

FIGURE 13 – Performance of the Random Forests model

Here is the table corresponding to the picture :

Classe	Précision	Rappel	F1-score	Support
0	0.96	0.98	0.97	14573
1	0.97	0.94	0.96	11403
Total	0.97	0.96	0.96	25976

We can see that Logistic regression model have 96% precision in total to predict who will be satisfied (1) and neutral or dissatisfied (0). It's better than logistic regression and same as S.V.M.

3.4 K-Neighbor

K-Nearest Neighbors is a supervised learning algorithm used for classification and regression. It classifies a data point according to the majority of categories among its k nearest neighbours in the feature space. For regression, the prediction is the average of the values of the k nearest neighbours. KNN is simple and efficient for small to medium-sized datasets, but can become slow and memory-intensive with large volumes of data.

Here is the confusion matrix for K-Neighbor model :

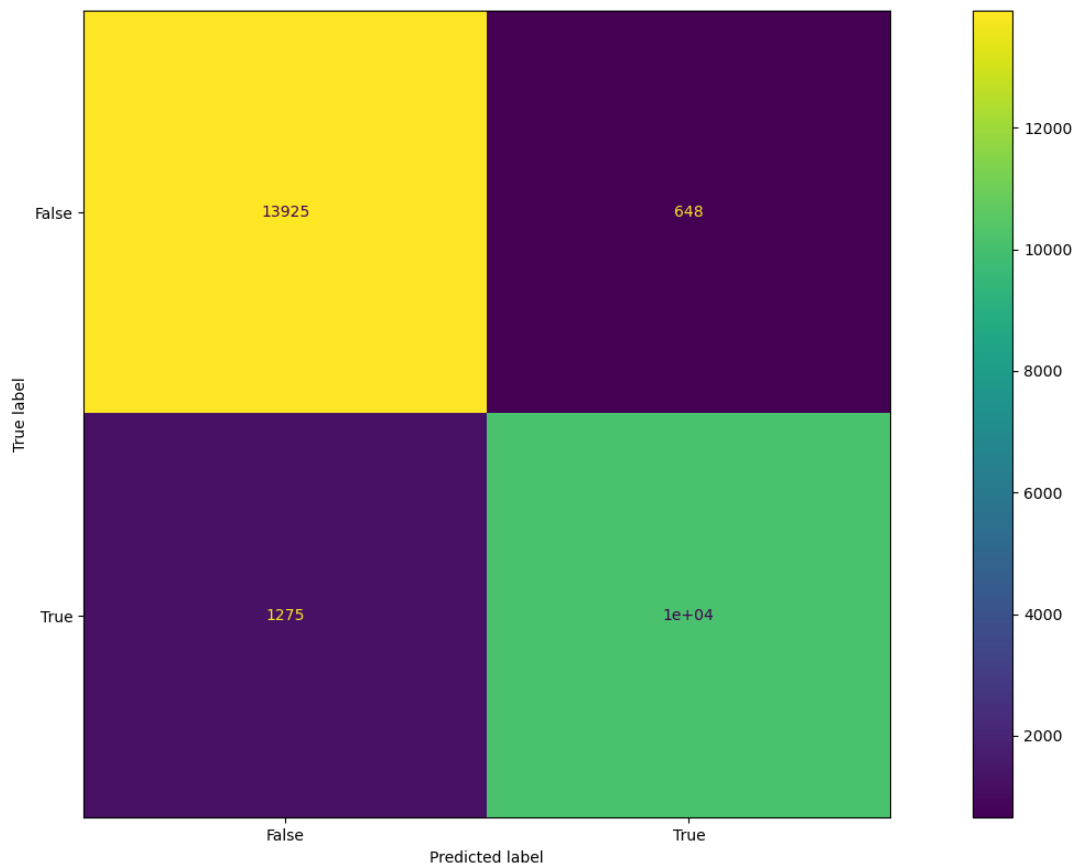


FIGURE 14 – Confusion matrix for K-Neighbor

True Negatives (False, False) : 13925 The model correctly predicted 13925 instances as being of class "False".

False Positives (False, True) : 648

The model incorrectly predicted 648 instances as being of class "True" when they were actually "False".

False Negatives (True, False) : 1275

The model incorrectly predicted 1275 instances as being of class "False" when they were "True".

True Positives : 10000

The model correctly predicted 10000 instances as being of the "True" class.

Now we can see the performance of the model.

```

Model: K_neighbor
Classification Report:
support
    0      0.92      0.96      0.94      14573
    1      0.94      0.89      0.91      11403

accuracy
macro avg      0.93      0.92      0.92      25976
weighted avg    0.93      0.93      0.93      25976

```

FIGURE 15 – Performance of the K-Neighbor model

Here is the table corresponding to the picture :

Classe	Précision	Rappel	F1-score	Support
0	0.92	0.96	0.94	14573
1	0.94	0.89	0.91	11403
Total	0.93	0.93	0.93	25976

We can see that Logistic regression model have 93% precision in total to predict who will be satisfied (1) and neutral or dissatisfied (0). It's still better than Logistic regression but not good as S.V.M and K-Neighbor.

4 Selection of the Best Model

After evaluating multiple training models, we want to select the best model based on performance metrics. The models we considered include Logistic Regression, Support Vector Machine (SVM), Random Forests, and K-Nearest Neighbors (KNN). Each model was evaluated based on its precision, recall, F1-score, and overall accuracy.

4.1 Selection Criteria

The selection of the best model is based on these criteria :

1. **Accuracy** : The proportion of correctly classified instances out of the total instances.
2. **Precision** : The proportion of true positive instances out of the total instances predicted as positive. High precision indicates a low false positive rate.
3. **Recall** : The proportion of true positive instances out of the total actual positive instances. High recall indicates a low false negative rate.
4. **F1-Score** : The harmonic mean of precision and recall.

5. **Confusion Matrix** : Analysis of true positives, true negatives, false positives, and false negatives.

Based on these criteria, we compared the performance of each model :

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.87	0.87	0.87	0.87
Support Vector Machine (SVM)	0.96	0.96	0.96	0.96
Random Forests	0.96	0.97	0.96	0.96
K-Nearest Neighbors (KNN)	0.93	0.93	0.93	0.93

TABLE 1 – Comparison of model performances

The final results on the test set confirmed that the Support Vector Machine (SVM) and Random Forests models performed the best, with both achieving an accuracy of 96%. Given the slight edge in precision for Random Forests and its interpretability, we select the **Random Forests** model as our best performing model.

To finish we can observe the most important components with two graphs.

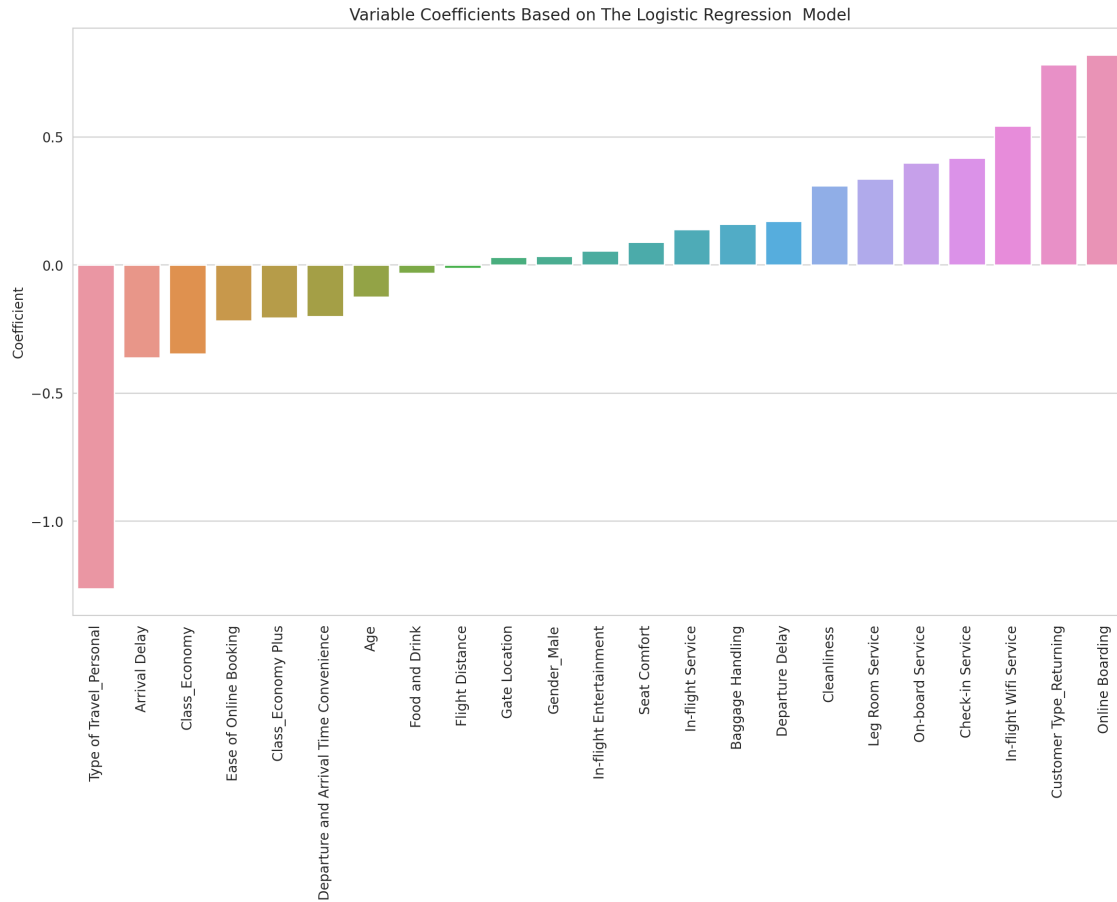


FIGURE 16 – Variable Coefficients Based on The Logistic Regression Model

This graph represent the impact, positive and negative of each components on the satisfaction. We can see that the most important component is the online boarding and the less important is the Type of travel.

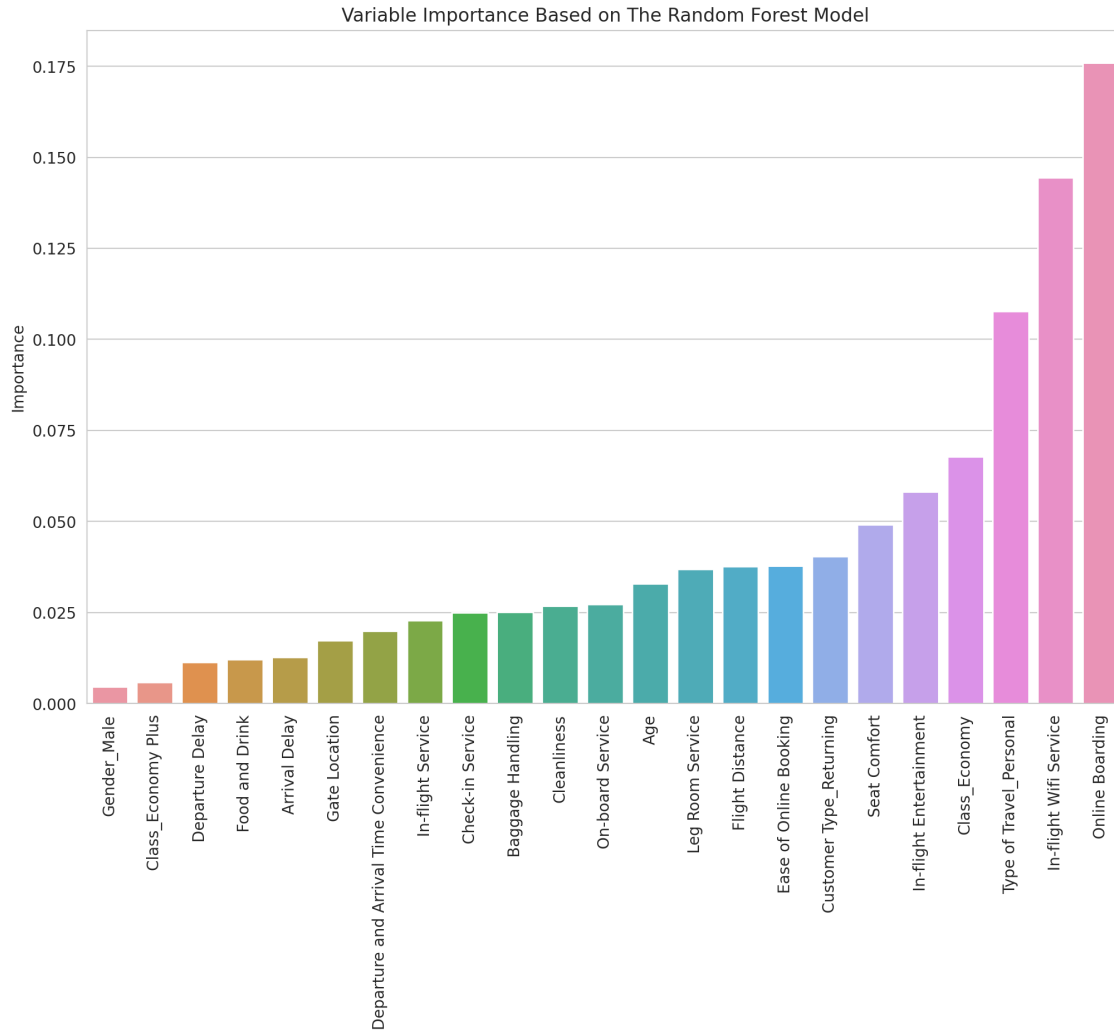


FIGURE 17 – Variable Importance Based on The Random Forest Model

This second graph show us the height of each components, positive or negative combined. That's why online boarding and type of travel are near and they are influent on passenger satisfaction.

5 Conclusion

In this study, we analyzed various factors affecting passenger satisfaction and evaluated multiple machine learning models to predict satisfaction levels. Through rigorous testing and evaluation, the Random Forests model emerged as the best performing model, providing high accuracy and balanced performance across precision, recall, and F1-score. Future work could involve further tuning of hyperparameters, exploring more advanced models, and integrating additional features to improve predictive performance.