Predictive Modeling of Airline Customer Satisfaction Using Machine Learning

Ilaria Gallo, Remo Irtuso

[1] Universidade de Aveiro
[2] Departamento de Eletrónica, Telecomunicações e Informática (DETI)
[3] Aprendizagem Automática, Professor: Petia Georgieva
[4] Work Load: Gallo Ilaria 50%, Irtuso Remo 50%

Abstract :

This study explores predictive modeling of airline customer satisfaction utilizing machine learning techniques. Given the intense competition within the aviation sector, airlines prioritize improving customer satisfaction to ensure loyalty and positive recommendations. The aim of this research is to develop a predictive model that accurately classifies airline customer satisfaction by analyzing various influencing factors. Several machine learning algorithms were employed, including Logistic Regression, Stochastic Gradient Descent (SGD), Decision Trees, Neural Networks, and ensemble methods like Bagging, Random Forest, Voting, and Boosting. The study also highlights the significance of factors such as travel class, departure and arrival delays, and loyalty in predicting customer satisfaction. By leveraging advanced machine learning techniques and thorough data analysis, this research offers actionable insights to enhance airline service quality and passenger experience.

Key Words: Airline Customer Satisfaction, Machine Learning, Classification Models, Data Preprocessing.

1 Introduction

The aviation sector is characterized by increasingly fierce competition between airlines, which strive to offer superior quality services to meet the needs of their customers. Passenger satisfaction has become a key factor in the success of any airline, as satisfied customers tend to become loyal customers and spread their appreciation towards the company. In the context of the growing importance of customer satisfaction, machine learning emerges as a crucial methodology to extract meaningful insights from the data collected by airlines. The main objective of this machine learning project is to develop a predictive model that can classify airline customer satisfaction.

The project involves using machine learning algorithms to train a classification model for airline customer satisfaction. Various variables and characteristics influencing customer satisfaction, as detailed in the dataset description, will be considered. The goal is to identify significant factors impacting airline customer satisfaction through data analysis. These insights can guide informed decision-making and corrective actions to enhance the overall passenger experience and boost loyalty to the airline. By leveraging data analysis and machine learning, we aim to develop an accurate predictive model that helps airlines better understand the factors affecting customer satisfaction and implement effective strategies to improve the travel experience.

2 State of the art

Predictive modeling of airline customer satisfaction has garnered significant attention in recent years, leveraging advancements in machine learning to achieve high levels of accuracy and insight. Numerous studies have explored a variety of methodologies, each focusing on different aspects of the customer experience and utilizing diverse datasets. These efforts aim to identify key factors that influence satisfaction and loyalty, enabling airlines to enhance their services strategically.

For instance, Bacani (2022) applied an Extreme Gradient Boosting (XGBoost) model, highlighting the critical role of model interpretability using SHAP values [**3**]. Shane (2023) employed Principal Component Analysis (PCA) alongside various machine learning models to underscore the importance of seat comfort and inflight entertainment [**2**]. Herawan (2021) achieved a remarkable 99% accuracy with a Random Forest model, emphasizing the significance of onboard wifi service. AlHabbal (2022) compared Decision Trees and Random Forests, advocating for hyperparameter optimization and the integration of deep learning models [**4**]. In another innovative approach, Mirthipati (2024) combined machine learning with causal analysis to enhance digital service improvements, thereby boosting customer satisfaction [**6**]. Ulkhaq and his collegue (2020) found neural networks to be more effective than logistic regression in predicting customer loyalty [**5**].

The techniques reviewed encompass a wide range of machine learning methods. Logistic regression, prized for its simplicity and efficacy, is often enhanced with LASSO and Ridge regularization. Stochastic Gradient Descent (SGD) is noted for its computational efficiency and adaptability. Decision trees are valued for their interpretability and robustness, though they require careful pruning to prevent overfitting. Neural networks, especially those utilizing hyperbolic tangent activation functions, are adept at capturing complex data patterns but demand extensive training and parameter tuning. Ensemble methods, such as Bagging, Random Forest, Voting, and Boosting, are particularly effective in combining multiple models to reduce variance and improve overall performance.

In essence, the application of machine learning techniques like logistic regression, SGD, decision trees, neural networks, and ensemble methods has proven highly effective in the realm of airline customer satisfaction prediction. Ensemble methods, particularly Bagging and Random Forest, are distinguished for their robust performance. Future research should prioritize advanced feature engineering, parameter optimization, and ongoing model evaluation to further enhance predictive accuracy and deliver actionable insights.

3 Dataset

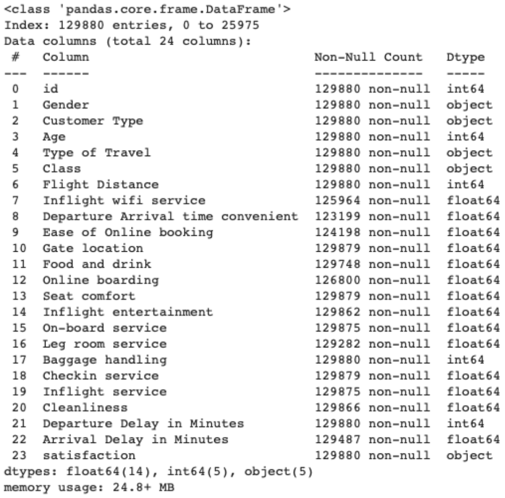
3.1 Dataset Description

The data used for this project was sourced from a dataset titled "Airline Passenger Satisfaction" [**1**]. This dataset is derived from surveys conducted with customers of an airline.

The dataset consists of 129,880 records and 25 attributes. The attributes are:

- Gender
- Customer Type: Indicates if the passenger is a loyal customer or a one-time passenger
- Age of the passenger
- Type of Travel: Personal Travel or Business Travel
- Class: The class of travel booked by the passenger
- Flight Distance
- Inflight Wifi Service: Indicates customer satisfaction regarding the inflight wifi service
- Departure/Arrival Time Convenience: Indicates customer satisfaction regarding the convenience of departure and arrival times
- Ease of Online Booking: Indicates customer satisfaction regarding the ease of online flight booking
- Gate Location: Indicates customer satisfaction regarding the location of the gate
- Food and Drink: Indicates customer satisfaction regarding the food and beverages offered on the plane
- Online Boarding: Indicates customer satisfaction regarding online boarding
- Seat Comfort: Indicates customer satisfaction regarding seat comfort
- Inflight Entertainment: Indicates customer satisfaction regarding inflight entertainment
- On-board Service: Indicates customer satisfaction regarding the services offered on board
- Leg Room Service: Indicates customer satisfaction regarding legroom between seats
- Baggage Handling: Indicates customer satisfaction regarding baggage handling
- Check-in Service: Indicates customer satisfaction regarding the check-in service (0: missing data, 1-5 measures the level of satisfaction).
- Inflight Service: Indicates customer satisfaction regarding the services offered during the flight
- Cleanliness: Indicates customer satisfaction regarding the cleanliness of the aircraft
- Departure Delay in Minutes
- Arrival Delay in Minutes
- Satisfaction: Indicates customer satisfaction (satisfied, neutral, or dissatisfied)

The dataset was then examined, revealing that only the attribute "Arrival Delay in Minutes" contains missing data, with 129,487 entries instead of 129,880, as shown in Figure 1. It is important to note that for attributes indicating customer satisfaction levels for various service categories, a value of 0 denotes missing data. To obtain more accurate information from the dataset, the values of 0 should be converted to NaN for all the relevant satisfaction attributes. After making this adjustment, it becomes evident that multiple columns now contain missing data. To gain a clearer understanding, the histogram 2 was created to indicate the fraction of missing data for each attribute. The missing data constitutes a very small fraction for each attribute. The attribute with the highest percentage of missing data is "Departure/Arrival Time Convenience" suggesting that fewer people provided a rating between 1 and 5 for this service category in the surveys.



```
<class 'pandas.core.frame.DataFrame'>
Index: 129880 entries, 0 to 25975
Data columns (total 24 columns):
 #   Column                             Non-Null Count   Dtype
---  ------                             --------------   -----
 0   id                                 129880 non-null  int64
 1   Gender                             129880 non-null  object
 2   Customer Type                      129880 non-null  object
 3   Age                                129880 non-null  int64
 4   Type of Travel                     129880 non-null  object
 5   Class                              129880 non-null  object
 6   Flight Distance                    129880 non-null  int64
 7   Inflight wifi service              125964 non-null  float64
 8   Departure Arrival time convenient  123199 non-null  float64
 9   Ease of Online booking             124198 non-null  float64
 10  Gate location                      129879 non-null  float64
 11  Food and drink                     129748 non-null  float64
 12  Online boarding                    126800 non-null  float64
 13  Seat comfort                       129879 non-null  float64
 14  Inflight entertainment             129862 non-null  float64
 15  On-board service                   129875 non-null  float64
 16  Leg room service                   129282 non-null  float64
 17  Baggage handling                   129880 non-null  int64
 18  Checkin service                    129879 non-null  float64
 19  Inflight service                   129875 non-null  float64
 20  Cleanliness                        129866 non-null  float64
 21  Departure Delay in Minutes         129880 non-null  int64
 22  Arrival Delay in Minutes           129487 non-null  float64
 23  satisfaction                       129880 non-null  object
dtypes: float64(14), int64(5), object(5)
memory usage: 24.8+ MB
```

Figure 1. Information about data type and non-null values.

Another analysis involves examining the characteristics of the numerical data. From this analysis, the following observations can be made:

- Passengers have an average age of 39 years, with a minimum age of 7 and a maximum age of 85.
- The average flight distance is nearly 1200 km, with distances ranging from 31 km to almost 5000 km.
- Both arrivals and departures have an average delay of about 15 minutes, with a maximum delay of 26.5 hours.
- For all other attributes indicating onboard services, the average satisfaction rating hovers around 3, with variations depending on the specific service.

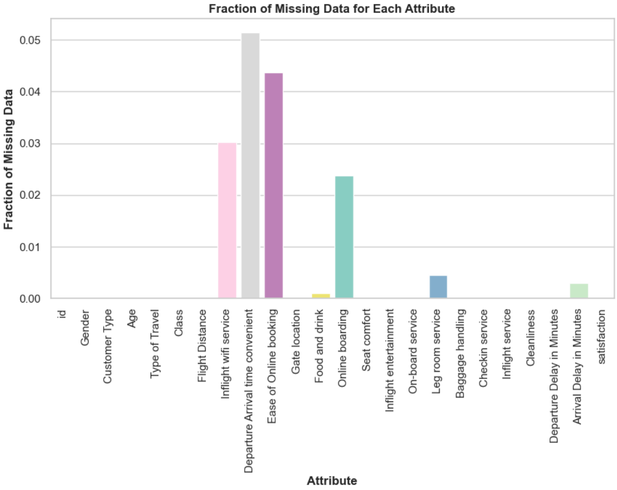


Figure 2. Fraction of missing data

.

3.2 Preliminary Analysis and Assumptions

Preliminary analysis of the dataset is crucial to understand its structure, identify potential issues, and formulate hypotheses that will guide subsequent modeling phases. Below are the main assumptions and analyses conducted.

3.2.1 Assumption 1: Customer Loyalty as an Indicator of Satisfaction It is assumed that "Loyal Customer" tend to be more satisfied compared to occasional customers, defined as "Disloyal Customer". This assumption is based on the premise that customer loyalty is a direct consequence of repeated positive experiences.

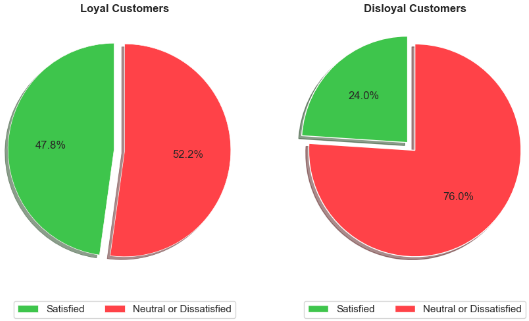


Figure 3. Customer loyalty as indicator of satisfaction.

.

Data analysis revealed that 52.2% of loyal customers are dissatisfied or neutral about the service received, while only 47.8% are satisfied. For occasional customers, 76% are dissatisfied or neutral, and only 24% are satisfied. These results, shown in Figure 3, suggest that although loyal customers are generally more satisfied compared to occasional customers, there is still a significant portion of loyal customers who are not completely satisfied.

3.2.2 Assumption 2: Travel Class Influences Satisfaction It is hypothesized that passengers traveling in Business class are more satisfied compared to those traveling in Economy or Economy Plus. This assumption is based on the premise that Business class offers higher quality services, contributing to greater customer satisfaction. The analysis confirms that 69.4% of Business class passengers are satisfied, compared to 18.8% and 24.6% for Economy and Economy Plus classes, respectively, as shown in Figure 4. This supports the assumption that the superior quality of services offered in Business class contributes to greater satisfaction.

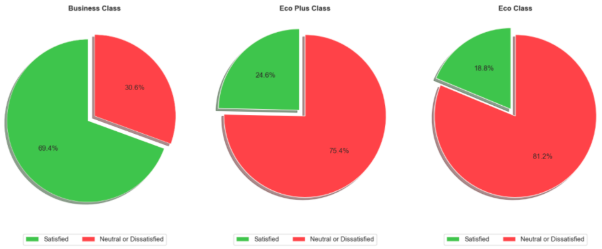


Figure 4. Satisfaction with respect to travel classes.

.

3.2.3 Assumption 3: Departure Delays Reduce Satisfaction It is assumed that flights departing with a delay greater than the average (15 minutes) reduce passenger satisfaction. The data shows that 62.9% of passengers on delayed flights are dissatisfied or neutral, while only 37.1% are satisfied. This result, shown in Figure 5, confirms that departure delays have a negative impact on customer satisfaction.

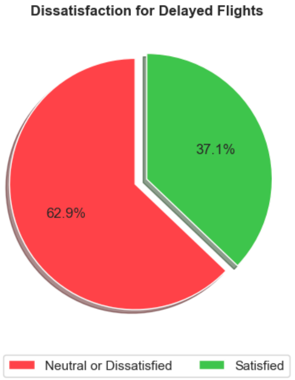


Figure 5. Dissatisfaction for delayed flights

.

3.2.4 Assumption 4: Arrival Delays Affect Satisfaction It is assumed that flights arriving with a delay greater than the average (15 minutes) negatively impact passenger satisfaction. Similar to departure delays, 64.2% of passengers on flights with arrival delays are dissatisfied or neutral, while 35.8% are satisfied. This confirms that arrival delays also negatively affect customer satisfaction. The chart which support our assumption is shown in Figure 6.

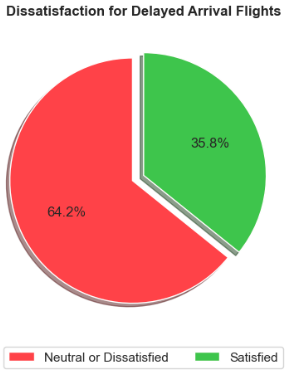


Figure 6. Dissatisfaction for delayed arrival flights

.

3.2.5 Recommendations for Enhancing Customer Satisfaction The preliminary analyses underscore travel class and delays as key factors influencing customer satisfaction, forming a solid foundation for developing predictive models and optimizing classification models. The findings suggest that enhancing loyalty programs, upgrading Economy class services, and minimizing delays are crucial steps for improving customer satisfaction. Personalized services and incentives tailored to loyal customers' needs can address their specific dissatisfaction. Improving seating comfort, in-flight entertainment, and food and beverage options in Economy and Economy Plus classes can elevate customer experience. Implementing efficient scheduling and operational strategies, alongside robust customer support during delays, such as timely updates and compensations, can mitigate dissatisfaction. By adopting these measures, airlines can significantly boost overall customer satisfaction, leading to increased loyalty and positive word-of-mouth promotion.

3.3 Correlation

After the preliminary assumptions, was conducted a study about the correlation between other attributes and the satisfaction class. In particular, the following attributes will be analyzed: id, Gender, Age, Type of Travel and Flight Distance.

To do this, the correlation matrix, represented in Figure 7, was visualized. This is useful for determining how two variables influence each other. It is observed that the correlation between satisfaction, Age, id, and Gender is very low, indicating that the target attribute is not influenced by the other three. However, there is a moderate correlation between the target variable and the attributes Type of Travel and Flight Distance.

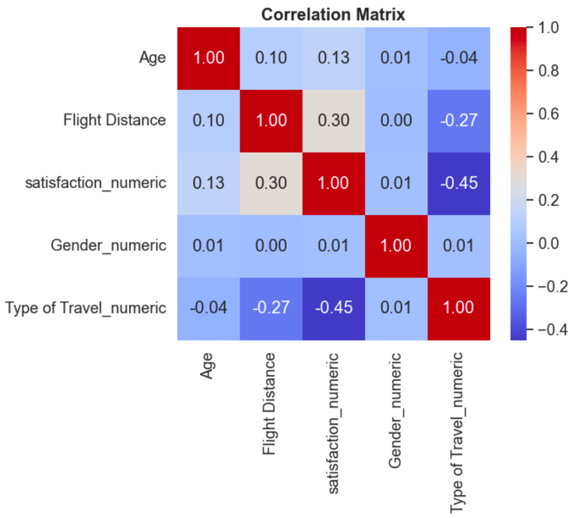


Figure 7. Correlation Matrix.

4 Preprocessing

4.1 Missing data management

It has been observed that some columns contain missing data. Concerning columns 7 to 20, the missing data is replaced with the median. These columns encompass a range of values from 1 to 5, indicating customer satisfaction in specific areas. Substituting missing data with the median is deemed the most impartial alternative. In the case of the 'Arrival Delay in Minutes' column, missing values are replaced with the mean value. This approach aims to mitigate the impact of missing data on classification algorithms and prevent the loss of information resulting from the deletion of rows with missing values.

4.2 Elimination of attributes not related to the target variable

In the previous section, it was deduced that the variables Age, id, and Gender are uncorrelated with respect to the target variable. Consequently, these columns can be removed as they will not significantly affect the analysis. The shape after removing uncorrelated attributes is (129880, 21).

4.3 Aggregation of the column relating to delays

Subsequently, the attributes related to delays, namely 'Arrival Delay in Minutes' and 'Departure Delay in Minutes', were consolidated into a single attribute. This consolidation was deemed advantageous for several reasons.

1. Reduction of Dataset Complexity and Dimensionality: Having a singular attribute to represent delays simplifies the dataset, thus reducing the complexity of analysis. Consolidating values into one attribute also diminishes the dataset's dimensionality, thereby facilitating the training of classification models.
2. Information Synthesis: The creation of the attribute 'Overall Delay' would provide a comprehensive overview of the experienced delay level, which could be relevant for evaluating the overall impact on customer satisfaction.

This consolidation serves to streamline the dataset while preserving crucial delay-related information, enhancing its usability for subsequent analysis and modeling processes.

4.4 Outliers

In the analysis of the flight delay dataset, it was observed that the maximum delay values are particularly high, indicating the presence of outliers. To identify such anomalous values, an upper limit was established based on an empirical threshold of 2 standard deviations beyond the mean of the overall delay. The calculated upper limit is 181.24 minutes. Records with an overall delay exceeding this value were classified as outliers, as shown in the Figure 8 .

The number of identified outliers was 5024, representing approximately 4% of the total records. The analysis revealed that about one-third of the customers whose flights experienced delays exceeding 3 hours still reported satisfaction, a situation that can be considered an anomaly.

When dealing with outliers, two options arise: removing the outliers to mitigate their impact on the analysis, albeit at the risk of losing relevant information, or retaining the outliers in the dataset to preserve all information, albeit at the cost of potential distortions in the analyses. However, outliers represent real-life situations that can occasionally occur. Therefore, retaining these data in the dataset could be useful to accurately represent the information contained therein. Furthermore, the presence of a significant percentage of satisfied customers despite the delays suggests that the analysis and modeling should account for these exceptions to improve the quality of service and predictions.

4.5 Data transformation

To ensure attribute balance and enhance the accuracy of predictive models, various data transformations have been applied.

Firstly, data normalization was performed using the mean normalization formula, which standardized the attribute value scales.

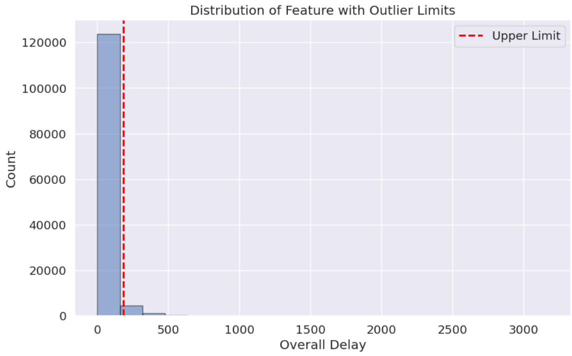


Figure 8. Distribution of attributes considering the outlier limit

.

The normalized attributes include Flight Distance and Overall Delay. Additionally, categorical attributes were converted into binary numeric vectors through One-Hot Encoding. This process involved attributes such as Customer Type, Type of Travel, and Class.

Finally, class balancing was checked, revealing a slight imbalance with a prevalence of dissatisfied customers, as shown in Figure 9. Although intervention was not necessary, in other contexts, oversampling or undersampling techniques could be applied to balance the classes. These transformations ensure that the data is adequately prepared for predictive analysis, reducing bias and improving the performance of machine learning models.

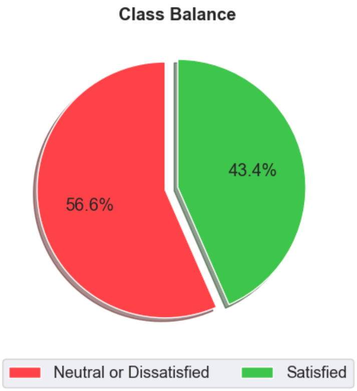


Figure 9. Class Imbalance graph.

5 Classification

After a preliminary phase in which the data was analyzed and transformed to meet the project requirements, the construction of classification models was initiated. To do this the dataset was divided into three distinct subsets: training, validation, and test sets. This division is crucial for developing a robust model and ensuring its generalizability to new, unseen data.

- **Training Set**: This subset, typically comprising 60% of the dataset, is used to train the machine learning models. The models learn from this data by identifying patterns and relationships between the input features and the target variable.
- **Validation Set**: Usually, 20% of the dataset is allocated to the validation set. This subset is used during the model development phase to tune hyperparameters and make decisions about model architecture. It helps in preventing overfitting by ensuring that the model's performance is optimized not only on the training data but also on a separate, unseen validation set.
- **Test Set**: The remaining 20% of the dataset is used as the test set. This subset is not used during the model training or validation phases. Instead, it is used to evaluate the final model's performance, providing an unbiased estimate of its accuracy, precision, recall, and other relevant metrics on completely new data.

The main objective of this phase was to train different models, compare their performances, and select the most suitable one for data classification. The models considered are:

- Logistic Regression
- Stochastic Gradient Descent (SGD)
- Decision Trees
- Multilayer Neural Networks

Ensemble classifiers were also considered in this project, specifically:

- Bagging
- Random Forest
- Voting
- Boosting (AdaBoost)

The process for each model in this project involves three main steps.

The first step is model training, where the chosen model is trained on the provided dataset to learn the underlying patterns and relationships.

Following training, the model's performance is evaluated using various metrics, including the Confusion Matrix, Accuracy, Precision, Recall, Specificity, F-measure, and ROC (Receiver Operating Characteristic). The Confusion Matrix provides a detailed breakdown of the model's predictions, helping to understand how well the model distinguishes between different classes. Accuracy measures the overall correctness of the model, while Precision and Recall provide insights into the model's performance with respect to false positives and false negatives, respectively. Specificity indicates how well the model identifies true negatives, and the F-measure offers a balance between Precision and Recall. The ROC curve visually represents the trade-off between sensitivity and specificity at different threshold levels.

The final step involves model optimization, where key parameters are adjusted to improve performance. By iterating through these steps, the models are fine-tuned to achieve optimal performance.

5.1 Logistic Regression

Logistic regression is a widely used machine learning algorithm for binary classification, chosen for its simplicity and effectiveness. It models the probability that a given input belongs to a particular class, using a logistic function to transform linear combinations of input features into a value between 0 and 1. This probability can then be used to classify the input into one of two categories. In this project, different logistic regression models were developed and compared, each with a different type of regularization:

- LASSO
- Ridge

The different parameter used for the logistic regression are shown in Table 1 and Table 2.

Parameter	Value
solver	liblinear
class weight	balanced
penalty	l1

Table 1
Parameters for Logistic Regression with LASSO regularization

Parameter	Value
solver	liblinear
class weight	balanced
penalty	l2

Table 2
Parameters for Logistic Regression with Ridge regularization

The results in Figure 16 in the appendix shows that both the logistic regression models with LASSO and Ridge regularization had very similar performance, with slight variations in evaluation metrics on both the training and test data. In Table 3 are shown the numeric result of the logistic regression with Ridge regularization, while in Table 4 are shown the results of the logistic regression with LASSO regularization.

Accuracy	Precision	Recall	Specificity	F-measure
0.88	0.85	0.88	0.88	0.86

Table 3
Logistic Regression (Ridge)

For Logistic Regression with LASSO, the training set accuracy was 0.8847, with a Precision of 0.8523, Recall of 0.8881, Specificity of 0.8821, and F-measure of 0.8698. The test set performance was similarly high, indicating a good ability of the model to generalize results to new data. The ROC curve showed an area under the curve (AUC) of 95%, highlighting the model's effectiveness in distinguishing between classes. All of this is shown in Figure 17 in the appendix. The Logistic Regression model with Ridge produced nearly identical results, as shown in Figure 18 in the appendix.

The analysis of regression coefficients, shown in figure 10 revealed that models with LASSO regularization tend to have more pronounced negative coefficients, indicating a greater influence of negative attributes on the target variable. This can be useful for deeper interpretations of the factors affecting customer satisfaction.

In conclusion, logistic regression with LASSO regularization was selected as the final model for its ability to maintain a good balance between accuracy and interpretability, proving effective in classifying records of satisfied and dissatisfied customers.

Accuracy	Precision	Recall	Specificity	F-measure
0.88	0.85	0.88	0.88	0.86

Table 4
Logistic Regression (Lasso)

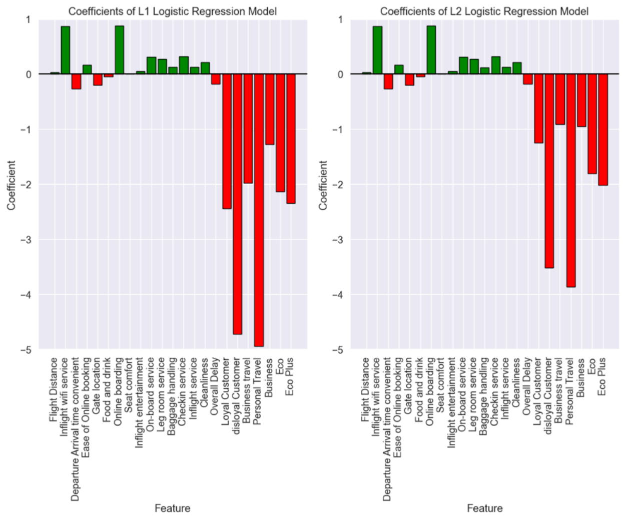


Figure 10. Logistic regression's coefficients.

5.2 SGD Classification

The Stochastic Gradient Descent (SGD) classification algorithm is a widely used optimization method for training classification models, renowned for its computational efficiency and adaptability. It works by iteratively adjusting model weights to minimize a specified loss function, updating the weights based on the gradient of the loss with respect to the model parameters for each training example. This makes SGD particularly effective for handling large datasets and noisy or nonlinear data, as it often converges faster and requires less memory compared to traditional gradient descent methods. The parameter found by the grid parameter used are shown in Table 5 and 6.

Parameter	Value
learning rate	constant
eta	0.01
loss	log_loss
penalty	l1

Table 5
Parameters for SGD with constant learning rate

In this project, a grid search was used to tune the SGD classifier, considering two scenarios: one with a *constant learning rate* and another with an *optimal learning rate* that adapts heuristically. The constant learning rate scenario used `log_loss` as the loss function and L1 (LASSO) regularization, while the optimal learning rate scenario used `log_loss` and L2 (Ridge) regularization.

Performance metrics for the first model are shown in Figure 19 and for the second model in Figure 20 in the appendix. The comparison between this models, shown in Figures 11, demonstrating that the constant learning rate model is faster to train and thus more suitable for the analysis. The numeric results of the two models are shown in Table 7 and Table 8.

5.3 Decision trees

Decision trees are a popular classification method that recursively splits the dataset into increasingly homogeneous subsets until a final classification is achieved. Decision trees offer significant advantages in binary classification, including high interpretability,

Parameter	Value
learning rate	optimal
eta	0.01
loss	log_loss
penalty	l2

Table 6
Parameters for SGD with optimal learning rate

Accuracy	Precision	Recall	Specificity	F-measure
0.87	0.92	0.77	0.95	0.84

Table 7
SGD costant learning

automatic handling of variables without complex data preparation, and robustness against highly correlated attributes.

Initially, a decision tree classifier was created without setting any attribute limits or depth restrictions. This model showed a tendency towards *overfitting*, with perfect training data fit as shown in Figure 21 in the appendix. To address this, a depth limit of 10 was imposed, reducing the model's complexity and overfitting tendency.

The performance metrics for the depth-limited decision tree model are shown in Figure 22 in the appendix. Despite the limitation, the metrics remained high, with an accuracy of 94.13% on the test set, precision of 95.42%, recall of 90.95%, specificity of 96.61%, and F-measure of 93.13%. The ROC curve area for this model was 98%, compared to 94% for the unrestricted model. The confusion matrix analysis revealed minimal misclassification of negative records (496) and relatively few misclassified positive records (1028). All the results are shown in Table 9 and 10.

A grid search identified the best parameters for the decision tree which are summarized in Table 11.

This optimized model showed improved classification of positive records, reducing false negatives by about 200. The performance metrics for the optimized model are also shown in figure 23 in the appendix, with an accuracy of 94.58%, precision of 95.03%, recall of 92.44%, specificity of 96.25%, and F-measure of 93.72% on the test set.

Comparing the depth-limited tree and the grid-searched tree, the depth-limited tree exhibited higher accuracy, recall, and specificity on the test data as shown in Figure 12. Its simplicity and interpretability make it a more suitable choice for the final model comparison. The depth-limited tree, with a maximum depth of 10, was ultimately chosen for its balance of performance and interpretability.

5.4 Multi-level Neaural Networks

In this project, different configurations of neural networks were explored for classification tasks. Neural networks, inspired by the structure and function of the human brain, consist of interconnected layers of nodes (neurons) that process input data. By adjusting the connections (weights) between these neurons through training, neural networks can learn to identify complex patterns and relationships within the data. Different network architectures, such as the number of layers and neurons per layer, activation functions, and learning rates, were experimented with to optimize performance. This approach is powerful for classification tasks, especially when dealing with large and intricate datasets, due to its ability to model nonlinear relationships and capture intricate patterns.

Accuracy	Precision	Recall	Specificity	F-measure
0.88	0.87	0.86	0.90	0.86

Table 8
SGD optimal learning rate

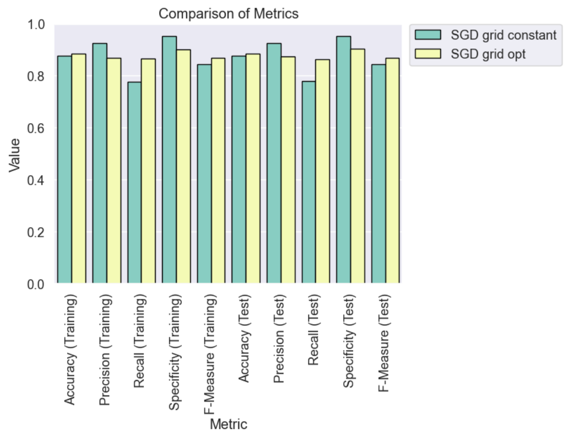


Figure 11. Comparison between SGD grid constant and SGD grid opt.

Two neural network models with two hidden layers, each consisting of 20 nodes, were compared. These models differed in their activation functions: the first used the *logistic sigmoid function*, while the second used the *hyperbolic tangent (tanh) function.* All the parameters are summarized in Table 12 adn Table 13.

Parameter	Value
learning_rate	adaptive
hidden_layer_sizes	(20,20)
activation	tanh
solver	sgd

Table 13
Parameters for NN with hyperbolic tangent activation function.

The results of the neural network with the logistic sigmoid activation function are shown in Figure 24 in the appendix. This model achieved a training accuracy of 92.43%, precision of 93.56%, recall of 88.67%, specificity of 95.33%, and F-measure of 91.05%. The test accuracy was 92.35%, precision 93.67%, recall 88.49%, specificity 95.36%, and F-measure 91.01%. The ROC curve for this model showed an area under the curve (AUC) of 0.98. The lowest metric was recall (88%), due to errors in classifying positive records (1307), while negative records were misclassified in only 679 instances. The second neural network model, using the hyperbolic tangent activation function, demonstrated superior performance, as shown in Figure 25 in the appendix. The training accuracy was 94.30%, precision 95.68%, recall 90.96%, specificity 96.85%, and F-measure 93.26%. The test accuracy was 94.12%, precision 95.75%, recall 90.58%, specificity 96.87%, and F-measure 93.09%. The ROC curve for this model showed an AUC of 0.99. There was also an improvement in classifying both positive and negative records. As showed before for the other classification models, the results are in Table 14 and Table 15.

The comparative analysis indicated that the neural network

Accuracy	Precision	Recall	Specificity	F-measure
0.94	0.95	0.90	0.96	0.93

Table 9
Decision Tree 10 deep

Accuracy	Precision	Recall	Specificity	F-measure
0.94	0.95	0.92	0.96	0.93

Table 10
Decision Tree Grid Search

with the hyperbolic tangent activation function had higher metrics across all categories compared to the network using the logistic sigmoid function, as shown in Figure 13. For the final model comparison, the neural network with the hyperbolic tangent activation function was selected due to its overall superior performance.

6 Ensemble Classifier

Ensemble classification techniques combine the results of multiple models to improve overall performance. These methods leverage the strengths of individual models to enhance accuracy and robustness. The project evaluated four ensemble classifiers: Bagging, Random Forest, Voting, and Boosting (specifically AdaBoost).

6.1 Bagging

Bagging involves training multiple independent models on randomly selected samples from the training dataset with replacement, creating diverse subsets known as bootstrap samples. Each model is trained separately on these samples, and their predictions are combined, typically by averaging for regression tasks or majority voting for classification tasks. This approach reduces variance and helps prevent overfitting, leading to improved overall performance and robustness of the model.

The Bagging parameters are shown in Table 16.

Parameter	Value
basic classifier	Decison Tree
max_dept	10
num_estimator	30

Table 16
Parameters for Bagging.

The Bagging model showed excellent performance, achieving a training accuracy of 94.94%, precision of 96.05%, recall of 92.12%, specificity of 97.10%, and F-measure of 94.05%, as shown in Figure 26 in the appendix. Test metrics mirrored these high values, demonstrating the model's precision and reliability. All the numeric results of the test performance are shown in Table 17.

6.2 Random Forest

Random Forest extends the Bagging approach by training each tree on a random subset of features, further reducing overfitting. This method decorrelates trees and improves generalization. The optimal number of features (max_features) was set to the square root of the total number of attributes. The parameter are show in Table 18.

Parameter	Value
max_depth	15
min_samples_leaf	7
min_samples_split	4

Table 11
Parameters for Decision Tree.

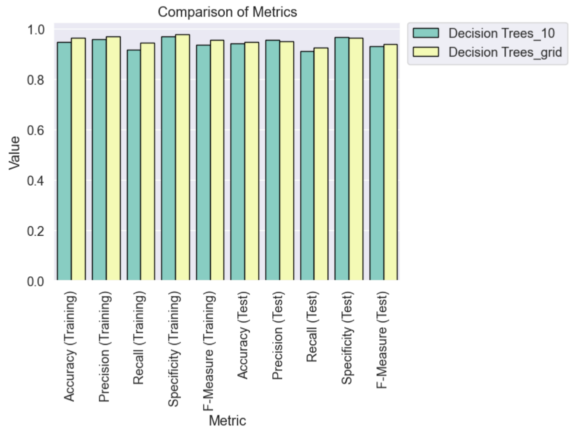


Figure 12. Comparison between decion trees.

Parameter	Value
n_estimators	30
max_depth	10
max_features	sqrt

Table 18
Parameter for Random Forest.

The Random Forest model achieved high performance metrics with a training accuracy of 94.22%, precision of 94.57%, recall of 91.95%, specificity of 95.95%, and F-measure of 93.24%, as shown in Figure 28 in the appendix. Test results also indicated strong performance, confirming its effectiveness in classifying both positive and negative records. As for the previous classifiers, the results are shown in Table 19.

6.3 Voting

The Voting ensemble combines the predictions of several independent classifiers trained on the same dataset. Final predictions are made either by majority vote or by averaging predicted probabilities. The classifiers used were Logistic Regression with LASSO regularization, SGD with a constant learning rate, a Decision Tree with a maximum depth of 10, and a Neural Network with a hyperbolic tangent activation function which are the best classifiers obtained in this study.

The Voting classifier showed robust performance, with a training accuracy of 93.77%, precision of 94.89%, recall of 90.51%, specificity of 96.27%, and F-measure of 92.65%, as shown in Figure 29 in the appendix. These metrics remained consistent in the test results, indicating reliability and robustness. In Table 20 is possible to see the results.

6.4 Boosting (AdaBoost)

Boosting assigns weights to records, initially equal, which are adjusted after each classification. Correctly classified records have

Parameter	Value
learning_rate	adaptive
hidden_layer_sizes	(20,20)
activation	logistic
solver	sgd

Table 12
Parameters for NN with Logistic Regression activation function.

Accuracy	Precision	Recall	Specificity	F-measure
0.92	0.93	0.88	0.95	0.91

Table 14
NN with Logistic activation function

their weights decreased, while misclassified records have their weights increased, focusing subsequent models on difficult examples. AdaBoost combines multiple weak classifiers to create a strong classifier. In this case 30 estimators were used.

The model showed good performance with a training accuracy of 90.82%, precision of 90.54%, recall of 88.03%, specificity of 92.95%, and F-measure of 89.27%. Test results were similar, though AdaBoost had the lowest area under the ROC curve (0.97) among the ensemble classifiers, with higher classification errors for both positive and negative records, as shown in Figure 27.

6.5 *Comparison of ensemble classifiers*

Overall, ensemble classifiers generally outperform single classifiers by reducing variance and bias. Bagging and Random Forest models showed particularly high performance, with all metrics exceeding 90%, as shown in Figure 14. The final model selection depended on factors such as training speed, error minimization, and model complexity.

7 Results

All techniques examined in the previous phases yielded good results. In this phase, a comprehensive comparison of the obtained models was conducted.

Initially, a comparison of "single" models was performed and is shown in Figure 30 in the appendix. The metrics for Logistic Regression with LASSO, SGD with a constant learning rate, Decision Trees with a maximum depth of 10, and Neural Networks using the hyperbolic tangent activation function were evaluated. The analysis revealed that the best-performing models were the Decision Tree with a maximum height of 10 and the Neural Network with the hyperbolic tangent activation function. These models showed high values across accuracy, precision, recall, specificity, and F-measure.

Subsequently, ensemble classification models were compared, as shown in Figure 14. Bagging, Random Forest and Voting exhibited the highest performance metrics.

By comparing the best selected results, the graph shown in Figure 15 is obtained.

The models in the previous graph all exhibit excellent performance, with the best-performing ones being Bagging and the neural network. It's challenging to definitively determine which of these models is most suitable for the problem. The choice of model depends on several factors:

Accuracy	Precision	Recall	Specificity	F-measure
0.94	0.94	0.90	0.96	0.93

Table 15
NN with Hyperbolic Tangent activation function

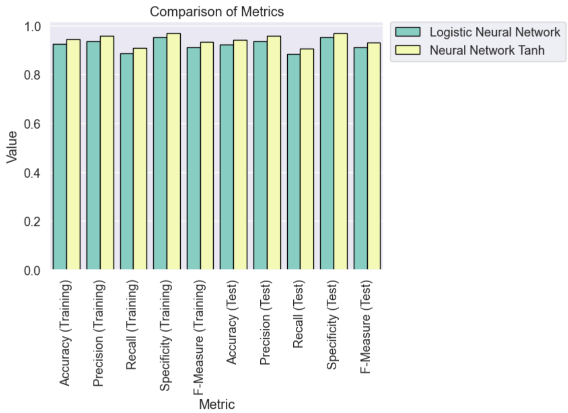


Figure 13. Comparison between the neural networks.

- Model training speed: If opting for models with reasonably fast training, the best choice is the selected tree model. Conversely, if training speed is not a significant concern, neural networks can be chosen.
- Error rate reduction: If the goal is to minimize classification error, the preference lies with ensemble algorithms as they combine results from multiple classification models, reducing the likelihood of erroneous classifications.
- Model complexity: If preferring lower complexity, simpler models like the decision tree can be chosen. This classifier is relatively easy to interpret and train, offering good performance across many scenarios.

In this specific case, the preference leans towards Bagging, as the objective is to reduce classification error for records pertaining to the 'neutral or dissatisfied' class.

8 Novelty and contributions

The exploration of airline passenger satisfaction is a crucial area of study in the highly competitive aviation industry. This research presents several novel contributions to the understanding and enhancement of passenger satisfaction through the application of advanced machine learning techniques. Furthermore, it contributes to the existing body of literature by comparing its findings with other works that have explored similar themes using different analytical approaches. This comparison not only validates the robustness of the findings but also highlights the unique insights derived from specific methodologies.

A comparative analysis between the results obtained in this study and those presented by Shane (2023) reveals numerous similarities and some significant differences. Both studies emphasize the importance of passenger satisfaction in the aviation sector, recognizing that various factors contribute to the overall flight experience. In this study, attributes such as seat comfort, inflight entertainment, and cleanliness were identified as critical determinants of passenger satisfaction. These findings align with Shane (2023), who highlighted similar variables through Principal Component Analysis (PCA). A significant point of convergence is the importance of customer loyalty, showing that loyal customers tend

Accuracy	Precision	Recall	Specificity	F-measure
0.94	0.95	0.91	0.96	0.93

Table 17
Bagging

Accuracy	Precision	Recall	Specificity	F-measure
0.93	0.94	0.91	0.95	0.92

Table 19
Random Forest

to report higher levels of satisfaction compared to non-loyal customers. This indicates that airlines should focus not only on improving services but also on loyalty strategies to maintain high levels of satisfaction among their regular customers [**2**].

In analyzing airline passenger satisfaction, this study compared its results with those of Chris Bacani's work "*Explaining airline passenger satisfaction using interpretable machine learning.*" Bacani utilized an Extreme Gradient Boosting (XGBoost) classification model, achieving a test accuracy of 95%, and emphasized model interpretability through SHAP. Various machine learning techniques were explored, including logistic regression, SGD, decision trees, multilayer neural networks, and ensemble methods such as Bagging, Random Forest, Voting, and Boosting. The best model, the Bagging classifier, achieved an accuracy of 94% and demonstrated excellent overall performance. This comparative analysis and the suggestions provided can contribute to enhancing predictive models in the context of airline passenger satisfaction [**3**].

In another study AlHabbal (2022) tried to predict airline passenger satisfaction with Decision Tree and Neural network reaching an accuracy level between 85% and 86% for both of the classifiers. In this study those classifiers reached better results between 93% and 94% [**4**]. Ulkhaq (2020) used neural networks to predict customer loyalty [**5**]. This study combined with the one conducted by Ulkhaq and his collegues could lead to a improvment on prediction customer satisfaction.

A classification model was developed using machine learning techniques and compared with the study by Mirthipati (2024), which used a combined approach of machine learning and causal analysis to improve customer satisfaction, focusing on the online check-in experience. Mirthipati demonstrated that improvements in digital services can significantly enhance overall customer satisfaction by applying machine learning models optimized through advanced parameter optimization techniques [**6**]. To further improve the performance of the classification model, several strategies can be considered. Integrating causal analysis techniques could provide deeper insights into the relationships between satisfaction factors and overall satisfaction. The use of ensemble learning, such as Bagging, Random Forest, and Boosting, has shown to reduce variance and improve model generalization, making them more robust. Parameter optimization through Grid Search could enhance model performance. Additionally, the use of feature engineering techniques and SHAP analysis can improve model interpretability, allowing for a better understanding of the importance of variables and their impact on prediction.

By comparing the study's findings with those of other notable works, such as those by Shane, Chris Bacani, AlHabbal, Ulkhaq, and Mirthipati, the robustness of the results was validated, and unique insights were derived. Future work could focus on a combination of all the results obtained by all the studies to give even better insights of the customers satisfaction.

Accuracy	Precision	Recall	Specificity	F-measure
0.93	0.95	0.90	0.96	0.92

Table 20
Voting

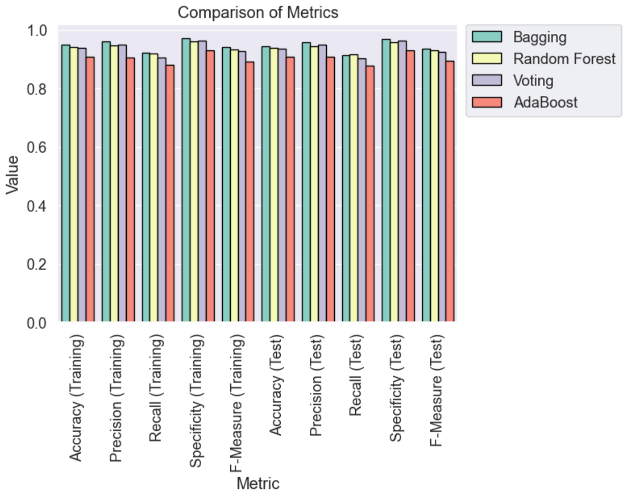


Figure 14. Comparison of metrics of the ensemble classifier.

9 Conclusion

The exploration of airline customer satisfaction using machine learning techniques revealed significant insights into the effectiveness and limitations of various predictive models. Each method showcased unique advantages and challenges, contributing to a comprehensive understanding of customer satisfaction factors.

The application of logistic regression, specifically with LASSO regularization, demonstrated a balance between accuracy and interpretability. This model effectively identified the most influential variables affecting customer satisfaction.

Stochastic Gradient Descent (SGD) classifiers highlighted their strength in computational efficiency and adaptability, particularly with large datasets. The grid search approach for tuning the learning rate showed that constant learning rates could offer quicker training, although potentially at the cost of slightly lower precision and recall compared to more sophisticated models.

Decision trees provided high interpretability and robustness against correlated attributes, making them a valuable tool for initial data exploration and straightforward predictive tasks. However, the tendency to overfit without careful parameter tuning was a notable disadvantage. Imposing depth limits and using grid search for optimization improved their performance but did not surpass ensemble methods.

Neural networks, particularly those using hyperbolic tangent activation functions, excelled in modeling nonlinear relationships within the data. These models achieved high accuracy and recall, although they required extensive training and parameter tuning. Their complexity and training time might be a trade-off for their superior performance in capturing intricate patterns.

Ensemble classifiers, including Bagging, Random Forest, Vot-

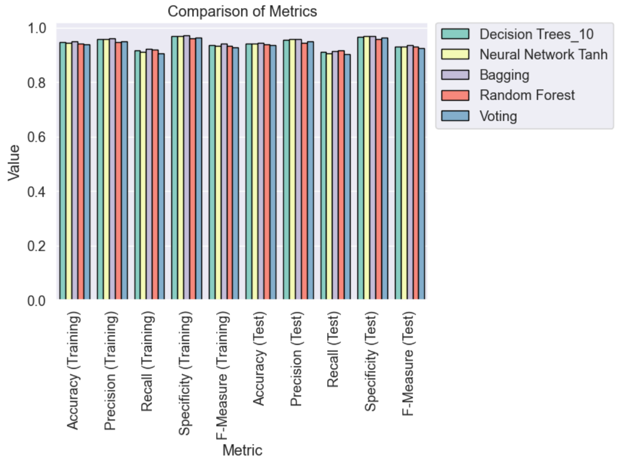


Figure 15. Comparison of metrics of best classifier.

ing, and AdaBoost, consistently outperformed single classifiers by combining multiple models to reduce variance and improve robustness. Bagging and Random Forest were particularly effective, with Random Forest benefiting from decorrelating trees through random feature subsets, enhancing generalization. Voting classifiers, leveraging diverse algorithms, provided robust performance across varied metrics. AdaBoost, while generally effective, showed slightly lower performance due to higher classification errors for both positive and negative records.

The critical evaluation of these methods revealed that ensemble techniques, especially Bagging and Random Forest, offer the best balance of accuracy, robustness, and error minimization. These methods mitigate the weaknesses of individual models, enhancing overall predictive performance.

Future research should focus on advanced feature engineering to uncover deeper insights into customer satisfaction. Incorporating causal analysis can provide a more profound understanding of the relationships between satisfaction factors and overall satisfaction. Additionally, further exploration of ensemble methods with advanced hyperparameter optimization and continuous model evaluation is essential to maintain and improve predictive accuracy.

Acknowledgment

This study is the result of the collaboration between both students, each of whom contributed significantly and equally to different phases of the process. In particular, Ilaria Gallo conducted a preliminary data analysis and formulated the related assumptions. Remo Irtuso then performed the data preprocessing based on his colleague's analysis. Subsequently, various classifiers were chosen and individually implemented by the students. Specifically, Ilaria Gallo worked on the Random Forest, Boosting, SGD and Neural Networks classifiers, while Remo Irtuso focused on Logistic Regression, Decision Trees and Voting. The comparison and study of the sources and existing papers were jointly conducted by both. Both students contributed equally to the writing of the paper, ensuring that every part of the work reflected their shared commitment.

References

[1] Klein, T. (2020, February 20). *Airline passenger satisfaction.* Kaggle. https://www.kaggle.com/datasets /teejmahal20/airline-passenger-satisfaction

[2] Shane, Hannah. "*36-315 Final Project: Airline Passenger Satisfaction.*" Cmu.edu, May 2023.

[3] Bacani, Chris. "*Explaining Airline Passenger Satisfaction Using Interpretable Machine Learning.*" Medium, 21 Nov. 2022,

[4] Ridwan, Mhd, AlHabbal. "*Predicting & Optimizing Airlines Customer Satisfaction Using Classification.*"

[5] Ulkhaq, M. Mujiya & Adyatama, Arga & Fidiyanti, Finsaria & Rozaq, Riyan & Raharjo, M.. (2020). "*An Artificial Neural Network Approach for Predicting Customer Loyalty: A Case Study in an Online Travel Agency*". International Journal of Machine Learning and Computing. 10. 283-289. 10.18178/ijmlc.2020.10.2.933.

[6] Mirthipati, Tejas. "*Enhancing Airline Customer Satisfaction: A Machine Learning and Causal Analysis Approach*", Georgia Institute Of Technology Atlanta, USA, year:2024.

10 Appendix

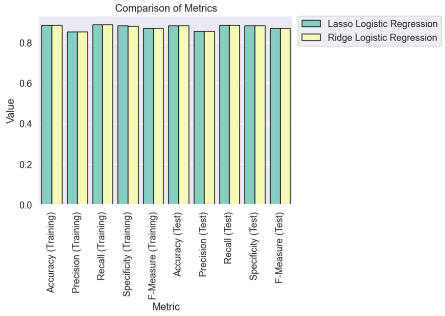
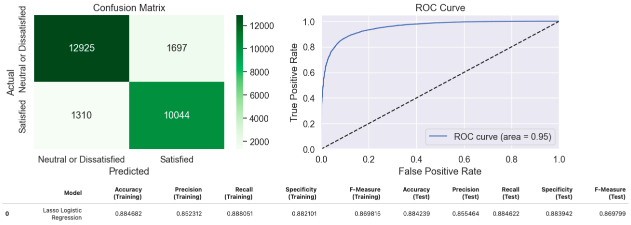
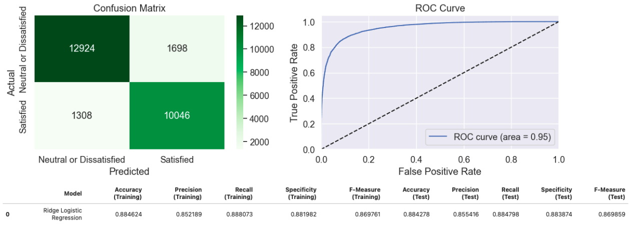


Figure 16. Comparison between Lasso Logistic Regression and Ridge Logistic Regression.



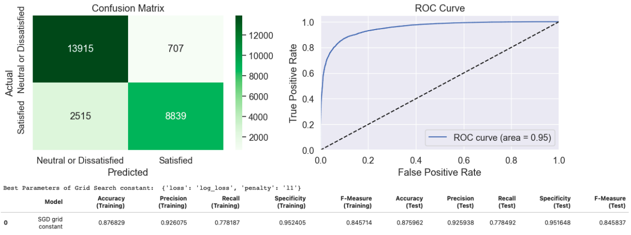
	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Lasso Logistic Regression	0.884682	0.852312	0.888051	0.882101	0.869815	0.884239	0.855464	0.884622	0.883942	0.869799

Figure 17. Lasso Logistic Regression.



	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Ridge Logistic Regression	0.884624	0.852189	0.888073	0.881982	0.869761	0.884278	0.855416	0.884798	0.883874	0.869859

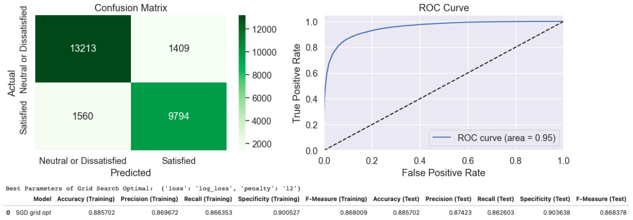
Figure 18. Ridge Logistic Regression.



```
Best Parameters of Grid Search constant:  {'loss': 'log_loss', 'penalty': 'l1'}
```

	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	SGD grid constant	0.876829	0.926075	0.778187	0.952405	0.845714	0.875962	0.925938	0.778492	0.951648	0.845837

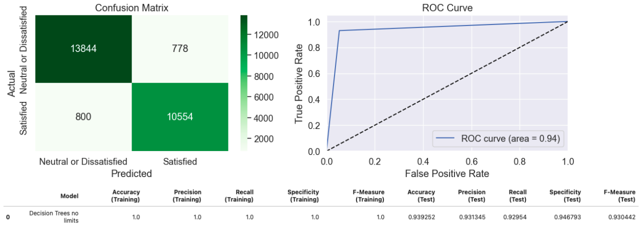
Figure 19. SGD grid constant.



```
Best Parameters of Grid Search Optimal:  {'loss': 'log_loss', 'penalty': 'l2'}
```

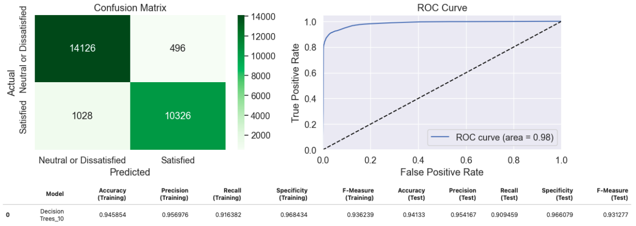
	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	SGD grid opt	0.885702	0.869672	0.866353	0.900527	0.868009	0.885702	0.87423	0.862603	0.903638	0.868378

Figure 20. SGD grid opt.



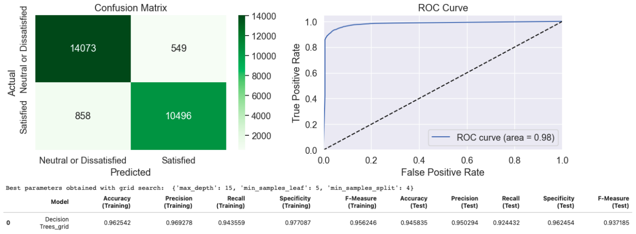
	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Decision Trees no limits	1.0	1.0	1.0	1.0	1.0	0.939252	0.931345	0.92954	0.946793	0.930442

Figure 21. Decision tree without limitations.



	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Decision Trees_10	0.945854	0.956976	0.916382	0.968434	0.936239	0.94133	0.954167	0.909459	0.966079	0.931277

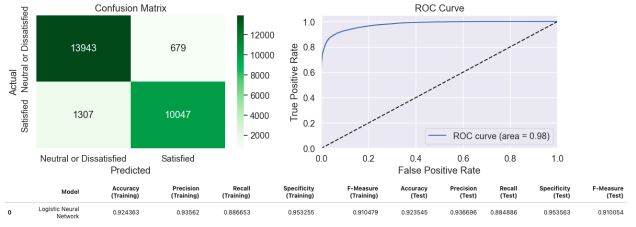
Figure 22. Decision tree with limitation.



```
Best parameters obtained with grid search:  {'max_depth': 15, 'min_samples_leaf': 5, 'min_samples_split': 4}
```

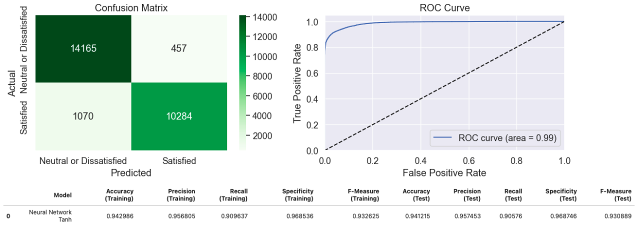
	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Decision Trees_grid	0.962542	0.969278	0.943559	0.977087	0.956246	0.945835	0.950294	0.924432	0.962454	0.937185

Figure 23. Grid-searched tree .



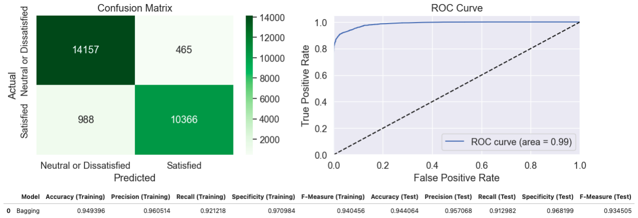
	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Logistic Neural Network	0.924363	0.93562	0.886653	0.953255	0.910479	0.923545	0.936696	0.884886	0.953563	0.910054

Figure 24. Neural network with logistic sigmoid function .



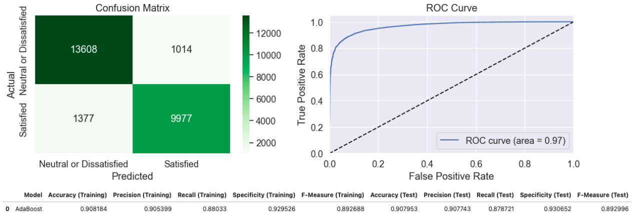
	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Neural Network Tanh	0.942986	0.956805	0.909637	0.968536	0.932625	0.941215	0.957453	0.90576	0.968746	0.930889

Figure 25. Neural network with hyperbolic tangent function.



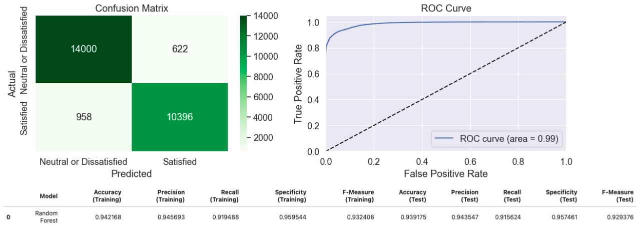
	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Bagging	0.949396	0.960514	0.921218	0.970984	0.940456	0.944064	0.957068	0.912982	0.968199	0.934505

Figure 26. Bagging classifier’s metrics.



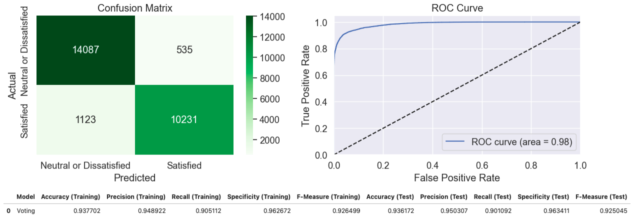
	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	AdaBoost	0.908184	0.905399	0.88033	0.929526	0.892688	0.907953	0.907743	0.878721	0.930652	0.892996

Figure 27. AdaBoost metrics.



	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Random Forest	0.942168	0.945693	0.919488	0.959544	0.932406	0.939175	0.943547	0.915624	0.957461	0.929376

Figure 28. Random Forest classifier’s metrics.



	Model	Accuracy (Training)	Precision (Training)	Recall (Training)	Specificity (Training)	F-Measure (Training)	Accuracy (Test)	Precision (Test)	Recall (Test)	Specificity (Test)	F-Measure (Test)
0	Voting	0.937702	0.948922	0.905112	0.962672	0.926499	0.936172	0.950307	0.901092	0.963411	0.925045

Figure 29. Voting classifier’s metrics.

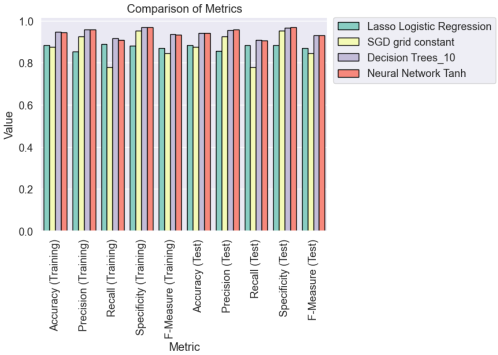


Figure 30. Comparison of metrics of the simple classifier.