US AIRLINE PASSENGER SATISFACTION

Aurora Toma
Department of Computer Science
University of Bari
Bari, Italy
a.toma21@studenti.uniba.it

September 29, 2024

Abstract

The primary objective of this study was to accurately predict passenger satisfaction for an American airline, utilizing survey data from approximately 130,000 passengers. The survey considered various factors influencing satisfaction. This prediction task is crucial for the airline as it allows for targeted improvements in service quality, operational efficiency, and customer experience. Indeed, by understanding the key drivers of satisfaction, the airline can implement strategic changes to enhance passenger comfort, streamline processes, and ultimately increase customer loyalty and retention.

Initially, a data pre-processing phase was conducted to clean the dataset of any noise. Moreover, mutual information was applied in order to select the most influent features to take into account during the training phase and then, various techniques such as cross-validation and grid search were employed to determine the optimal hyper-parameters for models deemed useful for the classification task. The robustness of these models was evaluated by applying metrics on means and standard deviations of each fold obtained by the cross validation step. Moreover, a T-test was also performed to define if any significant differences between models were present. Finally, the most performant models were selected with their tuned hyper-parameters, trained, and then evaluated exploiting confusion matrices, ROC and Precision-Recall curves to ensure robust performance.

In conclusion, to determine whether the hypothesis of considering only relevant features provided advantages, a comparative analysis was conducted between the application of this process on the dataset containing only the relevant features and the original, complete dataset.

***K*eywords** Classification · Decision Tree · Random Forest · Logistic regression · KNN

Introduction

In the highly competitive airline industry, understanding and enhancing passenger satisfaction is paramount for maintaining customer loyalty and achieving operational excellence. The main goal of this study is to accurately predict satisfaction levels of a US Airline based on a comprehensive set of influencing factors. Indeed, by leveraging advanced machine learning techniques survey data from approximately 130,000 passengers of an American airline was analyzed.

The motivation behind this work stems from the need to provide airlines with actionable insights that can drive improvements in service quality, operational efficiency, and overall customer experience. By accurately predicting passenger satisfaction, airlines can make informed decisions to enhance several aspects of their services.

To achieve these objectives, the study begins with a deep data pre-processing phase to cleanse the dataset of any noise, ensuring the reliability of the subsequent analysis. Following this, machine learning techniques such as cross-validation and grid search are employed to identify the optimal hyper-parameters for the classification models. An evaluation phase along with a t-test comparison are then applied. And finally, the most performant models are then selected, trained, and evaluated using confusion matrices, ROC curves, and Precision-Recall curves.

This document is organized as follows: Section 1 describes the dataset utilized and the pre-processing phase needed to clean it of any noise, ensuring the reliability of the subsequent analysis. Section 2 depicts the identification of relevant features. Section 3 employs hyper-parameters tuning within a cross validation step, useful to reduce the risk of overfitting and to gain a more accurate estimate of the expected performances of each model. It was also applied a T-test to statistically compare them and select the most promising ones for the next phase. Section 4 explores the most performing models defined previously, also discussing results gained throughout confusion matrices and the main curves. Then, section 5 describes the comparative analysis conducted between the application of this process on the dataset containing only the relevant features and the original, complete dataset. Finally, conclusions with a summary of the main findings will be discussed.

1 Dataset and Pre-Processing phase

In this section, the dataset used to achieve the main objective will be described along with the analysis needed to implement the pre-processing phase. It is available on Kaggle [1] and it consists of approximately 130,000 rows corresponding to responses from different passengers to a survey. The latter asked passengers to indicate their satisfaction levels regarding various aspects of the airline, as well as provide specific information about their flight, such as the travel class chosen and the expected flight distance, among others.

Features will be displayed by figure 1 with their distribution, and in this case the target one is 'satisfaction-v2', indicating the airline satisfaction level ('Satisfied', 'Neutral or Dissatisfied'). Therefore, the task corresponds to classify whether a passenger is satisfied or not with the airline, based on the responses provided in the survey.

To begin, the conversion of all nominal features into numerical representations was performed, using a suitable encoding method such as the label encoding one. This step is essential for most machine learning algorithms, which typically require numerical input, but also for the next step in which mutual information will be computed, requiring numerical inputs. Following this, a comprehensive search for missing values within the dataset was conducted. It was revealed that only 'Arrival Delay in Minutes' feature had missing values, and since its value could depend on different factors in the real life, it was chosen to drop those rows containing them instead of fulfilling them, also considering they're quite few. Next, the issue of inconsistent data types within the 'delay' columns was addressed. By converting all values of these columns to a single, compatible data type (integer), data consistency and facilitated numerical operations were ensured. Then, to enhance readability the 'Satisfaction-v2' column was renamed to a simpler label corresponding to 'Satisfaction' only.

Finally, features' distribution was visualized using histograms, as shown in figure 1, and it can be seen that the target class is quite balanced, indicating that further changes to dataset are not needed. Then, different factors are considered while providing an opinion of the overall airline passenger's satisfaction. Moreover, taking into account all the answers, in almost all cases delays were recorded as 0 minutes, indicating that the majority of flights arrived on time.

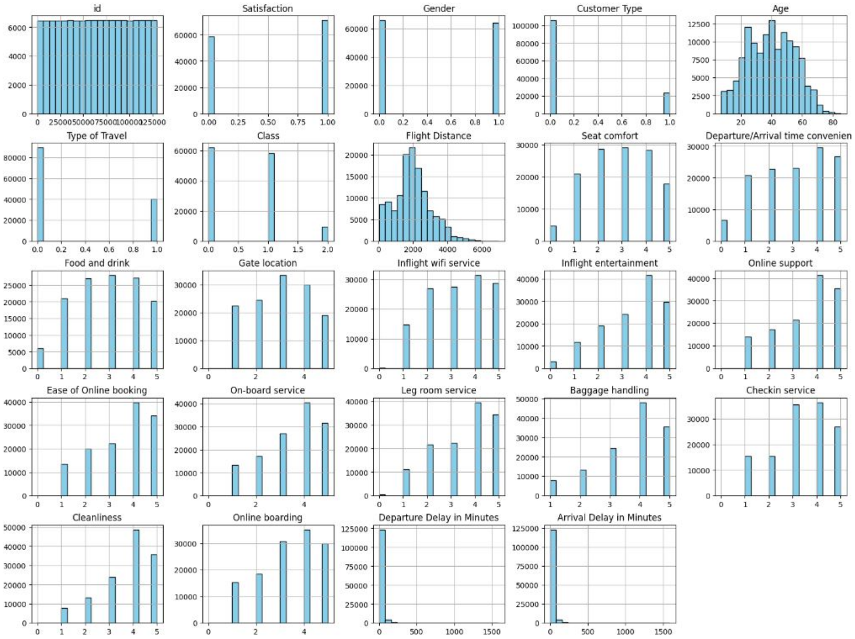


Figure 1: Features distribution.

2 Feature Selection

To identify the most informative features in the considered dataset, the Mutual Information between each feature and the target variable was computed. Mutual information quantifies the dependency between two variables, revealing how much knowledge of one variable reduces uncertainty about the other [2]. By calculating mutual information, a numerical score indicating features' relevance can be obtained. Indeed, higher scores indicates stronger relationships, suggesting that the corresponding feature is more informative for predicting the target. This method is valuable because it's distribution-free, capable of capturing both linear and non-linear relationships, and provides an intuitive interpretation. However, it's sensitive to missing data and can be computationally expensive for large datasets. By ranking features based on their mutual information scores, the top-performing ones were selected for subsequent analysis. This approach ensures the deemed models to focus on the most relevant aspects of the data, leading to improved performance and interpretability.

As depicted in the histogram in figure 2, the most influential factors contributing to passenger satisfaction are first of all 'Inflight entertainment', followed by 'Ease of Online booking', 'Online support' and 'Seat comfort'. Conversely, the remaining features, including various types of delays and the passenger's identification information, exhibit a diminishing impact on overall satisfaction.

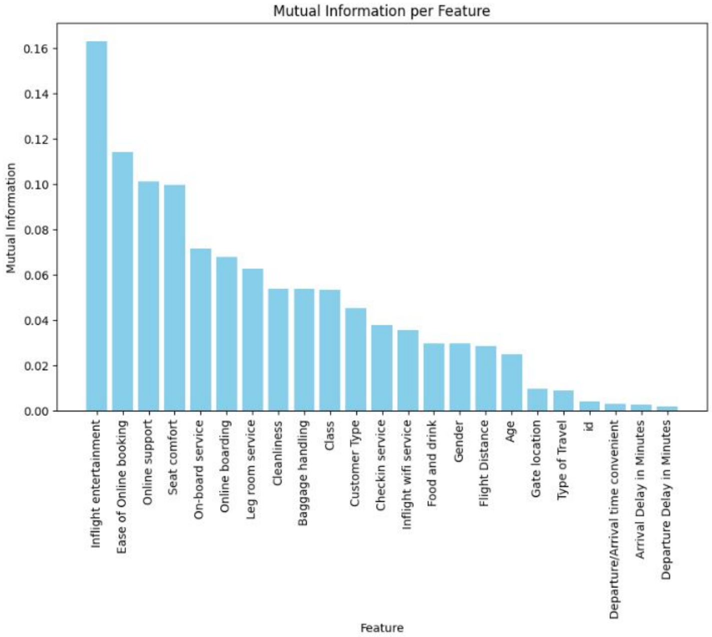


Figure 2: Mutual Information histogram per feature.

While it might seem counterintuitive to consider delays as non influent, the data suggests that the majority of flights operate without delays. Consequently, the presence or absence of delays provides limited discriminatory power in predicting passenger satisfaction, since people expect the flight to be in time.

To enhance the predictive accuracy of our model, the next analysis will focus exclusively on the most relevant features.

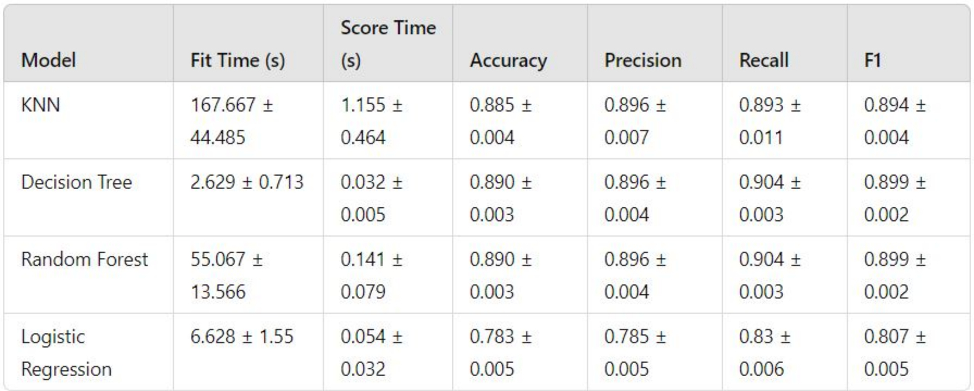
3 Hyper-parameter Tuning and Model Comparison

In this section, the optimization of the performance of the selected models will be discussed. Indeed, a Grid Search algorithm was employed to identify the optimal hyper-parameters for each of them. This hyper-parameter tuning process will be nested within a Stratified K-Fold Cross-Validation framework with k=10. The purpose of nested cross-validation is to mitigate the risk of overfitting by separating the hyper-parameter tuning and final model evaluation stages. Additionally, K-Fold CV provides a more reliable estimate of each model's expected performance, leading to a greater knowledge of the best possible model to use during the effective training and evaluation phase [3].

Models under consideration, along with their respective hyper-parameters to be tuned, are: K-Nearest Neighbors (KNN), having as hyper-parameters n-neighbors (Number of neighbors considered for predictions); weights (Weighting scheme for neighbor contributions); metric (Distance metric used for neighbor identification). Then, Decision Tree, along with max-depth (Maximum depth of the tree); criterion (Impurity measure for split evaluation). Next, Random Forest, considering n-estimators (Number of trees in the forest); max-depth (Maximum depth of individual trees). Finally, Logistic Regression, with C (Regularization strength); penalty (Type of regularization); solver (Optimization algorithm).

It's important to note that there have been different attempts to incorporate Support Vector Machines (SVM) into this analysis. However, even exploring various techniques to speedup and simplify the SVM process, the computational complexity proved excessive for the available resources, preventing its practical implementation.

Upon completion of the hyper-parameter tuning and cross-validation, the performance of each model was evaluated and interpreted. Indeed, Precision, Recall, F1 and Accuracy metrics were deemed on the mean and the standard deviation of all CV folds, leading to results showed by figure 3.



Model	Fit Time (s)	Score Time (s)	Accuracy	Precision	Recall	F1
KNN	167.667 ± 44.485	1.155 ± 0.464	0.885 ± 0.004	0.896 ± 0.007	0.893 ± 0.011	0.894 ± 0.004
Decision Tree	2.629 ± 0.713	0.032 ± 0.005	0.890 ± 0.003	0.896 ± 0.004	0.904 ± 0.003	0.899 ± 0.002
Random Forest	55.067 ± 13.566	0.141 ± 0.079	0.890 ± 0.003	0.896 ± 0.004	0.904 ± 0.003	0.899 ± 0.002
Logistic Regression	6.628 ± 1.55	0.054 ± 0.032	0.783 ± 0.005	0.785 ± 0.005	0.83 ± 0.006	0.807 ± 0.005

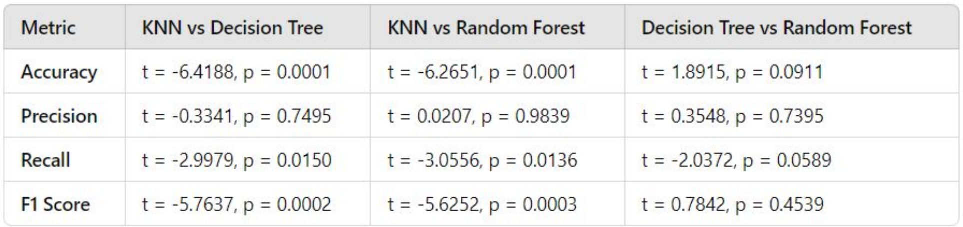
Figure 3: Metrics evaluation for each model considered.

The evaluation of KNN, Decision Tree, and Random Forest revealed that, beyond their computational time, these models demonstrated commendable performance, consistently achieving accuracy scores near 0.9 across most metrics. This indicates their proficiency in accurately classifying both positive and negative instances. Whereas, Logistic Regression under-performed with respect to the others, suggesting its limited capability in correctly classifying instances. This may be due to its simplicity, which can lead to data under-fitting.

Regarding model stability, most models exhibited low standard deviations across the cross-validation folds, indicating consistent performance across different data subsets. This suggests that these models are reliable and generalize well to unseen data.

Now, the most promising models will be subjected to a paired t-test to statistically assess if there are any significant differences between them. In machine learning, the t-test is used to determine whether the difference in performance between two models is statistically significant or merely due to random variation in the data. Specifically, a paired t-test is applied when comparing two models that have been evaluated on the same dataset, typically using cross-validation results [3].

The test evaluates the null hypothesis, which assumes that there is no significant difference between the models' performances. Then, if the p-value from the t-test is below a certain threshold (e.g., 0.05), the null hypothesis is rejected, concluding that the observed difference is statistically significant. This helps ensure that any observed performance difference is not just by chance, but likely reflects a true difference in the models' effectiveness. Results gained by the statistical T-test are described in figure 4.



Metric	KNN vs Decision Tree	KNN vs Random Forest	Decision Tree vs Random Forest
Accuracy	t = -6.4188, p = 0.0001	t = -6.2651, p = 0.0001	t = 1.8915, p = 0.0911
Precision	t = -0.3341, p = 0.7495	t = 0.0207, p = 0.9839	t = 0.3548, p = 0.7395
Recall	t = -2.9979, p = 0.0150	t = -3.0556, p = 0.0136	t = -2.0372, p = 0.0589
F1 Score	t = -5.7637, p = 0.0002	t = -5.6252, p = 0.0003	t = 0.7842, p = 0.4539

Figure 4: T-test applied on the most promising models.

Based on the results of the T-test, it can be concluded the following:

- Decision Tree vs KNN: Decision Tree is significantly better in accuracy, recall, and F1-score, but not in precision.

- Random Forest vs KNN: Random Forest outperforms KNN in accuracy, recall, and F1-score, but there is no difference in precision.

- Decision Tree vs Random Forest: No significant difference in accuracy, recall or F1-score, but Random Forest has better precision.

4 Model Performance Evaluation

This section discusses the performance of the selected models (Random Forest and Decision Tree), including results from Confusion Matrices and other evaluation metrics like ROC and Precision-Recall curves.

Figure 5 displays the confusion matrix obtained by the Decision Tree model. Whereas, figure 6 displays the one regarding Random Forest.

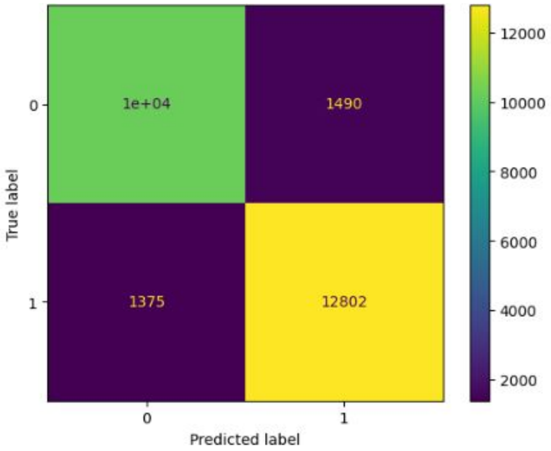


Figure 5: Decision Tree's confusion matrix.

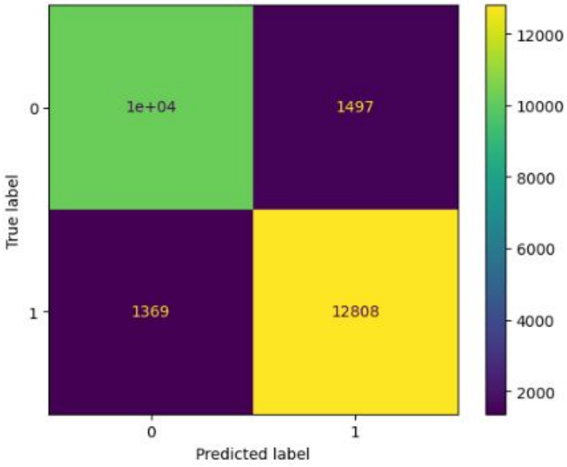


Figure 6: Random Forest's confusion matrix.

Looking at both confusion matrices compared they can be interpreted as follows: The number of False Positives (incorrect predictions of class 1 when it was 0) is very similar between the two models (1,490 for Decision Tree vs 1,497 for Random Forest). The number of False Negatives (incorrect predictions of class 0 when it was 1) is slightly better for Random Forest (1,369) compared to Decision Tree (1,375). Random Forest has slightly more True Positives (12,808 compared to 12,802 for Decision Tree), which suggests that Random Forest has a slight advantage in recognizing positive samples compared to the Decision Tree.

Therefore, overall, Random Forest seems to perform slightly better at correctly recognizing positive class samples compared to the Decision Tree.

Finally, when interpreting ROC curves, the ideal curve would be positioned as close as possible to the top-left corner of the plot. This indicates a high True Positive Rate (sensitivity) and a low False Positive Rate (specificity). During its computation, various probabilities were deemed in order to change the decision threshold of the model (the default is 50% - meaning that if a sample belongs to a class with a probability greater than that threshold, then it belongs to that class for sure), and then visualize any possible situation the model could achieve [3].

Whereas, when evaluating models using Precision-Recall curves, the goal is to maximize both Precision and Recall. Therefore, the ideal curve would be positioned as close as possible to the top-right corner of the plot [3].

In this analysis, both models exhibited identical performance across all possible thresholds. This suggests that the two models have learned essentially the same underlying function, as demonstrated in the comparison displayed in figures 7 and 8.

This situation is given by the fact that t-tests revealed no statistically significant differences between Decision Tree and Random Forest. Indeed, their performance metrics and confusion matrices were remarkably similar. Furthermore, the

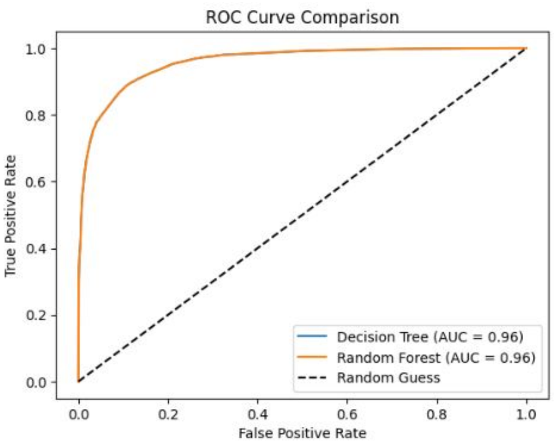


Figure 7: ROC curves of DT and RF models.

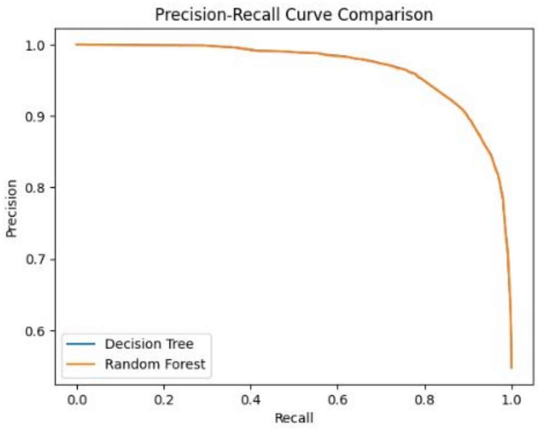


Figure 8: Precision-Recall curves of DT and RF models.

final learning curves confirmed that both models essentially learned the same underlying function, resulting in identical curves.

Despite their comparable performance, Decision Tree model is preferred thanks to its superior interpretability compared to Random Forest. Additionally, Decision Trees are generally less complex than Random Forests, making them more computationally efficient in certain scenarios.

5 Comparative Analysis of Feature-Selected vs. Complete Dataset Performance

This section aims to conduct the analysis described in the title, where a comparative study is performed between the application of the same process on a dataset containing only the most relevant features and the original, complete dataset. This analysis is useful to determine if considering relevant features only will lead this prediction task to better performances or not. The same pre-processing steps, as outlined in the section 2, will be applied here. However, the key difference lies in the fact that the entire dataset will be considered in this analysis. Specifically, tree-based algorithms (Decision Tree and Random Forest) will be applied, as they were identified as the top-performing models in the previous section.

Figure 9 displays the confusion matrix obtained by the Decision Tree model. Whereas, figure 10 displays the one regarding Random Forest.

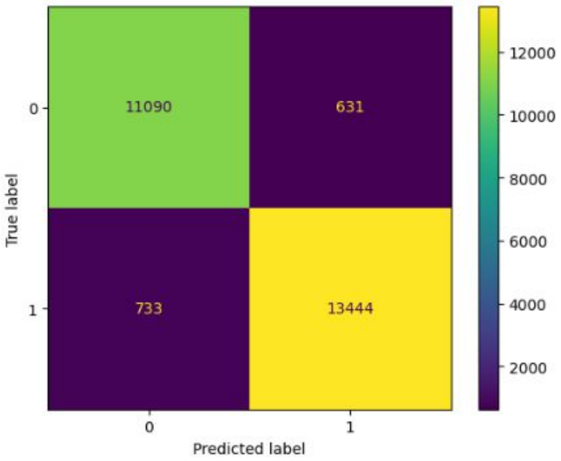


Figure 9: Decision Tree's confusion matrix.

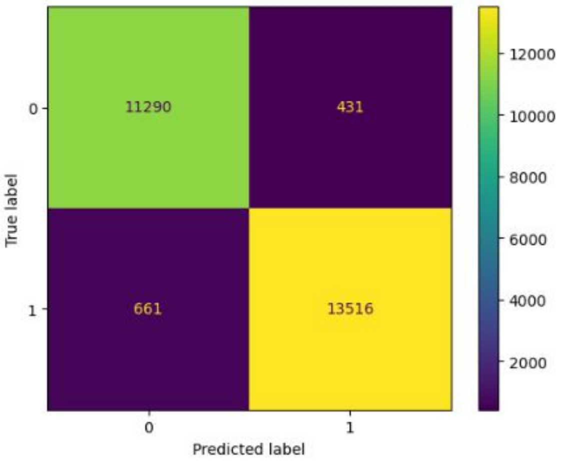


Figure 10: Random Forest's confusion matrix.

Looking at both confusion matrices compared they can be interpreted as follows: The number of False Positives is slightly better for Random Forest (431) compared to Decision Tree (631). The number of False Negatives is slightly better for Random Forest (611) compared to Decision Tree (733). Random Forest has slightly more True Positives (13,516 compared to 13,444 for Decision Tree), which suggests that Random Forest has a slight advantage in recognizing positive samples compared to the Decision Tree. Therefore, overall, Random Forest seems to perform slightly better at correctly recognizing positive class samples compared to the Decision Tree. Moreover, comparing them with respect to figures 5 and 6, it's evident that considering the whole dataset will lead to quite better performances for both models.

Finally, ROC and Precision-recall curves will be visualized by figures 11 and 12, exploiting their functionalities and then ending the comparison between the two models.

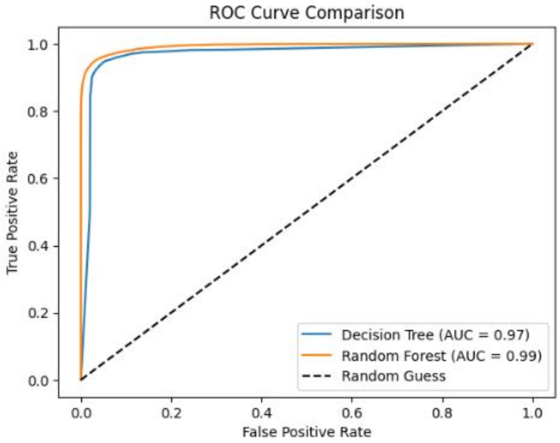


Figure 11: ROC curves of DT and RF models.

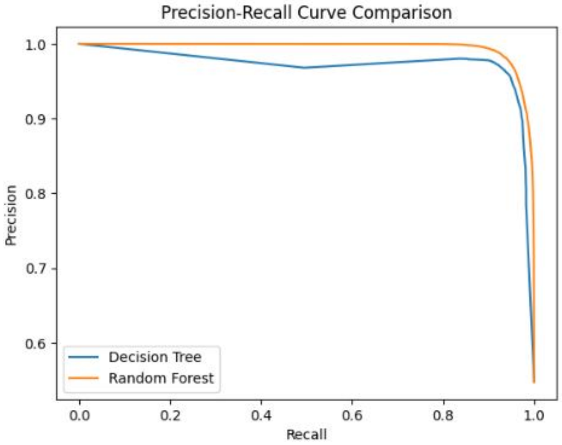


Figure 12: Precision-Recall curves of DT and RF models.

As shown, both models exhibited quite good performance across all possible thresholds. Nevertheless, Random Forest proved to be slightly better than Decision Tree in both curves.

To summarize, this analysis indicates that considering relevant features only will lead to performances that are slightly worse than those we would have while considering the whole dataset. Moreover, if we take into account the entire dataset we would prefer the Random Forest, instead of the Decision Tree, due to its evident superior capability of capturing data distribution.

Conclusions

In conclusion, the analysis of airline passenger satisfaction leveraged a comprehensive dataset from Kaggle [1], consisting of approximately 130,000 survey responses. After a thorough pre-processing phase, which included handling missing values, numerical encoding, and ensuring data consistency, mutual information was used to select the most relevant features for model training [2]. Key aspects such as 'Inflight entertainment', 'Ease of Online booking', 'Online support' and 'Seat comfort' emerged as the most influential predictors of passenger satisfaction. In contrast, delays were largely inconsequential due to the majority of flights arriving on time, providing limited discriminatory power for predicting satisfaction.

Next, to optimize model performance, a nested cross-validation strategy was employed, integrating a Grid Search algorithm within a Stratified K-Fold Cross-Validation framework (k=10) [3]. This approach effectively minimized the risk of overfitting and provided a reliable estimate of each model's expected performance. The selected models (K-Nearest Neighbors, Decision Tree, Random Forest, and Logistic Regression) were tuned based on their respective hyper-parameters, leading to high-performing results for the tree-based models.

Once the hyper-parameter tuning was complete, model evaluation based on Accuracy, Precision, Recall, and F1-score showed that the Decision Tree and Random Forest models consistently achieved accuracy scores near 0.9, far outperforming Logistic Regression, which demonstrated limitations in classifying the data effectively. The consistency of these results across cross-validation folds, as evidenced by low standard deviations, suggested that both tree-based models were robust and generalized well to unseen data.

To further assess whether the observed performance differences were statistically significant, a paired t-test was applied to the results [3]. The t-test revealed that the Decision Tree model outperformed KNN in accuracy, recall, and F1-score, but not in precision. Similarly, Random Forest showed significant superiority over KNN in the same metrics, with no difference in precision. However, when comparing Decision Tree and Random Forest, there was no significant difference in accuracy, recall or F1-score, but Random Forest achieved better precision. Furthermore, the lack of significant difference between these two models led both of them to learn the same underlying function, as resulted in identical curves computed.

To conclude, despite these comparable results, the Decision Tree model is preferred due to its superior interpretability and reduced computational complexity compared to Random Forest, making it a more practical choice in scenarios requiring transparent decision-making and efficiency.

A further analysis was undertaken to perform a comparative study between applying the same process to a dataset containing only the most relevant features and the original, complete dataset. This analysis is essential to determine whether focusing solely on the most relevant features yields improved performance in the prediction task. When comparing the results obtained in this section to those previously presented, it becomes clear that considering the entire dataset leads to better performance for both models. Furthermore, when the full dataset is used, the Random Forest emerges as the preferred model over the Decision Tree, owing to its clearly superior ability to capture the underlying data distribution.

In conclusion, while the hypothesis of utilizing only the most relevant features may seem intuitively correct, the results indicate that this approach has led to inferior performance compared to considering the entire dataset.

References

[1] https://www.kaggle.com/datasets/johndddddd/customer-satisfaction

[2] Claude Elwood Shannon, A mathematical theory of communication, October 1963.

[3] Witten & Frank, Data Mining: Practical Machine Learning Tool and Techiques, 2017 - 4th ed.