Bachelor's Thesis Computing Science

Radboud University Nijmegen

Exploring Feature Importances for Predicting Flight Delays Using Machine Learning Algorithms and Transfer Learning

---

Author:
Tygo Francissen
s1049742

First supervisor/assessor:
Dr Tom Claassen

Second assessor:
Dr Gabriel Bucur

July 4, 2023

Abstract

Flight delays have become a growing concern for passengers, airlines, and airports, and have significant environmental implications. This paper presents a comparative study of machine learning algorithms for predicting aircraft delays using historical flight data from the United States and Brazil. A wide range of algorithms is compared, including decision tree, random forest, gradient boosting tree, k-nearest neighbors, naive Bayes, logistic regression, neural network, and support vector machine, and evaluated by assessing their performance with metrics such as accuracy, precision, recall, F_1 score, and AUC score.

In this study, the most prominent factors contributing to aircraft postponements are also identified, which can be used by airlines and airports to take countermeasures against flight delays. Additionally, the use of transfer learning is considered to increase prediction accuracy and efficiency.

The results show that the random forest and gradient boosting tree classifiers achieve the highest performance, with potential accuracies of 81% and 86% respectively. These classifiers demonstrate a trade-off between accuracy and F_1 score. Furthermore, the analysis reveals that the departure month and hour are the key factors influencing flight delays in Brazil and the United States. Moreover, this research shows that flights scheduled in December and those departing later in the day have a significantly higher likelihood of experiencing delays.

From the obtained results, it can be concluded that leveraging a pre-trained model on the United States data set through transfer learning for predicting flight delays in the Brazilian data set results in an efficiency gain of around 6 times. This technique reduces training time and offers a scalable approach to model development and deployment.

Contents

Chapter 1

Introduction

In 2022, the average delay per flight in Europe, specifically the difference between scheduled and actual departure time experienced by passengers, airlines, and airports, was 17 minutes, an increase of 33% compared to 2019 [15]. In the United States, more than 20% of all flights in 2022 had a delay of more than 15 minutes[1]. Similar amounts of flight delays, also referred to as flight postponements, are experienced by passengers in different parts of the world (e.g., in 2020, the average delay of passenger flights in China was 9 minutes [7]).

High average delays can often translate into significant inconveniences for affected travellers, as they indicate the occurrence of substantial delays for a portion of flights. Moreover, if proactive measures are not implemented, they could potentially contribute to further disruptions in the global aviation system, including flight cancellations and other significant setbacks.

The worldwide outbreak of the coronavirus disease (COVID-19) in November 2019 could be one of the causes of the recent increase in aircraft delays. Research has found that the number of flights dropped to one-fourth of the original amount at the start of COVID-19 [55]. As a result of COVID-19's large flight cuts, approximately 24 million people lost their jobs in the aviation industry [38].

The number of flights is rapidly increasing now that coronavirus vaccines have been developed, but in the fourth quarter of 2022, the number of flight departures was still 16% lower than before the COVID-19 outbreak [9]. Since airlines, airports, and their suppliers severely downsized their workforces during the COVID-19 epidemic, they are not able to keep up with

[1]Bureau of Transportation Statistics. On Time Performance - Reporting Operating Carrier Flight Delays at a Glance, `https://www.transtats.bts.gov/homedrillchart.asp`

the current demand [20]. A significant portion of the discharged employees found another job, which leaves most airlines and airports with huge staff shortages.

The consequences of these shortages are noticeable to both passengers and airlines. Increased waiting lines at airports, cancelled flights, and flight postponements, among others. For example, aircraft operating in Europe experienced a 5-year high of 23 minutes of average delay per flight in the third quarter of 2022 [14]. This remarkable delay highlights the need for countermeasures to decrease aircraft postponements significantly, as the number of flight departures is slowly approaching its pre-COVID level.

Passengers often face the unpredictability of aircraft delays, leading them to allocate extra hours for their travel plans to ensure timely arrival at their destinations, incurring additional expenses as a result [3, 16]. Furthermore, the environmental impact of flight postponements is a complex issue [41]. Decreasing flight delays is not only convenient for passengers and the environment but also for airline carriers. For a carrier, flight delays namely cost billions of dollars each year [57]. As an example, the average cost of aircraft delay time for USA passenger airlines was estimated to be $101 per minute in 2022[2].

To reduce costs and increase passenger satisfaction, some actions have already been taken to decrease aircraft postponements (e.g., hiring new employees and creating optimal flight schedules). However, airports and airlines must find the most prominent causes of flight postponements and take countermeasures based on them.

One strategy to identify the most important causes is to use machine learning algorithms on historical flight data. Machine learning (ML) algorithms have various use cases like data mining, predictive analytics, image processing, etc [33]. In the context of flight delays, various ML algorithms can be utilized to predict whether a flight will have a delay and identify what the most important factors are for a delay.

A wide range of research studies has attempted to predict flight delays and their most prominent factors. However, existing research has (i) only considered one or two ML algorithms for use on their data or (ii) only looked at small-scale data (e.g., only several airports or small countries). To address these gaps in knowledge, this study employs multiple machine learning algorithms to predict flight delays and identify their underlying factors on a

[2]Airlines for America. U.S. Passenger Carrier Delay Costs, `https://www.airlines.org/dataset/u-s-passenger-carrier-delay-costs/`

large-scale data set. As a result, the conducted research provides valuable insights and answers to the following areas of investigation:

- Identification of influential factors in flight delays
  - Determining the best-performing machine learning algorithms for predicting flight delays
  - Uncovering the most significant factors contributing to aircraft postponements across different algorithms
- Evaluation of transfer learning in flight delay prediction

Machine learning algorithms are implemented on the collected data, trying to get the best classification of whether a flight gets delayed. More specifically, the ML algorithms that are compared are the decision tree, random forest, gradient boosting tree, k-nearest neighbors, naive Bayes, logistic regression, neural network, and support vector machine. When the algorithms are trained on the data set using the training data, each model's performance is evaluated by using it on the test data.

After evaluating all algorithms, the best-performing algorithms are selected, and for each of these algorithms, the most important factors for making a classification decision are investigated. All the most prominent factors are collected and summarized in a structured manner.

With these discovered factors, airports, airlines, and passengers can understand the causes of aircraft postponements and, more importantly, take countermeasures to decrease flight delays remarkably. With these measures, the upcoming rise in flight demand can be handled more efficiently without too many aircraft delays.

The impact of transfer learning in predicting aircraft delays is investigated as part of additional research. Transfer learning allows aviation companies to leverage existing knowledge and models to enhance their prediction systems. By adapting pre-trained models to specific datasets, transfer learning offers potential benefits in accuracy and efficiency.

It is, however, a challenge to address the difference between data sets and adapt the models accordingly. This research aims to handle that problem by carefully considering the similarities and differences between the source and target data sets. Transfer learning is carried out in order to find more techniques that aviation companies could use in their systems for predicting flight postponements.

The rest of this paper is organized in the following way: First of all, Chapter 2 introduces and explains the core concepts, terms, and theories that

are used in the rest of this thesis. After that, Chapter 3 summarizes other researchers' work on the same topic and how it relates to this research. Chapter 4 presents the methods and experiments of this paper. Furthermore, Chapter 5 contains the results, a discussion of the findings, as well as future research directions. Finally, the conclusions of this research are summarized in Chapter 6.

Chapter 2

Preliminaries

This chapter overviews the essential concepts, terms, and theories that form this thesis's basis. The focus is mainly on aviation terminology, general data mining strategies, an introduction to the dataset used in this study, preprocessing methods, machine learning algorithms, and evaluation metrics.

2.1 Aviation Terminology

The aviation sector is a complex distributed transportation system that must be coordinated accurately. It deals with various parties including passengers, airline carriers, airports, food suppliers, and other stakeholders. This study deals with various types of flights with the most prominent one being commercial flights.

In commercial aviation, passengers follow their itineraries while airlines plan schedules for aircraft and their staff. A typical operation of a commercial flight has been illustrated in Figure 2.1 [54]. All the presented stages in the figure are susceptible to different types of delays including weather conditions, mechanical problems, ground delays, air traffic control, runway queues and capacity constraints. This scheme is executed several times per day, for each flight.

2.1.1 Flight Delay

There is a difference between the departure delay and arrival delay of a flight. The departure delay refers to the amount of time an aircraft departs after its scheduled departure time, usually measured in minutes. On the other hand, arrival delay is the amount of time a plane arrives after its scheduled arrival time, also measured in minutes.

In general, there are multiple definitions of "delay" in the aviation sector

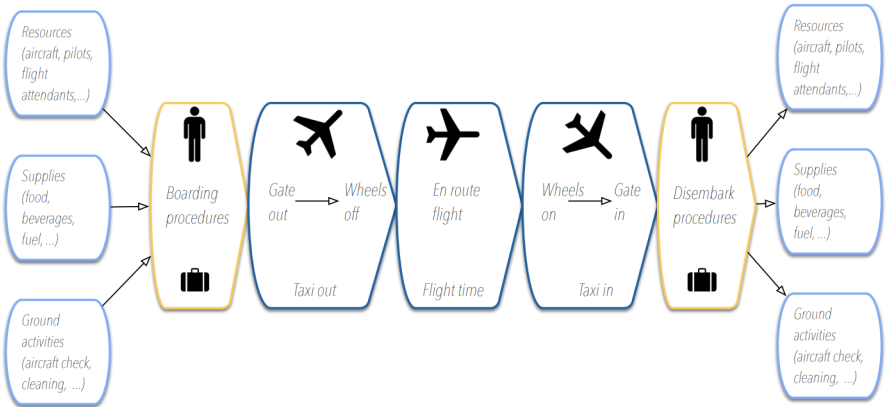


Figure 2.1: A typical operation of a commercial flight.

since different stakeholders may have different priorities when it comes to measuring delay. For a passenger, a delay is simply the difference between the scheduled arrival time (SAT) and the actual arrival time (AAT). However, delays are used to quantify the air traffic performance of the National Airspace System (NAS) and the aforementioned definition of "delay" has shortcomings for the NAS performance metric. It hides other factors like takeoff delays and ground delays. The difference between the SAT and AAT can therefore be defined as the "effective delay" [62].

Another definition of delay is the "technical delay", which represents the difference between the actual operating time and the aircraft's "optimal" operating time. The "optimal" operating time of an aircraft is the time the plane would take, from pushback at the origin airport to arrival at its destination airport, if there were no other aircraft ahead of it and all flow-control mechanisms are absent [62].

2.1.2 Specific Terms

The primary focus of this thesis will be on the prediction of effective delays of flights, which is referred to as "flight delay", "delay", or other similar terms throughout the study.

Another important term when considering the prediction of flight delays is the minimum delay time, which refers to the minimum amount of time a flight must be delayed before it is considered a delayed flight. According to the United States Federal Aviation Administration, a flight must be delayed

for at least 15 minutes to be considered a delayed flight[1].

In conclusion, the flight delay prediction problem can be defined as follows [32]: Given a flight record with all its available features, predict whether the flight will have a delay (of at least 15 minutes).

2.2 Data Mining Definition

Data mining is the process of discovering patterns, trends, and relationships in large data sets that may not be readily apparent [23]. In this study, data mining serves as a valuable tool to uncover interesting patterns and leverage them for classification purposes. By extracting feature importance values from the classification algorithms and applying transfer learning, the research aims to enhance the accuracy of classifying new, unseen instances.

However, a limitation of data mining techniques is the risk of overfitting [22]. Overfitting occurs when the model is trained extensively on the training data, but it performs poorly on new, unseen data. Given the high class imbalance in this research, overfitting is a concern. To address this, balancing techniques are utilized. Additionally, to mitigate the impact of an unrepresentative training set, a train-test split of 33% is employed, and a stratified KFold approach is used during the hyperparameter tuning phase, as explained in subsequent sections.

2.3 Flight Delay Data Sets

The data sets used in this study originate from the United States and Brazil, providing a comprehensive view of flight delays in these regions. The United States data set, obtained from the Bureau of Transportation Statistics (BTS), includes approximately 7 million flight records with 25 features. These features cover various aspects such as time period, airline, origin, destination, departure performance, arrival performance, and cancellations/diversions. On the other hand, the Brazilian data set, obtained from the Active Regular Flight Database (VRA), comprises around 900,000 flight records with 20 features, including origin, destination, and time period information.

Upon analysis, it is observed that variables such as time period, origin, destination, and flight number are shared by both data sets. However, the USA data set includes additional information such as distance, tail number,

[1]United States Federal Aviation Administration. Air Traffic Plans and Publications Order JO 7210.3CC, `https://www.faa.gov/air_traffic/publications/atpubs/foa_html/chap18_section_7.html`

and origin/destination city market, while the Brazilian data set includes details such as model equipment, number of seats, and line type code. Understanding these characteristics is particularly relevant for the subsequent transfer learning problem, where insights gained from one data set can be leveraged to improve predictions on the other.

Both data sets undergo rigorous data preprocessing steps, including the removal of cancelled, diverted, or uninformed flights. Missing values are handled by either removing the affected records or using alternative techniques. Additionally, irrelevant features and variables not known prior to departure are eliminated to focus on relevant attributes for predicting flight delays. The geographical distribution of airports revealed that the United States data set consists exclusively of domestic flight records, while the Brazilian data set includes flights that have either their origin or destination airport located in Brazil.

A more detailed explanation of the data sets, including attributes and insights, can be found in Chapter 4 of this thesis.

2.4 Data Preparation

Data preparation, known as data preprocessing, is a fundamental step in this research aimed at improving the accuracy and reliability of flight delay prediction. Flight data sets often exhibit characteristics such as incompleteness, inconsistencies, potential outliers, and wide attribute ranges, indicating the need for data preprocessing.

Assessing the data quality becomes essential during the initial stages of data collection. This can be done by looking at the data source, the completeness of the data set, and the related work that has been done with the same data. Once the data is collected, it undergoes data preprocessing steps to address the aforementioned issues and ensure the efficiency and accuracy of the subsequent data mining process. By employing techniques like data cleaning, integration, reduction, and transformation, the data is tailored to the specific requirements of the research, enabling the extraction of meaningful patterns and trends [22].

Without proper data preprocessing, the results of data mining, specifically flight delay prediction, could be unreliable and misleading. The following subsections focus on discussing the various steps of data preprocessing and their relevance to the domain of flight delay prediction. These steps are also summarized in Figure 2.2.

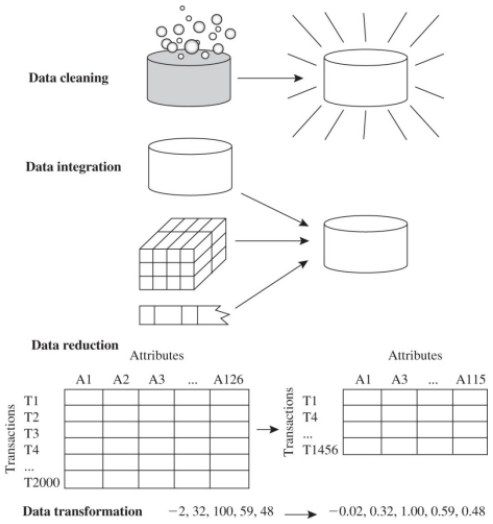


Figure 2.2: Forms of data preprocessing.

2.4.1 Data Cleaning

Data cleaning methods attempt to handle missing values, get rid of noise while identifying possible outliers, and correct inconsistencies in the data set [22]. There are several ways to tackle these issues, but only the ones that are most relevant to this research are discussed.

If multiple records of the data set have a missing value for one of the attributes, this is something that has to be taken care of. There are several possibilities to handle missing values [22]. These are the ones considered in this research:

1. Remove all rows with missing values: This is often done when the class label is missing when performing classification. This strategy is only effective when the row contains multiple attributes with missing values. When using this strategy, useful data could be removed.

2. Manually fill in the missing values: This method is time-consuming and is usually not feasible in a large data set with a lot of missing values.

3. Fill the missing values with a global value: This simple method replaces all missing values with a value like “unknown” or minus infinity. However, this approach may not be universally effective for data mining algorithms. While it provides a quick way to handle missing data, it does not consider the underlying patterns or relationships in the data. As a result, it may introduce biases or inaccuracies in the analysis.

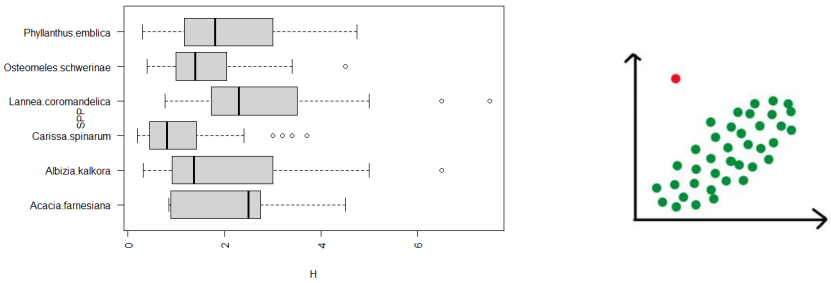


Figure 2.3: Visual representations of potential outliers, represented as circles in a box plot (left) and as the red circle in a scatter plot (right).

These methods are relevant to this research as they offer different approaches to handle missing values in flight data sets, each with its own advantages and considerations. Removing all rows with missing values can be effective when the proportion of missing values is small, as the classifier will still have sufficient data for analysis. Alternatively, missing values can be manually filled in using available historical flight data from multiple sources, enabling the restoration of missing information. Another approach is to replace the missing values with a global value, considering that flight delay data sets typically contain numerous attributes, and a few missing values are unlikely to significantly impact the analysis.

There are other methods to handle missing data including the use of the mean/median and the use of the most probable value based on prediction. However, these methods are not considered in this research for several reasons. Firstly, the mean/median imputation assumes that the missing values have a similar distribution to the observed values, which may not be accurate in the context of flight delay data. Secondly, predicting the most probable value introduces additional uncertainty and relies on the accuracy of the prediction model. It is noticeable that missing values are not always errors in the data, so missing values should first be investigated.

An outlier in a set of data is an observation or point that is considerably dissimilar or inconsistent with the remainder of the data [44]. Based on this definition, outliers may seem to have to be eliminated as quickly as possible, but the converse is true. Outliers can be detected in the data exploration phase by a lot of techniques including visualizations by using histograms, box plots and scatter plots as well as by using statistical methods [1]. Figure 2.3 gives a visual example representation of the detection of outliers.

In flight data sets, outliers can be relevant when considering attributes like the number of seats or flight distance. Data entry errors or anomalies, such

Method	Description
Correct	Rectify the outlier to its correct value
Remove	Eliminate the outlier
Study in detail	Conduct follow-up work to study the outlier in more detail
Keep	Threat the outlier as a normal data point
Have different versions	Report the findings with and without the outlier
Transform	Apply a deterministic mathematical function to each value, which also reduces the error variance and skew of data points
Modify	Change the outlier to another, less extreme value manually

Table 2.1: Overview of some outlier handling techniques based on Aguinis et al. [1]

as an operator that accidentally enters an excessively high value for the number of seats or an erroneous distance, can distort the data and impact the mean calculation and classification. Therefore, data points that are identified as potential outliers should be thoroughly investigated. If it turns out that the data point is considered an outlier, multiple outlier categories and techniques handle each category of outliers [1]. Some outlier handling methods that are relevant to this study are summarized in Table 2.1.

Other steps in the process of data cleaning besides handling missing values and outliers include correcting typos and fixing format errors. These are dataset-dependent and usually straightforward.

2.4.2 Data Integration

Merging multiple data stores - data integration - has to be executed carefully to reduce and avoid redundancies and inconsistencies in the final data set [22]. Merging data from multiple data sources often results in the problem of duplicate records, which can lead to data redundancy. However, in the context of this research, where data sets from different periods are being merged, the occurrence of duplicate records may not be prominent since each data source represents a distinct month of the year.

Consequently, the focus of data integration in this study primarily lies in harmonizing the formats of the different data stores to ensure compatibility and seamless merging of flight data. It is worth noting that in this research, the data sets to be merged are already in the same format, eliminating the

need for additional attention to formatting the different data stores.

2.4.3 Data Reduction

After carefully examining the size of the flight data sets and the distributions of variables, it has been determined that the current data sets used in this research, which cover a limited time period, do not necessitate extensive data reduction strategies. However, it is anticipated that for larger flight data sets that span multiple years of data, data reduction techniques such as dimensionality reduction, numerosity reduction, and data compression may become relevant to improve the efficiency of the analysis without compromising accuracy [22]. The decision to not delve into these techniques in depth for the current data sets is based on their size and scope, but future studies involving larger and more extensive flight data sets may benefit from exploring and implementing appropriate data reduction strategies.

2.4.4 Data Transformation

In the context of this research, the final step of data preprocessing is data transformation, where the flight data is transformed into suitable forms for data mining. This transformation enhances the efficiency of the mining process and facilitates a better understanding of the discovered patterns [22].

Among the various data transformation techniques available, the primary technique utilized in this study is normalization. By normalizing the data, all attributes are given equal weight and are transformed to a smaller range such as [-1, 1] or [0, 1]. This normalization process is particularly beneficial for distance-based classification algorithms and neural networks, as it not only accelerates the training phase but also prevents attributes with larger ranges from overshadowing those with smaller ranges [22]. Additionally, normalization ensures that the data is on a common scale, which is crucial for the correct functioning of many data mining algorithms. Without proper normalization, the results of data mining may lack reliability and could be misleading.

It should be noted that the four phases of data preprocessing — data cleaning, data integration, data reduction, and data transformation — exhibit significant overlap and employ similar strategies. Therefore, researchers often intertwine different preprocessing phases and strategies based on their specific research goals and intentions, tailoring the approach to the particular requirements of the analysis.

2.5 Feature Selection

In the context of this research, correlation analysis plays a crucial role in feature selection by examining redundancies within the collected data set. Given the nature of the data in this research, it is expected that collinearity exists between time-related attributes, such as departure and arrival times, and other flight-related information. By assessing the strength of the relationship between attributes based on the available data, correlation analysis helps identify redundant attributes that can be derived from other attributes in the data set.

This step is essential for improving the efficiency and accuracy of subsequent analyses, particularly in the application of machine learning algorithms. For example, suppose two attributes are found to be highly correlated. In that case, it indicates that they provide similar information, and including both of them in the analysis might introduce redundant or duplicate information, leading to biased results or overfitting.

In this study, the Pearson product-moment correlation coefficient (r_p) and the Spearman rank correlation coefficient (r_s) were utilized to measure the associations between attributes in the flight data sets. The choice of coefficient depends on the distribution characteristics of the data, with r_p being suitable for light-tailed distributions and r_s being preferable for heavy-tailed distributions or data with the presence of outliers. Given that the attributes in the flight data sets used in this research exhibit a combination of light-tailed and heavy-tailed distributions, both coefficients were employed.

Furthermore, by considering the significance level of the correlation, it is determined which attributes exhibit strong correlations that need to be addressed. This step enables the identification and management of potential multicollinearity issues, ensuring the integrity of subsequent analyses.

2.6 Data Mining Algorithms

In this study, the primary focus is on predicting flight delays using data mining algorithms. The prediction task is approached through the application of classification algorithms, which aim to assign instances to specific classes [28]. In the context of flight delay prediction, the two example classes are "delay" and "no delay." The objective is to categorize each instance accurately based on whether a delay is expected or not.

On the other hand, regression algorithms are designed to predict the exact numerical outcome of each instance [45]. In the case of flight delay

prediction, regression algorithms would estimate the duration of the delay in minutes. However, for the purposes of this research, the analysis is limited to classification algorithms, as they are more suitable for the specific prediction task at hand.

This section provides a theoretical background on the selected classification algorithms, along with discussions on dealing with categorical variables and class imbalance. Additionally, a brief overview of transfer learning is presented, offering insights into its potential application in the flight delay prediction domain.

2.6.1 Encoding

In both classification and regression analysis, categorical variables are widely used. However, only numerical values are accepted by most machine learning algorithms. Therefore, categorical variables have to be encoded into numerical values, representing each category with a corresponding number, before feeding them into the ML algorithms [43].

There exist several encoding techniques that can be used for the transformation of a data set with categorical variables. The techniques applied in this research are ordinal encoding, one-hot encoding, and target encoding. An example representation of the difference between ordinal and one-hot encoding is given in Figure 2.4. These techniques have been chosen based on their distinct strengths and weaknesses.

Ordinal encoding is a technique that assigns an integer to each category in a categorical column. It does not add extra columns to the data but does create a hierarchy in the variable that may not actually exist [59].

On the other hand, one-hot encoding represents each value in a categorical variable of cardinality d as a d-dimensional vector split into columns [51]. Each element of the vector indicates the presence (1) or absence (0) of the binary variable. A downside of this technique is that the cardinality of the variables can be large. Due to the creation of extra columns for each category, storing the data after applying one-hot encoding can become a problem [51].

The last encoding technique being used in this research is target encoding. Contrary to one-hot encoding, it is designed for high-cardinality categorical attributes. In short, this technique replaces the original categorical values with a blend of the posterior probability of the target value given the categorical value and the prior probability of the target value over all the data [35].

Figure 2.4: Graphical representation of the difference between ordinal and one-hot encoding.

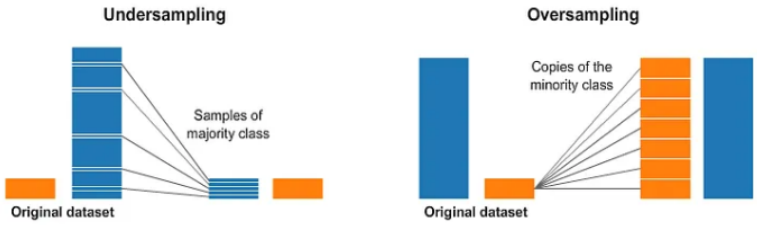


Figure 2.5: Illustration of undersampling (left) and oversampling (right).

2.6.2 Imbalance

Complications in machine learning arise when the data is imbalanced in the sense that one of the classes (e.g., the positive samples) is heavily underrepresented compared to the other class [29]. These complications include potential bias towards the majority class, therefore leading to poorer performance in the minority class. During the research, the issue of class imbalance came to light. To address this issue, both oversampling and undersampling techniques were employed.

Oversampling is the art of increasing the number of instances in the minority class by producing new instances or repeating certain instances. On the other hand, undersampling is the process of decreasing the number of instances in the majority class [36]. The difference between these two methods is illustrated in Figure 2.5.

2.6.3 Decision Tree

The decision tree algorithm, known for its ability to break down complex decision-making processes into simpler steps, is highly relevant to the flight delay prediction problem. It offers interpretability, allowing for a better understanding of the factors contributing to delays [49]. Decision trees ana-

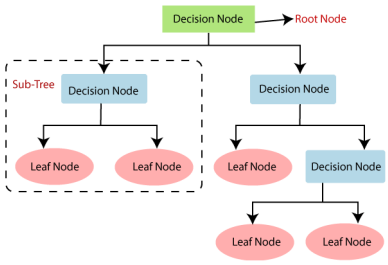


Figure 2.6: Diagram of a decision tree.

lyze flight attributes and provide a clear path to predict delays or non-delays.

However, they may be prone to overfitting in the presence of noisy or complex data, requiring careful consideration of data quality and pruning techniques [39]. The decision tree structure provides a visual representation of attribute importance, aiding in identifying influential factors affecting delays and enabling informed decision-making. An example structure of a decision tree is shown in Figure 2.6.

2.6.4 Random Forest

The random forest algorithm, which is an ensemble of multiple decision trees, is also highly relevant to the flight delay prediction problem. It combines the predictions of individual decision trees to make a final prediction by majority voting, providing a robust and accurate prediction mechanism [56]. The random forest's ability to handle complex relationships and capture the interactions between attributes makes it a powerful tool for identifying key factors contributing to flight delays.

Additionally, the random forest's ensemble approach helps mitigate overfitting and improves the overall predictive performance compared to a single decision tree. The information flow and composition of a random forest can be visualized in Figure 2.7, providing a clear understanding of its functioning in the context of flight delay prediction.

2.6.5 Gradient Boosting Tree

The gradient boosting tree, an ensemble of decision trees trained in sequence, plays a crucial role in the flight delay prediction problem. By iteratively fitting the negative gradients of the loss function, this algorithm creates a series of decision trees that collectively form a powerful predictive model [17].

The sequential nature of gradient boosting allows each tree to correct the errors made by previous trees, resulting in improved prediction accuracy.

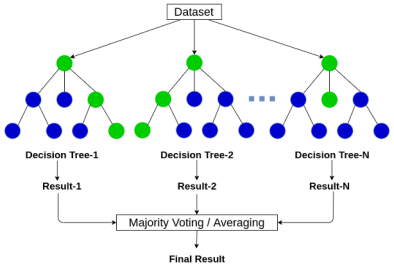


Figure 2.7: Information flow of a random forest.

One notable advantage of gradient boosting trees is their inherent capability to mitigate overfitting, ensuring reliable and robust predictions for flight delays.

2.6.6 K-Nearest Neighbor

The k-nearest neighbor (k-NN) classifier is a valuable algorithm employed in this study. By computing the distance between a test sample and the training samples, k-NN determines the class of the new sample based on the majority of labels from the k nearest training samples. Various distance metrics, including Euclidean and Manhattan distance, can be utilized. However, k-NN's sensitivity to feature scale necessitates data normalization to ensure optimal performance [42].

The choice of the value for k is a critical parameter in the k-NN classifier. It determines the number of neighbors considered for prediction. Selecting a smaller value results in a more localized model, while a larger value yields a more global model. The impact of different k values on the model's performance can be observed in Figure 2.8. Thus, in the flight delay prediction research, the appropriate selection of k plays a vital role in achieving accurate and reliable predictions based on the nearest neighbor relationships within the data.

2.6.7 Naive Bayes

The naive Bayes classifier is a frequently used algorithm in the flight delay prediction research. It determines the most likely class for a given instance's feature vector by assuming that all features are independent, which significantly speeds up the learning process. This can be expressed as

$$P(X|C) = \prod_{i=1}^{n} P(X_i|C) \tag{2.1}$$

where $X = (X_1, X_2, \ldots, X_n)$ represents the feature vector and C represents the class. Despite the unrealistic assumption of independence between vari-

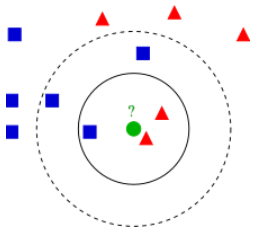


Figure 2.8: Example of k-NN classification where the green circle has to be classified according to its neighbors. It will be classified as a red triangle when a small number of neighbors is chosen (k=3, solid circle), whereas it will be classified as a blue square when a larger number of neighbors is chosen (k=5, dashed circle).

ables, empirical studies have demonstrated the effectiveness of the naive Bayes classifier in practical applications [46]. This algorithm's utilization enables the accurate classification of flight delay instances based on the independent probabilities of individual features.

2.6.8 Logistic Regression

The logistic regression classifier is a relevant algorithm in the context of flight delay prediction in this research. It aims to model the posterior probabilities of each class via linear functions in the input features while ensuring that they sum to one and remain in the range [0,1]. The classifier calculates the log-odds of the target variable being in the positive class. These log-odds are then transformed using the logistic function (also known as the sigmoid function[2]) to obtain the probabilities. Therefore, the model has the form

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{2.2}$$

where p is the probability of the positive class, $x_1, x_2, \ldots, x_n$ are the input features, and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients representing the impact of each feature on the log-odds [24]. The logistic regression classifier enables the estimation of the probability of flight delays based on the input features' coefficients and their corresponding weights. One limitation of logistic regression is its tendency to oversimplify complex relationships between variables [39].

2.6.9 Neural Network

A neural network, also known as an artificial neural network (ANN), is a machine learning algorithm inspired by the human brain. It mimics the way that biological neurons signal to one another. A neural network is built with

[2]The sigmoid function exhibits an S-shaped curve, see an in-depth explanation here.

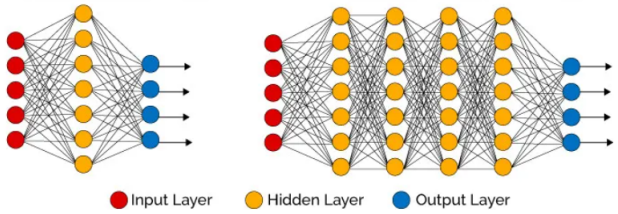


Figure 2.9: Examples illustrating the difference in the number of layers between an ANN (left) and a DNN (right).

interconnected neurons, consisting of an input layer, one or more hidden layers, and an output layer. Each neuron receives an input signal, applies a mathematical transformation, and produces an output signal. These signals are propagated through the network, where each neuron influences the final outcome based on its weights and activation function [6]. By employing neural networks, this research can effectively capture and learn the intricate relationships and patterns within the flight data to accurately predict flight delays.

It is noticeable that ANNs typically have only a few hidden layers, whereas deep neural networks (DNNs) have a larger number of hidden layers. This difference makes DNNs more suitable to deal with complex machine learning tasks [19]. Figure 2.9 provides a visual illustration of this difference.

2.6.10 Support Vector Machine

A support vector machine (SVM) aims to find the optimal hyperplane in a high-dimensional space that maximizes the margin between different classes, as shown in Figure 2.10. The hyperplane is constructed by taking the support vectors into account, which are a subset of training samples closest to the decision boundary [10].

One potential weakness of SVMs in this research is their sensitivity to the choice of kernel function and hyperparameters. The performance of an SVM model can heavily depend on these choices, and selecting the optimal combination may require careful experimentation and tuning. However, with proper parameter selection, SVMs can provide accurate and reliable predictions for flight delays.

2.6.11 Transfer Learning

The art of transfer learning is to leverage knowledge from one task to another related task. Consider an example of two people who want to become flight attendants. One of the two has no prior experience in aviation-related

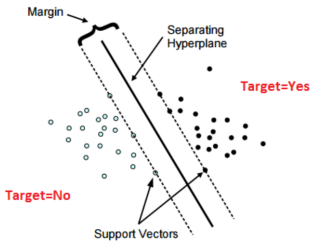


Figure 2.10: Illustration of support vector machine.

jobs, while the other person used to work as an information desk employee at an airport. The person with aviation knowledge can transfer and apply their background knowledge to the new job, giving them a head start, while the other person has to start completely from scratch.

A similar approach holds for the domain of machine learning, where a pre-trained model can reuse its knowledge for another related task. This method might result in higher performance and faster learning on the new task [61].

2.7 Data Analysis

Finally, after the creation of machine learning models, it is essential to evaluate their performance and effectiveness [24]. This can be done by utilizing different performance measures to assess the accuracy, precision, AUC score, and other metrics of the model. Using these measures, different machine learning algorithms and techniques can be compared.

Additionally, other techniques such as feature importance analysis, confusion matrices, and ROC/precision-recall curves are employed to gain deeper insights into each model's performance and reliability. This section provides a widespread explanation for each of these evaluation metrics.

2.7.1 Performance Measures

Performance measures such as accuracy, precision, recall, F_1 score, and AUC score are essential in this research as they allow for a comprehensive evaluation of the machine learning models used for flight delay prediction. These performance measures help in assessing the strengths and weaknesses of different machine learning algorithms and determining which ones are best suited for accurately predicting flight delays.

The accuracy of a model measures the overall correctness of its predictions. Accuracy is particularly useful in this research since it represents the overall

correctness of the model's predictions, It is calculated by dividing the number of correctly classified instances by the total number of instances, that is

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

where

- True Positive (TP) represents the number of instances that are correctly classified as positive.
- True Negative (TN) represents the number of instances that are correctly classified as negative.
- False Positive (FP) represents the number of instances that are incorrectly classified as positive.
- False Negative (FN) represents the number of instances that are incorrectly classified as negative.

Another important evaluation metric in this study is the precision of the classifier, which stands for the accuracy of the positive predictions. It is usually used along with the recall, which measures the ratio of positive instances that are correctly detected by the classifier [18]. The precision and recall are calculated as

$$precision = \frac{TP}{TP + FP} \tag{2.4}$$

$$recall = \frac{TP}{TP + FN} \tag{2.5}$$

Precision and recall can be combined into a single performance metric called the F_1 score, which provides a balanced evaluation of the model's performance. This metric is particularly useful in this paper due to its ability to compare two classifiers with one metric and to deal with imbalanced data sets. The F_1 score is the harmonic mean of precision and recall, being high only when both measures are high [18]. It is obtained by

$$F_1\ score = 2 \times \frac{precision \times recall}{precision + recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{2.6}$$

The F_1 score favours classifiers that have similar precision and recall. However, in practice, you might have a situation where you mostly care about either precision or recall. Therefore, based on the context, a decision has to be made according to the precision/recall trade-off.

For example, when a security system has to identify trespassers, it is favourable to have a classifier that has a few false accusations (low precision) but makes sure that almost all trespassers will get caught (high recall). On the contrary, when a classifier has to detect damaged products in a factory process,

it is preferable to reject many undamaged products (low recall) while keeping only the best products (high precision). In a situation where a balance between precision and recall is favoured, like in flight delay prediction, the F_1 score is used.

The Area Under the ROC Curve (AUC) score is a performance measure that is also highly relevant in this research. It measures the model's ability to distinguish between positive and negative instances. A perfect classifier will have an AUC score equal to 1, whereas a purely random classifier will have an AUC score equal to 0.5 [18].

2.7.2 Feature Importance

In this research, feature importance analysis plays a critical role as it helps to identify the most influential features in the predictive model for flight delay prediction. By understanding which features have the most significant impact on the model's decisions, valuable insights can be gained to achieve the research goals effectively.

Various methods have been utilized to retrieve feature importances, such as coefficient analysis in linear models, information gain in tree-based algorithms, and permutation importance[3] in support vector machines [2]. The resulting feature importances can be used to perform feature selection: a method where the less important features are discarded in order to gain a performance increase and speedup.

2.7.3 Confusion Matrix

A confusion matrix is utilized in this research because it serves as an essential evaluation metric for assessing the performance of the classification model in flight delay prediction. It shows the number of TN, FP, FN, and TP as can be seen in Table 2.2. Each row represents an actual class, while each column represents a predicted class.

The matrix can be used to gain insights into the errors made by the classifier, such as misclassifying negative instances as positive (FP) or not identifying positive instances (FN) [18]. With the values from the confusion matrix, one can calculate the accuracy, precision, recall, and other metrics as described in Section 2.7.1.

[3]Method that measures the effect of different features on the model's performance by randomly removing features and measuring the change in model accuracy.

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

Table 2.2: Confusion matrix showing the position of true negatives (top left), false positives (top right), false negatives (bottom left), and true positives (bottom right).

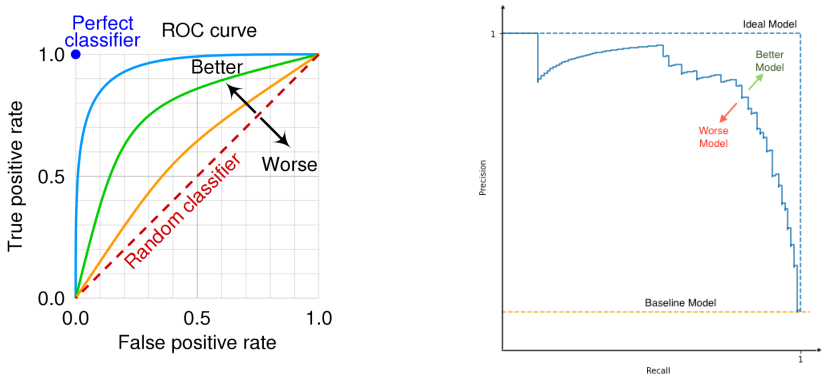


Figure 2.11: Graphical representation of ROC curves (left) and precision/recall curves (right).

2.7.4 Evaluation Curves

The Receiver Operating Characteristic (ROC) curve and precision/recall curve are utilized in this research to assess the performance of each classifier in terms of its sensitivity, specificity, precision, and recall. These curves offer additional diagnostic information beyond the traditional performance measures, allowing for a more detailed evaluation of the classifier's predictive capabilities in the context of flight delay prediction.

The ROC curve plots the true positive rate (TPR/recall) against the false positive rate (FPR/fall-out). The TPR is calculated using Equation 2.5. Similarly, the FPR is calculated as the ratio of negative instances that are incorrectly classified as positive, that is

$$FPR = \frac{FP}{FP + TN} \tag{2.7}$$

The curve provides a visual depiction of the trade-off: the higher the recall (TPR), the more false positives (FPR) the classifier produces. The dotted line in Figure 2.11 represents the ROC curve of a completely random classifier. In the same plot, a classifier with a ROC curve closer to the top-left

corner indicates higher performance [18]. The area under the ROC curve is called the AUC score, as described in Section 2.7.1.

On the other hand, the precision/recall curve shows the trade-off between precision and recall for different thresholds. A larger area under the curve stands for high recall and high precision, as can be seen in Figure 2.11. The precision/recall curve allows for the selection of a threshold that balances precision and recall according to the specific needs of the problem [18].

Chapter 3

Related Work

The prediction and prevention of aircraft postponements in the aviation industry is a major issue due to the inconvenience that flight delays cause to airlines, airports, and passengers. In recent decades, numerous studies have investigated flight data sets using machine learning and data mining approaches. These studies vary their focus from airport capacity to taxi-out time, including the focus of this paper: flight delay. This chapter describes the state-of-the-art research done for flight data sets and, in particular, predicting flight delays by using different strategies.

3.1 Global Flight Delay Data Sets

Several data sets have been used to conduct research on flight data pertaining to aircraft postponements. The data sets that are used for data mining algorithms to predict aircraft delays vary a lot. The flight data sets mainly differ in the region of the world they cover. Some researchers focused on a single airport or route, while others examined the big picture for an entire country or continent.

3.1.1 United States

The most common research on flight delays has been done using flight data from the United States (US). The data is widely used due to its public availability brought by the United States Department of Transportation [1], primarily through the Federal Aviation Administration [2] and the Bureau of Transportation Statistics databases [3] [54].

The majority of aircraft postponement research with US flight data sets has

[1]DOT. United States Department of Transportation, `https://www.transportation.gov/`

[2]FAA. Federal Aviation Administration, `https://www.faa.gov/`

[3]BTS. Bureau of Transportation Statistics, `https://www.bts.gov/`

covered the entire country to get an overall flight delay prediction for the United States [5, 31, 58, 64]. Their main findings include high performances of the random forest classifier, a strong correlation between month, travel day, and departure delay, and the months with the least amount of flight delays to be September and October. Other researchers decided to look at a smaller scale in the US by selecting one or more airports [13, 27, 52]. An airport that is frequently looked at is the John F. Kennedy International Airport (JFK) in New York.

3.1.2 China

The Chinese airspace has also proven to be an attractive region for predicting aircraft delays. Researchers have used different sources to gather useful data. These sources include historical flight data obtained through websites [11, 32], airports [48, 63], the Civil Aviation Administration of China[4] [65], and self-gathered data [21].

Predicting flight postponements using the Chinese airspace in its entirety is a common strategy, with several researchers having successfully applied this method [11, 21, 32]. The main results of their research showed the effectiveness of the long short-term memory method on Chinese flight data. However, predicting flight delays in China is often also done by looking at a single airport in the country [48, 63, 65].

3.1.3 Brazil

Research in flight delay prediction has also been done in South America, mainly in Brazil. One of the most widely used data sources is the Brazilian National Civil Aviation Agency[5] [8]. They offer the Active Regular Flight Database (VRA)[6] as a public flight data set.

Despite the high availability of reliable flight data in Brazil through the VRA, not much research on flight delays has been done in this region. Only a few researchers examined the historical flight data of the country, most of them looking at the complete country [37, 53], with some exceptions [40].

In their analysis, Sternberg et al. [53] identified meteorological conditions as the most influential factors contributing to flight delays. They also observed that flight delays were more prevalent during vacation months and on Fridays. Additionally, they found that departures in the early or mid-morning

[4]CAAC. Civil Aviation Administration of China, `http://www.caac.gov.cn/`

[5]ANAC. National Civil Aviation Agency of Brazil, `https://www.gov.br/anac/`

[6]VRA. Active Regular Flight Database, `https://sas.anac.gov.br/sas/bav/view/frmConsultaVRA`

had a lower likelihood of being delayed. Notably, their study revealed that Brazil experienced delay propagation, with late evening and night being the most critical periods associated with significant flight delays.

3.1.4 Europe

Eurocontrol[7] provides a database of European historical flight data, made available by an intergovernmental organization in Europe [54]. Not a lot of researchers have investigated this data set, but some have managed to perform small-scale flight delay predictions on data immediately gathered from their target airport [47, 67].

3.1.5 Weather

Besides historical flight data, other typical data sources to predict aircraft postponements include weather databases [8]. Commonly used databases for weather factors are provided by the National Oceanic and Atmospheric Administration[8], the United States Department of Transportation[9], and The Weather Company[10] [8, 54].

3.2 Machine Learning Algorithms

The main objective of this research is to address the problem of flight delay prediction using machine learning techniques. While other categories such as probabilistic models and statistical analysis are mentioned to provide a broader understanding, the focus remains on classification algorithms. Regression, as a common type of probabilistic model, also plays a significant role in predicting flight delays.

3.2.1 Classification

As explained in Chapter 2, classification is the art of putting all instances of a data set in a certain class. Further, the output of the instances in classification admits only discrete, unordered values [28]. In the flight delay prediction problem, classification algorithms are often used to classify each individual flight into a class. The possible classes are frequently "delay" and "no delay," but several authors created multiple classes, for example: "no delay," "1-15 minutes delay," "16-30 minutes delay," etc.

[7]EUROCONTROL. European Organisation for the Safety of Air Navigation, `https://www.eurocontrol.int/`

[8]NOOA. National Oceanic and Atmospheric Administration of the United States, `https://www.noaa.gov/`

[9]DOT. United States Department of Transportation, `https://www.transportation.gov/`

[10]TWC. The Weather Company, `https://www.ibm.com/weather`

Venkatesh et al. [58] compared Artificial Neural Networks (ANN) and Deep Belief Networks (DBN) and found that ANN performed the best for binary classification of aircraft delay. In another study [40], ANN was used in combination with the Random Search technique. This resulted in a correct predictive capacity of 90% for their data set. Moreira et al. [37] compared a broader scale of classifiers: Neural Networks (NN), Random Forests (RF), Support Vector Machines (SVM), Naive-Bayes (NB), and k-Nearest Neighbors (k-NN). Although their best-performing classifier was NN, it achieved slightly lower performance compared to other researchers, which can be attributed to differences in data regions and timeframes. However, the largest contribution of their work was the comparison of data preprocessing methods, especially focusing on the unbalanced distribution of the classes of delay. Their results indicated the need for balanced training data when predicting flight postponements.

Instead of using binary classification to predict flight postponements, multiclass classification can be used to get more accurate predictions. Esmaeilzadeh et al. [13] analyzed factors contributing to flight delay by applying multiclass SVM. They found the departure demand-capacity level, weather activity, and TMIs to be the main factors with the highest impacts on departure delay. Other researchers [21] have applied RF and long short-term memory (LSTM) to a self-gathered aviation data set. Their findings show that RFs performed better on their data set.

Besides the aforementioned classification algorithms, researchers often use unfamiliar algorithms or modify existing ones. For example, Yazdi et al. [64] created the Stack Denoising Autoencoder-Levenberg Marquart method, which achieved an extremely high accuracy of 96% on their balanced flight data set.

In conclusion, previous studies have consistently shown that neural network and random forest classifiers outperform other methods in accurately classifying flight delays. Additionally, weather conditions have been identified as a significant factor influencing delays. Given these findings, it becomes intriguing to investigate flight delay prediction without considering meteorological data in this study. By examining the impact of non-weather-related factors, we can gain a deeper understanding of their relevance and potential contribution to accurate delay prediction.

3.2.2 Regression

Regression is a similar prediction strategy to classification. However, the goal here is not to put every instance in a certain class but to predict the

exact numerical outcome of each instance [45]. The strategy can be used for problems including the prediction of exam scores, forecasting in sales, and flight delay prediction, among others.

In their paper, Shao et al. [52] compared Linear Regression (LR), Support Vector Regressor (SVR), Multilayer Perceptron (MLP), and LightGBM as regressors on historical flight data. They show that LightGBM outperforms the other regressors and that the airport situational awareness map plays a more important role in departure delay time prediction than weather conditions. Ye et al. [65] confirmed the effectiveness of LightGBM with high performances on their particular data set. The SVR has also been used in research that created a prediction model for Beijing Capital International Airport, which proved to be effective and accurate [63]. Other forms of regressors are the Random Forest Regressor (RFR) and Mixture Density Networks (MDN), which have been compared and proved to be successful in predicting future flight delays at Rotterdam The Hague Airport [67].

Like classification algorithms, researchers also tend to create regression algorithms that are modifications of existing ones. Khanmohammadi et al. [27] designed an ANN for defect of module prediction, DMP-ANN, which is suitable for the prediction of defects such as delays in operations.

From the findings of these previous studies, it becomes evident that no single regressor stands out as dominant in accurately predicting the exact amount of flight delays. This observation highlights the inherent challenge in accurately forecasting the precise duration of flight delays.

3.2.3 Combination

There is a huge space of machine learning algorithms that can be used for both classification and regression. Therefore, researchers often use a combination of classification and regression algorithms in their work.

Li et al. [31] have used RF for both classification and regression, which resulted in a validation of their hypotheses. Other researchers have used the classification-regression combination to compare a new model with NB and the C4.5 approach [11]. Lastly, in recent research, Li et al. [32] created an impressive ST-Random Forest algorithm, based on spatial features and temporal properties. The model achieved a high AUC score, implying a high prediction performance.

3.2.4 Other Techniques

Besides classification and regression, other methods for the flight delay prediction problem are frequently used. Consider Baluch et al. [5], who used a combination of the Decision Tree (DT) algorithm and a clustering algorithm to determine the best day and month to travel on. A completely different approach was taken by Sternberg et al. [53]. They focused on pattern mining through the use of association rules learning, which showed that flight delays in Brazil are mostly caused by adverse meteorological conditions.

Statistical analysis is also regularly used to predict aircraft postponements. This method analyzes the data to find patterns and trends. A commonly applied method of statistical analysis is to use Bayesian Networks (BN). Rodríguez-Sanz et al. [47] created a tool for AMAN systems[11] by laying out a causal model based on a BN approach. This model can be applied in a forward analysis or in an inverse analysis. Rong et al. [48] have also utilized BN, but in combination with the Gaussian mixture model - expectation maximization algorithm. Their model is different from other works in the sense that it is based on random flight points.

3.3 Other Flight Data Research

Nowadays, flight data has a plethora of applications in the context of data mining, where it is often considered a valuable resource [66]. Combining aircraft data with data mining gives rise to great opportunities as well as multiple challenges. Different areas of flight data research, including ground delay [30], airplane control [25], airport capacity [60], and taxi-out time [3, 4], provide valuable insights into the factors that can contribute to flight delays. Ground delay research examines the causes and effects of delays that occur while the aircraft is on the ground, while airplane control research focuses on optimizing flight routes and navigation systems. Airport capacity research explores limitations and identifies potential bottlenecks, while taxi-out time research investigates factors that can impact the time it takes for an aircraft to taxi for takeoff.

One specific area in the research of flight data is airline passenger satisfaction. Research has found that the most influential factors for passenger satisfaction on an airline are online boarding, type of travel, and inflight wi-fi services [39]. It also revealed that flight delays are among the top five most influential factors influencing passenger satisfaction in the United States, excluding non-service factors such as type of travel and customer type. The paper mentioned above obtained these results by using a large number of

[11] Arrival Manager planning systems

ML algorithms to find the best-performing classifier, a strategy that is also implemented in this paper. Their findings were confirmed by research that found approximately the same influential factors for passenger satisfaction, but then by only considering American Airlines [26].

Chapter 4

Methodology

This chapter outlines the methodology employed to predict flight delays using multiple classification algorithms. It covers the key steps involved in the research, including data collection, preprocessing, exploration, model implementation, and analysis. The focus is on presenting the techniques utilized to achieve accurate and reliable predictions.

4.1 Data Collection

Based on all the data sources for flight postponements that have been used by inspiring previous researchers, this study uses data sets that have their origin in the United States and Brazil. The historical flight data for China has not been utilized due to limited accessibility. Furthermore, European historical flight data has been excluded since the accessible data sources only contain aggregated data.

The data sets for the United States and Brazil were obtained through reliable sources, specifically the Bureau of Transportation Statistics (BTS)[1] and the Active Regular Flight Database (VRA)[2]. Flight records in the time period from January 1st 2022 up to and including December 31st 2022 are considered, to ensure a wide variety of flights and situations. Restricting the analysis to one year of historical flight data accounts for the evolving circumstances in the aviation industry, including regulatory changes and COVID-19 measures.

Both data sets contain the response variable of a flight delay, measured in minutes. However, for the purpose of this research, the focus is on binary classification. In this context, a flight is classified as delayed if its delay is at

[1] BTS. Bureau of Transportation Statistics, `https://www.bts.gov/`

[2] VRA. Active Regular Flight Database, `https://sas.anac.gov.br/sas/bav/view/frmConsultaVRA`

least 15 minutes, otherwise, it is classified as non-delayed (see Section 2.1.2 for more details). This method is chosen to facilitate a better comparison of different classifiers. By framing the problem as a binary classification task, it simplifies the problem and allows for better precision in identifying the delay status, providing valuable insights for airlines and passengers.

4.1.1 United States Data Set

The airline on-time performance data from the United States of America are reported by US-certified air carriers and published through the BTS. It is noteworthy that carriers are required to report on-time data for domestic flights they operate, which also contributes to the trustworthiness and accuracy of the data.

The data set comprises individual flight records encompassing a comprehensive range of 109 fields, including information about the time period, airline, origin, destination, departure performance, arrival performance, cancellations/diversions, flight summaries, cause of delay, gate return information, and diverted airport information. A selection of these fields has been chosen in accordance with the research objectives[3], which leaves the data set with approximately 7 million records and 25 features.

4.1.2 Brazilian Data set

In contrast to the extensive research on flight delays in other regions, there has been relatively limited investigation into flight delays in Brazil. Recognizing this research gap, this study obtained the second data set from the VRA, a governmental institution in Brazil. The VRA collects and provides flight data, which is solely the responsibility of the airlines operating in the country.

The Brazilian data set contains roughly 900.000 flight records from 2022 and 20 features including the origin, destination, and time period. No features were removed before preprocessing the data as no features seemed to be redundant at first glance.

4.1.3 Comparison

It is worth noting that the two data sets share similarities and differences. Both data sets measure time period attributes in the same units, allowing for potential transfer learning. The origin/destination airport attributes

[3]Attributes that lack useful information for this research (e.g., gate return information and diverted airport information) are not utilized. This also holds for fields that can be derived from other variables and fields that are unknown prior to departure.

also align, although the chances of a significant overlap are small due to the different regions covered. However, there are notable differences between the data sets. The USA data set includes additional attributes such as distance, tail number, and city markets, while the Brazilian data set contains attributes like model equipment, number of seats, and line type code.

From previous studies discussed in Chapter 3, it is evident that time period attributes are suggested to be highly relevant in flight delay prediction. On the other hand, attributes like flight number and airline codes may be less informative for predicting flight delays, as they are treated as IDs and have differences in coding conventions between the data sets. However, if it is found that these factors significantly influence the classification, careful considerations will be made.

4.2 Data Preprocessing

To prepare the US and Brazilian data sets for the machine learning phase, a series of data preprocessing steps are performed, as outlined in Section 2.4. The first step in data preparation for both data sets is to remove the flight records that were either cancelled, diverted or uninformed. This has been done because the aim of this research is to predict flight delays rather than cancellations or diversions.

Next, to handle missing values and columns without entries, cleaning techniques were applied. Since the portion of flight records with missing values was relatively small ($< 3\%$), these records were removed. The variables that contained missing values in the data sets were the scheduled/actual departure/arrival times, which are crucial for determining whether a flight was delayed. These variables also play a key role in creating the response variable that classifies flights as delayed or non-delayed. Other techniques could have been used as well in order to ensure complete information. Furthermore, empty features were removed since they have no use. Finally, variables that are not known prior to the departure of a flight were identified and subsequently removed from the data sets, as they are not relevant for predicting flight delays.

4.3 Data Exploration

Various visualization methods are utilized to enhance the interpretability of the data. Histograms and boxplots are often used to illustrate the distributions of individual variables, while the correlation plot provides insights into their impact on the response variable.

Each data set contains both numerical and categorical attributes. The descriptive statistics for the numerical attributes in the American and Brazilian data set are shown in Table 4.1 and Table 4.2 respectively. Note that the delays (departure delay, arrival delay, and overall delay) are binary, representing whether the delay was at least 15 minutes (1) or not (0). Furthermore, the average and standard deviation are not calculated for variables that are IDs or where the range of values is meaningless.

Upon examining the numerical attributes in both data sets by using the aforementioned tables, some values stand out. In the US data set, it is noticeable that the number of flights attribute is essentially redundant as every record has a value of 1. As a result, this feature is removed during the data exploration phase. Furthermore, the elapsed time attribute has entries that are below 0 minutes, which is logically impossible. On the other hand, the numerical attributes in the Brazilian data set exhibit no noticeable anomalies in the table.

Similar to the numerical attributes, the descriptive statistics for the categorical attributes in both data sets are shown in Table 4.3 and Table 4.4. Note that the attributes starting status and arrival status are equivalent to the departure and arrival delay from the previous table. Additionally, the code DI attribute stands for the type of authorization for each flight stage, while the line type code represents the type of operation performed in the flight[4]. In these tables, none of the attributes have notable anomalies.

4.3.1 Airport Locations

The main difference between flights is their origin and destination airport. A world map has been created to get a clear overview of the locations of airports represented in the data sets. Figure 4.1 shows all the airports that have been used either as origin or destination airport in the data, for the USA and Brazil separately.

The world maps provide a visual representation of where the airports are geographically distributed. After observing the world map representing the United States data set, it becomes apparent that all origin and destination airports are situated within the borders of the United States. This observation confirms that the data set only comprises domestic flight records.

Although the world map might indicate possible outliers for flight records in the USA, it is important to note that the airports situated outside the continental United States are still considered part of US territory. These

[4]For more information, visit ANAC's website with a description for each variable here.

Numerical attribute	min	average	max	SD
Flight number	1	-	9562	-
Origin airport ID	10135	-	16869	-
Origin city market ID	30070	-	36101	-
Destination airport ID	10135	-	16869	-
Destination city market ID	30070	-	36101	-
Departure delay	0	0.21	1	0.4
Arrival delay	0	0.2	1	0.4
Elapsed time (minutes)	-85	143	690	72.58
Number of flights	1	1	1	0
Distance (miles)	31	816.96	5095	597.19
Delayed	0	0.25	1	0.43
Departure month	1	6.58	12	3.39
Departure week	1	26.74	52	14.76
Departure day	1	15.73	31	8.76
Departure hour	0	13.02	23	4.88
Departure minute	0	26.95	59	18.2
Departure day of week	0	2.98	6	2
Arrival hour	0	14.6	23	5.1
Arrival minute	0	29.22	59	17.8

Table 4.1: Descriptive statistic of numerical attributes for the US data set.

Numerical attribute	min	average	max	SD
Number of seats	0	157.61	515	65.91
Delayed	0	0.18	1	0.39
Departure month	1	6.72	12	3.45
Departure week	1	27.39	52	15.06
Departure day	1	15.75	31	8.79
Departure hour	0	12.99	23	5.75
Departure minute	0	26.35	59	17.59
Departure day of week	0	2.92	6	1.98
Arrival hour	0	13.07	23	6.01
Arrival minute	0	27.12	59	17.4

Table 4.2: Descriptive statistic of numerical attributes for the Brazilian data set.

Categorical attribute	unique values	most occurring + count
Airline carrier	17	WN (1261865)
Tail number	5872	N475HA (3033)

Table 4.3: Descriptive statistic of categorical attributes for the US data set.

Categorical attribute	unique values	most occurring + count
Airline carrier	100	AZU (275721)
Flight number	5399	702 (981)
Code DI	8	0 (796480)
Line type code	5	N (695924)
Model equipment	57	A320 (128316)
Origin airport	301	SBGR (111062)
Destination airport	306	SBGR (111208)
Starting status	2	0 (685284)
Arrival status	2	0 (683405)

Table 4.4: Descriptive statistic of categorical attributes for the Brazilian data set.

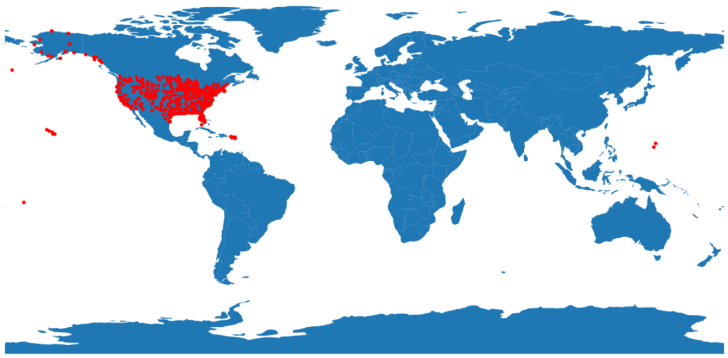
locations include states such as Alaska and Hawaii, as well as territories like Puerto Rico, the U.S. Virgin Islands, Guam, the Northern Mariana Islands, and American Samoa[5].

The world map visualization of the Brazilian data set highlights the global distribution of airports, with a concentration in Brazil. This observation aligns with the data set, as each flight record has either its origin or destination airport located in Brazil. The presence of airports from various countries on the world map indicates that Brazilian air travel has numerous flight connections with other parts of the world.

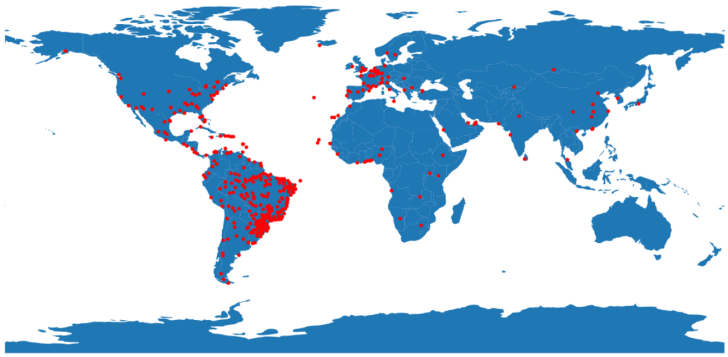
4.3.2 Distributions

Visualizing variable distributions is crucial for gaining insights into their central tendencies and identifying potential outliers. Histograms serve as effective tools for illustrating these distributions. For the USA data set, the histograms show that most variables have a relatively normal distribution, while certain variables show a significant influence from specific values, which heavily impact the overall distribution. Two of these histograms are shown in Figure 4.2.

[5]More information about the US territory can be found here.



(a)



(b)

Figure 4.1: World map that illustrates the locations of airports used as origin or destination in the data sets for the USA (a) and Brazil (b).

Analyzing the figure, it becomes challenging to derive the most occurring values for the tail number due to a large number of unique values. On the other hand, it is evident that the number of flights is considerably lower during nighttime compared to daytime.

Similarly, the histograms for two variables from the Brazilian data set are plotted in Figure 4.3, mirroring the analyses observed in Figure 4.2. The complete set of histograms depicting the distributions of variables in both data sets can be found in Appendix A, Section A.1.

In order to address the challenge of retrieving values from the histogram peaks, new histograms have been developed that display the top 10 most frequent values for each variable. These additional histograms can also be found in Section A.1 of Appendix A.

However, the aim of this research is to predict flight delays. Therefore, the

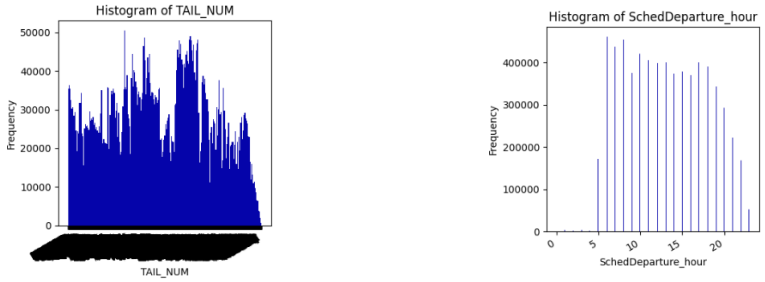


Figure 4.2: Histograms that show the distribution of two variables in the USA data set.

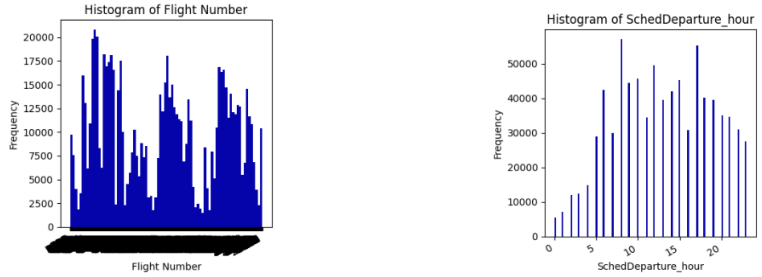


Figure 4.3: Histograms that show the distribution of two variables in the Brazilian data set.

proportion of delayed flights amongst different attribute values is important. Figure 4.4 and Figure 4.5 present the proportion of delayed instances among the top 10 most recurring values per variable for both data sets.

Upon examining Figure 4.4, several observations can be made. Note that the red part of the bars represents delayed flights, while the green part represents non-delayed flights. Southwest Airlines (WN) operates the most domestic flights within the United States, closely followed by American Airlines (AA) and Delta Air Lines (DL). However, it is noteworthy that Southwest Airlines has a relatively high delay rate of 44%. In contrast, the other major airlines show lower delay rates, with the exception of JetBlue Airways (B6), which experiences a 41% delay rate.

Other remarkable observations include the prominence of Denver International (11292), Dallas/Fort Worth International (11298), and Hartsfield-Jackson Atlanta International (10397) as the most frequently used airports. Moreover, New York City (31703) stands out as the city market with the highest number of operating flights. However, it also experiences a significantly higher delay rate compared to other frequent city markets.

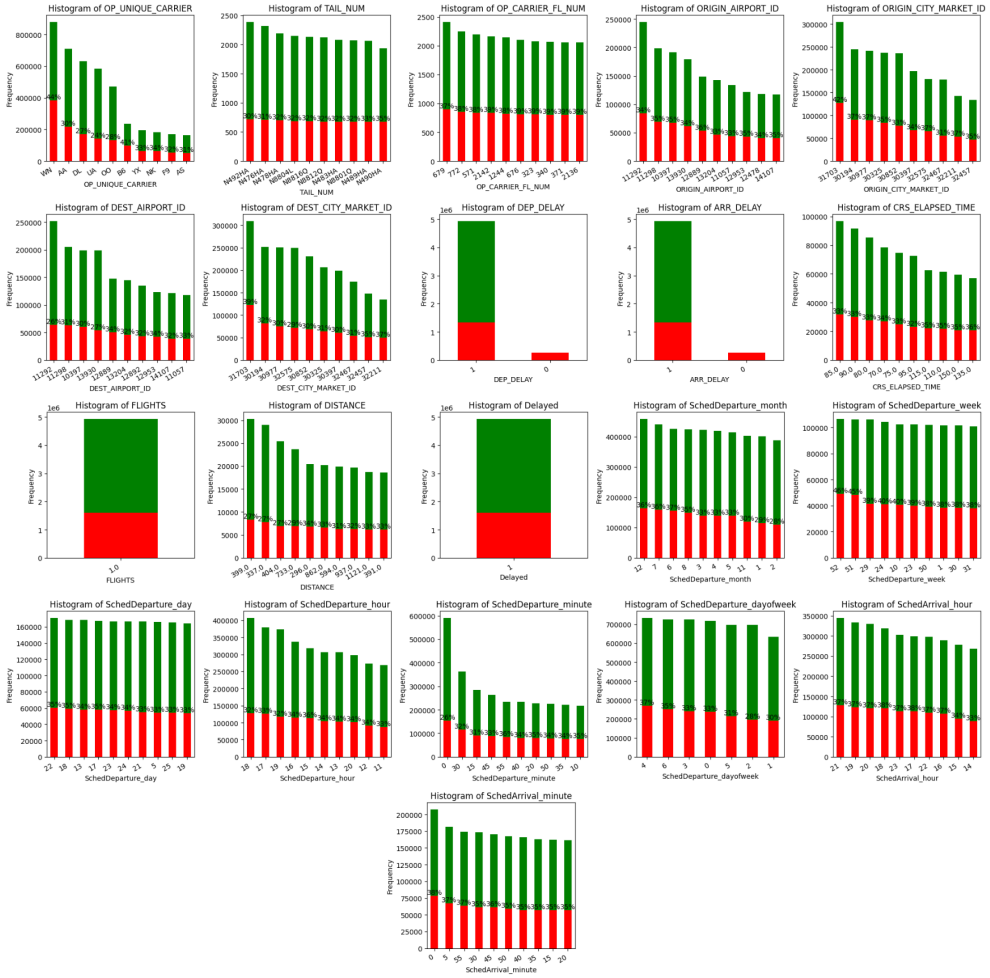


Figure 4.4: Histograms that show the proportion of delayed flights in the top 10 most occurring values of each variable in the USA data set.

Regarding the flight period, it is evident that flights during the final two weeks of the year, encompassing the Christmas and New Year holidays, exhibit the highest delay rates. Additionally, there is a noticeable preference for flights departing and arriving at rounded hours such as 8:00 AM or 9:00 AM, which occur more frequently compared to flights scheduled at unconventional times like 4:15 AM or 6:15 AM.

On the other hand, Figure 4.5 depicts the proportion of delayed flights in the top 10 most occurring values of each variable in the Brazilian data set. Gol Intelligent Airlines S.A. (GLO), Azul Brazilian Airlines (AZU), and LATAM Airlines Brasil (TAM) dominate the Brazilian airspace, with a significant lead in terms of the number of flights operated. Furthermore,

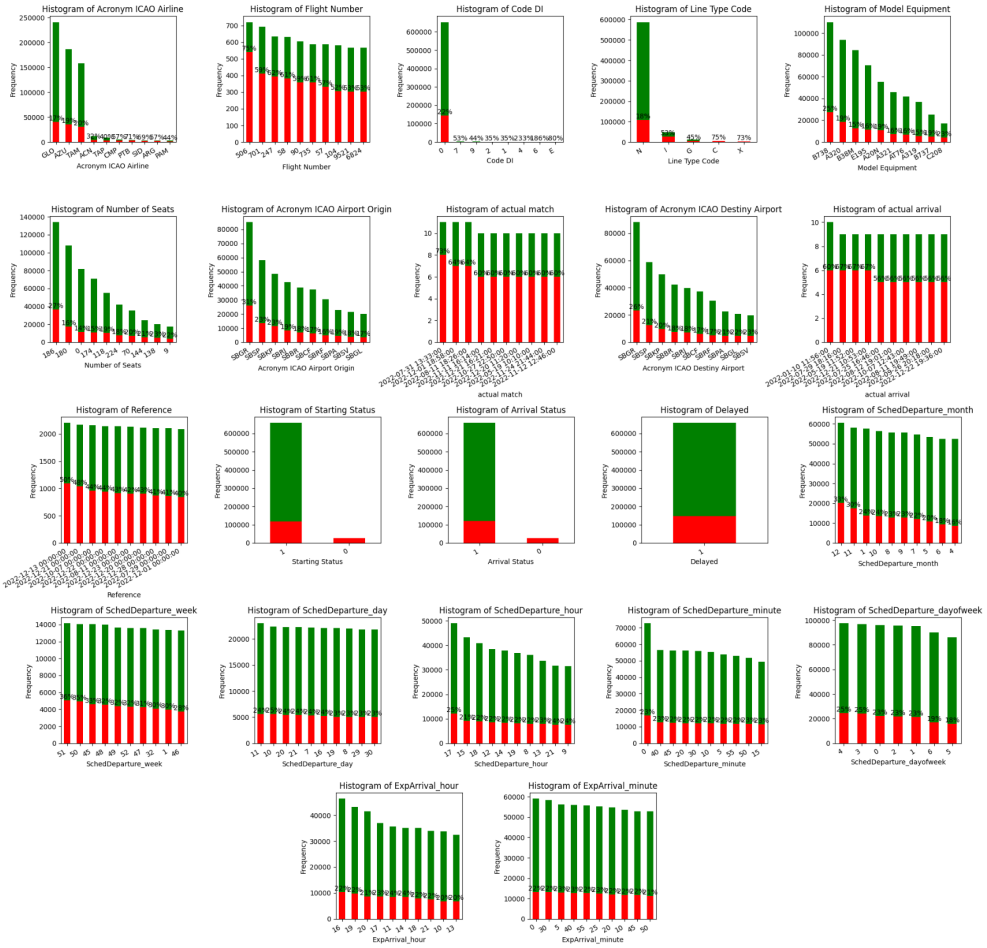


Figure 4.5: Histograms that show the proportion of delayed flights in the top 10 most occurring values of each variable in the Brazilian data set.

one notable flight plan stands out, which is the route from Frankfurt to São Paulo Guarulhos, operated by Lufthansa (506). Surprisingly, this frequently travelled flight experiences a delay in 75% of its occurrences, which is significantly higher compared to other commonly travelled routes.

In addition to the aforementioned observations, it was found that the code DI of 0 dominates over other values. The line type code N is the most frequently occurring and has the lowest delay rate. Moreover, it is common for flights to utilize the B738 plane model or a plane with 186 seats, both of which exhibit a high delay rate. Among the airports in Brazil, São Paulo/Guarulhos International Airport (SBGR) emerges as the most frequently utilized airport in 2022. However, it also exhibits the highest delay rate compared to other commonly used airports. Lastly, the flight period

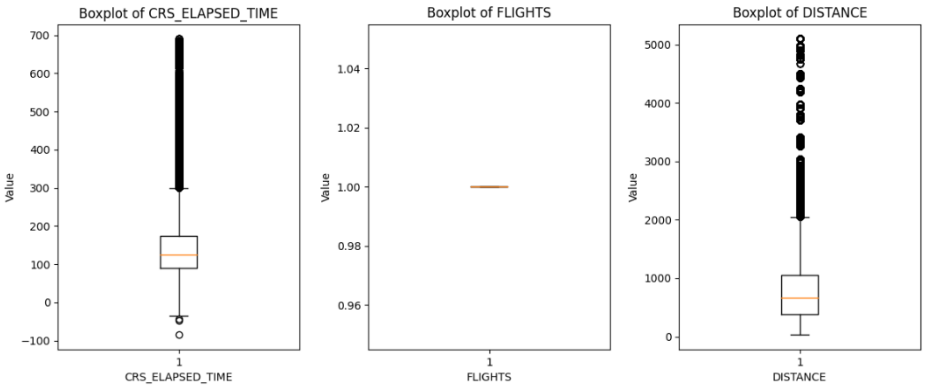


Figure 4.6: Boxplots showing the distributions of selected numerical attributes in the USA data set.

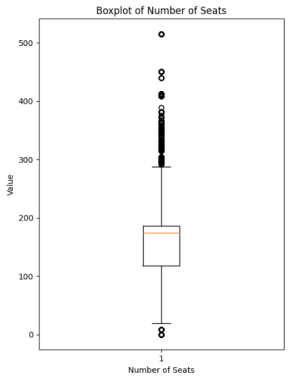


Figure 4.7: Boxplot showing the distributions of a selected numerical attribute in the Brazilian data set.

delay rates for flights in Brazil demonstrate a similar pattern to those observed in the USA, indicating consistency in delay patterns across regions in terms of the time period.

4.3.3 Outlier Detection

To identify potential outliers in the data sets, boxplots have been created for the numerical attributes in both the USA and Brazilian data sets, which can be observed in Figure 4.6 and Figure 4.7 respectively. The points labeled as "outliers" in a boxplot indicate potential outliers, but further analysis is needed to confirm their status as actual outliers. Note that the boxplots for numerical attributes related to the time period can be found in Section A.2 of Appendix A, as they do not contain any outliers.

Several potential outliers are revealed in Figure 4.6, which introduces the

need for further investigation into these points of the USA data set. It is noticeable that the flight number, origin/destination airport ID, and origin/destination city market ID are not presented in boxplots. This is based on the fact that they are numerical attributes but should be treated as categorical variables, as they represent distinct categories or identifiers.

The boxplot for the number of flights attribute reveals that it contains a single unique value, which again emphasizes the need to exclude this variable from the analysis. Furthermore, the elapsed time and distance attributes also exhibit potential outliers in their boxplots. However, the elapsed time attribute has entries that are below 0 minutes, which is logically impossible. As a result, these entries are removed from the data set. The other outliers in the elapsed time variable have values exceeding 300 minutes, but upon further investigation do not appear to be outliers after all. For example, the long-haul flight between Boston and Honolulu typically takes around 10 hours and covers a distance of more than 5000 miles. Similarly, this example also eliminates the potential removal of outliers in the distance attribute.

The Brazilian data set only contains a single numerical attribute with potential outliers, depicted in a boxplot in Figure 4.7. The potential outlier values are present in the number of seats attribute, with values ranging from 0 to approximately 500 seats. These values are, however, not true outliers. The data includes cargo planes, which may have zero passenger seats, as well as planes like the A388, which can accommodate up to 516 seats. Thus, these variations in the number of seats should not be considered as outliers.

4.3.4 Delay Proportions and Trends

A comprehensive understanding of the factors that significantly impact delay rates can be obtained by examining the delay proportions associated with each value within an attribute. This provides a clear overview of which values have the highest and lowest delay rates. The top 10 values with the highest and lowest proportional delays are presented in a histogram in Appendix A, Section A.3. Additionally, the most substantial attributes and their top three values with the highest and lowest delay rates are summarized in Table 4.5.

Although the table might seem relatively different between the two countries, the analysis of all the values with the highest and lowest delay rates reveals certain trends that are consistent in both the USA and Brazil, and potentially applicable to other regions as well. It is observed that airlines with the highest and lowest delay rates are often those that operate a relatively small number of flights in either country. A high delay rate can be attributed to the fact that smaller airlines may have fewer resources, making

	United States		Brazil	
Variable \ Delay rate	Highest	Lowest	Highest	Lowest
Airline carrier	JetBlue Airways	Horizon Air	ITA Airways	Air Canada
	Frontier Airlines	Endeavor Air	Air Atlanta Icelandic	Cargojet Airways
	Allegiant Air	SkyWest Airlines	Atlas Air	Amaszonas
Origin airport	Ogden-Hinckley	Francisco C. Ada Saipan International	Eirunepé	Lorenzo
	New Castle	Cedar City Regional	Jomo Kenyatta International	Garanhuns
	Milton J. Ferguson Field	Pocatello Regional	Piarco International	Pampulha
Destination airport	Pago Pago International	Francisco C. Ada Saipan International	Jomo Kenyatta International	Vancouver International
	Punta Gorda Airport	Gustavus Airport	Amsterdam Airport Schiphol	Lorenzo
	Rafael Hernandez	Pitt Greenville	Paris Orly Airport	Piloto Osvaldo Marques Dias
Month	December	October	December	February
	June	September	November	March
	July	November	January	April
Departure hour	0	5	0	5
	21	6	1	4
	20	7	18	6
Day of week	Friday	Tuesday	Friday	Saturday
	Sunday	Wednesday	Thursday	Sunday
	Saturday	Monday	Monday	Tuesday

Table 4.5: Summary of the top three values with the highest and lowest delay rates for the most substantial attributes in both data sets.

them more vulnerable to delays. Conversely, their lower flight volume may contribute to a lower overall delay rate.

Regarding the origin and destination airports, regional airports exhibit both the lowest and highest proportional delays. This can be attributed to the limited number of flights operating at these airports, where on-time or delayed arrivals significantly impact the proportional delay. However, some larger airports also experience notable delay rates, such as Amsterdam Airport Schiphol (high delay rate) and Vancouver International Airport (low delay rate).

The analysis of flight dates and times reveals that the month of December consistently exhibits the highest delay rates in both data sets. This can be associated with the increased travel during the holiday season, specifically around Christmas and New Year. the summer period in the United States and the months surrounding December in Brazil tend to experience higher proportional delays. Conversely, lower delay rates are apparent during the spring season in Brazil and the fall season in the USA.

As for the departure hour, flights in both data sets tend to have lower proportional delays at the beginning of the day and higher proportional delays towards the end of the day. This pattern can be attributed to several factors such as accumulated delays throughout the day, weather conditions, and operational challenges. Flights operating during the early hours of the day typically benefit from a fresh start and less congested airspaces, while flights in the later hours can experience drawbacks from earlier delayed flights and increased operational demands.

Finally, the day of the week does not significantly influence the delay rate in the Brazilian data set. However, in the United States, flights during the weekend tend to experience a significantly higher delay rate compared to flights on weekdays. This observation can be associated with the fact that weekends are generally affiliated with increased travel demand, as people have more leisure time.

4.3.5 Correlation

Assessing the relationships between variables is a critical step in data preprocessing and exploration to identify potential collinearity in the data and ensure reliable data analysis. The correlation between variables can be computed by various methods such as Pearson's correlation coefficient and Spearman's rank correlation coefficient (refer to Section 2.4.2 for more details).

In this research, both correlation coefficients were utilized as the data sets both contain attributes with light-tailed and heavy-tailed distributions. Pearson's correlation coefficient assumes linear relationships, while Spearman's rank correlation coefficient is more robust to non-linear relationships. By using both coefficients, this study was able to capture a broader range of associations between variables and gain a comprehensive understanding of their relationships.

The correlation results are visualized using heatmaps, with Figure 4.8 displaying the heatmaps based on Pearson's correlation coefficient. The heatmaps resulting from Spearman's rank correlation coefficient, which showed no significant differences, are provided in Appendix A, Section A.4.

A correlation significance level of 0.7 was selected, as it is a widely used threshold for correlation coefficients [12]. The correlation heatmaps in Figure 4.8 reveal notable relationships between certain pairs of variables, characterized by high positive or negative correlations exceeding the threshold (>0.7). It is important to note that collinearity among the response variables (delayed, departure delay, arrival delay) does not pose a concern, as only one of them is utilized in the machine learning models.

To address collinearity issues and ensure reliable classification models, correlated variables were removed from the analysis. The United States data set shows strong collinearity between the distance and elapsed time attributes. This is to be expected, as the flight time is largely determined by the flight distance. As a result, the elapsed time attribute was excluded from the analysis. This decision was made because the elapsed time is vulnerable to

external factors such as weather conditions and technical issues, whereas the distance attribute remains constant.

Moreover, there is a noticeably high correlation between the departure hour and arrival hour, as well as between the departure week and month. Consequently, the arrival hour and departure week attributes are removed from the analysis. This decision is made to avoid multicollinearity, which could result in unreliable classification models. It is worth noting that the origin/destination city market and the corresponding origin/destination airport show a relatively high correlation of 0.64. However, as their correlation does not exceed the threshold and the city market and airport can provide distinct and complementary information, both attributes are retained in the analysis.

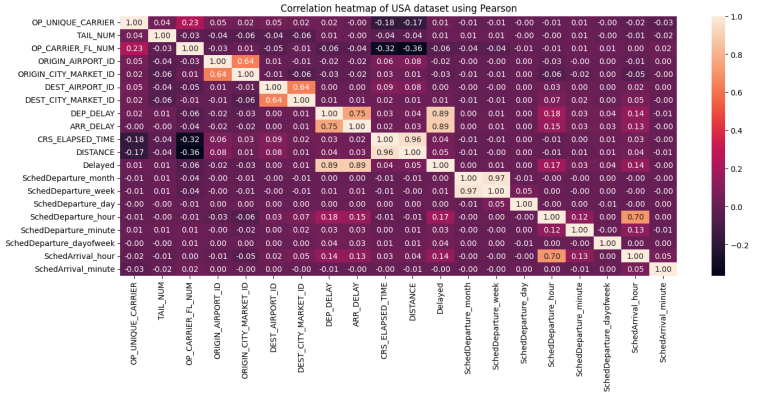
On the other hand, in the Brazilian data set, there is only one attribute pair that exhibits high collinearity, which is the departure week and month. Similar to the approach taken in the USA data set, the departure week was excluded from the analysis. The decision to remove the week attribute was based on the consideration that the month captures the seasonal variations in flight patterns and delays more comprehensively.

To explore the presence of non-linear relationships and heteroscedasticity, scatter plots were created and are presented in Section A.5 of Appendix A. However, the analysis of these plots did not reveal any significant additional insights beyond what was already observed through the correlation analysis.

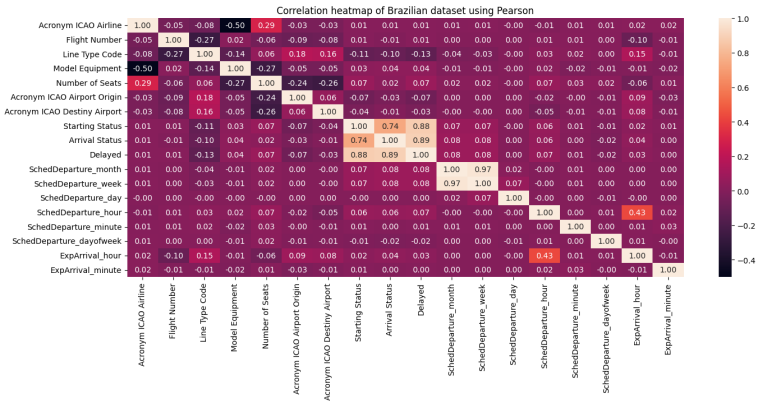
4.4 Classification Algorithms

This study compares eight supervised machine learning algorithms including decision tree (DT), random forest (RF), gradient boosting tree (GBT), k-nearest neighbors (k-NN), naive Bayes (NB), logistic regression (LR), neural network (NN), and support vector machine (SVM). This diverse set of algorithms covers a wide spectrum of classification algorithms, ranging from tree-based methods to probabilistic classifiers and neural networks. These particular classifiers have been selected based on their widespread usage in related studies. It is important to note that this set of classifiers is not exhaustive, but rather represents a comprehensive set of popular and widely-used machine learning algorithms.

Specific hyperparameter choices and limitations are examined for each algorithm, aiming to identify their strengths and weaknesses. Furthermore, a range of scaling and encoding techniques is examined to preprocess the



(a)



(b)

Figure 4.8: Pearson correlation heatmap for the USA data set (a) and the Brazilian data set (b).

input features for the machine learning algorithms. This analysis includes evaluating the effects of normalizing the data and encoding techniques (such as one-hot encoding, ordinal encoding, and target encoding). The goal is to identify the most suitable preprocessing approach that enhances the performance of the algorithms in predicting flight delays.

The primary objective of this paper is to gain insights into their comparative performance and identify the best-performing approaches for both data sets. To achieve this, various performance measures have been employed to evaluate and compare the performance of these algorithms. The specific performance measures used in this study are thoroughly discussed and substantiated in Section 4.6.

4.5 Experimental Setup

The experiments in this project were conducted using python, which was chosen due to its extensive support for machine learning libraries that are highly relevant to this research. The implementation of this paper can be found in the GitHub repository, which provides detailed code accompanied with clear documentation for reproducing the results.

In the implementation of this research, several well-known Python libraries for machine learning were utilized. A wide range of machine learning algorithms was provided by the scikit-learn[6] library, which were used for classification tasks in this study. Moreover, Pandas[6] was used for preprocessing tasks, as it offers efficient data structures and functions for handling large data sets. For data visualization purposes, Matplotlib[6] and Seaborn[6] were employed to create insightful plots and figures. NumPy[6] was utilized for numerical computations and array operations. Finally, TenserFlow[6] and Keras[6] were used for building and training neural networks, including the implementation of transfer learning techniques.

4.6 Model Creation And Evaluation

To create and evaluate the machine learning models, the preprocessed data sets were encoded using various techniques, including one-hot encoding, ordinal encoding, and target encoding (Section 2.6.1). The impact of scaling the data sets was also investigated. This was done by assessing the performance of the classifiers on both the unscaled data set and the data set with all numerical variables scaled.

After selecting the best encoding technique and scaling approach, the next step involved tuning the hyperparameters for each algorithm. This iterative process allowed for the identification of the best configuration for each model. The algorithms, along with their optimized hyperparameters, were then re-evaluated to assess their performance. A summary of this process is provided in Figure 4.9.

4.6.1 Evaluation Methods

The evaluation of algorithms involved assessing their performance using several performance measures, including accuracy, AUC score, precision, recall, F_1 score, efficiency, feature importances, and confusion matrices. Moreover, ROC curves and precision-recall curves were plotted to provide a visual

[6]Refer to the documentation of scikit-learn, Pandas, Matplotlib, Seaborn, NumPy, TenserFlow, and Keras.

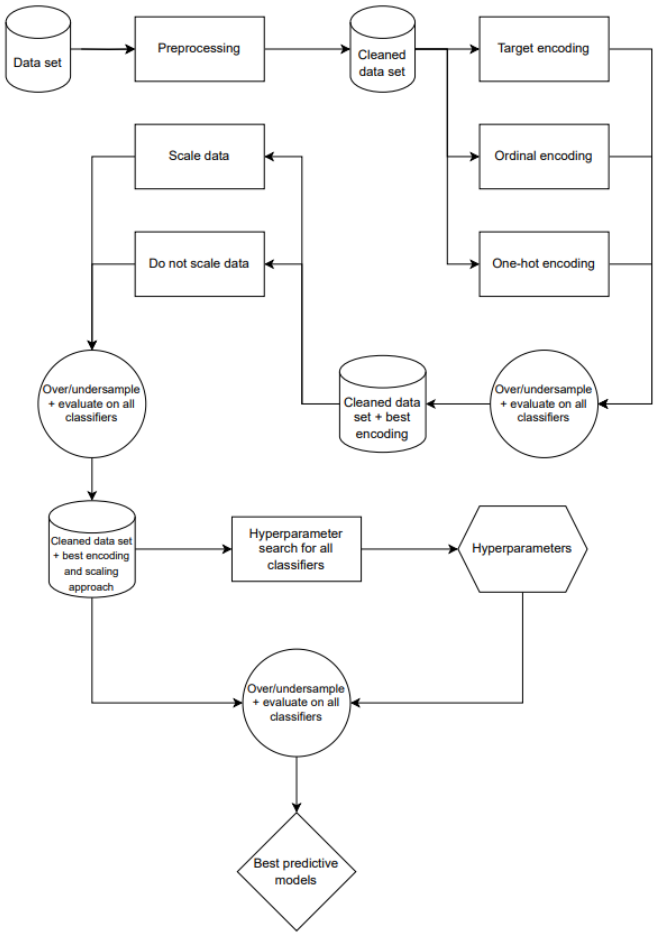


Figure 4.9: Workflow for creating and evaluating machine learning models in this research. The process starts with the original data set in the top left corner. The workflow is indicated with arrows leading to the next step. Each rectangle indicates an experiment, each data store represents the new data set and each circle represents an evaluation.

comparison between classifiers. Additionally, the impact of oversampling and undersampling was considered during evaluation, as only approximately 30% of flights in the USA and 25% in Brazil were delayed.

The chosen performance measures were specifically selected to address the class imbalance issue present in the data sets. While accuracy is an important measure, it can be misleading in imbalanced datasets where a classifier can achieve high accuracy by simply predicting the majority class. Hence, additional measures were employed.

The precision, recall, and F_1 score are particularly valuable for their ability to provide a clear understanding of how well the classifier performs in correctly identifying instances of the minority class (delayed flights) while minimizing false positives. Additionally, other metrics are utilized to gain insights into the classifiers' decision-making process, examine their computational efficiency (which is crucial for airline operations), and enable visual comparisons.

To conduct the evaluation, the data sets were divided into a training set and a test set. Subsequently, each classifier underwent three separate evaluations: one using the training set with oversampling, one with undersampling, and one without any sampling techniques applied. The classifiers were trained on the (modified) training data and then evaluated using the test data. By employing this approach, the aforementioned performance measures were computed, enabling a comprehensive understanding of the classifiers' performance while considering the impact of the imbalanced class distribution.

4.6.2 Hyperparameter Tuning

The machine learning models were optimized through a process called hyperparameter tuning. Hyperparameters are options of an algorithm that determine how it learns and makes predictions. In this study, an iterative approach was employed to explore different combinations of hyperparameter values for each classifier. The evaluation process utilized a nested cross-validation framework to ensure reliable performance assessment. An illustration of cross-validation is given in Figure 4.10.

The data set was divided into training and test sets using a stratified k-fold strategy within the outer loop. To address the class imbalance, the training data was adjusted based on the specified balance method. Within each inner fold, the model was evaluated using stratified cross-validation on the modified training data. The performance measures, including accuracy and F_1 score, were computed for each run. Additionally, the model trained

Figure 4.10: The k-fold cross-validation strategy to split data into training data and testing data.

on the training data was evaluated using the independent test data to assess its generalization ability.

By averaging the performance measures from each run and the test data evaluation, a comprehensive performance overview for each hyperparameter value was obtained. These performance measures were used to plot the relationship between the hyperparameter values and the model's performance. The chosen performance measures of accuracy and F_1 score provide meaningful insights into the overall performance of the models, capturing both the ability to make correct predictions and the balance between precision and recall.

However, it is important to notice that the implementation of this hyperparameter tuning method has its limitations. Firstly, due to computational constraints, the search space for hyperparameters was limited, and better combinations may exist beyond the explored range. Secondly, the hyperparameter tuning process only utilized a fraction of the data (10%), leading to a reduced sample size for evaluating and optimizing the models. This choice was made as it allows for quicker experimentation, provides valuable insights into performance trends, and helps avoid overfitting to the specific dataset. Nonetheless, it provided a solid foundation for model evaluation and comparison.

4.6.3 Ethical Considerations

It is important to highlight that both data sets contain flight records that do not include any personally identifiable information. This careful consideration was made to prioritize and protect individuals' privacy rights and adhere to ethical guidelines in data handling and analysis.

Consequently, the absence of personally identifiable information in the data

sets allows the creation and evaluation of machine learning models to be conducted in compliance with privacy regulations and ethical standards.

4.7 Transfer Learning

In this research, a transfer learning approach was employed as an attempt to enhance the performance of the model on the Brazilian dataset by leveraging knowledge learned from the USA dataset. A sequential neural network model was trained on the United States dataset. It consisted of multiple dense layers and was trained for 5 epochs. The evaluation of the model's performance on the USA dataset was based on the accuracy metric.

Subsequently, a new model was created by transferring the learned weights from the pre-trained model to a modified architecture suitable for the Brazilian dataset. The transferred model was then trained on the Brazilian dataset for 10 epochs. The training progress was tracked using accuracy and loss metrics, and the resulting curves were plotted to visualize the training performance.

Additionally, a separate model was trained on the Brazilian dataset without employing transfer learning. This model served as a baseline for comparison. The training and evaluation of both models provide insights into the effectiveness of transfer learning in improving the performance of the model on the Brazilian dataset. Through the utilization of transfer learning, this research aimed to leverage the knowledge gained from the United States dataset and apply it to the Brazilian dataset, thereby potentially improving the model's performance on the latter dataset. The complete workflow of the transfer learning process in this research is presented in Figure 4.11

To address the challenges of transferring knowledge from the USA dataset to the Brazilian dataset, the research considered various factors. One important aspect was ensuring consistent data processing techniques across both data sets. By using the same steps and formats from the start, the research aimed to establish a common framework where transfer learning could be seamlessly applied.

To achieve this, similar attributes in both data sets were converted to the proper format to ensure compatibility. This involved standardizing data types, units, and representations to facilitate knowledge transfer. Additionally, attribute names were renamed, if necessary, to align with the terminology used in the target dataset. These steps aimed to create a harmonized data structure that would enable effective transfer of knowledge.

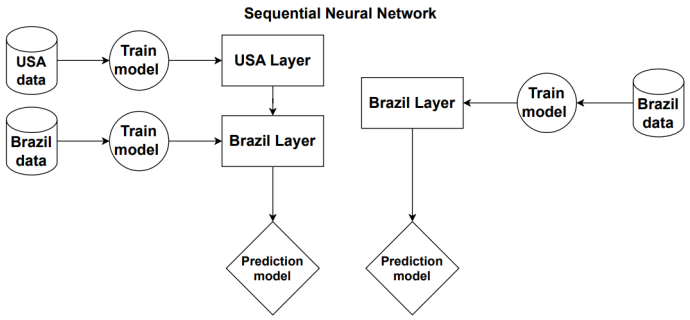


Figure 4.11: Workflow of the transfer learning process. It can be seen that the NN created on the left side used transfer learning, while the NN on the right side is only trained on the Brazilian data set.

The learning effect was considered by initially experimenting with a subset of the data sets to assess transfer learning performance. The sample data allowed for an evaluation of knowledge transfer while mitigating limitations from limited target dataset size. Subsequently, the full data sets were utilized to maximize available information and fine-tune the transfer learning process. This approach provided insights into the learning effect and allowed for a comprehensive analysis of the model's performance on the target dataset.

Noticeably, transfer learning has already been used for the aircraft delay prediction problem by McCarthy et al. [34], but only for small-scale data. Furthermore, they are using the LSTM method, whereas this paper uses a sequential neural network. Their findings demonstrated the effectiveness of transfer learning in leveraging knowledge from a large airline carrier to a smaller one for predicting flight delays. The transfer learning approach resulted in a significant decrease of approximately 5 to 10% in the mean absolute error compared to the base model.

Chapter 5

Results

The outcomes of the machine learning models and transfer learning techniques are presented in this chapter, highlighting their performance in relation to the research objectives. The results are interpreted and analyzed, drawing comparisons with relevant studies. Additionally, the chapter concludes by suggesting potential directions for future research in the field.

5.1 Encoding Techniques

Various encoding techniques have been extensively evaluated in this research to handle the categorical variables present in both data sets. By systematically running all classifiers with different balancing techniques and their default parameter settings, the performance of each encoding technique was assessed. Subsequently, the most effective encoding technique was selected for each data set based on the evaluation results.

5.1.1 One-hot Encoding

The one-hot encoding technique was found to be impractical for both data sets due to the presence of categorical attributes with a large number of unique values, exceeding 50.000. The resulting data frames from one-hot encoding would be too large to create and store, making it infeasible to employ this technique for encoding the categorical variables. As a result, alternative encoding techniques had to be explored.

5.1.2 Ordinal Encoding

The results of the performance evaluation for the machine learning algorithms applied to the USA and Brazilian data sets are presented in Table 5.1 and Table 5.2 respectively. The evaluation is conducted using ordinal encoding for the categorical variables. These tables provide insights into the

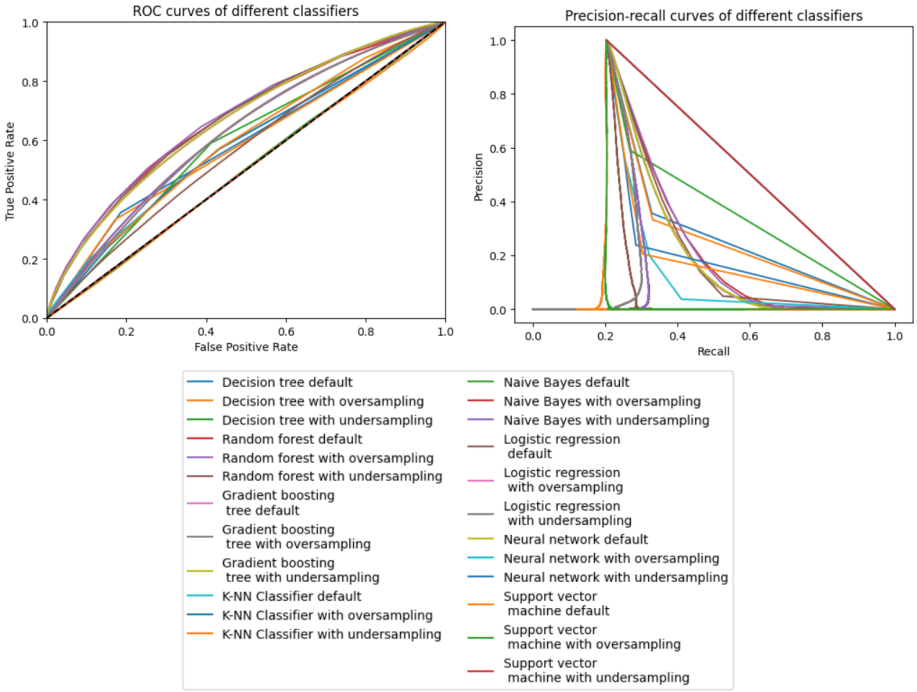


Figure 5.1: ROC curves (top left) and precision-recall curves (top right) of classifiers with different balancing techniques using ordinal encoding on the USA data set, accompanied with their legend (bottom).

accuracy, precision, recall, F_1 score, and AUC score achieved by each classifier under different scenarios: default data, undersampled training data, and oversampled training data.

From Table 5.1, it is evident that the performance of most classifiers on the USA data set is not exceptional, as indicated by the low F_1 scores. This suggests that accurately classifying flight records into one of the two classes is a challenging task. Additionally, the ROC curves and precision-recall curves of the classifiers can be observed in Figure 5.1.

The analysis of AUC scores and ROC curves reveals that the random forest and gradient boosting tree classifiers exhibit higher performance compared to other classifiers, while the neural network and support vector machine classifiers perform similarly to random guessing. The precision-recall curves illustrate relatively low precision and recall values for most classifiers, except for the random forest classifier.

The ROC curves and precision-recall curves of the classifiers applied to the Brazilian data set are visualized in Figure 5.2. The corresponding perfor-

ML Algorithm	Accuracy	Precision	Recall	F_1 score	AUC score
DT	72.11%	58.06%	58.53%	58.27%	0.59
+ oversampling	72.75%	58.03%	58.05%	58.04%	0.58
+ undersampling	58.84%	55.86%	58.94%	53.18%	0.59
RF	79.26%	64.66%	56.19%	56.63%	0.67
+ oversampling	77.66%	62.60%	58.71%	59.69%	0.67
+ undersampling	66.62%	58.77%	62.38%	58.55%	0.67
GBT	79.63%	69.44%	50.01%	44.35%	0.66
+ oversampling	59.96%	57.60%	61.63%	54.83%	0.66
+ undersampling	60.00%	57.59%	61.62%	54.85%	0.66
k-NN	75.15%	56.79%	54.61%	54.89%	0.58
+ oversampling	63.82%	54.58%	56.30%	54.08%	0.58
+ undersampling	57.03%	54.58%	57.01%	51.48%	0.59
NB	79.63%	39.81%	50.00%	44.33%	0.62
+ oversampling	57.58%	55.97%	59.17%	52.61%	0.62
+ undersampling	57.55%	55.97%	59.17%	52.60%	0.62
LR	79.63%	39.81%	50.00%	44.33%	0.57
+ oversampling	58.95%	55.99%	59.13%	53.32%	0.62
+ undersampling	58.94%	55.98%	59.12%	53.31%	0.62
NN	79.63%	39.81%	50.00%	44.33%	0.50
+ oversampling	20.37%	10.19%	50.00%	16.93%	0.50
+ undersampling	79.63%	39.81%	50.00%	44.33%	0.50
SVM	46.88%	49.90%	49.85%	43.48%	0.50
+ oversampling	51.93%	50.43%	50.66%	46.39%	0.50
+ undersampling	53.02%	50.84%	51.28%	47.15%	0.50

Table 5.1: Performance measures of classifiers with different balancing techniques using ordinal encoding on the USA data set.

ML Algorithm	Accuracy	Precision	Recall	F_1 score	AUC score
DT	78.57%	59.39%	60.24%	59.77%	0.60
+ oversampling	79.43%	59.77%	59.86%	59.81%	0.60
+ undersampling	61.02%	55.99%	61.50%	52.49%	0.61
RF	84.94%	68.32%	58.31%	60.17%	0.70
+ oversampling	83.17%	64.71%	60.94%	62.28%	0.70
+ undersampling	69.90%	58.86%	65.32%	58.58%	0.71
GBT	85.11%	75.00%	50.60%	47.28%	0.69
+ oversampling	67.69%	57.89%	64.08%	56.90%	0.70
+ undersampling	67.44%	57.86%	64.11%	56.78%	0.69
k-NN	83.18%	63.50%	58.53%	59.91%	0.64
+ oversampling	73.94%	57.90%	61.34%	58.54%	0.64
+ undersampling	63.08%	56.55%	62.37%	53.86%	0.66
NB	82.00%	58.22%	54.53%	55.08%	0.64
+ oversampling	75.44%	56.84%	58.57%	57.37%	0.64
+ undersampling	75.62%	56.77%	58.34%	57.27%	0.64
LR	85.02%	69.30%	50.10%	46.20%	0.57
+ oversampling	54.99%	53.49%	56.85%	47.78%	0.59
+ undersampling	55.38%	53.18%	56.22%	47.76%	0.58
NN	84.95%	65.27%	50.91%	48.12%	0.67
+ oversampling	65.55%	57.17%	63.17%	55.44%	0.68
+ undersampling	19.88%	53.33%	51.46%	19.22%	0.58
SVM	55.41%	51.25%	52.42%	46.41%	0.49
+ oversampling	56.16%	50.14%	50.27%	45.86%	0.50
+ undersampling	62.94%	52.06%	53.64%	50.02%	0.50

Table 5.2: Performance measures of classifiers with different balancing techniques using ordinal encoding on the Brazilian data set.

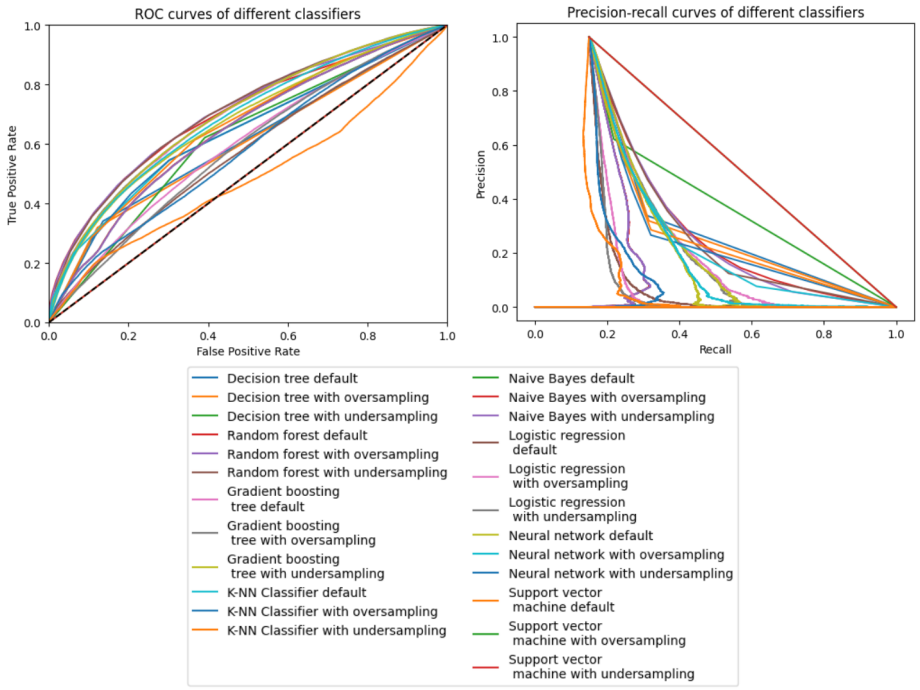


Figure 5.2: ROC curves (top left) and precision-recall curves (top right) of classifiers with different balancing techniques using ordinal encoding on the Brazilian data set, accompanied with their legend (bottom).

mance evaluation in Table 5.2 demonstrates that the classifiers applied to the Brazilian data exhibit similar effectiveness to those applied to the USA data set. Notably, the classifiers applied to the Brazilian data set achieve higher accuracies, F_1 scores, and AUC scores. Moreover, the relative ranking of classifier performance remains largely consistent across both data sets. These observations suggest that the chosen classifiers exhibit consistent behaviour and performance across different data sets.

5.1.3 Target Encoding

Similar to the performance evaluation of ordinal encoding, the results of the performance evaluation for the machine learning algorithms applied to the USA and Brazilian data sets, using target encoding for categorical variables, are presented in Appendix B, Section B.1. These tables provide a comprehensive overview of the accuracy, precision, recall, F_1 score, and AUC score achieved by each classifier across various scenarios: default data, undersampled training data, and oversampled training data.

Additionally, the ROC curves and precision-recall curves of the classifiers applied to the USA and Brazilian data sets are visualized in Appendix B,

Section B.1. Comparing the performance evaluations of ordinal encoding and target encoding, it is observed that the encoding techniques yield similar performance measures, ROC curves, and precision-recall curves for both data sets. Notably, the neural network and support vector machine classifiers show improved performance, which can be attributed to the use of a maximum number of iterations in these experiments.

5.1.4 Comparison

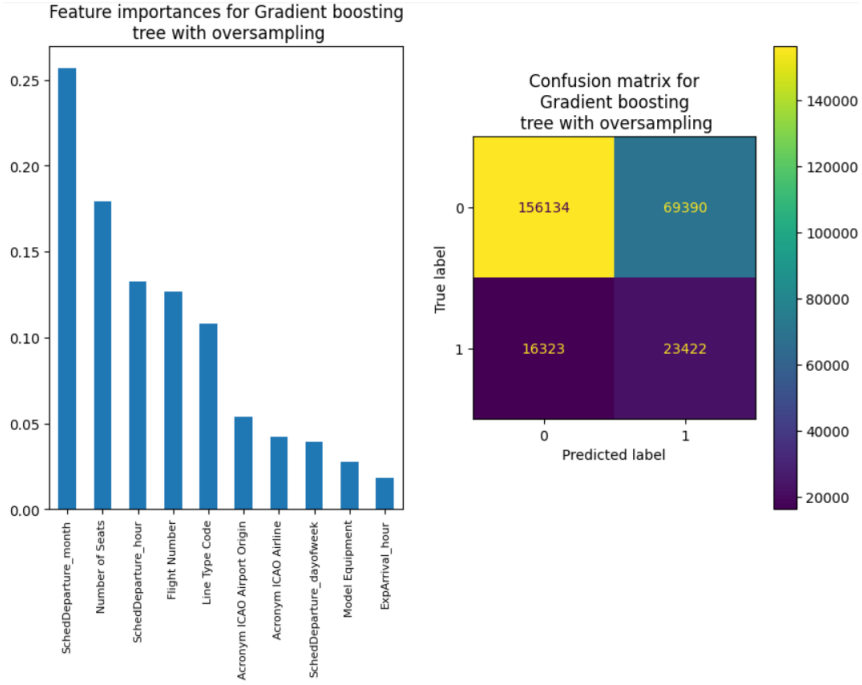
To decide which encoding technique performs better for this paper's use case, it is crucial to look at the decision-making of the classifier instead of only looking at performance measures. Namely, the performance metrics achieved by classifiers using ordinal encoding were comparable to those using target encoding. The feature importances and confusion matrix of the GBT classifier with oversampling applied to the Brazilian data set using both encoding techniques are shown in Figure 5.3.

The figure clearly illustrates that the utilization of target encoding introduced a noticeable bias in the decision-making process of the classifiers. Notably, the flight number feature exhibited a disproportionately strong influence on the predictions of the classifiers, indicating overfitting. This observation raises valid concerns regarding the generalizability and fairness of the models, as they heavily rely on a single feature rather than taking into account a wider range of variables. It is reasonable to expect overfitting when target encoding is employed, as this encoding technique replaces values with a metric linked to the target variable. On the other hand, ordinal encoding treated all categorical variables equally, ensuring a more balanced and unbiased approach to feature representation.

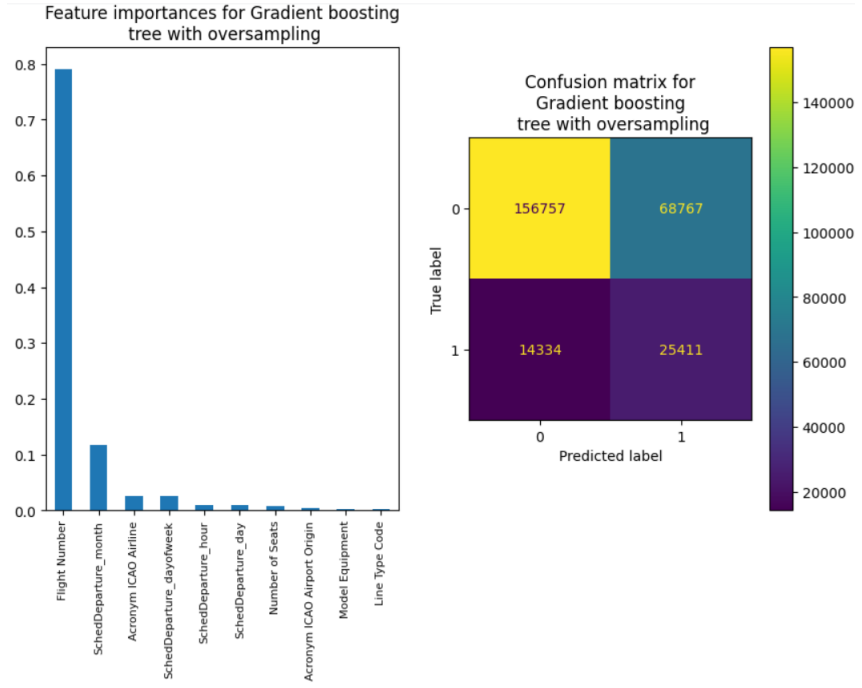
Based on the findings and comparison of the performance evaluations, the decision has been made to utilize ordinal encoding for categorical variables in the subsequent experiments. This encoding technique demonstrated more reliable and interpretable results compared to target encoding.

5.2 Data Scaling

To explore the potential impact of scaling on the performance of distance-based classifiers, the data sets were scaled and evaluated using the same methodology as employed in finding the best encoding techniques. The results of this performance evaluation for the machine learning algorithms applied to the scaled USA and Brazilian data sets are presented in Section B.2 of Appendix B. Additionally, the ROC curves and precision-recall curves of the classifiers are visualized in Figure 5.4 and Figure 5.5. It is worth noting that the categorical variables in the data sets were encoded using ordinal



(a)



(b)

Figure 5.3: Feature importances and confusion matrix for the GBT classifier with oversampling applied to the Brazilian data set using ordinal encoding (a) and target encoding (b). It is evident that the classifier is overfitting on the flight number in (b), when target encoding is applied.

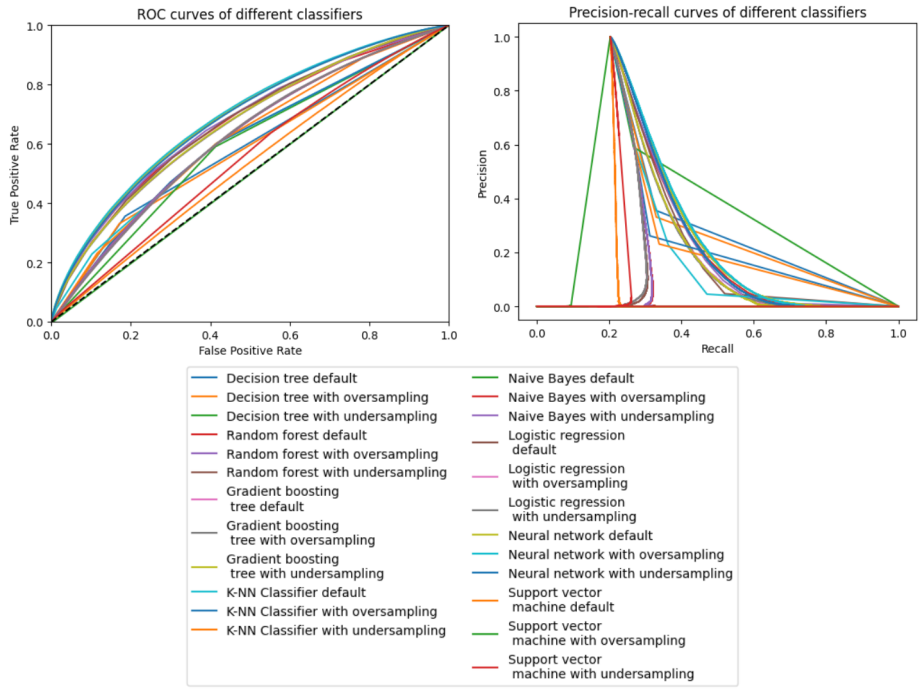


Figure 5.4: ROC curves (top left) and precision-recall curves (top right) of classifiers with different balancing techniques using ordinal encoding on the scaled USA data set, accompanied with their legend (bottom).

encoding, as Section 5.1 demonstrated its effectiveness as the most suitable encoding technique for this research.

After a thorough analysis of the results obtained from scaled and unscaled data sets, it was determined that utilizing unscaled data sets produces more reliable and interpretable results. This decision was based on two key factors. Firstly, scaling the data could potentially remove important domain-specific knowledge, as certain features may have distinct absolute values or specific ranges that carry significant information. For instance, the number of seats variable has a wide range compared to other variables, and scaling could obscure this valuable information.

Secondly, scaling can impact the range and distribution of feature values, which in turn affects the calculation of feature importance. When features are scaled, their relative importance may change, making it harder to directly interpret and compare the importance values across different features. Additionally, scaling can mask the true impact of certain features on the model's predictions. It is important to note that during the evaluation of the performance metrics in Appendix B, Section B.2, no significant differ-

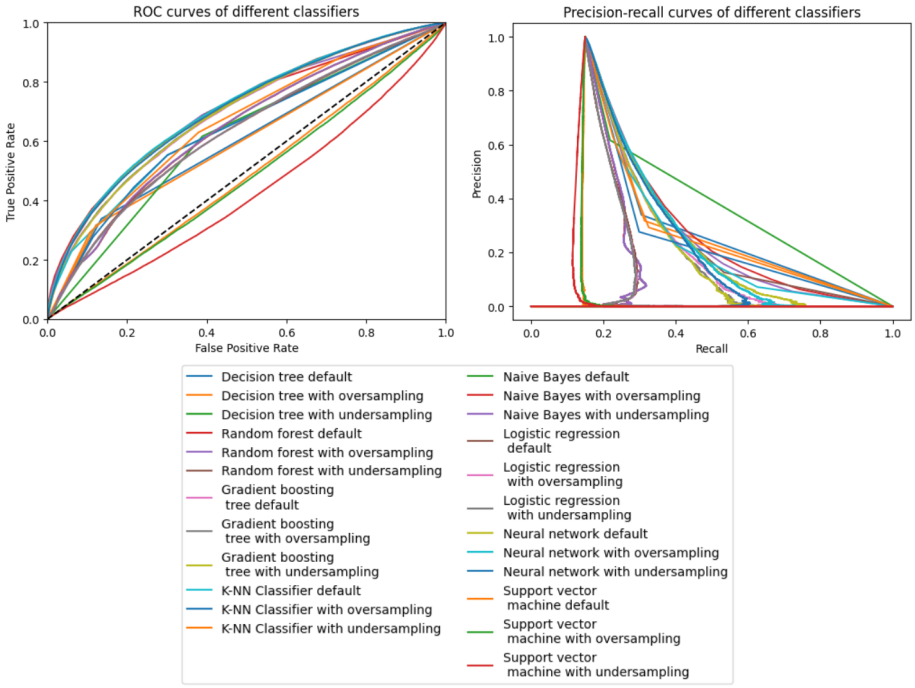


Figure 5.5: ROC curves (top left) and precision-recall curves (top right) of classifiers with different balancing techniques using ordinal encoding on the scaled Brazilian data set, accompanied with their legend (bottom).

ences were observed between the models trained on scaled and unscaled data. Despite this, the decision to utilize unscaled data sets for the subsequent analyses in this research was primarily driven by the consideration of domain-specific knowledge and interpretability, as discussed earlier.

5.3 Hyperparameter Tuning

To optimize classifier performance, hyperparameter tuning was conducted on each classifier, and the results are presented in Appendix B, Section B.3. The plots depict the relationship between parameter values and performance metrics (accuracy and F_1 score), enabling the identification of optimal parameter settings. Four of these plots are shown in Figure 5.6. The accuracy and F_1 score were selected as the evaluation metrics due to their ability to generalize the model's performance.

After conducting the hyperparameter tuning process for each classifier, specific hyperparameter values were selected based on their impact on the accuracy and F_1 score. The F_1 score was considered more relevant when a tradeoff had to be made between accuracy and the F_1 score. For instance,

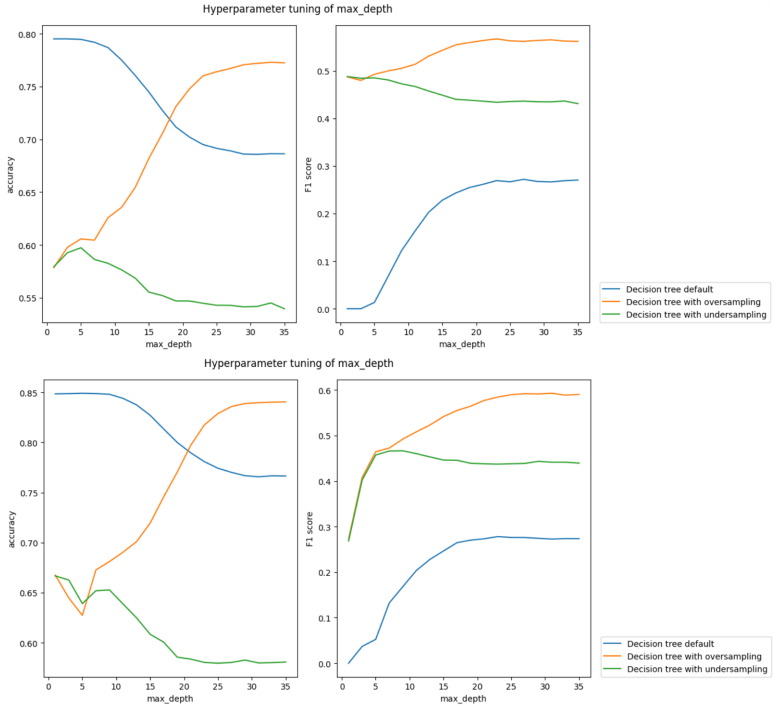


Figure 5.6: Plots illustrating the impact of varying the `max_depth` parameter values on the accuracy and F_1 score of the decision tree classifier applied to the USA (left) and Brazilian (right) data set. It can be seen that for both data sets, the accuracy only increases for an oversampled training data set when the `max_depth` increases. The F_1 score increases in all scenarios when the `max_depth` increases, but stagnates around a `max_depth` of 25.

in Figure 5.6, a `max_depth` value of 30 was chosen for the decision tree classifier applied to both data sets, as it exhibited a significant increase in the F_1 score while sacrificing some accuracy.

The chosen hyperparameter values aim to optimize the classifiers' performance in predicting flight delays. A summary of the selected hyperparameter values for each classifier applied to the USA and Brazilian data sets can be found in Table 5.3 and Table 5.4 respectively.

5.4 Interpretation and Comparison

Following a thorough evaluation of different encoding and scaling techniques, as well as the hyperparameter tuning phase, the classifiers employed in this study were ready to be compared in terms of their performance on each data set. This comparative analysis allows for the selection of the top-performing models, which can then be utilized to derive the feature importances.

Classifier	Hyperparameter Values
Decision Tree	max_depth=30, min_samples_split=5, min_samples_leaf=1
Random Forest	n_estimators=18, max_depth=30, min_samples_split=5, min_samples_leaf=5
Gradient Boosting Tree	n_estimators=15, learning_rate=0.2, max_depth=15
K-Nearest Neighbors	n_neighbors=3, weights='distance'
Naive Bayes	var_smoothing=10^{-9}
Logistic Regression	max_iter=800, solver='liblinear'
Neural Network	hidden_layer_sizes=(50, 50), alpha=0.0001, activation='logistic', max_iter=350
Support Vector Machine	kernel="linear", max_iter=400, probability=True, C=10, gamma='auto', shrinking=True

Table 5.3: Selected hyperparameter values for each classifier applied to the USA data set.

Classifier	Hyperparameter Values
Decision Tree	max_depth=30, min_samples_split=5, min_samples_leaf=1
Random Forest	n_estimators=15, max_depth=30, min_samples_split=20, min_samples_leaf=5
Gradient Boosting Tree	n_estimators=15, learning_rate=0.2, max_depth=17
K-Nearest Neighbors	n_neighbors=3, weights='distance'
Naive Bayes	var_smoothing=10^{-9}
Logistic Regression	max_iter=800, solver='liblinear'
Neural Network	hidden_layer_sizes=(50, 50), alpha=0.01, activation='logistic', max_iter=350
Support Vector Machine	kernel="linear", max_iter=500, probability=True, C=10, gamma='scale', shrinking=True

Table 5.4: Selected hyperparameter values for each classifier applied to the Brazilian data set.

5.4.1 Performance Evaluation

Table 5.5 presents the results of the performance evaluation for the USA data set, providing a detailed analysis of the accuracy, precision, recall, F_1 score, and AUC score achieved by each classifier. These metrics are examined under different scenarios, including default data, undersampled training data, and oversampled training data. The evaluation is conducted on the unscaled data set using ordinal encoding, as this has been identified as the most suitable technique. Furthermore, the parameter values employed for each classifier are the ones determined during the hyperparameter tuning phase.

Likewise, Table 5.6 presents the performance evaluation results for the Brazilian data set. This table offers insights into the classifier performance metrics under the same scenarios as the USA data set. Similar to the USA data set, the evaluation for the Brazilian data set is performed on the unscaled data using ordinal encoding. The hyperparameters used for each classifier are also the ones determined in the hyperparameter tuning phase.

Alongside the performance evaluations, the ROC curves and precision-recall curves for each classifier applied to the USA and Brazilian data set are visualized in Figure 5.7 and Figure 5.8 respectively. These curves provide

ML Algorithm	Accuracy	Precision	Recall	F_1 score	AUC score
DT	74.22%	59.14%	58.45%	58.74%	0.60
+ oversampling	71.11%	58.02%	59.10%	58.39%	0.59
+ undersampling	60.29%	56.14%	59.27%	54.05%	0.60
RF	80.66%	71.64%	55.58%	55.40%	0.72
+ oversampling	74.74%	62.74%	64.28%	63.35%	0.71
+ undersampling	65.96%	60.35%	65.29%	59.55%	0.71
GBT	81.22%	73.82%	57.13%	57.79%	0.74
+ oversampling	71.22%	62.69%	67.35%	63.31%	0.73
+ undersampling	67.79%	61.59%	66.92%	61.24%	0.73
k-NN	73.82%	56.06%	54.77%	55.07%	0.58
+ oversampling	65.64%	54.60%	55.97%	54.50%	0.58
+ undersampling	56.70%	54.36%	56.68%	51.18%	0.59
NB	79.63%	39.81%	50.00%	44.33%	0.62
+ oversampling	57.59%	55.97%	59.17%	52.62%	0.62
+ undersampling	57.59%	55.97%	59.17%	52.62%	0.62
LR	79.63%	39.81%	50.00%	44.33%	0.62
+ oversampling	58.99%	56.01%	59.16%	53.35%	0.62
+ undersampling	59.00%	56.01%	59.16%	53.36%	0.62
NN	79.63%	39.81%	50.00%	44.33%	0.51
+ oversampling	79.63%	39.81%	50.00%	44.33%	0.50
+ undersampling	79.63%	39.81%	50.00%	44.33%	0.50
SVM	42.65%	50.14%	50.20%	40.94%	0.51
+ oversampling	51.23%	50.40%	50.62%	46.03%	0.50
+ undersampling	43.11%	51.28%	51.83%	41.62%	0.48

Table 5.5: Performance measures of classifiers with different balancing techniques using ordinal encoding and selected hyperparameter values on the USA data set.

ML Algorithm	Accuracy	Precision	Recall	F_1 score	AUC score
DT	80.64%	60.75%	59.78%	60.21%	0.62
+ oversampling	78.37%	59.19%	60.10%	59.58%	0.60
+ undersampling	63.22%	56.18%	61.60%	53.64%	0.63
RF	86.18%	76.91%	57.03%	58.71%	0.75
+ oversampling	78.60%	62.90%	67.19%	64.25%	0.75
+ undersampling	70.48%	60.18%	67.86%	59.95%	0.74
GBT	85.41%	70.33%	59.77%	62.03%	0.73
+ oversampling	79.98%	63.31%	65.96%	64.35%	0.72
+ undersampling	69.14%	59.44%	66.84%	58.80%	0.72
k-NN	82.55%	62.46%	58.68%	59.87%	0.64
+ oversampling	74.89%	58.00%	60.92%	58.70%	0.64
+ undersampling	62.38%	56.24%	61.85%	53.31%	0.66
NB	82.00%	58.22%	54.53%	55.08%	0.64
+ oversampling	75.47%	56.83%	58.52%	57.34%	0.64
+ undersampling	75.51%	56.75%	58.37%	57.25%	0.64
LR	84.86%	53.52%	50.08%	46.30%	0.62
+ oversampling	61.01%	55.12%	59.75%	51.84%	0.63
+ undersampling	60.97%	55.14%	59.79%	51.83%	0.63
NN	84.96%	65.39%	50.80%	47.86%	0.65
+ oversampling	65.88%	56.76%	62.27%	55.23%	0.67
+ undersampling	73.21%	56.78%	59.69%	57.24%	0.64
SVM	57.51%	50.36%	50.69%	46.58%	0.49
+ oversampling	52.99%	49.51%	49.06%	44.05%	0.50
+ undersampling	50.31%	49.33%	48.69%	42.72%	0.51

Table 5.6: Performance measures of classifiers with different balancing techniques using ordinal encoding and selected hyperparameter values on the Brazilian data set.

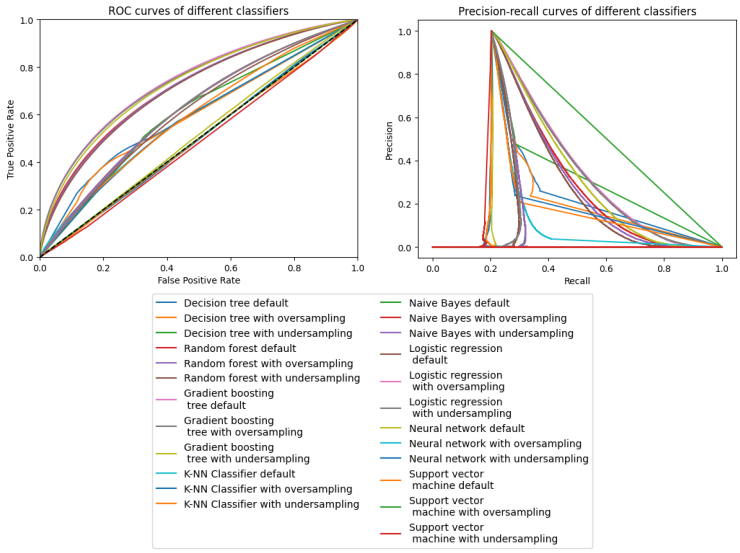


Figure 5.7: ROC curves (top left) and precision-recall curves (top right) of classifiers with different balancing techniques using ordinal encoding and selected hyperparameter values on the USA data set, accompanied with their legend (bottom).

further insights into the classifiers' performance, allowing for a comprehensive analysis of their predictive capabilities.

From the comprehensive analysis of the performance evaluation, ROC curves, and precision-recall curves of the classifiers applied to the USA data set, several conclusions can be drawn.

In the evaluation of the decision tree classifier, it achieved an accuracy of ∼74%, an F_1 score of ∼59%, and an AUC score of 0.6 on the unbalanced data set. When oversampling was applied to the training data, the performance slightly decreased, but the F_1 score remained relatively stable at around 58%. However, when undersampling was performed on the training data, the performance significantly deteriorated, resulting in an accuracy of approximately 60% and an F_1 score of 54%.

The random forest classifier exhibited better performance with an accuracy of approximately 80% and an AUC score of 0.72 on the unbalanced data set. When applied to the oversampled training data, the random forest classifier achieved an F_1 score of 63.35%, indicating its improved ability to balance precision and recall. However, this improvement came at the cost of a slight decrease in accuracy.

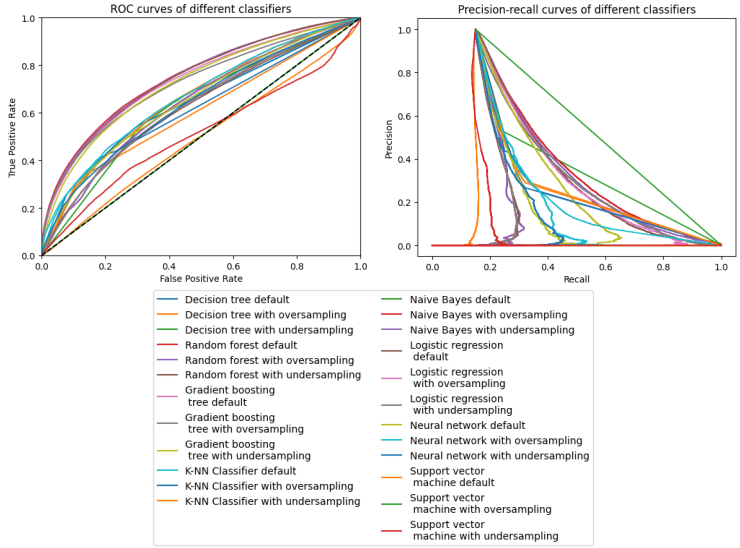


Figure 5.8: ROC curves (top left) and precision-recall curves (top right) of classifiers with different balancing techniques using ordinal encoding and selected hyperparameter values on the Brazilian data set, accompanied with their legend (bottom).

Similarly, the gradient boosting tree classifier performed best on the unbalanced training data with an accuracy of 81.22% and an AUC score of 0.74. However, its F_1 score was relatively lower at around 58% compared to the gradient boosting tree classifier applied to the oversampled or undersampled training data set.

The K-nearest neighbor classifier showed slightly lower performance compared to the tree-based classifiers, achieving an accuracy of approximately 74% and an F_1 score of 55%. The naive Bayes classifier, logistic regression classifier, and neural network performed even worse. Although their accuracy of nearly 80% may appear high, they tend to predict every instance as non-delayed, as evident from their confusion matrices. Figure 5.9 illustrates one of these matrices. While the naive Bayes and logistic regression classifiers improved their F_1 score when trained on oversampled/undersampled data, the neural network consistently predicted every instance as non-delayed.

Lastly, the support vector machine classifier exhibited the poorest performance among the classifiers applied to the USA data set. Its highest accuracy was just above 50%, while the F_1 score did not exceed 46%. These results suggest that the support vector machine struggles to identify the correct patterns in the data set, performing worse than random guessing.

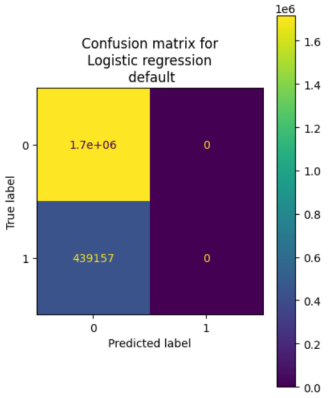


Figure 5.9: Confusion matrix of the logistic regression classifier applied to USA data set, which predicts every instance as non-delayed to get a high accuracy at the cost of a low F_1 score.

Together with the ROC curves and precision-recall curves in Figure 5.7, it can thus be concluded that the random forest and gradient boosting tree classifiers demonstrated the highest performance in predicting flight delays in the USA data set, outperforming other classifiers in terms of accuracy, precision, recall, F_1 score, and AUC score. However, it is worth noting that there was a trade-off between F_1 score and accuracy when oversampling was applied. While the F_1 score increased, the accuracy decreased. Therefore, depending on the specific requirements and considerations of the problem at hand, one can choose the random forest classifier with the unbalanced data for higher accuracy or the oversampled data for improved F_1 score and a better balance between precision and recall.

Upon analysis of the performance evaluation, ROC curves, and precision-recall curves of the classifiers applied to the Brazilian data set, additional conclusions can be drawn. The decision tree classifier exhibits an accuracy of approximately 80% and an F_1 score of 60%. The impact of a balanced training data set does not significantly influence the classifier's performance, as the differences observed are only a few percentage points.

Both the random forest and gradient boosting tree classifiers demonstrate similar patterns, with accuracies of around 86% and 85%, respectively. The gradient boosting tree classifier shows a slightly higher F_1 score, indicating a better balance between precision and recall. The influence of balanced training data on these classifiers aligns with the findings of the classifiers applied to the USA data set, highlighting the trade-off between accuracy and F_1 score.

The K-nearest neighbors classifier exhibits performance similar to the tree-based classifiers. Similarly, the naive Bayes classifier achieved comparable accuracies, but at the cost of a lower F_1 score of around 57%.

The logistic regression classifier performs suboptimally, with F_1 scores barely surpassing the 50% mark. In contrast, the neural network achieves better overall performance, with relatively high accuracies and AUC scores. However, the balance between precision and recall is still far from optimal. Once again, the support vector machine classifier proves to be the worst-performing classifier, with a maximum F_1 score of 46.58% on the Brazilian data set.

In conclusion, the random forest and gradient boosting tree classifiers consistently exhibit the highest performance in predicting flight delays, this time on the Brazilian data set. This observation is reinforced by the analysis of the ROC curves and precision-recall curves in Figure 5.8. However, it is worth noting that the choice between unbalanced and balanced training data still depends on the specific requirements and trade-offs, particularly between accuracy and the F_1 score.

5.4.2 Precision-recall Curves

The observed patterns in the precision-recall curves from Figure 5.7 and Figure 5.8 align with the findings from feature importance analysis and performance metrics such as precision, recall, and AUC score. The presence of a single sharp bend in some curves corresponds to the classifiers' heavy reliance on a particular feature or criterion, as indicated by the feature importance analysis. This behavior is reflected in the abrupt transition from high precision to high recall.

Similarly, the steepness of certain curves, combined with deviations towards the left, reflects classifiers that exhibit high confidence but fail to capture a significant number of positive instances, consistent with low recall and high precision. The presence of no slope around precision 0 suggests an inability to distinguish between positive and negative instances, resulting in a high number of false positives, while drastic veering to the left indicates a classifier struggling to differentiate between positive and negative instances, leading to poor overall performance.

5.4.3 Feature Importances

In order to accomplish the goals of this paper and provide insights to passengers and aviation companies regarding the most influential factors of flight delays, it is crucial to examine the feature importance values of each classi-

	1	2	3	4	5
DT	Tail Number	Departure Day	Flight Number	Departure Hour	Distance
RF	Tail Number	Departure Hour	Departure Day	Flight Number	Departure Month
GBT	Departure Hour	Departure Day	Departure Month	Tail Number	Flight Number
k-NN	Airline	Departure Day Of Week	Departure Month	Departure Hour	Departure Minute
NB	Departure Hour	Flight Number	Departure Day Of Week	Distance	Departure Minute
LR	Departure Hour	Departure Day Of Week	Airline	Departure Minute	Arrival Minute
NN	Origin City Market	Destination City Market	Origin Airport	Destination Airport	Flight Number
SVM	Departure Month	Departure Minute	Tail Number	Departure Day	Departure Day Of Week

Table 5.7: Top 5 most important features for each classifier applied to the USA data set.

	1	2	3	4	5
DT	Departure Day	Flight Number	Departure Month	Departure Day Of Week	Origin Airport
RF	Flight Number	Departure Month	Departure Day	Origin Airport	Departure Hour
GBT	Departure Day	Flight Number	Departure Month	Departure Day Of Week	Origin Airport
k-NN	Line Type Code	Number Of Seats	Flight Number	Airline	Departure Month
NB	Number Of Seats	Line Type Code	Origin Airport	Destination Airport	Departure Month
LR	Line Type Code	Departure Month	Departure Day Of Week	Departure Hour	Arrival Hour
NN	Departure Month	Departure Hour	Airline	Number Of Seats	Model Equipment
SVM	Departure Day	Departure Month	Line Type Code	Model Equipment	Departure Day Of Week

Table 5.8: Top 5 most important features for each classifier applied to the Brazilian data set.

fier. These values indicate the major factors considered by the classifiers to classify flight records.

The top 5 most important features for each classifier applied to the USA and Brazilian data sets are summarized in Table 5.7 and Table 5.8 respectively. They were obtained by analyzing the feature importances in different scenarios: one with unbalanced data, one with oversampled training data, and one with undersampled training data. The importance values were aggregated, resulting in the top 5 lists presented in the tables.

The tables of the top-performing classifiers, specifically decision tree and gradient boosting tree, identified in Section 5.4.1, reveal that the most important features for predicting flight delays in the United States dataset are the departure hour, day, month, and tail number. On the other hand, in the Brazilian dataset, there is a noticeable shift towards the significance of the departure day, month, and flight number as the primary factors.

To gain a more comprehensive understanding of the influential factors across all classifiers, it is important to consider the feature importance values collectively within each dataset. This approach allows for an overall impression of the most influential factors contributing to flight delays. Since different classifiers analyze distinct aspects of the data, an aggregate summary provides a more representative view of the key factors impacting flight delays.

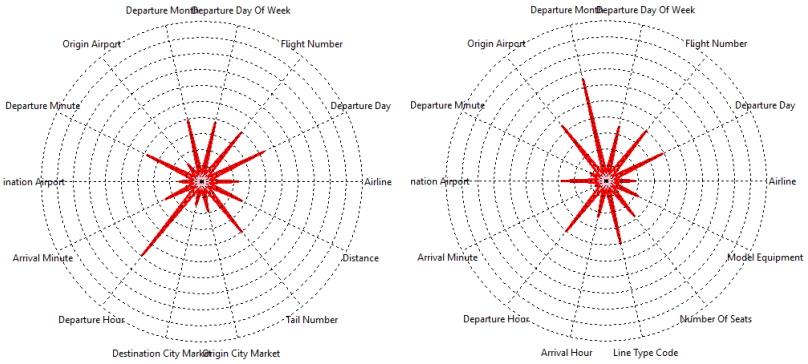


Figure 5.10: Stellar charts summarizing the key factors impacting flight delays in the United States (left) and Brazilian (right) data sets, aggregated across the importance scores from each classifier.

In order to obtain this summary, the importance scores of all relevant attributes from each classifier were aggregated for both the United States and Brazilian data sets. The resulting summary, represented by the weightage of attributes within each classifier[1], is visualized in the stellar charts shown in 5.10. These stellar charts provide a concise overview of the significant factors contributing to flight delays in each dataset.

To further analyze and compare the most influential factors contributing to flight delays in the United States and Brazil, a radar chart has been generated and presented in Figure 5.11. This chart offers a visual representation of the variations in key factors between the two data sets, allowing for a comprehensive understanding of the distinct patterns in each country.

The stellar charts and radar chart show that, in the USA, the most influential factor by a significant margin is the departure hour, indicating that the time of day plays a crucial role in flight delays. Additionally, the departure day, tail number, and departure month also contribute to the prediction of flight delays in the USA.

On the other hand, in Brazil, the departure month emerges as the dominant factor affecting flight delays. This suggests that certain months may experience higher levels of delays compared to others. The origin airport, departure hour, flight number, and line type code also exhibit notable importance in predicting flight delays in Brazil.

[1]The more important a feature is for a classifier, the higher the weight it gets in the summary.

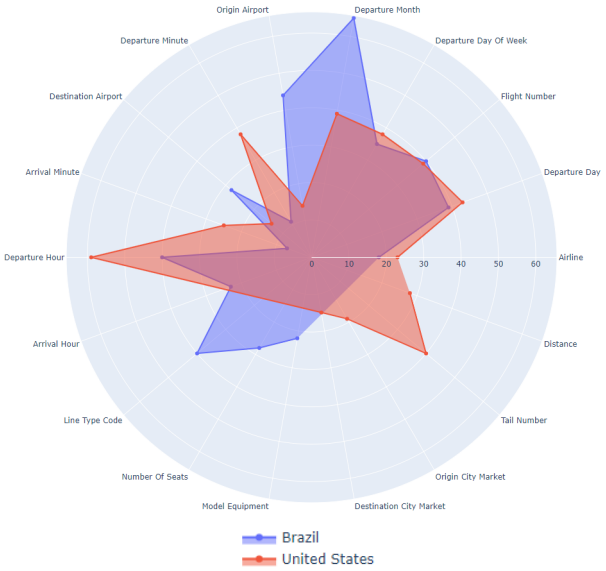


Figure 5.11: Radar chart presenting the most influential factors of flight delays in the United States and Brazil.

Conversely, the origin/destination airport and origin/destination city market appear to have less impact on flight delays in the USA. Similarly, in Brazil, factors such as arrival minute, departure minute, and airline seem to have relatively lower significance in predicting flight delays.

By understanding these specific factors of flight postponements, aviation companies and passengers can make more informed decisions and take necessary precautions to mitigate the risk of delays.

5.4.4 Related Work Comparison

In comparison to related works, this thesis has made significant advancements by exploring a wider range of machine learning algorithms and incorporating diverse data sets. Notably, Moreira et al. [37] achieved an accuracy of 78% using a neural network, which serves as a relevant benchmark for comparison.

In this study, classifiers were developed and evaluated using the USA and Brazilian data sets, resulting in accuracies of 81% and 86% respectively. It should be noted that the variations in accuracy can be attributed to differences in data sets used by different researchers. This research leveraged unique data sets specific to each country, which likely contributed to the improved performance of the classifiers.

Regarding feature importance values, Esmaeilzadeh et al. [13] identified the departure demand-capacity level, weather activity, and TMIs as key factors for predicting flight delays. However, it is important to consider that this study specifically focused on stateless flight record data with the aim of making predictions well in advance. As a result, factors such as the departure month and hour emerged as the most influential in this research.

However, the results of this study roughly match the findings by Sternberg et al. [53]. They found meteorological conditions as the most influential factors contributing to flight delays in Brazil, while this paper purposely did not consider weather conditions. However, their results also showed that flight delays were more prevalent during vacation months and on Fridays. Additionally, their study revealed that Brazil experienced delay propagation, with late evening and night being the most critical periods associated with significant flight delays. These results match the data exploration conclusions from this research. Lastly, they found that departures in the early or mid-morning had a lower likelihood of being delayed, which is supported by the conclusions of this study.

5.4.5 Practical Use

Based on the findings of this thesis, several recommendations can be made for both passengers and aviation companies. The random forest and gradient boosting tree classifiers have demonstrated the highest performance for flight records in the USA and Brazil, achieving potential accuracies of 81% and 86% respectively. However, when choosing a classifier, aviation companies need to carefully consider their specific use cases and the trade-off between accuracy and the F_1 score.

If an airline prioritizes accuracy and is willing to accept a higher number of false negatives (flights predicted as non-delayed but actually delayed), then using an unbalanced data set without data balancing techniques would be sufficient. On the other hand, if the aim is to achieve a high F_1 score and the airline is less concerned about a few additional false positives (flights predicted as delayed but actually not delayed), oversampling the training data can be an appropriate approach. The impact of the balancing technique on the number of false positives and false negatives is evident when examining the confusion matrices of the random forest classifier, as illustrated in Figure 5.12.

In practice, most airlines are likely to prioritize minimizing false positives. This is driven by the fact that allocating extra resources for flights that are predicted to be delayed but turn out to be on time can be very costly, whereas dealing with unexpected delays is already a common occurrence.

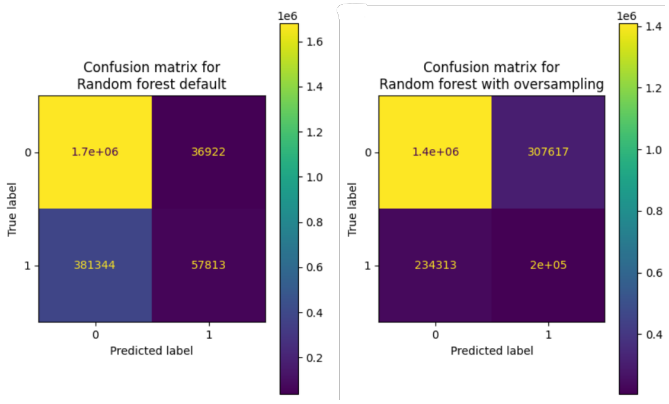


Figure 5.12: Confusion matrix of the random forest classifier applied to the USA data set without balanced data (left) and oversampled training data (right).

Furthermore, the findings of this thesis shed light on the key factors contributing to flight delays in both the USA and Brazil. In Brazil, the departure month emerges as the most influential factor. Analyzing the distributions and proportions of delayed flights in Section 4.3.2, patterns become evident. Travelling during the months around December, including October, November, December, and January, carries a higher risk of experiencing flight delays. Conversely, the spring months of February, March, and April exhibit lower chances of delays. These insights suggest that airlines could allocate more staff and take additional precautions during the busy end-of-year months to mitigate flight postponements.

Another significant factor influencing flight delays in both the USA and Brazil is the departure hour. It is observed that flights scheduled for later hours in the day have a higher likelihood of experiencing delays, potentially due to delay propagation throughout the day. As a result, it is advisable for passengers to consider booking flights that depart early in the morning to minimize the risk of delays. This finding also implies that airlines should allocate additional staff resources during the later hours of the day to better manage and mitigate potential delays.

By considering these key factors and incorporating them into their operational strategies, airlines can optimize their scheduling, allocate resources more effectively, and improve overall on-time performance. Passengers can also make more informed decisions about their travel plans, taking into account the identified influential factors and selecting travel times and months with a lower likelihood of flight delays.

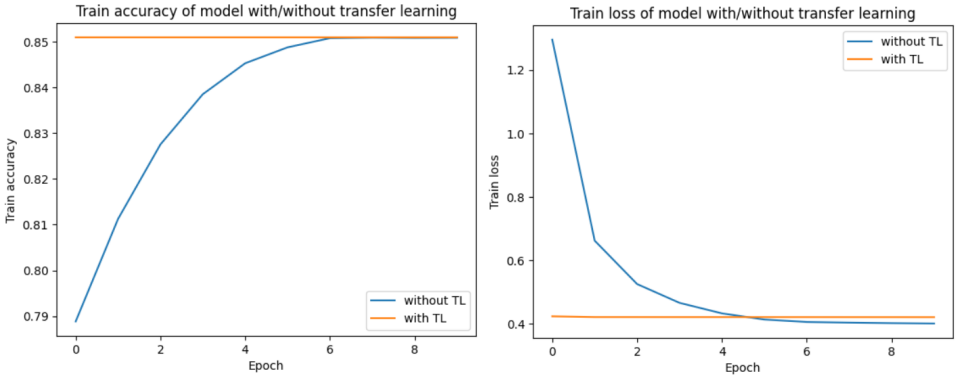


Figure 5.13: Comparison of train accuracy (left) and train loss (right) over time for a pre-trained model on the USA data set applied to the Brazilian data set and a model trained exclusively on the Brazilian data set.

5.5 Transfer Learning

The train accuracy and loss over time are depicted in Figure 5.13 for two scenarios: (1) a pre-trained sequential neural network initially trained on the United States data set and then applied to the Brazilian data set, and (2) a sequential neural network trained exclusively on the Brazilian data set. The left plot showcases the train accuracy of both the pre-trained model on the Brazilian data and the model solely trained on the Brazilian data. Meanwhile, the right plot illustrates the train loss of both models.

From the plots, it is evident that both models eventually end up with the same performance. However, the pre-trained model almost directly reaches the maximum level, while the fresh model needs around six epochs to catch up. While this might not seem like a significant difference, it has important implications, especially when dealing with smaller data sets.

The efficiency gains achieved through transfer learning are notable, as leveraging knowledge from a large flight records data set to another data set allows for faster convergence and reduced training time. This becomes increasingly crucial in practical applications where computational resources are limited. In today's context, where efficiency holds increasing importance over accuracy [50], this finding highlights the practical value of transfer learning in the field of flight delay prediction.

The optimal performance achieved by the Brazilian model pretrained on US data after a single epoch can be attributed to the transfer of general knowledge and shared patterns between the two datasets. The pretrained model has already learned generic features and representations from the

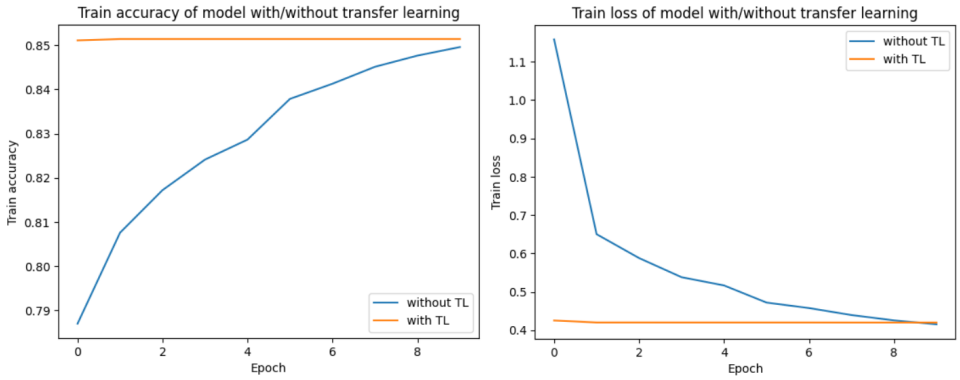


Figure 5.14: Comparison of train accuracy (left) and train loss (right) over time for a pre-trained model on 50% of the USA data set applied to 50% of the Brazilian data set and a model trained exclusively on 50% of the Brazilian data set.

US dataset, allowing it to quickly adapt and leverage this knowledge when training on the Brazilian dataset. This accelerates the learning process and enables the model to achieve optimal performance early on by focusing on learning task-specific information rather than generic features.

Additionally, the pre-trained model demonstrates the ability to transfer learned patterns and representations from one data set to another, showcasing the potential for generalization and adaptability. These insights emphasize the practical value of transfer learning in the context of flight delay prediction, offering a more efficient and scalable approach to model training and deployment.

The train accuracy and loss plots in Figure 5.14 and Figure 5.15 provide insights into the learning effect and the impact of transfer learning. With only 10% of the available data, transfer learning shows an improvement of approximately 3% in accuracy compared to the model without transfer learning. As the data set size increases, both models eventually converge to similar performance levels. However, it is worth noting that the model with transfer learning is approximately 10 times faster than the model without transfer learning when trained on 50% of the data set, and approximately 6 times faster when trained on the complete data set.

5.6 Future Research Directions

The field of predicting flight delays and cancellations is a dynamic area of research with ongoing developments. This thesis has identified important factors influencing flight postponements, but several promising avenues for

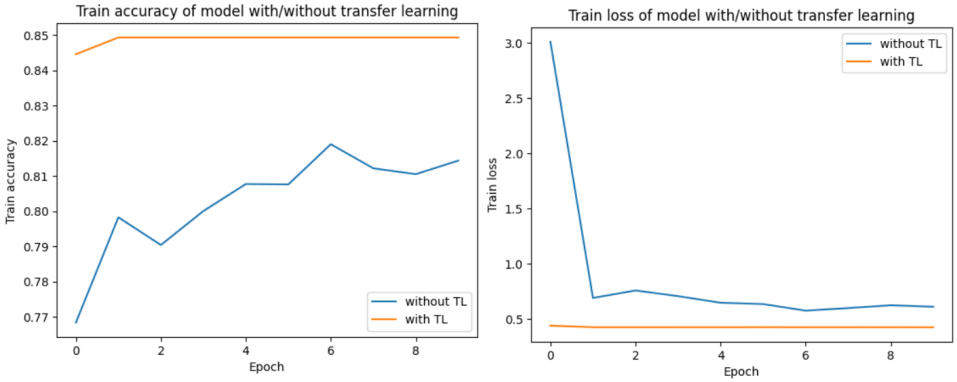


Figure 5.15: Comparison of train accuracy (left) and train loss (right) over time for a pre-trained model on 10% of the USA data set applied to 10% of the Brazilian data set and a model trained exclusively on 10% of the Brazilian data set.

future research can enhance the understanding of this complex phenomenon. The key areas for future investigation include the impact of COVID-19 on flight delays, the incorporation of additional data sets, such as weather data, and the exploration of advanced prediction techniques.

5.6.1 Influence of COVID-19

One significant area for future research is investigating the impact of the coronavirus disease on flight delays, particularly its effect on the key factors contributing to aircraft postponements. This can be achieved by comparing the most influential factors of flight postponements before, during, and after the COVID-19 outbreak. Conducting additional research in this context will provide a comprehensive understanding of how the coronavirus disease has influenced the aviation industry and shed light on the evolving dynamics of flight delays.

5.6.2 Data Sets

Another area of future research involves the incorporation of additional data sets to further enhance the prediction models for flight delays. While this thesis has not included weather data, it is recognized that weather conditions play a crucial role in flight operations and can have a significant impact on flight schedules. Numerous related works have shown the importance of weather data in predicting flight delays. Therefore, integrating weather data into the predictive models can provide a more comprehensive understanding of the factors influencing flight delays and improve the accuracy of delay predictions.

Furthermore, expanding the scope of the analysis to include data from other regions around the world, such as Europe, Asia, Oceania, and Africa, would be valuable. Investigating flight delays in these regions and comparing them to the findings from this thesis would shed light on the differences and similarities in factors influencing flight delays across various geographic areas. Additionally, utilizing transfer learning techniques to leverage knowledge gained from one region and apply it to another can provide insights into the transferability of predictive models and highlight regional-specific factors that contribute to flight delays.

5.6.3 Prediction Techniques

Lastly, future research can focus on exploring advanced prediction techniques to improve the accuracy and efficiency of flight delay predictions. Traditional machine learning algorithms have been widely used in this field, but emerging techniques, such as deep learning and ensemble methods, offer new possibilities for more sophisticated modelling and prediction. Investigating the application of these advanced techniques and comparing their performance with traditional approaches can lead to enhanced prediction capabilities and more accurate identification of critical factors impacting flight delays.

Chapter 6

Conclusions

In conclusion, this thesis aimed to identify the most influential factors in flight delays through the application of various machine learning algorithms to extensive datasets from the United States and Brazil. Techniques such as ordinal encoding and hyperparameter tuning were successfully employed to draw meaningful conclusions. The random forest and gradient boosting tree classifiers emerged as the top performers, achieving potential accuracies of 81% and 86% respectively.

While considering the trade-off between accuracy and the F_1 score, the findings of this thesis highlight the dominance of the departure month in Brazil and the departure hour in both countries as key factors in predicting flight delays, with flights scheduled for December and those departing later in the day showing a significantly higher likelihood of experiencing delays.

These findings offer valuable insights for both passengers and airlines. Passengers can leverage this knowledge to optimize their travel plans by selecting optimal departure times and days that minimize the risk of delays. Meanwhile, airlines can strategically allocate additional resources during peak delay-prone periods to enhance operational efficiency.

Additionally, transfer learning was employed to potentially enhance performance on the Brazilian data set using a pre-trained model from the United States data set, resulting in a efficiency gain of around 10 times with 50% of the data and approximately 6 times with the complete data. Despite no accuracy improvement, these findings highlight the practical value of transfer learning in enhancing efficiency for flight delay prediction.

Looking ahead, future research directions should include investigating the impact of COVID-19 on flight delays, incorporating additional datasets such as weather data, and exploring advanced prediction techniques.

Bibliography

[1] Herman Aguinis, Ryan K Gottfredson, and Harry Joo. Best-practice recommendations for defining, identifying, and handling outliers. *Organizational Research Methods*, 16(2):270–301, 2013.

[2] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[3] Poornima Balakrishna, Rajesh Ganesan, and Lance Sherry. Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of tampa bay departures. *Transportation Research Part C: Emerging Technologies*, 18(6):950–962, 2010.

[4] Poornima Balakrishna, Rajesh Ganesan, Lance Sherry, and Benjamin S Levy. Estimating taxi-out times with a reinforcement learning algorithm. In *2008 IEEE/AIAA 27th Digital Avionics Systems Conference*, pages 3–D. IEEE, 2008.

[5] Megan Baluch, Tristan Bergstra, and Mohamad El-Hajj. Complex analysis of united states flight data using a data mining approach. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1–6. IEEE, 2017.

[6] Chris M Bishop. Neural networks and their applications. *Review of scientific instruments*, 65(6):1803–1832, 1994.

[7] CAAC. Statistical bulletin of civil aviation industry development in 2020. Technical report, 2020.

[8] Leonardo Carvalho, Alice Sternberg, Leandro Maia Goncalves, Ana Beatriz Cruz, Jorge A Soares, Diego Brandão, Diego Carvalho, and Eduardo Ogasawara. On the relevance of data science for flight delay research: a systematic review. *Transport Reviews*, 41(4):499–528, 2021.

[9] CIRIUM. The on time performance review 2022. Technical report, 2023.

[10] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[11] Yi Ding. Predicting flight delay based on multiple linear regression. In *IOP conference series: Earth and environmental science*, volume 81, page 012198. IOP Publishing, 2017.

[12] Carsten F Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R García Marquéz, Bernd Gruber, Bruno Lafourcade, Pedro J Leitão, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

[13] Ehsan Esmaeilzadeh and Seyedmirsajad Mokhtarimousavi. Machine learning approach for flight departure delay prediction and analysis. *Transportation Research Record*, 2674(8):145–159, 2020.

[14] EUROCONTROL. Coda digest - all-causes delays to air transport in europe quarter 3 2022. Technical report, 2022.

[15] EUROCONTROL. 2022 - the year european aviation bounced back, despite war & omicron/covid. Technical report, 2023.

[16] Pablo Fleurquin, Bruno Campanelli, Victor M Eguiluz, and José J Ramasco. Trees of reactionary delay: Addressing the dynamical robustness of the us air transportation network. *transportation*, 11:12, 2014.

[17] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[18] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc.", 2022.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT press, 2016.

[20] Stefan Gössling and Nadja Schweiggart. Two years of covid-19 and tourism: What we learned, and what we should have learned. *Journal of Sustainable Tourism*, 30(4):915–931, 2022.

[21] Guan Gui, Fan Liu, Jinlong Sun, Jie Yang, Ziqi Zhou, and Dongxu Zhao. Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, 69(1):140–150, 2019.

[22] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques third edition. *University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University*, 2012.

[23] David J Hand. Principles of data mining. *Drug safety*, 30:621–622, 2007.

[24] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[25] Saed Hussain, Maizura Mokhtar, and Joe M Howe. Aircraft sensor estimation for fault tolerant flight control system using fully connected cascade neural network. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2013.

[26] Xuchu Jiang, Ying Zhang, Ying Li, and Biao Zhang. Forecast and analysis of aircraft passenger satisfaction based on rf-rfe-lr model. *Scientific Reports*, 12(1):11174, 2022.

[27] Sina Khanmohammadi, Salih Tutun, and Yunus Kucuk. A new multilevel input layer artificial neural network for predicting flight delays at jfk airport. *Procedia Computer Science*, 95:237–244, 2016.

[28] Sotiris B Kotsiantis, Ioannis Zaharakis, P Pintelas, et al. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160(1):3–24, 2007.

[29] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, page 179. Citeseer, 1997.

[30] Deepak Kulkarni, Yao Wang, and Banavar Sridhar. Data mining for understanding and improving decision-making affecting ground delay programs. In *2013 IEEE/AIAA 32nd Digital Avionics Systems Conference (DASC)*, pages 5B1–1. IEEE, 2013.

[31] Qiang Li and Ranzhe Jing. Generation and prediction of flight delays in air transport. *IET Intelligent Transport Systems*, 15(6):740–753, 2021.

[32] Qiang Li and Ranzhe Jing. Flight delay prediction from spatial and temporal perspective. *Expert Systems with Applications*, 205:117662, 2022.

[33] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9:381–386, 2020.

[34] Nicholas McCarthy, Mohammad Karzand, and Freddy Lecue. Amsterdam to dublin eventually delayed? lstm and transfer learning for predicting delays of low cost airlines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9541–9546, 2019.

[35] Daniele Micci-Barreca. A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 3(1):27–32, 2001.

[36] Roweida Mohammed, Jumanah Rawashdeh, and Malak Abdullah. Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)*, pages 243–248. IEEE, 2020.

[37] Leonardo Moreira, Christofer Dantas, Leonardo Oliveira, Jorge Soares, and Eduardo Ogasawara. On evaluating data preprocessing methods for machine learning models for flight delays. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2018.

[38] Godwell Nhamo, Kaitano Dube, David Chikodzi, Godwell Nhamo, Kaitano Dube, and David Chikodzi. Covid-19 and implications for the aviation sector: A global perspective. *Counting the cost of COVID-19 on the global tourism industry*, pages 89–107, 2020.

[39] Tri Noviantoro and Jen-Peng Huang. Investigating airline passenger satisfaction: Data mining method. *Research in Transportation Business & Management*, 43:100726, 2022.

[40] Daniel Alberto Pamplona, Li Weigang, Alexandre Gomes de Barros, Elcio Hideiti Shiguemori, and Claudio Jorge Pinto Alves. Supervised neural network with multilevel input layers for predicting of air traffic delays. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2018.

[41] Tamara Pejovic, Robert B Noland, Victoria Williams, and Ralf Toumi. A tentative analysis of the impacts of an airport closure. *Journal of Air Transport Management*, 15(5):241–248, 2009.

[42] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[43] Kedar Potdar, Taher S Pardawala, and Chinmay D Pai. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, 175(4):7–9, 2017.

[44] Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000.

[45] Susmita Ray. A quick review of machine learning algorithms. In *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pages 35–39. IEEE, 2019.

[46] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46, 2001.

[47] Álvaro Rodríguez-Sanz, Fernando Gómez Comendador, Rosa Arnaldo Valdés, Javier Pérez-Castán, Rocío Barragán Montes, and Sergio Cámara Serrano. Assessment of airport arrival congestion and delay: Prediction and reliability. *Transportation Research Part C: Emerging Technologies*, 98:255–283, 2019.

[48] Fei Rong, Li Qianya, Hu Bo, Zhang Jing, and Yang Dongdong. The prediction of flight delays based the analysis of random flight points. In *2015 34th Chinese Control Conference (CCC)*, pages 3992–3997. IEEE, 2015.

[49] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[50] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[51] Cedric Seger. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing, 2018.

[52] Wei Shao, Arian Prabowo, Sichen Zhao, Siyu Tan, Piotr Koniusz, Jeffrey Chan, Xinhong Hei, Bradley Feest, and Flora D Salim. Flight delay prediction using airport situational awareness map. In *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 432–435, 2019.

[53] Alice Sternberg, Diego Carvalho, Leonardo Murta, Jorge Soares, and Eduardo Ogasawara. An analysis of brazilian flight delays based on frequent patterns. *Transportation Research Part E: Logistics and Transportation Review*, 95:282–298, 2016.

[54] Alice Sternberg, Jorge Soares, Diego Carvalho, and Eduardo Ogasawara. A review on flight delay prediction. *arXiv preprint arXiv:1703.06118*, 2017.

[55] Xiaoqian Sun, Sebastian Wandelt, and Anming Zhang. How did covid-19 impact air transportation? a first peek through the lens of complex networks. *Journal of Air Transport Management*, 89:101928, 2020.

[56] Vladimir Svetnik, Andy Liaw, Christopher Tong, J Christopher Culberson, Robert P Sheridan, and Bradley P Feuston. Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of chemical information and computer sciences*, 43(6):1947–1958, 2003.

[57] Trefis Team. What is the impact of flight delays? 2016.

[58] Varsha Venkatesh, Arti Arya, Pooja Agarwal, S Lakshmi, and Sanjay Balana. Iterative machine and deep learning approach for aviation delay prediction. In *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, pages 562–567. IEEE, 2017.

[59] Alexander Von Eye and Clifford C Clogg. *Categorical variables in developmental research: Methods of analysis*. Elsevier, 1996.

[60] Yao Wang. Prediction of weather impacted airport capacity using ruc-2 forecast. In *2012 IEEE/AIAA 31st Digital Avionics Systems Conference (DASC)*, pages 3C3–1. IEEE, 2012.

[61] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016.

[62] Frederick Wieland. Limits to growth: results from the detailed policy assessment tool [air traffic congestion]. In *16th DASC. AIAA/IEEE Digital Avionics Systems Conference. Reflections to the Future. Proceedings*, volume 2, pages 9–2. IEEE, 1997.

[63] Guan Xiangmin and Ma Li. Departure capacity prediction for hub airport in thunderstorm based on data mining method. In *2017 29th Chinese Control And Decision Conference (CCDC)*, pages 6004–6009. IEEE, 2017.

[64] Maryam Farshchian Yazdi, Seyed Reza Kamel, Seyyed Javad Mahdavi Chabok, and Maryam Kheirabadi. Flight delay prediction based on deep learning and levenberg-marquart algorithm. *Journal of Big Data*, 7:1–28, 2020.

[65] Bojia Ye, Bo Liu, Yong Tian, and Lili Wan. A methodology for predicting aggregate flight departure delays in airports based on supervised learning. *Sustainability*, 12(7):2749, 2020.

[66] Xin-bin Zhao, Bin Li, and Cheng-guo Wang. There is a gold mine in flight data: A framework of data mining in civil aviation. *DEStech Transactions on Social Science, Education and Human Science*, 2017.

[67] Micha Zoutendijk and Mihaela Mitici. Probabilistic flight delay predictions using machine learning and applications to the flight-to-gate assignment problem. *Aerospace*, 8(6):152, 2021.

Appendix A

Data Exploration

This appendix contains supplementary figures to showcase parts of the data exploration phase in more detail.

A.1 Variable Distribution Histograms

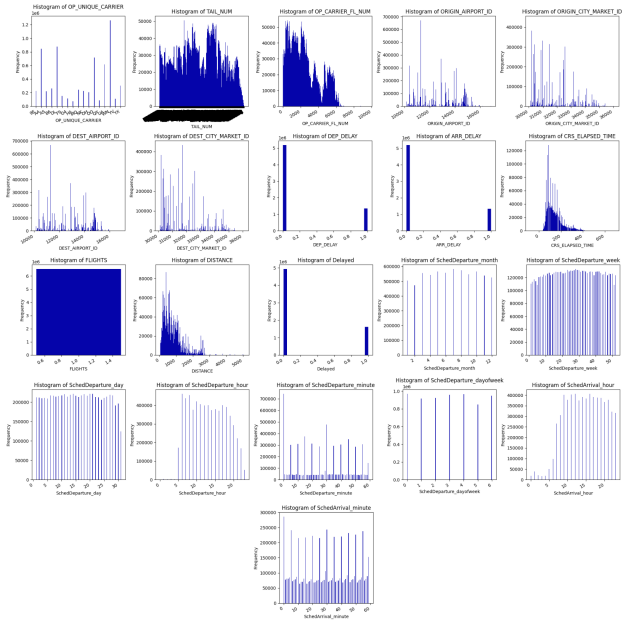


Figure A.1: Histograms for the distributions of each variable in the USA data set. It can be seen that most variables follow a normal distribution, with some exceptions.

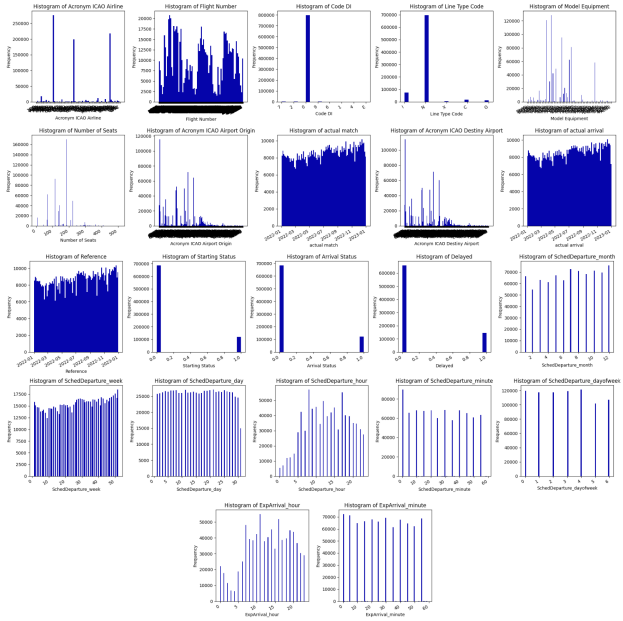


Figure A.2: Histograms for the distributions of each variable in the Brazilian data set. It can be seen that most variables follow a normal distribution, with some exceptions.

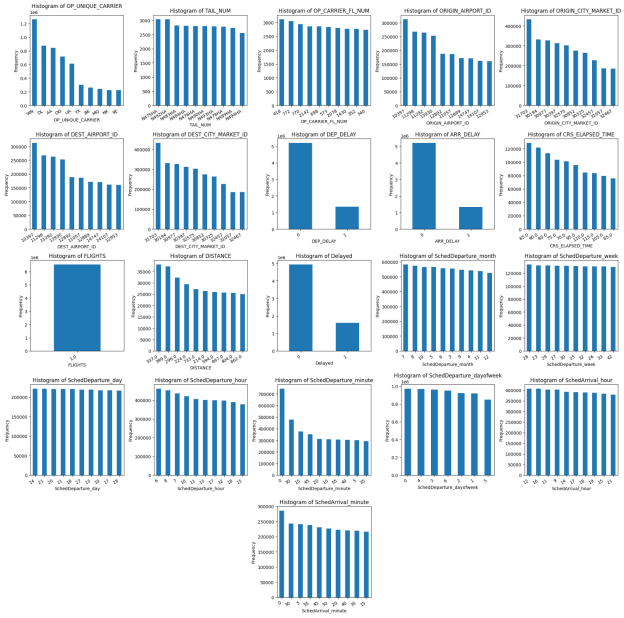


Figure A.3: Histograms that show the top 10 most occurring values of each variable in the USA data set. It is evident that some airlines and airports are observed to be more prominent in the data set.

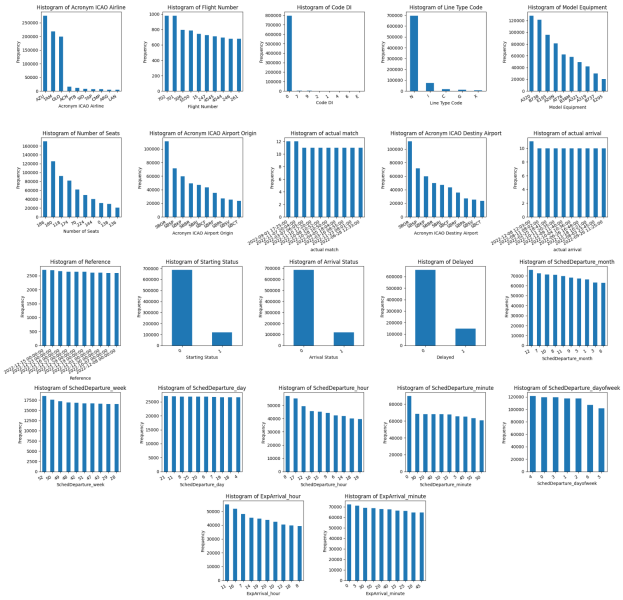


Figure A.4: Histograms that show the top 10 most occurring values of each variable in the Brazilian data set. It is evident that some airlines, airports, code DIs, and line type codes are observed to be more prominent in the data set.

A.2 Numerical Attributes Boxplots

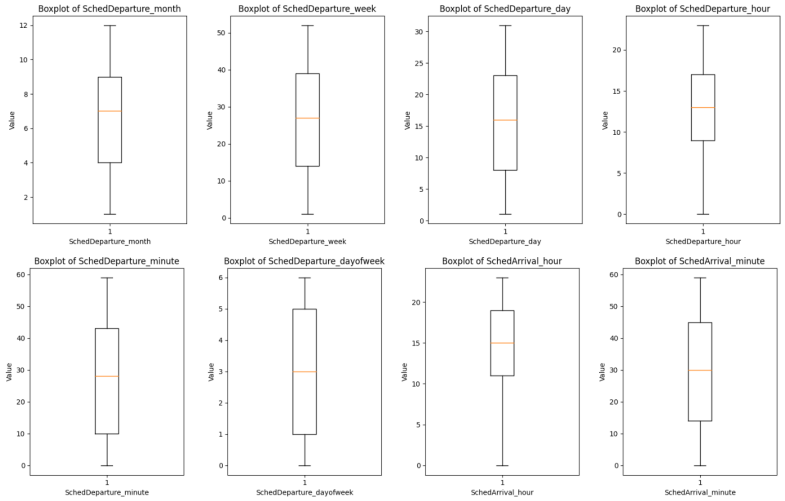


Figure A.5: Boxplots showing the distributions of numerical attributes related to the time period in the USA data set, where no boxplots indicate potential outliers.

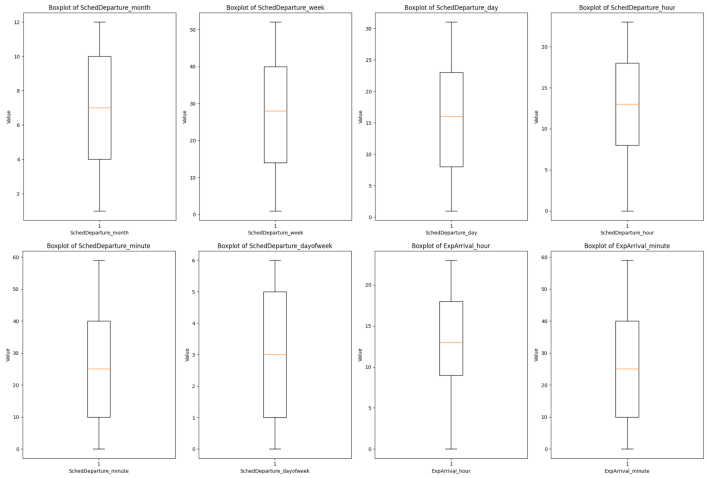


Figure A.6: Boxplots showing the distributions of numerical attributes related to the time period in the Brazilian data set, where no boxplots indicate potential outliers.

A.3 Delay Proportion Histograms

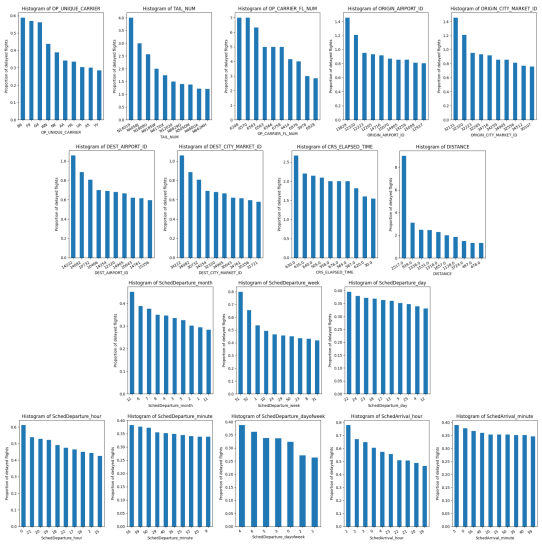


Figure A.7: Histograms showing the top 10 values with the highest proportional delay for each attribute in the USA data set. The plots show that some values experience a higher proportion of delays than others.

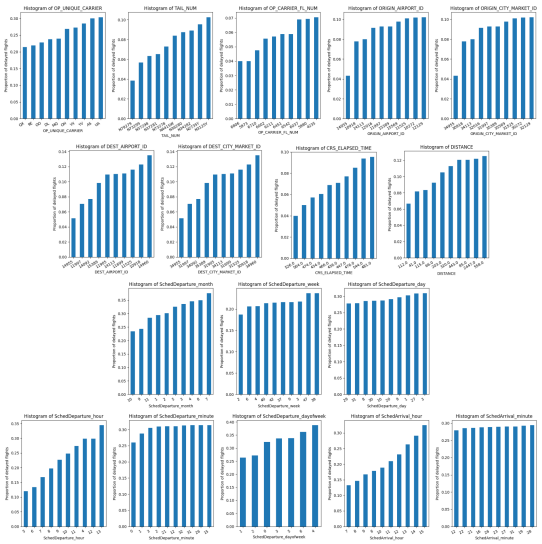


Figure A.8: Histograms showing the top 10 values with the lowest proportional delay for each attribute in the USA data set. The plots show that some values experience a lower proportion of delays than others.

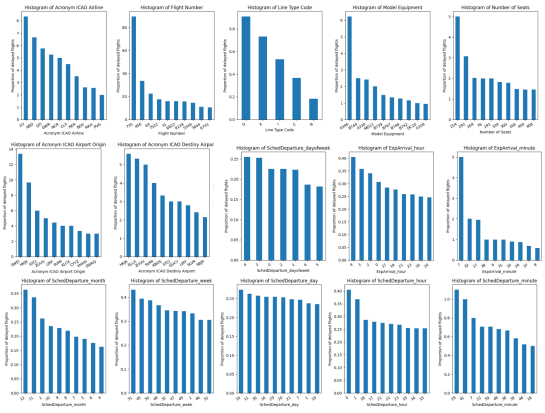


Figure A.9: Histograms showing the top 10 values with the highest proportional delay for each attribute in the Brazilian data set. The plots show that some values experience a higher proportion of delays than others.

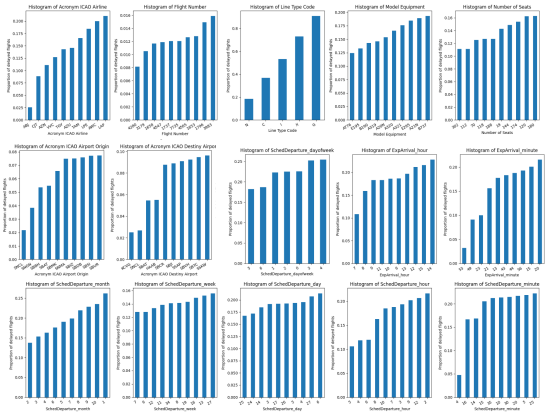
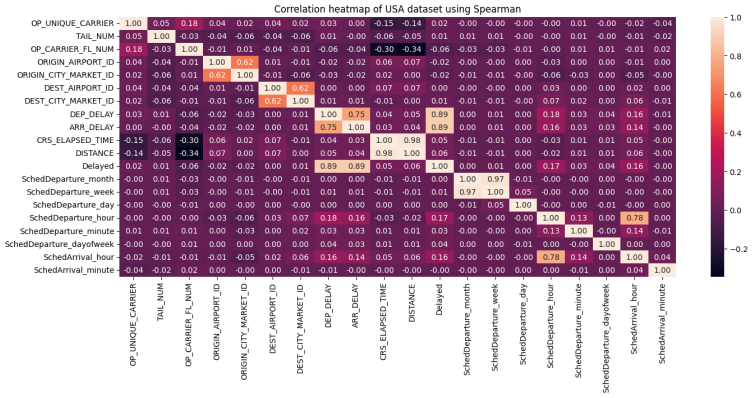
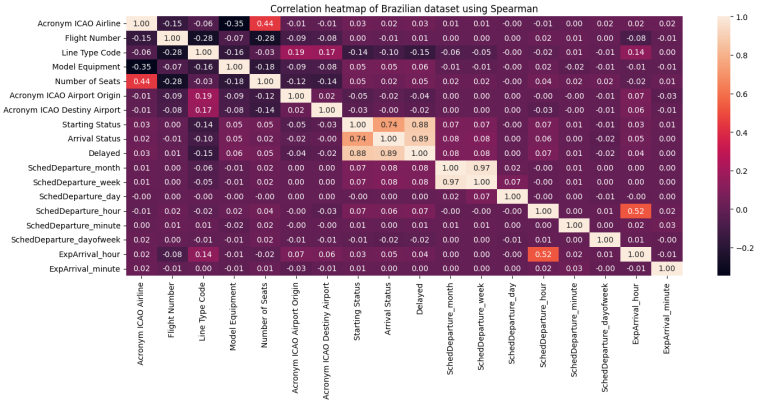


Figure A.10: Histograms showing the top 10 values with the lowest proportional delay for each attribute in the Brazilian data set. The plots show that some values experience a lower proportion of delays than others.

A.4 Correlation Heatmaps



(a)



(b)

Figure A.11: Spearman rank correlation heatmap for the USA data set (a) and the Brazilian data set (b). It can be seen that the response variables experience high collinearity, as well as the variables mentioned in this thesis.

A.5 Attribute Scatter Plots

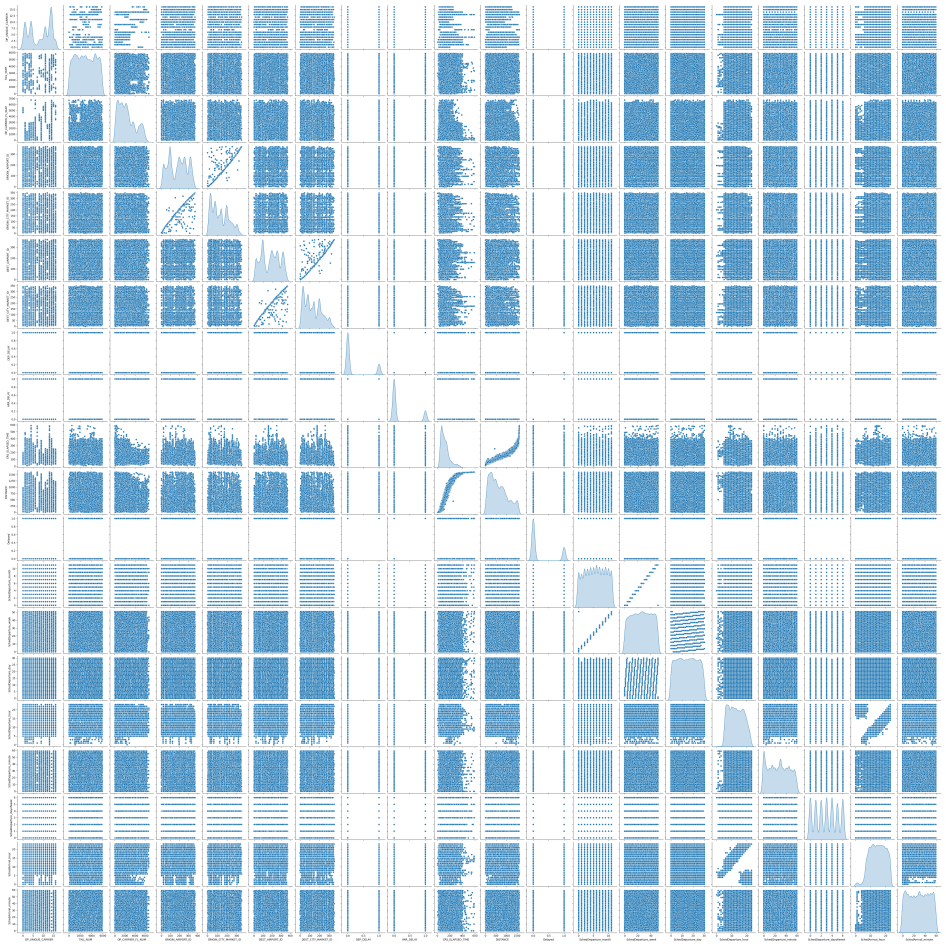


Figure A.12: Scatter plots showing the relationships between different variables in the United States data set. It can be seen that some attributes have a high correlation, since they have a linear or monotonic relationship.

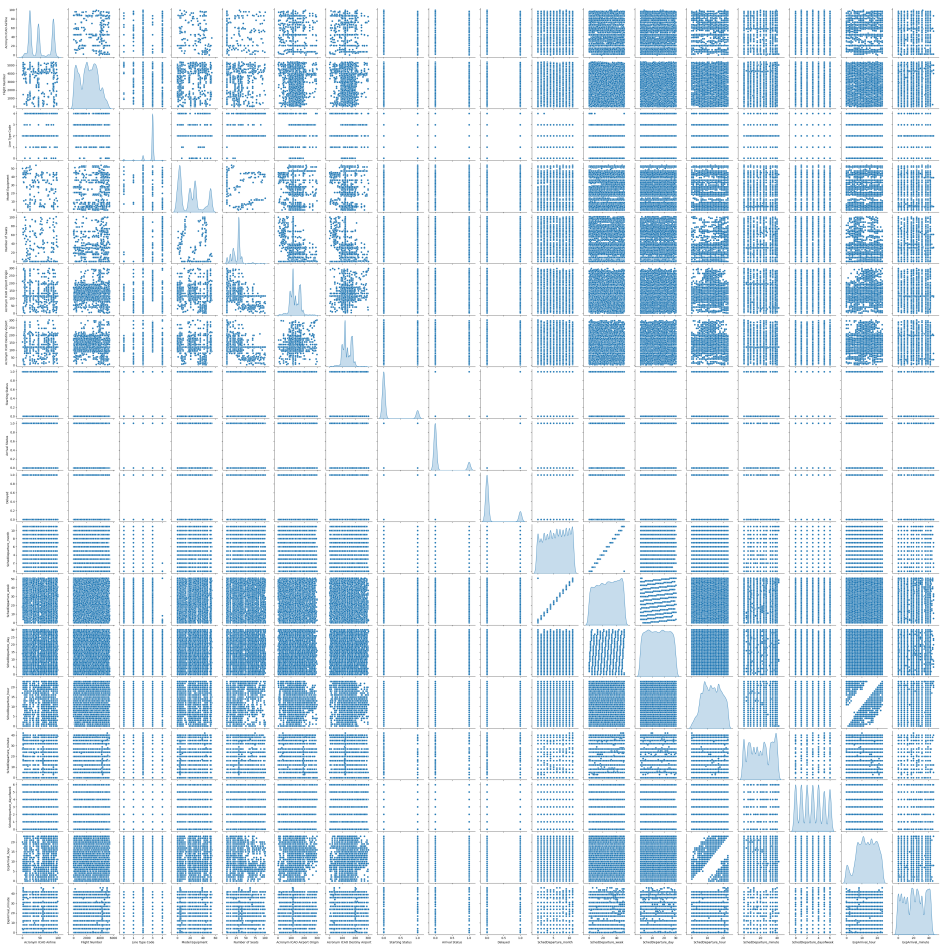


Figure A.13: Scatter plots showing the relationships between different variables in the Brazilian data set. It can be seen that some attributes have a high correlation, since they have a linear or monotonic relationship.

Appendix B

Machine Learning Models

This appendix contains supplementary figures and tables to showcase the performance evaluation of the machine learning models in more detail.

B.1 Target Encoding

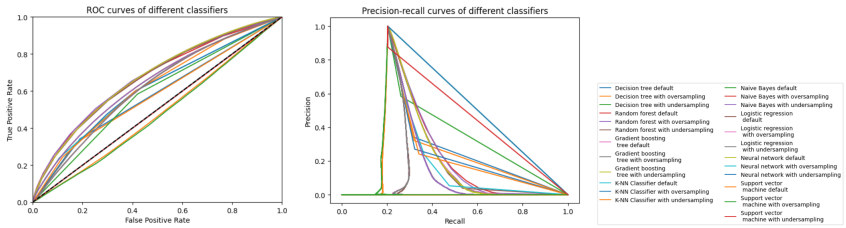


Figure B.1: ROC curves (left) and precision-recall curves (middle) of classifiers with different balancing techniques using target encoding on the USA data set, accompanied with their legend (right). It can be seen that these figures are similar to the ones when applying ordinal encoding.

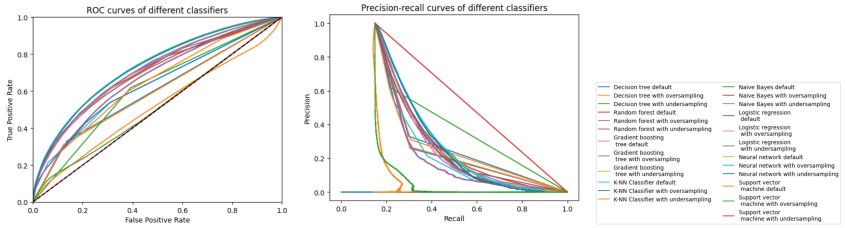


Figure B.2: ROC curves (left) and precision-recall curves (middle) of classifiers with different balancing techniques using target encoding on the Brazilian data set, accompanied with their legend (right). It can be seen that these figures are similar to the ones when applying ordinal encoding.

ML Algorithm	Accuracy	Precision	Recall	F_1 score	AUC score
DT	71.48%	57.19%	57.64%	57.38%	0.58
+ oversampling	72.23%	57.29%	57.32%	57.30%	0.57
+ undersampling	58.04%	55.34%	58.15%	52.46%	0.58
RF	79.48%	65.21%	55.68%	55.85%	0.67
+ oversampling	78.12%	63.10%	58.30%	59.31%	0.67
+ undersampling	66.76%	58.75%	62.30%	58.58%	0.67
GBT	79.72%	71.43%	50.48%	45.46%	0.67
+ oversampling	63.15%	58.40%	62.58%	56.91%	0.67
+ undersampling	63.17%	58.40%	62.58%	56.92%	0.67
k-NN	76.03%	59.28%	56.59%	57.19%	0.61
+ oversampling	65.69%	56.48%	58.86%	56.27%	0.61
+ undersampling	59.32%	56.20%	59.45%	53.64%	0.62
NB	79.55%	63.69%	51.15%	47.28%	0.64
+ oversampling	62.54%	56.94%	60.28%	55.60%	0.64
+ undersampling	62.54%	56.94%	60.27%	55.60%	0.64
LR	79.63%	39.81%	50.00%	44.33%	0.62
+ oversampling	58.89%	55.97%	59.10%	53.27%	0.62
+ undersampling	58.91%	55.97%	59.10%	53.29%	0.62
NN	79.63%	39.81%	50.00%	44.33%	0.50
+ oversampling	20.37%	10.19%	50.00%	16.93%	0.50
+ undersampling	20.37%	10.19%	50.00%	16.93%	0.50
SVM	35.65%	51.53%	51.72%	35.57%	0.48
+ oversampling	30.41%	51.65%	51.33%	30.24%	0.47
+ undersampling	50.29%	50.47%	50.72%	45.61%	0.50

Table B.1: Performance measures of classifiers with different balancing techniques using target encoding on the USA data set. These results also show similar results compared to the results from ordinal encoding.

ML Algorithm	Accuracy	Precision	Recall	F_1 score	AUC score
DT	78.32%	58.91%	59.69%	59.25%	0.60
+ oversampling	79.20%	59.29%	59.35%	59.32%	0.59
+ undersampling	60.94%	55.88%	61.30%	52.37%	0.61
RF	84.88%	68.08%	58.34%	60.19%	0.70
+ oversampling	82.95%	64.24%	60.74%	61.99%	0.70
+ undersampling	69.78%	58.89%	65.42%	58.55%	0.71
GBT	85.52%	74.81%	53.49%	52.97%	0.73
+ oversampling	68.67%	59.30%	66.72%	58.50%	0.73
+ undersampling	69.14%	59.38%	66.72%	58.76%	0.73
k-NN	83.02%	62.60%	57.59%	58.84%	0.63
+ oversampling	73.41%	57.48%	60.85%	58.02%	0.63
+ undersampling	61.31%	55.68%	60.85%	52.39%	0.64
NB	81.91%	60.37%	57.03%	57.98%	0.67
+ oversampling	80.49%	59.40%	57.92%	58.50%	0.67
+ undersampling	80.52%	59.42%	57.90%	58.50%	0.67
LR	85.39%	75.63%	52.37%	50.87%	0.69
+ oversampling	69.94%	58.66%	64.88%	58.40%	0.71
+ undersampling	70.54%	59.00%	65.33%	58.92%	0.71
NN	85.54%	72.88%	54.81%	55.26%	0.73
+ oversampling	71.23%	59.84%	66.78%	59.92%	0.73
+ undersampling	65.95%	58.81%	66.48%	56.91%	0.73
SVM	84.92%	55.24%	50.08%	46.24%	0.51
+ oversampling	84.44%	53.97%	50.33%	47.28%	0.51
+ undersampling	79.33%	52.72%	51.83%	51.82%	0.50

Table B.2: Performance measures of classifiers with different balancing techniques using target encoding on the Brazilian data set. These results also show similar results compared to the results from ordinal encoding.

B.2 Scaled Data

ML Algorithm	Accuracy	Precision	Recall	F_1 score	AUC score
DT	72.10%	58.05%	58.53%	58.26%	0.59
+ oversampling	72.73%	57.99%	57.99%	57.99%	0.58
+ undersampling	58.78%	55.83%	58.89%	53.14%	0.59
RF	79.25%	64.65%	56.26%	56.74%	0.67
+ oversampling	77.62%	62.47%	58.56%	59.52%	0.67
+ undersampling	66.61%	58.79%	62.42%	58.57%	0.67
GBT	79.63%	69.44%	50.01%	44.35%	0.66
+ oversampling	59.90%	57.54%	61.54%	54.77%	0.66
+ undersampling	59.94%	57.60%	61.64%	54.83%	0.66
k-NN	76.14%	59.18%	56.32%	56.89%	0.61
+ oversampling	65.38%	56.25%	58.58%	55.99%	0.61
+ undersampling	58.91%	56.12%	59.35%	53.38%	0.62
NB	79.63%	39.81%	50.00%	44.33%	0.62
+ oversampling	57.57%	55.97%	59.17%	52.61%	0.62
+ undersampling	57.60%	55.97%	59.17%	52.62%	0.62
LR	79.63%	39.81%	50.00%	44.33%	0.62
+ oversampling	59.03%	56.04%	59.20%	53.39%	0.62
+ undersampling	59.02%	56.04%	59.21%	53.39%	0.62
NN	79.98%	69.96%	52.25%	49.36%	0.69
+ oversampling	64.12%	59.31%	63.92%	57.98%	0.69
+ undersampling	62.81%	58.84%	63.36%	57.05%	0.69
SVM	45.54%	51.28%	51.90%	43.27%	0.53
+ oversampling	57.50%	53.01%	54.53%	50.62%	0.50
+ undersampling	35.09%	47.09%	46.31%	34.61%	0.55

Table B.3: Performance measures of classifiers with different balancing techniques using ordinal encoding on the scaled USA data set. These results show similar results compared to the results with unscaled data.

ML Algorithm	Accuracy	Precision	Recall	F_1 score	AUC score
DT	78.56%	59.35%	60.17%	59.71%	0.60
+ oversampling	79.39%	59.68%	59.76%	59.72%	0.60
+ undersampling	61.18%	55.94%	61.39%	52.53%	0.61
RF	85.03%	68.69%	58.41%	60.31%	0.70
+ oversampling	83.00%	64.36%	60.81%	62.09%	0.70
+ undersampling	69.71%	58.80%	65.26%	58.45%	0.71
GBT	85.11%	75.00%	50.60%	47.28%	0.69
+ oversampling	67.66%	57.92%	64.16%	56.92%	0.70
+ undersampling	67.63%	57.87%	64.06%	56.86%	0.70
k-NN	83.32%	63.71%	58.41%	59.82%	0.65
+ oversampling	73.27%	57.72%	61.38%	58.26%	0.64
+ undersampling	62.22%	56.57%	62.53%	53.48%	0.66
NB	82.00%	58.23%	54.55%	55.10%	0.64
+ oversampling	75.49%	56.82%	58.49%	57.33%	0.64
+ undersampling	75.41%	56.90%	58.68%	57.44%	0.64
LR	85.00%	59.73%	50.04%	46.08%	0.63
+ oversampling	61.07%	55.14%	59.79%	51.88%	0.63
+ undersampling	60.98%	55.14%	59.79%	51.83%	0.63
NN	85.34%	71.86%	53.49%	53.03%	0.71
+ oversampling	65.41%	58.06%	65.04%	56.09%	0.71
+ undersampling	66.35%	58.01%	64.72%	56.46%	0.70
SVM	50.68%	50.63%	51.24%	43.76%	0.48
+ oversampling	50.88%	50.87%	51.71%	44.01%	0.47
+ undersampling	49.04%	52.71%	55.27%	44.07%	0.42

Table B.4: Performance measures of classifiers with different balancing techniques using ordinal encoding on the scaled Brazilian data set. These results show similar results compared to the results with unscaled data.

B.3 Hyperparameter Tuning

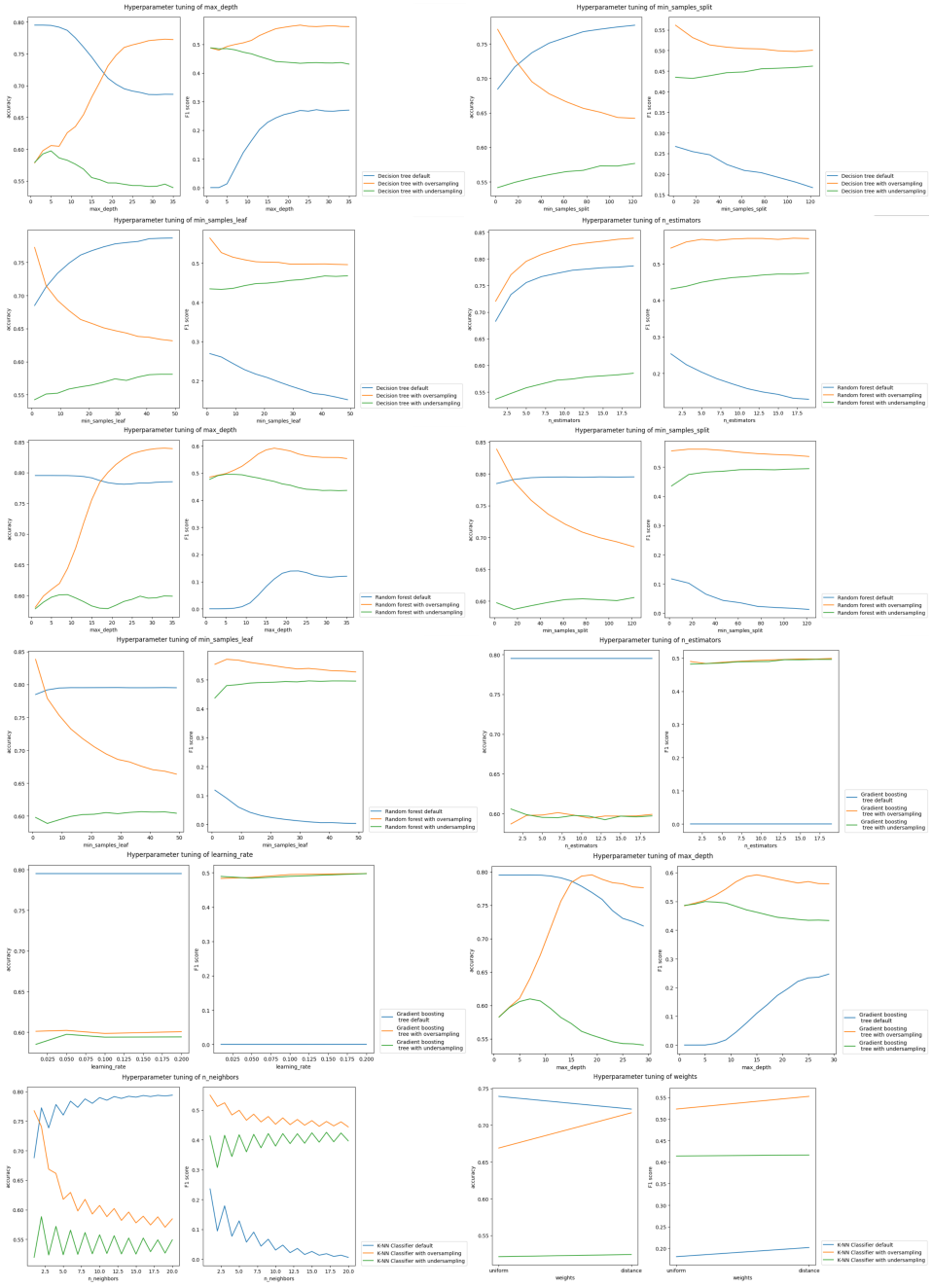


Figure B.3: Plots illustrating the impact of varying parameter values on the accuracy and F_1 score of different machine learning models applied to the USA data set. The optimal parameter values are chosen based on the highest F_1 score/accuracy at the point where the performance seems to stagnate.

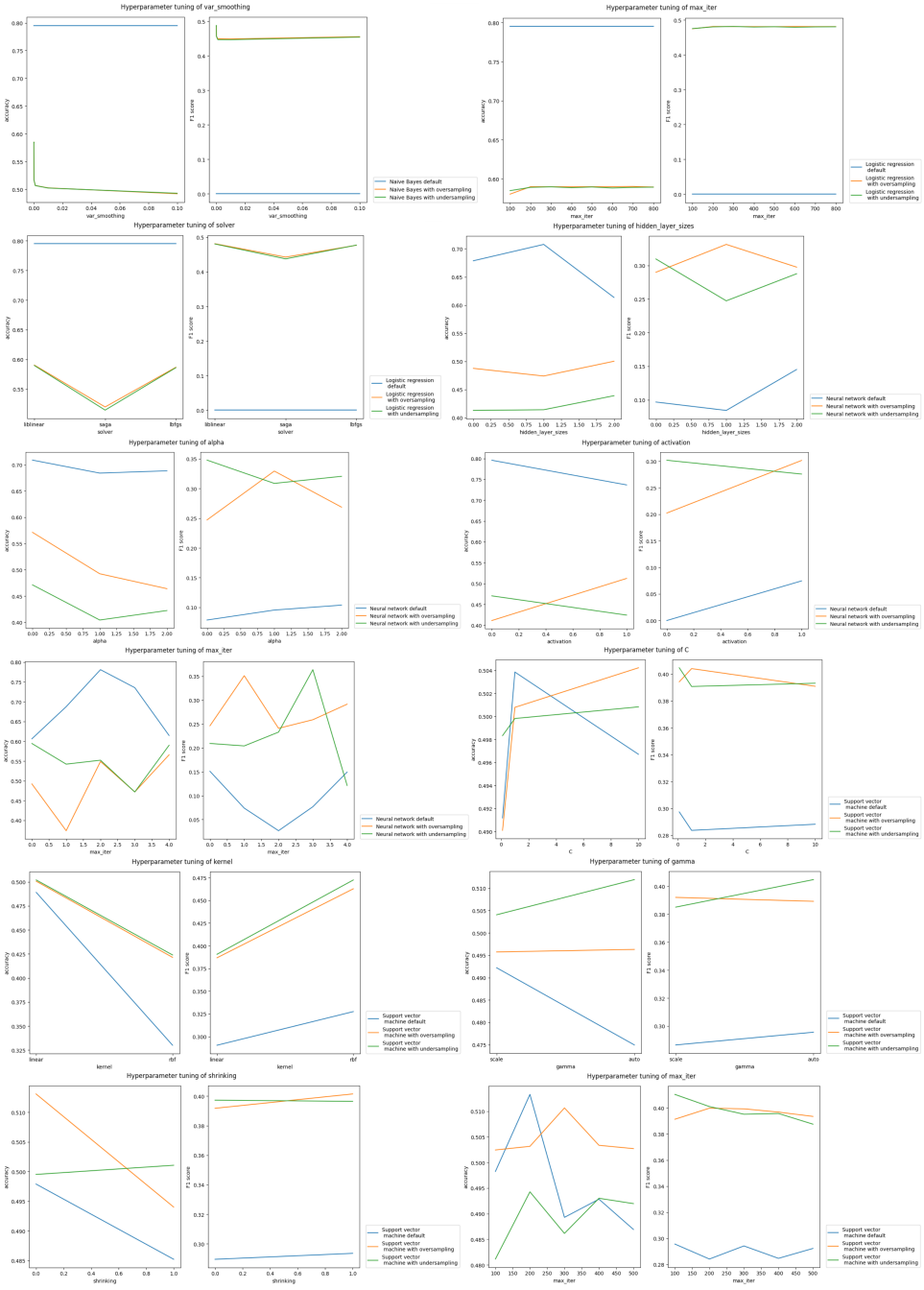


Figure B.4: Plots illustrating the impact of varying parameter values on the accuracy and F_1 score of different machine learning models applied to the USA data set. The optimal parameter values are chosen based on the highest F_1 score/accuracy at the point where the performance seems to stagnate.

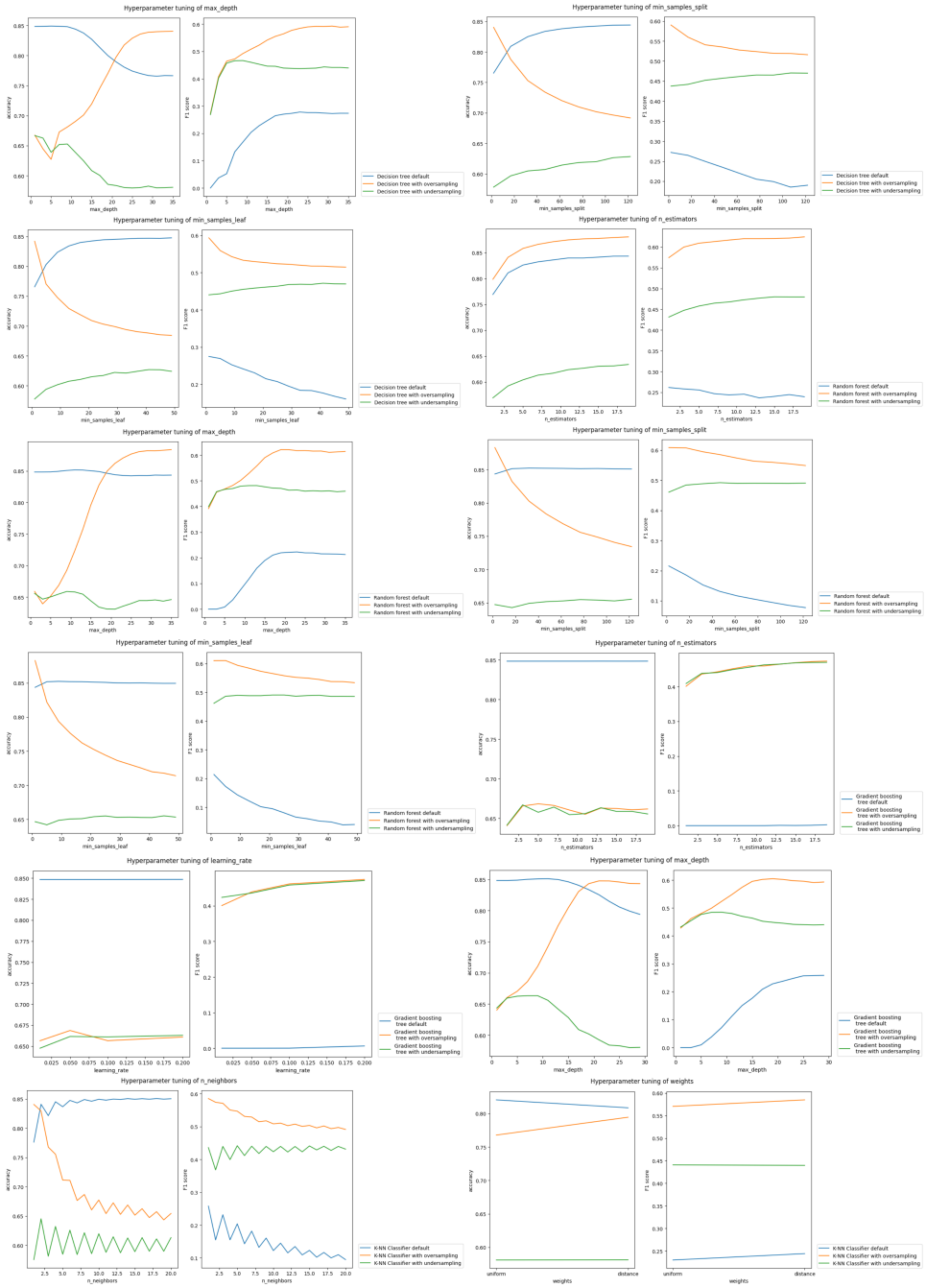


Figure B.5: Plots illustrating the impact of varying parameter values on the accuracy and F_1 score of different machine learning models applied to the Brazilian data set. The optimal parameter values are chosen based on the highest F_1 score/accuracy at the point where the performance seems to stagnate.

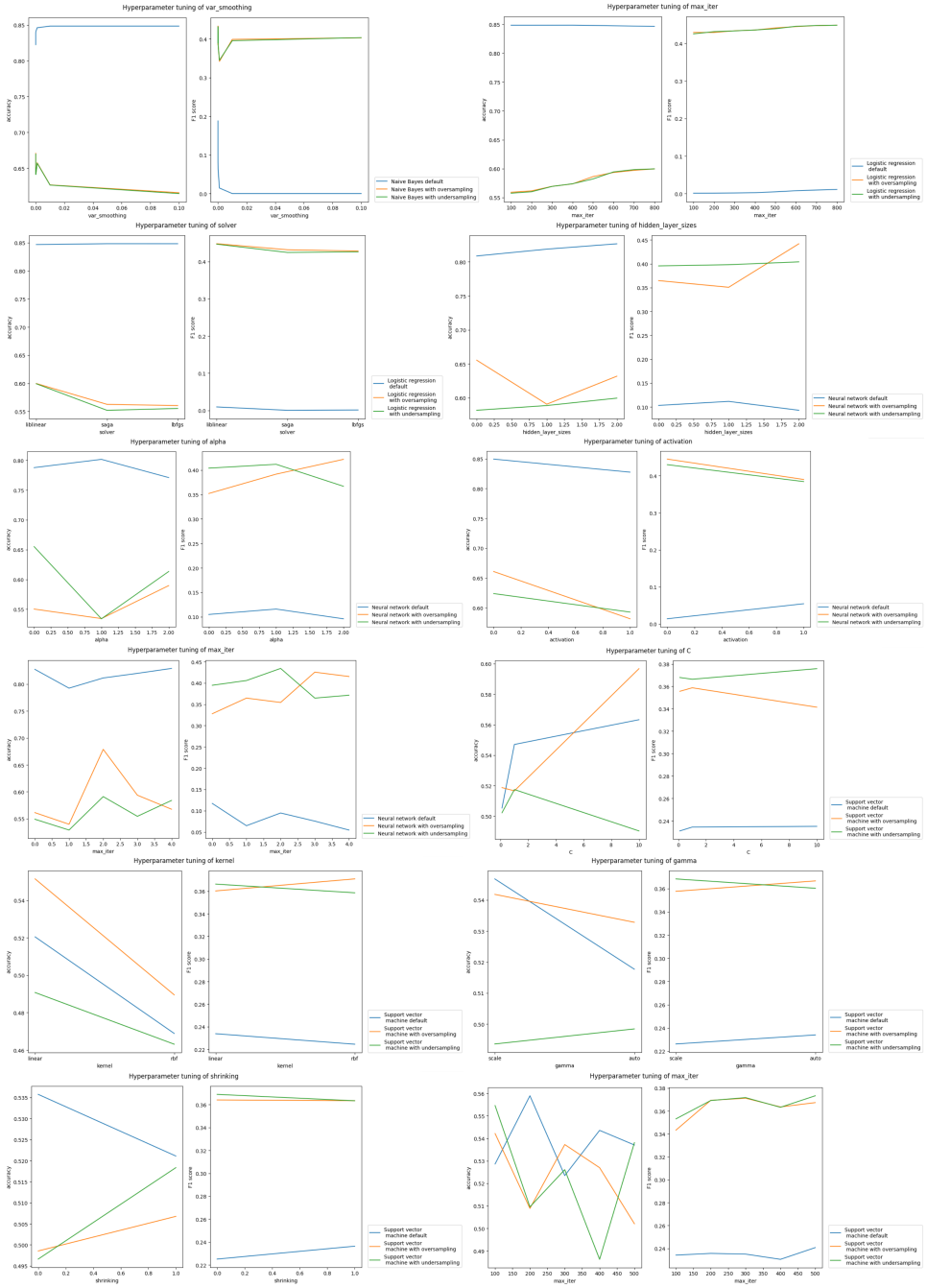


Figure B.6: Plots illustrating the impact of varying parameter values on the accuracy and F_1 score of different machine learning models applied to the Brazilian data set. The optimal parameter values are chosen based on the highest F_1 score/accuracy at the point where the performance seems to stagnate.