

**MARMARA UNIVERSITY**

**FACULTY OF TECHNOLOGY**

**DEPARTMENT OF COMPUTER ENGINEERING**

**COMPLETION PROJECT**

Flight Cancellation Delay Prediction and Smart Flight Ticketing System with Machine Learning

**PROJECT AUTHOR**

Eda Nur Mutlu - 170421843

Nuri Can Birdemir - 171421013

**ADVISOR**

Prof. Dr. Lecturer. Member EYUP EMRE ÜLKÜ

**PROVINCE, THESIS YEAR**

**İstanbul, 2025**



**MARMARA UNIVERSITY**

**FACULTY OF TECHNOLOGY**

**DEPARTMENT OF COMPUTER ENGINEERING**

**COMPLETION PROJECT**

Flight Cancellation Delay Prediction and Smart Flight Ticketing System with Machine Learning

**PROJECT AUTHOR**

Eda Nur Mutlu - 170421843

Nuri Can Birdemir - 171421013

**ADVISOR**

Prof. Dr. Lecturer. Member EYUP EMRE ÜLKÜ

**PROVINCE, THESIS YEAR**

**İstanbul, 2025**

**CONTENTS**

Page

[CHAPTER 1. INTRODUCTION 8](#_Toc196668572)

[1.1. Purpose and Importance of Project Work 10](#_Toc196668573)

[1.2. Literature Review 11](#_Toc196668574)

[CHAPTER 2 – MATERIALS AND METHODS 14](#_Toc196668575)

[2.1. Data Set 15](#_Toc196668576)

[2.2. Data Preproccessing 20](#_Toc196668577)

[2.3. Supervised Learning Models 23](#_Toc196668578)

[CHAPTER 3 – FINDINGS AND DISCUSSION 28](#_Toc196668579)

[3.1. Flight Cancellation Forecasting Model Results 28](#_Toc196668580)

[3.2. Cancellation Code Forecasting Model Results 28](#_Toc196668581)

[3.3. Flight Delay Estimation Model Results 29](#_Toc196668582)

[3.4. Model Optimization and Evulation Techniques 31](#_Toc196668583)

[3.4.1. Cross-validation Results 31](#_Toc196668584)

[3.4.2. Training and Test Accuracy Comparison 31](#_Toc196668585)

[3.4.3. XGBoost Regularization Optimization 32](#_Toc196668586)

[3.5.Literature Review and Comparative Analysis of Project Success Values 32](#_Toc196668587)

[3.5.1. Results of the Success Values of the Developed Model 32](#_Toc196668588)

[3.5.2. Comparison with the Success Values of Similar Projects in the Literature 35](#_Toc196668589)

[CHAPTER 4 – CONCLUSIONS 40](#_Toc196668590)

[CHAPTER 5 – FUTURE WORK 41](#_Toc196668591)

[SOURCES 42](#_Toc196668592)

**FIGURE LIST**

Page

[**Figure 1** Delay Status Graph](#_heading=h.1fob9te) 17

[**Figure 2** Chart of Delay Status by Airline](#_heading=h.4d34og8) 18

[**Figure**](#_heading=h.3rdcrjn) **3** [Delay Factor Distribution](#_heading=h.3rdcrjn) 19

[**Figure 4** Heatmap of Delay Factors](#_heading=h.lnxbz9) 19

[**Figure**](#_heading=h.44sinio) **5** [Flight Cancellation Code Prediction Model Feature Importance](#_heading=h.44sinio) 20

[**Figure**](#_heading=h.44sinio) **6** [Flight Cancellation Prediction Model Feature Importance](#_heading=h.44sinio) 26

[**Figure**](#_heading=h.44sinio) **7** [Flight Cancellation Code Prediction Model Feature Importance](#_heading=h.44sinio) 26

[**Figure**](#_heading=h.lnxbz9) **8** [Flight Delay Prediction Model Feature Importance](#_heading=h.lnxbz9) 27

**Figure 9** Flight Delay Prediction Model Complexity Matrix 30

**Figure 10** Performance Results of Khaksar and Sheikholeslami Study 35

**Figure 11** Performance Results of M.Kurt's Study 36

**Figure 12** Performance Results of Y.Yanying's Study 37

**Figure 13** Flight Delay Model of Giarmas' Study Performance Results 38

**Figure 14** Flight Cancellation Model Performance Results of Giarmas' Study 38

**Figure 15** Performance Results of Ansari's Study 39

**Figure 16** Mass Flight Cancellation Prediction Model Performance Comparison 40

**Figure 17** Batch Flight Delay Prediction Model Performance Comparison 40

**TABLE LIST**

Page

[**Table 1** Structure of the Original Data Set](#_heading=h.17dp8vu) 16

[**Table**](#_heading=h.35nkun2) **2** [Structure of the Model Training Dataset](#_heading=h.35nkun2)  21

**Table 3** Cancel Code Prediction Model Output Classes and Description 29

**Table 4** Flight Cancellation Prediction Model Performance Results 32

**Table 5** Cancellation Code Prediction Model Performance Results 33

**Table 6** Flight Delay Prediction Model Performance Results 34

**ABSTRACT**

This study aims to develop a machine learning model capable of predicting flight cancellations and delays. In air transportation, unexpected cancellations and delays cause significant problems for both passengers and airline companies. In this context, prediction models have been created using historical flight data, weather information, and flight details.

Initially, open flight data provided by the U.S. Department of Transportation (DOT) was examined, and relevant datasets were compiled. Various data preprocessing techniques were applied, including data cleaning, handling missing values, adjusting time formats, and feature engineering.

During the machine learning phase, XGBoost, Random Forest, Decision Tree, KNN, and Gradient Boosting algorithms were used to develop prediction models. To improve model performance, the SMOTE method was applied to balance the dataset, and evaluation metrics such as accuracy, precision, recall, f1-score, and ROC-AUC were calculated. Finally, the effectiveness of the developed model in predicting flight cancellations and delays was assessed.

As a result of this study, a successful model capable of predicting flight cancellations and delays was created. Future work includes integrating this model into a web platform to assist passengers in making informed flight booking decisions.

**ABSTRACT**

This study aims to develop a machine learning model capable of predicting flight cancellations and delays. In air transportation, unexpected cancellations and delays cause significant problems for both passengers and airline companies. In this context, prediction models have been created using historical flight data, weather information, and flight details.

Initially, open flight data provided by the U.S. Department of Transportation (DOT) was examined, and relevant datasets were compiled. Various data preprocessing techniques were applied, including data cleaning, handling missing values, adjusting time formats, and feature engineering.

During the machine learning phase, XGBoost, Random Forest, Decision Tree, KNN, and Gradient Boosting algorithms were used to develop prediction models. To improve model performance, the SMOTE method was applied to balance the dataset, and evaluation metrics such as accuracy, precision, recall, f1-score, and ROC-AUC were calculated. Finally, the effectiveness of the developed model in predicting flight cancellations and delays was assessed.

As a result of this study, a successful model capable of predicting flight cancellations and delays was created. Future work includes integrating this model into a web platform to assist passengers in making informed flight booking decisions.

# CHAPTER 1. INTRODUCTION

Although airline transportation is a rapidly growing industry on a global scale, flight delays remain one of the most important problems faced by the industry. Flight delays directly affect passenger satisfaction and cause economic losses to airlines and airports due to operational disruptions. In addition, delays congest airport traffic, causing chain delays and complicating airspace management [1].

According to data published by the US Department of Transportation in 2023, only 78% of domestic flights of major airlines in the US were able to depart on time. In other words, approximately 22% of flights departed 15 minutes or more later than scheduled [3].

Similarly, flight delays are a serious problem in the UK. In 2023, 34% of passengers traveling from the UK experienced flight delays or cancellations. In total, 45 million passengers were affected by such disruptions, with 3.8 million passengers having their flights canceled altogether. London Gatwick Airport in particular stood out as the airport with the lowest on-time performance with a 42% delay rate [4]. This situation reveals that flight delays are still a widespread problem in the sector that needs to be solved.

The large fluctuations in the demand for airline transportation during the pandemic period have once again highlighted the importance of analysis and forecasting studies for flight delays and cancellations. In order to solve these complex problems, it has become inevitable to go beyond traditional approaches and use advanced technologies such as artificial intelligence and machine learning. Intelligent systems to be developed in this context will provide strategic benefits that will both increase the individual satisfaction of passengers and improve the operational efficiency of airline companies.

In this study, a machine learning-based model that can predict flight delays and cancelations has been developed. One of the most important innovative aspects of the model is the integration of meteorological data with flight information, which is often ignored in traditional studies. In particular, by incorporating weather parameters (temperature, precipitation, wind direction/speed, etc.) obtained through the Meteostat API into the model, the accuracy and reliability of the forecasts are significantly improved. While only historical flight data or operational variables are usually used in the literature, this study presents a more holistic and realistic approach by modeling the impact of environmental conditions in detail. Thanks to this unique integration, both passenger satisfaction-oriented and operational efficiency-oriented predictions have been achieved.

The developed model aims to improve the user experience by integrating flight delay predictions directly into the ticket sales platform. When passengers enter their flight information, they will be able to receive predictions about the delay risk of their flight and optimize their travel planning accordingly. Value will be added to passengers by providing alternative flight suggestions, ticket price analysis and better planning options based on delay risk. At the same time, airlines will be able to make their operational processes more efficient, better manage sources of delay and minimize revenue losses.

Machine learning methods are a powerful tool for extracting meaningful information from large and complex data sets [5]. In this study, XGBClassifier algorithm is preferred to predict flight delays and cancelations. XGBClassifier is characterized by its high accuracy and generalizability [6]. In addition, the flexible nature of this algorithm allows the model to be continuously updated with new data, improving its future performance and providing a dynamic solution. Compared to traditional statistical methods, XGBClassifier performs better in multifactor problems such as flight delays and cancellations due to its ability to model more complex relationships.

The study aims to create value for three main target audiences:

1. Passengers: Receive early warnings about the risk of delays and will be able to plan their trips more consciously and access alternative flight options.
2. Airline Companies: By better analyzing the sources of delays, they will be able to optimize their operational processes and increase customer satisfaction.
3. Airport Management and Civil Aviation Authorities: They will be able to make flight planning more efficient based on delay forecasts.

As a result, this study utilizes the innovative opportunities provided by machine learning technologies in the search for solutions to flight delays. The study aims to provide a model that will contribute to global airline transportation and respond to the operational challenges [8], especially after the pandemic period.

## Purpose and Importance of Project Work

The main objective of this study is to develop a machine learning model that can predict flight delays, cancellations and reasons for cancellations in airline transportation. Within the scope of the study, a large number of factors affecting flight operations are systematically analyzed and the complex relationships between these factors are modeled through machine learning algorithms. In particular, meteorological parameters obtained from Meteostat API are integrated with flight data in order to investigate the impact of weather data in depth. In this way, the accuracy and generalizability of the prediction model have been significantly improved and a decision system that can both increase passenger satisfaction and contribute to the operational decision-making processes of airline companies has been created.

In the technical aspect of the study, the performances of different machine learning algorithms were evaluated comparatively. In this context, XGBClassifier, Random Forest Classifier, Decision Tree Classifier, KNN Classifier and Gradient Boosting Classifier algorithms were used. As a result of the performance comparison of the algorithms, the model with the highest success was selected and optimization studies were carried out on this model. Preliminary evaluations showed that the XGBClassifier algorithm can effectively learn patterns in complex data sets and quickly adapt to new data. The developed model was evaluated with various performance metrics to continuously improve the prediction accuracy.

In the practical application dimension of the study, the developed prediction model was integrated into ticket sales platforms. Through this integration, delay risk assessment is provided to passengers at the flight selection stage, alternative flight routes are suggested and dynamic pricing strategies based on delay risk are developed. While this approach enables passengers to make their travel plans more consciously, it is also planned to contribute to airline companies to increase customer satisfaction.

For airlines, the study aims to create a decision support system to improve operational efficiency. By analyzing the root causes of delays, this system will enable early interventions to optimize operational costs and prevent delays. Thus, airline companies will be able to improve their operational processes and gain competitive advantage by making data-driven decisions.

This study makes concrete contributions to the academic literature, especially in areas such as the integration of meteorological data into flight forecasting models and the prediction of multi-class cancellation reasons. Weather variables, which are often ignored in the literature, are extensively treated in this project and evaluated in a way that directly affects model performance. Moreover, with its ability to classify the reasons for cancellation, the study paves the way for artificial intelligence-based systems that can predict not only the presence of a delay but also its cause. From an industry perspective, the developed model enables airline companies to improve their delay risk analysis, provide more transparent information to passengers and use it as a decision support system in flight planning. In these respects, the study is a strong reference for both academic research and digital transformation applications in the aviation industry.

## Literature Review

Today, the rapidly increasing demand for air transportation has made the analysis of flight delays an important area of research [2]. Researchers widely utilize machine learning and data mining techniques to predict flight delays. Studies in the literature have generally focused on factors such as the location of airport facilities, weather and airport capacity. Machine learning techniques make significant contributions in this field by enabling the storage and processing of large-scale data sets [9]. However, most of the existing research has focused on a specific geographical region or a limited number of factors, and has been lacking in the development of holistic approaches to delay prediction.

This study aims to provide a new perspective on forecasting models for flight delays. In order to fill the gap in the literature, a comprehensive dataset integrating various factors such as flight data as well as meteorological data is used. In particular, the use of comprehensive meteorological data sources such as Meteostat allows for more precise modeling of weather impacts. As emphasized in Delahaye and Puechmorel's study, weather factors, which are often neglected or superficially addressed in the literature, are analyzed in depth in this study to increase the accuracy and reliability of the model by analyzing the impact of weather factors on flight delays. [10].

As Chawla et al. have shown, the problem of data imbalance is an important issue in the literature due to the fact that delays and cancelations contain rare observations [11]. This study aims to develop more balanced, generalizable and highly accurate forecasting models by addressing the data imbalance problem in a systematic way.

In their study, Khaksar and Sheikholeslami apply various machine learning algorithms to predict airline delays [12]. Methods such as Decision Trees, Random Forest, Clustering and Bayesian classification were tested on the data obtained from the US and Iran flight networks and focused on predicting the causes of delays more accurately. In particular, it was reported that visibility and wind speed significantly affected the delays in the US flight network, while in the Iranian flight network, factors such as fleet age and aircraft type came to the fore. In the predictions obtained, a hybrid classification with Decision Trees and Clustering methods was used and an accuracy rate of approximately 70% was achieved.

Bojia Ye et al. propose a methodology using supervised learning methods to predict flight delays at airports [13]. In the study, operational flight data and weather information from Nanjing Lukou International Airport are processed and four types of airport-related features are generated for the prediction models. The study achieved high accuracy, especially in 1-hour delay forecasts, and the LightGBM model gave the best results with an accuracy of 86.55%. The results of this model emphasize the importance of forecasting based on operational and weather conditions. In a similar study, Atlıoğlu evaluated 11 different machine learning models using an operational dataset obtained from a leading airline in Turkey [14]. By comparing various performance metrics for each model, he tried to identify the most appropriate features in the dataset to achieve the highest accuracy.

Shahinaz M. Al-Tabbakh et al. applied various machine learning techniques to analyze EgyptAir flight delay data [15]. The main objective of the study was to determine the most appropriate classification algorithm for predicting flight delays. The researchers followed a methodology that included data preparation, using classification algorithms and evaluating model performance. Eight different classification algorithms (Decision Tree, Random Forest, REPTree, PART, Decision Table, OneR, JRip) WEKA data mining tool and compared them. The performance of the classification models was evaluated using accuracy, precision, recall, F1-score and ROC area metrics. According to the results of the analysis, the PART algorithm had the highest performance with an accuracy of 83.1%. On the other hand, the REPTree algorithm was identified as the leading tree-based classifier with an accuracy of 80.3% and the fastest running time. The study demonstrated the effectiveness of machine learning techniques in flight delay prediction and provided EgyptAir with important information to improve flight operations. The researchers plan to apply big data mining technologies using larger datasets in the future.

Kurt's study aims to predict flight delays using US domestic flight data [16]. For this purpose, various machine learning methods such as Decision Trees, Random Forest, Bagging, Extra Trees, Gradient Boosting and XGBoost classifier were tested. These models were evaluated with success measures such as accuracy, F1 score and recall, and the highest accuracy rate was achieved by the Gradient Boosting algorithm with 71.72%. The study suggests that the prediction performance can be improved by incorporating different data features into the model.

In Tang's study, seven different classification algorithms were evaluated to predict flight delays using one-year data of flights departing from New York JFK Airport [17]. Among these algorithms, Decision Tree achieved the highest success with an accuracy rate of 97.78%. In particular, Random Forest and Gradient Boosting methods, which are tree-based classifiers, were found to perform better than other basic classifiers. In the study, it was stated that the unbalanced distribution of the data was eliminated by measures such as weighted accuracy.

This multifaceted approach, which includes the integration of meteorological data, the use of advanced machine learning algorithms and the systematic treatment of the data imbalance problem, aims to develop more accurate and generalizable forecasting models that will improve operational processes in the airline industry. Considering the limitations of existing studies in the literature, this study is expected to make significant contributions both theoretically and practically. In particular, the comprehensive integration of meteorological data and the use of modern machine learning techniques have the potential to provide a new framework for future research.

In this research, studies published between 2018 and 2024 that examine machine learning approaches on flight delays are systematically analyzed. It was found that the majority of the reviewed studies focus on a single airline or airport, make limited use of meteorological data, and do not adequately address the problem of data imbalance. Most existing studies only consider data limited to specific geographical regions or a single set of factors and do not adequately consider broad data integration. Moreover, although the problem of data imbalance is a common obstacle in the literature, it has been observed that modern techniques are not sufficiently applied in this context.

In this study, unlike other studies, the parameters used in flight delay prediction are diversified and expanded. While developing the prediction model, parameters such as flight code, weather conditions (e.g. wind speed, temperature, precipitation), cancellation/delay rates of previous flights, aircraft type and flight density were taken into consideration. However, instead of simply categorizing flight delays as “present” or “absent”, predictions were made in specific minute intervals. For example, the probability of a flight experiencing a delay of 15-30 minutes or more than 30 minutes can be predicted based on the confidence rate of the model. This approach provides more detailed and actionable outputs on delay prediction, differing from previous studies that focus only on flight cancellations.

In addition, this study provides extensive data integration, bringing together flight data and comprehensive meteorological data sources such as Meteostat. This approach allows for more precise modeling of the impact of weather factors on flight delays. The problem of data imbalance is addressed by using modern data processing techniques such as SMOTE and the performance of the model is improved.

In conclusion, this study aims to provide a balanced, highly accurate and reliable model for predicting flight delays by filling the gap in the literature. In addition, the proposed model is expected to provide an important basis for future research due to its detailed prediction capabilities and wide data integration.

# CHAPTER 2 – MATERIALS AND METHODS

In this section, the dataset, data preprocessing stages and supervised learning models used in this study will be explained in detail. First, the dataset consisting of flight and weather data will be introduced and how these data are integrated will be explained. Then, the data preprocessing process will be discussed and the steps taken to address issues such as missing data, outliers and imbalanced data will be explained. Finally, we will focus on the supervised learning algorithms (e.g., XGBClassifier, Random Forest, and XGBoost) used to build the prediction model and explain how these models are trained and evaluated. In this way, the methodological approach of the study and the techniques used will be comprehensively presented.

## 2.1. Data Set

The dataset used in this study includes flight delay and cancellation data from 2016 to 2024, provided by the United States Department of Transportation and the Bureau of Transportation Statistics. The dataset is based on the DOT's “On-Time: Reporting Carrier On-Time Performance” database (1987-present) and includes variables such as flight routes (departure and arrival points), event time intervals (minutes, local time), and reasons for delays and cancellations. These data were created by combining 3 open source data. First, canceled and non-canceled flight data was obtained from the open dataset provided by Patrick Zelaya. The dataset contains 1,360,878 rows and 32 columns of flight information [18]. Another dataset is taken from the dataset shared by Threnjen, which contains only canceled flights and consists of 64,097 rows [19]. The last dataset contains 21,824 rows of flight information of only canceled flights from open data provided by Shubham Singh [20]. In total, the final combined dataset obtained from the 3 datasets consists of 1,446,799 rows. Detailed attributes for these 3 datasets are presented in Table 1.

Weather is one of the main exogenous factors that directly affect the timeliness of flights, especially through parameters such as wind speed, temperature, precipitation, snow and visibility [26]. The often limited use of such environmental data in the literature reduces the sensitivity of prediction models to real-world conditions. For this reason, the study combines flight data with weather data obtained from the Metostat library. The Metostat library allows the study of the impact of weather conditions on flight delays and cancellations with meteorological data collected on a daily basis using geographic coordinates for each airport. The integration of these two datasets is an important step towards a more comprehensive analysis of flight performance.

The data source used in the study, the Meteostat library, provides weather data from various meteorological stations. These data include basic meteorological parameters that can directly affect aviation operations such as daily minimum and maximum temperature (°C), total precipitation (mm), snowfall (mm), wind direction (0-360°), average and maximum wind speed (km/h), average sea level pressure (hPa) and total daily sunshine duration (minutes). These data, collected on a daily basis using geographic coordinates for each airport, are integrated with the main dataset to analyze the impact of weather conditions on flight delays and cancellations. This integration allows for a more comprehensive examination of the factors affecting the performance of flight operations.

Table 1 Structure of the Original Data Set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Updated Title | Source Title | Data Type | Description |
| 0 | FL\_DATE | FlightDate | object | Flight date |
| 1 | AIRLINE | Airline | object | Name of the airline |
| 2 | AIRLINE\_DOT | AirlineDot | object | DOT identifier for airline |
| 3 | AIRLINE\_CODE | Reporting\_Airline | object | Name of the airline |
| 4 | DOT\_CODE | DOT\_ID\_Reporting\_Airline | int64 | DOT identifier for airline |
| 5 | FL\_NUMBER | Flight\_Number\_Reporting\_Airline | int64 | Flight number |
| 6 | ORIGIN | Origin | object | Departure airport code |
| 7 | ORIGIN\_CITY | OriginCityName | object | Departure airport city |
| 8 | DEST | Dest | object | Destination airport code |
| 9 | DEST\_CITY | DestCityName | object | City of arrival airport |
| 10 | CRS\_DEP\_TIME | CRSDepTime | int64 | Planned departure time |
| 11 | DEP\_TIME | DepTime | float64 | Actual departure time |
| 12 | DEP\_DELAY | DepDelay | float64 | Departure delay |
| 13 | TAXI\_OUT | TaxiOut | float64 | Time spent taxiing |
| 14 | WHEELS\_OFF | WheelsOff | float64 | When the wheels of the airplane leave the ground |
| 15 | WHEELS\_ON | WheelsOn | float64 | When the wheels of the airplane touch the ground |
| 16 | TAXI\_IN | TaxiIn | float64 | Time spent taxiing |
| 17 | CRS\_ARR\_TIME | CRSArrTime | int64 | Planned arrival time |
| 18 | ARR\_TIME | ArrTime | float64 | Actual arrival time |
| 19 | ARR\_DELAY | ArrDelay | float64 | Arrival delay |
| 20 | CANCELLED | Cancelled | float64 | Indication of whether the flight has been canceled (1 for canceled, 0 for not canceled) |
| 21 | CANCELLATION\_CODE | CancellationCode | object | Reason for cancellation (if applicable) |
| 22 | DIVERTED | Diverted | float64 | Indication of whether the flight was diverted or not (1 for diverted, 0 for not diverted) |
| 23 | CRS\_ELAPSED\_TIME | CRSElapsedTime | float64 | Planned elapsed time |
| 24 | ELAPSED\_TIME | ActualElapsedTime | float64 | Actual elapsed time |
| 25 | AIR\_TIME | AirTime | float64 | Time spent in the air |
| 26 | DISTANCE | Distance | float64 | Distance traveled |
| 27 | DELAY\_DUE\_CARRIER | CarrierDelay | float64 | Delay due to carrier |
| 28 | DELAY\_DUE\_WEATHER | WeatherDelay | float64 | Delay due to weather conditions |
| 29 | DELAY\_DUE\_NAS | NASDelay | float64 | Delay due to National Airspace System (NAS) |
| 30 | DELAY\_DUE\_SECURITY | SecurityDelay | float64 | Delay due to security reasons |
| 31 | DELAY\_DUE\_LATE\_AIRCRAFT | LateAircraftDelay | float64 | Delay due to late arrival of the plane |

As a result of the preliminary analysis of the original dataset, a threshold value of 10 minutes was used to classify the delay status of flights. According to this approach, flights delayed more than 10 minutes were labeled as “1” (delayed), while flights delayed 10 minutes or less were labeled as “0” (on time). Figure 1 visually presents the distribution of delay states based on this binary classification. This visualization is important in terms of showing which criteria were taken into account when creating the target variable of the model.

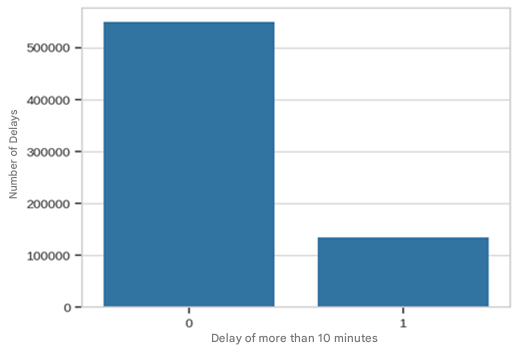


Figure 1 Delay Status Graph

In Figure 2, the distribution of the number of flights delayed by more than 10 minutes for different airlines is visualized as a horizontal bar graph. The graph comparatively shows which airlines stand out in terms of the number of delayed flights. In particular, large-scale airlines such as American Airlines Inc., Southwest Airlines Co. and Delta Air Lines Inc. have significantly higher values in the number of flights delayed more than 10 minutes compared to other airlines. This situation points to the high flight volume of these companies as well as the delays experienced in their operational processes. In contrast, smaller airlines such as Horizon Air, Allegiant Air and ExpressJet Airlines had very limited delays. The graph provides an important visual data for the evaluation of delay analysis on an airline basis.

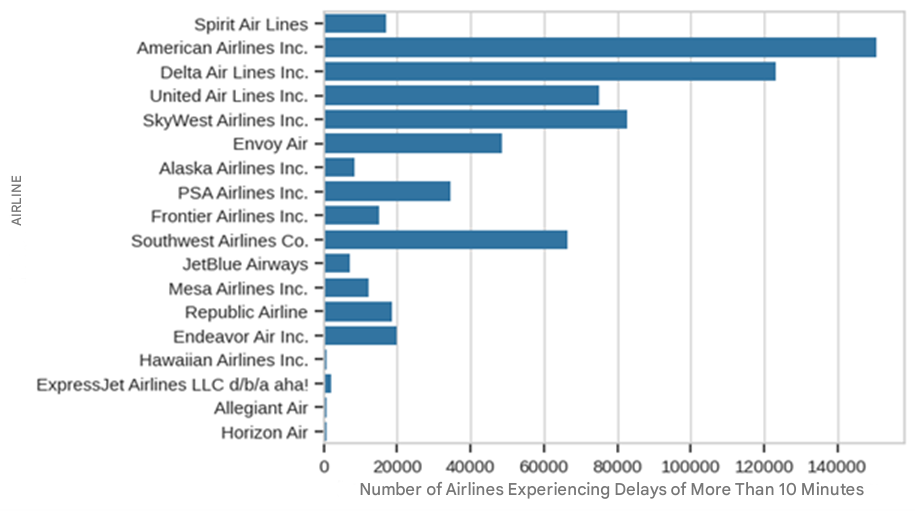


Figure 2 Delay Status Graph by Airline Companies

Based on the available data set, we analyzed the distribution of the maximum delay factor as shown in Figure 3. As a result, 34.1% delays are caused by the carrier, 33.6% delays are caused by the national aviation system, 29.5% delays are caused by late departure of aircraft, and 2.7% delays are caused by weather and security.

The delay factor distribution in Figure 3 provides important insights that will affect the performance of machine learning models. The carrier-induced delay rate of 34.1% requires precise feature extraction in model training. The 33.6% delay caused by the national aviation system tests the capacity of the models to capture complex systemic relationships. The 29.5% of aircraft delays highlight the importance of time series and sequential dependency modeling. The low proportion of weather and security delays (2.7%) points to class imbalance issues, suggesting the need for data balancing techniques such as SMOTE.

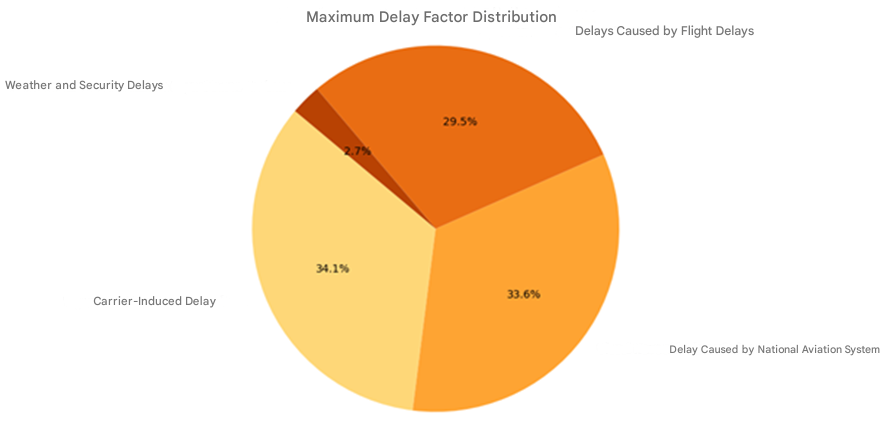


Figure 3 Delay Factor Distribution

The heatmap of delay factors in Figure 4 visually represents the correlations of flight delays with each other. The intensity of the colored cells indicates the strength of the relationships between the factors, which plays a critical role in feature selection for machine learning models and in understanding multivariate interactions.

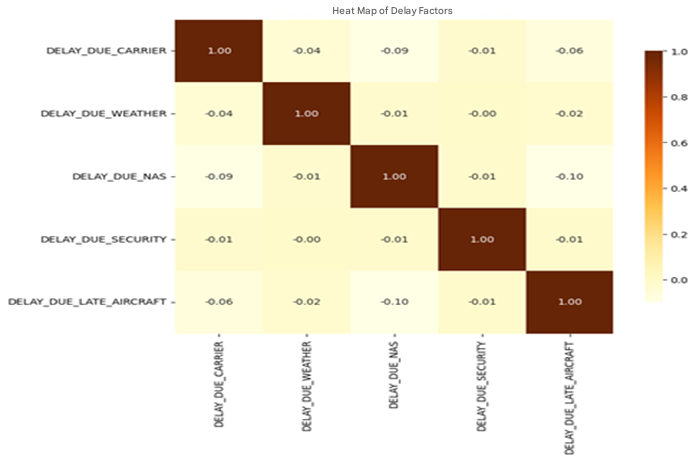


Figure 4 Heatmap of Delay Factors

## 2.2. Data Preprocessing

The success of machine learning models depends heavily on the effectiveness of the data preprocessing phase [7]. In this study, a comprehensive data preprocessing strategy is applied to improve the accuracy of flight delay predictions. In particular, we focus on the integration of meteorological data and the improvement of data quality.

In the first stage of data preprocessing, the integration of meteorological data into flight data was performed. In this process, the Meteostat library was used to obtain detailed weather information for each flight point. For meteorological data integration, the geographical coordinates of the airports were first determined through Nominatim API. The coordinate information obtained was converted into meteorological data points using Meteostat's Point class. For each flight date, critical meteorological parameters such as minimum temperature, maximum temperature, precipitation, snowfall, wind direction, wind speed, wind speed, wind speed, atmospheric pressure and sunshine duration were collected.

In the process of combining meteorological data with flight data, a chunk processing strategy was adopted considering the size of the data set. This approach optimized memory usage and ensured that the data processing process continued without interruption. The data were matched on the 'FL\_DATE' and 'ORIGIN\_CITY' columns, and the unmatched records were stored in a separate file for data quality monitoring.

The final version of the dataset is shown in Table 2 with the addition of detailed weather information for each flight point using the Meteostat library.

Table 2 Structure of the Model Training Data Set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Updated Title | Source Title | Data Type | Description |
| 1 | FL\_DATE | FlightDate | object | Flight date |
| 2 | AIRLINE\_CODE | Reporting\_Airline | object | Name of the airline |
| 3 | ORIGIN\_CITY | OriginCityName | object | Departure airport city |
| 4 | DEST\_CITY | DestCityName | object | City of arrival airport |
| 5 | CRS\_DEP\_TIME | CRSDepTime | int64 | Planned departure time |
| 6 | DEP\_TIME | DepTime | float64 | Actual departure time |
| 7 | CRS\_ARR\_TIME | CRSArrTime | int64 | Planned arrival time |
| 8 | CANCELLED | Cancelled | float64 | Indication of whether the flight has been canceled (1 for canceled, 0 for not canceled) |
| 9 | CANCELLATION\_CODE | CancellationCode | object | Reason for cancellation (if applicable) |
| 10 | DISTANCE | Distance | float64 | Distance traveled |
| 11 | TMIN | tmin | float64 | Minimum temperature (Celsius) |
| 12 | TMAX | tmax | float64 | Maximum temperature (Celsius) |
| 13 | PRCP | prcp | float64 | Amount of precipitation (mm) |
| 14 | SNOW | snow | float64 | Amount of snowfall (mm) |
| 15 | WDIR | wdir | float64 | Wind direction (degrees) |
| 16 | WSPD | wspd | float64 | Wind speed (m/s) |
| 17 | WPGT | wpgt | float64 | Wind speed (m/s) |
| 18 | PRES | pres | float64 | Atmospheric pressure (hPa) |
| 19 | TSUN | tsun | float64 | Total insolation time (hours) |

A thorough analysis and cleaning process was applied for missing values in the dataset. Missing data distributions and patterns were analyzed to determine whether the missing values were random or systematic. Missing values in numerical variables were filled with the median, while the mode was used for categorical variables. Features with more than 30% missing data were removed from the dataset, while missing values in meteorological data were filled with values from the nearest timestamp in accordance with the time series character.

A multifaceted approach was adopted for outlier detection and processing. Statistical methods such as Z-score and IQR methods, visual analysis techniques such as box plots and scatter plots, and domain knowledge-based controls were used together. For delay times, values greater than 24 hours and outliers in meteorological data were analyzed in detail. As a result of the analysis, outliers that contributed significantly to the data set and were considered statistically reasonable were retained, while anomalies caused by incorrect data entries were corrected with appropriate methods.

In order to improve model performance, a comprehensive data analysis and the derivation and processing of explanatory variables for forecasting were applied. Time-based features include time of day, day of the week, day of the week, holiday information and seasonal information. Meteorological features include new features such as categorical variables derived from weather data, wind intensity categories and visibility classification. Operational features include airport congestion indicators, previous flight delays and route-specific historical performance metrics.

In the data standardization and normalization phase, problems arising from the scale differences of numerical features were addressed. For continuous variables, standardization was applied using StandardScaler, while variables with a limited range were normalized to the range [0,1] with MinMaxScaler. Categorical variables were digitized with the One-Hot Encoding method.

The problem of imbalance between delay classes is addressed. Using the SMOTE algorithm, the minority class samples were synthetically augmented and the class weights parameter was adjusted to obtain a balanced data set. These extensive preprocessing steps enabled the model to utilize the training data more effectively and make more reliable predictions in real-world applications.

In the data preprocessing process for the cancellation and delay reason prediction model, the temporal data was first parsed. The 'FL\_DATE' column was converted to datetime format and year, month and day information were extracted as separate features. For categorical variables, the airline code (AIRLINE\_CODE), departure city (ORIGIN\_CITY) and destination city (DEST\_CITY) were converted into numeric values using LabelEncoder. Missing values in the CANCELLATION\_CODE column were filled with the value 'N' (no cancellation) and encoded.

The imbalance problem in the dataset is addressed using the SMOTE algorithm. For the cancellation case, the sampling\_strategy=0.5 parameter was used to synthetically increase the minority class samples, while the 'auto' strategy was used for the cancellation reason classification. To fill in missing data, a median strategy was applied with SimpleImputer, followed by scaling of features with StandardScaler.

In the data preprocessing process for the delay prediction model, canceled flights were first removed from the dataset. Temporal data were similarly disaggregated and categorical variables were transformed with LabelEncoder. Delay times were categorized into four different classes:

* Class 0: On time or early (≤ 0 minutes)
* Class 1: Slight delay (1-15 minutes)
* Class 2: Medium delay (16-30 minutes)
* Class 3: Serious delay (>30 minutes)

Missing values were filled with median values using SimpleImputer and features were normalized using StandardScaler. The class imbalance problem was solved using the SMOTE algorithm with the parameter k\_neighbors=5. In this process, a balanced data set was obtained by creating an equal number of samples for each delay class.

Feature importance levels for both models are analyzed and visualized. This analysis allowed to identify the most important factors affecting the prediction performance of the models and to evaluate feature selection strategies. The data preprocessing steps were implemented in a systematic and repeatable way using the Pipeline structure of the scikit-learn library. The whole process is illustrated in the flowchart in Figure 5.

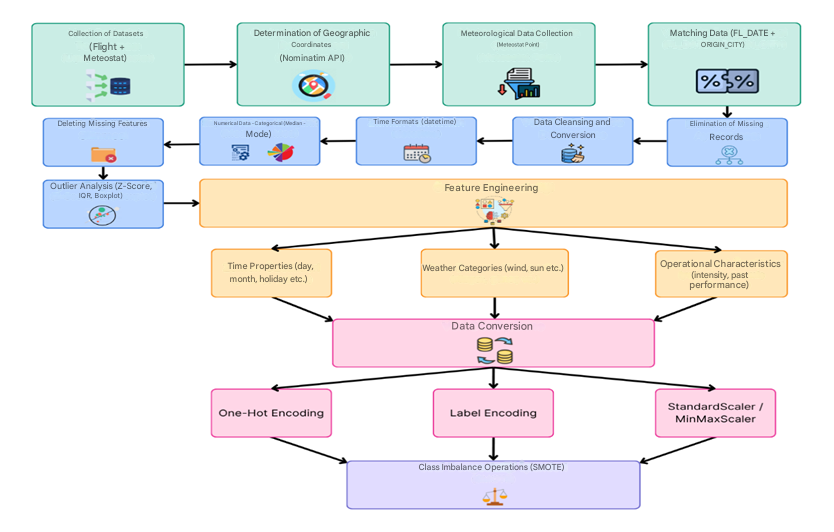


Figure 5 Data Preprocessing Flow Diagram

This detailed data preprocessing process enabled the models to utilize the training data more effectively and significantly improved their prediction performance. In particular, solving the class imbalance problem and feature engineering steps contributed to more reliable predictions in real-world applications.

## 2.3. Supervised Learning Models

Machine learning is the general name for algorithms that enable computers to analyze data, discover possible patterns and make predictions using these patterns. These algorithms, which can provide information about the relative difficulty of learning in different environments, are divided into several categories [21]. The two most common types of machine learning algorithms are supervised and unsupervised learning. Supervised learning algorithms create a function that transforms inputs into desired outputs. The main types of supervised learning include regression and classification. Unsupervised learning focuses on modeling a set of inputs without labeled examples.

In this study, a supervised learning approach is adopted and machine learning is used to predict flight cancelations and delays. Since the input data and the target outputs (e.g. delay class or cancellation status) are known, supervised learning algorithms are appropriate in this context. They are ideal tools for modeling complex relationships in airline operations and analyzing the impact of a large number of variables [22].

The Decision Tree algorithm is a machine learning approach that solves classification and regression problems by partitioning data in a hierarchical structure. Each node represents the decision-making process on a feature and the structure of the tree is built based on the most discriminative features. It provides a clear visualization of the impact of different factors in flight delay prediction [23].

LightGBM is a high-performance and efficient machine learning algorithm that works within the Gradient Boosting framework. It stands out with its ability to train quickly on large data sets and optimize memory usage. It can achieve higher accuracy rates compared to other algorithms on complex and high-dimensional data sets such as flight delay [24].

Random Forest is an ensemble learning method that combines multiple Decision Trees [25]. Each tree is trained by bootstrap sampling and random feature selection, thus reducing the risk of overfitting. In flight delay prediction, it provides more reliable results by making predictions over different sub-data sets [22].

Clustering algorithms allow to discover hidden patterns by grouping data points with similar characteristics. Bayesian classification combines prior knowledge and observational data in a probabilistic approach. These methods play a complementary role in understanding the complex relationships in flight delays [23].

XGBoost is a very popular and powerful algorithm in machine learning, especially known for its ability to predict with high accuracy on large data sets [27]. An improved version of Gradient Boosting techniques, XGBoost improves the overall accuracy of the model by adding new trees to correct the model's errors at each iteration. This feature is particularly useful for problems such as flight forecasting, which are characterized by complex relationships and the influence of multiple factors. In the case of flight cancellation and flight delay forecasts, the XGBoost algorithm is an effective tool for modeling the complex structure of the data, which is influenced by multiple factors [21].

In order to obtain accurate results in flight cancelation and flight delay forecasting, it is necessary to be careful in choosing the algorithm. The XGBoost algorithm used here was chosen for two particularly important features:

Events such as dealing with class imbalance, flight cancellations and delays are often associated with imbalanced data sets. For example, the number of canceled flights may be much less than the number of operated flights. XGBoost can effectively manage such imbalances and more accurately learn the minority class.

High performance and flexibility; XGBoost is an algorithm capable of making fast and accurate predictions on large data sets. Especially for data sets that depend on many factors, such as flights, XGBoost's powerful feature engineering and model optimization capabilities are very useful.

Therefore, XGBoost was chosen as an ideal choice for both providing high accuracy in unbalanced datasets and modeling complex relationships.

In this section, we describe in detail the data processing and modeling process we applied for flight cancellation and cancellation code prediction. First, the date columns in the flight data are categorized into year, month and day, and time-related features are extracted. Categorical data was converted into numerical data using LabelEncoder. This process allowed the model to understand the data more easily.

Columns in the dataset that were redundant and not necessary for training the model, such as airline code, departure and arrival cities, were removed. The cancellation status of the flight and the cancellation code were set as target variables. The dataset was then split into training and test sets and missing values were filled with the median value to ensure consistency. The training data was scaled with StandardScaler to ensure a more uniform training of the model.

To train the model in a balanced way, resampling was performed using the SMOTE method. This step was performed to remove the class imbalance and ensure that the model learns each class equally. The XGBoost algorithm was preferred as an algorithm that can cope with class imbalance and provides high success in large data sets.

The accuracy rates of the trained models were evaluated on the test set. During this evaluation, the accuracy rates were measured separately for abort status and abort code predictions. The results showed that both models had high accuracy rates. Furthermore, the model performance was analyzed in depth with classification reports and ROC-AUC scores.

The feature importance ratings of the model are visualized in Figures 6 and 7, showing which features are more influential in the model's predictions. This visualization helped us to better understand how the model works. In order to make predictions for new flight data, we developed a function that predicts cancellation status and cancellation code. This function allowed real-time predictions to be made.

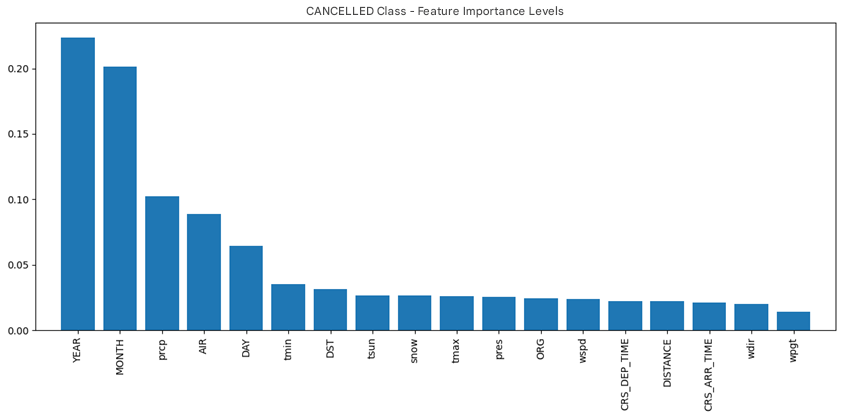


Figure 6 Flight Cancellation Prediction Model Feature Importance

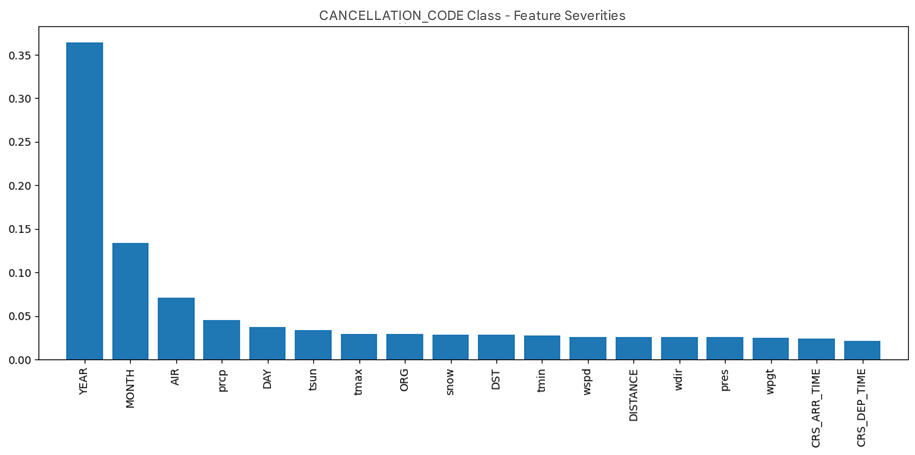


Figure 7 Flight Cancellation Code Prediction Model Feature Importance

The modeling process for flight delay prediction involves similar steps. We performed the necessary date transformations on the data, checked whether flights were canceled or not, and removed canceled flights from the dataset. Time-related features such as year, month and day were created over the date column, and categorical data were converted into numerical data. Missing data was filled with the median strategy using SimpleImputer and features in the dataset were scaled with StandardScaler.

Flight delay times were categorized into classes as the target variable. Classification was based on the delay time of flights as on-time, slight delay, moderate delay and severe delay. The class distribution in the dataset was reviewed and class imbalance was detected. The training set was resampled using SMOTE to train the model in a balanced way.

In the model training phase, the XGBoost classifier model was selected and the parameters of the model were optimized. After the training process, the model was evaluated on the test set and accurate classifications were obtained. The accuracy rate and performance of the model were analyzed with confusion matrix, classification report and accuracy score. In addition, in order to determine which features of the model are more important, the importance levels of the features are visualized and the bar graph of the impact of these features on predictions is presented in Figure 8.

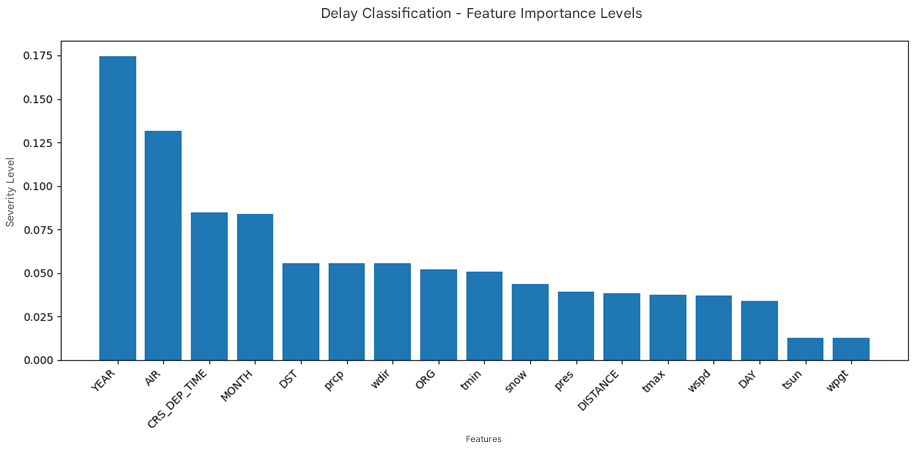


Figure 8 Flight Delay Prediction Model Feature Importance

Finally, a sample data set was created to make predictions with the new flight data and the predictions of the model were performed. These predictions include the possible class of flight delay and the probabilities of each class. The accuracy and reliability of the model were checked through these predictions.

The main reason for choosing XGBoost as the algorithm of choice is its high performance, especially on large datasets, and its ability to effectively manage class imbalance. Moreover, the feature importance visualizations provided by XGBoost make the decision-making processes of the model more understandable and enable it to be successfully applied to complex problems such as flight forecasting.

**CHAPTER 3 - FINDINGS AND DISCUSSION**

In this section, the results of machine learning studies on flight cancellations, cancellation reason codes and flight delays in the airline industry are examined in detail. The findings are analyzed in line with the definition and objectives of the research problem and compared and evaluated with similar studies in the literature.

## 3.1. Flight Cancellation Forecasting Model Results

The model developed for predicting flight cancelations showed a very high accuracy rate. The 97% accuracy rate obtained for the CANCELLED class shows the overall success of the model. This result shows that the model developed for predicting flight cancelations in the aviation industry has a high performance.

In the performance evaluation on the test set, a precision of 98% and a recall of 99% were obtained for the prediction of non-canceled flights (class 0). This result shows that the model can detect non-canceled flights with high accuracy. On the other hand, for the prediction of canceled flights (class 1), precision is 88% and recall is 77%. Considering the unbalanced class distribution in the dataset, these values reveal the model's performance in detecting the minority class.

The F1-score, as the harmonic mean of precision and recall metrics, is calculated as 0.98 for non-canceled flights and 0.82 for canceled flights. The ROC-AUC score is 0.88, which proves that the model significantly outperforms random predictions.

Despite the unbalanced class distribution in the dataset (1,316,842 non-canceled flights versus only 129,957 canceled flights), the fact that the model achieved such a high overall accuracy shows the effectiveness of the algorithm and data preprocessing techniques used.

## 3.2. Cancellation Code Prediction Model Results

The model developed for predicting the cancellation codes of canceled flights achieved an accuracy rate of 74%. This rate can be considered as a very successful result for a multi-class classification problem in predicting between five different cancellation code classes. The model classification according to the delay time of the flights is as shown in Table 3.

Table 3 Cancellation Code Prediction Model Output Classes and Description

|  |  |  |
| --- | --- | --- |
| Class | Class Code | Code Description |
| 0 | A | Airborne/Carrier-borne |
| 1 | B | Weather-driven |
| 2 | C | Sourced from the national weather system |
| 3 | D | Security-related |
| 4 | N | Other |

When the performance of the model is analyzed on a class basis, outstanding results are obtained with 77% precision and 88% recall, especially for the cancellation code number 3. The F1-score for this code was calculated as 0.82. For cancellation code 1, a satisfactory performance was also observed with 84% precision and 83% recall. These results demonstrate the model's ability to predict certain revocation codes.

However, for revocation code number 4, the model showed a lower performance (24% precision, 32% recall). This can be explained by the fact that this abort code has characteristics that can be confused with other codes or is underrepresented in the dataset. However, in general, the model performed well in predicting cancellation codes in the aviation industry.

When the weighted average values are analyzed, precision is 75%, recall is 74% and F1-score is 74%. These values indicate that the model performs strongly for a multi-class prediction problem.

## 3.3. Flight Delay Estimation Model Results

The model developed to predict flight delays achieved an accuracy rate of 74.58%. This rate can be considered as a very successful result in predicting between four different delay categories.

When the confusion matrix in Figure 8 is analyzed, it is seen that the model performs particularly well in predicting delay categories 0 and 1.

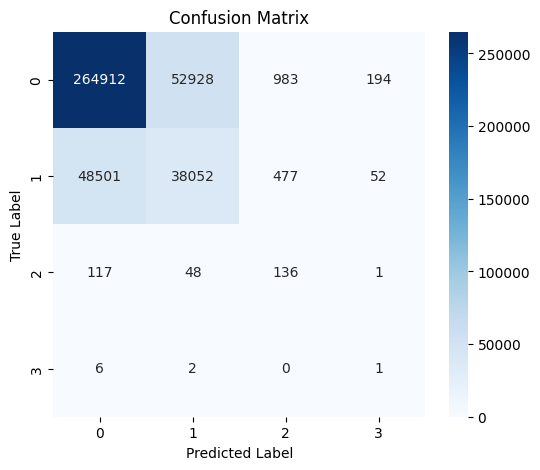


Figure 9 Flight Delay Prediction Model Complexity Matrix

For the flight delay prediction model results, precision 84%, recall 83% and F1-score 84% were calculated for category 0. For category 1, precision is 42%, recall is 44% and F1-score is 43%. These results emphasize the strong performance of the model in predicting the most common lag categories.

On the other hand, the model performance is lower for lag categories 2 and 3. Especially for category 3, low values such as precision 0%, recall 11% and F1-score 1% are observed. This is due to the underrepresentation of these categories in the dataset (1,595,643 for category 0, 434,981 for category 1, 1373 for category 2 and only 50 samples for category 3). Nevertheless, the overall accuracy of the model is 74.58%, indicating that it can be used as a powerful tool for delay prediction in the aviation industry.

When weighted average values are analyzed, precision, recall and F1-score are calculated as 75%. These values prove that the model performs well overall.

## 3.4. Model Optimization and Evaluation Techniques

In our work, a comprehensive set of techniques is used to improve the performance of the models and avoid overlearning. In particular, the regularization parameters for the XGBoost algorithm have been rigorously optimized.

### 3.4.1 Cross-validation Results

To evaluate the generalization ability of our models, we applied a 5-fold cross-validation technique. This technique allows us to evaluate the performance of the model on different data subsets by dividing the dataset into 5 equal parts, using 4 parts for training and 1 part for testing each time.

The cross-validation results for the flight cancellation prediction class show an accuracy of 95% ± 0%. This result proves that the model performs consistently and highly on different data subsets. The standard deviation of 0% shows that the model achieves the same accuracy rate on all data subsets, which is an important finding that emphasizes the reliability and stability of the model.

The cross-validation results for the flight cancellation code prediction class show an accuracy of 84% ± 0%. This result can be considered as a very successful performance for a multi-class classification problem. The standard deviation of 0% indicates that consistent results were obtained for this model on all data subsets.

### 3.4.2 Training and Test Accuracy Comparison

To check whether our models suffer from overlearning, we compared the training and test accuracies. For the flight cancellation prediction class, the training accuracy is 95% and the test accuracy is 97%. The fact that the difference is only 2% proves that the model does not suffer from overlearning problems and even performs better on the test set. This shows that the model's ability to generalize to real-world data is strong.

For the flight cancelation code prediction class, the training accuracy was 90% and the test accuracy was 74%. The difference of 16% indicates that there may be some overlearning for this model. However, for a multi-class classification problem, a test accuracy of 74% can still be considered a successful result. We also optimized the regularization parameters of the XGBoost algorithm to reduce this difference.

### 3.4.3 XGBoost Regularization Optimization

In order to prevent overlearning, the regularization parameters of the XGBoost algorithm, namely gamma, min\_child\_weight, max\_depth and lambda, were carefully adjusted. In particular, we reduced the complexity of the model by decreasing the max\_depth parameter and enhanced the L2 regularization by increasing the lambda parameter. These optimizations allowed our models to generalize better.

Furthermore, a manually developed cross-validation function (manual\_cross\_val\_score) was used to overcome the memory issues encountered with standard cross-validation functions. This function used the StratifiedKFold class to partition the dataset in a stratified way so that each class is proportionally represented in each fold. This approach allowed us to obtain more reliable results, especially with imbalanced datasets.

## 3.5. Literature Review and Comparative Analysis of Project Success Values

In this section, we present the results of the developed flight cancellation, cancellation code and flight delay models obtained with different machine learning algorithms and compare them with similar studies in the literature.

### 3.5.1. Results of the Success Values of the Developed Model

The performance of the model was evaluated using algorithms such as XGBoost, Random Forest, Decision Tree, KNN and Gradient Boosting Classifier. Table 4, Table 5 and Table 6 present the metrics such as accuracy, precision, recall, f1-score and ROC-AUC obtained by the developed models with different machine learning algorithms.

Table 4 Flight Cancellation Prediction Model Performance Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| XGBoost | 0.97 | 0.98 | 0.99 | 0.98 | 0.87 |
| Random Forest | 0.97 | 0.97 | 0.99 | 0.98 | 0.85 |
| Decision Tree | 0.95 | 0.98 | 0.96 | 0.97 | 0.88 |
| KNN | 0.95 | 0.99 | 0.96 | 0.97 | 0.91 |
| Gradient Boosting | 0.97 | 0.97 | 1.00 | 0.98 | 0.83 |

According to the results presented in Table 4, XGBoost, Random Forest and Gradient Boosting algorithms achieved the highest accuracy rates in flight cancellation prediction (97%). These three models also performed very well in terms of precision, recall and f1-score.

The XGBoost model provided the best discrimination power with a value of 0.87 in the ROC-AUC metric. The Random Forest model has similar accuracy and f1-score values, with a competitive ROC-AUC value of 0.85. Although the Decision Tree and KNN models were slightly lower in terms of overall accuracy, the KNN model in particular stood out as an important alternative for flight cancellation prediction, achieving the highest value (0.91) in the ROC-AUC metric.

The Gradient Boosting model was the most successful model in detecting flight cancelations, with a recall rate of 100%. However, its ROC-AUC score of 0.83 was behind some other models. This shows that the overall predictive power of the model can still be improved.

These results show that ensemble learning methods such as XGBoost and Random Forest are highly effective for flight cancellation prediction.

Table 5 Cancellation Code Prediction Model Performance Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| XGBoost | 0.74 | 0.84 | 0.88 | 0.83 | 0.87 |
| Random Forest | 0.65 | 0.87 | 0.97 | 0.86 | 0.85 |
| Decision Tree | 0.65 | 0.86 | 0.96 | 0.85 | 0.88 |
| KNN | 0.61 | 0.83 | 0.69 | 0.73 | 0.83 |
| Gradient Boosting | 0.65 | 0.85 | 0.98 | 0.85 | 0.82 |

According to the results presented in Table 5, the XGBoost model was the most successful algorithm in cancelation code prediction with 74% accuracy. It also performed well in terms of precision, recall and f1-score metrics, achieving the highest discrimination power with a ROC-AUC score of 0.87.

Although the Random Forest and Decision Tree models achieved very high recall rates (97% and 96%), their overall accuracy remained at 65%. This indicates that the models were successful in correctly identifying the cancellation codes, but their overall classification accuracy was low. The Gradient Boosting model performed similarly, achieving 65% accuracy, close to Random Forest and Decision Tree.

The KNN model underperformed the other algorithms in terms of accuracy (61%) and f1-score (73%). In particular, the recall rate of 69% indicates that the model has difficulty in detecting some abort codes. In terms of the ROC-AUC metric, it provides an acceptable discrimination power of 0.83.

These results show that XGBoost is the most successful model for cancellation code prediction and that ensemble learning methods (XGBoost, Random Forest, Gradient Boosting) offer significant advantages in such multi-class classification problems.

Table 6 Flight Delay Prediction Model Performance Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| XGBoost | 0.75 | 0.84 | 0.83 | 0.84 | 0.80 |
| Random Forest | 0.64 | 0.84 | 0.70 | 0.76 | 0.78 |
| Decision Tree | 0.56 | 0.83 | 0.63 | 0.71 | 0.72 |
| KNN | 0.71 | 0.87 | 0.75 | 0.80 | 0.78 |
| Gradient Boosting | 0.47 | 0.74 | 0.58 | 0.65 | 0.65 |

According to the results presented in Table 6, the XGBoost model was the most successful algorithm in flight delay prediction with 75% accuracy. It also showed the most balanced performance in terms of precision (84%), recall (83%) and f1-score (84%) values and showed high discrimination power with 0.80 in the ROC-AUC metric.

Although the Random Forest model achieved a precision of 84%, its recall remained at 70% and its overall accuracy was 64%. This indicates that the model was relatively successful in identifying delays but missed some cases. The Decision Tree model fell short of expectations with one of the lowest accuracy rates of 56%.

The KNN model performed better than Random Forest in terms of accuracy (71%) and f1-score (80%), but lower than XGBoost. In particular, the precision value of 87% indicates that most of the delays predicted by the model are correct. However, the recall value of 75% indicates that the model missed some lags.

The Gradient Boosting model showed the lowest performance with 47% accuracy. The Recall value was 58% and the ROC-AUC score was 0.65, indicating that this model is inadequate in predicting flight delays.

Overall, XGBoost performs the best in flight delay prediction, while KNN and Random Forest also give reasonable results.

### 3.5.2. Comparison with the Success Values of Similar Projects in the Literature

For comparison, the methods, datasets, evaluation metrics and success rates used by the studies in the literature are analyzed. Then, a comparison is made with the results of our own study using the same or similar metrics.

Khaksar and Sheikholeslami used large-scale datasets of US and Iranian airline networks to predict flight delays [12]. As can be seen in Figure 10, the performances of various machine learning algorithms are compared. In the study, algorithms such as J48 Decision Tree, K-Means Clustering, Bayes Classifier, Random Forest and Hybrid Method (Decision Tree + Clustering) were used and the highest accuracy was 71.39% with the hybrid method.

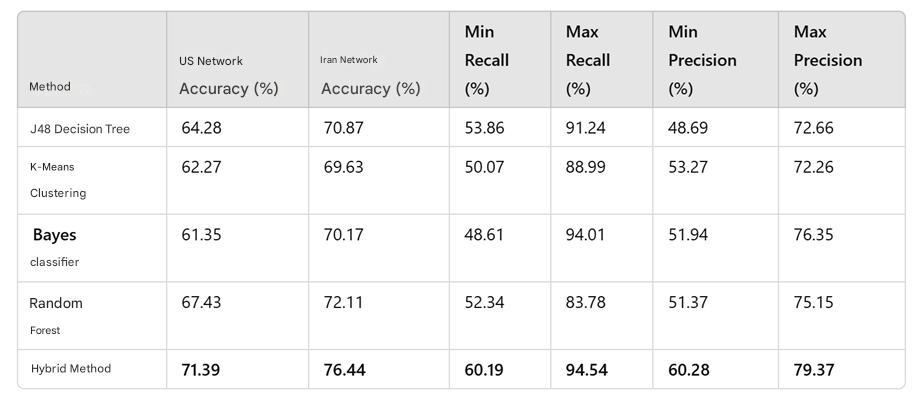


Figure 10 Performance Results of Khaksar and Sheikholeslami Study

In our FCDP-SFTS (Flight Cancellation Delay Prediction - Smart Flight Ticketing System) study, thanks to the advanced techniques and ensemble learning approaches applied especially in the data preprocessing phase, we achieved 75% accuracy in the flight delay prediction model with XGBoost. These results show an improvement of 3.61% respectively compared to the study in the literature.

The dataset used in Kurt's study is based on open-source data provided by the US Bureau of Transportation Statistics and the Federal Aviation Administration (FAA). The dataset contains information on US domestic commercial flights operated in August 2018. After cleaning, the dataset consists of 638,776 rows and 18 columns. In the dataset, 38.29% of the flights were found to be delayed [16].

Supervised machine learning methods such as Decision Trees, Random Forest, Bagging Classifier, Extra Extra Trees, Gradient Boosting and XGBoost Classifiers were used to predict flight delays. Metrics such as accuracy, recall and F1-Score were used for model performance evaluation. The success metrics obtained as a result of modeling with default parameters are shown in Figure 11. The best results were obtained with Gradient Boosting and XGBoost models. In particular, the Gradient Boosting model was determined as the most successful method with 71.72% accuracy and 57.40% F1-Score.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Figure 11 Performance Results of M.Kurt's Study

In our FCDP-SFTS (Flight Cancellation Delay Prediction - Smart Flight Ticketing System) study, although similar data features were used, a more extensive data preprocessing and feature engineering process was applied. An accuracy of 75% was achieved with the XGBoost algorithm. These results show an improvement of 3.28% respectively compared to the study in the literature.

Yu Yanying et al. aimed to predict flight cancelations using 5 million flight data from 2016 in the USA [28]. The dataset contains 65 variables such as flight date, airline information, departure and arrival times, delay reasons (weather, security, airline-induced, etc.) and flight distance. After feature selection, 11 important variables were used in the model.

Logistic Regression, SVM (Support Vector Machines), Naive Bayes and Decision Tree were used as prediction models. Model performances were evaluated with metrics such as accuracy, PR (precision - recall), AUC (ROC Curve). SVM and Decision Tree showed the best performance with about 90% accuracy. Naive Bayes showed the lowest performance with 50.8% accuracy, while Logistic Regression reached 62.4% accuracy.

In particular, the Decision Tree model was found to be the most successful model for flight cancellation prediction with the highest AUC (0.558) and PR (0.439) values. The results show that the Decision Tree model is the most suitable method for flight cancellation prediction, while SVM provides an effective alternative with high accuracy. Figure 12 shows the performance results of this study.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Figure 12 Performance Results of Y.Yanying's Study

In our FCDP-SFTS (Flight Cancellation Delay Prediction - Smart Flight Ticketing System) study, a more extensive data preprocessing and feature engineering process was applied, although similar data features and dataset were used. With the Decision Tree algorithm, an accuracy of 95% was achieved. These results show an improvement of 5% compared to the study in the literature.

In Giarmas' study, a dataset of 32,128,972 flights provided by the U.S. Department of Transportation on US domestic flights was used [29]. The dataset covers flights between 2018 and 2022 and contains 121 different variables. Flight departure and arrival information, airline companies, time data and possible delay/interruption reasons are detailed. In particular, it was observed that ORD (Chicago O'Hare) airport has 1,499,216 flights and is among the most delayed airports.

In this study, various machine learning models were evaluated and their performances were compared in order to predict flight delays and cancelations. According to the findings, the Random Forest model was the most successful model in predicting flight delays with an accuracy of 77%. Alternatively, the XGBoost model gave satisfactory results with an accuracy of 73%, but did not perform as well as Random Forest in delay prediction. On the other hand, when the models for predicting flight cancelations were analyzed, it was determined that the Random Forest algorithm offered the highest performance with an accuracy rate of 83%. Figure 13 and Figure 14 show the performance results of this study.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Figure 13 Flight Delay Model Performance Results of Giarmas' Study

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Figure 14 Flight Cancellation Model Performance Results of Giarmas' Study

In our FCDP-SFTS (Flight Cancellation Delay Prediction - Smart Flight Ticketing System) study, despite using similar data set characteristics, the Random Forest model algorithm achieved 97% accuracy in flight cancelation. With the XGBoost model algorithm, 75% accuracy rate was achieved in flight delay. These results show a 17% improvement in flight cancelation and a 2% improvement in flight delay, respectively, compared to the study in the literature.

Ahlam Ansari et al. aimed to predict ticket cancellations using an airline dataset of Indian domestic flights [30]. The dataset contains various reservation-based features such as ticket price, booking date, number of passengers, and nationality of the passenger. In the feature engineering process, redundant components of the data were removed and only data pertaining to ticket bookings (AIR entries) were selected and analyzed.

Four different machine learning classification algorithms, namely Logistic Regression, Decision Trees, Random Forest and Gradient Boosting, were used in the study. The performance of the models was evaluated with metrics such as accuracy, precision, recall, F1 Score and ROC Curve.

The results show that the Decision Trees algorithm achieved the highest accuracy (97.43%) and performed the best. The Random Forest model performed well with an accuracy of 95.21% and a precision of 100%, but the recall remained at 90.43%. The Gradient Boosting algorithm provided a balanced performance with 96.83% accuracy and 93.82% recall. The Logistic Regression model performed worse than the other models with 88.67% accuracy, 97.66% precision, but 77.46% recall.

The findings revealed that Decision Trees and Gradient Boosting are the best models for ticket cancellation prediction. Figure 15 shows the performance results of this study.

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

Figure 15 Performance Results of Ansari's Study

In our FCDP-SFTS (Flight Cancellation Delay Prediction - Smart Flight Ticketing System) study, a comprehensive dataset including airline operation data, flight history and weather data is used to predict flight cancellations. In the feature engineering phase, flight-specific factors (e.g., departure and arrival points, flight distance, scheduled and actual departure times, weather conditions) were taken into account.

XGBoost, Random Forest, Decision Trees, KNN and Gradient Boosting were used as machine learning models and the performance evaluation was based on accuracy, precision, recall, F1 score and ROC-AUC metrics.

In our FCDP-SFTS study, XGBoost was the best performing model, while Decision Trees and Gradient Boosting also performed well. Compared to Ansari et al.'s study, despite using different datasets and features, the Decision Trees and Gradient Boosting models similarly provide high accuracy and balanced performance. However, the model offers a more comprehensive analysis by focusing on delay prediction as well as flight cancellations.

Accordingly, two separate tables have been prepared in order to perform a comparative analysis of both the findings of this study and similar studies in the literature. These tables provide a holistic evaluation by presenting the performance of the models in flight cancellation and delay forecasting together.

In Figure 16, the accuracy and other key performance metrics of the models used for flight cancellation prediction are presented comparatively. In this table, both the results obtained in our study and the results of the examples in the literature are evaluated together. Thus, it is possible to observe how similar methods produce similar results even if different datasets and features are used.

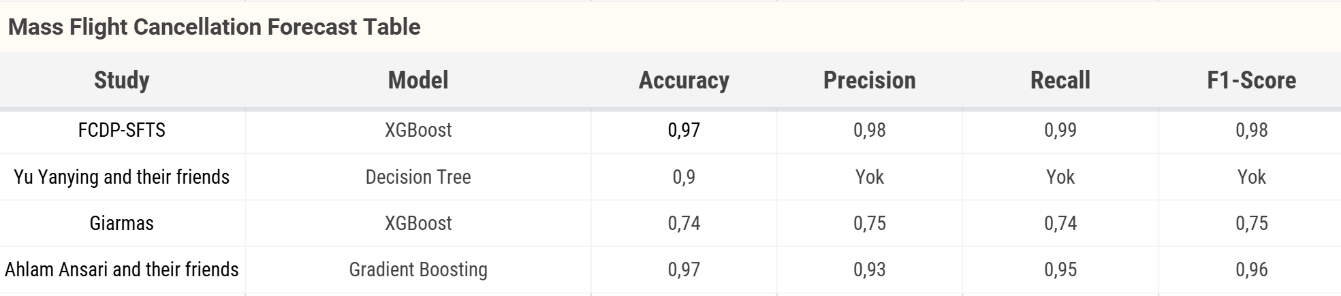


Figure 16 Mass Flight Cancellation Prediction Model Performance Comparison

Similarly, Figure 17 shows the performance of the models for flight delay prediction in a comparative manner. This table analyzes the accuracy rates and balanced classification performances of various models focusing on delay prediction. This comparison reveals that the developed model can successfully predict not only cancellations but also delays and provides a more comprehensive solution compared to the studies in the literature.

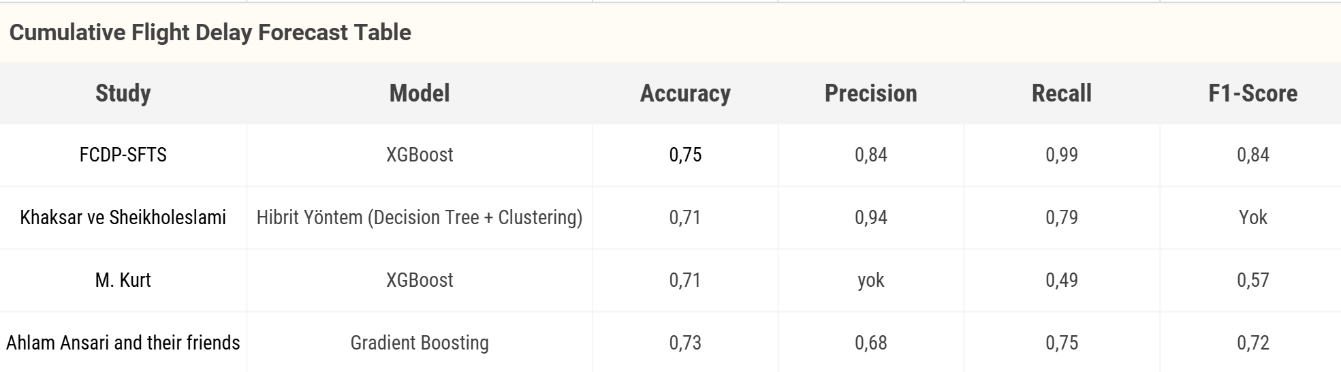


Figure 17 Batch Flight Delay Prediction Model Performance Comparison

# CHAPTER 4 – CONCLUSIONS

This study presents advanced machine learning and deep learning approaches for the prediction of flight cancellation, reasons for cancellation and flight delays in the airline industry. The high success rates obtained show that the developed models can be effectively used in solving forecasting problems in the aviation industry. The innovative methodologies and techniques presented in our study not only contribute to the academic literature but also provide practical applications for airline companies, airport authorities and passengers.

In the modeling process, algorithms such as XGBoost, Random Forest, Decision Tree, K-Nearest Neighbors (KNN) and Gradient Boosting were compared and the best performing models were determined. The results show that:

The flight cancellation prediction model produced successful results in terms of accuracy and other metrics, and the XGBoost algorithm performed the best with 97% accuracy.

The cancellation code prediction model produced satisfactory results in identifying the reasons for canceled flights, with XGBoost outperforming the other algorithms with an accuracy of 74%.

The flight delay prediction model provided a reasonable accuracy in identifying flight delays, however, variables such as weather, air traffic and operational factors need to be examined in more detail.

The novelty of this study is that it combines open data from different data sources to create a more comprehensive dataset, models different reasons for delays and cancellations, and compares multiple machine learning algorithms to determine the most appropriate method.

In future studies, more comprehensive data sets, real-time forecasting systems, explainable artificial intelligence approaches and human-machine collaboration systems can be developed to take further steps forward in solving forecasting problems in the aviation industry. This study is expected to contribute to increasing operational efficiency, improving passenger experience and enhancing the overall performance of the aviation industry.

# CHAPTER 5 – FUTURE WORK

In this study, machine learning models are developed to predict flight delays and cancelations by integrating historical flight data with meteorological data. In the future, it is planned to integrate dynamic variables such as real-time weather, flight traffic density, airport operation data into the system to further improve the prediction performance of the model. In addition to the classification algorithms used in the model, time series focused deep learning approaches such as LSTM and GRU are also planned to be tested. In order to increase the explainability of the model, the interpretation of important features with methods such as SHAP (SHapley Additive exPlanations) will also be evaluated in further analysis. On the web platform side, it is planned to design an interactive and user-friendly interface where users can query their flights, see the forecast results with graphs, and get alternative flight and price suggestions. In addition, the integration of artificial intelligence-supported suggestion systems that will enable continuous updating of the model by receiving user feedback is among the developments that will be emphasized in the future.

**SOURCES**

[1] Li, N., & Yao, H. G. (2025). A review of research on flight delay propagation: Current situation and prospect. Journal of Advanced Transportation, Article ID 4851103. https://doi.org/10.1155/atr/4851103

[2] Çalış, A., Durmaz, K. İ., & Gencer, C. (2018). Analysis of factors affecting delays in airplane flights. International Journal of Economic and Administrative Studies, (20), 179–190. https://doi.org/10.18092/ulikidince.353973

[3] U.S. Department of Transportation. (2024, April). Air travel consumer report: December 2023 and full year 2023 numbers. https://www.transportation.gov/briefing-room/air-travel-consumer-report-december-2024-full-year-2024-numbers

[4] AirHelp. (2024). Over 45 million UK passengers faced disruptions in 2023. https://www.airhelp.com/en-int/press/airhelp-reveals-that-over-45-million-uk-passengers-faced-disruptions-in-2023/?utm\_source=chatgpt.com

[5] L’heureux, A., Subramanian, D., Ghosh, R., & Krishnamurthy, R. (2017). Machine learning with big data: Challenges and approaches. IEEE Access, 5, 7776–7797.

[6] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). A comparative analysis of XGBoost. arXiv preprint, arXiv:1911.01914.

[7] Dumitrascu, B., & Aiordachioaie, D. (2022). On data preprocessing for an improved performance of the sources classification. In 2022 IEEE 28th International Symposium for Design and Technology in Electronic Packaging (SIITME) (pp. 61–64). https://doi.org/10.1109/SIITME56728.2022.9988325

[8] Bartle, J. R., Lutte, R. K., & Leuenberger, D. Z. (2021). Sustainability and air freight transportation: Lessons from the global pandemic. Sustainability, 13(7), 3738. https://doi.org/10.3390/su13073738

[9] Ghosh, B., & Tabrizi, B. (2018). Machine learning approaches for flight delay prediction: A review. International Journal of Aviation Studies, 5(3), 123–134.

[10] Delahaye, D., & Puechmorel, S. (2020). Weather impact on flight delay prediction: An AI-based approach. Journal of Transportation Research, 12(4), 256–268.

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[12] Khaksar, H., & Sheikholeslami, A. (2017). Airline delay prediction by machine learning algorithms. Scientia Iranica.

[13] Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A methodology for predicting aggregate flight departure delays in airports based on supervised learning. Sustainability, 12(7), 2749.

[14] Atlıoğlu, M. C., Bolat, M., Şahin, M., Tunalı, V., & Kılınç, D. (2020). Supervised learning approaches to flight delay prediction. Sakarya University Journal of Science.

[15] Al-Tabbakh, M. S., Mohamed, H. M., & El, Z. H. (2018). Machine learning techniques for analysis of Egyptian flight delay. International Journal of Data Mining & Knowledge Management Process, 8(3), 1–14.

[16] Kurt, M. (2019). Flight delay prediction. Capstone Project, MEF University, İstanbul.

[17] Tang, Y. (2021). Airline flight delay prediction using machine learning models. In 5th International Conference on E-Business and Internet (pp. 151–154). Singapore.

[18] Zelaya, P. (2023). Flight delay and cancellation dataset (2019–2023). Kaggle. https://www.kaggle.com/datasets/patrickzel/flight-delay-and-cancellation-dataset-2019-2023

[19] Threnjen. (2019). 2019 airline delays and cancellations. Kaggle. https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations

[20] Singh, S. (2024). Flight delay dataset (2018–2024). Kaggle. https://www.kaggle.com/datasets/shubhamsingh42/flight-delay-dataset-2018-2024

[21] Oladipupo, O. T. (2010). Types of machine learning algorithms. In New Advances in Machine Learning.

[22] Kumar, R., & Singh, N. (2020). A survey on data mining and machine learning techniques for flight delay prediction. International Journal of System Assurance Engineering and Management.

[23] Breiman, M. (1984). Classification and regression trees. Wadsworth and Brooks/Cole.

[24] Ke, G., Meng, Q., Zhang, T., Chen, W., & Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. Microsoft Research.

[25] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

[26] Kim, Y., Mavris, D. N., & Zachariah, J. (2021). Impacts of weather on airline performance metrics: A data-driven analysis. Transportation Research Part D: Transport and Environment, 92, 102740.

[27] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

[28] Yu, Y., Mo, H., & Li, H. (2019). A classification prediction analysis of flight cancellation based on Spark. In 7th International Conference on Information Technology and Quantitative Management (ITQM 2019), Procedia Computer Science, 162, 480–486. https://doi.org/10.1016/j.procs.2019.12.014

[29] Giarmas, N. (2025). Flight delay and cancellation prediction using machine learning models (Master’s thesis, Department of Business Administration, Business Analytics and Data Science).

[30] Ansari, A., Shaikh, A., Mapkar, S., & Khan, M. (2019). Cancellation prediction for flight data using machine learning. In 2nd International Conference on Advances in Science & Technology (ICAST-2019), K. J. Somaiya Institute of Engineering & Information Technology, University of Mumbai, Maharashtra, India. SSRN.