

Jerzy MANEROWSKI

Air Force Institute of Technology
 e-mail: jerzy.manerowski@itwl.mil.pl, ORCID: 0000-0003-3556-3885

Krzysztof CUR

Polish Air Force University
 e-mail: k.cur@law.mil.pl; ORCID: 0000-0003-4552-445X

Paweł GOŁDA

Polish Air Force University
 e-mail: p.golda@law.mil.pl; ORCID: 0000-0003-4066-7814

Karol PRZANOWSKI

SGH Warsaw School of Economics
 e-mail: kprzan@sgh.waw.pl; ORCID: 0000-0002-3507-7492

DOI: 10.55676/asi.v4i2.79

PREDICTIVE MODELING OF FLIGHT DELAYS USING DECISION TREE

MODELOWANIE PREDYKCYJNE OPÓŹNIEŃ LOTÓW Z WYKORZYSTANIEM DRZEW DECYZYJNYCH

Abstract

Nowadays, although technology has developed on an unimaginable scale, there are still factors that can disrupt the safe and smooth functioning of many areas of daily life. One such factor are delays. Unquestionably, they are an undesirable and, in some cases, even dangerous element. A particular case in point may be air traffic, which is one of the most technologically advanced areas. However, air traffic delays, which occur quite frequently, have made it desirable to study this area based on airport capacity modelling and machine learning methods, with the main focus on decision tree algorithms. Based on these decision tree methods, the result of acquiring and processing data and variables has been the creation of specific models that can support air traffic management and, consequently, the levelling of the resulting delays.

Keywords: air transport, decision tree, delays of aircraft

Streszczenie

Współcześnie, choć technologia rozwinęła się na niewyobrażalną skalę, wciąż istnieją czynniki, które mogą zakłócić bezpieczne i sprawne funkcjonowanie wielu obszarów codziennego życia. Jednym z nich są opóźnienia. Niewątpliwie są one elementem niepożądanym, a w niektórych przypadkach nawet niebezpiecznym. Szczególnym przypadkiem może być ruch lotniczy, który jest jednym z najbardziej zaawansowanych technologicznie obszarów. Jednak występujące dość często opóźnienia w ruchu lotniczym sprawiły, że pożądanym stało się badanie tego obszaru w oparciu o modelowanie przepustowości lotnisk i metody uczenia maszynowego, z głównym naciskiem na algorytmy drzew decyzyjnych. W oparciu o te metody drzew decyzyjnych, wynikiem pozyskiwania i przetwarzania danych i zmiennych było stworzenie konkretnych modeli, które mogą wspierać zarządzanie ruchem lotniczym, a w konsekwencji niwelowanie powstałych opóźnień.

Słowa kluczowe: transport lotniczy, drzewa decyzyjne, opóźnienia samolotów

1. INTRODUCTION

1.1. OVERVIEW OF CHALLENGES IN AIR TRAFFIC MANAGEMENT, FOCUSING ON FLIGHT DELAYS

An overview of challenges in air traffic management, focusing on flight delays, reveals that they significantly impact airlines, passengers, and overall operational efficiency. Various factors contribute to flight delays, including airport operations, weather conditions, air traffic control constraints, and airline scheduling. Research has focused on understanding the patterns and interdependencies of flight delays to develop effective strategies for mitigation.

Studies¹ highlight the importance of recognizing the characteristics and propagation effects of flight delays to improve scheduling and operational decision-making. Additionally, another research² emphasizes the development of predictive models and optimization approaches to address air traffic delays at a systemic level.

Furthermore, the work³ introduces multi-airport ground-holding strategies to optimize airport capacity allocation and reduce delays. Operational factors⁴ also play a role in causing flight delays, with air traffic control facing coordination challenges, especially during peak traffic periods.

Collaborative decision-making and accurate delay prediction are essential for effective air traffic scheduling and resource management⁵. Understanding the dynamic nature of air traffic flow is crucial for managing congestion and minimizing delays⁶.

In conclusion, addressing flight delays in air traffic management requires a comprehensive approach that considers various factors such as airport operations, weather conditions, air traffic control constraints, and airline scheduling. By developing predictive models, optimization strategies, and collaborative decision-making processes, the aviation industry can work towards reducing flight delays and improving overall system efficiency.

1.2. THE ROLE OF TECHNOLOGY IN MANAGING AND PREDICTING FLIGHT DELAYS

The role of technology in managing and predicting flight delays is crucial for enhancing operational efficiency and passenger satisfaction in air traffic management. Advanced technologies, particularly machine learning and deep learning algorithms, have been increasingly utilized to develop accurate and reliable methods

¹ Wu, Lin, and Ji, 'An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics'.

² Moreno Rebollo and Balakrishnan, 'Characterization and Prediction of Air Traffic Delays'; Bertsimas, Lulli, and Odoni, 'An Integer Optimization Approach to Large-Scale Air Traffic Flow Management'.

³ Jiang and Xie, 'Study of the Multi-Airport Ground-Holding Strategy Model and Application'.

⁴ Zámková, Prokop, and Stolín, 'Factors Influencing Flight Delays of a European Airline'.

⁵ Zhang, 'Spatio-Temporal Data Mining for Aviation Delay Prediction'.

⁶ Li et al., 'Fast Evaluation of Aircraft Icing Severity Using Machine Learning Based on XGBoost'.

for predicting flight delays⁷. Studies such as those by⁸ emphasize the maturity and effectiveness of flight delay prediction methods based on machine learning and artificial intelligence technologies. These technologies enable the classification and prediction of flight delays, thereby preventing and reducing economic losses associated with delays. Furthermore, research by other authors⁹ highlights the application of deep learning algorithms, such as CNN-based models, for predicting flight delay propagation and classifying delays effectively. Additionally, other studies¹⁰ demonstrates the successful application of machine learning approaches, including artificial neural networks, for flight departure delay prediction and analysis.

Moreover, technology plays a significant role in optimizing air traffic flow and scheduling. Studies¹¹ focus on probabilistic flight delay predictions using machine learning and trajectory-based allocation models for collaborative air traffic management programs. These technologies contribute to improving the accuracy of delay predictions and enhancing decision-making processes in air traffic management.

By leveraging advanced technologies, the aviation industry can enhance operational efficiency, reduce economic losses, and improve overall air traffic management systems.

1.3. CONTEXT OF AIR TRAFFIC DELAYS AND THEIR IMPACT ON OPERATIONS AND SAFETY

Air traffic delays significantly impact aviation operations and safety, affecting productivity, passenger loyalty, and overall system efficiency. Studies highlight cascading effects such as decreased aircraft productivity and disrupted crew scheduling¹². Managing delays is complex, especially when sector capacities and traffic demand are not aligned¹³. Network resilience is crucial for maintaining safety and efficiency during disturbances¹⁴. Departure delays often have a significant influence on arrival delays, creating a ripple effect throughout the system¹⁵. Effective decision-making for rerouting and holding aircraft must balance flight safety with operational costs¹⁶.

⁷ Ziółkowski, Małachowski, Oszczypała, Szkutnik-Rogoż, Konwerski, 'Simulation model for analysis and evaluation of selected measures of the helicopter's readiness'.

⁸ Jia et al., 'Flight Delay Classification Prediction Based on Stacking Algorithm'; Qu, Wu, and Zhang, 'Flight Delay Propagation Prediction Based on Deep Learning'; Zhang, 'Spatio-Temporal Data Mining for Aviation Delay Prediction'.

⁹ Bisandu et al., 'A Deep Feedforward Neural Network and Shallow Architectures Effectiveness Comparison: Flight Delays Classification Perspective'.

¹⁰ Esmailzadeh and Mokhtarimousavi, 'Machine Learning Approach for Flight Departure Delay Prediction and Analysis'.

¹¹ Zoutendijk and Mitici, 'Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem'.

¹² Wong and Tsai, 'A Survival Model for Flight Delay Propagation'.

¹³ Palopo, Chatterji, and Lee, 'Interaction of Airspace Partitions and Traffic Flow Management Delay'.

¹⁴ Xu and Zhang, 'Statistical Analysis of Resilience in an Air Transport Network'.

¹⁵ Zhen et al., 'A Deep Learning Approach for Short-Term Airport Traffic Flow Prediction'.

¹⁶ Zhang, 'Spatio-Temporal Data Mining for Aviation Delay Prediction'.

Controlling delays is essential to minimize economic impacts¹⁷. Efficient air traffic control systems are necessary to mitigate delays and their repercussions.

Technological advancements have also greatly influenced air traffic management, improving efficiency, safety, and sustainability. Innovative technologies have transformed the control, monitoring, and optimization of air traffic. Advanced communication and surveillance systems, like System-Wide Information Management (SWIM), enhance data exchange, situational awareness, and decision-making¹⁸. Predictive modeling and optimization tools, using machine learning and deep learning techniques, forecast delays and optimize scheduling¹⁹. New concepts such as Distributed Air/Ground Traffic Management leverage automation and AI to streamline operations²⁰. Digital twins and simulation tools allow for testing and validating strategies, improving operational resilience.

In conclusion, air traffic delays and the evolution of air traffic management technology significantly impact aviation operations and safety. Understanding delay propagation, optimizing airspace management, and leveraging advanced technologies are crucial for mitigating delays and ensuring safe, efficient air travel.

2. INTRODUCTION TO DECISION TREE ALGORITHMS

Decision tree algorithms are powerful tools in machine learning used for classification and prediction tasks. They segment the predictor space into simple regions, making them interpretable and suitable for feature selection and classification²¹. These algorithms are particularly relevant due to their ability to handle complex decision-making processes and extract valuable insights from data.

Decision trees automatically select predictive variables and accurately classify data²². They have been applied in various domains, including air quality classification²³, botnet detection²⁴, and traffic flow prediction. Their importance in predicting flight delays is significant, as they help extract relevant features, identify patterns, and make accurate predictions²⁵.

¹⁷ Chen et al., 'An Empirical Study on the Indirect Impact of Flight Delay on China's Economy'.

¹⁸ Graupl, Mayr, and Rokitansky, 'A Method for SWIM-Compliant Human-in-the-Loop Simulation of Airport Air Traffic Management'.

¹⁹ Izdebski, Gołda, Zawisza, 'The Use of Simulation Tools to Minimize the Risk of Dangerous Events on the Airport Apron'.

²⁰ Bilmoria et al., 'FACET: Future ATM Concepts Evaluation Tool'.

²¹ Kalliguddi and Leboulluec, 'Predictive Modeling of Aircraft Flight Delay'.

²² Khan et al., 'An Adaptive Multi-Layer Botnet Detection Technique Using Machine Learning Classifiers'.

²³ Hamami and Dahlan, 'Air Quality Classification in Urban Environment Using Machine Learning Approach'.

²⁴ Duan et al., 'A Novel and Highly Efficient Botnet Detection Algorithm Based on Network Traffic Analysis of Smart Systems'.

²⁵ Khan et al., 'An Adaptive Multi-Layer Botnet Detection Technique Using Machine Learning Classifiers'.

Methods like the J48 Decision Tree, combined with K-means clustering, enhance flight dataset training and prediction outcomes²⁶. Decision trees, alongside other machine learning algorithms, have been effective in predicting flight time deviations²⁷. Compared with other classifiers, decision trees are scalable and efficient, making them valuable for large-scale air traffic scenarios²⁸. They offer high prediction accuracy and valuable insights into the decision-making process, making them essential for analyzing complex datasets.

While decision trees are a primary focus, other predictive modeling techniques also play a crucial role in air traffic management. Some authors emphasized precise delay classification and value prediction for operational decision-making. Hybrid ARIMA-LR models have been highlighted for improved predictive accuracy. The effectiveness of deep learning techniques like ANN, LSTM, GRU, and CNN has been demonstrated for short-term airport traffic flow prediction. Safety performance functions have been developed to forecast separation minima infringements, enhancing safety measures. A predictive autoregression model combining SVM with polynomial and autoregression models has been introduced to improve traffic flow prediction accuracy.

3. IDENTIFICATION OF GAPS IN EXISTING RESEARCH THAT THE CURRENT STUDY AIMS TO ADDRESS

Identification of gaps in existing research that the current study aims to address involves focusing on the limitations and unexplored areas in the field of flight delay prediction. While previous studies have made significant progress in areas such as empirical analysis, causal relations, delay propagation modeling, and prediction methods²⁹, there are still gaps to be addressed. For instance, some research has focused on the delay of previous flights without considering the effects of inter-aircraft propagated delays (network effects)³⁰. Additionally, existing centrality and causality metrics have been found inadequate in characterizing the effect of delays in the air traffic system. Moreover, while some studies have successfully predicted aggregate flight departure delays based on supervised learning methods, there is a need to delve deeper into individual flight delays and consider factors like aircraft types and carriers that may impact delays³¹. Furthermore, the impact of external factors such as the COVID-19 pandemic on flight delays remains an area that requires more exploration. Therefore, the current study aims to bridge these gaps by developing more accurate and reliable methods for predicting flight delays, incorporating network effects, refining causality metrics, and considering individual flight characteristics to enhance the effectiveness of delay prediction models.

²⁶ Jia et al., 'Flight Delay Classification Prediction Based on Stacking Algorithm'.

²⁷ Stefanovič, Štrimaitis, and Kurasova, 'Prediction of Flight Time Deviation for Lithuanian Airports Using Supervised Machine Learning Model'.

²⁸ Tong, Qu, and Prasanna, 'High-Throughput Traffic Classification on Multi-Core Processors'.

²⁹ Tang, Kay, and He, 'Toward Optimal Feature Selection in Naive Bayes for Text Categorization'.

³⁰ Li et al., 'Fast Evaluation of Aircraft Icing Severity Using Machine Learning Based on XGBoost'.

³¹ Wang, 'A Note on Logistic Regression and Logistic Kernel Machine Models'.

4. CLASSIFICATION OF FLIGHT DELAYS

Due to the lack of comprehensive regulations in U.S. federal law regarding compensation for delayed/canceled flights, there are no standardized definitions for different categories of delays based on measurable values such as time or distance. Therefore, this study uses regulations from the European Union, specifically Regulation (EC) No 261/2004 of the European Parliament and of the Council of 11 February 2004, which establishes common rules on compensation and assistance to passengers in the event of denied boarding, cancellation, or long delays of flights. Based on these regulations and literature, the following categories of delays were identified:

- **Early Arrival:** The aircraft lands before the estimated arrival time.
- **On Time:** The aircraft arrives on time or with a delay of up to 15 minutes.
- **Delayed:** The aircraft arrives at the destination airport with a delay ranging from 15 minutes to 2 hours.
- **Significantly Delayed:** The aircraft arrives at the destination airport with a delay of more than 2 hours.

From these categories, we can identify four basic levels:

- Level 1: Early arrival
- Level 2: On time
- Level 3: Delayed
- Level 4: Long delayed

Additionally, an extra level is defined for exceptional situations:

- Level 5: Extraordinary situations, such as massive delays exceeding at least 6 hours beyond the estimated arrival time, etc.

Types of Delay Reasons

Unlike delay classification, the reasons for delays have been directly classified by the U.S. Department of Transportation. The department defines five possible types of delays:

- **Carrier Delay:** Delays within the control of the airline, such as aircraft cleaning, maintenance, loading baggage, legal rest for crew, waiting for connecting passengers and baggage, damage from hazardous materials, fueling, catering, slow boarding, overbooking, and documentation issues.
- **Late Arrival Delay:** Delays resulting from the late arrival of the previous flight, causing a ripple effect on subsequent flights.
- **National Airspace System (NAS) Delay:** Delays due to air traffic control, weather conditions (excluding extreme weather), airport operations, and high air traffic volumes.
- **Security Delay:** Delays related to security issues, including terminal evacuations, re-boarding due to security breaches, malfunctioning security equipment, and long security check lines exceeding 29 minutes.
- **Weather Delay:** Delays due to extreme weather conditions at the departure airport, along the flight route, or at the arrival airport.

These delay types can occur independently, as factors like bad weather do not influence overbooking by an airline. Therefore, knowing the specific delay values allows for the calculation of the total delay value based on the sum of the individual components. The total delay can be represented by the following formula:

$$O = P + PP + K + B + A \quad O = P + PP + K + B + A,$$

where:

O = Total delay;

P = Carrier delay;

PP = Late arrival delay;

K = Air traffic control delay;

B = Security delay;

A = Weather delay.

Understanding the individual components of delays and being able to determine their total value is essential for data analysis and building a model to predict the likelihood and level of delays. Given that flight information has been regularly collected and published since 1987 and the U.S. has the most airports globally, accounting for about 30% of the world's passenger air transport, the dataset provides an impressive amount of information on hundreds of millions of flights over the years. Therefore, selecting an appropriate time range was crucial for proper and efficient analysis.

For this study, data from January 2018 to December 2022 was selected. During this period, the number of recorded flights exceeded 30 million. The data files downloaded from the Bureau of Transportation Statistics website were in CSV format, requiring proper formatting for analysis. The files were processed using Excel, and due to Excel's limitation of approximately one million rows per sheet, a data model was created to allow simultaneous analysis of over 30 million cases.

5. DATA ANALYSIS AND MODEL DEVELOPMENT

5.1. DATA SET

This analysis used data from the United States Department of Transportation, covering flights operated by American carriers domestically. According to US regulations, aircraft crews report detailed flight information, which is published monthly by the Bureau of Transportation Statistics.

The data includes time-related details (year, month, day, etc.), airline information (name, DOT and IATA codes, tail number, flight number), and cancellation/diversion data (cancellation status and reason, diversion status). It also covers departure and arrival performance (scheduled and actual times, delays, taxi times) and flight-specific details (elapsed flight times, air time, number of flights, distance, distance groups).

This data was crucial for analyzing flight delays, diversions, and airline performance, requiring careful processing and elimination of correlated information based on the analysis's purpose.

5.2. MODEL DEVELOPMENT

To develop a decision tree model, start by collecting and preparing your data, ensuring it's clean and properly encoded. Split the data into training and testing sets. Select important features through analysis and statistical methods. Build the model using a decision tree algorithm and train it on the training data. Optimize the model by pruning to avoid overfitting and tuning hyperparameters. Evaluate its performance on the test set using appropriate metrics. Visualize the decision tree and analyze feature importance to interpret the model's decisions. Finally, implement the model in production, monitor its performance, and document the process thoroughly.

5.3. TECHNIQUES FOR MODEL EVALUATION AND VALIDATION

The target function is `default60`, defined as follows: `default60=(ArrDelayMinutes>60)`, so delay of a particular airplane is greater than 60 minutes. There were tested also other various benchmarks for delay, like 120, 30 or 90 minutes. All options resulted similar tree decision structure and variable importance, so 60 minutes is used as an example of delay event.

The tree decision model is built in SAS Viya 3.5 by application Model Studio. It is the tool with In-Memory data-processing technology based on dedicated server called CAS (Cloud Analytical Services) – The Architecture of the SAS® Cloud Analytic Services in SAS® Viya™.

The model is built on random sample with 492 850 rows, what is 0.002398377 share of the whole data. Even if it is small sample, the model is stable over time on don sample datasets and has a good predictive power calculated on the whole data.

In general the statistics AUC is equaled to 96.33%, what means that if you take two airplanes from the available data, where one is delayed and the other is not, then the model with the about 96.33% probability is able to identify what is more likely to delay, so there is only $100\% - 96.33\% = 3.67\%$ of uncertainty in identification of delayed airplanes.

The main parameters of the tree: splitting criterion – Entropy, the depth of three – 6, min number of cases per leaf – 200, subtree method – C4.5 with p-value – 0.25.

In the model variables there are included information only available up to the moment of departers of airplanes, so based on the model we can calculate a probability that an airplane will be delayed more than 60 minutes, so after taking off an airline can incorporate some actions to minimize a cost of delay.

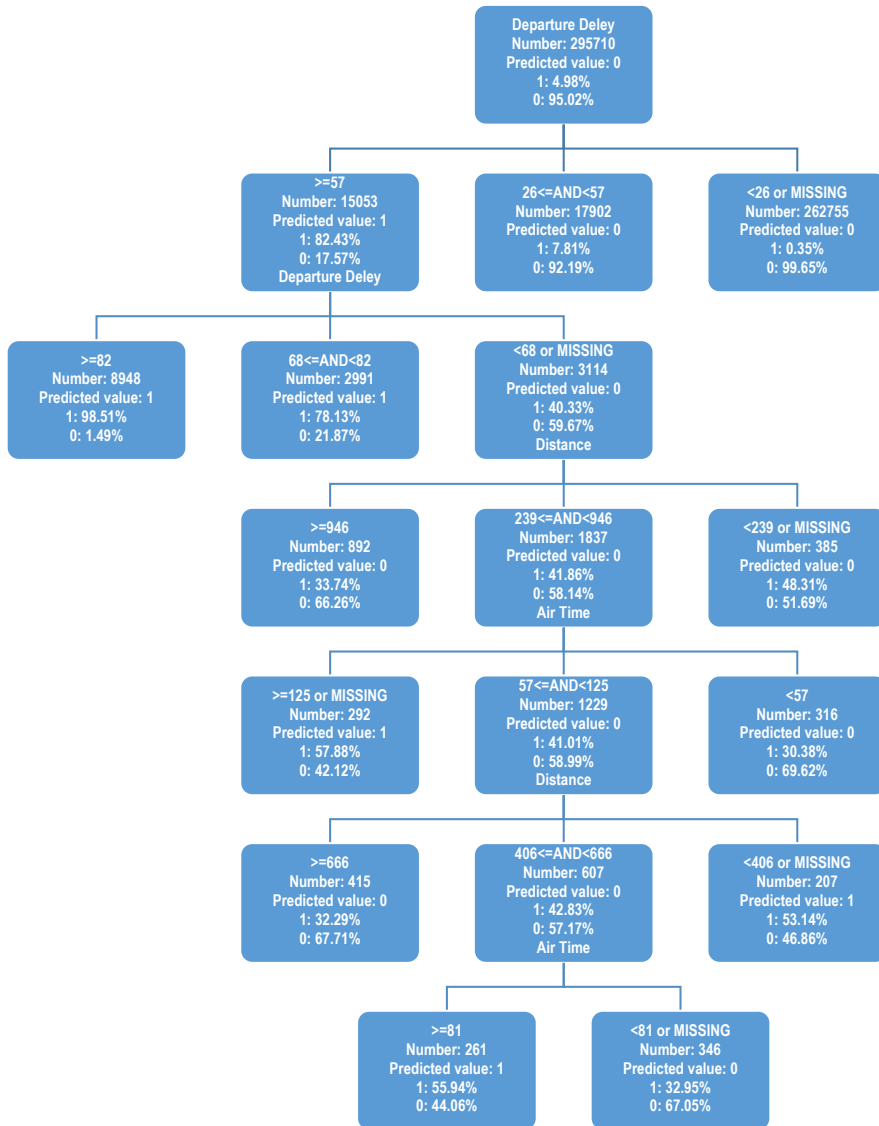


Figure 1. The tree decision with leaves identification [own elaboration]

The tree decision with leaves identification is presented in Figure 1.

Table 1. The main model measures on modelling data sets [own elaboration]

Modeling data set	Size	Gini	AUC
Train	295 710	92.45%	96.23%
Validate	147 854	92.56%	96.28%
Test	49 286	92.91%	96.46%
Whole data	205 493 100	92.65%	96.33%

The table 1 provides a comprehensive summary of the modelling data set’s performance metrics, including size, Gini coefficient, and Area Under the Curve (AUC) for the Train, Validate, Test, and Whole data subsets.

The training set, consisting of 295 710 observations, is used to train the predictive model. The validation set, which includes 147 854 observations, helps in tuning the model and selecting the best parameters to prevent overfitting. The test set, containing 49 286 observations, is utilized to evaluate the model’s performance on new, unseen data. The entire data set, comprising 205 493 100 observations, provides a comprehensive evaluation of the model.

The Gini coefficient is a measure of model performance, with higher values indicating better predictive accuracy. The Gini coefficients for the training, validation, test, and whole data sets are 92.45%, 92.56%, 92.91%, and 92.65%, respectively. These consistently high values indicate that the model performs well across all subsets, with a slight increase in the Gini coefficient from training to test sets, suggesting good generalization to new data.

The AUC measures the model’s ability to distinguish between classes, with values closer to 1 indicating higher effectiveness. The AUC values for the training, validation, test, and whole data sets are 96.23%, 96.28%, 96.46%, and 96.33%, respectively. These high AUC values across all subsets demonstrate the model’s excellent discriminatory power and confirm its robustness and accuracy in predicting outcomes.

Overall, the table indicates that the predictive model is highly accurate and reliable, with consistently high Gini coefficients and AUC values across all data subsets. The model performs well on both training and validation data, maintaining strong performance on the test set, which demonstrates its ability to generalize to unseen data. The metrics for the whole data set further reinforce the model’s overall effectiveness and robustness in predicting the target variable.

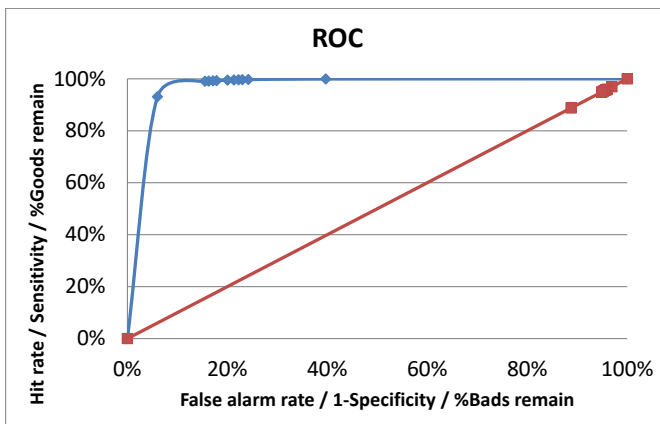


Figure 2. The ROC curve on whole data [own elaboration]

Table 2. The importance of variables [own elaboration]

Variable	Relative importance
DepDeleay	99.83%
AirTime	0.11%
Distance	0.06%


The visualization of variable importance presents that the most important variable is DepDeleay what is, of course, obvious, but due to the tree decision now it is shown by numbers and after tree decision study we can see, that airplane with delay ≥ 57 minutes has only 17.57% chance to be not delayed more than 60 minutes. Further investigation should focus on more advance models and data, especially it would be a valuable to use weather data or traffic and capacity of airport. Based on extended data the problem could be defined in better way, and especially the variable DepDeleay could be less important. Nevertheless, based on the current data can be observed that the biggest reason of air delays is delay of departer and that factor can be treated as a separate case in modelling.

Finally in the tree model there are used also variables like: Distance and AirTime.

It is also possible to prepare some additional reports called prediction plots, separate graphs for every variable used in the tree, there are not presented due to very intuitive results. On every graph can be observe the evolution of Delay rate over all possible values of variables. In case of variables like AirTime and Disnance plots are almost flat. For AirTime can be observed monotonic relation between time of travel and probability of delay, longer time – bigger change of delay, But after 150 minutes the plot is stabilized and probability of delay is constant and equal about 5%.

For DepDeleay variable there is a big shift in prediction around 23 minutes from almost zero delay rate up to 8%. That information gives a business factor for further investigation, namely to define some service level agreement of departer delays, to not exceed 23 minutes.

Table 3. The measures of every tree leaf [own elaboration]

 Tree leaf id	Size	Share	Delay rate	Cumulative delay rate	Average probability of delay
4	6 253 646	3.04%	98.25%	98.25%	98.51%
5	2 009 737	0.98%	78.55%	93.46%	78.13%
10	222 471	0.11%	54.32%	92.43%	57.88%
15	142 721	0.07%	54.02%	91.80%	53.14%
16	195 462	0.10%	48.86%	90.85%	55.94%
9	282 129	0.14%	45.45%	89.44%	48.31%
7	614 910	0.30%	36.40%	86.08%	33.74%
12	222 170	0.11%	33.05%	84.90%	30.38%
13	270 127	0.13%	32.99%	83.53%	32.29%
17	234 465	0.11%	31.99%	82.37%	32.95%
2	12 590 415	6.13%	7.70%	41.56%	7.81%
3	182 454 847	88.79%	0.34%	4.96%	0.35%

As can be seen based on measures on every leaf (Table 3), selection of first or two leaves identify 3% or 4% of all airplanes with the biggest probability of delayed, more than 93%. It gives us a possibility to identify airplanes with the delay and incorporate some actions to minimize a cost of potential delay. Dependently on the target group - selected number of leaves can be organized dedicated campaign to flight customers to decrease costs of delays, the cumulative delay rate informs about possible share of delayed flights in the target group.

The shares of leaves are not similar, and this is the most weakly point of tree decision techniques, but in our case we focus only on first leaves to identify the most delayed airplanes, what is enough to use it in Airline business optimization. Further our analyses in next articles, where more different models are built, like Logistic regression or Risk Scorecards will present solutions with better splitting into delay rate groups.

The main limitation of current approach is that the model is built with the assumption that at the moment of flight departer there is the best possible time to investigate some actions, campaigns to the most probable delay airplanes to traveling customers to minimize costs of delays. In case of more data, like weather or traffics and capacity on airports, the business case can be moved into earlier time points, maybe before entering customers into an airplane. It requires more research, and it will be considered in next articles.

6. INTERPRETATION OF RESULTS IN THE CONTEXT OF AIR TRAFFIC MANAGEMENT

Interpreting the results of decision trees in air traffic management, particularly regarding limiting aircraft delays, involves analyzing the outcomes of various decisions to optimize the flow of air traffic. The decision tree model developed in this study uses a 60-minute delay threshold to predict potential delays. This model was built using SAS Viya 3.5 with a random sample of 492.850 rows from a comprehensive dataset of over 205 million records. The model's high performance, with an AUC of 96.33%, indicates its robustness and reliability in predicting delays.

The results show that the decision tree model performs consistently well across training, validation, and test data sets, with Gini coefficients and AUC values demonstrating its excellent predictive accuracy. The high AUC values, above 96% for all subsets, suggest that the model is highly effective at distinguishing between delayed and non-delayed flights. This high level of accuracy is crucial for practical applications in air traffic management, where minimizing delays is a primary objective.

In the realm of air traffic flow management, decision support tools are vital for limiting controller workload and complexity while enhancing air traffic throughput. The decision tree model serves as a decision support tool by identifying the most significant variables affecting delays, such as Departure Delay (DepDelay), which has a relative importance of 99.83%. This insight allows air traffic controllers and airline operators to focus on key factors that influence delays and implement strategies to mitigate them.

For example, the model shows that an airplane with a departure delay of 57 minutes has only a 17.57% chance of not being delayed by more than 60 minutes. This information can be used to prioritize interventions for flights that are already significantly delayed, potentially reallocating resources or adjusting schedules to minimize the impact of these delays on overall operations.

Furthermore, integrating deterministic integer programming models and decision support systems can assign delays to aircraft under capacity constraints, allowing for reactive adjustments to uncertainties in the system. This approach can be enhanced by the decision tree model's ability to predict delays accurately, enabling more effective traffic flow management and equitable distribution of delays among stakeholders.

Additionally, the implementation of new strategies, such as rerouting management during adverse weather conditions, requires decision support tools that can translate weather information into anticipated operational impacts on air traffic³². The decision tree model can aid in evaluating the impact of new aircraft separation minima and implementing arrival traffic strategies, thus minimizing delays and enhancing the efficiency of air traffic operations.

The model's insights also support the use of multi-agent systems to resolve conflicts through local speed regulation and departure delay adjustments, ultimately reducing the number of conflicts and simplifying traffic management for controllers. Furthermore, utilizing genetic algorithms for en-route airspace capacity enhancement provides a flexible framework for dynamic route utilization, contributing to more efficient air traffic and capacity management.

In conclusion, interpreting the results of the decision tree model in air traffic management involves leveraging decision support tools to optimize traffic flow, minimize delays, and enhance operational efficiency. By utilizing advanced modelling techniques, integrating integer programming models, and implementing innovative strategies, air traffic controllers can effectively manage traffic flow, reduce delays, and ensure the safe and efficient operation of air traffic systems. The high accuracy and reliability of the decision tree model developed in this study provide a strong foundation for these efforts, offering valuable insights and actionable information to improve air traffic management practices.

7. SUMMARY

The study on predictive modelling of flight delays using decision tree algorithms highlights the significant impact that delays have on air traffic management. By leveraging advanced machine learning techniques, particularly decision trees, the research provides a robust framework for understanding and mitigating flight delays. The decision tree model developed in this study, utilizing a 60-minute delay threshold, has

³² Zhang, Bianco, and Beck, 'Solving Job-Shop Scheduling Problems with QUBO-Based Specialized Hardware'.

demonstrated high accuracy and reliability, with an AUC of 96.33%, indicating its effectiveness in predicting delays.

The model's ability to identify key factors influencing delays, such as departure delay (DepDelay), and its application in real-time decision-making, offers substantial benefits for air traffic controllers and airline operators. These insights enable the implementation of targeted interventions to reduce delays, improve scheduling, and enhance overall operational efficiency. Furthermore, the integration of additional data, such as weather conditions and airport capacity, could further refine the model, providing even more accurate predictions and effective management strategies.

The consistent performance of the model across training, validation, and test data sets underscores its robustness and potential for practical application in the aviation industry. The study's findings emphasize the importance of predictive modelling in addressing the complexities of air traffic management, ensuring safe, efficient, and timely air travel.

In conclusion, the use of decision tree algorithms in predictive modelling of flight delays represents a significant advancement in air traffic management. The high predictive accuracy of the model, combined with its interpretability and practical applicability, makes it a valuable tool for optimizing air traffic flow, minimizing delays, and enhancing the efficiency of air traffic operations. Future research should focus on incorporating more comprehensive data and exploring additional modelling techniques to further improve delay predictions and management practices in the aviation industry.

REFERENCES

- Bertsimas D., Lulli G., Odoni A., 'An Integer Optimization Approach to Large-Scale Air Traffic Flow Management'. *Operations Research* 59, no. 1 (February 2011): 211–27. <https://doi.org/10.1287/opre.1100.0899>.
- Bilmoria K.D., Sridhar B., Chatterji G.B., Sheth K., Grabbe S., 'FACET: Future ATM Concepts Evaluation Tool'. *Air Traffic Control Quarterly*, 2001. <https://doi.org/10.2514/atcq.9.1.1>.
- Bisandu D.B., Homaid M.S., Moulitsas I., Filippone S., 'A Deep Feedforward Neural Network and Shallow Architectures Effectiveness Comparison: Flight Delays Classification Perspective', 2021. <https://doi.org/10.1145/3505711.3505712>.
- Chen Y., Jiang Y., Tsai S.-B., Zhu J., 'An Empirical Study on the Indirect Impact of Flight Delay on China's Economy'. *Sustainability*, 2018. <https://doi.org/10.3390/su10020357>.
- Duan L., Zhou J., You W., Xu W. 'A Novel and Highly Efficient Botnet Detection Algorithm Based on Network Traffic Analysis of Smart Systems'. *International Journal of Distributed Sensor Networks*, 2022. <https://doi.org/10.1177/15501477211049910>.

- Esmailzadeh E., Mokhtarimousavi S., 'Machine Learning Approach for Flight Departure Delay Prediction and Analysis'. *Transportation Research Record Journal of the Transportation Research Board*, 2020. <https://doi.org/10.1177/0361198120930014>.
- Graupl T., Mayr M., Rokitansky C.-H., 'A Method for SWIM-Compliant Human-in-the-Loop Simulation of Airport Air Traffic Management'. *International Journal of Aerospace Engineering*, 2016. <https://doi.org/10.1155/2016/6806198>.
- Hamami F., Dahlan I.A., 'Air Quality Classification in Urban Environment Using Machine Learning Approach'. *Iop Conference Series Earth and Environmental Science*, 2022. <https://doi.org/10.1088/1755-1315/986/1/012004>.
- Izdebski M., Gołda P., Zawisza T., The Use of Simulation Tools to Minimize the Risk of Dangerous Events on the Airport Apron, *Lecture Notes in Networks and Systems*, 2023, 604 LNNS, pp. 91–107. https://doi.org/10.1007/978-3-031-22359-4_6.
- Jia Y., Zhang H., Líu H., Zhong G., Li G., 'Flight Delay Classification Prediction Based on Stacking Algorithm'. *Journal of Advanced Transportation*, 2021. <https://doi.org/10.1155/2021/4292778>.
- Jiang X., Xie Y., 'Study of the Multi-Airport Ground-Holding Strategy Model and Application', 2016. <https://doi.org/10.2991/i3csee-16.2016.4>.
- Kalliguddi A.M., Leboulluec A.K., 'Predictive Modeling of Aircraft Flight Delay'. *Universal Journal of Management*, 2017. <https://doi.org/10.13189/ujm.2017.051003>.
- Khan R.U., Xiaosong Zhang X., Kumar R., Sharif A., Golilarz N.A., Alazab M., 'An Adaptive Multi-Layer Botnet Detection Technique Using Machine Learning Classifiers'. *Applied Sciences*, 2019. <https://doi.org/10.3390/app9112375>.
- Li S., Qin J., He M., Paoli R., 'Fast Evaluation of Aircraft Icing Severity Using Machine Learning Based on XGBoost'. *Aerospace*, 2020. <https://doi.org/10.3390/aerospace7040036>.
- Moreno R., Luis J., Balakrishnan H., 'Characterization and Prediction of Air Traffic Delays'. *Transportation Research Part C Emerging Technologies*, 2014. <https://doi.org/10.1016/j.trc.2014.04.007>.
- Palopo K., Chatterji G.B., Lee H.-T., 'Interaction of Airspace Partitions and Traffic Flow Management Delay', 2010. <https://doi.org/10.2514/6.2010-9295>.
- Qu J., Wu S., Zhang J., 'Flight Delay Propagation Prediction Based on Deep Learning'. *Mathematics*, 2023. <https://doi.org/10.3390/math11030494>.
- Stefanovič P., Štrimaitis R., Kurasova O., 'Prediction of Flight Time Deviation for Lithuanian Airports Using Supervised Machine Learning Model'. *Computational Intelligence and Neuroscience*, 2020. <https://doi.org/10.1155/2020/8878681>.
- Tang B., Kay S., He H., 'Toward Optimal Feature Selection in Naive Bayes for Text Categorization'. *Ieee Transactions on Knowledge and Data Engineering*, 2016. <https://doi.org/10.1109/tkde.2016.2563436>.
- Tong D., Qu Y.R., Prasanna V.K., 'High-Throughput Traffic Classification on Multi-Core Processors', 2014. <https://doi.org/10.1109/hpsr.2014.6900894>.

Wang R., 'A Note on Logistic Regression and Logistic Kernel Machine Models', 2011. <https://doi.org/10.48550/arxiv.1103.0818>.

Wong J.-T., Chang Tsai S., 'A Survival Model for Flight Delay Propagation'. *Journal of Air Transport Management*, 2012. <https://doi.org/10.1016/j.jairtraman.2012.01.016>.

Wu Z., Lin W., Ji Y., 'An Integrated Ensemble Learning Model for Imbalanced Fault Diagnostics and Prognostics'. *IEEE Access* 6 (2018): 8394–8402. <https://doi.org/10.1109/ACCESS.2018.2807121>.

Xu G., Zhang X., 'Statistical Analysis of Resilience in an Air Transport Network'. *Frontiers in Physics*, 2022. <https://doi.org/10.3389/fphy.2022.969311>.

Zámková M., Prokop M., Stolin R., 'Factors Influencing Flight Delays of a European Airline'. *Acta Universitatis Agriculturae Et Silviculturae Mendelianae Brunensis*, 2017. <https://doi.org/10.11118/actaun201765051799>.

Zhang J., Bianco G.L., Beck J.Ch., 'Solving Job-Shop Scheduling Problems with QUBO-Based Specialized Hardware'. *Proceedings of the International Conference on Automated Planning and Scheduling* 32 (13 June 2022): 404–12. <https://doi.org/10.1609/icaps.v32i1.19826>.

Zhang K., 'Spatio-Temporal Data Mining for Aviation Delay Prediction', 2021. <https://doi.org/10.48550/arxiv.2103.11221>.

Zhen Y., Yang H., Li F., Lin Y., 'A Deep Learning Approach for Short-Term Airport Traffic Flow Prediction'. *Aerospace*, 2021. <https://doi.org/10.3390/aerospace9010011>.

Ziółkowski J., Małachowski J., Oszcypała M., Szkutnik-Rogoż J., Konwerski J., Simulation model for analysis and evaluation of selected measures of the helicopter's readiness, *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 2022, 236(13), pp. 2751–2762. <https://doi.org/10.1177/09544100211069180>.

Zoutendijk M., Mitici M., 'Probabilistic Flight Delay Predictions Using Machine Learning and Applications to the Flight-to-Gate Assignment Problem'. *Aerospace* 8, no. 6 (June 2021): 152. <https://doi.org/10.3390/aerospace8060152>.