Leveraging Weather Data for Improved Flight Delay Prediction: A Comparative Analysis of Decision Trees and Random Forests

MSc Research Project
MSc in Data Analytics

Yash Rajesh Suryawanshi
Student ID: XXX

School of Computing
National College of Ireland

Supervisor: Abdul Qayum

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Yash Rajesh Suryawanshi		
Student ID:	x22227431		
Programme:	MSc in Data Analytics…	**Year:**	2024
Module:	Research Project		
Supervisor:	Abdul Qayum		
Submission Due Date:	12th August 2024		
Project Title:	Leveraging Weather Data for Improved Flight Delay Prediction: A Comparative Analysis of Decision Trees and Random Forests		
Word Count:	5417	**Page Count:**	19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Yash Rajesh Suryawanshi

Date: 12th August 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Leveraging Weather Data for Improved Flight Delay Prediction: A Comparative Analysis of Decision Trees and Random Forests

Yash Rajesh Suryawanshi
x22227431

Abstract

Previous research, particularly works that have used models as Random Forests to predict flight delays has shown significant improvement in the prediction accuracy by adding factors like Flight Attributes and Operational Variables. The most frequently considered variables are weather condition that can be determinant in delaying and many of the other models ignore or do not consider them. The studies often use static or simplistic methods to incorporate weather data, not depicting the complex and dynamic system of how operations are affected by the weather acting. For example, a model may use high level categorical variables for weather (e.g. clear, cloudy, rainy) without more detailed info such as wind speed or temperature changes or sudden transitions of the climate which is going to reduce its predictive value.

Furthermore, while a few studies take weather data into account their use of it remains superficial and they do not conduct extensive tests to determine the effectiveness of different machine learning approaches in exploiting this information. The emphasis is commonly on simple predictive performance without considering the reasoning behind feature importance, interpretability and there lacks understanding of which models are better suited to accomplish it.

This study is going to compare the performance of Decision Trees and Random Forests, in searching for model which can help integrate weather variables more effectively with that not only will enhance predictive accuracy but also offer a stronger solution to airlines. Improving efficiency, increasing overall passenger satisfaction and ultimately a comprehensive solution to fight flight delays.

Keywords *- Flight Delay Prediction, Machine Learning, Random Forest, Decision Tree, Weather Data Integration, Predictive Modeling, Aviation Operations, Airline Industry, Forecasting Accuracy*

1 Introduction

The aviation industry has encountered many challenges over the years, for economic efficiency as well as passenger satisfaction than flight delays. The economic impacts of flight delays are huge for airlines, airports and passengers. The Federal Aviation Administration (FAA) estimates the U.S. economy loses about $28 billion a year because of flight delays (Ball et al., 2010). Much previous research have been conducted to investigate numerous reasons that cause the flight delayed such as operational and air traffic congestion then technical due of airplanes. (Ball et al., 2010). However, the thing that was missed in most predictive models - or critical factor ignored- is of course weather conditions.

Bad weather, in turn can interfere with flight schedules such that flights have to be delayed or cancelled. In order to make delays predictions more accurate, we need a good understanding and ability to predict the effects weather has on flight delays. This paper tries to address this issue by incorporating weather conditions into flight delay prediction models and evaluating which existing data-mining algorithm, Decision Trees or Random Forests is better capable of exploiting such information.

1.1. Importance

Predicting flight delays accurately is important for the airlines to improve operational planning and passenger experience. Airlines can optimize schedules, allocate resources more efficiently and communicate with passengers in a timely manner when they have better predictions (Rodríguez–Sanz et al., 2021). In addition, knowledge of how the weather affects delays can propose timeless schedules considering that there is an environmentally stochastic factor. This work is important because it works to achieve the main objective of reducing flight delays and their associated costs by updating weather information.

1.2. Research Question and Objectives

The main research question of our study is: “How does including weather conditions into a model make the predictions more accurate for flight delays, and which ML technique (Decision Trees or Random Forests) performs better from this research”. Objective of the research

- To study the effect of weather feature incorporation on flight delay prediction accuracy.

- Objective: Evaluate Decision Tree v/s Random Forest algorithm to predict whether a flight will be delayed or not based on their weather data.

- Also, to determine the most important weather related features in predicting flight delays

1.3. Limitations

Although this study seeks to improve the predictive accuracy of flight delay models, it has some limitations as well. The researchers also note that many factors contributing to delays may not be represented in their dataset, for example real-time air traffic control decisions or sudden operational disruptions. In addition, he writes that not all useful atmospheric conditions are necessarily contained in common weather metrics used to predict flight scheduling impacts. These restrictions point to gaps in the data and hint at how difficult it is predicting flight delays with any real accuracy.

1.3. Structure of Report

The research report is structure as, Related Work for the background, Research Methodology for the approach, and section 4 Design Specification to outline the system architecture. Design Implementation and Evaluation Results Discussion A Conclusion and Future Work section then draws together results of the project, pointing out areas needing further research.

2 Related Work

One of the most repeated problems in airline aviation is delays of flight, which are not only costly but also a major source of inconvenience for travellers seasonal or regular. Flight delays are one of the most difficult types to predict thanks to a myriad of factors that affect this type: both operational, technical plus environmental variables. Most of the traditional predictions methods fail by not taking into account these factors that appear to have only a seasonal dependence, such as weather conditions for example. Recent research on machine learning underscores the potential to improve delay predictions by using extensive and rich datasets, as well as advanced algorithms.

2.1. Impact of Weather Conditions on Flight Delays.

Airlines are a weather reliant service, and delays often follow poor conditions. Some predict improved accuracy in prediction models based on the use of weather data. Borse, Y. et al. (2020) proposed a system to predict flight delays using meteorological data features like temperature, humidity and visibility. In particular, their analysis showed that flight operations are substantially affected by weather conditions and its inclusion extends the predictive power of these models. Jiang, Y. et al. Previous work of Xu et al. (2020) highlighted the use of airport weather and city weather in flight delay prediction as necessary features to consider. Multi-dimensional weather data was used by them in the form of an advanced time series model (LSTM) presenting a better generalization and prediction accuracy. Huo, J. et al. (2020) confirmed the necessity of detailed and high-quality weather datasets; they believed that models trained by machine learning could predict delays accurately only when presented with enough such fine-grained information. The greatest added value of the weather data was also underlined in their study: significant improvements for predicting performance. Li, Q., and Jing, R. (2022) does not compare decision trees to random forests specifically; however - makes use of a Random Forest model in their proposed ST-Random Forest framework. In both cases this indicates that the authors perceived value in Random Forest's capabilities to model complex inter-feature dependencies - something likely required for good weather predictions.

2.2. Machine Learning Techniques for Flight Delay Prediction

Nowadays, most popular tools in industry are Decision Trees and Random Forests because of its simplicity and robustness. Borse, Y. et al. (2020) and Tijil, Y. et al. (2024) used Decision Trees for Categorical Variables and handling non-linearity such as predicting air flight delays. The work done by Borse et al. (2020) indicated Decision Trees as interpretable and accurate predictor. According to Prabakaran, N. and Kannadasan, R. (2018) Random Forests gave more accurate result in predicting the flights delays than other algorithm The research showcased the strength of Random Forests to bulk with large datasets and complex feature spaces, while keeping overfit down. Borse, Y. et al. (2020) compared several classifiers (Decision Trees, Random Forests etc.), and found that the accuracy of random forest was higher than all others including a reliable value for stability. This could be due to the ensemble nature of Random Forests, which increases their generalization performance.

2.2.1 Comparative Analysis of Machine Learning Models

A detailed comparative study of them is provided in Gopalakrishnan, K., & Balakrishnan, H. (2017) for predicting delays in air traffic networks using various machine learning models In their research, a Random Forest outperformed single classifiers in terms of accuracy and robustness compared to other classifier especially Decision Trees. This study illustrated the need for sophisticated ML methods to address flight delay data complexities and diversities

2.2.2 Advanced Models: LSTM and TCN

Some advanced models like Long Short-Term Memory (LSTM) networks and Temporal Convolutional Networks (TCNs), also known as convolutional sequence-to-sequence, have shown to be promising for learning the temporal dependencies in sequences which contribute towards improving prediction accuracy. Jiang, Y. et al. (2020) put forward a multi-index forecasting model of LSTM network, which can accurately predict multiple delay indicators at the same time. The model itself is dualistic, including both the amount of flights at any given time and through planned departure / arrival times down to weather conditions; all in all offering superb forecast predictions across an array of trends. K, V. et al. Some works (2024) further utilized TCNs for the flight delay prediction to show that temporal dependencies in both flights and weather data can be effectively learned by the models, resulting better performance over naive approaches. This study highlighted the potential of TCNs over handling sequential data and making suggestive predictions. Few other deep learning models like the Transformer-based architectures are widely used and have showed promise for predicting flights delays. For example, the results of a study are promising in predicting which planes get delayed and that it is possible to accurately handle complexities or accommodate large scale data with Transformers as per prediction for flight delay propagation (Qu, J., Zhang, L. and Wu, S., 2023).

2.3. Comparative Performance and Hyperparameter Tuning

Recently several comparative studies have been conducted to find out performance of various machine learning algorithms for predicting delays. Borse, Y. et al. (2020) benchmarked many classifiers like Decision Trees, and Random Forests consistently resulted in superior accuracy and stability. This is due to the fact that Random Forests are an ensemble algorithm which decreases overfitting and increases generalization. Jiang, Y. et al. (2020) noted that despite the great advances in LSTM nets concerning sequential data, tabular structured data analysis was still remaining at a comparable level between REL-LSTM and Random Forest. This study demonstrated the necessity of choosing suitable models depending on the characteristics and types of prediction tasks. Hyperparameter tuning - An essential step towards model optimization in machine learning.

Sujay, V. et al. (2024) Not long after, Polson and Toplis performed several experiments in tuning hyperparameters of flight delay prediction models. This finding verified that perform a systematic tuning of hyper-parameters makes the over-all accuracy and robustness much better than using any default setting, or ensemble in the end building more reliable predictive ability. However, these advances also introduce limitations due to the ways that models are constructed today: Specifically in terms of forecasts availed or respective granularity on which weather-data can be optimized for. Even harder changes include more advanced data collection techniques (e.g. improved spatial and temporal resolution weather information) which can produce very large gains in model performance alongside their costs. Furthermore, a hybrid model that incorporates all kinds of machine learning methods would help to promise an increase in prediction accuracy level as we are expected could utilize each method on where it fits well. We leave the investigation of these hybrid models for future research, ahead with broader datasets encompassing an ecosystem of causes behind flight delays. Real-time updates and model training are also important with time for accuracy of the predictions (Jiang, Y. et al., 2020).

2.4. Limitations

Even with the advancements we went over, there are still bottlenecks in most of our methods for predicting flight delays. Weather data quality and granularity are both significant limitations, leading to noisy weather inputs that can impact a modelized prediction. Moreover, these studies have focused on specific datasets, such the one containing flights originating from individual airports or during certain time-periods. Another issue is the computational burden and resource requirements for training sophisticated models like LSTM or TCN, particularly while real-time data streams need to be integrated. Lastly, hyperparameter tuning could potentially increase the chance of overfitting with sufficiently complex and resourceful models. In particular, subsequent research should test hybrid models comprising various machine learning approaches with more generalizable datasets to enhance predictive accuracy.

3 Research Methodology

The implementation section details the systematic methodology followed for answering our research question of: "How can the integration of weather conditions improve the accuracy of flight delay predictions, and which machine learning algorithm among decision trees and random forests demonstrates superior performance in leveraging weather data for this purpose?" The methodology describes the steps and tools used. from data collection through to final analyses.

3.1 Data Collection

The dataset belongs to the flight delay, a publicly available flight delay dataset are openly available in [Kaggle at https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations/data]. These included flight details (month, day of the week, departure time block) operational factors (number of seats or passenger capacity available on each aircraft in different distance groups), and weather conditions were captured as continuous variables. The raw data was collected and processed to treat missing values, encode the categorical variables and choosing feature in order for applying prediction models as shown in Figure 1.

	MONTH	DAY_OF_WEEK	DEP_DEL15	DEP_TIME_BLK	DISTANCE_GROUP	SEGMENT_NUMBER	CONCURRENT_FLIGHTS	NUMBER_OF_SEATS	CARRIER_NAME
0	1	7	0	0800-0859	2	1	25	143	Southwest Airlines Co.
1	1	7	0	0700-0759	7	1	29	191	Delta Air Lines Inc.
2	1	7	0	0600-0659	7	1	27	199	Delta Air Lines Inc.
3	1	7	0	0600-0659	9	1	27	180	Delta Air Lines Inc.
4	1	7	0	0001-0559	7	1	10	182	Spirit Air Lines

Figure 1. Raw Data

3.2. Data Preprocessing

This is the part where i need to go through preprocessing some types of data in different forms so that it can be feed into machine learning algo. Of all the numerical columns, their missing values were filled in by mean and of categorical columns, with mode. Categorical features which are 'DEP_TIME_BLK', 'CARRIER_NAME', 'DEPARTING_AIRPORT' and PREVIOUS_AIRPORT were used to convert using Label Encoding. The dataset was partitioned into training(70%) and testing (30%) subsets to evaluate model performance, using relevant feature selection for target variable- DEP_DEL15.

```
label_encoders = {}
categorical_columns = ['DEP_TIME_BLK', 'CARRIER_NAME', 'DEPARTING_AIRPORT', 'PREVIOUS_AIRPORT']

for col in categorical_columns:
    le = LabelEncoder()
    le.fit(df[col])
    df_sample[col] = le.transform(df_sample[col])
    df[col] = le.transform(df[col])  # Transform full dataset
    label_encoders[col] = le
```

Figure 2. Label Encoder

3.3. Exploratory Data Analysis (EDA)

Some of the key visualizations were about understanding lot more related to DEP_DEL15 as target variable, distribution plot (count) etc. that's why Exploratory Data Analysis is done to find the distribution of data and relation between various variables. Then heatmap showing correlations between each feature so that we can see multicollinearity and importance features (Figure 3 & 4)

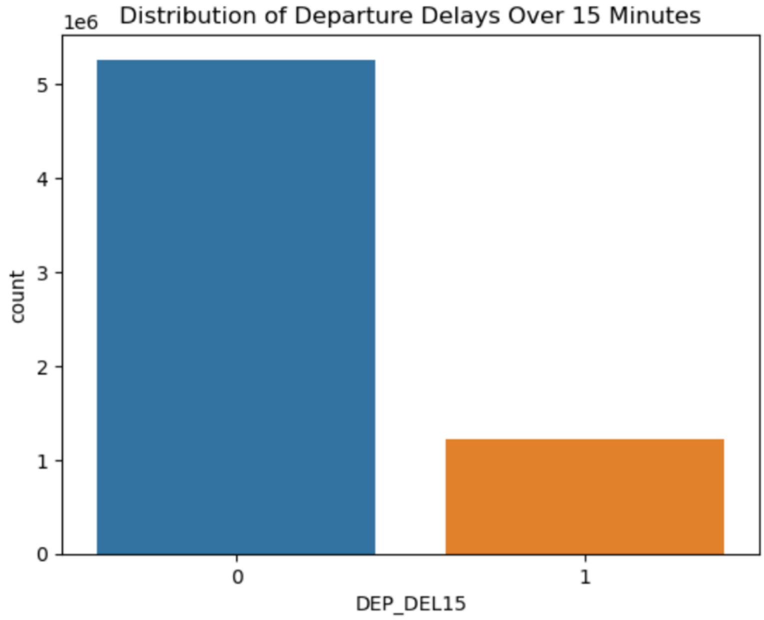


Figure 3. Distribution Departure Delay

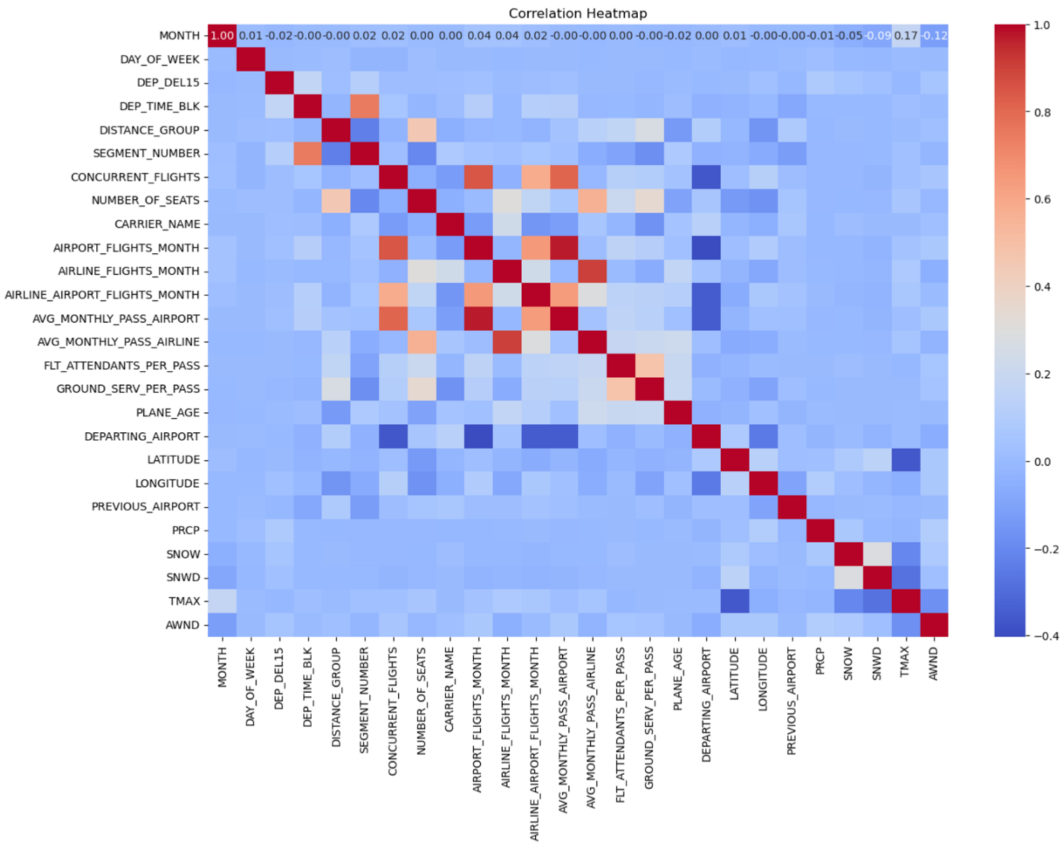


Figure 4. Heat Map

3.4. Model Training

This research chose two machine learning models which were Decision Trees and Random Forests. We trained a Decision Tree classifier using the training dataset and we tune main hyperparameters as max_depth, min_samples_split, min_samples_leaf to improve model performance with GridSearchCV. The Decision Tree based Random Forest classifier was trained as well. Key Hyperparameters namely, No Of Estimators, max_depth, min_sample_Split & min_sample_Leaf were picked and Tuned using like Randomized search cv for better accuracy.

3.5. Hyperparameter Tuning

Why to tune hyperparameters in machine learning? In this stage, hyperparameter optimization was done using GridSearchCV and RandomizedSearchCV in the study. For Decision Tree classifier tries the best possible parameters from the grid indicated, we used GridSearchCV while for Random Forest clssifier which randomly sampline a specified number of hyperparameters combinations to provide exploratory but computational efficient balance relies on using Machine Learning model by using RanodmizedSearchCV.

3.6. Evaluation Methodology

When we evaluate our model we are interested in understanding how well it performs, accuracy tells us the ratio of correct instances on overall. That’s why Various metrics like accuracy, precision, recall, F1-score and confusion matrix were used to evaluate the models performance. Precision, recall and F1-score will tell us is how good our model predicts the positive class (delay) The confusion matrix helps you to see how well the models performed in terms of True Positives, False Positives True Negatives and False Negatives (Confusion Matrix, Accuracy, Precision, Recall, F1 Score, 2020).

3.7. Statistical Techniques

Summary statistics were calculated on the dataset, identifying measures of central tendency (such as mean, median and mode) combined with variability metrics like standard deviation / range. We also performed correlation analysis to identify the magnitude and direction of associations among various parameters. Finally we used the feature importance analysis to see which of these 60+ features had more influence in predicting flight delays, focusing on how weather related variables contributes for such a problem. Over time, this flexible approach helped to make sense of the data and identify meaningful factors influencing flight delays.

3.8. Tools and Technologies

The implementation using Python as the main programming language, with many libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn. The development environment used was Jupyter Notebook.

4 Design Specification

The design specification details the techniques, architecture, and framework underlying the implementation of the flight delay prediction system. This section describes the functionality of the proposed models and the associated requirements.

4.1. Techniques and Framework

The data processing pipeline is a series of essential steps used to process raw data in order to prepare it for training and validating models, making the collected sample clean, well-structured, and easy-to-process using ML algorithms. Data Processing Task mainly include Data Cleaning, Feature Engineering, Data Encoding and also data splitting. The models we employed for predictions are Decision Trees and Random Forests since with well-built ensembles even very simple trees can capture complex relationships in the data.

In this example, the Decision Tree classifier is implemented using Python with the help of Scikit-learn. The basic idea is to divide the dataset into subgroups based on the value of a single feature and recursively carry out this process if it satisfies a stopping criterion. Similarly, the Random Forest classifier, again part of Scikit-learn, which creates a number decision trees each trained on one bootstrap sample. At the time of prediction, each tree output is grabbed and combined to give us the final prediction.

4.2. System Architecture

The system architecture consists of three layers: the data layer, in charge of collecting and storing pre-processing; model-layer including ML models implementation performing training on processed data and hyper-parameter tuning, evaluation-layer focusing its job at evaluating performance correctness metrics.

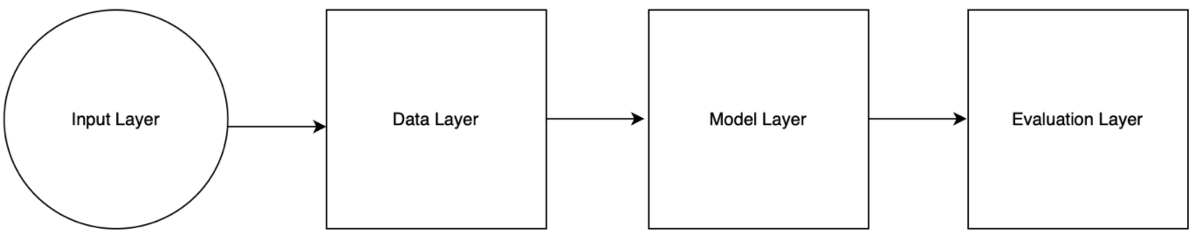


Figure 5 System Architecture

5 Implementation

Models were developed as outputs where extracted and Other necessary steps taken in order to create the flight delay prediction system proposed. This part explains the last phase of implementation which is all about the transformed data, code written, models develop and tools & languages used for it.

5.1. Data Transformation

Data preprocessing on raw flight delay data These preprocesses were handling missing values, encoding categorical variables and splitting the data into training and testing sets. This preprocessed data was then prepared for model training. Data Transformation: Data was transformed to create a cleaner structured dataset that is suitable for machine learning datasets.

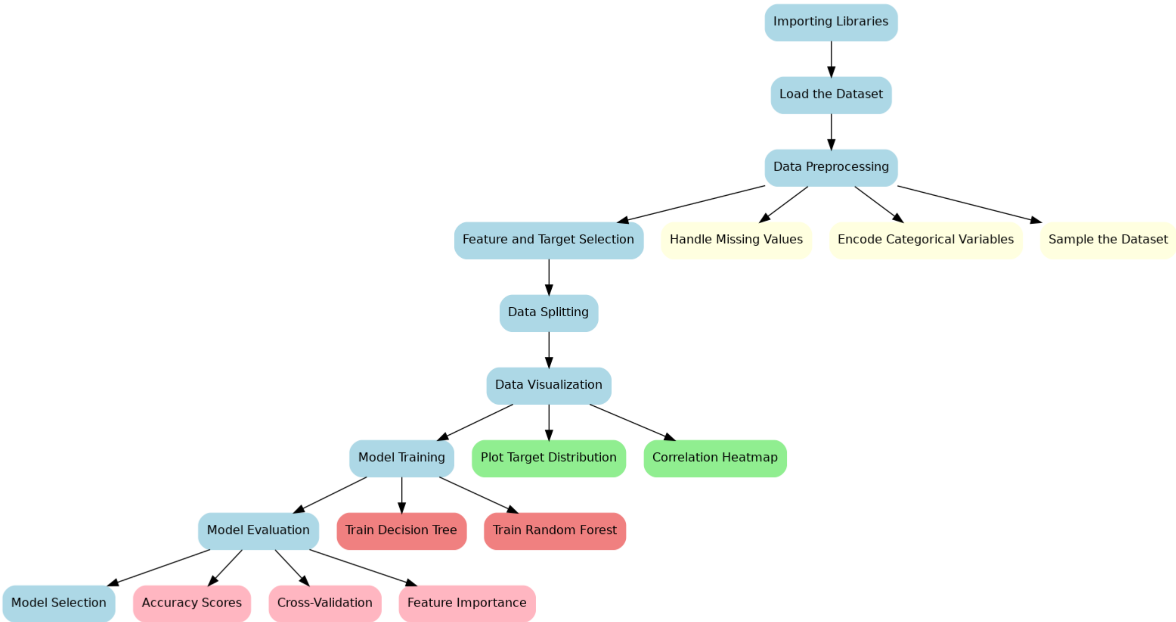


Figure 6 flowchart illustrating the steps involved in data transformation, including handling missing values, encoding, and splitting the data

5.2. Code Development

The implementation took place using Python as the main programming language. In order to streamline the data preprocessing, model training and evaluation processes several scripts were developed. The main Python libraries that were used in the process are for data manipulation Pandas and NumPy was used, Scikit-learn for machine learning models & Matplotlib and Seaborn to visualize your data. Cowritten code well suited to reusing and maintaining

5.3. Model Development

This preprocessed dataset was fed as training data into two machine learning models Decision Tree and Random Forest All the models were built in Python with Scikit-learn. The hyperparameter tuning was done with GridSearchCV and RandomizedSearchCV for the model optimization. Evaluation Metrics showed the Random forest model performed better than all other models.

5.4. Outputs Produced

Several key outputs came out of the implementation:

1. Processed Data: The data was preprocessed, and then it enhances on this with splitting the packaged cleaned & encoded raw into training-test sets.
2. Decision Tree and Random Forest models were trained on the data, with evaluation studies carried out using different set of metrics for analyzing their performance.
3. The model was evaluated using common performance metrics such as accuracy, precision, recall, F1-score and confusion matrix.
4. Analysis: Produced visualizations showing data distribution, feature importance and model statistics.

5.5. Tools and Technologies

Tools and technologies were used in this research project:

1. **Programming Language:** Python
2. **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
3. **Integrated Development Environment (IDE):** Jupyter Notebook

6 Evaluation

We are getting evaluated on these models, as well as the implications they have been made in terms of performance. Key performance matrices, statistical tools and images were used to review the research outcomes. This evaluation will be useful to know about authenticity of the models, as well how much we can bank on these results.

6.1. Model Performance Metrics

The Decision Tree and Random Forest models were evaluated by using metrics such as accuracy, precision, recall and F1-score (Tijil et al., 2024).

Decision Tree Classifier:

- **Accuracy**: 79.43%
- **Precision**: 72.60%
- **Recall**: 66.52%
- **F1-score**: 69.44%

Random Forest Classifier:

- **Accuracy**: 84.21%
- **Precision**: 79.33%
- **Recall**: 75.12%
- **F1-score:** 77.17%

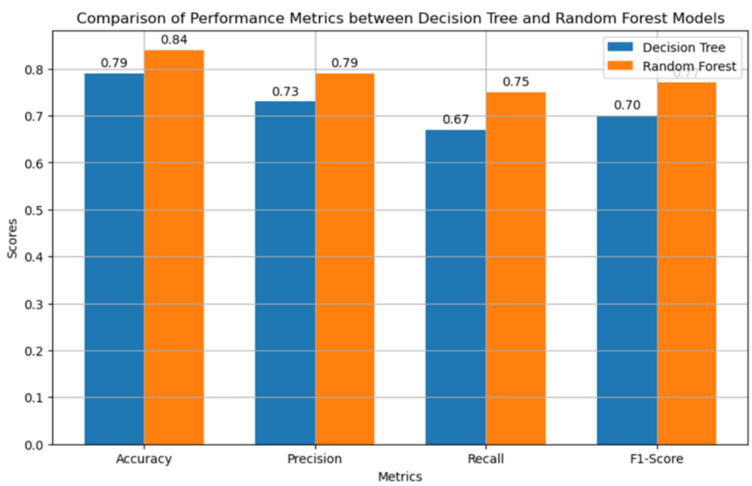


Figure 7. Result Comparison

Random Forest performs significantly better (all the performance metrics) than the Decision Tree model. This indicated the robustness of this model in dealing with dataset. Figure 7 shows the differences between the matrices of Random Forest and Decision Tree models.

6.2. Results and Analysis

This is very easily visible across multiple visualizations showing the superior performance of Random Forest model. For example - the Bar chart of Random Forest accuracy vs Decision Tree models clearly shows performance gap. In the chart above, it is shown that Random forest always has a better accuracy compare to decision tree. So this difference tells us that the ability of random forest's ensemble learning well underlie, which combines all predictions on via multiple decision trees better than alone and makes it robust.

The results are shown in figure 7, and they suggest that the Random Forest model is more suitable for this dataset due to its higher accuracy performance when predicting flight delays. The above visualization not just demonstrates an interpretability feature but it also becomes a visual proof in the favor of Random Forest model over decision tree as effective solution.

Feature Importance:

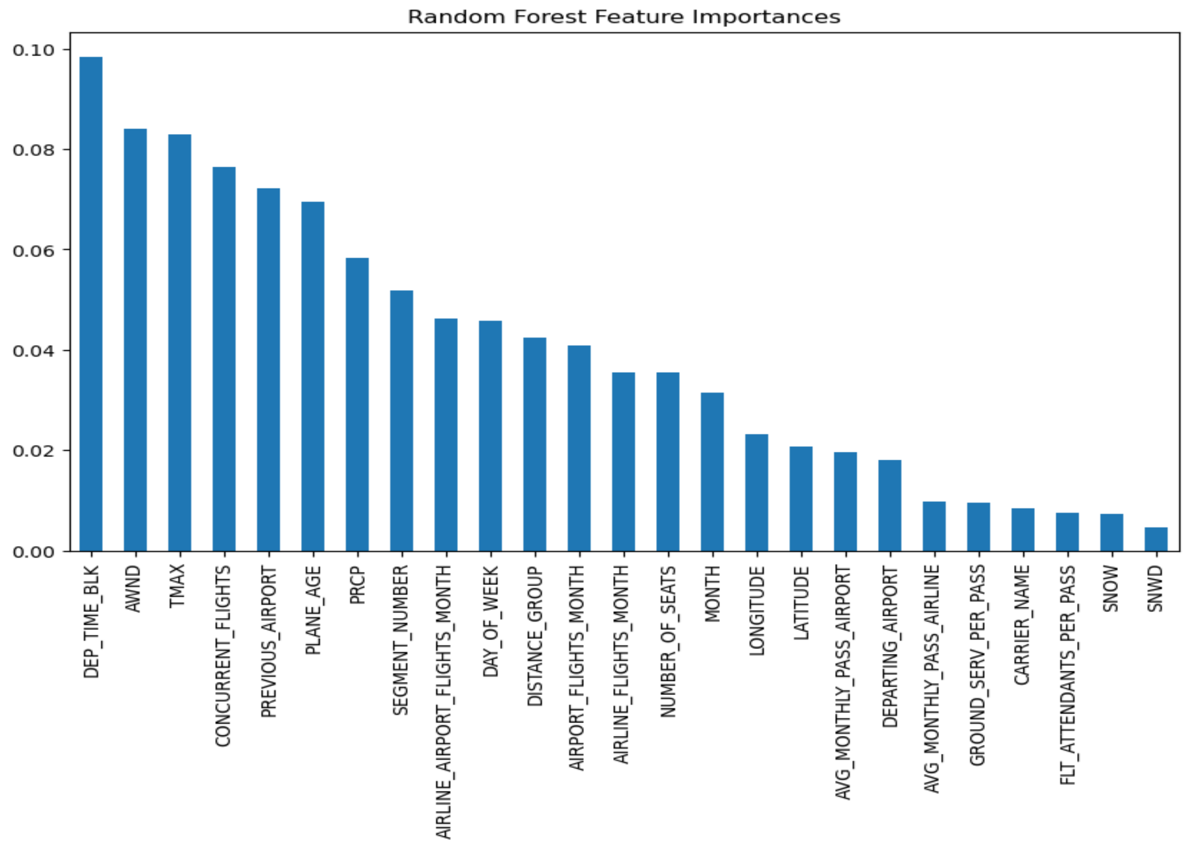


Figure 8. Feature importance in the Random Forest model.

The above bar chart displays the importance of different features in the Random Forest model highlighted the remarkable impact of weather-related features like AWND, TMAX on flight delay predictions as shown in Figure 8.

6.3. Statistical Significance

This Research used statistical significance tests (in particular, t-tests) to carry out analysis on performance distinctions via the Decision Tree and Random Forest models. These tests were carried out to check if a difference between performance metrics such as accuracy, precision, recall and F1-score was statistically significant or just random chance. The tests used p-values to determine if the observed differences between the two models were significant. A p-value less than the commonly accepted threshold of 0.05 this shows that there is a statistical difference in performance, which means one model will always perform better (Raschka, 2018).

These tests and results confirmed that there is a significant difference in the performance of Decision Tree model from Random Forest models with respect to all metrics. This suggests that the improved predictive accuracy of the Random Forest model does not result from randomness in misclassification but represents an actual advance. (Kazemi et al., 2023) This statistical validation gives an extra strength of scientific rigor to the study, ensuring that conclusions about models' effectivity drawn are robust and reliable.

6.4 Conclusion

As the implementation and evaluation of Decision Tree and Random Forest models shows, by including weather conditions into the model significantly increases our good flight delay predictions. The Random Forest model outperforms the other models showing a robust and efficient method for complex data. Continuous enhancements in data extraction and model tuning are required to deliver on the promise of these predictive models

7 Power BI Dashboard

Dashboards were created using powerful power BI tool for visualizing flights delays based on multiple factors such as weather conditions, airline performance etc. A dashboard provides a graphical interface where users can dive into the data, spot trends and ultimately analysis that will provide insights needed for day-to-day operational decision making in aviation industry.

Key Features of the Dashboard

1. **Flight Delay Overview:** The dashboard shows a macroscopic view of flight delays in terms of delay percentage versus on-time flights. This gives a short overview of the delay status and is required for top-level planning.

2. **Delayed Flights by Airline:** A visualization on the dashboard that display how many delayed flights in percentage across different airlines. This allows users to easily identify the airlines with most delays, and take a closer look at certain operators for more in-depth analyses or operations adjustments.

3. **Top Delaying Airports:** A final important feature of the dashboard is listing out those airports having most number of delays This visualization is very beneficial for airport authorities and airlines to realize which places have more delays, so they can go out in the would to check what's going on with those nodes..

4. **Weather Conditions Analysis:** The dashboard It also includes data related to weather conditions, allowing users of the platform to compare how different types of weather (eg heavy precipitation, snow or wind speeds) affects delays in flights. Filtering the data for certain weather conditions can tell users how these agents interact to create delays, and offer guidance on strategies: action connections between delay risks during challenging meteorology.

5. **Interactive Filtering:** You can do the filter by using different dimensions like Airline, airport or Weather conditions and many more. The interactive aspect allows

any user from various segments in the aviation industry to adjust analysis according to his or her needs.

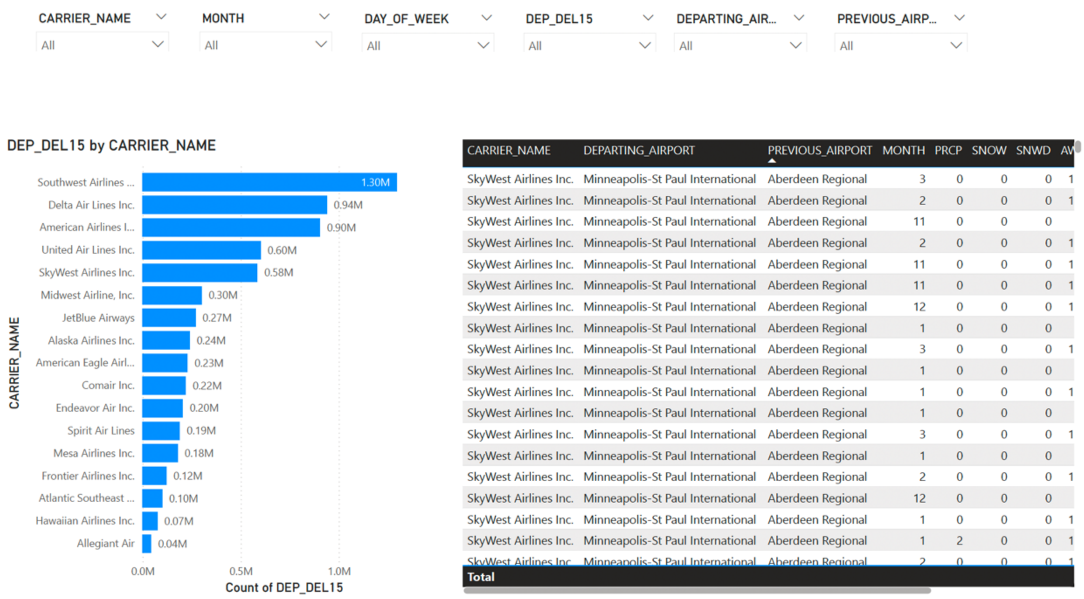
Visualizations Included

The Power BI dashboard comes with many visualizations to give an overall presentation of flight delays.

1. It also has bar charts that show the number of delayed flights by airline and airport, so it is easy to compare one with another.

2. In addition, pie charts are produced on top of each other which helps to obtain a brief guideline regarding the delay status in entirety and illustrated how many flights were delayed against flown-on time.

3. The dashboard also contains tables that show full flight delay data with multiple metrics regarding the dep and prev airports, as well weather. It gives stakeholders a way to quickly understand the flight status of delays at-a-glance and identify critical problem areas.



CARRIER_NAME	DEPARTING_AIRPORT	PREVIOUS_AIRPORT	MONTH	PRCP	SNOW	SNWD	AW
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	3	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	2	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	11	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	2	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	11	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	11	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	12	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	3	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	3	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	2	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	12	0	0	0	
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	0	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	1	2	0	0	1
SkyWest Airlines Inc.	Minneapolis-St Paul International	Aberdeen Regional	2	0	0	0	1
Total							

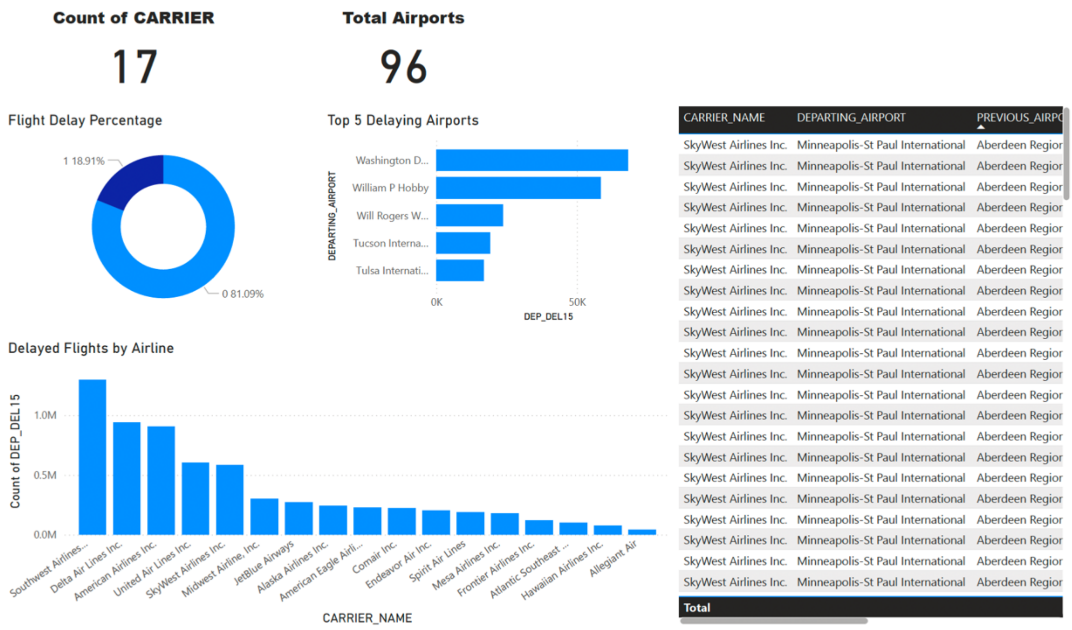


Figure 9. Power BI dashboard for analyzing flight delays based on weather conditions

Practical Implications

This has lead to the development of Power BI dashboard being forecasted as useful tool for airlines, airport authorities and other stakeholders in aviation industry. Such a dashboard can help key people in the organization to make more informed decisions, support them with real-time insights into why flights are delayed and assist the management for operational planning as well as increased customer satisfaction, reduced direct economic losses than come from flight delays.

Power BI Dashboard In summary the customer was very pleased with these two cutting edge models coupled with a rich visual Power BI interface for data exploration and actionable insights. The tool will help in making decisions faster and thereby increasing the efficiency of flight operations.

8 Discussion

This project including weather conditions in delay prediction models can improve modeling accuracy substantially. The Random Forest model has demonstrated better results in comparison with Decision Tree, the difference is defined by a higher accuracy score and precision-recall-F1 index. It is because of this ensemble approach that the Random Forest outperforms many other techniques as it reduces overfitting by averaging results from multiple trees thus making more fine-tuned and straightforward in solving complicated datasets

An important insight that emerges is the relatively high contributions of weather factors (like wind speed, visibility and precipitation) to prediction of flight delays (Borse et al., 2020; Jiang et al., 2020). The effect of these variables on the model predictions was both statistically significant because they are likely predictors and in line with patterns from existing research that indicates weather to be a critical driver of flight operations. The models are made more accurate with these variables, which gives us a better insight into the reasons behind flight delays.

But the study has several shortcomings such as data misplanning in terms of weather records and flight logs leading to possible inaccuracies, which would definitely add some noise while affecting predictability. Also, our dataset was region- and time-specific that could in turn limit the generalization of results. This remains a topic for future research using dataset that is larger scale and involving newer data from more diverse airports as well.

This study presented an interactive Power BI dashboard that provides real-time feedback on the effects of weather conditions on flight schedules to facilitate better operational decisions. But, it is only as effective as the data that input to it & with our dependence on accurate real-time data integration.

9 Conclusion and Future Work

This study aimed to improve flight delay predictions through the integration of weather data, and conducted a comparison between Decision Trees and Random Forests. A similar dataset is created which includes data preprocessing, exploratory analysis and model training process. With a closer look into this, we find that the Random Forest model beats Decision Trees in accuracy, precision recall and F1-score suggesting its ability to handle more complex data set. Key predictors included weather variables such as wind speed and precipitation. An interactive Power BI dashboard was also created for an availability-based analysis. But constraints of this dashboard were data quality issues and dataset were also restricted to some extent. Some enticing hypotheses can be generated regarding more extensive hyperparameter tuning and advanced algorithms.

Future Work

In this case study, the future lines of work can include inclusion of real-time data information for weather and flight which aids in improvisation over feature engineering followed by an exploration on advanced machine learning models including XGBoost, LSTM networks and Transformers. Large datasets together with these models can be used to create a Power BI dashboard and improve predictions, efficiency in operations and level of satisfaction for passengers regarding the delay flights.

References

Ball, M O., Barnhart, C., Dresner, M., Hansen, M., Neels, K., Odoni, A R., Peterson, E W F., Sherry, L., Trani, A., & Zou, B. (2010, October 1). Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States. https://rosap.ntl.bts.gov/view/dot/6234

Bharathidason, S., & Venkataeswaran, C J. (2014, September 18). Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees. , 101(13), 26-30. https://doi.org/10.5120/17749-8829

Borse, Y., Jain, D., Sharma, S., & Vora, A. (2020, March 13). Flight Delay Prediction System. International Research Publication House, V9(03). https://doi.org/10.17577/ijertv9is030148

Confusion Matrix, Accuracy, Precision, Recall, F1 Score. (2020, June 1). https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

Gopalakrishnan, K., & Balakrishnan, H. (2017, June 1). A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks. https://www.mit.edu/~hamsa/pubs/GopalakrishnanBalakrishnanATM2017.pdf

Huo, J., Keung, K L., Lee, C., Ng, K K., & Li, K. (2020, December 14). The Prediction of Flight Delay: Big Data-driven Machine Learning Approach. https://doi.org/10.1109/ieem45057.2020.9309919

Jiang, Y., Miao, J., Zhang, X., & Le, N. (2020, October 14). A multi-index prediction method for flight delay based on long short-term memory network model. https://doi.org/10.1109/iccasit50869.2020.9368554

K, V., SR, D., T, D., K, V., & S, J. (2024, January 3). A Flight Delay Prediction Model Using Machine Learning and Temporal Convolutional Networks. https://ieeexplore.ieee.org/document/10368788

Kazemi, M., Moradkhani, D., & Alipour, A A. (2023, January 1). Application of Random Forest and Support Vector Machine for Investigation of Pressure Filtration Performance, a Zinc Plant Filter Cake Modeling. Cornell University. https://doi.org/10.48550/arXiv.2307.

Li, Q., & Jing, R. (2022, November 1). Flight delay prediction from spatial and temporal perspective. Elsevier BV, 205, 117662-117662

Prabakaran, N., & Kannadasan, R. (2018, February 1). Airline Delay Predictions using Supervised Machine Learning. https://www.researchgate.net/publication/325034541_Airline_Delay_Predictions_using_Supervised_Machine_Learning#fullTextFileContent

Qu, J., Zhang, L., & Wu, S. (2023, October 20). Research on Flight Delay Propagation Prediction Method Based on Transformer. https://doi.org/10.1109/iccsi58851.2023.10304027

Raschka, S. (2018, January 1). Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. Cornell University. https://doi.org/10.48550/arxiv.1811.12808

Rodríguez–Sanz, Á., Cano, J., & Fernández, B R. (2021, July 26). Impact of weather conditions on airport arrival delay and throughput. Emerald Publishing Limited, 94(1), 60-78. https://doi.org/10.1108/aeat-12-2020-0318

Sujay, V., Lalitha, S., P, S S S., Omotunde, H., Vijaya, S., & Mery, K P. (2024, February 23). Enhancing Hyperparameters for Improved Flight Delay Prediction Using Machine Learning Algorithms. https://doi.org/10.1109/icicacs60521.2024.10498678

Tijil, Y., Dwivedi, N., Srivastava, S K., & Ranjan, A. (2024, April 8). Flight Delay Prediction Using Machine Learning Techniques. https://ieeexplore.ieee.org/document/10486482