Applied Information Technology and Computer Science Vol. 5 No. 1 (2024) 475-492


**AITCS**

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/aitcs
e-ISSN :2773-5141

# Flight Data Analysis and Delay Prediction System

**Loy Dong Xuan**[1]

[1]Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, 86400, MALAYSIA

DOI: https://doi.org/10.30880/aitcs.2024.05.01.027
Received 24 June 2023; Accepted 18 May 2024; Available online 30 August 2024

**Abstract**: This project focuses on flight delay prediction using data analysis techniques. Various machine learning algorithms were benchmarked to determine the most effective method. Through experimentation, it was discovered that the random forest algorithm yielded the best results. The developed system was successfully deployed for real-time flight delay prediction, providing valuable insights for airlines and passengers alike.

**Keywords**: flight delay prediction, data analysis, machine learning algorithms, benchmarking, random forest, system deployment

## 1. Introduction

Flight delays can cause significant disruptions to air travel operations, resulting in inconvenience for passengers and financial losses for airlines. To address this issue, accurate prediction of flight delays has become a critical area of research. In this project, a flight delay prediction system was developed using data analysis and machine learning techniques. The objective was to identify the best algorithm for predicting flight delays among several benchmarked methods.

The project involved a comprehensive analysis of historical flight data, including attributes such as arrival time, actual elapsed time, origin, destination, and delay occurrence. Various machine learning algorithms, including random forest, were benchmarked to determine their performance in predicting flight delays. The evaluation process involved training models on a subset of the dataset and testing their accuracy using an evaluation set. The results of the benchmarking process revealed that the random forest algorithm outperformed other methods in predicting flight delays. This research study contributes to the field of flight delay prediction by demonstrating the effectiveness of data analysis and machine learning algorithms. The findings highlight the significance of accurate prediction models in the aviation industry and lay the foundation for further research in this area.

**Corresponding author: ai200236@siswa.uthm.edu.my*

## 2. Literature Review

### 2.1 Machine Learning

Machine learning means "the programming of a digital computer to behave as human while involving the process of learning [7]. Machine learning are mainly divided into four categories which are supervised learning, unsupervised learning, semi supervised learning and enhanced learning. Supervised learning which is used in this project requires manually giving input and required output, in addition to providing feedback about the prediction accuracy in the training process.

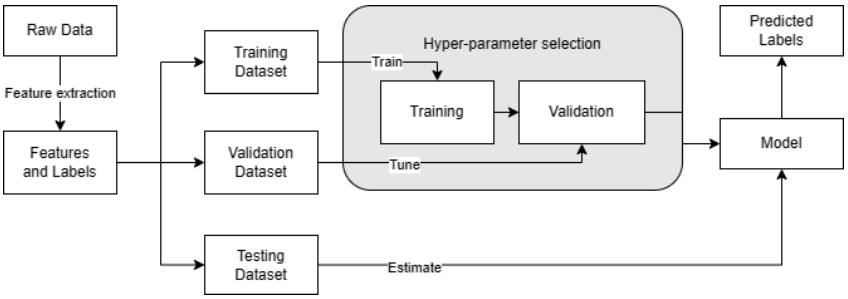



**Figure 2(a): The workflow diagram of machine learning algorithm.**

#### 2.1.1 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification tasks. It gives a prediction model in the form of a set of weak prediction models, usually a decision tree. The resulting algorithm, known as gradient-boosted trees, typically outperforms random forest when a decision tree is the weak learner. It trains each model to correct the errors made by the previous ones, resulting in improved accuracy and robustness.

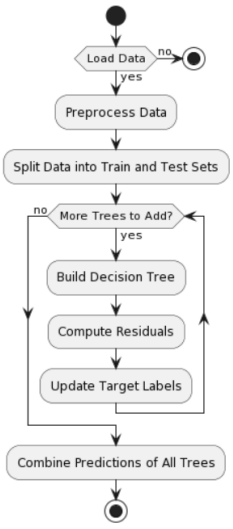



**Figure 2(b): The algorithm of gradient boosting**

### 2.1.2 Multinomial Logistic Regression

Multinomial logistic regression is a classification algorithm used when the target variable has more than two categories. It estimates the probabilities of each category using a logistic regression model and assigns the class with the highest probability as the predicted class.

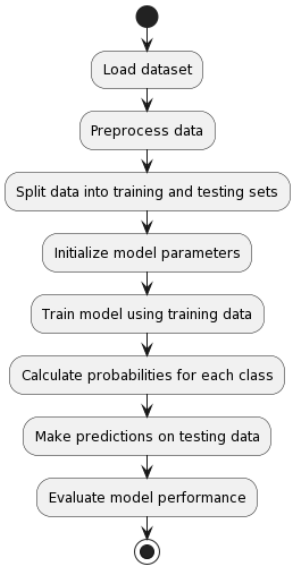



**Figure 2(c): The algorithm of Multinomial Logistic Regression**

### 2.1.3 Naïve Bayes Classification

Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that the features are conditionally independent, given the class, resulting in a simple and efficient model. It calculates the probability of each class based on the feature values and assigns the class with the highest probability as the predicted class.

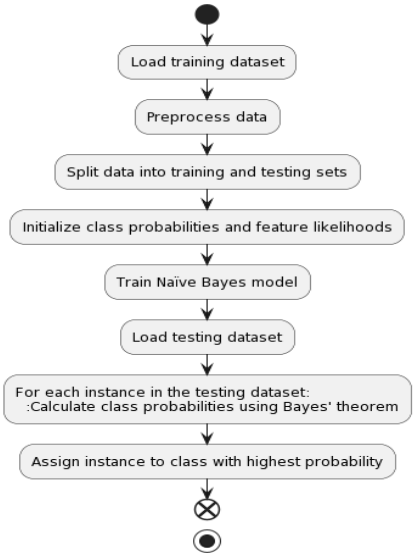



**Figure 2(d): The flowchart of algorithm of Naïve Bayes Classification**

### 2.1.4 Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to create a robust and accurate model. It randomly selects subsets of features and data samples to build individual decision trees. Each tree independently predicts the outcome, and the final prediction is determined by aggregating the predictions of all trees. Random Forest is known for its ability to handle high-dimensional data, handle missing values, and mitigate overfitting.

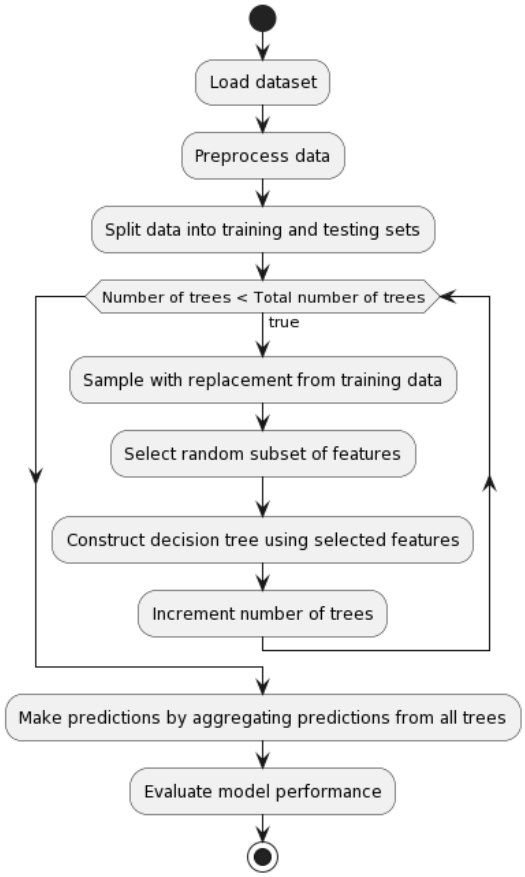



**Figure 2(e): The flowchart of random forest algorithm**

.

## 2.3 Confusion Matrix

A confusion matrix, also known as an error matrix, is a particular table layout that enables visualization of the performance of an algorithm, typically a supervised learning one, in the field of machine learning and more specifically the problem of statistical classification. In this project, the confusion matrix is derived as a measure to evaluate the performance of the trained models in predicting flight delays. The confusion matrix is a 2x2 matrix that shows the counts of true positives, true negatives, false positives, and false negatives. These values are obtained by comparing the predicted flight delay labels with the actual flight delay labels in the validation dataset

## 2.4 Comparison in existing techniques

**Table 2.6.1 shows the comparison between the existing technique and proposed system**

| Author and Year | Method/ Framework | Considered factors |
|---|---|---|
| (Qiang Li et al., 2022) [6] | ST-Random Forest | Spatial features of the aviation network, the temporal correlations of weather conditions and airport/aviation network crowdedness, delay of previous flights, scheduled turnaround time, average turnaround time, distance, scheduled fly time and average fly time |
| (Belcastro et al. 2016) [4] | Random forest | Origin and destination airport, scheduled departure and arrival time, weather condition in origin and destination airports |
| (Thiagarajan et al. 2017) [8] | Levenberg-Marquart (LM) algorithm | Origin and destination airport, quarter of year, month, time-of-day, day-of-week, scheduled departure and arrival time, weather condition at destination airport |
| (Lambelho et al. 2020) [5] | LightGBM, multilayer perceptron and random forest | Airline, route type code, departure airport, arrival airport, scheduled day and hour of departure, scheduled day and hour of arrival |
| (Yu et al., 2019) [9] | Deep belief network | Time-of-day, day-of-week, month-of-year, number of passengers, aircraft capacity, air route situation, airline properties, boarding option, origin or pass-by flight, flight terminal, gap between check-in time, scheduled departure time, closing time of gate, closing time of cargo-hold door, and ready time of shuttles or jet bridge etc. |
| Proposed System | Gradient Boosting, Multinomial Logistic Regression, Naïve Bayes Classification, Random Forest | Departure time, scheduled arrival time, scheduled elapsed time, origin airport and destination airport |

## 3. Methodology

### 3.1 Research Framework

The research process is shown in Appendix A to understand the flow of research activities.

#### 3.1.1 Data collection phase

In this phase, we import and choose samples of flight data from 2004 to 2006 [1][2][3] from Harvard Dataverse. The historical flight delay data is gathered for analysis and model development. The sub-activities include identifying relevant data sources and obtaining access to the data. Then, the necessary data attributes, such as departure time, origin, destination, and elapsed time, are extracted. The data is cleaned by removing irrelevant or inconsistent entries, and it is stored in a suitable format for further analysis. The figure of the raw data of flight historical data is in Appendix C.

#### 3.1.2 Data Pre-processing phase

The data pre-processing phase aims to prepare the collected data for analysis. The sub-activities in this phase involve cleaning the data by handling missing values, outliers, and inconsistencies. Techniques such as data imputation, outlier detection, and data normalization are employed. Tools like RStudio or software like Excel are used for data pre-processing. The figure of pre-processing the flight historical data is in Appendix D.

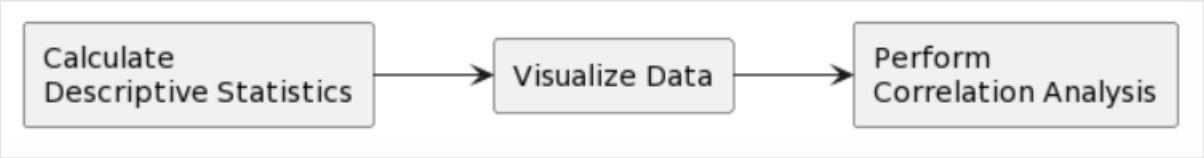



**Figure 3(a): Flowchart of data data pre-processing phase**

#### 3.1.3 Exploratory Data Analysis Phase

In the exploratory data analysis phase, the focus is on understanding the characteristics and patterns in the data. The sub-activities include calculating descriptive statistics, visualizing data using graphs and charts, and performing correlation analysis. Tools like Jupyter Notebook and RStudio are used for data visualization and exploration. The figure of visualizing the data is in Appendix E.

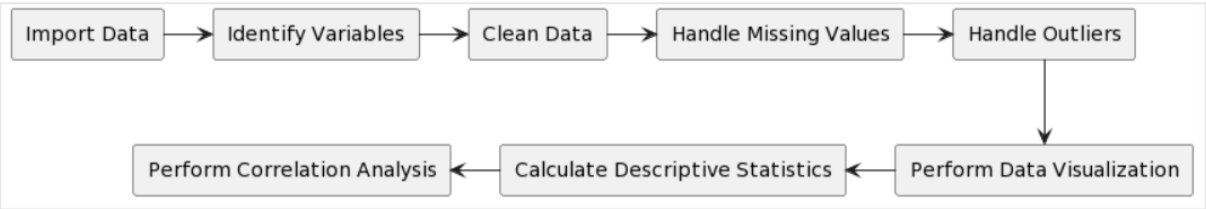



**Figure 3(b): Flowchart of Exploratory Data Analysis Phase**

### 3.1.4 Algorithm Selection and Training Phase

In the algorithm selection and training phase, the aim is to identify and train machine learning algorithms for flight delay prediction. The sub-activities include benchmarking and evaluating various algorithms such as Naïve Bayes, decision trees, random forests, or support vector machines. Tools like Python's scikit-learn library are used.

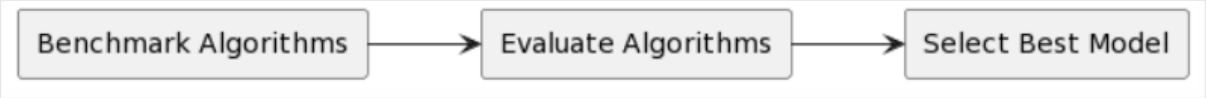



**Figure 3(c): The flowchart of Algorithm Selection and Training Phase**

### 3.1.5 Model Validation Phase

The model validation phase focuses on assessing the performance and accuracy of the trained models. The sub-activities include evaluating the models using validation techniques like cross-validation, confusion matrices, and accuracy metrics. Tools such as Python's scikit-learn library or statistical software packages are used for model validation.


**Figure 3(d): The flow chart of model validation phase**

### 3.1.6 Evaluation Phase

In this phase, we evaluate the outcome of the data. After determining whether the outcome met the expectation, we will proceed to the finding of the results. During the conclusion and report writing process, the final output will be presented. If the outcome does not meet the expectations, we will restart the research process again. The figure of testing accuracy of trained model using Excel is in Appendix F.

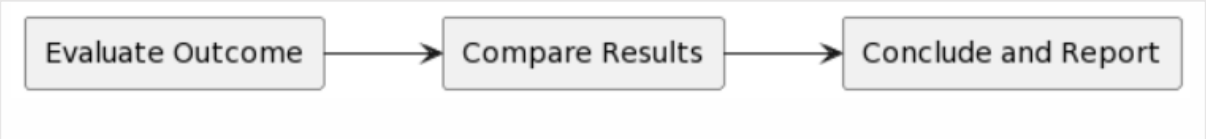



**Figure 3(e): The flowchart of evaluation phase**

### 3.1.7 Implementation Phase

Several tasks were accomplished in developing the flight delay prediction system using Microsoft Studio and integrating Microsoft Azure Machine Learning Studio. Firstly, the development environment was set up in Microsoft Studio, which involved installing the necessary software components and configuring the development environment for coding and model building. Next, the flight delay prediction system was implemented using appropriate programming languages, such as Python or R, within Microsoft Studio. This included writing code to pre-process the data, train the machine learning models, and perform predictions on new data. The figure of implanting the flight delay prediction system using Microsoft Studio is in Appendix G.

### 3.17 Testing Phase

In the testing phase, various activities were conducted to ensure the functionality and reliability of the flight delay prediction system. Firstly, the system was deployed to a web host to make it accessible to users. This involved configuring the necessary hosting environment, setting up the required infrastructure, and ensuring proper connectivity between the system and the web host. Once the system was deployed, rigorous testing was performed to validate its output. Test scenarios were designed to cover different use cases and input scenarios, including various flight parameters and historical data. The system's response to different inputs was observed and compared against expected outcomes. The figure of deploying the data to web hosting server is in Appendix H.

### 3.2 Data Selection

In this research study, dataset from flight data from 2004 to 2006 in Harvard Dataverse is used. These files are part of "Data Expo 2009: Airline on time data". Each of the database file are consists of over millions of datasets of historical flight records. Each record have variety of information such as date of flight, scheduled arrival time, origin airport, destination airport, distance of flight, cancelled flight and etc.

### 3.3 Data Pre-processing

In the data pre-processing phase, various procedures are performed to prepare the dataset for analysis and model training in the flight delay prediction system. These procedures include data cleaning, smoothing, and grouping. Data cleaning involves identifying and removing rows with missing data. Smoothing techniques are applied to eliminate noise or inconsistencies in the data. Data grouping involves categorizing the data into bins or categories. Furthermore, the dataset is modified to classify flight arrival delays as either delayed or non-delayed, with specific criteria for labeling each category. Overall, data pre-processing aims to improve the accuracy and quality of the dataset before applying machine learning algorithms.

### 3.3 Parameter Selection

To implement the proposed system methodology, there are a few parameters of the dataset to be considered in the research experiment. The parameters used are the scheduled arrival time, scheduled elapsed time, origin airport, and destination airport. We also created a new parameter “isDelayed”. The parameters and the value set of “isDelayed” and the benchmark grid setting parameters and values are shown in Table 3(a) and 3(b) below.

**Table 3(a): Benchmark grid setting parameters and values**

| Parameters | Values |
|---|---|
| Value of cancelled and diverted flight | NULL |
| Arrival delay time less than 0 | 1 |
| Arrival delay time greater than 0 | 0 |

**Table 3(b): Benchmark grid setting parameters and values**

| Parameters | Value |
|---|---|
| Seed of gradient boosting | 100 |
| Ranger of number of trees | >50 && <100 |
| Random forest resampling | 3 |
| Number of evaluation | 20 |

3.4 Software and Hardware Specification

The data analysis software used in this research study is RStudio. RStudio refers to the integrated development environment (IDE) of R language. R is the language and operating environment for statistical analysis and drawing. R is a free, free and open-source software belonging to GNU system. It is an excellent tool for statistical calculation and statistical mapping. The requirement for RStudio has been specified for developing the system as shown in Table 3.4.1

**Table 3(c): RStudio system requirement**

| Configuration | Recommended requirement |
|---|---|
| Operating System | Microsoft Windows 7 and above |
| Processor | 2.65 GHz and above |
| Ram | 2G of RAM and above |
| CPU | 2 core and above |
| Core | 2 core and above |
| Storage | 200 GB disk and above |

## 4. Research Design and implementation

4.1 Proposed Solution

The flight data analysis and delay prediction system is proposed to achieve the objectives of this research study. The aim of this paper is to develop a prediction model with practical data. The flight data of the flight database that we have obtained does not provide real-time prediction, such as weather. Therefore, no one has used its analysis to find flights that may be delayed. One delayed flight seems to have nothing to do with other delayed flights, but in fact there is some pattern hidden. Other variables, such as scheduled arrival time, scheduled elapsed time, origin airport, and destination airport, can be used as reference data for predicting cascading flight delay.

### .4.2 Experiment design

This paper presents an experiment design that focuses on training a machine learning model for an effective flight prediction system. The experiment follows standard procedures shown in Figure 4.3.1 and utilizes historical flight data obtained from Harvard Dataverse. With a dataset of 800,000 flight samples, a pipeline is established to connect essential procedures for a comprehensive machine learning training experiment. The experiment involves steps such as connecting a two-class boosted decision tree algorithm, data splitting (80% train set and 20% test set), and evaluating the model's accuracy using an evaluation dataset. The final step involves evaluating the machine learning model's performance based on the historical flight data.

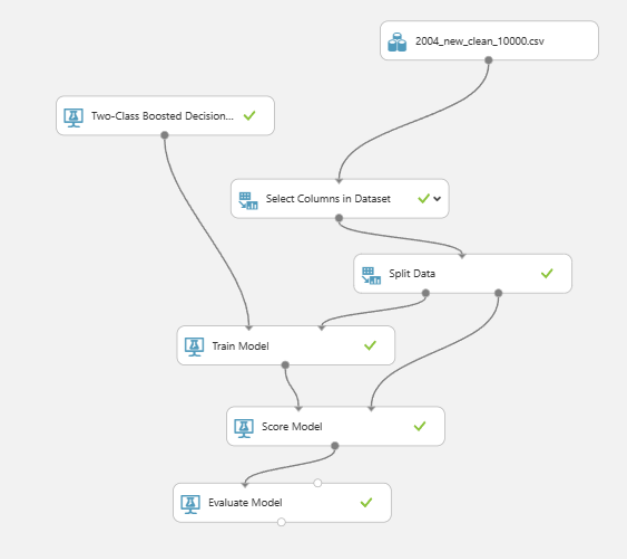



**Figure 4(a): The flight delay prediction model training experiment design.**

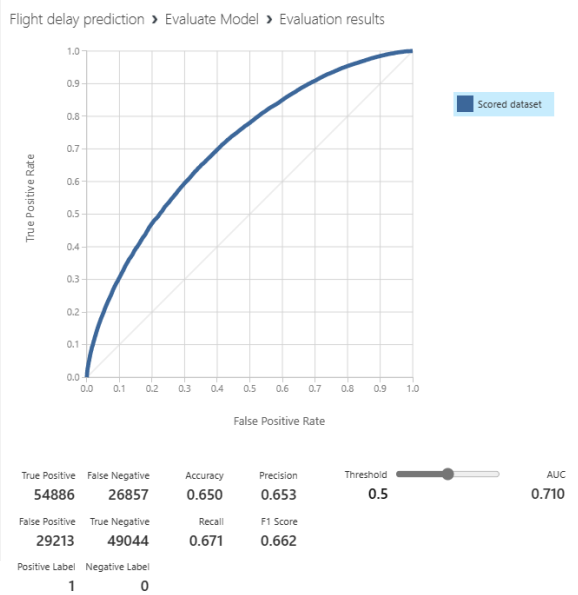



**Figure 4(b): The evaluation results in the training experiment**

### 4.3 Parameter and Testing Methods

In this research study, several important parameters are considered and tested using historical flight data and evaluation data. These parameters include arrival time, actual elapsed time, origin airport, destination airport, and flight delay occurrence. After setting up the training experiment, a predictive experiment is also conducted. In this predictive experiment, the parameter "IsDelayed" is removed from the dataset to be used for testing with the 20% evaluation set. A total of 800,000 flight data samples are extracted from the dataset for the experiment.

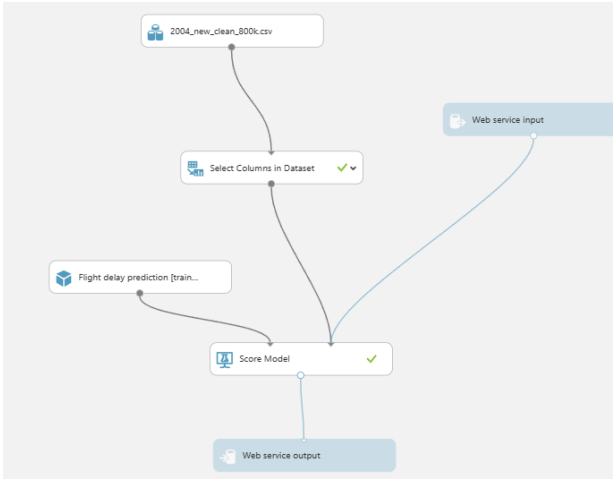



**Figure 4(c): The flight delay prediction in predication experiment design.**

Flight delay prediction [Predictive Exp.] › Score Model › Scored dataset

rows 799999 columns 6

| ArrTime | ActualElapsedTime | Origin | Dest | Scored Labels | Scored Probabilities |
|---|---|---|---|---|---|
| 901 | 98 | ORD | CLT | 1 | 0.634497 |
| 911 | 110 | ORD | CLT | 1 | 0.561695 |
| 920 | 107 | ORD | CLT | 1 | 0.634497 |
| 859 | 92 | ORD | CLT | 1 | 0.634497 |
| 918 | 103 | ORD | CLT | 1 | 0.634497 |
| 905 | 97 | ORD | CLT | 1 | 0.634497 |
| 944 | 114 | ORD | CLT | 1 | 0.561695 |
| 904 | 97 | ORD | CLT | 1 | 0.634497 |
| 903 | 100 | ORD | CLT | 1 | 0.634497 |
| 901 | 95 | ORD | CLT | 1 | 0.634497 |
| 923 | 112 | ORD | CLT | 1 | 0.561695 |
| 901 | 93 | ORD | CLT | 1 | 0.634497 |
| 939 | 134 | ORD | CLT | 1 | 0.531244 |
| 906 | 98 | ORD | CLT | 1 | 0.634497 |
| 859 | 95 | ORD | CLT | 1 | 0.634497 |
| 904 | 97 | ORD | CLT | 1 | 0.634497 |
| 910 | 100 | ORD | CLT | 1 | 0.634497 |
| 1753 | 283 | ORD | PDX | 0 | 0.102701 |

**Figure 4(d): The evaluation results in the training experiment**

After training the model, the next step involves testing its accuracy using the evaluation dataset. We deployed the experiment design to a predictive web service and established a connection with Microsoft Excel. In Excel, we imported 160,000 evaluation data sets and tested the accuracy of the trained model.

The figure below displays the results of the predicted score labels.

| ActualEla | Origin4 | Dest5 | Scored | Actual | Scored Labels | Combined |
|---|---|---|---|---|---|---|
| 128 | MEM | EWR | 0.434927 | 1 | 0 | 10 |
| 62 | LGA | BOS | 0.797897 | 0 | 1 | 01 |
| 80 | CLE | MDW | 0.510925 | 1 | 1 | 11 |
| 235 | EWR | OKC | 0.359446 | 0 | 0 | 00 |
| 80 | IAH | PNS | 0.529821 | 1 | 1 | 11 |
| 88 | ABE | CLE | 0.469045 | 1 | 0 | 10 |
| 91 | CLE | CLT | 0.690636 | 1 | 1 | 11 |
| 88 | HPN | CLE | 0.469045 | 1 | 0 | 10 |
| 177 | CLE | AUS | 0.449596 | 1 | 0 | 10 |
| 81 | IAH | MOB | 0.539781 | 0 | 1 | 01 |
| 186 | IAH | CVG | 0.150184 | 0 | 0 | 00 |
| 71 | IAH | SHV | 0.435259 | 0 | 0 | 00 |
| 99 | IAH | MAF | 0.360321 | 0 | 0 | 00 |
| 67 | CLE | CMH | 0.593956 | 0 | 1 | 01 |
| 101 | CLT | CLE | 0.472036 | 0 | 0 | 00 |
| 69 | CLE | ABE | 0.618866 | 1 | 1 | 11 |
| 81 | BNA | CLE | 0.483499 | 1 | 0 | 10 |
| 71 | SDF | CLE | 0.448109 | 1 | 0 | 10 |
| 62 | EWR | BUF | 0.682404 | 1 | 1 | 11 |
| 93 | IAH | TUL | 0.532693 | 0 | 1 | 01 |
| 76 | AEX | IAH | 0.366906 | 0 | 0 | 00 |
| 109 | ATL | CLE | 0.312276 | 0 | 0 | 00 |
| 67 | IAH | LFT | 0.562615 | 0 | 1 | 01 |
| 99 | IAH | ICT | 0.365047 | 1 | 0 | 10 |
| 58 | EWR | DCA | 0.42765 | 0 | 0 | 00 |
| 167 | PSP | IAH | 0.440688 | 1 | 0 | 10 |
| 101 | CLT | EWR | 0.587611 | 1 | 1 | 11 |
| 31 | BPT | IAH | 0.897687 | 1 | 1 | 11 |
| 131 | IAH | ELP | 0.440271 | 0 | 0 | 00 |
| 117 | GSP | EWR | 0.377421 | 0 | 0 | 00 |

| | Actually Delayed (1) | Actually Not Delayed (0) | Total |
|---|---|---|---|
| Predicted Delayed (1) | 50127 | 29683 | 79810 |
| Predicted Not Delayed (0) | 28302 | 51888 | 80190 |
| Total | 78429 | 81571 | 160000 |

Sheet1

**Figure 4(e): The predicted score label of the experiment**

Subsequently, we merged the actual "IsDelayed" values with the predicted labels and computed the accuracy using the confusion matrix. The table below illustrates the confusion matrix:

**Table 4(a): The confusion matrix to measure the accuracy of the trained model**

| | **Actually Dealyed (1)** | **Actually Not Delayed(0)** | **Total** |
|---|---|---|---|
| **Predicted Delayed (1)** | 50127 | 29683 | 79810 |
| **Predicted Not Delayed (0)** | 28302 | 51888 | 80190 |
| **Total** | 78429 | 81571 | 160000 |

4.4 Result and Discussion

We compared the confusion matrix table results with the evaluation results in the training experiment in Table 4(b). The table provides various performance measures for the trained model in predicting flight delays. Sensitivity represents the proportion of correctly predicted positive cases, while specificity represents the proportion of correctly predicted negative cases. Precision indicates the accuracy of positive predictions, while the negative predictive value represents the accuracy of the negative predictions.

The false positive rate and false negative rate show incorrect prediction proportions. Accuracy reflects the overall correctness of predictions, while the F1 score balances precision and recall. The Matthews Correlation Coefficient quantifies the overall performance of the model, with a value closer to 1 indicating better performance.

**Table 4(b): The overview of the measure in confusion matrix**

| **Measure** | **Value** |
|---|---|
| Sensitivity | 0.6391 |
| Specificity | 0.6361 |
| Precision | 0.6281 |
| Negative Predictive Value | 0.6471 |
| False Positive Rate | 0.3639 |
| False Negative Rate | 0.3609 |
| Accuracy | 0.6376 |
| F1 Score | 0.6336 |
| Matthews Correlation Coefficient | 0.2572 |

.We noticed that the difference between the accuracy, precision, recall, and F1 score of the evaluation results in training experiment and the result from the confusion table are all smaller than 0.04

**Table 4(b): The comparison between both result is shown in Figure 4.5.2.**

| | Training experiment result | Confusion Matrix Table | Difference |
|---|---|---|---|
| **Accuray** | 0.65 | 0.6376 | 0.1124 |
| **Preicsion** | 0.653 | 0.6281 | 0.0249 |
| **Recall** | 0.671 | 0.6391 | 0.0319 |
| **F1 Score** | 0.662 | 0.6336 | 0.0284 |

Hence, it can be observed that the evaluation results using the data sets exhibit only minor variations compared to the evaluations conducted with the two-class boosted decision tree algorithm. The accuracy of the trained model is determined by utilizing a ratio of 6:2:2 for the training dataset, validation dataset, and evaluation dataset, respectively. This ratio ensures a comprehensive assessment of the model's performance across different datasets.

## 5. Conclusion

### 5.1 Discussion

Upon reviewing the objectives, it can be concluded that the research study has successfully achieved its goals. The proposed flight prediction system utilizes multiple supervised machine learning algorithms to find out the best model for the flight delay prediction system. The parameters relevant to the flight prediction system have been thoroughly studied and measured through visualizations using tables and charts. Additionally, the chosen flight data from 2004 to 2006 has been effectively utilized to test the prediction model for a delay prediction system by assigning the data to clusters in the experiment. Overall, the objectives of this research have been mostly accomplished.

### 5.2 Future Work Improve computing efficiency

. In future work, an additional phase will be incorporated into the proposed system research framework, focusing on performance evaluation. This phase will assess the effectiveness of visualization techniques and decision-making processes. Moreover, the prediction accuracy of the system will be thoroughly tested and evaluated. Furthermore, there are plans to deploy the developed system on a web server, enabling users to input parameters and predict the occurrence of flight delays. It is also worth considering the integration of deep learning algorithms, such as neural networks, to further enhance the accuracy and efficiency of the system.

## References

[1] 2008. (n.d.). 2004.csv.bz2 [Data file]. Data Expo 2009: Airline on time data. Harvard Dataverse. https://doi.org/10.7910/DVN/HG7NV7/CCAZGT

[2] 2008. (n.d.). 2005.csv.bz2 [Data file]. Data Expo 2009: Airline on time data. Harvard Dataverse. https://doi.org/10.7910/DVN/HG7NV7/JTFT25

[3] 2008. (n.d.). 2006.csv.bz2 [Data file]. Data Expo 2009: Airline on time data. Harvard Dataverse. https://doi.org/10.7910/DVN/HG7NV7/EPIFFT

[4] Belcastro, L., Marozzo, F., Talia, D., & Trunfio, P. (2016). Using scalable data mining for predicting flight delays. ACM Transactions on Intelligent Systems and Technology, 8(1), 1–20. https://doi.org/10.1145/2888402.

[5] Lambelho, M., Mitici, M., Pickup, S., & Marsden, A. (2020). Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *Journal of Air Transport Management*, *82*, 101737. https://doi.org/10.1016/j.jairtraman.2019.101737.

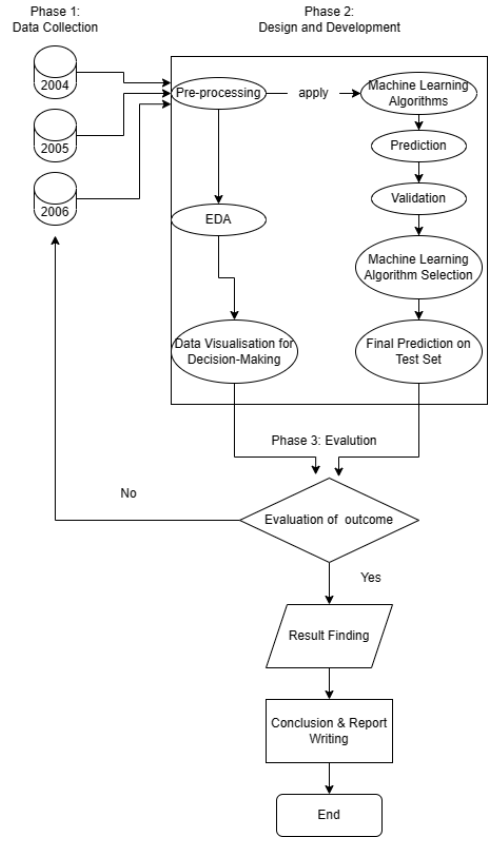
[6] Li, Q., & Jing, R. (2022). Flight delay prediction from spatial and temporal perspective. Expert Systems with Applications, 205, 117662. https://doi.org/10.1016/j.eswa.2022.117662

[7] Samuel, A. L. (1988). Some studies in machine learning using the game of checkers. I. *Computer Games I*, 335–365. https://doi.org/10.1007/978-1-4613-8716-9_14.

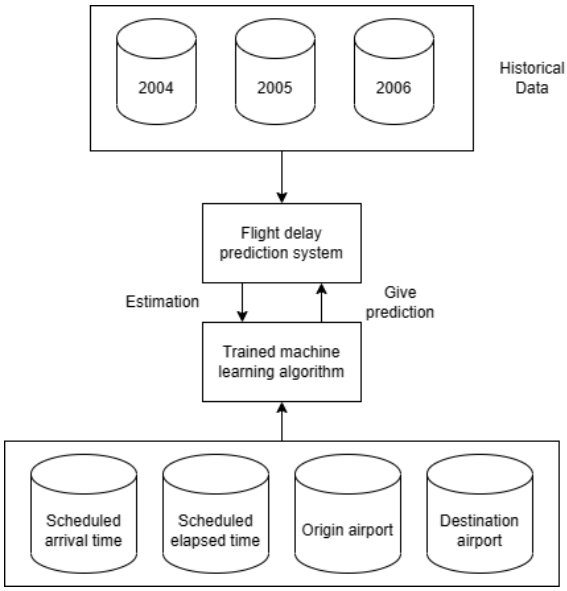
[8] Thiagarajan, B., Srinivasan, L., Sharma, A. V., Sreekanthan, D., & Vijayaraghavan, V. (2017). A machine learning approach for prediction of on-time performance of flights. *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*. https://doi.org/10.1109/dasc.2017.8102138.

[9] Yu, B., Guo, Z., Asian, S., Wang, H., & Chen, G. (2019). Flight delay prediction for commercial air transport: A deep learning approach. *Transportation Research Part E: Logistics and Transportation Review*, *125*, 203–221. https://doi.org/10.1016/j.tre.2019.03.013 .

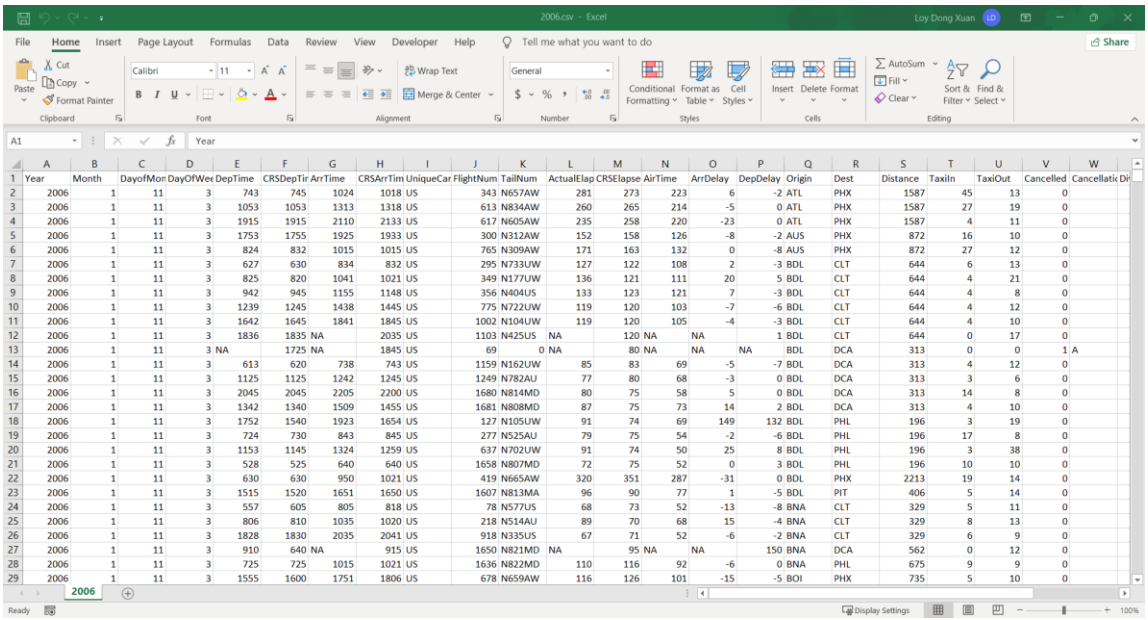
## Appendixes




**Appendix A: Figure of the research framework of the project**

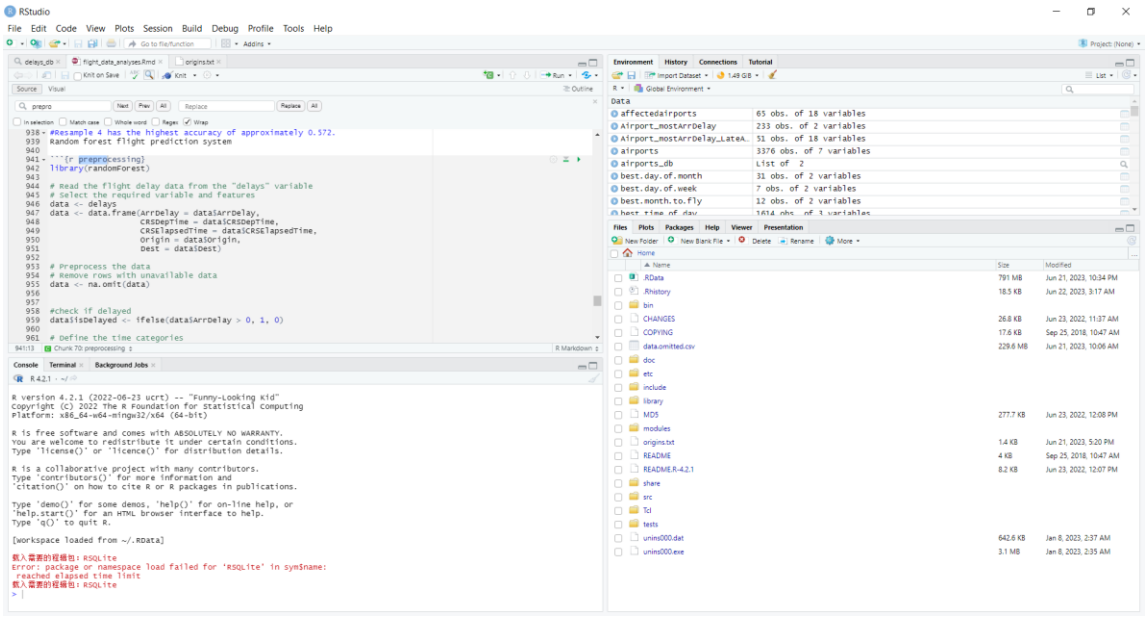



**Appendix B: Figure of the architecture of flight delay prediction system**

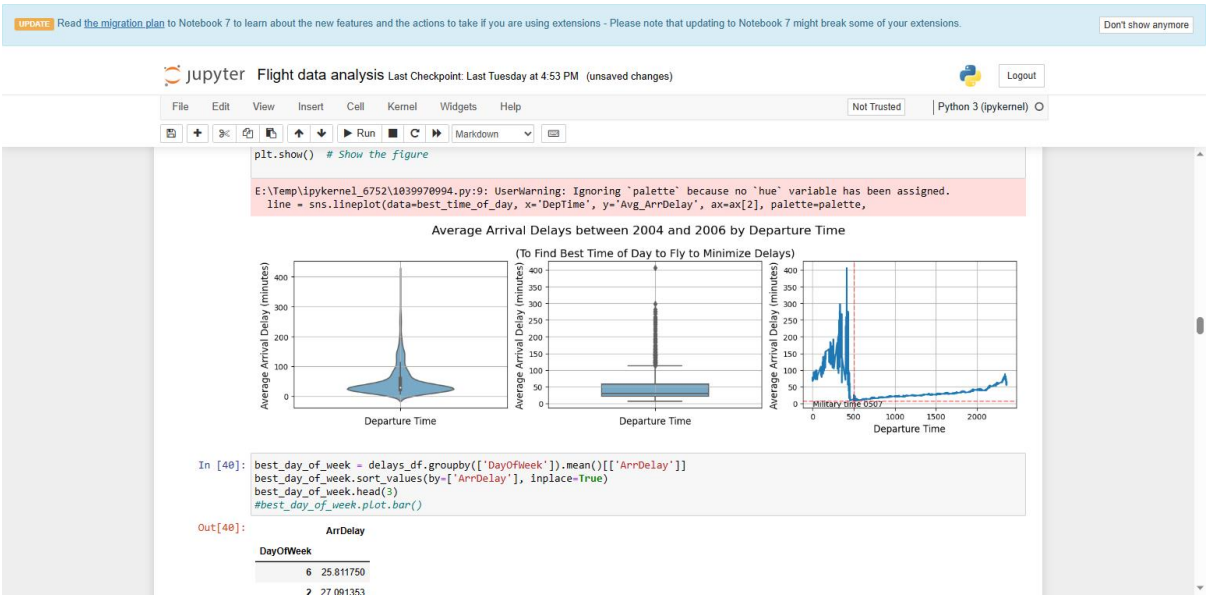


| Year | Month | DayofMon | DayOfWee | DepTime | CRSDepTir | ArrTime | CRSArrTim | UniqueCar | FlightNum | TailNum | ActualElap | CRSElapse | AirTime | ArrDelay | DepDelay | Origin | Dest | Distance | TaxiIn | TaxiOut | Cancelled | Cancellatic |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2006 | 1 | 11 | 3 | 743 | 745 | 1024 | 1018 | US | 343 | N657AW | 281 | 273 | 223 | 6 | -2 | ATL | PHX | 1587 | 45 | 13 | 0 | |
| 2006 | 1 | 11 | 3 | 1053 | 1053 | 1313 | 1318 | US | 613 | N834AW | 260 | 265 | 214 | -5 | 0 | ATL | PHX | 1587 | 27 | 19 | 0 | |
| 2006 | 1 | 11 | 3 | 1915 | 1915 | 2110 | 2133 | US | 617 | N605AW | 235 | 258 | 220 | -23 | 0 | ATL | PHX | 1587 | 4 | 11 | 0 | |
| 2006 | 1 | 11 | 3 | 1753 | 1755 | 1925 | 1933 | US | 300 | N312AW | 152 | 158 | 126 | -8 | -2 | AUS | PHX | 872 | 16 | 10 | 0 | |
| 2006 | 1 | 11 | 3 | 824 | 832 | 1015 | 1015 | US | 765 | N309AW | 171 | 163 | 132 | 0 | -8 | AUS | PHX | 872 | 27 | 12 | 0 | |
| 2006 | 1 | 11 | 3 | 627 | 630 | 834 | 832 | US | 295 | N733UW | 127 | 122 | 108 | 2 | -3 | BDL | CLT | 644 | 6 | 13 | 0 | |
| 2006 | 1 | 11 | 3 | 825 | 820 | 1041 | 1021 | US | 349 | N177UW | 136 | 121 | 111 | 20 | 5 | BDL | CLT | 644 | 4 | 21 | 0 | |
| 2006 | 1 | 11 | 3 | 942 | 945 | 1155 | 1148 | US | 356 | N404US | 133 | 123 | 121 | 7 | -3 | BDL | CLT | 644 | 4 | 8 | 0 | |
| 2006 | 1 | 11 | 3 | 1239 | 1245 | 1438 | 1445 | US | 775 | N722UW | 119 | 120 | 103 | -7 | -6 | BDL | CLT | 644 | 4 | 12 | 0 | |
| 2006 | 1 | 11 | 3 | 1642 | 1645 | 1841 | 1845 | US | 1002 | N104UW | 119 | 120 | 105 | -4 | -3 | BDL | CLT | 644 | 4 | 10 | 0 | |
| 2006 | 1 | 11 | 3 | 1836 | 1835 | NA | 2035 | US | 1103 | N425US | NA | 120 | NA | NA | 1 | BDL | CLT | 644 | 0 | 17 | 0 | |
| 2006 | 1 | 11 | 3 | NA | 1725 | NA | 1845 | US | 69 | 0 | NA | 80 | NA | NA | NA | BDL | DCA | 313 | 0 | 0 | 1 | A |
| 2006 | 1 | 11 | 3 | 613 | 620 | 738 | 743 | US | 1159 | N162UW | 85 | 83 | 69 | -5 | -7 | BDL | DCA | 313 | 4 | 12 | 0 | |
| 2006 | 1 | 11 | 3 | 1125 | 1125 | 1242 | 1245 | US | 1249 | N782AU | 77 | 80 | 68 | -3 | 0 | BDL | DCA | 313 | 3 | 6 | 0 | |
| 2006 | 1 | 11 | 3 | 2045 | 2045 | 2205 | 2200 | US | 1680 | N814MD | 80 | 75 | 58 | 5 | 0 | BDL | DCA | 313 | 14 | 8 | 0 | |
| 2006 | 1 | 11 | 3 | 1342 | 1340 | 1509 | 1455 | US | 1681 | N808MD | 87 | 75 | 73 | 14 | 2 | BDL | DCA | 313 | 4 | 10 | 0 | |
| 2006 | 1 | 11 | 3 | 1752 | 1540 | 1923 | 1654 | US | 127 | N105UW | 91 | 74 | 69 | 149 | 132 | BDL | PHL | 196 | 3 | 19 | 0 | |
| 2006 | 1 | 11 | 3 | 724 | 730 | 843 | 845 | US | 277 | N525AU | 79 | 75 | 54 | -2 | -6 | BDL | PHL | 196 | 17 | 8 | 0 | |
| 2006 | 1 | 11 | 3 | 1153 | 1145 | 1324 | 1259 | US | 637 | N702UW | 91 | 74 | 50 | 25 | 8 | BDL | PHL | 196 | 3 | 38 | 0 | |
| 2006 | 1 | 11 | 3 | 528 | 525 | 640 | 640 | US | 1658 | N807MD | 72 | 75 | 52 | 0 | 3 | BDL | PHL | 196 | 10 | 10 | 0 | |
| 2006 | 1 | 11 | 3 | 630 | 630 | 950 | 1021 | US | 419 | N665AW | 320 | 351 | 287 | -31 | 0 | BDL | PHX | 2213 | 19 | 14 | 0 | |
| 2006 | 1 | 11 | 3 | 1515 | 1520 | 1651 | 1650 | US | 1607 | N813MA | 96 | 90 | 77 | 1 | -5 | BDL | PIT | 406 | 5 | 14 | 0 | |
| 2006 | 1 | 11 | 3 | 557 | 605 | 805 | 818 | US | 78 | N577US | 68 | 73 | 52 | -13 | -8 | BNA | CLT | 329 | 5 | 11 | 0 | |
| 2006 | 1 | 11 | 3 | 806 | 810 | 1035 | 1020 | US | 218 | N514AU | 89 | 70 | 68 | 15 | -4 | BNA | CLT | 329 | 8 | 13 | 0 | |
| 2006 | 1 | 11 | 3 | 1828 | 1830 | 2035 | 2041 | US | 918 | N335US | 67 | 71 | 52 | -6 | -2 | BNA | CLT | 329 | 6 | 9 | 0 | |
| 2006 | 1 | 11 | 3 | 910 | 640 | NA | 915 | US | 1650 | N821MD | NA | 95 | NA | NA | 150 | BNA | DCA | 562 | 0 | 12 | 0 | |
| 2006 | 1 | 11 | 3 | 725 | 725 | 1015 | 1021 | US | 1636 | N822MD | 110 | 116 | 92 | -6 | 0 | BNA | PHL | 675 | 9 | 9 | 0 | |
| 2006 | 1 | 11 | 3 | 1555 | 1600 | 1751 | 1806 | US | 678 | N659AW | 116 | 126 | 101 | -15 | -5 | BOI | PHX | 735 | 5 | 10 | 0 | |

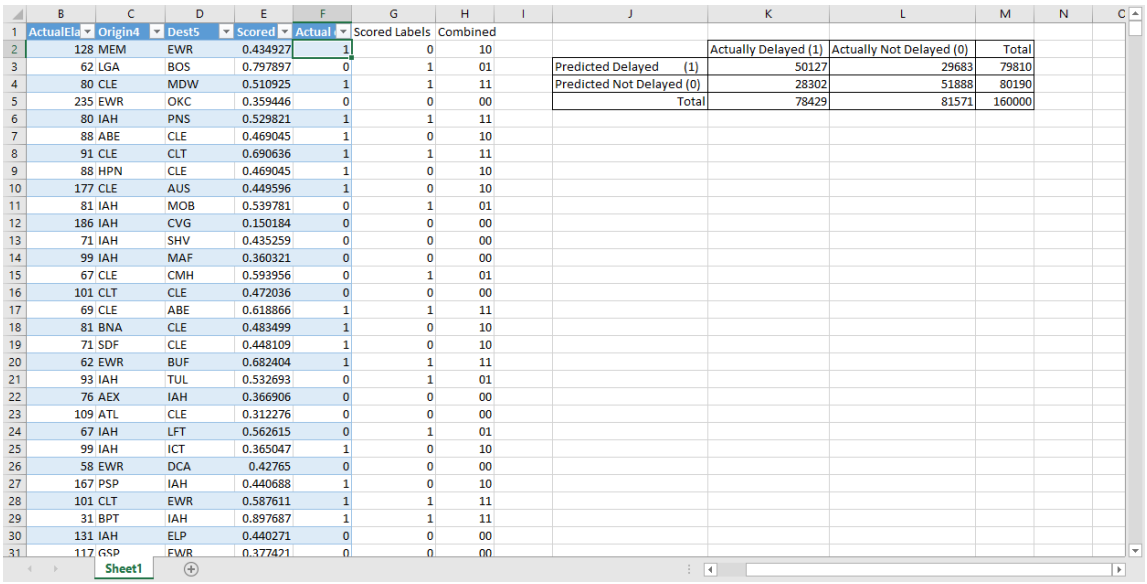
**Appendix C: The flight historical data of 2006 from Dataverse**



**Appendix D: Figure of pre-processing the flight historical data**




**Appendix E: Visualizing the data using Jupyter Notebook**

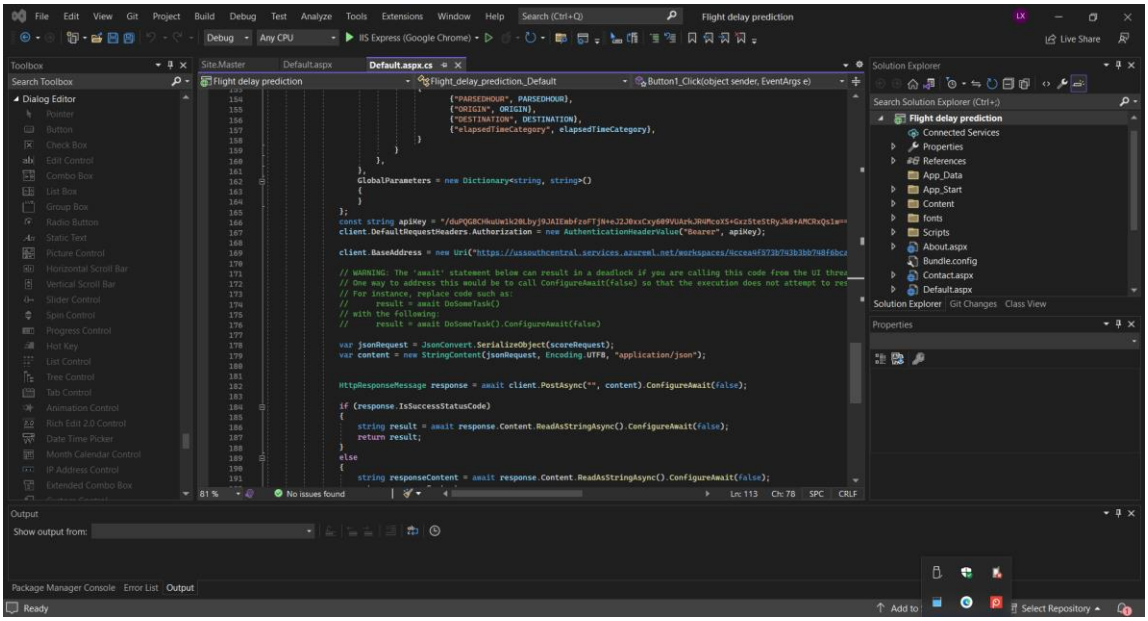


| | ActualEla | Origin4 | Dest5 | Scored | Actual | Scored Labels | Combined |
|---|---|---|---|---|---|---|---|
| 2 | 128 | MEM | EWR | 0.434927 | 1 | 0 | 10 |
| 3 | 62 | LGA | BOS | 0.797897 | 0 | 1 | 01 |
| 4 | 80 | CLE | MDW | 0.510925 | 1 | 1 | 11 |
| 5 | 235 | EWR | OKC | 0.359446 | 0 | 0 | 00 |
| 6 | 80 | IAH | PNS | 0.529821 | 1 | 1 | 11 |
| 7 | 88 | ABE | CLE | 0.469045 | 1 | 0 | 10 |
| 8 | 91 | CLE | CLT | 0.690636 | 1 | 1 | 11 |
| 9 | 88 | HPN | CLE | 0.469045 | 1 | 0 | 10 |
| 10 | 177 | CLE | AUS | 0.449596 | 1 | 0 | 10 |
| 11 | 81 | IAH | MOB | 0.539781 | 0 | 1 | 01 |
| 12 | 186 | IAH | CVG | 0.150184 | 0 | 0 | 00 |
| 13 | 71 | IAH | SHV | 0.435259 | 0 | 0 | 00 |
| 14 | 99 | IAH | MAF | 0.360321 | 0 | 0 | 00 |
| 15 | 67 | CLE | CMH | 0.593956 | 0 | 1 | 01 |
| 16 | 101 | CLT | CLE | 0.472036 | 0 | 0 | 00 |
| 17 | 69 | CLE | ABE | 0.618866 | 1 | 1 | 11 |
| 18 | 81 | BNA | CLE | 0.483499 | 1 | 0 | 10 |
| 19 | 71 | SDF | CLE | 0.448109 | 1 | 0 | 10 |
| 20 | 62 | EWR | BUF | 0.682404 | 1 | 1 | 11 |
| 21 | 93 | IAH | TUL | 0.532693 | 0 | 1 | 01 |
| 22 | 76 | AEX | IAH | 0.366906 | 0 | 0 | 00 |
| 23 | 109 | ATL | CLE | 0.312276 | 0 | 0 | 00 |
| 24 | 67 | IAH | LFT | 0.562615 | 0 | 1 | 01 |
| 25 | 99 | IAH | ICT | 0.365047 | 1 | 0 | 10 |
| 26 | 58 | EWR | DCA | 0.42765 | 0 | 0 | 00 |
| 27 | 167 | PSP | IAH | 0.440688 | 1 | 0 | 10 |
| 28 | 101 | CLT | EWR | 0.587611 | 1 | 1 | 11 |
| 29 | 31 | BPT | IAH | 0.897687 | 1 | 1 | 11 |
| 30 | 131 | IAH | ELP | 0.440271 | 0 | 0 | 00 |
| 31 | 117 | GSP | EWR | 0.377421 | 0 | 0 | 00 |

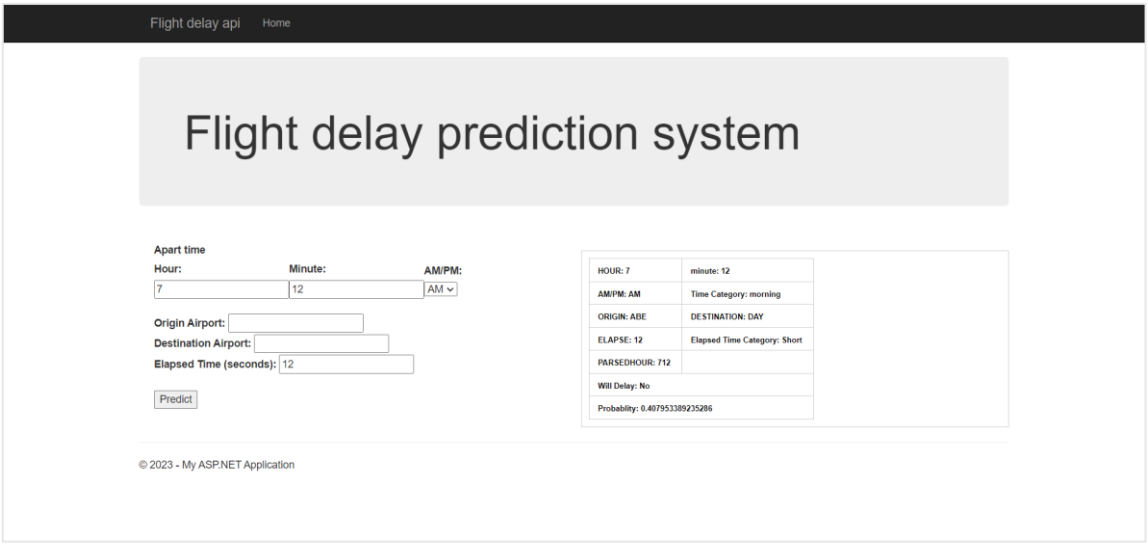
| | Actually Delayed (1) | Actually Not Delayed (0) | Total |
|---|---|---|---|
| Predicted Delayed (1) | 50127 | 29683 | 79810 |
| Predicted Not Delayed (0) | 28302 | 51888 | 80190 |
| Total | 78429 | 81571 | 160000 |

Sheet1

**Appdedix F: Figure of Confusion matrix for testing accuracy of trained model using Excel**




**Appendix G: Implementing delay prediction system in Microsoft Studio**




**Appendix H: Testing deploying delay prediction system to web host.**