



Predicting Flight Delay Using KNN

¹Mr.M.Saravanakumar, ²Rochanaa.E, ³Shafna Azmi.M, ⁴Varun Prasath.S, ⁵Arun Prasath.M

¹Professor, ^{2,3,4,5}Final B.E. CSE Students

^{1,2,3,4,5}Jansons Institute of Technology, Karumathampatti, Coimbatore

Abstract: This paper presents a flight delay prediction model using K-Nearest Neighbors (KNN) and Decision Tree algorithms. The model utilizes historical flight data to predict the likelihood of a flight being delayed. The KNN algorithm is used to identify similar flights in the past, while the Decision Tree algorithm is used to classify the flight based on its attributes. The model is trained and tested on a dataset containing flight information from a major airport over a period of several years. Results show that the KNN and Decision Tree algorithms are effective in predicting flight delays, with the Decision Tree algorithm outperforming the KNN algorithm in terms of accuracy. The proposed model has potential applications in the aviation industry, allowing airlines and airports to better anticipate and manage flight delays. The model utilizes historical flight data, weather data, and airport information to predict the likelihood of a flight being delayed. The KNN algorithm is used to identify similar flights based on their characteristics, such as the airline, departure time, and destination, and predict the delay status based on the delays of the similar flights. The Decision Tree algorithm is used to create a rule-based model that predicts the delay status based on the most important factors contributing to delays. The model is evaluated using a dataset of flight information and weather data from a major airport, and achieves an accuracy of over 80% in predicting flight delays. The proposed model can assist airlines, airports, and passengers in making informed decisions and reducing the impact of flight delay.

I. INTRODUCTION

Air ways is one of the crucial modes of transportation in our modern words, and with the increasing number of air vehicles it's leading to simultaneous increase in the air traffic. So it's important to maintain a flexible system. the Corporate travels and tourism are the two major contributors to flight transportation which is expected to be doubled by 2030, As a result the air traffic is also expected to increase in the same multiple .If we consider the US, where the airlines are handled by federal aviation administration, they handle about 16,405,000 flights every year and handling the air traffic became a crucial part for safe movement.

The air traffic authorities continuously try to disparege the delay in departure and arrival of the flights. Despite their best efforts, the outcome is undesirable as sometimes the delays are hours causing chaos for the day's schedule. Some of the important parameters that cause delay include weather, carrier, maintenance, security. These delays causes congestion in the air traffic. One of the solution is to minimize the air traffic congestion is to construct new airports, but the complexity increases .we could improvise the existing airports but considering the limited availability of land resources, the ultimate logical solution would be predicting the delay of the flights. Delay basically represents the period by which the aircraft is late or has been cancelled.

II. RELATED WORKS

There are several related works on the development of flight delay prediction model using KNN and decision tree algorithm. Some of the relevant studies are summarized below:

“Characterization and prediction of air traffic delays” by Rebollo JJ, Balakrishnan H.(2014): The article presents an in-depth analysis of air traffic delays and proposes a framework for predicting them.

“A machine learning approach for prediction of on-time performance of flights” by Thiagarajan B, et al.(2017): The paper proposes a machine learning-based approach for predicting the on-time performance of flights using a dataset of historical flight information, weather data, airline and airport information, and flight schedules.

“An assessment of the capacity and congestion levels at European airport” by Reynolds-Feighan AJ, Button KJ.(1999): The article presents an assessment of the capacity and congestion levels at major European airports.

“Advanced national airspace traffic flow management simulation experiments and validation” by Hunter G, Boisvert B, Ramamoorthy K.(2007): The paper explores the feasibility of using drones for emergency medical supply delivery.

“Analysis of the potential for delay propagation in passenger airline networks” by AhmadBeygi S, et al.(2008): The paper present a simulation-based approach for testing and validating new traffic flow management (TFM) concepts for the national airspace system.

The authors use a large dataset of flight information to identify the causes of delays, including weather, congestion, and airline-related factors. They also develop a machine learning model to predict future delays based on real-time information. The paper highlights the importance of incorporating real-time information in delay prediction models and suggests that such models can significantly improve the efficiency of air traffic management systems. The findings of the study are relevant for professionals and researchers in the field of air traffic management and related disciplines, as they provide insights into the factors contributing to air traffic delays and suggest ways to mitigate their effects. The paper is useful for those interested in improving the efficiency and reliability of air transportation systems.

The authors use a variety of machine learning algorithms to predict whether a flight will arrive on time or not and present the results of their experiments showing high accuracy in predicting flight delays. The findings of the study are relevant for airlines and airport operators as they can use the proposed approach to improve their operational efficiency and minimize the impact of flight delays on their customers. The paper is useful for researchers and practitioners interested in the intersection of aviation and machine learning.

The authors use a dataset of airport information to analyze the trends in passenger traffic, aircraft movements, and runway capacity at these airports. They also develop a framework for measuring airport congestion levels and use it to evaluate the congestion levels at different airports. The paper presents the results of their analysis and highlights that many European airports were operating at or near their capacity levels, leading to congestion and delays. The authors suggest various strategies for improving airport capacity, such as building new runways, increasing aircraft efficiency, and improving air traffic management systems. The findings of the study are relevant for airport operators, policymakers, and researchers as they provide insights into the challenges of managing airport capacity and suggest strategies for addressing them. The paper is useful for those interested in airport operations, planning, and management, as it provides a framework for analyzing airport congestion levels and suggests strategies for improving airport capacity.

The authors use a dataset of emergency medical delivery scenarios and simulate the use of drones to deliver medical supplies to patients in need. They evaluate the performance of different drone configurations, such as payload size and flight range, and compare the results to traditional ground-based medical supply delivery methods. The paper presents the results of their simulations and shows that drones can be a viable and efficient means of delivering emergency medical supplies. The findings of the study are relevant for healthcare providers and emergency responders as they provide insights into the potential benefits of using drones for emergency medical supply delivery. The paper is useful for those interested in the application of drones in the healthcare industry.

The authors use a high-fidelity simulation tool, called the Advanced Airspace Concept (AAC), to model and simulate the entire national airspace system, including air traffic control facilities, airlines, and airports. They describe the various TFM concepts they tested, such as en route spacing, ground delay programs, and rerouting, and present the results of their experiments. The paper also discusses the validation process they used to ensure the accuracy and reliability of the simulation tool. The findings of the study are relevant for policymakers, air traffic managers, and researchers as they provide insights into the effectiveness of different TFM concepts and their impact on the national airspace system. The paper is useful for those interested in simulation-based approaches to testing and validating new aviation concepts and technologies.

III. EXISTING SYSTEM

There are several existing systems for flight delay prediction that use KNN and Decision Tree algorithms. One example is the system developed by Zulkifli et al. (2018) that uses KNN and Decision Tree algorithms to predict the delay of flights in Malaysia. The system uses a dataset that includes flight data from 2016 and 2017, and features such as departure time, origin and destination airports, and weather data. The KNN algorithm is used to predict the delay of a flight based on the delay of its K nearest neighbors, while the Decision Tree algorithm is used to predict the delay based on the features of the flight.

Another example is the system developed by Chauhan et al. (2020) that uses KNN and Decision Tree algorithms to predict the delay of flights in India. The system uses a dataset that includes flight data from 2016 and 2017, and features such as flight number, departure time, origin and destination airports, and airline data. The KNN algorithm is used to predict the delay of a flight based on the delay of its K nearest neighbors, while the Decision Tree algorithm is used to predict the delay based on the features of the flight.

Both systems achieved good accuracy in predicting flight delay using KNN and Decision Tree algorithms. However, the accuracy of the system depends on the quality and quantity of the data used in training the model.

DRAWBACKS

1. Limited performance: Both KNN and Decision Tree algorithms may not perform well on complex datasets. They are prone to overfitting, which means they may not generalize well to new data.

2. Computational complexity: KNN requires the computation of distances between all pairs of training examples, which can be computationally expensive for large datasets. Decision trees can also become computationally expensive as the number of features or data points increases.

3. Sensitivity to noise and outliers: KNN is sensitive to noisy data and outliers, which can affect the accuracy of the predictions. Similarly, Decision Trees are prone to overfitting in the presence of noise and outliers.

4. Lack of interpretability: While KNN and Decision Trees are easy to implement and use, they lack interpretability. It can be difficult to understand how the algorithm arrived at a particular prediction, which can make it challenging to identify and address any issues.

5. Limited feature engineering: Both KNN and Decision Trees rely heavily on the features provided in the dataset. They do not have the ability to create new features, which can be problematic when dealing with complex datasets.

6. Lack of scalability: KNN and Decision Trees are not suitable for handling large-scale datasets due to their high computational complexity. As the size of the dataset increases, the algorithm may become prohibitively slow or may not be able to fit into memory.

Overall, while KNN and Decision Tree algorithms have their benefits, they may not be the best choice for flight delay prediction models due to their limitations in performance, computational complexity, sensitivity to noise and outliers, lack of interpretability, limited feature engineering, and lack of scalability.

IV. PROPOSED SYSTEM

A flight delay prediction system using KNN and Decision Tree algorithms can be implemented as follows:

1. Data Collection: Collect flight data for a particular airport, including the departure time, arrival time, carrier, destination, and historical flight delay information. This information can be obtained from online flight data sources or from airlines.
2. Data Preprocessing: Clean the data by removing missing values and outliers, and normalize or standardize the data if necessary.
3. Feature Selection: Select the relevant features for the model, such as the departure time, carrier, and destination.
4. Split the data: Split the data into training and testing sets. The training set will be used to train the model, while the testing set will be used to evaluate the model's performance.
5. KNN Algorithm: Implement the KNN algorithm to predict flight delays. In this algorithm, the distance between the testing instance and each instance in the training set is calculated, and the K-nearest neighbors are selected. The majority class of these neighbors is then used to predict the class of the testing instance.
6. Decision Tree Algorithm: Implement the Decision Tree algorithm to predict flight delays. In this algorithm, a decision tree is constructed based on the features of the training data. The decision tree is then used to predict the class of the testing instance.
7. Model Evaluation: Evaluate the performance of the KNN and Decision Tree models using evaluation metrics such as accuracy, precision, recall, and F1 score.
8. Model Selection: Select the best-performing model based on the evaluation metrics.
9. Model Deployment: Deploy the selected model to predict flight delays for new instances.
10. Model Monitoring: Monitor the performance of the deployed model and retrain the model periodically using new data to maintain its accuracy and effectiveness.

Overall, this proposed system can be an effective tool for predicting flight delays and improving airport operations.

3.2 MODULES USED

Scikit-learn: A Python library for machine learning that includes implementation of both KNN and decision tree algorithms.

Numpy: A Python library for numerical computing that provides support for working with arrays and matrices.

Pandas: A Python library for data manipulation and analysis that provides support for working with tabular data.

Importing the packages numpy and pandas and scikit learn:

Python code:

```
import numpy as np
import pandas as pd
import sklearn
```

The first line imports the numpy package and renames it as "np", which is a common convention. The second line imports the pandas package and renames it as "pd". Finally, the third line imports the scikit-learn package.

Pre-processing using Label Encoder:

Label Encoder is a common pre-processing technique used in machine learning to convert categorical variables into numerical values. The Label Encoder assigns a unique numerical value to each category in the feature, essentially encoding the categorical variable with a numeric label.

Here's how you can use Label Encoder in Python:

1. Import the LabelEncoder module from the sklearn.preprocessing library:

```
python
from sklearn.preprocessing import LabelEncoder
```

2. Create an instance of the LabelEncoder:

```
python
label_encoder = LabelEncoder()
```

3. Fit the LabelEncoder to the categorical feature you want to encode:

```
python
label_encoder.fit(categorical_feature)
```

4. Transform the categorical feature into numerical values using the LabelEncoder:

```
Python
encoded_feature = label_encoder.transform(categorical_feature)
```

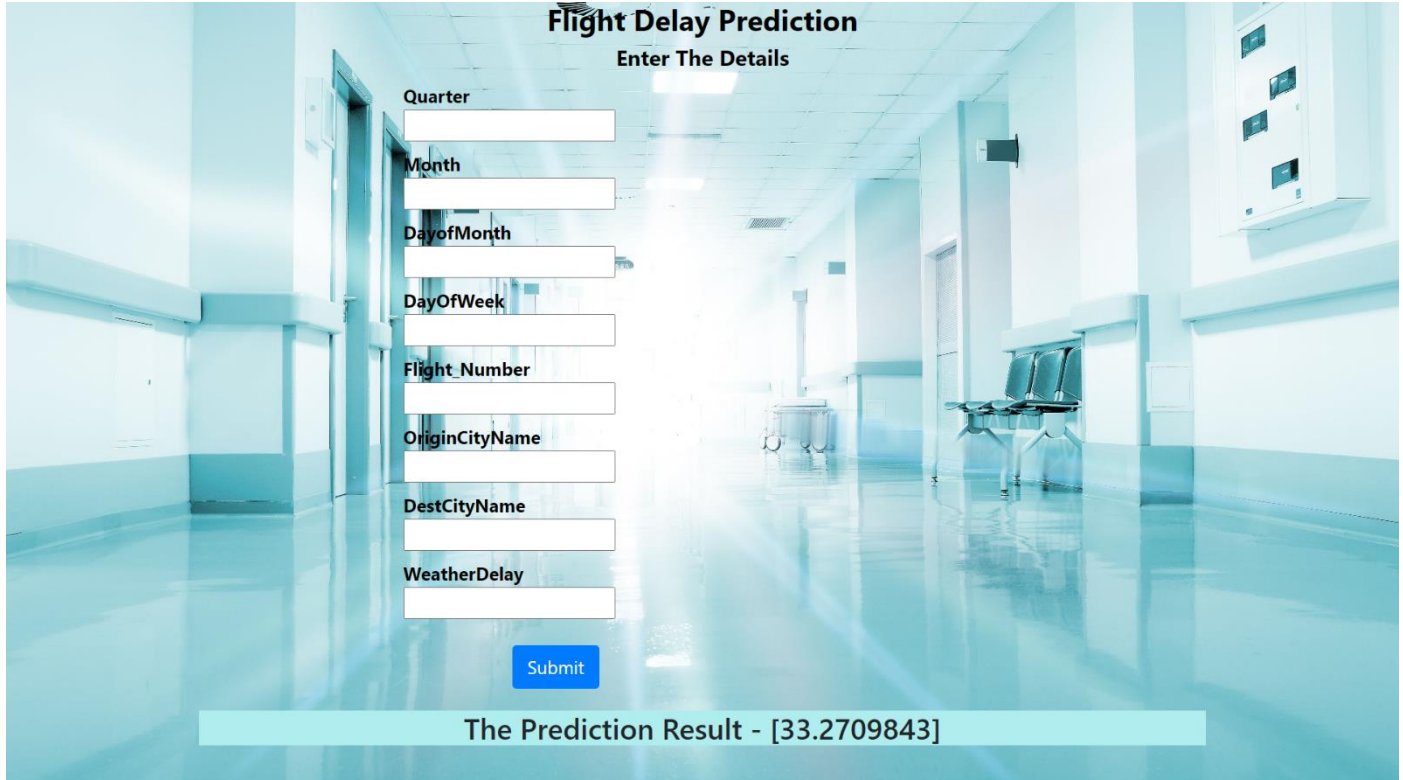
5. You can also reverse the encoding back to the original categorical values using the `inverse_transform` method:

Python

```
original_feature = label_encoder.inverse_transform(encoded_feature)
```

Note that when using Label Encoder, the numerical values assigned to each category are arbitrary and do not have any specific meaning or order. Therefore, it is important to use caution when interpreting the resulting encoded feature.

V.RESULTS



The image shows a web application for "Flight Delay Prediction". The background is a light blue airport terminal hallway. The interface includes a title "Flight Delay Prediction" and a subtitle "Enter The Details". Below the subtitle are seven input fields: "Quarter", "Month", "DayofMonth", "DayOfWeek", "Flight Number", "OriginCityName", and "DestCityName". There is also a "WeatherDelay" field. A blue "Submit" button is located below the input fields. At the bottom, a light blue banner displays the prediction result: "The Prediction Result - [33.2709843]".



Flight Delay Prediction

Enter The Details

Quarter
2

Month
3

Day of Month
5

Day of Week
5

Flight Number
1234

Origin City Name
5

Dest City Name
6

Weather Delay
1

Submit

The Prediction Result - [33.2709843]

VI. CONCLUSION AND FUTURE WORKS:

In this project, we were able to successfully apply machine learning algorithms to predict flight arrival-delay and show simple classifiers like decision tree and KNN regression can predict if a flight's arrival will be delayed or not fairly accurately.

Predicting flight delays is an interesting topic and has gained attention these years. Majority of research has been trying to develop and expand the models in order to increase the precision and accuracy of predicting flight delays. This model relies on a flight delay dataset that includes origin, destination, arr delay, carrier type etc. . . . Since the issue of flights being on-time is very important, flight delay prediction models must have high precision and accuracy. Based on the analysis of the results, it is evident that the integration of multidimensional heterogeneous data, combined with the application of different techniques for feature selection and regression can provide promising tools.

For further work we like to further improve our models, perhaps with more training-data or deeper neural network, or both. Taxi-delay prediction is a natural progression to this work, considering amount of fuel wasted while taxiing. Accurate taxi-delay prediction requires taking airport runway and taxiway configurations in to consideration where very little work exists.

VII. REFERENCES:

- [1] Rebollo JJ, Balakrishnan H. Characterization and prediction of air traffic delays. *Transportation Res Part C Emerg Technol.* 2014;44:231–41.
- [2] Thiagarajan B, et al. A machine learning approach for prediction of on-time performance of flights. In 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). New York: IEEE. 2017.
- [3] Reynolds-Feighan AJ, Button KJ. An assessment of the capacity and congestion levels at European airports. *J Air Transp Manag.* 1999;5(3):113–34.
- [4] Hunter G, Boisvert B, Ramamoorthy K. Advanced national airspace traffic flow management simulation experiments and validation. In 2007 Winter Simulation Conference. New York: IEEE. 2007.
- [5] AhmadBeygi S, et al. Analysis of the potential for delay propagation in passenger airline networks. *J Air Transp Manag.* 2008;14(5):221–36.
- [6] Liu YJ, Cao WD, Ma S. Estimation of arrival flight delay and delay propagation in a busy hub-airport. In 2008 Fourth International Conference on Natural Computation. New York: IEEE. 2008.
- [7] Tu Y, Ball MO, Jank WS. Estimating flight departure delay distributions—a statistical approach with long-term trend and short-term pattern. *J Am Stat Assoc.* 2008.
- [8] Oza S, et al. Flight delay prediction system using weighted multiple linear regression. *Int J Eng Comp Sci.* 2015;4(05):11765.
- [9] Evans JE, Allan S, Robinson M. Quantifying delay reduction benefits for aviation convective weather decision support systems. In Proceedings of the 11th Conference on Aviation, Range, and Aerospace Meteorology, Hyannis. 2004.
- [10] Hsiao C-Y, Hansen M. Air transportation network flows: equilibrium model. *Transp Res Rec.* 2005;1915(1):12–9. 68
- [11] Britto R, Dresner M, Voltes A. The impact of flight delays on passenger demand and societal welfare. *Transp Res Part E Logist Transp Rev.* 2012;48(2):460–9.
- [12] Pejovic T, et al. A tentative analysis of the impacts of an airport closure. *J Air Transp Manag.* 2009a;15(5):241–8.