|  |
| --- |
| **SYNOPSIS** |
| 1. INTRODUCTION |
| 1. PROBLEM DEFINITION& DESIGN THINKING |
| 1. RESULT |
| 1. ADVANTAGES & DISADVANTAGES |
| 1. APPLICATIONS |
| 1. CONCLUSION |
| 1. FUTURE SCOPE |
| 1. APPENDIX |

**FLIGHT DELAY PREDICTION AVIATION INDUSTRY USING MACHINE LEARNING**

**FLIGHT DELAY PREDICTION AVIATION INDUSTRY USING MACHINE LEARNING**

1. **INTRODUCTION**

**1.1 Overview**

OVER the last twenty years, air travel has been increasingly preferred among travelers, mainly because of its speed and in some cases comfort. This has led to phenomenal growth in air traffic and on the ground. An increase in air traffic growth has also resulted in massive levels of aircraft delays on the ground and in the air. These delays are responsible for large economic and environmental losses. According to, taxi-out operations are responsible for 4,000 tons of hydrocarbons, 8,000 tons of nitrogen oxides and 45,000 tons of carbon monoxide emissions in the United States in 2007. Moreover, the economic impact of flight delays for domestic flights in the US is estimated to be more than $19 Billion per year to the airlines and over $41 Billion per year to the national economy In response to growing concerns of fuel emissions and their negative impact on health, there is active research in the aviation industry for finding techniques to predict flight delays accurately in order to optimize flight operations and minimize delays.

**1.2 Purpose**

Nowadays, the demand for airline transportation is increasing significantly. Analysis of flight delay, therefore, has become a popular research area. Various researchers used different techniques of machine learning and data mining to conduct the investigation. They were interested in different aspects such as airport facility location, weather condition, and airport capacity.

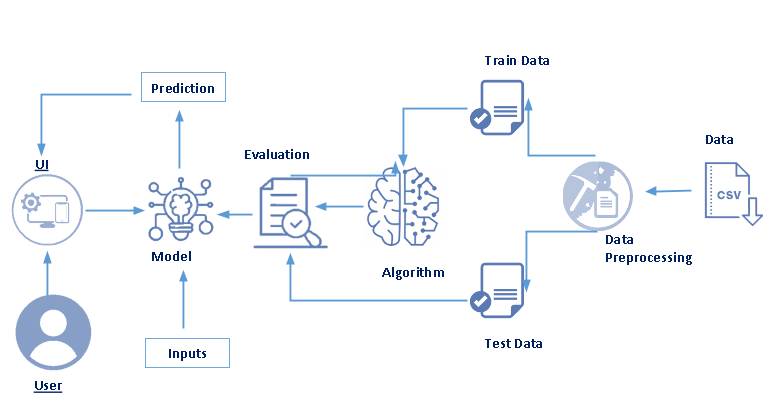
Using machine learning allows researchers to handle large quantities of flight data for storing and processing.

The study conducted by Khaksar and Sheikholeslami used Bayesian modeling, decision tree, cluster classification, random forest to estimate the occurrences and magnitude of delay in the US and Iranian airline network. They determined that the main parameters that affected the airline network of the US are visibility, wind, and departure time, and that of Iran are fleet age and aircraft type.

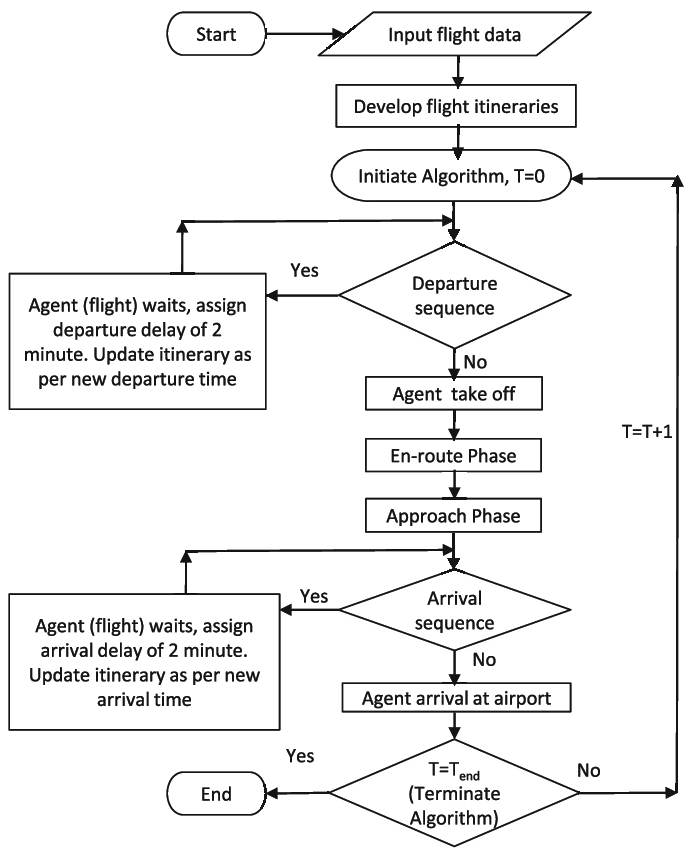
Esmaeilzadeh and Mokhtarimousavi used the support vector machine (SVM) to investigate the causes and patterns of air traffic delay at three major New York City airports. Several explanatory variables were tested to discover their association with flight delay, airport operation, and flow management. The probabilities of them causing the delay were calculated and compared to understand the causes of departure delays better.

1. **PROBLEM DEFINITION & DESIGN THINKING**

**2.1 Empathy Map**

****

* 1. **Ideation & Brainstorming Map**



1. **RESULT**

The result shows that the highest values of accuracy, precision, recall, and f1-score are generated by the Decision Tree model (accuracy: 0.9778; precision: 0.9777; recall: 0.9778; f1-score: 0.9778). Such high values indicate that the Decision Tree performs well when predicting flight delays in the data set.

1. **ADVANTAGES & DISADVANTAGES**

**Advantages:**

By providing passengers with real-time flight information, FIDS **reduces stress for passengers associated with air travel, improves communication between airlines and passengers, and helps airlines and airports better manage their operations**.

* High Speed.
* Fast Service.
* Send almost everywhere your freight.
* High Standard of Security.
* Natural Route.
* There is less need for heavy packaging.

**Disadvantages:**

Flight delays not only irritate air passengers and disrupt their schedules but also **cause a decrease in efficiency, an increase in capital costs, reallocation of flight crews and aircraft, and additional crew expenses**

1. **APPLICATIONS**

Machine learning algorithms and IoT can help carriers curb costs of unplanned maintenance by monitoring aircraft and detecting anomalies. The machine learning algorithm will track the real-time technical condition of the aircraft and notify technicians of possible malfunctions.

Before we begin the process of training the models, it is essential to perform some preprocessing steps. The techniques and methodologies used for preprocessing are mentioned below :

1.Handling missing values –The dataset contains small percentage of missing values for certain columns like Departure delay, taxi out and so on. These rows containing missing values are dropped as they make up a very small portion of the dataset.

2.Formatting times –Initially the times in the dataset are in the form of 4 digit numbers which are not of much use, so these are transformed into HH:MM format. New columns which have the formatted time are created for Departure time, Scheduled arrival, Scheduled departure and arrival time.

3.Feature selection - Some of the features listed above are not really needed for the prediction of delays, so the following were dropped only the following features were kept for the prediction purposes Airline operator, Origin Airport, Destination Airport, Distance, Actual Departure, Date, Day, Scheduled Departure, Departure Delay, Actual Arrival, Scheduled Arrival, Arrival Delay, Scheduled Time, Elapsed Time, Air Time, Taxi In, Taxi Out, Diverted.

4.Label Encoding –Some of the features are in the form of a string these were converted to number values using Label encoder and were assigned number beginning with zero, this is done so that the dataset is more machine learning friendly as models tend to not perform well with strings as features.

1. **CONCLUSION**

If the project, we can see that it is possible to predict flight delay patterns from just the volume of concurrently published tweets, and their**.** In this project, we use flight data, weather, and demand data to predict flight departure delay. Our result shows that the Random Forest method yields the best performance compared to the SVM model. Somehow the SVM model is very time consuming and does not necessarily produce better results. In the end, our model correctly predicts 91% of the non-delayed flights. However, the delayed flights are only correctly predicted 41% of time. As a result, there can be additional features related to the causes of flight delay that are not yet discovered using our existing data sources. In the second part o sentiment and objectivity. This is not unreasonable; people tend to post about airport delays on Twitter; it stands to reason that these posts would become more frequent, and more profoundly emotional, as the delays get worse. Without more data, we cannot make a robust model and find out the role of related factors and chance on these results. However, as a proof of concept, there is potential for these results. It may be possible tweets to routinely use to ascertain an understanding of concurrent airline delays and traffic patterns, which could be useful in a variety of circumstances

1. **FUTURE SCOPE**

This project is based on data analysis from year 2008. A large dataset is available from 1987-2008 but handling a bigger dataset requires a great amount of preprocessing and cleaning of the data. Therefore, the future work of this project includes incorporating a larger dataset. There are many different ways to preprocess a larger dataset like running a Spark cluster over a server or using a cloud-based services like AWS and Azure to process the data. With the new advancement in the field of deep learning, we can use Neural Networks algorithm on the flight and weather data. Neural used in cluster analysis. Neural Network offers distributed computer architecture with important learning abilities to represent nonlinear relationships. Also, the scope of this project is very much confined to flight and weather data of United States, but we can include more countries like China, India, and Russia. Expanding the scope of this project, we can also add the flight data from international flights and not just restrict our self to the domestic flights.

Network works on the pattern matching methodology. It is divided into three basic parts for data modelling that includes feed forward networks, feedback networks, and selforganization network. Feed-forward and feedback networks are generally used in the areas of prediction, pattern recognition, associative memory, and optimization calculation, whereas self-organization networks are generally

1. **APPENDIX**

**SOURCE CODE**

****