Flight Delay Prediction for aviation Industry using Machine Learning

1. INTRODUTION:

Air ways is one of the crucial modes of transportation in our modern words, and with the increasing number of air vehicles its leading to simultaneous increase in the air traffic. So its 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 airtraffic authorities continuously try to disparage 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 days 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. The delay results in complexity in air traffic and dissatisfaction of customers and increase in costs for the company .If a flight is delayed by 10 minutes the flight is considered delay. In the USA from march 2021 to march 2022 itself the delay in flights is about 20.29 % of which air carrier delay is 7.04%, delay due weather is 0.72% , delay due to navigation system is 4.26 and delay due to aircraft arriving late and security delay is 6.12% .we cannot exactly predict the reason for the delay but after the arrival we can predict the delay time for reaching the destination

1.1 Overview :

The goal in this challenge is to predict the flight delays.

You should predict ARIVAL_DELAY column. Here there are explanations for some data rows:

- **YEAR, MONTH, DAY, DAY_OF_WEEK**: dates of the flight
- **AIRLINE**: An identification number assigned by US DOT to identify a unique airline
- **ORIGIN_AIRPORT, DESTINATION_AIRPORT**: code attributed by IATA to identify the airports

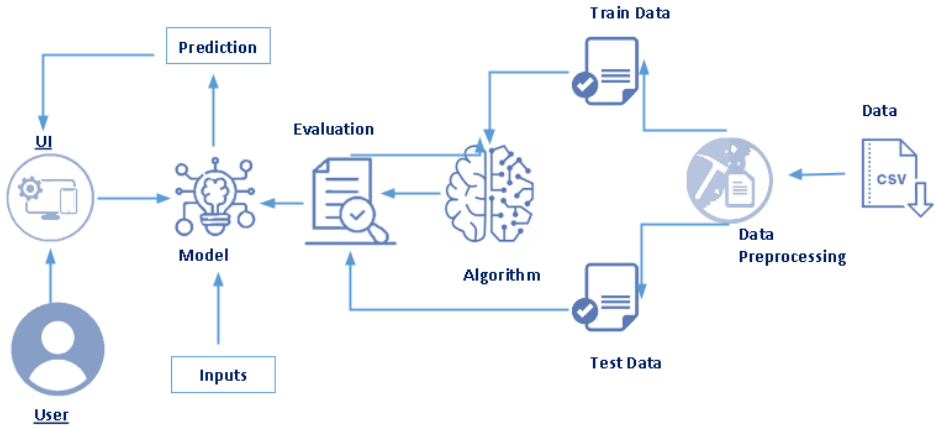
- **SCHEDULED_DEPARTURE, SCHEDULED_ARRIVAL**: scheduled times of take-off and landing
- **DEPARTURE_TIME, ARRIVAL_TIME:**: real times at which take-off and landing took place
- **DEPARTURE_DELAY, ARRIVAL_DELAY**: difference (in minutes) between planned and real times
- **DISTANCE**: distance (in miles)

1.2 Purpose :

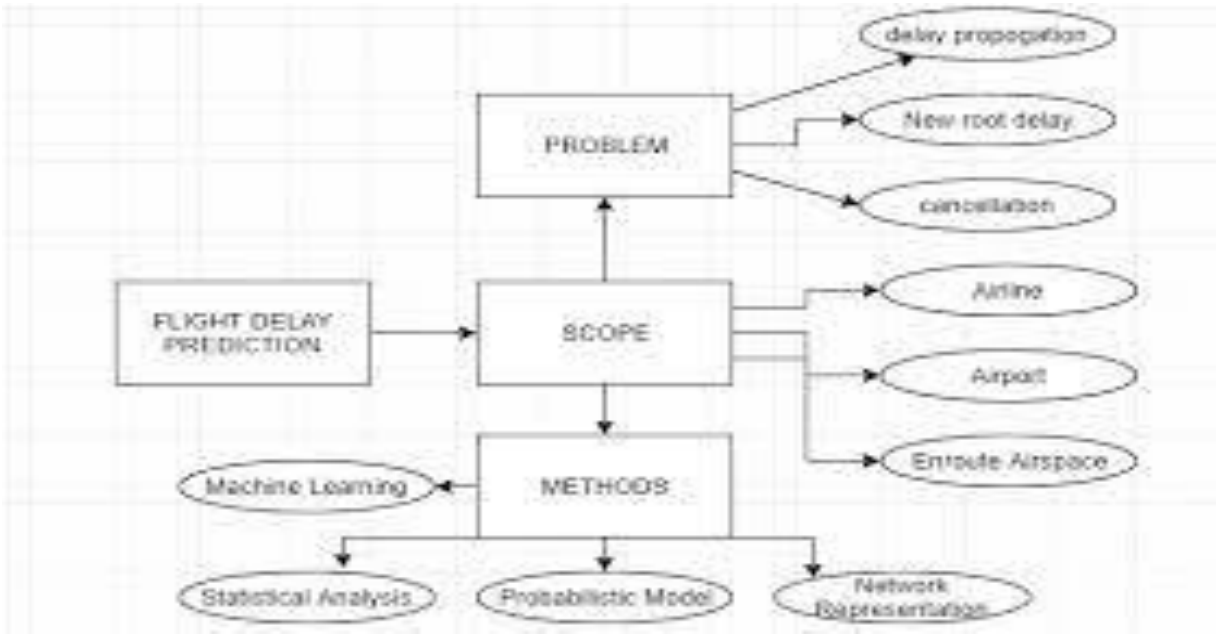
Therefore, predicting flight delays can **improve airline operations and passenger satisfaction**, which will result in a positive impact on the economy. In this study, the main goal is to compare the performance of machine learning classification algorithms when predicting flight delays.

2. Problem Definition & Design Thinking :

2.1 Empathy Map :



2.2 Ideation & Brainstorming Map :



3. Result :

4. ADVANTAGES & DISADVANTAGES :

Advantages :

When you delay, you can decide to take action later. The benefit of this is that **you give yourself time to think about what the best course of action would be**. You'll have more information about yourself and your situation and will have more options to consider when you do decide to act

Disadvantages :

Carriers attribute flight delays to several causes such as bad weather conditions, airport congestion, airspace congestion, and use of smaller aircraft by airlines. These delays and cancellations tarnish the airlines' reputation, often resulting in **loss of demand by passengers**

5. APPLICATION :

- **Arilines predict the delay of the flight.**

- **Customers can predict the flight delay.**

6. CONCLUSION :

Predicting flight delays is on interesting research topic and required many attentions these years. Majority of research have tried to develop and expand their models in order to increase the precision and accuracy of predicting flight delays. 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 their 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 inference in the cancer

domain. Regardless of the type of prediction task at hand; regression or classification. It has become the state-of-the-art machine learning algorithm to deal with structured data. Compare to all algorithms MLP algorithm gives high accuracy that is 82%

7. FUTURE SCOPE :

Based on the analysis of their 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 inference in the cancer domain. The XGBoost is used in the analysis of this paper because XGBoost is one of the most popular machine learning algorithms these days. Regardless of the type of prediction task at hand; regression or classification. It has become the state-of-the-art machine learning algorithm to deal with structured data.

8. APPENDIX :

Runway: a paved strip of ground on a landing feld for the landing and takeof of aircraft.

En-route: In aviation, an en-route chart is an aeronautical chart that guides pilots fy ing under instrument fight rules (IFR) during the en-route phase of fight.

Random Forest: It is an ensemble learning method, which uses decision tree as sub classifers, and introduces random attributes selection into the decision tree.

LSTM: LSTM network is one of most powerful RNNs with more complex cell struc ture, and overcomes the gradient vanishing problem in RNNs.

Autoencoder: An autoencoder is a type of artifcial neural network used to learn ef cient data coding in an unsupervised manner. Te aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal “noise”.

Denoising autoencoder: Denoising autoencoders are an extension of the basic autoen coder, and represent a stochastic version of it. Denoising autoencoders attempt to address identity-function risk by randomly corrupting input (i.e. introducing noise) that the autoencoder must then reconstruct, or denoise.

Weight: Weights in an ANN are the most important factor in converting an input to impact the output. Tis is similar to slope in linear regression, where a weight is multi plied to the input to add up to form the output. Weights are numerical parameters which determine how strongly each of the neurons afects the other.

Bias: is the confict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training.

Cost function: A cost function is a measure of “how good” a neural network did with respect to it is given training sample and the expected output. It also may depend on variables such as weights and biases.

Activation function: An activation function determines the output behavior of each node, or “neuron” in an artifcial neural network.

Overftting: A model overfts the training data when it describes features that arise from noise or variance in the data, rather than the underlying distribution from which the data were drawn. Overftting usually leads to loss of accuracy on out-of-sample data.

Dropout: Dropout changed the concept of learning all the weights together to learning

a fraction of the weights in the network in each training iteration.

Epoch: in neural networks generally, an epoch is a single pass through the full training set.

Supervised learning: Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input–output pairs.

Unsupervised learning: Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision.

Fine-tune: Fine tuning is a process to take a network model that has already been trained for a given task, and make it perform a second similar task.

Precision: precision is the ration of system generated results the correctly predicted positive observations (True Positive) to the system's total predicted positive observa tions, both correct (True positive) and incorrect (False Positives).

Recall: Recall is the ratio of system generated results that correctly predicted positive observations (True positives) to all observations in the actual malignant class (Actual positives).

Accuracy: Accuracy is the most intuitive performance measure and is simply a ratio of the correctly predicted classifcations (both True Positives+True Negatives) to the total Test Dataset.

Fi measure: the F1 Score is the weighted average (or harmonic mean) of Precision and Recall. Terefore, this score takes both False Positives and False Negatives into account to strike a balance between precision and Recall.

Specifcity: Specifcity (also called the true negative rate) measures the proportion of actual negatives that are correctly identifed as such (e.g., the percentage of healthy peo ple who are correctly identifed as not having the condition)