Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

Flight Delay Prediction System

Mrs Yogita Borse*, Dhruvin Jain#, Shreyash Sharma#, Viral Vora#, Aakash Zaveri#
*Assistant Professor , #UG students,
Department of Information Technology, K.J Somaiya College of Enginerring, Mumbai, India)

***Abstract*— Flight Planning is one of the challenges in industrial world which faces many uncertain conditions. One such condition is delay occurrence, which stems from various factors and imposes considerable costs on airlines, operators, and travelers. Delays in departure can occur due to bad weather conditions, seasonal and holiday demands, airline policies, technical issue such as problems in airport facilities, luggage handling and mechanical apparatus, and accumulation of delays from preceding flights. Here in flight delay prediction system based on the weather parameters which can result in delays. The system considers the temperature, humidity, rain in mm, visibility and month number as important parameters for prediction of delay.**

Keywords*— *Flight delay, weather, supervised machine learning, Naïve Bayes

I. INTRODUCTION

One of the key business issues that airlines face is that the vital prices that are related to flights being delayed because of natural occurrences and operational shortcomings that is an upscale affair for the airlines, making issues in scheduling and operations for the endusers therefore inflicting unhealthy name and client disco ntent. As we all know that we have a tendency to not get the flight delay before departure as customers of the Airline Company neither the airline company's ground staff gets the airline delay prediction supported varied conditions. However, we all know that one in all the most reasons for delay in flights is that the weather. This motivates us to use the live weather knowledge in conjunction with different metrics to calculate the delay on the wing before departure.

Indian state of affairs, in 2017, in line with the reports by the directorate General of Civil Aviation (DGCA), between January and April, close to 5.12 hundred thousand domestic passengers in India faced issues because of airline corporations denying boarding, moreover as flight cancellations and delays [2]. Airline corporations had to pay the passengers compensations of over Rs. twenty five crore for varied inconveniences throughout the first four months of this year. Hence, the prediction analysis retrieved from this project can contribute within the form of a prototype in helping to identify operational variables that contribute to delays in any country scenario[2]

The main issues associated with flight delay prediction are known and arranged in taxonomy. It includes the problem that causes the flight delay, the range of institution it affects, and ways that of handling flight delay prediction downside. It considers flight domain options, like problem and scope. Major problem which causes delay in flights can be delay propagation, delay caused on the departure point or the root of the flight, and cancellation of flights. These problems cannot be eliminated forever, but a delay prediction tool will allow the operator and the administrators to take the concerned actions for smooth operation. This problem that is causes delay affects Airline, Airport and the enroute airspace which are independent entities which works in synchronization and hence delay in flight causes issues in all the sectors. Various methods that can be used to develop a system which predicts the delay in flights can be Machine Learning, Probabilistic models, Statistical analysis or Network Representations.

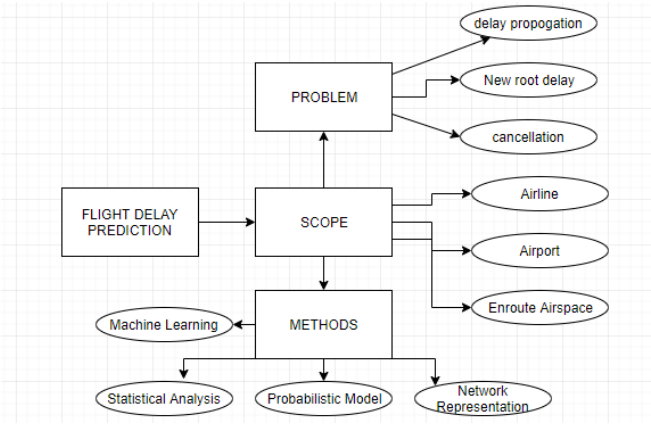

Fig. 1: Taxonomy of Flight Delay Prediction Problem

II. LITERATURE SURVEY

As discussed, considering the standard taxonomy of the flight delay and its problems, one will contemplate the scope of prediction to be one in every of these factors or combination of those factors[3]. The models developed during this system may be applied to predict the incidence of flight delay at airports. Such prognosticative capabilities would facilitate traffic managers and airline dispatchers to organize mitigation methods for reducing traffic disruptions.
This issue can be reduced by developing the flight delay prediction tool which can be developed using following methods.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

Statistical analysis

Statistical model requires the use of correlation analysis, parametric and non parametric tests, multivariate analysis and econometric models. Government agencies have invested in these econometric models to understand the relationship between delay and Passenger demand, fare, size of aircraft etc

Probabilistic models

Probabilistic model requires analysis tools that estimates the probability of an event based on the historic data. The estimated outcome is given in form of a distribution function of the probability. The factor of randomness always makes an impact on the decision or the outcome produced by the probabilistic model.

Machine Learning

Supervised Machine learning could be a task where the dataset input and also the output are recognized, then many algorithms are used to analyze this data to map new examples. Here in this case is that the prediction of delay in flight.

Supervised Learning problems can be further categorized into following problems

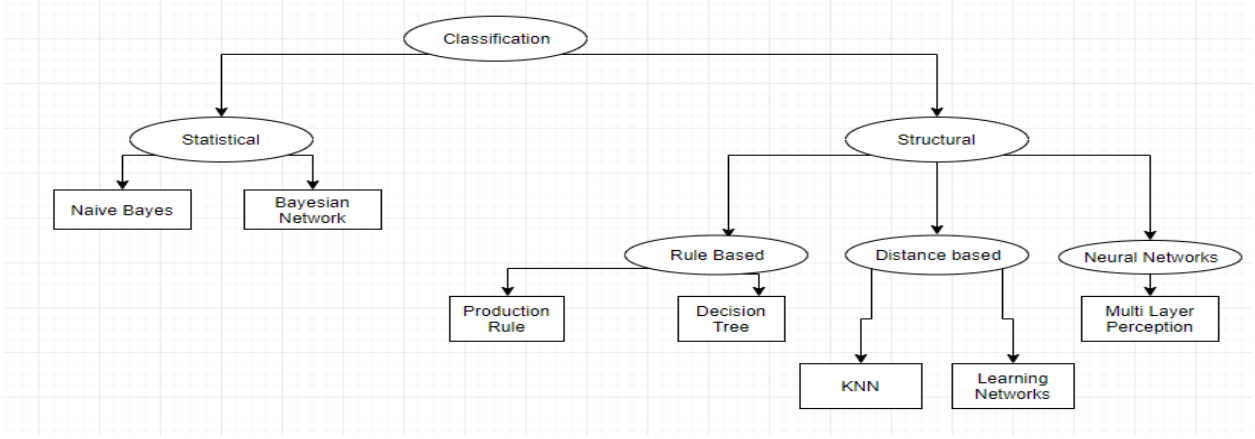

Fig 2. Overview of Classification approach

• Classification – It is a type problem in which the output variable is an entire category itself, such as “Win” or “Lose”, the entire input data is classified into the category variables; it is generally used largely for recommendation problems

• Regression – It is a type of problem is which the output variable is a real value, such as few raw data values related to something.

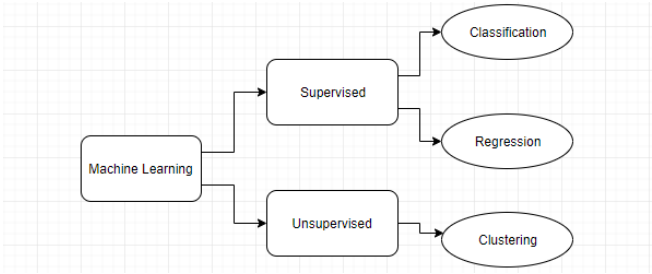

Fig 3. Overview and classification of machine learning

III. CLASSIFICATION APPROACH

Naïve Bayes Classification

Naive Bayes classifiers comes under the family of straightforward "probabilistic classifiers” in machine learning domain ,the attributes used in the naïve bayes classifier is assumed to be independent from one another. 1950s was the time when naïve bayes was studied more than ever before and it has had a special name since that time in the text retrieval problem solving community during the 1960s and it is still considered as a root method for text categorization, the matter of judgment documents as turns out into one class or the opposite (such as spam or legitimate, sports or politics, etc.) with word frequencies because options. Once the threshold and rules has been provided beforehand, it's competitive during this section with a lot of advanced ways as well as support vector machines. It additionally finds application in automatic diagnosis.

Naive Bayes classifiers are extremely scalable, it needs number of variety of maximum-likelihood training will be done by evaluating a closed-form expression,[1] that takes linear time, instead

of by pricy repetitious approximation as used for several alternative forms of classifiers. In the statistics and engineering literature, naive bayes models are best-known below a range of names, together with easy Bayes and independence Bayes all these names reference the utilization of theorem within

the classifier's call rule, however naive bayes isn't (necessaril y) a theorem technique

Bayesian Network (BN) Algorithm

B Network is a math supervised learning algorithm that represents the connections between variables. The theorem network may well be a directed acyclic graph that consists of nodes and edges. it's supported applied math. once constructing best BN, it performs classification task using probabilistic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

Naive Bayes classifier performed surprisingly well achieving error rates between 15 and 18 percent. However, closer investigation of the classifiers showed that they were also mostly predicting that flights would not be delayed. The precision was also fairly high around 75% but
the recall varied widely, from 37.32% on Newark International flights to 7.6% on the entire dataset.[7]

IV. OTHER METHODOLOGY

Various methodology can be applied to implement the system that predicts the delay in flight. Few of those methodology are discussed below.

Decision Tree: As the name suggest the main idea behind decision tree algorithm is to make a tree like structure and get the answers in form of true of false. The model begins from a root node and ends on the decision. Each node receives a Yes No question and answer is passed on to the next node. Root node gets all the input of the training dataset.
The challenge to assembling such a tree is that question to ask at a node and when. To do this, decision tree algorithmic program uses accepted indices like entropy or Gini-impurity to quantify an uncertainty or impurity related to an explicit node. Equations (1) and (2) show however entropy and Gini impurity are calculated, severally, for a setoff information. Within the equations, C is that the variety of classes[1].

$$H(s) = -\sum_{c \in C} p(c) log p(c)$$

$$H(s) = 1 - \sum_{c \in C} p(c)^2$$

Logistic Regression: Logistic regression an algorithm that performs classification using,

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

Which in turn represents the maximum likelihood of estimation and gradient ascent
Logistic regression is the applicable multivariate analysis to conduct once the variable is divided (binary). Like all regression analyses, the logistical regression is a predictive analysis. logistic regression is employed to explain data and to explain the relationship between one dependent binary variable and one or additional nominal, ordinal, interval or ratio-level independent variables.

Sometimes logistic regressions are tough to interpret; the Intellects Statistics tool simply permits you to conduct the analysis, then in plain English interprets the output

Neural Network: Neural Network is made by stacking along multiple neurons in layers to provide a final output. Initial layer is that the input layer and therefore the last is that the output layer. All the layers in between is named hidden layers. every nerve cell has an activation function. a number of the popular activation functions are Sigmoid, ReLU, tanh etc. The parameters of the network are the weights and biases of each layer. The goal of the neural network is to search out the network parameters specified the expected outcome is that an equivalent as the ground truth. Back-propagation on loss-function is employed to search out the network parameters [1]

Algorithm	Decision Tree	Logistic Regression	Neural Network
Precision	.93	.92	.91

Classification Report of Decision tree, logistic regression and Neural Network Classifiers.

Number of test samples used to generate the reports is 15001.
Data patameters used for the algorithms are Month, Day, Day of the week, Flight Number, Origin airport, Destination Airport, Scheduled departure, departure delay, taxi-out, distance, Scheduled Arrival.
These data features are the ones which are usually known beforehand.

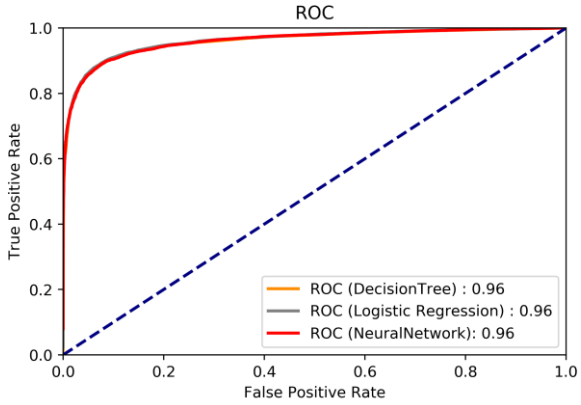

Fig 4[1]: Receiver Operating Curves for Decision Tree, Logistic Regression and Neural Network models

Figure 4 shows the receiver operating curves (ROC) for all three classifiers with an area under the Curve. The observation here is that decision tree classifier turns out to be better at predicting on time flights whereas performance of neural network has be better at delayed flight's prediction.The difference is, however, very small[1].
For a balanced dataset. whereas the accuracy of the most effective algorithmic rule for a two hour prediction with a sixty minute threshold is 93.7%, even a naïve classifier that continually predicts a delay below the edge can provide an accuracy of 93.5% [4].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

V. PROPOSED SOLUTION

As discussed, weather condition plays an important role in proper and timely functioning of flights. We propose a flight delay prediction system which focuses mainly on predicting delay of a flight based on the weather situation. To make the system more scalable it is necessary to choose an algorithm which considers all the parameters to be independent. Supervised learning as the name indicates a presence of supervisor as teacher. Essentially supervised learning could be a learning that within which we tend to teach or train the machine exploitation data which is well tagged which means some data is already labeled with correct answer. After that, machine is given new set of examples(data) so supervised learning algorithmic rule analyses the coaching knowledge(set of training examples) and produces an correct outcome from tagged data Using supervised machine learning approach, the labeled data gives it authenticity. Naïve bayed model is one of the algorithm which is proven to be efficient for real time prediction as well as the fact that it considers every attribute to be independent from each other makes it an apt algorithm for the concerned project

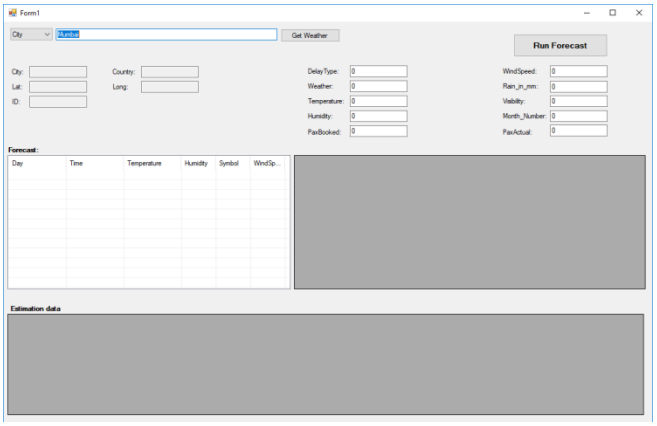

Fig 5: System User Interface

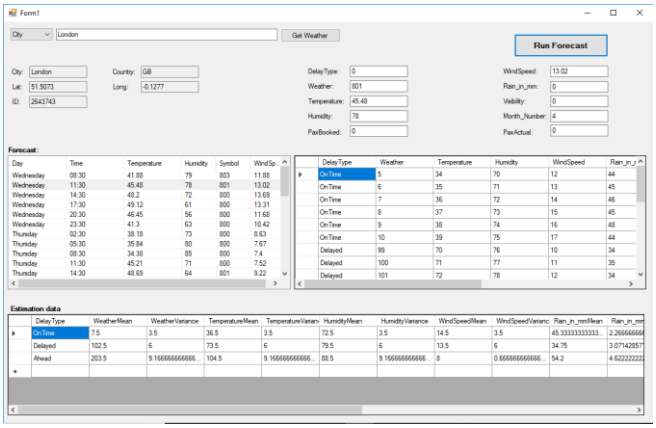

Fig 6: System User Interface with results

The proposed system takes the city of departure as its input in a textbox field as shown in fig 5 and 6. It then returns the predicted weather data using an API (Application Program Interface) and passes the data into the algorithm. The attributes considered for calculations and taken by the API are as follows weather, temperature, humidity, Rain in mm, Visibility and Month number. As discussed that supervised machine learning is based on having a set of correct labeled data form which the algorithm bases its prediction. We use a CSV file for storing that data as a flat file format is easier to edit , update and retrieve it for calculations.

VI. FUTURE SCOPE

Further supportive study is required to correlate all the problem, scope and method for getting most accurate result. Although weather conditions are the major reasons for flight delay, other unprecedented events such as major calamities , natural or man-made can cause major delay in flight.

CONCLUSION

This paper presented the need to develop a system to predict the delay in flights along with its methodology. The paper gives details about the range of different methodology that is used or can be used to find out the delay in flights. As flight delay cost a lot to the airlines as well as passangers in financial and environmental terms, flight delay is a the talk of the hour.Flight delay causes surging of prices by costing a lot on operational purpose They may increase prices to customers and operational prices to airlines. As the outcome is directly associated with the passanger and the airlines which inturn is liked to another set of airline and pasaangers it is very crucial to get real time delay for each player within the air transport system. hence there is a requirement to develop a system to predict the delay in flights to scale back monetory loss and for the higher and smooth operation. Classification or reggrerssion ways are often accustomed determine the delay which includes Feed forward network, Neural Network, Random Forrest, decision tress, Naïve Bayes Classification Tree, Regression Tree, etc. As seen from the articles and papers these methodologies offer virtually identical accuracy however we want an algorithmic rule that is good with real world prediction and analysis and thus: naïve-Bayes. except being smart with real time prediction algorithmic rule that considers or assumes independence among predictors that makes the system scalable as other independent attribute may be superimposed up to the algorithmic rule for computation of the delay. the expected delay can thus facilitate the ground employees for creating correct and smooth operation plans and therefore the data if sent to the passengers will profit the airlines also because the passengers

REFERENCES

[1] Kuhn, Nathalie and Navaneeth Jamadagni. “Application of Machine Learning Algorithms to Predict Flight Arrival Delays.” (2017).
[2] N, Prabakaran & Kannadasan, Rajendran. (2018). Airline Delay Predictions using Supervised Machine Learning. International Journal of Pure and Applied Mathematics. 119.
[3] A Review on Flight Delay Prediction Alice Sternberg, Jorge Soares, Diego Carvalho, Eduardo Ogasawara _ CEFET/RJ Rio de Janeiro, Brazil November 6, 2017

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 03, March-2020

[4] Gopalakrishnan, Karthik and Hamsa Balakrishnan. "A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks." Air Traffic Management Research and Development Seminar, June 2017, Seattle, Washington, USA, ATM Seminar, June 2017 © 2017 ATM Seminar

[5] Rebollo, Juan Jose and Balakrishnan, Hamsa. "Characterization and Prediction of Air Traffic Delays." Transportation Research Part C: Emerging Technologies 44 (July 2014): 231–241 © 2014 Elsevier LtdA model for acurracy prediction using geoRSS using naive bayes

[6] A Model for Accurate Prediction in GeoRSS Data Using Naive Bayes Classifier K Netti* and Y Radhika CSIR-National Geophysical Research Institute, Uppal Road, Hyderabad-500007, India Received 17 April 2016; revised 15 October 2016; accepted 13 February 2017

[7] Naive baye's classification algorithm in prediction of Flight delays using MR Ujwalla Urkurde and Prathiba Richariya IJRIT April 2016.

[8] On-Time Flight Departure Prediction System Using Naive Bayes Classification Method (Case Study: XYZ Airline) IJCTT december 2017

[9] Predicting Flight Delays Dieterich Lawson jdlawson@stanford.edu William Castillo will. castillo@stanford.edu

[10] https://en.wikipedia.org/wiki/Naive_Bayes_classifier [Dated:-14/1/2019 11:08:00].

[11] http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/[Dated:-3/2/2019 12:26:00].

[12] https://acadgild.com/blog/naive-bayesian-model. [Dated:-3/2/2019 13:20:00].

[13] https://www.quora.com/In-what-real-world-applications-is-Naive-Bayes-classifier-used[Dated:-3/2/2019 14:05:00].

[14] https://blog.statsbot.co/machine-learning-algorithms-183cc73197c[Dated:-3/2/2019 14:18:00].

[15] https://www.statisticssolutions.com/what-is-logistic-regression/[Dated:-3/2/2019 14:40:00].