

Flight Delay Prediction Based On Aviation Big Data And Machine Learning

Naveen Kumar¹, Hanumanth², Nagareddy³, Jyothi Patil⁴

¹Student, Department of Computer Science, PDA College of Engineering, Kalaburagi, India.
naveenindurkar232@gmail.com

²Student, Department of Computer Science, PDA College of Engineering, Kalaburagi, India.
hanumanthmatagar42@gmail.com

³Student, Department of Computer Science, PDA College of Engineering, Kalaburagi, India.
nagareddyatil265@gmail.com

⁴Professor, Department of Computer Science, PDA College of Engineering, Kalaburagi, India.

ABSTRACT

Among the many difficult scenarios in the business world is flight planning, which must account for a wide range of uncertainties. This situation in delay incidence results from a number of causes and places significant financial burdens on airlines, operators, and passengers. Airport infrastructure, luggage handling, and mechanical equipment, as well as the cumulative delays from prior flights, may all contribute to departure delays, along with severe weather, peak travel times, airline rules, and technical issues. In This Aviation Data-driven algorithm forecasts potential aircraft delays. The system takes into account a number of factors. This system employs the algorithms Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM).

Keywords: Prediction, Aviation, Machine Learning

I. INTRODUCTION

Vital costs associated with flight delays due to natural occurrences as well as operational shortcomings are an expensive affair for airlines, causing scheduling and operational issues for end users, which in turn leads to a negative reputation and customer dissatisfaction. Knowing that both the passengers and the ground crew of the departing airline get a prediction of delay from the supporting airline under varying circumstances, aircraft delays are quite rare. Of course, we all know that bad weather is a major contributor to flight delays, and that unforeseen difficulties, such as mechanical failure, may also put passengers at danger. Because of this, we determined wing lag in advance of takeoff using real-time weather data and other indicators. According to DGCA data, between January and April of 2017, about 5.12 million Indian domestic passengers had difficulties due to airline firms not boarding, as well as flight cancellations and delays. Over the first four months of that year, airline companies paid out compensations to customers totaling over 25cr due to a variety of problems. Therefore, the prototype prediction analysis developed via this study may help pinpoint the operational elements that cause delays in any given scenario. Predicting flight delays is possible for a variety of reasons, the most common of which are recognized and classified in a taxonomy. This comprises the root cause of the flight delay, the scope of its effects, and potential solutions to the issue of flight delay prediction. It takes into account potentials in the field of airlines, such as issues and answers. Major issues that cause flight delays include the spread of delays, delays at the airport, and aircraft cancellations. While there is no permanent solution to these issues, a delay predictor will let operators and administrators take corrective measures. There are several parties involved in this delay problem, including airlines, airports, and the airspace used for rerouting flights. As a result, flight delays are an issue in every respect. There are a number of ways to build a system that can anticipate flight delays, including using machine learning, probabilistic models, statistical analysis, or network representations.

II. LITERATURE SURVEY

Considering the wide-ranging effects that aviation delays may have, it's no surprise that delay prediction models have advanced significantly during the 1990s. The length of the delay impacted the efficacy of marketing plans. If a domestic flight is late taking off or landing, it might disrupt an international flight's schedule. For airport industries, even a little reduction in the delay value may have a huge impact. Predicting the frequency of aircraft delays at airports is possible with the help of the models established throughout this system. Such predictive skills would aid in the organization of mitigation efforts by traffic managers and airline dispatchers to reduce traffic interruptions. Making use of the following strategies, one may reduce this challenge by developing a flight delay prediction tool.

[1] Guan Gui and Fan Liu, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning", IEEE 2020. To create a more productive airline company, precise forecast of flight delays is essential. Applying machine learning techniques to forecast aircraft delays has been the subject of recent research. Prior techniques of forecasting often only test hypotheses in a particular environment, such as an airport or specific route. In this study, we look at a wider range of causes that might contribute to a flight being late, and we evaluate multiple machine learning-based models using artificially created extended flight delay prediction tasks. By collecting, preprocessing, and combining data sources including weather forecasts, aircraft schedules, and airport details, the suggested technique may provide a useful dataset. Numerous classification problems, in addition to a regression challenge, make up the specified prediction tasks. The acquired airplane sequence data is well within the capabilities of long short-term memory (LSTM), as shown experimentally; yet, the overfitting issue arises in our little dataset. The suggested random forest-based model is superior to earlier methods in both prediction accuracy and avoiding overfitting.

[2] N Lakshmi Kalyani and Bindu Sri Sai U, "Machine Learning Model – based Prediction of Flight Delay", IEEE 2020. Predicting when a flight will arrive is important for passengers and airlines alike, as the former stand to lose a great deal of money due to delays while the latter risk having their years-long reputation and customer base damaged. Our paper's goal is to use existing data to forecast the arrival delay of a planned individual aircraft at the destination airport. Using supervised machine learning methods, the model shown here can predict when flights will arrive late. Information on domestic flights in the United States, as well as meteorological conditions, during the period of July 2019 through December 2019 were collected for use in training the prediction model. The predictive model used to anticipate aircraft delays was built using XGBoost and linear regression methods. Algorithms were evaluated based on how well they performed. The model was given flight data and meteorological data to analyze. A binary classifier trained using XGBoost was used to determine whether there would be a delay in the flight's arrival, and a linear regression model was used to estimate how long the delay would last.

[3] Jiage Huo and K.L. Keung, "The Prediction of Flight Delay: Big Data-driven Machine Learning Approach", IEEE 2020. The Hong Kong International Airport is now experiencing overcrowding and saturation. Increased passenger and freight traffic at Hong Kong International Airport, without a corresponding increase in runway capacity, has led to serious complications, such as difficulty in choosing taxiways and a lengthening of the lead time at the runway holding position. The primary focus of this research is on making accurate flight delay predictions using machine learning techniques. Using actual data from the Hong Kong International Airport, we compare and fully assess the prediction outcomes of numerous machine learning algorithms. The aviation and insurance sectors may benefit from this paper's findings and suggestions. Predicting aircraft delays allows airports to better organize their infrastructure.

III. EXISTING SYSTEM

The findings of a reliable flight delay prediction may be used to boost airline agency profits and customer happiness. Many studies have been conducted to model and forecast aircraft delays, with most of them focusing on extracting significant characteristics and most connected aspects to better predict flight delays. Massive data volumes, dependencies, and an excessive number of factors, however, render most suggested approaches inaccurate.

3.1 Disadvantages:

- The research precision of flight delays is lower.
- The necessary criteria for determining flight delay are missing.

IV. PROPOSED SYSTEM

This proposed work uses machine learning algorithms to predict flight delays, which is important because the aviation industry is a major contributor to many countries' economies and because air travel is the most convenient and time-efficient mode of transportation available. The results of this simulation show the time, day, weather, etc., that major airports are most likely to have delays, and therefore the volume of delay should be minimal based on the developed model.

4.1 Advantages:

– Due to the unpredictability of delays, this study looks at qualitative prediction of airline delays in order to make the required adjustments and improve the customer experience.

V. ARCHITECTURE

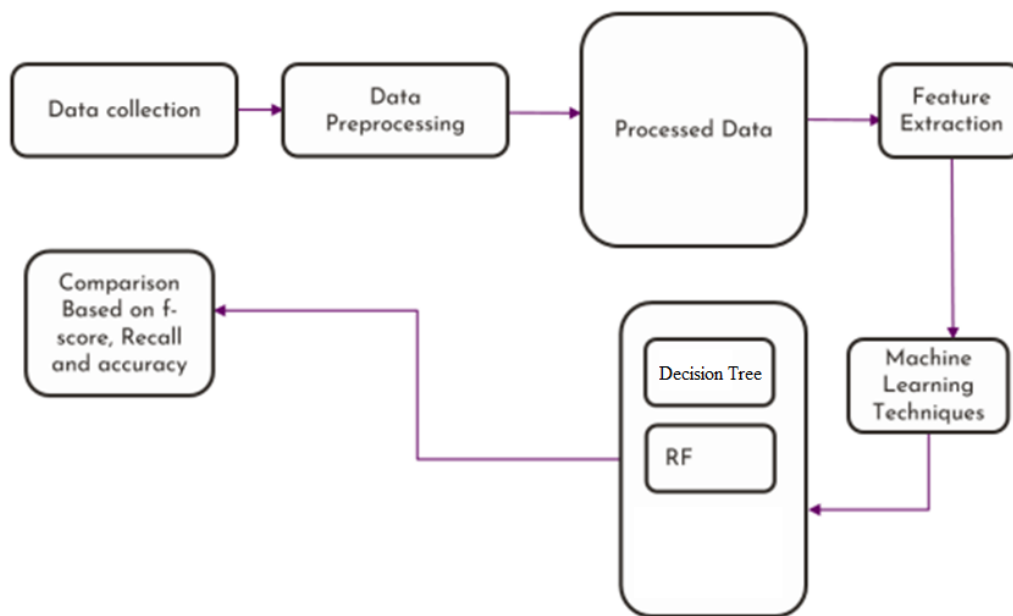


Fig 1: System Architecture

VI.SYSTEM REQUIREMENTS

HARDWARE AND SOFTWARE REQUIREMENTS

6.1 HARDWARE REQUIREMENTS:

- Processor : Intel Core I3 and above
- Processor Speed : 1.0GHZ or above
- RAM : 4 GB RAM or above
- Hard Disk : 500 GB hard disk or above

6.2 SOFTWARE REQUIREMENTS:

- Operating System : Windows 7/10 or above
- Front End : Python
- Back End : SQLite3

VII. FEASIBILITY STUDY

Preliminary research looks at whether or not the project is doable and whether or not the system will be valuable to the company. The primary purpose of a feasibility study is to examine the practicality of implementing new features and fixing bugs in an existing system from a technical, operational, and financial perspective. If there had limitless time and materials, any system would be doable. The feasibility analysis is a managerial task. Finding out whether an information system project is feasible and offering potential alternatives are the goals of a feasibility study.

There are aspects in the feasibility study portion of the preliminary investigation:

- ✓ Technical Feasibility
- ✓ Operational Feasibility
- ✓ Economical Feasibility

7.1 Operational Feasibility

It alludes to how likely it is that the product will really work. Some goods may function perfectly in the lab but fall short when put to the test in the real world. Examining the need for, and the availability of, more technical personnel is part of this process. It entails making assumptions about the system's potential uptake based on the skillsets of the people currently working on the project. It's a tally of how effectively a proposed system addresses issues and makes the most of possibilities discovered in scope definition, as well as how well it meets needs discovered in analysis. It takes a look at whether or not the company is willing to back the new system. The level of management support for the proposed project is a key consideration in making this determination.

7.2 Technical Feasibility

The question is whether or not the currently available software provides enough support for the given application. It analyzes the benefits and drawbacks of implementing a certain piece of software into the development process. It also investigates the supplementary instruction required of the workforce for the application to succeed. The next step is to examine how well the company's technical resources meet the technological needs. If the organization's current level of technical expertise is enough to meet the needs of the systems project, then it is regarded technically viable. An improvement or addition to existing technological resources may be possible, and the analyst must determine whether this is the case in order to proceed with the request at hand.

7.3 Economic Feasibility

It's the ratio of the product's overall advantages or outcomes to its entire cost to produce. It is not practical to build the product if it is functionally equivalent to the previous system. An alternative name for economic analysis is cost-benefit analysis. It is the gold standard for measuring how well a new system performs. The standard practice in economic analysis is to weigh the potential gains and savings of a potential system against its total price tag. There will be a choice to build and deploy the system if the advantages exceed the expenses. Before taking any kind of action, a business owner has to carefully consider the costs and advantages.

VIII. SYSTEM TESTING

Introduction:

The software testing process is the last check on the specifications, designs, and codes that make up a piece of software. The goal of testing is to detect bugs in a software by running it. Testing involves running the code in question alongside a collection of test cases and analyzing the results to see whether they conform to specifications.

8.1 Testing Objectives:

- The term "testing" refers to the process of running a program with the intention of discovering bugs.
- A well-designed set of test cases should be able to unearth previously unknown bugs.
- A successful test is one that reveals a previously unknown flaw.
- A radical shift of perspective is required to achieve the aforementioned goals. The only thing that software testing can prove is that there are bugs in the system.

8.2 The following are the Testing methodologies:

Unit Testing:

When verifying software, unit testing isolates and evaluates individual modules. This test makes sure each component works as intended before moving on to the next. Unit testing describes the practice of testing individual modules.

Integration Testing:

It's a methodical approach to building independent pieces of code into a larger, more functional whole. By validating the complete module, this test reveals any flaws during its execution.

Output Testing:

It is possible to test the output to see whether it is correct.

Validation Testing:

The program is complete and meets all requirements once it has undergone integration testing. However, verification against the specification is necessary to find any unforeseen future flaws and boost its dependability.

Software Testing Strategies:

The software engineer has a road plan thanks to the software testing technique they used. You may prepare for and carry out tests in a methodical manner. This is why it's important to develop a software testing framework, or a series of phases into which we can slot various test case creation techniques. The following elements are necessary for every software testing strategy:

Starting with individual modules, testing progresses "outward" to cover the whole computerized system.

1. There are times and places for each possible kind of testing.
2. Both the software's creator and an external testing team are responsible for the process.
3. Although testing and debugging are distinct processes, any good testing plan will include debugging..

IX. DESIGN & ANALYSIS

Modules

- Load dataset: Our dataset will be uploaded to the application.
- Data Preprocessing: Before using our methods, it is important to verify the accuracy of the input data.
- Feature Extraction: Processing raw data by transforming it into numerical characteristics while keeping the original data set intact.
- Generate models: To construct our algorithms, we first must extract characteristics from the dataset.
- Accuracy Graph: We will create a graph comparing the precision of each method.

9.1 DATAFLOW

Introduction:

In the design phase, you'll work out the kinks in how you're going to implement the fixes outlined in the requirements papers. In this stage, we take our first baby steps toward solving the issue. In other words, design leads us to how to define requirements once we start with what is required.

UML Diagrams:

The software engineer may express an analytical model in the unified modeling language using the modelling notation, which is subject to a standard set of syntactic, semantic, and pragmatic criteria.

9.2 Data Flow Diagram (DFD):

Data Flow Diagrams (DFDs) are graphical representations of how data moves through a system, including its entry and exit points and its storage locations. A data flow diagram is a visual depiction of a business process.

DFD's benefits

All parties participating in the system and with the users will have no trouble grasping these straightforward notations. Users may participate in DFD research to improve its precision. Inspecting charts and beginning to take precautions early on helps reduce the likelihood of a failed system.

The goal of the "bubble Chart" representation of a DFD is to define the system's requirements and to identify the main changes that will become the design's programs. Therefore, it is the first step in designing with the smallest possible increments of detail. A DFD depicts a system's data flows as a sequence of bubbles connected by lines.

The notations used to draw DFD are as follows:

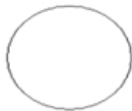



Name	Symbol	Meaning
process		Transforms of incoming data flows(s) to outgoing data flows(s).
Data Store		A repository of data that is to be store for use by one or more processes.
Data Flow		Movement of the data in the system.
External Entity		Sources and Destination outside the specified system boundary.

Table: 1 symbols used in DFD

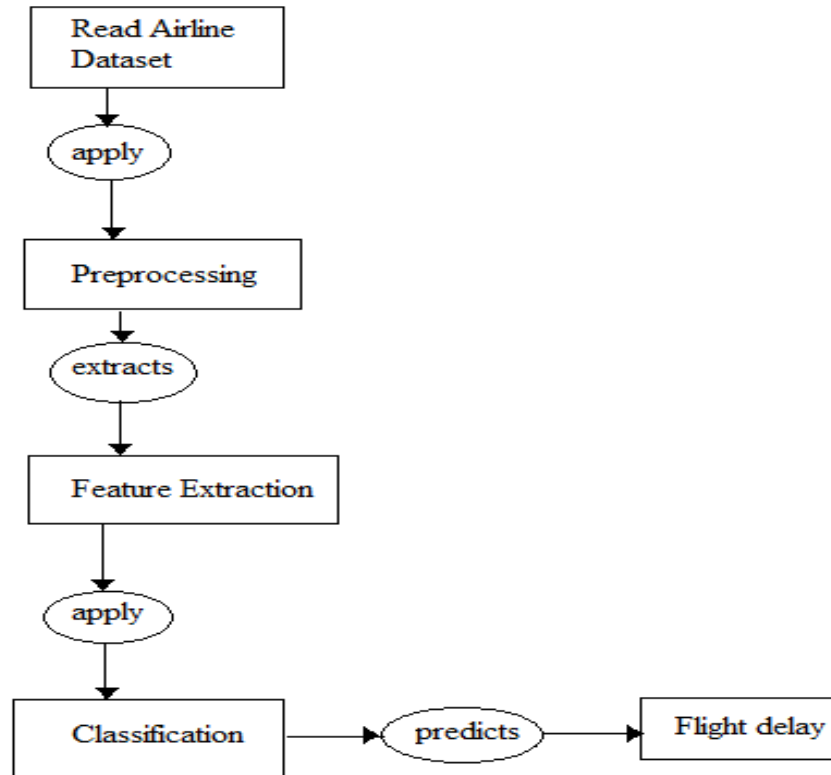


Fig:2 Data Flow Diagram

9.3 USE CASE

The dynamic behavior of a system is the most crucial feature to represent. When we talk about a system's dynamic behavior, we're referring to how it acts when it's really in use.

To accurately represent a system, one must take into account both its static and dynamic behaviors. Use case diagram is one of five such diagrams provided by UML for modeling such dynamism. Since the use case diagram is inherently dynamic, we must now consider what internal or external elements contribute to the formation of the interaction.

Actors refer to both internal and external factors. Actors, use cases, and their connections make up use case diagrams. The diagram represents a representation of an application's system or subsystem. Each use case diagram represents a specific feature of the system.

Therefore, several use case diagrams are used to represent the whole system.

X. IMPLEMENTATION & RESULT ANALYSIS

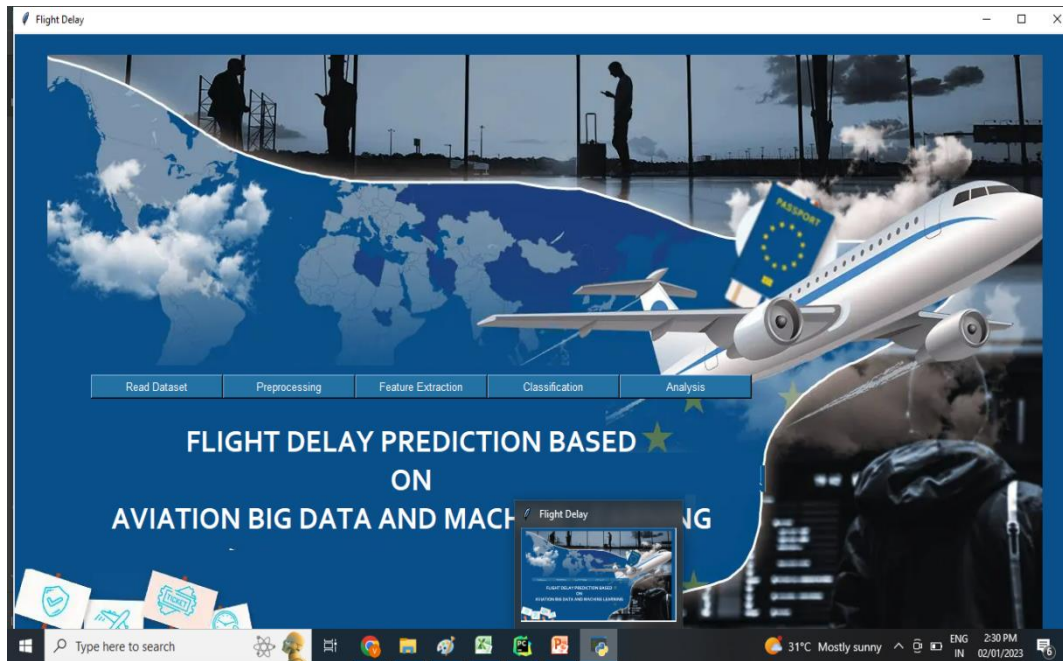
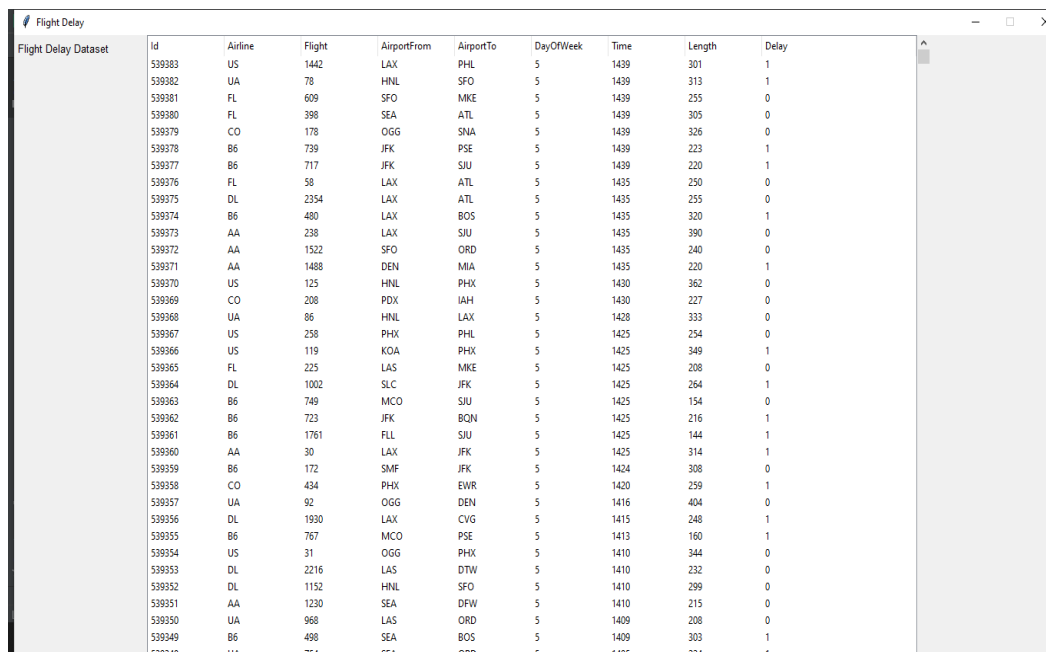


Fig 3: Main



Id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay
539383	US	1442	LAX	PHL	5	1439	301	1
539382	UA	78	HNL	SFO	5	1439	313	1
539381	FL	609	SFO	MKE	5	1439	255	0
539380	FL	398	SEA	ATL	5	1439	305	0
539379	CO	178	OGG	SNA	5	1439	326	0
539378	B6	739	JFK	PSE	5	1439	223	1
539377	B6	717	JFK	SIU	5	1439	220	1
539376	FL	58	LAX	ATL	5	1435	250	0
539375	DL	2354	LAX	ATL	5	1435	255	0
539374	B6	480	LAX	BOS	5	1435	320	1
539373	AA	238	LAX	SIU	5	1435	390	0
539372	AA	1522	SFO	ORD	5	1435	240	0
539371	AA	1488	DEN	MIA	5	1435	220	1
539370	US	125	HNL	PHX	5	1430	362	0
539369	CO	208	PDX	IAH	5	1430	227	0
539368	UA	86	HNL	LAX	5	1428	333	0
539367	US	258	PHX	PHL	5	1425	254	0
539366	US	119	KOA	PHX	5	1425	349	1
539365	FL	225	LAS	MKE	5	1425	208	0
539364	DL	1002	SLC	JFK	5	1425	264	1
539363	B6	749	MCO	SIU	5	1425	154	0
539362	B6	723	JFK	BCN	5	1425	216	1
539361	B6	1761	FLL	SIU	5	1425	144	1
539360	AA	30	LAX	JFK	5	1425	314	1
539359	B6	172	SMF	JFK	5	1424	308	0
539358	CO	434	PHX	EWR	5	1420	259	1
539357	UA	92	OGG	DEN	5	1416	404	0
539356	DL	1930	LAX	CVG	5	1415	248	1
539355	B6	767	MCO	PSE	5	1413	160	1
539354	US	31	OGG	PHX	5	1410	344	0
539353	DL	2216	LAS	DTW	5	1410	232	0
539352	DL	1152	HNL	SFO	5	1410	299	0
539351	AA	1230	SEA	DFW	5	1410	215	0
539350	UA	968	LAS	ORD	5	1409	208	0
539349	B6	498	SEA	BOS	5	1409	303	1
539348	DL	762	SEA	ORD	5	1409	272	1

Fig 4: Dataset

	id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay
0	1	CO	269	SFO	IAH	3	15	205	1
1	2	US	1558	PHX	CLT	3	15	222	1
2	3	AA	2400	LAX	DFW	3	20	165	1
3	4	AA	2466	SFO	DFW	3	20	195	1
4	5	AS	108	ANC	SEA	3	30	202	0

```
#      Column      Non-Null Count  Dtype
---  -
0      id           539383 non-null  int64
1      Airline        539383 non-null  object
2      Flight         539383 non-null  int64
3      AirportFrom    539383 non-null  object
4      AirportTo      539383 non-null  object
5      DayOfWeek      539383 non-null  int64
6      Time           539383 non-null  int64
7      Length         539383 non-null  int64
8      Delay          539383 non-null  int64
dtypes: int64(6), object(3)
memory usage: 37.0+ MB
```

Fig 5: Preprocessing

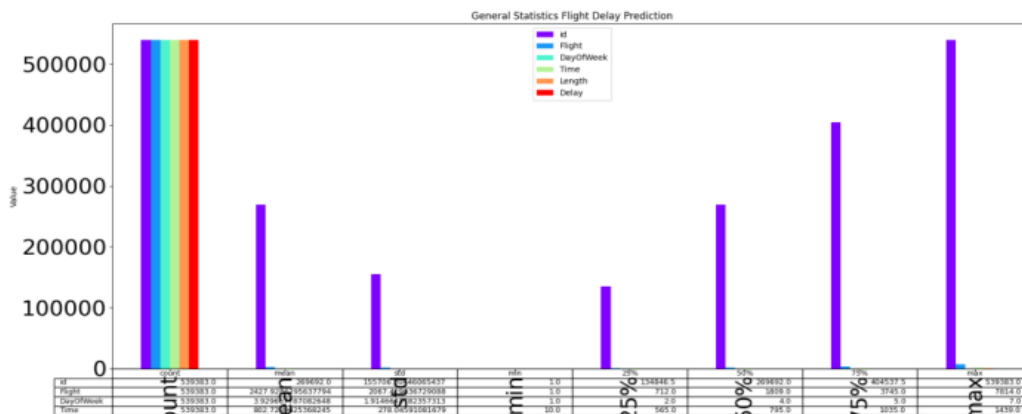
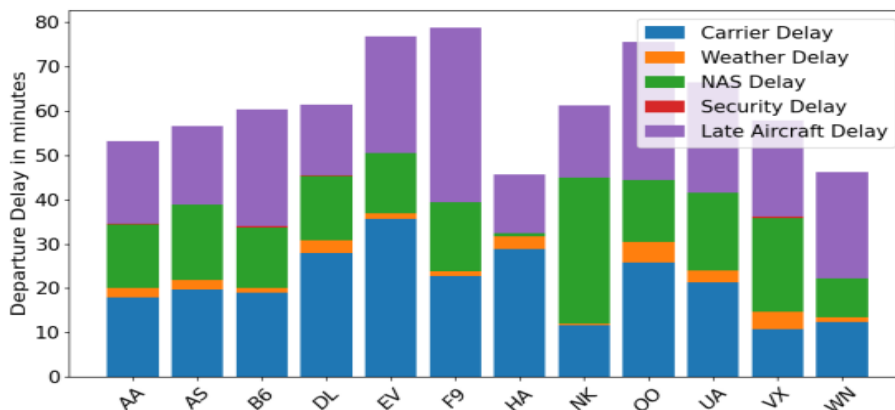


Fig 6: Statistics



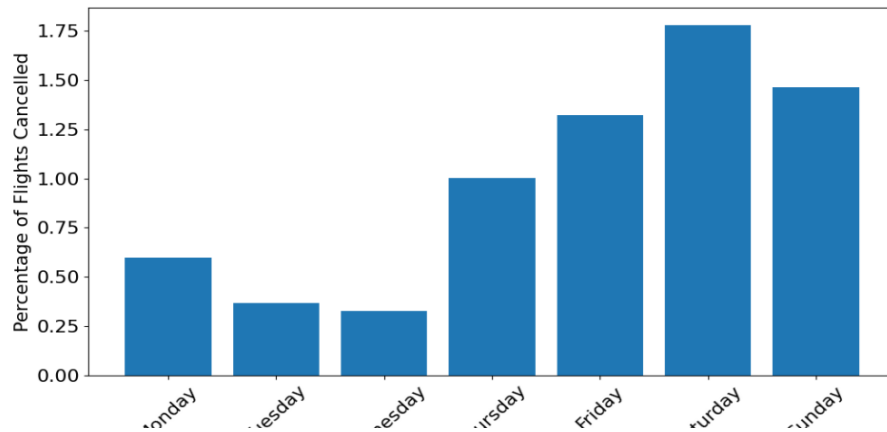


Fig 7 (a,b): Classification

XI. CONCLUSION

As a result, it is possible to anticipate flight delays using machine learning and deep learning algorithms. The goal of this categorization and analysis is to determine how long it takes to accomplish the many goals of humanity as well as to identify the causes of this delay. In order to construct the proposed system, we used SVM, RF, and KNN. Using the aforementioned methods, we may determine the overall precision, recall, and accuracy. According to the published materials Nearly as accurate as these methods are, however, is naive-Bayes's algorithmic rule, which excels in the prediction and analysis of the actual world. The only way to make the system scalable is to use an algorithmic approach for making predictions in real time that takes into account or assumes independence among forecasts. The anticipated delay can help the ground crew implement accurate and simple operation plans, and the facts, if sent to the passengers, will benefit the airlines and the passengers. This is because other independent attributes can be superimposed up to the algorithmic rule for computation of the delay.

REFERENCES

- [1] Guan Gui and Fan Liu, "Flight Delay Prediction Based on Aviation Big Data and Machine Learning", IEEE 2020.
- [2] N Lakshmi Kalyani and Bindu Sri Sai U, "Machine Learning Model – based Prediction of Flight Delay", IEEE 2020.
- [3] Jiage Huo and K.L. Keung, "The Prediction of Flight Delay: Big Data-driven Machine Learning Approach", IEEE 2020.
- [4] Vijayarangan Natarajan and Shubham Sinha, "A Novel Approach: Airline Delay Prediction Using Machine Learning", IEEE 2018.
- [5] Sabina Anjum and Asra Fatima, "Predictive Analytics For FIFA Player Prices: An ML Approach", *JSRT*, vol. 1, no. 6, pp. 204–212, Sep. 2023.
- [6] Tianyi Wang, Samira Pouyanfar, Haiman Tian, Miguel Alonso Jr., Steven Luis and Shu-Ching Chen "A Framework for Airfare Price Prediction: A Machine Learning Approach", IEEE 2020.
- [7] Viet Hoang Vu, Quang Tran Minh, Phu H. Phung, "An Airfare Prediction Model to Developing Market", IEEE 2020.
- [8] K. Tziridis, Th. Kalampokas, G.A. Papakostas, K.I. Diamantaras "Airfare Prices Prediction Using Machine Learning Techniques", IEEE 2020.
- [9] Gaurav Prajapati, Avinash, Lav Kumar, and Smt. Rekha S Patil, "Road Accident Prediction Using Machine Learning", *JSRT*, vol. 1, no. 2, pp. 48–59, May 2023.
- [10] Hao li, Yu Xiong, "Dynamic Pricing of Airline Tickets in Competitive Markets", IEEE 2008.