# DEVELOPING A FLIGHT DELAY PREDICTION

USING MACHINE LEARNING

*A Project report submitted in partial fulfilment of 7th semester in degree of*

BACHELOR OF ENGINEERING

IN

# COMPUTER SCIENCE AND ENGINEERING

***Submitted by***

Team ID: PNT2022TMID49886

|  |  |
| --- | --- |
| S.PRIYADHARSHINI | 950619104050 |
| S.SUJITHA | 950619104066 |
| P.MURUGESWARI | 950619104042 |
| S.DIVYA | 950619104501 |



## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

## EINSTEIN COLLEGE OF ENGINEEING,

**TIRUNELVELI-627 012**

## BONAFIDE CERTIFICATE

Certified this Report ”**DEVELOPING A FLIGHT DELAY PREDICTION”**, for the project, is the bonafied work of **S.PRIYADHARSHINI**(950619104050),  **S.SUJITHA**(950619104066), **P.MURUGESWARI**(950619104042), and **S.DIVYA(950619104501)** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here in does not form part of any other thesis or dissertation on the basis of which a degree or award was co-offerred on the earlier occasion on this or any other candidate.

## SIGNATURE SIGNATURE

**Dr. T.SURESH THANKRISHNAN Dr. T.SURESH THANKRISHNAN**

**HEAD OF THE DEPARTMENT HEAD OF THE DEPARTMENT**

**COMPUTER SCIENCE AND ENGINEERING COMPUTER SCIENCE AND ENGINEERING EINSTEIN COLLEGE OF ENGINEEING EINSTEIN COLLEGE OF ENGINEERING TIRUNELVELI-12 TIRUNELVELI-12**

**ACKNOWLEDGEMENT**

We have successfully completed the project with blessings showered onus by god, the almighty, A project of this nature needs co-opreation and support from many for successful completion.

We express our heartfelt thanks to **Mr.A.MATHIVANAN, BE., M.sc.,(Agri),** Managing Trustee of Einstein college of Engineering, Tirunelveli, for his mortal support and device.

Our thanks to **Prof.A.AMUTHAVANAN, BE., M.S (USA).B.L,**Chairman of our college for making necessary arrangements to do this project.

Our heartly thanks to **Prof.EZHILVANAN, MBA.,** Secretary of our college for making necessary arrangements to do this project.

We wish to express our gratitude to **Dr.VELAYUTHAM,M.E,Ph.D.,FIE.** Principal for the support he provided us to carry out this project successfully.

We are very much thankful to **Dr. T.SURESH THANKRISHNAN**, Head of the Department , Electronics and communication Engineering who is always a constant of inspiring us.

We are extend our sincere thanks to all teaching and non-teaching staff members and our family members, friends for their help in completing this project.

**TABLE OF CONTENTS**

**CHAPTER NO. TITLE PAGE NO.**

1. **INTRODUCTION** 06
   1. Project Overview 07
   2. Purpose 07
2. **LITERATURE SURVEY** 11
   1. Existing Problem
   2. References 13
   3. Problem Statement Definition
3. **IDEATION & PROPOSED SOLUTION** 
   1. Empathy Map Canvas
   2. Ideation & Brainstorming
   3. Proposed Solution
   4. Problem Solution Fit
4. **REQUIREMENT ANALYSIS** 
   1. Functional Requirements
   2. Non – Functional Requirements
5. **PROJECT DESIGN**
   1. Data Flow Diagrams
   2. Solution & Technical Architecture
   3. User Stories
6. **PROJECT PLANNING & SCHEDULING**
   1. Sprint Planning & Estimation
   2. Sprint Delivery Schedule
   3. Reports From JIRA
7. **CODING & SOLUTIONING** 
   1. Feature 1
   2. Feature 2
   3. Database Schema
8. **TESTING** 
   1. Test Cases
   2. User Acceptance Testing
9. **RESULTS** 
   1. Performance Metrics
10. **ADVANTAGES & DISADVANTAGES**
11. **CONCLUSION**
12. **FUTURE SCOPE**
13. **APPENDIX**

**CHAPTER 1**

**INTRODUCTION**

In the present world, the major components of any transportation system include passenger airline, cargo airline, and air traffic control system. With the passage of time, nations around the world have tried to evolve numerous techniques of improving the airline transportation system. This has brought drastic change in the airline operations. Flight delays occasionally cause inconvenience to the modern passengers [1]. Every year approximately 20% of airline flights are canceled or delayed, costing passengers more than 20 billion dollars in money and their time. Average aircraft delay is regularly referred to as an indication of airport capacity. Flight delay is a prevailing problem in this world. It's very tough to explain the reason for a delay. A few factors responsible for the flight delays like runway construction to excessive traffic are rare, but bad weather seems to be a common cause. Some flights are delayed because of the reactionary delays, due to the late arrival of the previous flight. It hurts airports, airlines, and affects a company's marketing strategies as companies rely on customer loyalty to support their frequent flying programs.

* 1. **PROJECT OVERVIEW**

Commercial aviation is a complex distributed transportation system. It deals with valuable resources, demand fluctuations, and a sophisticated origin-destination matrix that need orchestration to provide smooth and safety operations. Furthermore, individual passenger follows her itineraries while airlines plan various schedules for aircrafts, pilots and flight attendants. Figure 1 illustrates a typical operation of a commercial flight. Stages can take place at terminal boundaries, airports, runways, and airspace, being susceptible to different kinds of delays. Some examples include mechanical problems, weather conditions, ground delays, air traffic control, runway queues and capacity constraints [103, 63, 3]

This scheme is repeated several times throughout the day for each flight in the system. Pilots, flight attendants and aircrafts may have different schedules due to legal rests, duties, and maintenance plans for airplanes. So, any disruption in the system can impact the subsequent flights of the same airline [2]. Moreover, disturbances may cause congestion at airspace or other airports, creating queues and delaying some flights from other carriers [106, 123]. In this way, the prediction of flight delays is an essential subject for airlines, airports, Air Navigation Service Providers (ANSP), and network managers, like FAA [52] and Eurocontrol [46]. The flight delay prediction problem can be treated by different points of view: (i) delay propagation, (ii) root delay and cancellation. In delay propagation, one study how delay propagates through the network of the transportation system. On the other hand, considering that new problems may happen eventually, it is also important to predict further delays and understand their causes. Such occurrences, in this paper, are named as a root delay problem. Finally, under specific situations, delays can lead to cancellations, forcing airlines and passengers to reschedule their itineraries. So, researchers focused on cancellation analysis try to figure out which conditions lead to cancellations. Moreover, it explores the airlines’ decision-making process for choosing the flights to be canceled.

* 1. **PURPOSE**

Average aircraft delay is regularly referred to as an indication of airport capacity. Flight delay is a prevailing problem in this world. It's very tough to explain the reason for a delay. A few factors responsible for the flight delays like runway construction to excessive traffic are rare, but bad weather seems to be a common cause. Some flights are delayed because of the reactionary delays, due to the late arrival of the previous flight. It hurts airports, airlines, and affects a company's marketing strategies as companies rely on customer loyalty to support their frequent flying programs.

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 EXISTING PROBLEM**

Problem is the core feature in domain taxonomy. As seen in Section 2, there are three major concerns regarding the flight delay prediction problem: delay propagation, root delay and cancellation. Depending on the emphasis of the research, authors select one of these lines to develop their models..

* 1. **REFERENCES**

[1] K. F. Abdelghany, S. S. Shah, S. Raina, and A. F. Abdelghany. A model for projecting flight delays during irregular operation conditions. Journal of Air Transport Management, 10(6):385–394, Nov. 2004. ISSN 0969- 6997.

[2] S. AhmadBeygi, A. Cohn, Y. Guan, and P. Belobaba. Analysis of the potential for delay propagation in passenger airline networks. Journal of Air Transport Management, 14(5):221–236, Sept. 2008. ISSN 0969-6997

[3] S. Ahmadbeygi, A. Cohn, and M. Lapp. Decreasing airline delay propagation by re-allocating scheduled slack. IIE Transactions (Institute of Industrial Engineers), 42(7):478–489, 2010.

[4] C. Ariyawansa and A. Aponso. Review on state of art data mining and machine learning techniques for intelligent Airport systems. In Proceedings of 2016 International Conference on Information Management, ICIM 2016, pages 134–138, 2016.

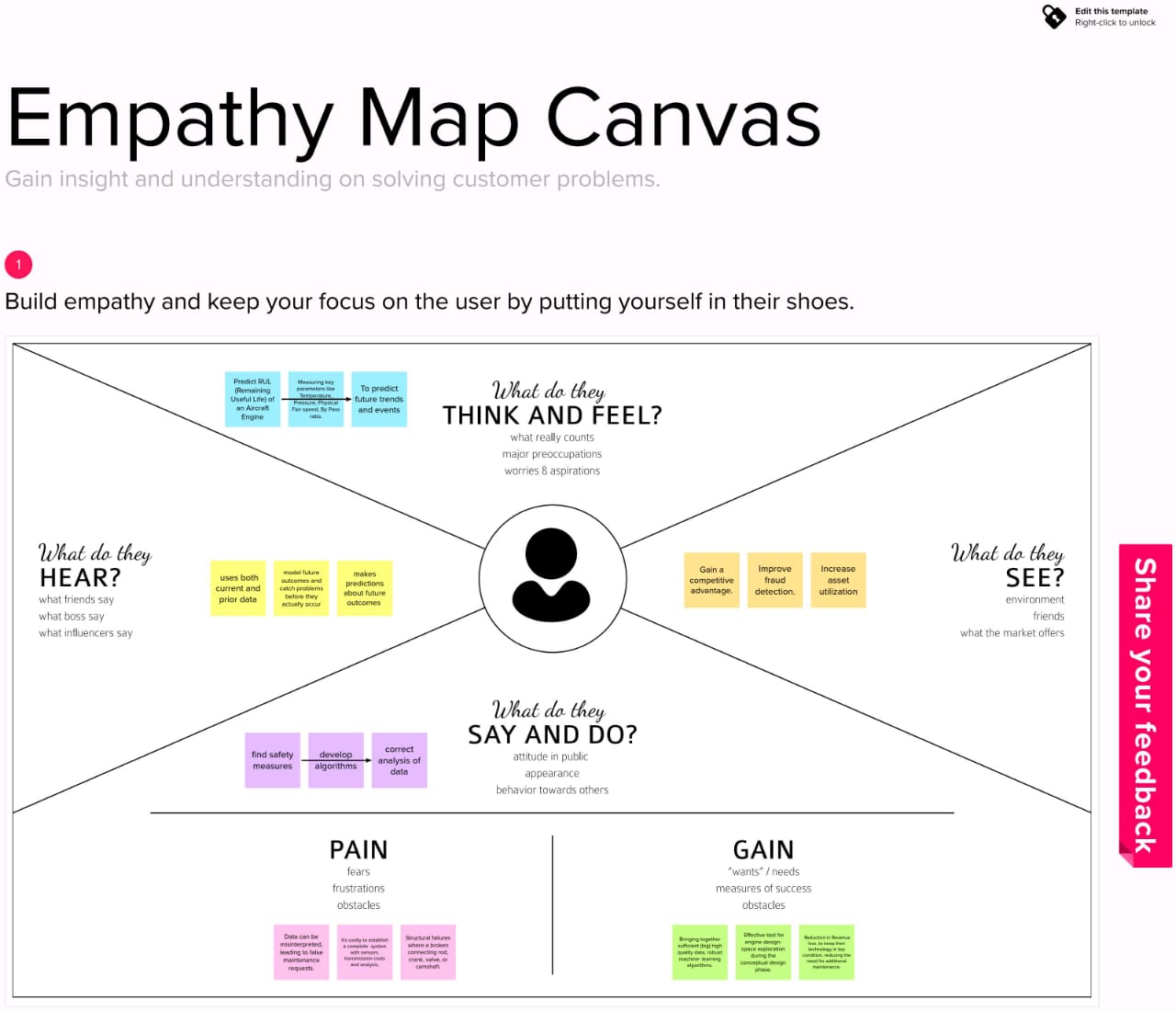
* 1. **PROBLEM STATEMENT DEFINITION**

My case study was about LaGuardia Airport in New York, Logan International Airport in Boston, San Francisco International Airport in San Francisco, and O’Hare International Airport in Chicago, which are four major airports in the United States of America. But we focused the idea and research on LaGuardia International Airport. Compared with the data produced by all airports in USA, the data which we gathered was very limited, but it gave us a great direction on how weather plays a part in flight delays. In this project, the goal is to use exploratory analysis and to build machine learning models to predict airline departure and arrival delays.

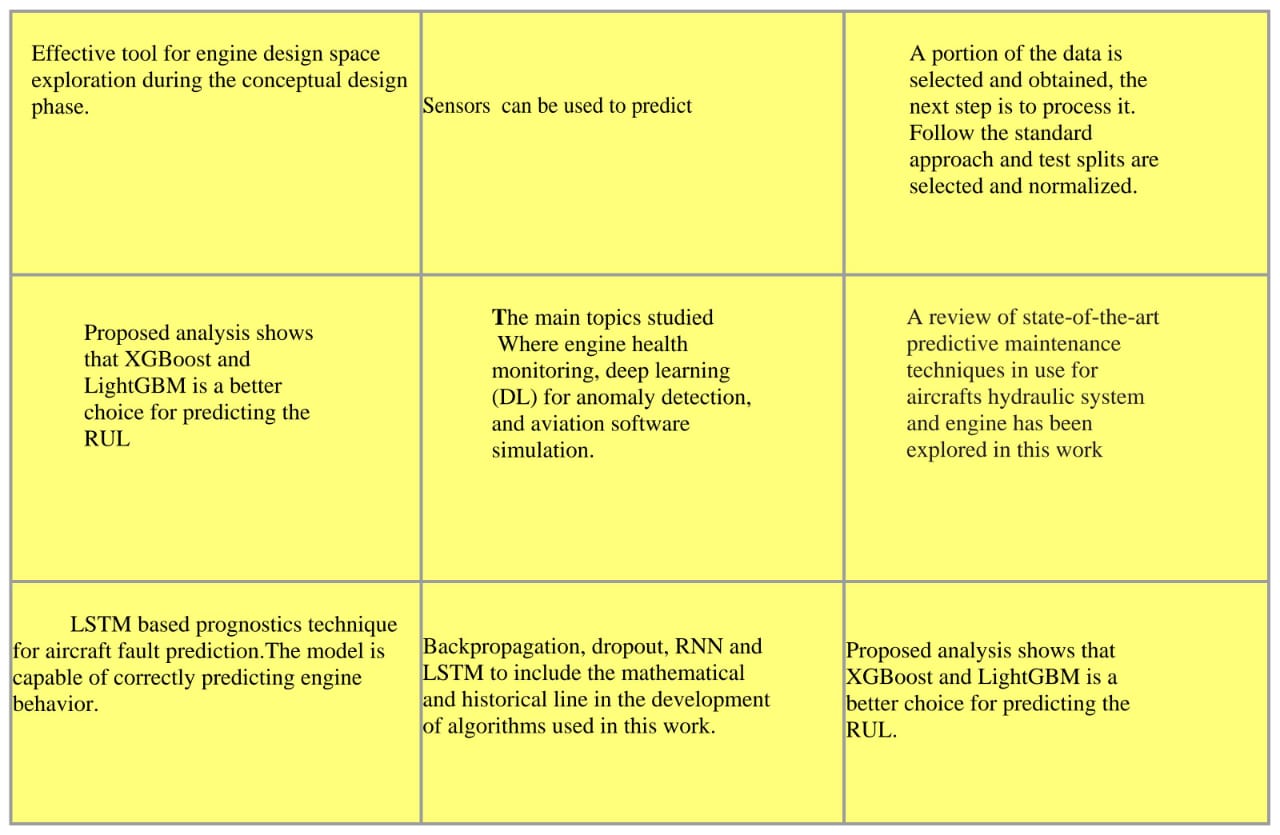
**CHAPTER 3**

**IDEATION & PROPOSED SOLUTION**

**3.1 EMPATHY MAP CANVAS**

****

**3.2 IDEATION & BRAINSTORMING**

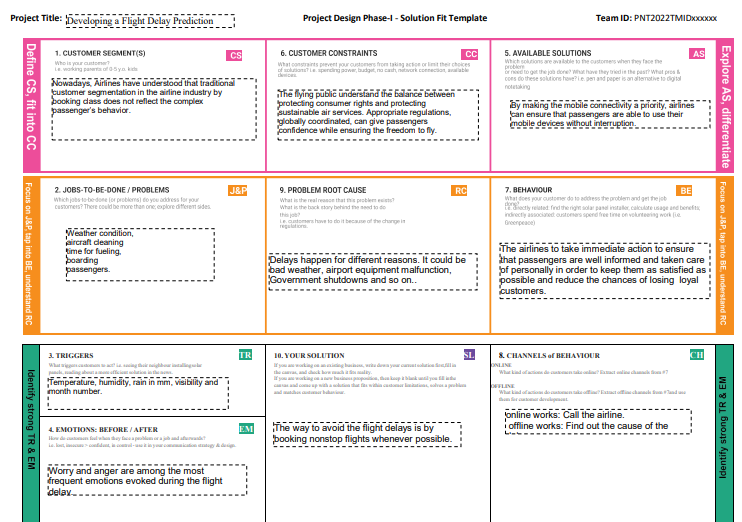
****

* The predictive maintenance strategies are based on real time data for diagnosis of impending failures and prognosis of machine health.
* The model is capable of correctly predicting engine behavior.
* This also presents an LSTM based prognostics technique for aircraft fault prediction.
* It can be an effective tool for engine design space exploration during the conceptual design phase.
* This model aims to predict the fuel mass flow rate having as an input

**3.3 PROPOSED SOLUTION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No.** | **Parameter** | **Description** | | |
| 1. | Problem Statement (Problem to be solved) | Engine failure is highly risky and needs a lot of time for repair. Unexpected failure leads to loss of money and time. Predicting the failure prior will save time, effort, money and sometimes even lives. The failure can be detected by installing the sensors and keeping a track of the values. The failure detection and predictive maintenance can be for any device, out of which we will be dealing with the engine failure for a threshold number of days. . | | |
| 2. | Idea / Solution description | The failure can be detected by installing the sensors and keeping a track of the values. The failure detection and predictive maintenance can be for any device, out of which we will be dealing with the engine | | |
|  | failure for a threshold number of days.Based on machine learning-applied data science it |  |
|  |  |  | aims to predict the failure of an engine.th sustainability of waste services. | e |
| 3. | Novelty / Uniqueness |  | Predicting the failure prior will save time, effort, money and sometimes even |  |
| lives.Through its predictive analysis techn iques we can easily detect the failures and take appropriate measures. |
| 4. | Social Impact / Customer  Satisfaction | Using IoT and smart sensors, waste management companies can increase | | |
| 5. | Business Model (Revenue Model) | Machine Learning generates revenue through the provision of various aircraft management and other services and provide solutions to residential, commercial, industrial, and all other aircraft managements. The Company derives its revenue in the form of various fees associated with its service offerings. | | |
| 5. | Scalability of the Solution | It is found that scaling down of engines is detrimental to SFC and fuel burn, mainly due to the Reynolds number effect. The more scaling done, the more prominent the effect. It is determined that new technology such as higher TIT, OPR and turbomachinery [eta]poly’s for aircraft engines enable the operation of larger bypass ratios, which is the most influential parameter to SFC and fuel bum. The increase of bypass ratio up to a value of 8 is found to be effective for such improvement. SFC decrease from the current to mid-term model is found to be ~20% and ~9% from mid-term to far-term. Range and endurance improvements are found to be ~30% and ~10% respectively for the mission examined. Finally, the mid-term engine model has performance comparable to that of a current, larger state-of-the-art engine, thus suggesting that improvement in small gas turbine technology in the next 10 years will make the application of commodity thrust or distributed propulsion an attractive option for future aircraft. | | |

**3.4 PROBLEM SOLUTION FIT**

****

**CHAPTER 4**

**REQUIREMENT ANALYSIS**

**4.1.Functional Requirements**

Following are the functional requirements of the proposed solution.

|  |  |  |
| --- | --- | --- |
| **FR No.** | **Functional Requirement (Epic)** | **Sub Requirement (Story / Sub-Task)** |
| FR-1 | User Registration | Registration through Form  Registration through Gmail  Registration through LinkedIN |
| FR-2 | User Confirmation | Confirmation via Email Confirmation via OTP |
| FR-3 | User dashboard | Can access the dashboard |

**4.2.Non-functional Requirements:**

Following are the non-functional requirements of the proposed solution.

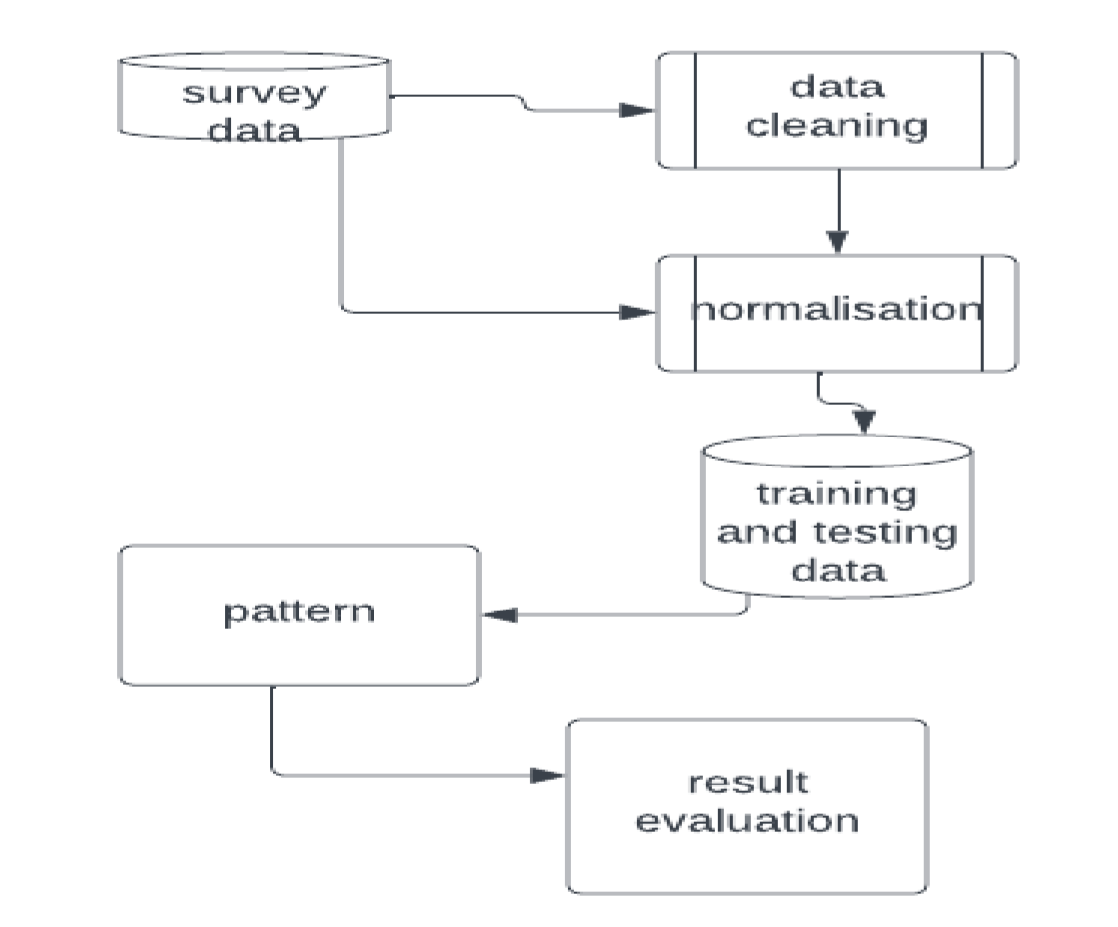
|  |  |  |
| --- | --- | --- |
| **FR No.** | **Non-Functional Requirement** | **Description** |
| NFR-1 | **Usability** | For detecting the faults occurance in aircraft engine |
| NFR-2 | **Security** | Manage the safety and security of passengers life |
| NFR-3 | **Reliability** | The system is more reliable because of it is system adequacy and system security. |
| NFR-4 | **Performance** | High performance because of the system efficiency. |
| NFR-5 | **Availability** | High availability because it allows continuous functioning. |
| NFR-6 | **Scalability** | More number of aircrafts can be secured |

**CHAPTER 5**

**PROJECT DESIGN**

**5.1.Data flowDiagrams**

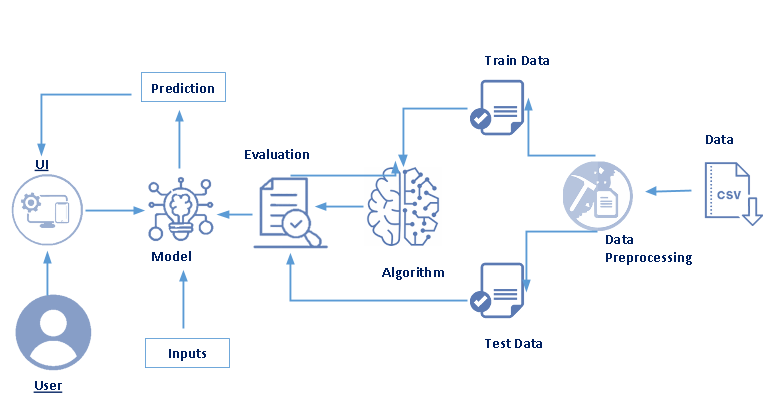
A Data Flow Diagram (DFD) is a traditional visual representation of the information flows within a system. A neat and clear DFD can depict the right amount of the system requirement graphically. It shows how data enters and leaves the system, what changes the information, and where data is stored.



.

**5.2.Solution and Technical Architecture**

# Predictive Analysis for aircraft engine using machine learning



**5.3.User Stories**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **User Type** | **Functional**  **Requirement (Epic)** | **User Story**  **Number** | **User Story / Task** | **Acceptance criteria** | **Priority** | **Release** |
| Customer ( user) | Registration | USN-1 | As a user, I can register for the application by entering my email, password, and confirming my password. | I can access my account / dashboard | High | Sprint-1 |
|  |  | USN-2 | As a user, I will receive confirmation email once I have registered for the application | I can receive confirmation email & click confirm | High | Sprint-1 |
|  |  | USN-3 | As a user, I can register for the application through Facebook | I can register & access the dashboard Facebook Login | Low | Sprint-2 |
|  |  | USN-4 | As a user, I can register for the application through Gmail |  | Medium | Sprint-1 |
|  | Login | USN-5 | As a user, I can log into the application by entering email & password |  | High | Sprint-1 |
| Administrator | Maintain data |  | Have to calculate the data according to ddefault values |  | High |  |

**CHAPTER 6**

**Project Planning and Scheduling**

**6.1.Sprint Planning &Estimation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sprint** | **Functional**  **Requirement(Epic)** | **User Story**  **Number** | **User Story/Task** | **Story point** | **Priority** | **Team Members** |
| Sprint-1 | Registration | USN-2 | As a user, I can register for the application by entering my email, password, and confirming my password. | 2 | High | PRIYADHARSHINI .S |
| Sprint-1 |  | USN-2 | As a user, I will receive confirmation email once I have register for the application | 1 | High | SUJITHA S |
| Sprint-2 |  | USN-3 | As a user, I can register for the application through Facebook | 2 | Low | MURUGESWARI P |
| Sprint-1 |  | USN-4 | As a user, I can register for the application through Gmail | 2 | Medium | DIVYA S |
| Sprint-1 | Login | USN-5 | As a user, I can log into the application by entering email & password | 1 | High | PRIYADHARSHINI S |

**6.2.Sprint Delivery Schedule**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sprint** | **Total Story point** | **Duration** | **Sprint Start Date** | **Sprint End Date (Planned)** | **Story Point Completed**  **(as on Planned End**  **Date)** | **Sprint Release Date (Actual)** |
| Sprint-1 | 20 | 6 Days | 26 Oct 2022 | 31 Oct 2022 | 20 | 31 Oct 2022 |
| Sprint-2 | 20 | 6 Days | 1 Nov 2022 | 6 Nov 2022 | 20 | 6 Nov 2022 |
| Sprint-3 | 20 | 6 Days | 7 Nov 2022 | 12 Nov 2022 | 20 | 12 Nov 2022 |
| Sprint-4 | 20 | 6 Days | 13 Nov 2022 | 18 Nov 2022 | 20 | 18 Nov 2022 |

**CHAPTER 7**

**Coding &Solutioning**

**CHAPTER 8**

**TESTING**

**8.1.Test Cases**

|  |  |  |
| --- | --- | --- |
| Test case ID | Feature Name | Actual Result |
| Test Case\_1 | Set1 | Above the range ,  predicted to be fault |
| Test Case\_2 | Set2 | Below the range ,  Predicted to be correct range |
| Test Case\_3 | Set 3 | At the equal range ,  Predicted to be in dangerous condition |

**8.2.User Acceptance Testing**

**1.Purpose of Document**

The purpose of this document is to briefly explain the flight delay prediction And delay of the flights using Machine Learning

**2.Defect Analysis**

In delay propagation, the primary objective is to understand how delay propagates through airlines and airports based on the assumption that an initial delay has already occurred in the transportation system. A particular scenario happens when delays are spread to other flights of the same airline as chain reactions [24, 16, 2, 118]. Under this situations, it is important to measure how stable and reliable carriers can be to recover from delay propagation [119, 41]. Also, a delay may continue to propagate due to the scheduling of critical resources or retentions in other airports [59]. When scheduled time for take-off or landing is not fulfilled, flights need new slots that may be unavailable. In this scenario, it is important to understand the effects that a root delay in flight may produce to both departure and arrival airports [123, 100, 61]. Such phenomenon may increase the number of flights at some period, generating capacity problems and queues.

**CHAPTER 9**

**RESULTS**

Gaussian Naive Bayes Classifier Gaussian Naive Bayes Classifier was used and an accuracy 0.820 was obtained. This was a considerably good result to start with. Logistic Regression Logistic Regression with default parameters was also tried on the given data and an accuracy of 0.864 was achieved. Decision Tree Decision Tree with default parameters was tried on the given data and an accuracy of 0.806 was achieved. XGBoost Classifier The XGBoost Classifier gave a base accuracy of 0.881. It was further tuned by testing different values for Hyperparameters class weight, learning rate, max depth and n estimators using Grid-Search CV. Figure 4. Tuning XGB Classifier The best accuracy obtained after tuning was 0.884. Optimal Hyper-Parameters • class weight: None • learning rate: 0.1 • max depth: 7 • n estimators: 300

**10.ADVANTAGES**

* Reduction or near elimination of unscheduled equipment downtime caused by equipment or system failure;
* Increased labor utilization
* Increased production capacity
* Reduced maintenance costs
* Increased equipment lifespan
* Maintenance frequency is minimal and reliability is as high as possible
* Eliminating unnecessary costs.

**DISADVANTAGES**

* Required correct datasets
* System will provide inaccurate results if data entered incorrectly
* Datasets varies according to the aircrafts

**11.CONCLUSION**

Flight delays are an important subject in the literature due to their economic and environmental impacts. They may increase costs to customers and operational costs to airlines. Apart from outcomes directly related to passengers, delay prediction is crucial during the decision-making process for every player in the air transportation system. In this context, researchers created flight delay models for delay prediction over the last years, and this work contributes with an analysis of these models from a Data Science perspective. We developed a taxonomy scheme and classified models in respect of detailed components.

.

**12.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 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 self- organization 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 used in cluster analysis. Neural Network offers distributed computer architecture with important learning abilities to represent nonlinear relationships.

**13.APPENDIX :**

**import pandas as pd**

**import numpy as np**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**flights=pd.read\_csv('/home/vaishali/projects/python\_proj/Flight/flights.csv')**

**flights=flights.sample(n=100000)**

**flights.head()**

**flights.shape**

**flights.isnull().values.any()**

**flights.isnull().sum()**

**sns.countplot(x='CANCELLATION\_REASON',data=flights)**

**sns.countplot(x="MONTH",hue="CANCELLATION\_REASON",data=flights)**

**plt.figure(figsize=(10, 10))**

**axis = sns.countplot(x=flights['ORIGIN\_AIRPORT'], data =flights, order=flights['ORIGIN\_AIRPORT'].value\_counts().iloc[:20].index)**

**axis.set\_xticklabels(axis.get\_xticklabels(), rotation=90, ha="right")**

**plt.tight\_layout()**

**plt.show()**

**axis = plt.subplots(figsize=(10,14))**

**Name = flights["AIRLINE"].unique()**

**size = flights["AIRLINE"].value\_counts()**

**plt.pie(size,labels=Name,autopct='%5.0f%%')**

**plt.show()**

**axis = plt.subplots(figsize=(20,14))**

**sns.heatmap(flights.corr(),annot = True)**

**plt.show()**

**corr=flights.corr()**

**corr**

**variables\_to\_remove=["YEAR","FLIGHT\_NUMBER","TAIL\_NUMBER","DEPARTURE\_TIME","TAXI\_OUT","WHEELS\_OFF","ELAPSED\_TIME","AIR\_TIME","WHEELS\_ON","TAXI\_IN","ARRIVAL\_TIME","DIVERTED","CANCELLED","CANCELLATION\_REASON","AIR\_SYSTEM\_DELAY", "SECURITY\_DELAY","AIRLINE\_DELAY","LATE\_AIRCRAFT\_DELAY","WEATHER\_DELAY","SCHEDULED\_TIME","SCHEDULED\_ARRIVAL"]**

**flights.drop(variables\_to\_remove,axis=1,inplace= True)**

**flights.columns**

**airport = pd.read\_csv('/home/vaishali/projects/python\_proj/Flight/airports.csv')**

**airport**

**flights.loc[~flights.ORIGIN\_AIRPORT.isin(airport.IATA\_CODE.values),'ORIGIN\_AIRPORT']='OTHER'**

**flights.loc[~flights.DESTINATION\_AIRPORT.isin(airport.IATA\_CODE.values),'DESTINATION\_AIRPORT']='OTHER'**

**flights**

**print(flights.ORIGIN\_AIRPORT.nunique())**

**print(flights.DESTINATION\_AIRPORT.nunique())**

**print(flights.AIRLINE.nunique())**

**flights=flights.dropna()**

**flights**

**flights.shape**

**df=pd.DataFrame(flights)**

**df['DAY\_OF\_WEEK']= df['DAY\_OF\_WEEK'].apply(str)**

**df["DAY\_OF\_WEEK"].replace({"1":"SUNDAY", "2": "MONDAY", "3": "TUESDAY", "4":"WEDNESDAY", "5":"THURSDAY", "6":"FRIDAY", "7":"SATURDAY"},inplace=True)**

**flights**

**dums = ['AIRLINE','ORIGIN\_AIRPORT','DESTINATION\_AIRPORT','DAY\_OF\_WEEK']**

**df\_cat=pd.get\_dummies(df[dums],drop\_first=True)**

**df\_cat**

**df\_cat.columns/**

**df.columns**

**flights.columns**