

Flight Delay Prediction Using Machine Learning Model

Project Report

Date	18 November 2022
Team ID	PNT2022TMID53082
Project Name	Developing a Flight Delay Prediction Model using Machine Learning.

CONTENTS

1. INTRODUCTION

1.1 Project Overview

1.2 Purpose

2. LITERATURE SURVEY

2.1 Existing problem

2.2 References

2.3 Problem Statement Definition

3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

3.2 Ideation & Brainstorming

3.3 Proposed Solution

3.4 Problem Solution fit

4. REQUIREMENT ANALYSIS

4.1 Functional requirement

4.2 Non-Functional requirements

5. PROJECT DESIGN

5.1 Data Flow Diagrams

5.2 Solution & Technical Architecture

5.3 User Stories

6. PROJECT PLANNING & SCHEDULING

6.1 Sprint Planning & Estimation

6.2 Sprint Delivery Schedule

6.3 Reports from JIRA

7. CODING & SOLUTIONING (Explain the features added in the project along with code)

7.1 Feature 1

7.2 Feature 2

7.3 Database Schema (if Applicable)

8. TESTING

8.1 Test Cases

8.2 User Acceptance Testing

9. RESULTS

9.1 Performance Metrics

10. ADVANTAGES & DISADVANTAGES

11. CONCLUSION

12. FUTURE SCOPE

13. APPENDIX

Source Code

GitHub & Project Demo Link

1.INTRODUCTION

1.1 Project Overview

Air travel has become widely common and preferred among travelers over the years due to its comfort and the time of travel. This has in a way led to a lot of air traffic and on ground and hence resulting in massive levels of aircraft delays in the air and on ground. The delays are a cause of environmental and economic losses. The proposed system helps to predict flight delay to optimize flight operations and minimize delays in a most accurate manner.

1.2 Purpose

The system aims to help predict flight delays thereby easing the task of a passenger. It prevents a lot of discomfort for a passenger. The flight operations are optimized and more reliable.

14. LITERATURE SURVEY

14.1 Existing problem

Yuemin Tang from University of Southern California proposed a paper with the main goal to compare the performance of machine learning classification algorithms when predicting flight delays. The data set used for analysis contains data about flights leaving from JFK airport between one year from November 2019 to December 2020. In this study, classification models were selected and trained using seven algorithms: Logistic Regression, K-Nearest Neighbor (KNN), Gaussian Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, and Gradient Boosted Tree. The value of each evaluation measure for every algorithm is presented in table as result. The result shows that decision Tree performs well when predicting flight delays in the data set. Other tree-based ensemble classifiers also show good performance. Random Forest and Gradient Boosted Tree have an accuracy of 0.9240 and 0.9334, significantly higher than the rest of the models. The other four base classifiers Logistic Regression, KNN, Gaussian Naïve Bayes, and SVM, are not tree-based and did not show good

performance. The KNN model is the worst performed since its precision and f1-score are the lowest among the seven models.[1] Bhuvan Bhatia used Python based Logistic Regression along with Support Vector Machine to predict flight delays and compared the results with other models such as Random Forest Classifier and derive the best classifier to solve the problem. The dataset focused on LaGuardia International Airport. The result shows that the Random Forest method yields the best performance compared to the SVM model.[2]

14.2 References

[1] Yuemin Tang. 2021. Airline Flight Delay Prediction Using Machine Learning Models. In 2021 5th International Conference on E-Business and Internet (ICEBI 2021), October 15-17, 2021, Singapore, Singapore. ACM, New York, NY, USA, 7 Pages.

<https://doi.org/10.1145/3497701.3497725>

[2] 99257830966401671; <https://hdl.handle.net/10211.3/202926>

14.3 Problem Statement Definition

The proposed system aims to build a machine learning model for flight delay prediction to optimize flight operations and minimize delays in a most accurate manner. Using the machine learning model, we can predict flight arrival delays.

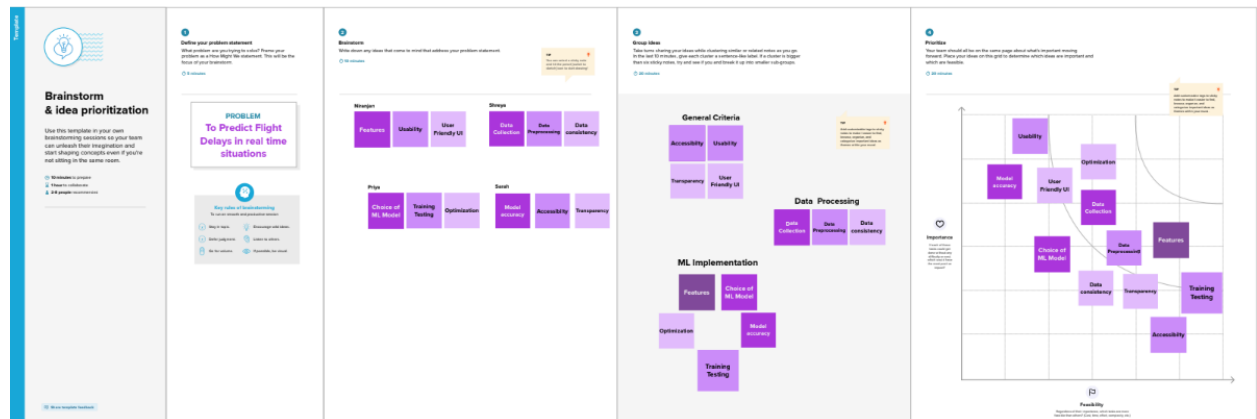
3. IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas

Build empathy and keep your focus on the user by putting yourself in their shoes.



3.2 Ideation & Brainstorming



3.3 Proposed Solution

S.No.	Parameter	Description
-------	-----------	-------------

1.	Problem Statement	<p>The dataset is analysed and preprocessed to remove/replace unnecessary entries using standard techniques. Various classification models have been built to consider factors like arrival date, departure date, flight number, distance etc. Models like regression, KNN, Decision tree etc have been trained and an ensemble of the best performing models is built to further improve accuracy of the predictions. The models have been trained and tested successfully using the given dataset. A UI layer is built using Python Flask application and details like Departure date, arrival date and distance is taken from the user and the data is fed into the final proposed ML model. The predictions are evaluated by the model and is returned to user.</p>
2.	Idea / Solution description	<p>Ensembling of many base classifier models make the final model robust and more accurate. Thus all stakeholders can have their requirements met</p>
3.	Social Impact / Customer Satisfaction	<p>Accurate predictions of flight delays can lead to customer satisfaction, improves business for the airline and also improves productivity of aviation systems. This leads to a positive impact on a country's economy.</p>
4.	Scalability of the Solution	<p>The model can be further scalable by training the proposed model on a larger dataset collected from one or more airports. Also, more powerful tools/computers can be used to predict results more effectively in lesser time.</p>

3.4 Problem Solution fit



4 REQUIREMENT ANALYSIS

4.1 Functional requirement

FR No.	Functional Requirement	Sub Requirement(Story/Sub-Task)
--------	------------------------	---------------------------------

FR-1	User Login	an user/passenger should be able to log in to the system using his/her credentials.
FR-2	The System should contain User details.	The System should contain information about the passenger and his/her light details.
FR-3	The System should allow a passenger to search for his/her flight details .	The results for which will be fetched based on the entered flight details. There should not be any ambiguity in the fetched results.
FR-4	The arrival and departure of the flights should be tracked continuously and updated in the system from time to time.	The System predicts the flight delay using the tracked information .
FR-5	When a flight is delayed , the passenger should be notified about the same if he/she is supposed to be flying in it.	The passenger can view this delay.
FR-6	The system must allow the user to submit queries through a form.	

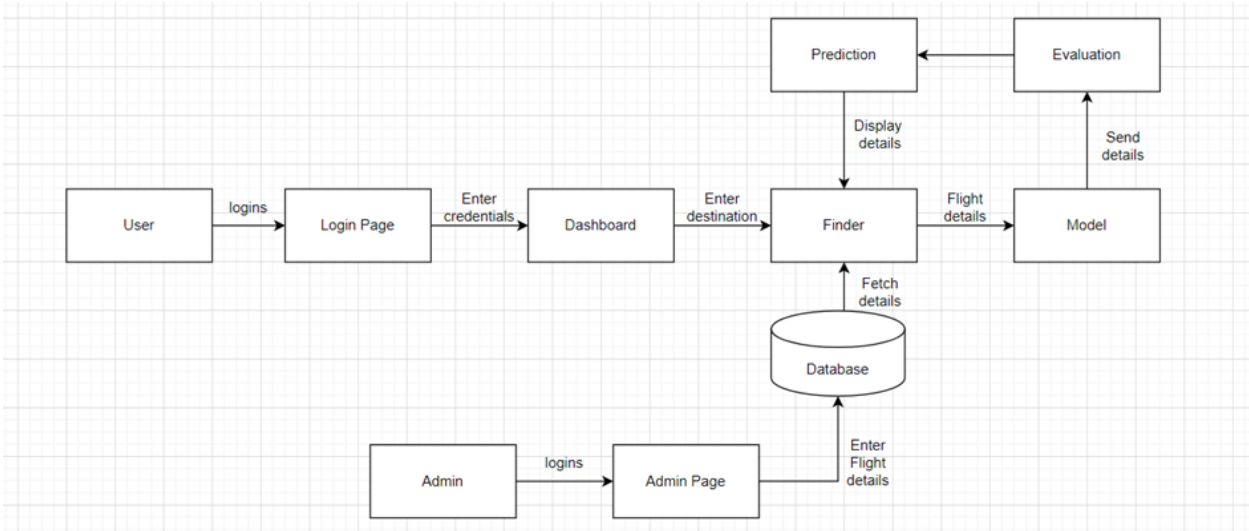
4.2 Non-Functional requirements

NFR No.	Non Functional Requirement	Sub Task/Story
1	Security	The product stores flight timing information and user data which are secure by firewalls and encryption in the backend to maintain data integrity.
2	Capacity	Storage requirements are 15 gb rom and 6-8 gb ram for proper functioning of the system
3	Compatibility	Any python based IDE should suffice the system requirements for compiling and execution of source code. End product is hosted in docker.
4	Reliability and Availability	The system accurately predicts the delay in flight arrival and departure and is immune to crashing as it is hosted in docker and is yet to be scaled.
5	Usability	System is easily understandable in terms of it's

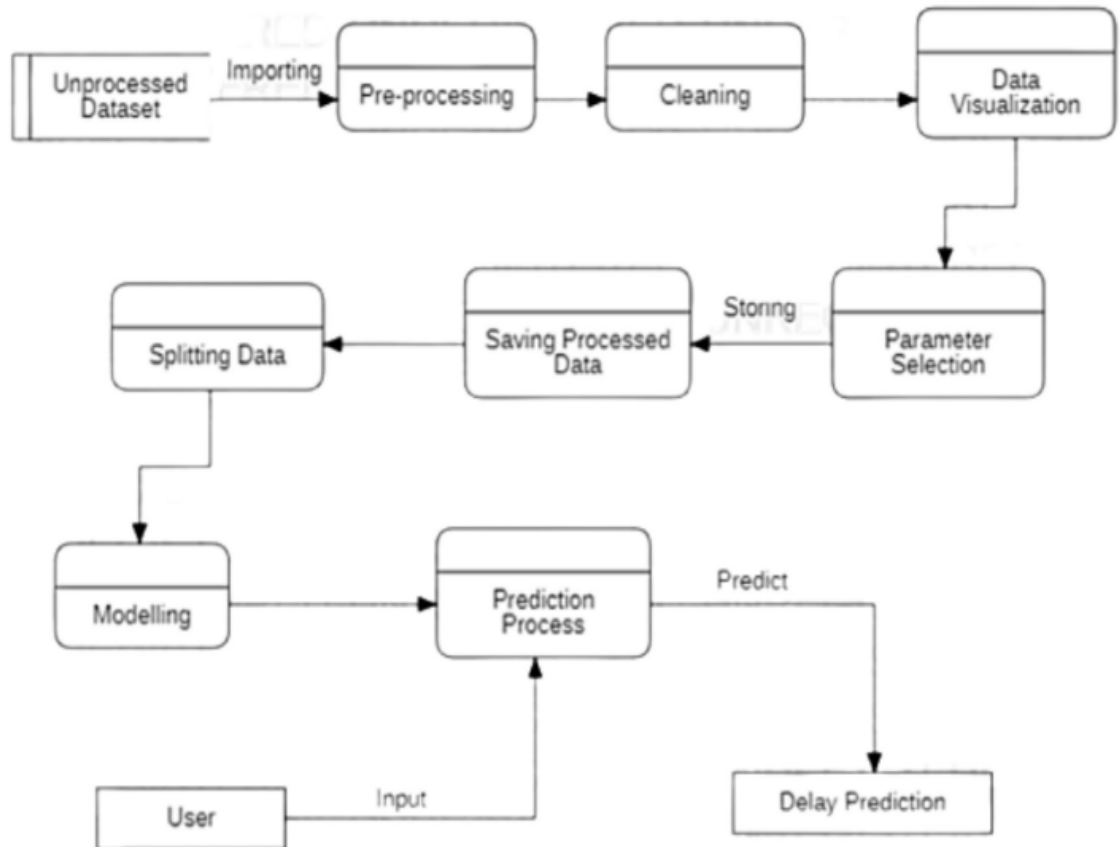
		functionality and User Interface.
--	--	-----------------------------------

5 PROJECT DESIGN

5.1 Data Flow Diagrams



5.2 Solution & Technical Architecture



The deliverable includes architecture diagram and information as per table 1 and 2

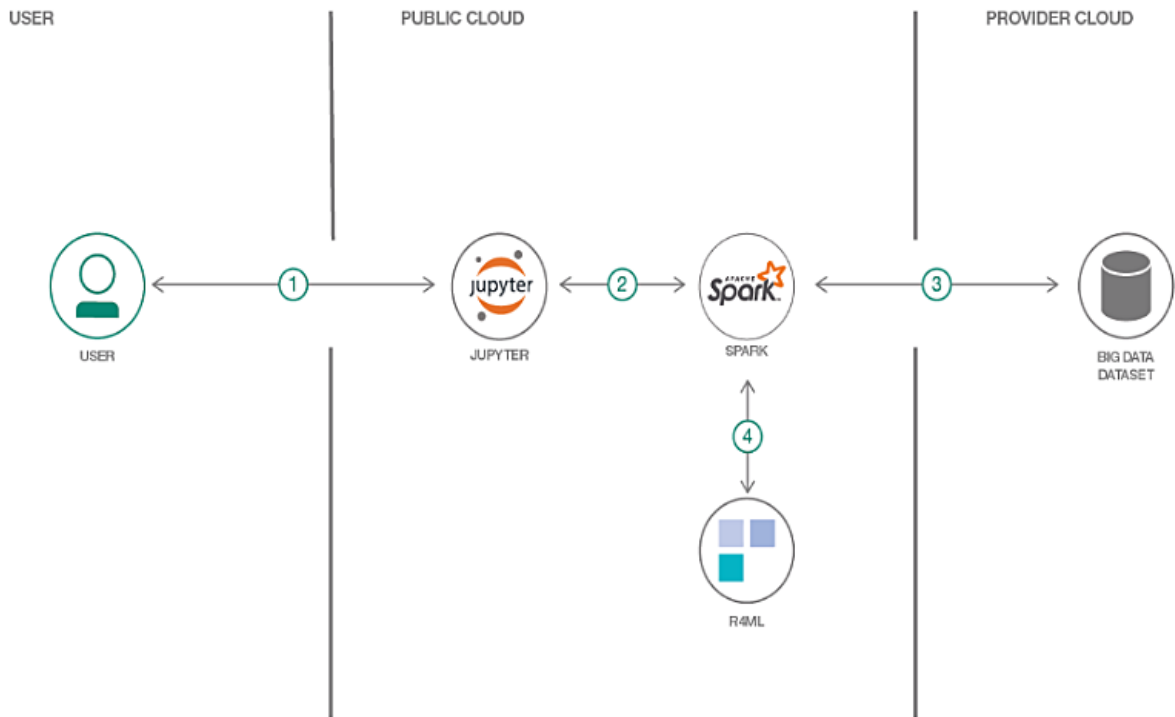


Table 1: Components and Technologies

SNo	Component	Description	Technology
1	User interface	How user interacts with application	HTML, CSS, JS, Flask
2	Application Logic - 1	Logic for process in application	Python
3	Application Logic - 2	Logic for process in application	IBM Watson STT service
4	Data processing	To clean data	Pandas, numpy, matplotlib etc
5	Database	Store data	MySQL
6	File storage	Storing files	IBM Block storage

7	External API-1	External API used in the system	IBM Weather API
8	External API-2	External API used in system	Email API
9	Machine Learning model	Purpose of ML model	Evaluation and prediction models
10	Infrastructure (Server/Cloud)	Application deployment	IBM Cloud

Table 2: Application Characteristics

S No	Characteristics	Description	Technology
1	Open Source Frameworks	Open source frameworks used	Python-Flask
2	Security Implementations	Security/access controls implemented, firewalls used	Encryptions, SHA2
3	Scalable Architecture	Scalability of architecture	Python
4	Avalability	Availability of the application	IBM Cloud
5	Performance	Consideration for the performance of the application	Python, Flask

5.3 User Story

User Type	Functional Requirement (Epic)	User Story Number	User story/Task	Acceptance Criteria	Priority	Release
Customer (Web 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 with 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 app by entering email and password	I can access the dashboard	High	Sprint-1
	Dashboard	USN-6	As a user, I can navigate through different pages using the dashboard	I can various pages	High	Sprint-2
	Search	USN-7	As a user, I can search for flights with destination location.	I can receive information on various flights.	High	Sprint-2
	View	USN-8	As a user, I can view the details of the flights.	I get the information such as flight no, departure and arrival time, etc.	High	Sprint-3
	Receive notifications	USN-9	As a user, I will receive notifications about the flight.	I get frequent updates of the flight's location.	Low	Sprint-3

Admin	GPS	USN-10	As an admin, I need the location of flights	I can track the flights	High	Sprint-4
	Analyse	USN-11	As an admin, I will analyse the given dataset.	I can analyse the dataset.	High	Sprint-2

6.1 Sprint Planning and Estimation

Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data Collection , Cleaning and Preprocessing	USN-1	As a user, I can't interact with anything. Waiting is the user's task.	2	High	Niranjan, Priya
Sprint-1	Model Building	USN-2	As a user, I predict flight delay by using the ML models developed.	1	High	Sarah, Shreya
Sprint-1	Model Evaluation	USN-3	As a user, I can evaluate my models' accuracy.	2	Low	Niranjan
Sprint-2	Model Deployment and integration with Flask.	USN-4	As a user, I can access the models on integration with flask.	2	Medium	Sarah, Shreya
Sprint-2	Sign-up	USN-5	As a user, I can register into the application by entering	1	High	Priya

			email & password and confirming my password.			
Sprint-2	login	USN-6	As a user, I can log into the application by entering email & password	1	high	Shreya
Sprint-2	Dashboard	USN-7	As a user, I can explore the various services in the dashboard.	1	high	Sarah
Sprint-3	Raise query complaint and give feedback	USN-8	As a user , I can raise complaints or provide feedback.	1	low	Niranjan
Sprint 3	Improve Model accuracy.	USN -9	As a user I will get better accurate predictions for my query.	3	high	Shreya,Sarah
Sprint-4	Deployment and Testing	USN-10	As a user , I can access the web-app and use it	11	meduim	Priya,Shreya
Sprint-4	Improvements if any.	USN-11		1	low	Niranjan..

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	20	6 Days	24 Oct 2022	29 Oct 2022	20	31 Oct 2022
Sprint-2	20	6 Days	31 Oct 2022	05 Nov 2022	20	05 Nov 2022
Sprint-3	20	6 Days	07 Nov 2022	12 Nov 2022	20	12 Nov 2022
Sprint-4	20	6 Days	14 Nov 2022	19 Nov 2022	20	19 Nov 2022

14.4 Reports from JIRA

Projects / FlightDelayPrediction

FLIG Sprint 1

To Develop a basic ML model to predict flight delay.

SN

Epic

GROUP 8

TO DO

IN PROGRESS

DONE 4 ISSUES

Download Dataset

FLIG-9

Data Pre-Processing

FLIG-10

Build Model

FLIG-11

Evaluate the predictions

FLIG-12

FLIG Sprint 1

24 Oct – 19 Nov

(4 issues)

000

Complete sprint

To Develop a basic ML model to predict flight delay.

FLIG-9

Download Dataset

DONE

FLIG-10

Data Pre-Processing

DONE

FLIG-11

Build Model

DONE

FLIG-12

Evaluate the predictions

DONE

+ Create issue

▼ FLIG Sprint 2 28 Oct – 5 Nov (3 issues)

To integrate with Flask and build frontend for the System.

000

Complete sprint

...

FLIG-17

Create Database for Users and Integrate

DONE ✓

FLIG-14

Model Integration with flask

DONE ✓

FLIG-15

Build UI for the system

DONE ✓

+ Create issue

Projects / FlightDelayPrediction

FLIG Sprint 2

To integrate with Flask and build frontend for the System.

0 days remaining

Complete sprint

...

Q

SN

Epic ▼

Sprint 1 ▼

Clear filters

GROUP BY

None ▼

Insights

TO DO

IN PROGRESS

DONE 3 OF 7 ISSUES ✓

Create Database for Users and Integrate

FLIG-17

✓

Model Integration with flask

FLIG-14

✓

Build UI for the system

FLIG-15

✓

Projects / FlightDelayPrediction

FLIG Sprint 3

Improve model Accuracy and User Interfaces.

0 days remaining

Complete sprint

...

Q

SN

Epic ▼

Sprint 1 ▼

Clear filters

GROUP BY

None ▼

Insights

TO DO

IN PROGRESS

DONE 2 OF 11 ISSUES ✓

Improve Model Accuracy

FLIG-18

✓

Enhance User Interface

FLIG-19

✓

FLIG Sprint 4

Deploy and test the System

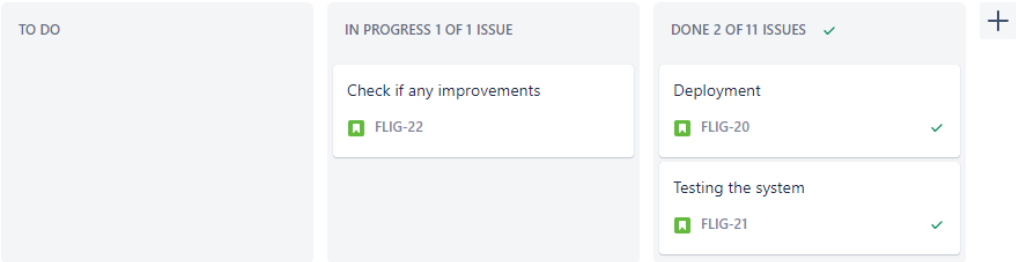
2 days remaining Complete sprint

Epic

Sprint **1**

Clear filters

GROUP BY None Insights



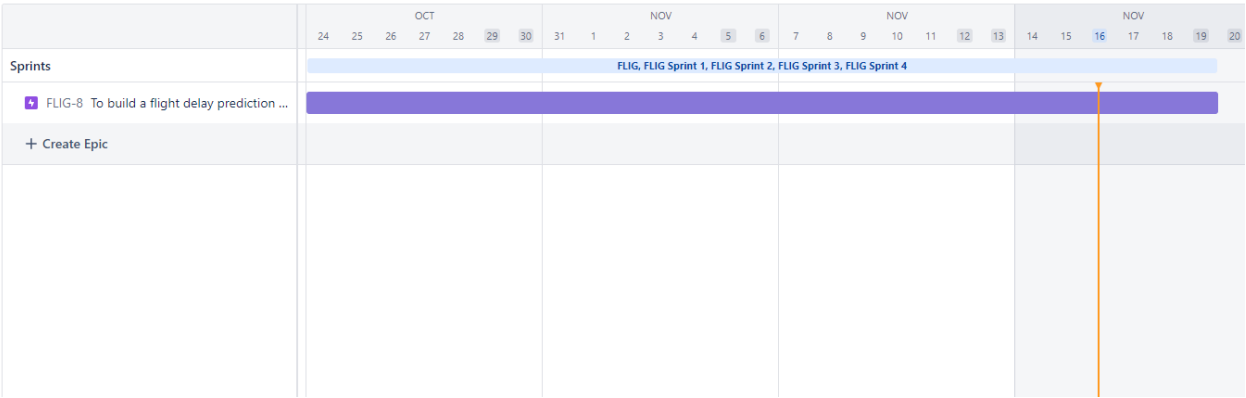
FLIG Sprint 4 15 Nov – 19 Nov (3 issues) 0 0 0 Complete sprint

Deploy and test the System

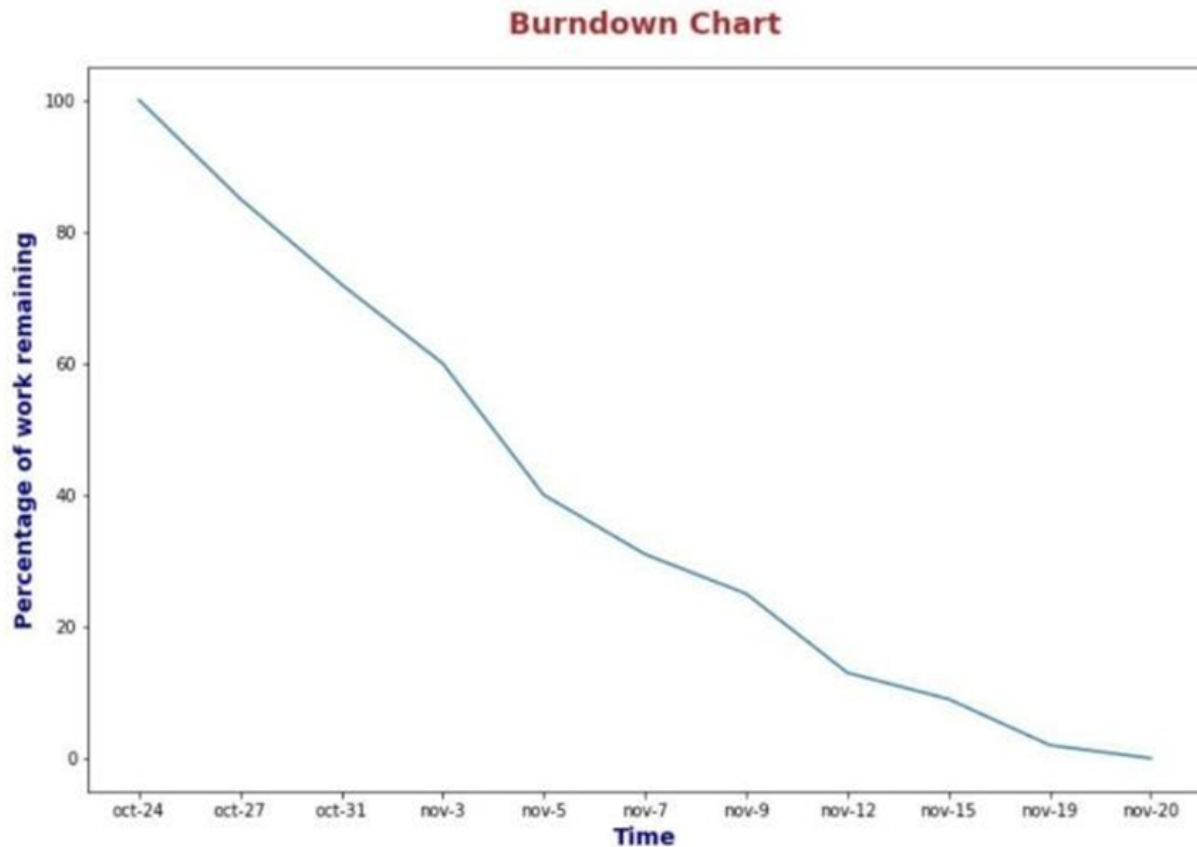
FLIG-20	Deployment	DONE	
FLIG-21	Testing the system	DONE	
FLIG-22	Check if any improvements	IN PROGRESS	

Create issue

RoadMap



Burndown Chart



15. CODING & SOLUTIONING (Explain the features added in the project along with code)

15.1 Feature 1

15.2 Feature 2

15.3 Database Schema (if Applicable)

MySQL was used to save user data like credentials and emailID .

SCHEMA

```
mysql> describe accounts;
```

Field	Type	Null	Key	Default	Extra
id	int	NO	PRI	NULL	auto_increment
name	varchar(50)	NO		NULL	
password	varchar(10)	NO		NULL	
email	varchar(100)	NO		NULL	

```
4 rows in set (0.07 sec)
```

TESTING

15.4 Test Cases

S.NO	Test scenario	User Input	Expected output	actual output	result
01	User Signs Up	User enters correct Username , Password , emailId	User Registered	User Registered	Pass
02	User Sign Up	User enters an invalid email ID	User prompted with a message	User not registered .	Pass
03	User Sign Up	User enters an already registered emailID and Username	User prompted with a message	Message : Already Exists.	Pass
04	User Log In	User enters Username and password.	Logged in	User logged in	pass

[illegible]

15.5 User Acceptance Testing

Acceptance Testing UAT Execution & Report Submission

Date	03 November 2022
Team ID	PNT2022TMID53082
Project Name	Project - Flight Time Delay Prediction

1. Purpose of User Acceptance Testing

The purpose of this document is to briefly explain the test coverage and open issues of the [ProductName] project at the time of the release to User Acceptance Testing (UAT).

2. Defect Analysis

This report shows the number of resolved or closed bugs at each severity level, and how they were resolved

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	10	4	2	3	20
Duplicate	1	0	3	0	4
External	2	3	0	1	6
Fixed	11	2	4	20	37
Not Reproduced	0	0	1	0	1
Skipped	0	0	1	1	2
Won't Fix	0	5	2	1	8
Totals	24	14	13	26	77

3. Test Case Analysis

This report shows the number of test cases that have passed, failed, and untested

Section	Total Cases	Not Tested	Fail	Pass
Print Engine	7	0	0	7
Client Application	51	0	0	51
Security	2	0	0	2
Outsource Shipping	3	0	0	3
Exception Reporting	9	0	0	9
Final Report Output	4	0	0	4
Version Control	2	0	0	2

16. RESULTS

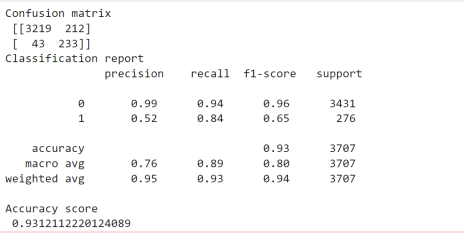
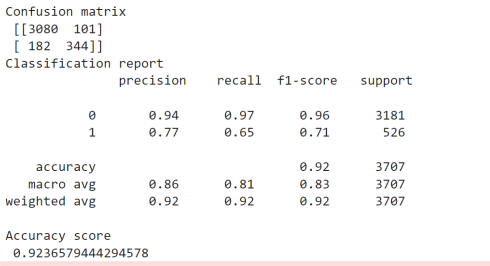
The System developed helps predict flight delays , the proposed model works around an accuracy of 94% in predicting the delay. The user can successfully sign-up into the system using his/her emailId. The user logs into the system and can predict flight delay if any. The user query is fetched and executed with an accuracy of 94%. Hence , the System helps to predict flight delays accurately.

16.1 Performance Metrics

Project Development Phase Model Performance Test

Date	10 November 2022
Team ID	PNT2022TMID53082
Project Name	Project - Flight time delay prediction

Model Performance Testing:

S.No.	Parameter	Values	Screenshot
1.	Accuracy score	Classification Model: KNN	 <pre>Confusion matrix [[3219 212] [43 233]] Classification report precision recall f1-score support 0 0.99 0.94 0.96 3431 1 0.52 0.84 0.65 276 accuracy 0.93 3707 macro avg 0.76 0.89 0.80 3707 weighted avg 0.95 0.93 0.94 3707 Accuracy score 0.9312112220124089</pre>
2.	Accuracy score	Classification Model: Gaussian Naive Bayes	 <pre>Confusion matrix [[3080 101] [182 344]] Classification report precision recall f1-score support 0 0.94 0.97 0.96 3181 1 0.77 0.65 0.71 526 accuracy 0.92 3707 macro avg 0.86 0.81 0.83 3707 weighted avg 0.92 0.92 0.92 3707 Accuracy score 0.9236579444294578</pre>

3.	Accuracy score	Classification Model: Logistic Regression	<pre> Confusion matrix [[3225 182] [37 263]] Classification report precision recall f1-score support 0 0.99 0.95 0.97 3407 1 0.59 0.88 0.71 300 accuracy 0.94 3707 macro avg 0.79 0.91 0.84 3707 weighted avg 0.96 0.94 0.95 3707 Accuracy score 0.9409225789047747 </pre>
----	----------------	--	--

17. ADVANTAGES & DISADVANTAGES

The system eases out the task of a user by predicting flight delays. The discrepancies and problems faced by a user can be avoided. The users can have a better traveling experience with these predictions. The user needn't face anxiety or feel helpless when traveling through flights. The system is user friendly and a person with no experience of softwares can also have a good user experience.

When it comes to the disadvantages involved, the system is only 94% accurate with its predictions. The prediction model is based on static data from a particular time frame and cannot accommodate user queries beyond the stored data, i.e cannot work on real time data.

18. CONCLUSION

The system that has been developed is a user friendly software that is accurately predicting the delay in flight time given the corresponding details taken as input. The overall aim of the project has been achieved and the software system is in line with the objectives set by IBM.

19. FUTURE SCOPE

The scope of the project can be extended to include prediction from real time data and dataset extension.

20. APPENDIX

Source Code - [IBM-EPBL/IBM-Project-27630-1660061379: Developing a Flight Delay Prediction Model using Machine Learning \(github.com\)](https://github.com/IBM-EPBL/IBM-Project-27630-1660061379)

[IBM-EPBL/IBM-Project-27630-1664792961: Developing a Flight Delay Prediction Model using Machine Learning \(github.com\)](#)

Github and Project Demo link - <https://drive.google.com/drive/folders/18hAlhoYr6YNo41X-N2Bj9eVfqVsK4pOx>