PROJECT REPORT

Developing a Flight Delay Prediction Model using Machine Learning

PNT2022TMID35530

AKASHRAM J	**2019115011**
GAYATHRI P	**2019115033**
THARANYAA R	**2019115113**
VIJAY S	**2019115120**

CHAPTER 1

INTRODUCTION

Delay is one of the best-known performance indicators in any transportation system. Civil aviation officials in particular understand delay as the time at which a flight is delayed or rescheduled. Delay can therefore be expressed as the difference between the scheduled flight time and the actual departure or arrival time. National regulators have a number of indicators relating to acceptable levels of flight delays. Flight delays are a major problem associated with air transportation systems.

Analysts and data scientists are immersed in this vast amount of data generated by sensors and IoT, enhancing their computational and data management skills to extract useful information from each data. In this context, the process of understanding domains, managing data, and applying models is called data science, a trend for solving challenging big data-related problems. In this project, extensive data analysis was performed to extract the key attributes/factors responsible for flight delays. In addition, there are other factors that can affect flight delays, such as: These factors, such as climate, natural disasters, pandemics, or technical problems with aircraft, vary from place to place and are not considered in this project as such problems rarely occur.

1.1 PROJECT OVERVIEW

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. Using this algorithm model is built to predict the flight delay.

1.2 PURPOSE

Flight delays result in significant financial and other losses to airlines, airports and passengers. Predictions are important in the decision-making process of all parties in the aviation industry. Therefore, predicting potential delays based on flight characteristics bridges an important information asymmetry between airlines and passengers. The main use cases for this algorithm are to forecast of possible delays on a given day for a given airport and airline.

CHAPTER 2

LITERATURE SURVEY

2.1 EXISTING PROBLEM

Many existing flight delay prediction methods are based on small samples and/or are complex to interpret with little or no opportunity for machine learning deployment. The proposed model gains insight into factors causing flight delays, cancellations and the relationship between departure and arrival delay using exploratory data analysis.

2.2 REFERENCES

[1] Jiang, Yushan, et al. "Applying machine learning to aviation big data for flight delay prediction." 2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). IEEE, 2020.

[2] Liu, Fan, et al. "Generalized flight delay prediction method using gradient boosting decision tree." 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring). IEEE, 2020.

[3] Ding, Yi. "Predicting flight delay based on multiple linear regression." IOP Conference Series: Earth and Environmental Science. Vol. 81. No. 1. IOP Publishing, 2017.

[4] Thiagarajan, Balasubramanian, et al. "A machine learning approach for prediction of on-time performance of flights." 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC). IEEE, 2017.

[5] Kim, Young Jin, et al. "A deep learning approach to flight delay prediction." 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC). IEEE, 2016.

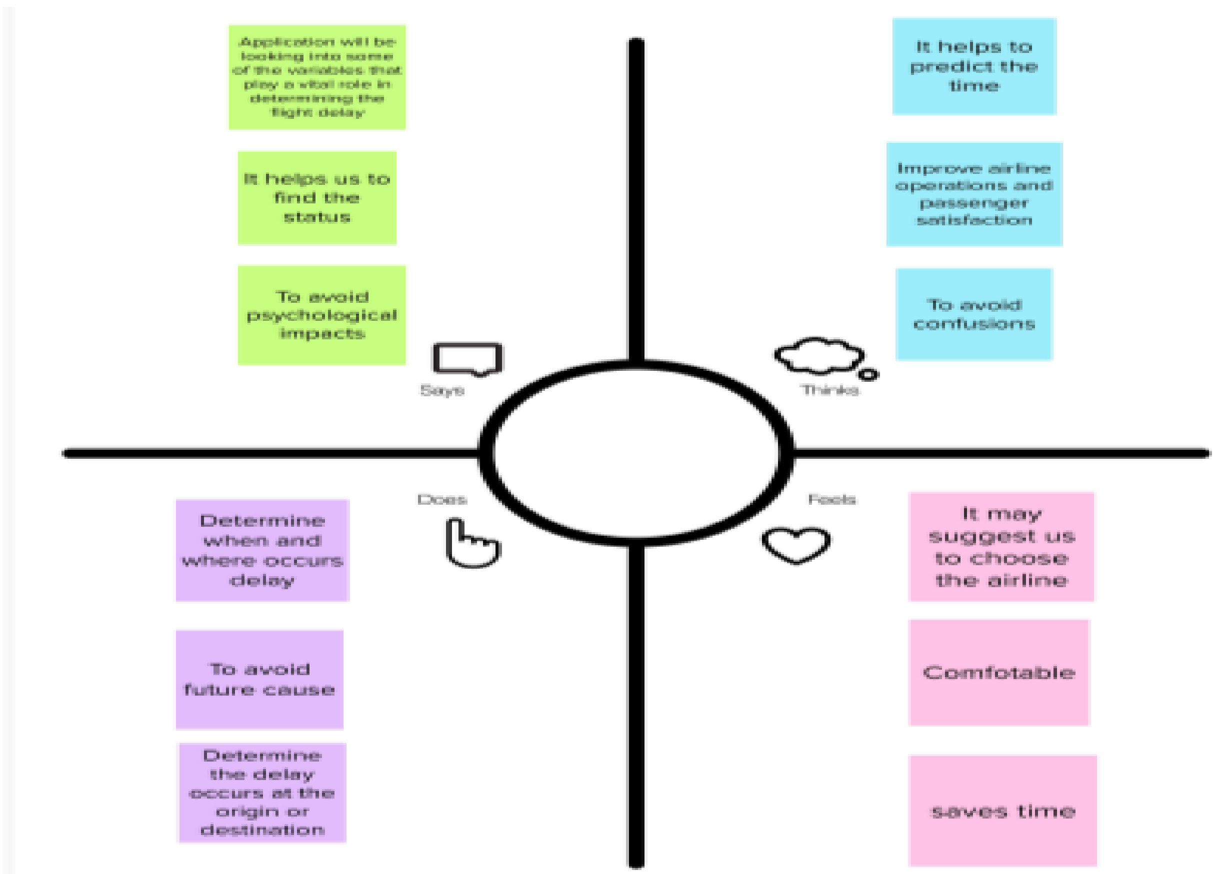
2.3 PROBLEM STATEMENT DEFINITION

Flight delays in air transportation are a major concern that has adverse effects on the economy, the passengers, and the aviation industry. This matter critically requires an accurate estimation for future flight delays that can be implemented to improve airport operations and customer satisfaction. Having said that, a massive volume of data and an extreme number of parameters have restricted the way to build an accurate model. Many existing flight delay prediction methods are based on small samples and/or are complex to interpret with little or no opportunity for machine learning deployment.

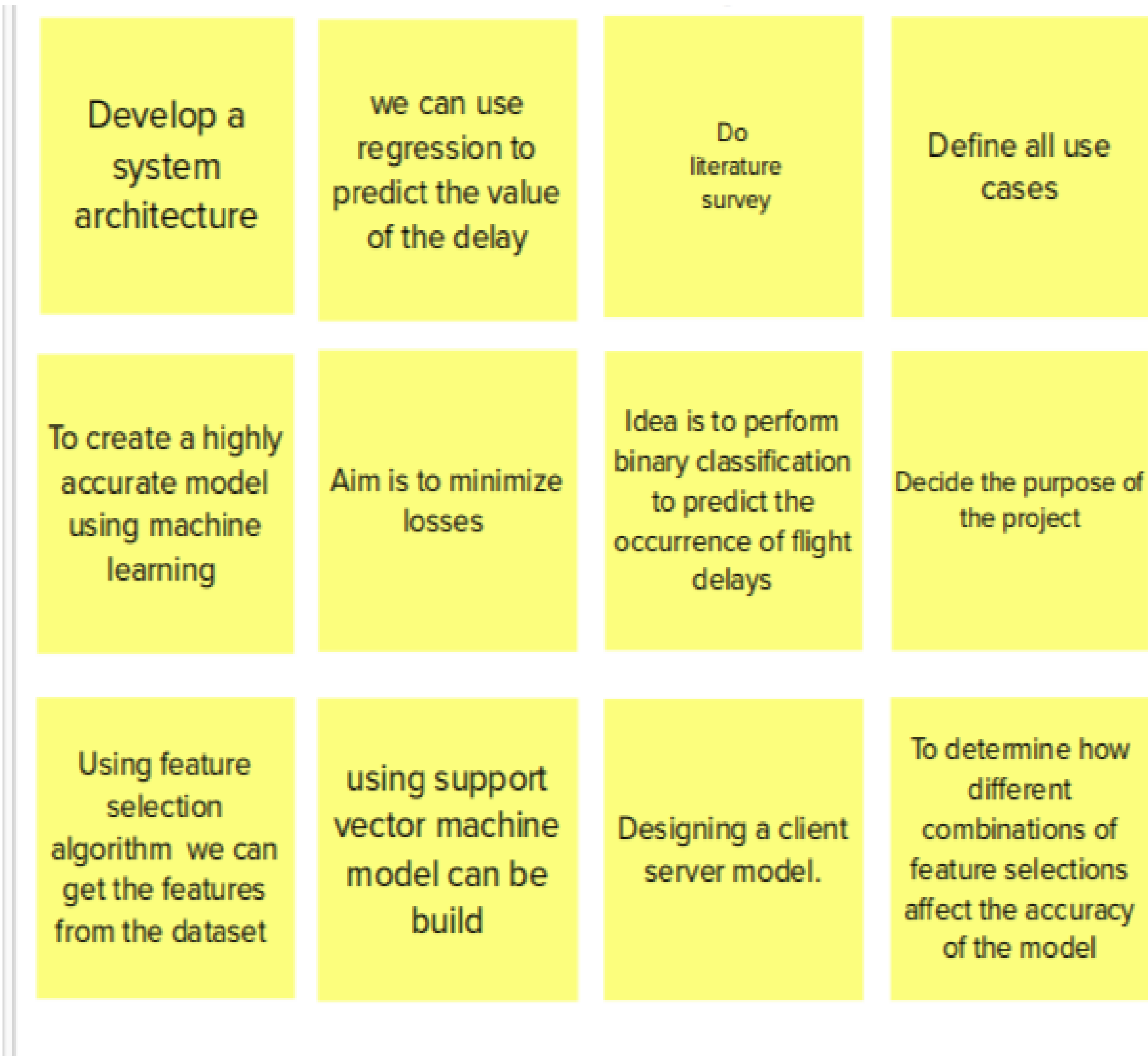
CHAPTER 3

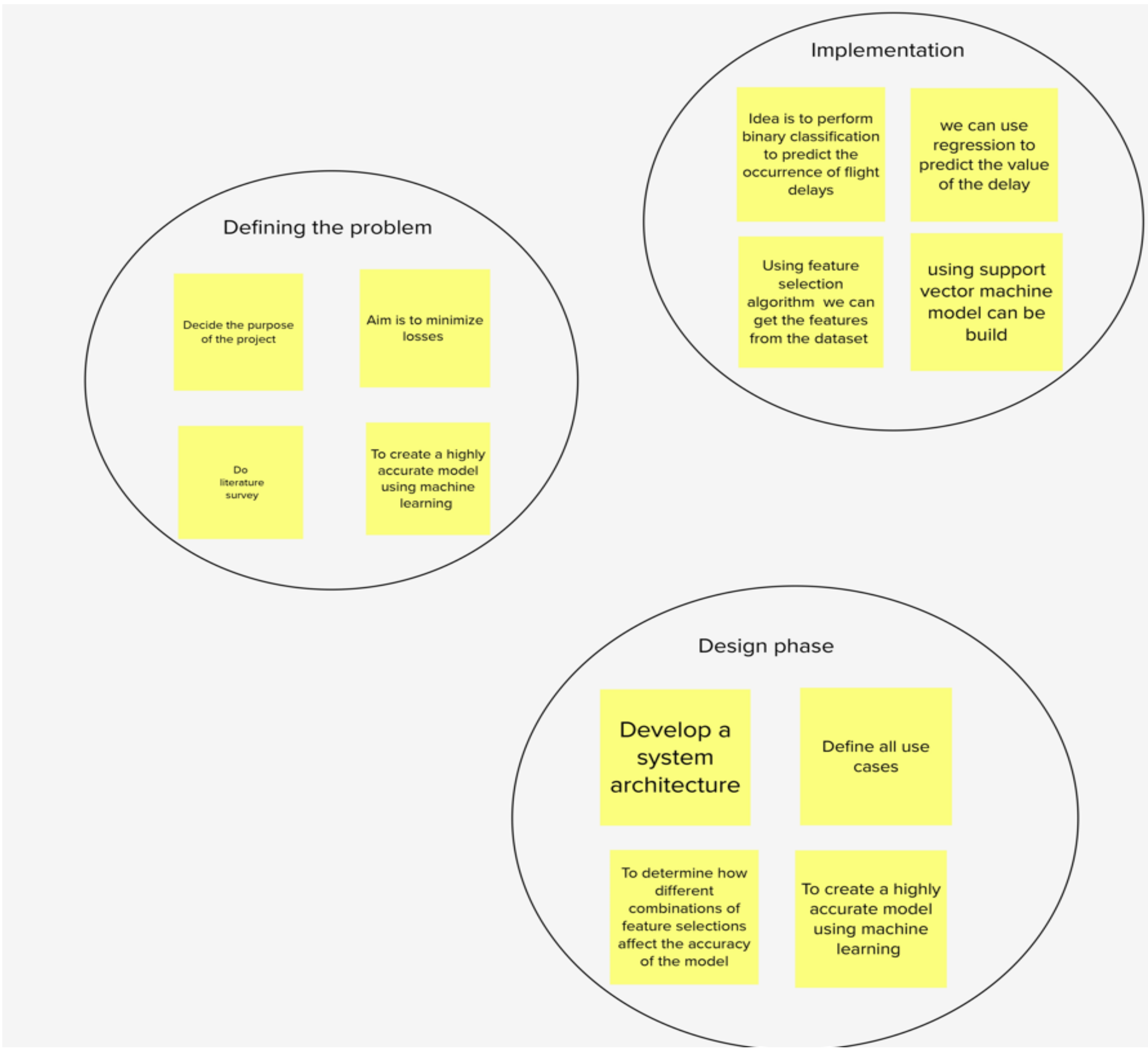
IDEATION AND PROPOSED SOLUTION

3.1 EMPATHY MAP CANVAS



3.2 IDEATION AND BRAINSTORMING




Implementation
Idea is to perform binary classification to predict the occurrence of flight delays
we can use regression to predict the value of the delay
Using feature selection algorithm we can get the features from the dataset
using support vector machine model can be build
Defining the problem
Decide the purpose of the project
Aim is to minimize losses
Do literature survey
To create a highly accurate model using machine learning
Design phase
Develop a system architecture
Define all use cases
To determine how different combinations of feature selections affect the accuracy of the model
To create a highly accurate model using machine learning


Idea is to perform binary classification to predict the occurrence of flight delays
Decide the purpose of the project
Idea is to perform binary classification to predict the occurrence of flight delays
Do literature survey
Develop a system architecture
we can use regression to predict the value of the delay
Define all use cases
using support vector machine model can be build
Using feature selection algorithm we can get the features from the dataset
To determine how different combinations of feature selections affect the accuracy of the model
To create a highly accurate model using machine learning
Develop a system architecture

3.3 PROPOSED SOLUTION

S.No.	PARAMETER	DESCRIPTION
1	Problem Statement (Problem to be solved)	Flight delays in air transportation are a major concern that has adverse effects on the economy, the passengers, and the aviation industry. This matter critically requires an accurate estimation for future flight delays that can be implemented to improve airport operations and customer satisfaction. Having said that, a massive volume of data and an extreme number of parameters have restricted the way to build an accurate model. Many existing flight delay prediction methods are based on small samples and/or are complex to interpret with little or no opportunity for machine learning deployment.

2	Idea / Solution description	The proposed model gains insight into factors causing flight delays, cancellations and the relationship between departure and arrival delay using exploratory data analysis. In addition, Random Forest (RF) algorithm is used to train and test the big dataset to help the model development. A web application has also been developed to implement the model and the testing results are presented with the limitation discussed
3	Novelty / Uniqueness	Many existing flight delay prediction methods are based on small samples and/or are complex to interpret with little or no opportunity for machine learning deployment. The proposed model gains insight into factors causing flight delays, cancellations and the relationship between departure and arrival delay using exploratory data analysis.
4	Social Impact / Customer Satisfaction	An accurate estimation of flight delay is critical for airlines because the results can be applied to increase customer satisfaction and incomes of airline agencies. Predicting flight delays can improve airline operations and passenger satisfaction, which will result in a positive impact on the

		economy
5	Business Model (Revenue Model)	A web application has been developed to provide the end-users an interface to help predict flight delays. In future, we can implement the subscription plan for the prediction process and also if our model predicts well, we can sell it airlines, so they can prior inform the passenger.

6	Scalability of the Solution	The proposed combined method of delay analysis and its prediction can also be further explored in other studies and also can extend the application in more comfortable with the end user. In the situation of airline, they can develop this system and make the passenger feels good and inform prior.

3.4 PROBLEM SOLUTION FIT

Define CS, fit into CC

1. CUSTOMER SEGMENT(S) CS

Passenger as well as airlines

6. CUSTOMER CONSTRAINTS CC

No proper access to application.

5. AVAILABLE SOLUTIONS AS

The solution is to deliver timely message and keep passengers informed about the status.

Explore AS, differentia

Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS

Displaying the delaying time to the passengers is not done on time.

9. PROBLEM ROOT CAUSE RC

Flights being delayed due to natural occurrences like weather delay, extreme heat and severe storms and operational shortcomings, which is an expensive affair for the airlines, creating problems in scheduling and operations for the end-users thus causing bad reputation and customer dissatisfaction.

7. BEHAVIOUR BE

Agree to a new connection

Call the airline

Focus on J&P, tap into BE, understand RC

3. TRIGGERS TR

Waste of time due to flight delay.

10. YOUR SOLUTION SL

We implemented flight delay prediction through the proposed approaches that were based on machine learning algorithms.

8. CHANNELS of BEHAVIOUR CH

8.1 ONLINE

Check For Reimbursements

8.2 OFFLINE

Agree to A New Connection, call airline

4. EMOTIONS: BEFORE / AFTER EM

BEFORE: Many people consider these delays to be a waste of time, and no customers enjoy waiting for long periods due to delays.

AFTER: People can plan accordingly once they known the delay.

CHAPTER 4

REQUIREMENT ANALYSIS

4.1 FUNCTIONAL REQUIREMENTS

FR No.	**Functional requirements (epic)**	**Sub Requirement (Story / Sub-Task)**
FR-1	Details	Getting input like current year, month, date, selecting the airline and airport details from user.
FR-2	Data processing	Given data is fed to the model, using the algorithm it predicts
FR-3	Output	Displaying the result as delayed or not delayed.

4.2 NON-FUNCTIONAL REQUIREMENTS

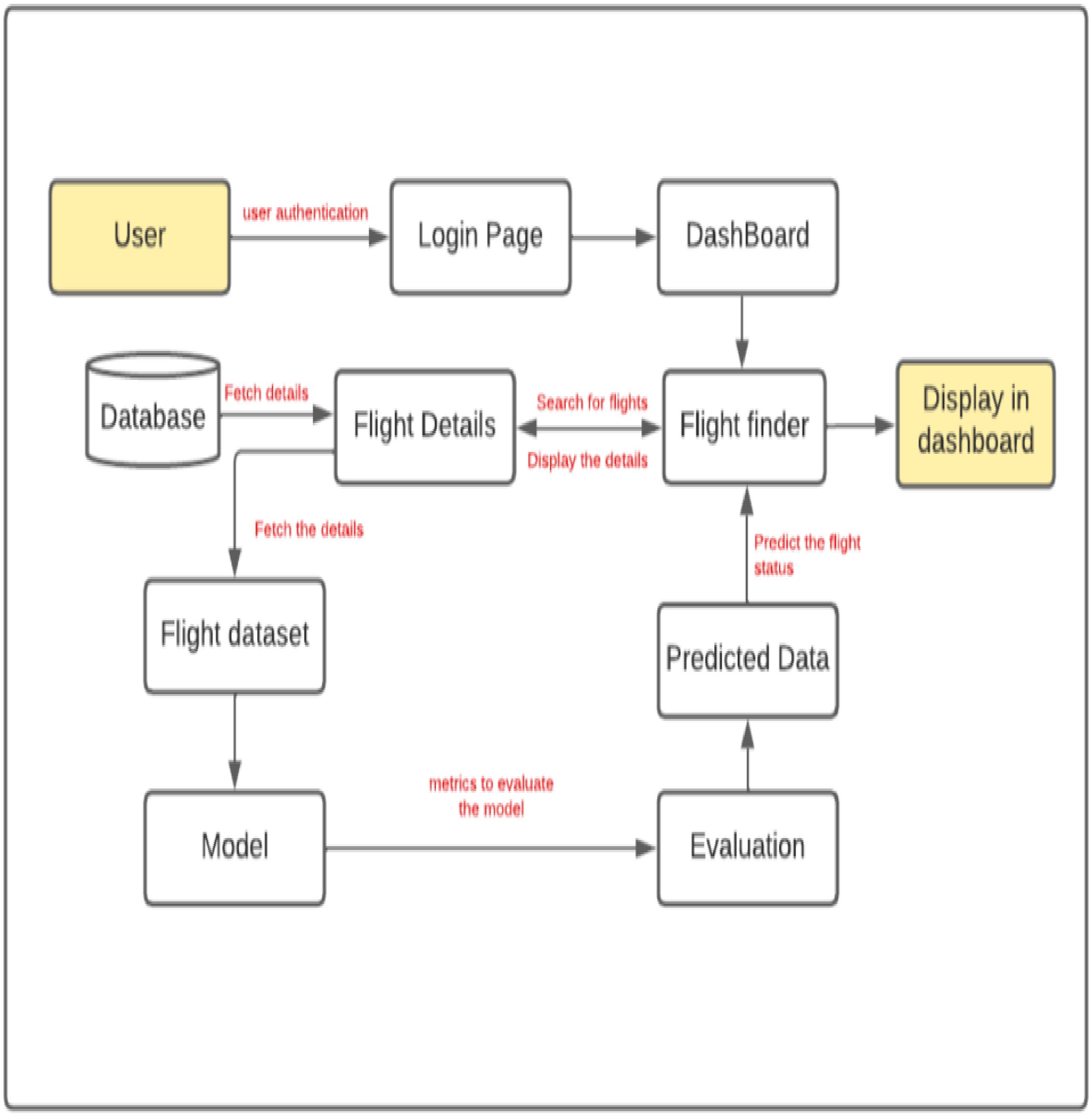
FR No.	**Non-Functional Requirement**	**Description**
NFR-1	**Usability**	User interface is very effective to use when compared with others.

NFR-2	**Security**	The data collected from the user will be stored securely in the cloud
NFR-3	**Reliability**	The user can trust the results from the application and they can check their flight status
NFR-4	**Performance**	Accurate prediction can be achieved.
NFR-5	**Availability**	Available if the network bandwidth of the user is of good range
NFR-6	**Scalability**	This application can be accessed from any place.

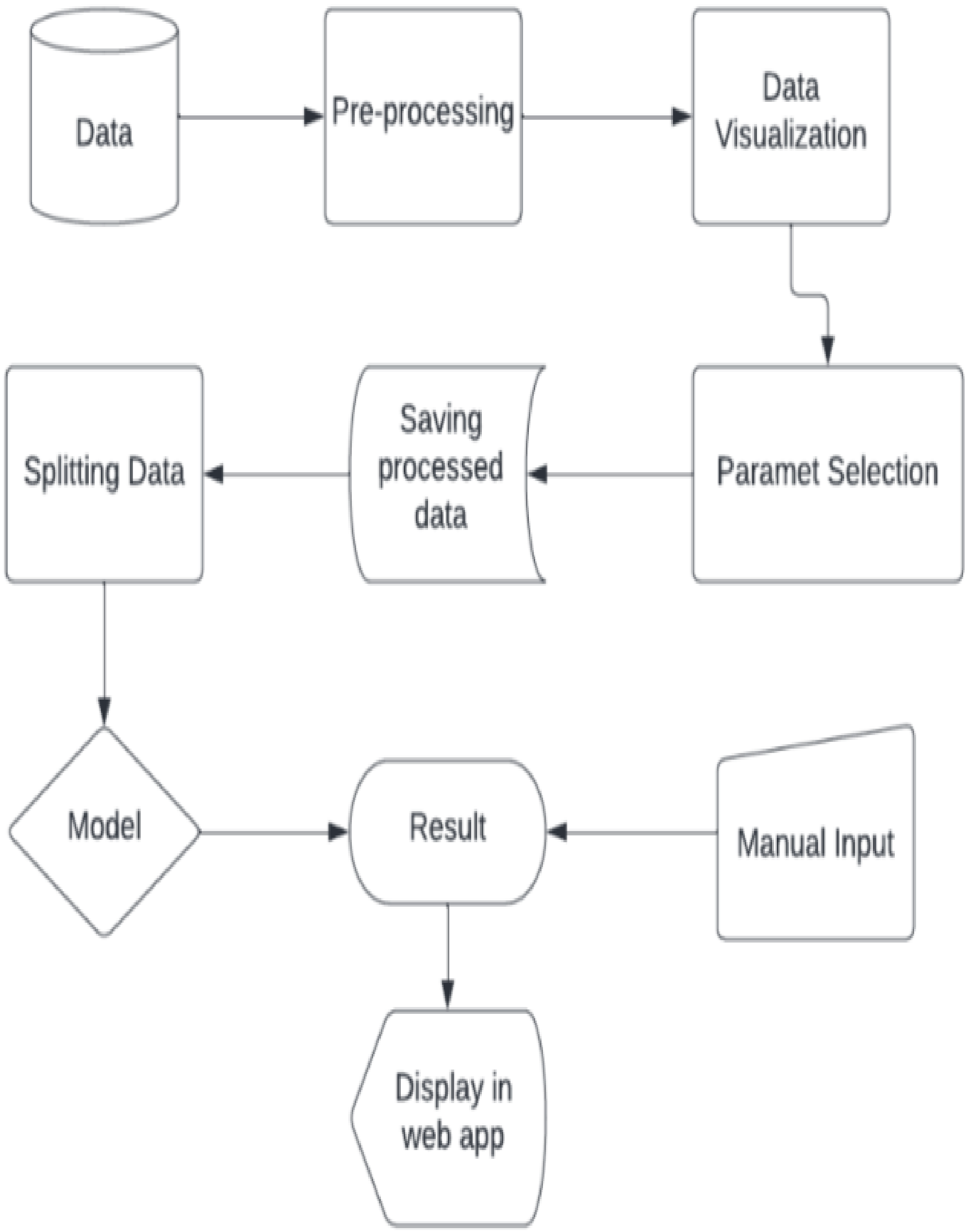
CHAPTER 5

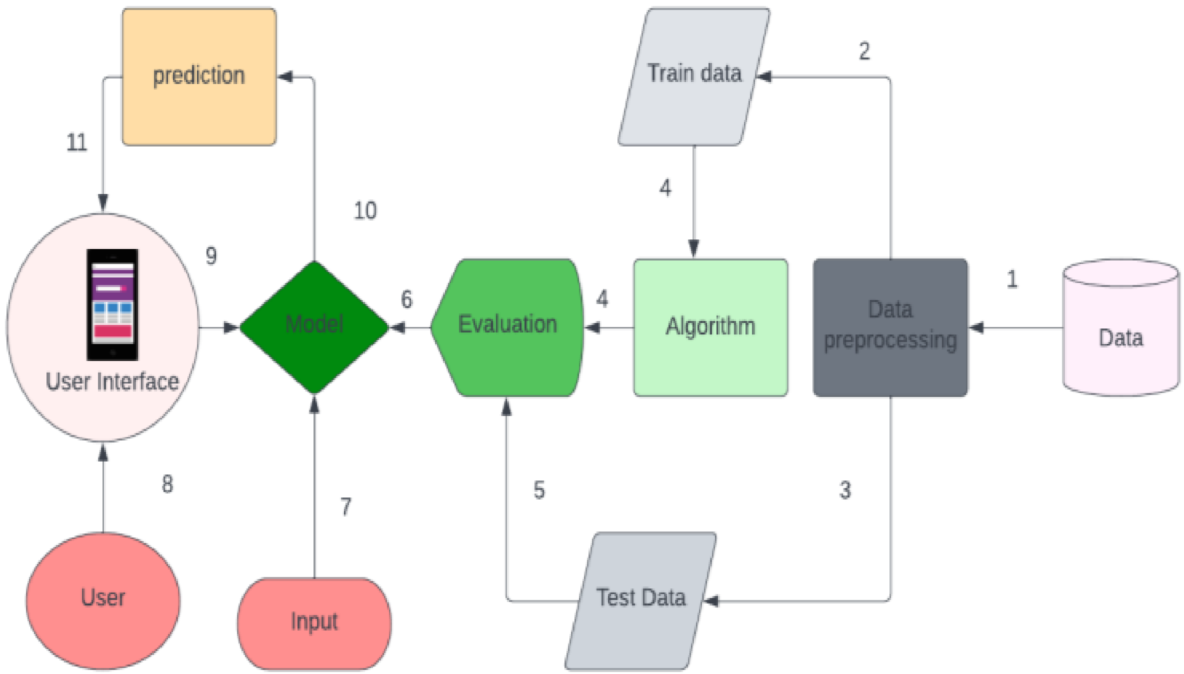
PROJECT DESIGN

5.1 DATA FLOW DIAGRAM



5.2 SOLUTION AND TECHNICAL ARCHITECTURE





5.3 USER STORIES

USER TYPE	USER STORY NUMBER	USER STORY/TASK	ACCEPTANCE CRITERIA	PRIORITY	RELEASE
Customer	USN-1	I can use this web app for flight delay prediction	I am getting the result	Medium	Sprint1
	USN-2	As a tourist	As a user I can	High	Sprint2

		person, I can able to get the accurate result.	able to access the dashboard.		
Customer (Web user)	USN-3	As a user, I can use the web application virtually anywhere	I can use the application in any device with a browser.	Medium	Sprint3
Administrative Management	USN-4	As an administrative I would provide all the IT support	Allows growth and success of the website	High	Sprint-3

CHAPTER 6

PROJECT PLANNING AND SCHEDULING

6.1 SPRINT PLANNING AND ESTIMATION

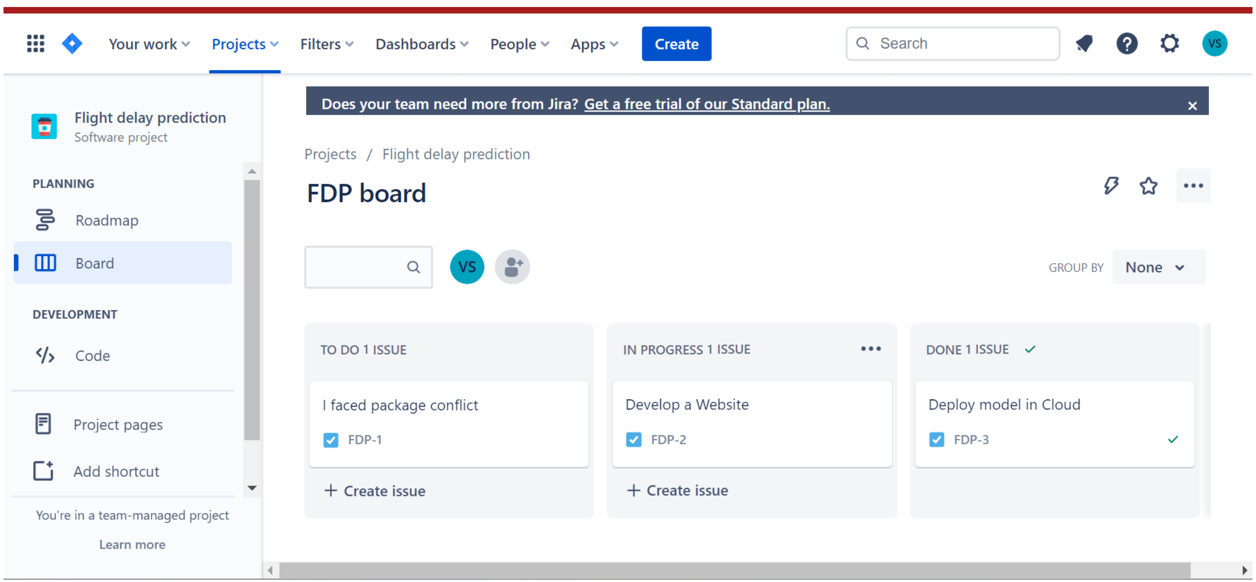
Sprint	Functional Requirement (Epic)	User Story Number	User Story / Task	Story Points	Priority	Team Members
Sprint-1	Data Engineering	USN-1	Data Collection, Data Pre-processing and Feature Extraction	4	High	Tharanyaa R
Sprint-2	Machine Learning Prediction Model	USN-2	Building a Machine Model for Flight Delay Prediction.	4	High	Vijay S
Sprint-3	Flask Web Page	USN-3	Building Home Page.	4	Medium	Gayathri P
Sprint-4	Integration.	USN-4	Integrating the flask pages with the ML Model and IBM Cloud Deployment	4	Medium	Akashram J

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	29 Oct 2022	20	29 Oct 2022

			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

6.3 REPORTS FROM JIRA



CHAPTER 7

CODING AND SOLUTIONING

We completed four sprints—Sprint 1, Sprint 2, Sprint 3 and Sprint 4—during the project development phase

7.1 Sprint 1

The dataset has been downloaded. The features are analysed and visualized and data has been cleaned and pre-processed. The independent and dependent variables are then identified and the dataset is split into train and test sets.

7.2 Sprint 2

Several machine learning algorithms have been applied for classification like logistic regression, K means, naïve bayes and random forest classifier and it is found that logistic regression gives the highest accuracy, so it is used for deployment.

7.3 Sprint 3

We had done building HTML files, written Python code, and running the application during Sprint 2.

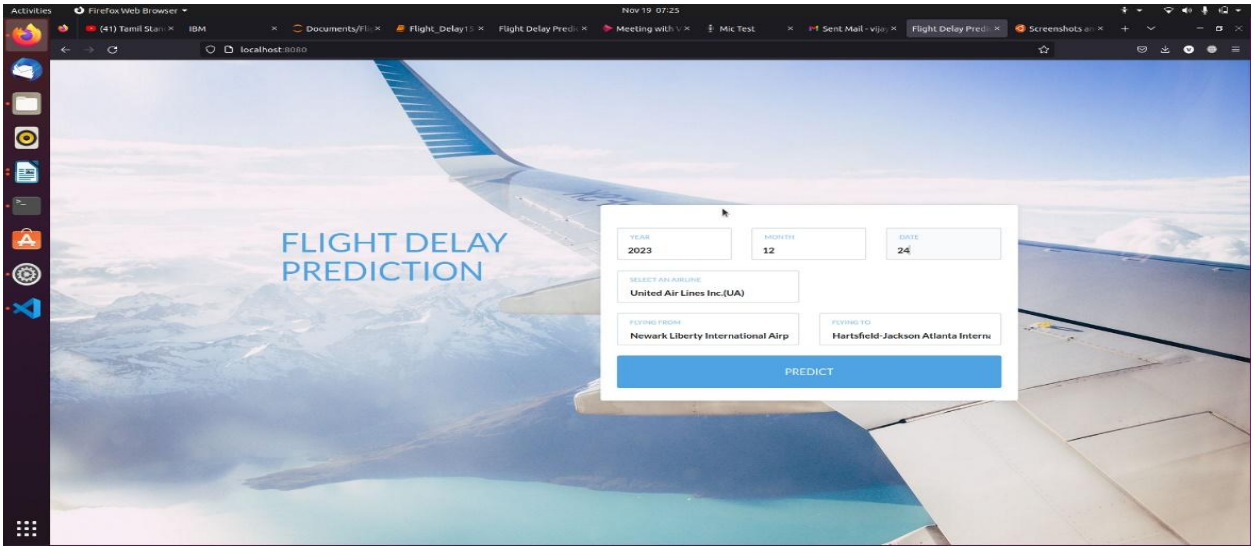
7.4 Sprint 4

We trained the ML model on IBM and integrated the flask. Registered on IBM cloud and activated Watson machine learning, cloud storage and Watson studio then trained the ML model on IBM using API KEY during sprint 4.

CHAPTER 8

TESTING

8.1 TEST CASES



8.2 USER ACCEPTANCE TESTING

Resolution	Severity 1	Severity 2	Severity 3	Severity 4	Subtotal
By Design	1	0	1	0	2
Duplicate	0	1	0	0	1
External	0	0	2	0	2
Fixed	4	1	0	0	5
Not Reproduced	0	0	1	1	2
Skipped	0	0	0	1	1
Won't Fix	1	0	0	0	1
Totals	6	2	4	2	14

Section	Total Cases	Not Tested	Fail	Pass
Client Application	9	0	1	8
Security	2	0	0	2
Exception Reporting	4	0	1	3
Performance	4	0	0	4

CHAPTER 9

RESULTS

9.1 PERFORMANCE METRICS

Model: Logistic Regression performance values

There is no big variation in the training and testing accuracy. Therefore, the Logistic Regression model is not overfit or underfit.

```
In [24]: print("Train set Acuracy: ", metrics.accuracy_score(y_train, LR.predict(X_train)))
         print("Test set Accuracy: ", metrics.accuracy_score(y_test, LR.predict(X_test)))

         Train set Acuracy:  0.72045166414639
         Test set Accuracy:  0.7202230029020925
```

Model: Naive Bayes performance values

There is no big variation in the training and testing accuracy

```
In [13]: print("Train set Acuracy: ", metrics.accuracy_score(y_train,gnb.predict(X_train)))
         print("Test set Accuracy: ", metrics.accuracy_score(y_test, y_pred))

         Train set Acuracy:  0.7207321224393608
         Test set Accuracy:  0.7196448209848886
```

Model: K means performance values

There is no big variation in the training and testing accuracy

```
In [14]: k = 6
         neigh6 = KNeighborsClassifier(n_neighbors=k).fit(X_train, y_train)
         yhat6 = neigh6.predict(X_test)
         print("Train set Acuracy: ", metrics.accuracy_score(y_train, neigh6.predict(X_train)))
         print("Test set Accuracy: ", metrics.accuracy_score(y_test, yhat6))

         Train set Acuracy:  0.7792924895752188
         Test set Accuracy:  0.7262257522529403
```

Model: Random Forest performance values

There is slight variation in the training and testing accuracy

```
In [21]: print("Train set Acuracy: ", metrics.accuracy_score(y_train,clf.predict(X_train)))
         print("Test set Accuracy: ", metrics.accuracy_score(y_test,clf.predict(X_test)))

         Train set Acuracy:  0.8363318656342538
         Test set Accuracy:  0.7019318889988636
```

On comparing the four models built, based on the performance metrics it is clear that random forest gives the highest performance. Hence, that model is chosen for deployment.

CHAPTER 10

10.1 ADVANTAGES

- The application is fast and offers great accuracy in predicting the flight delay.
- Less maintenance is required.
- It is user friendly.
- It helps in reducing the tension of the passengers in knowing how long they will have to wait and lets passengers plan their schedule accordingly, thus in a way saving their time

10.2 DISADVANTAGES

- It requires an internet connection for the website to work.

CHAPTER 11

CONCLUSION

From this study, we have developed a web application model that shows the flight delay prediction. In particular, by applying random forest algorithm to the prediction model, a reliable delay status of a single day could be acquired. Once the model was built it was integrated along with the Flask framework so that the users can enter their flight details and see if the flight would be on time or get delayed. Then this model is trained and deployed in the IBM Cloud.

As a result, anticipating delays can enhance airline operations and passenger satisfaction, which will be benefit the economy and bring a positive impact.

CHAPTER 12

FUTURE SCOPE

The next steps are to apply other algorithms to the prediction and analyse the task of flight delays. It may yield important patterns and accuracy in flight delay data.

Web application can further be improved in which notification is sent via message or mail and allowing administrators to verify the identity of the user.

A section where the users can give their feedback can also be implemented.

CHAPTER 13

APPENDIX

GITHUB LINK

https://github.com/IBM-EPBL/IBM-Project-7979-1658904775.git

PROJECT DEMO LINK

https://drive.google.com/file/d/1YxQk_UI7crlYlI4yUGX3uFkZ6InghEUu/view?usp=share_link