Airplane Crash Analysis and Prediction

Submitted in partial fulfillment of the requirements
of the degree of
Bachelor of Engineering in Information Technology

By

Likita J. Raikar (Roll No. 17101A2012)

Sayali Pardeshi (Roll No. 17101A2008)

Pritam Sawale (Roll No. 16101A0010)

Under the Guidance of

Prof. Neha Kudu

Department of Information Technology

Vidyalankar Institute of Technology

Wadala(E), Mumbai-400437

University of Mumbai

2019-20

CERTIFICATE OF APPROVAL

This is to certify that the project entitled

"Airplane Crash Analysis and Prediction"

is a bonafide work of

Likita J. Raikar (Roll No. 17101A2012)

Sayali Pardeshi (Roll No. 17101A2008)

Pritam Sawale (Roll No. 16101A0010)

submitted to the University of Mumbai in partial fulfilment of the requirement for the award of the

degree of

Undergraduate in **"INFORMATION TECHNOLOGY"**.

Guide	Head of Department	Principal
(Prof. Neha Kudu)	(Prof. Deepali Vora)	(Dr. S.A. Patekar)

Project Report Approval for B. E.

This project report entitled ***Air plane crash Analysis and prediction*** by

Likita J. Raikar (Roll No. 17101A2012)

Sayali Pardeshi (Roll No. 17101A2008)

Pritam Sawale (Roll No. 16101A0010)

is approved for the degree of ***Bachelor of Engineering in Information Technology.***

Examiners

1._____-

2._____-

Date:

Place:

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Name of student	Roll No.	Signature
1. Likita J. Raikar	**(Roll No. 17101A2012)**	
2. Sayali Pardeshi	**(Roll No. 17101A2008)**	
3. Pritam Sawale	**(Roll No. 16101A0010)**	

Date:

ACKNOWLEDGEMNT

We are honoured to present "**Air plane Crash Analysis and Prediction**" as our B.E Final Year Project. We are using this opportunity to express our profound gratitude to our principal "**Dr. Sunil Patekar**" for providing us with all proper facilities and support.

We express our deepest gratitude towards our HOD "**Prof. Deepali Vora**" for her valuable and timely advice during the various phases in our project. We would like to thank our project guide **"Prof. Neha Kudu"** for her support, patience and faith in our capabilities and for giving us flexibility in terms of working and reporting schedules. Finally, we would like to thank everyone who have helped us directly or indirectly in our project.

We would also like to thank our staff members and lab assistant for permitting us to use computer in the lab as when required. We thank our college for providing us with excellent facilities that helped us to complete and present this project.

Likita Raikar (Roll No. 17101A2012)
Sayali Pardeshi (Roll No. 17101A2008)
Pritam Sawale (Roll No. 16101A0010)

ABSTRACT

Airplanes are the most frequent mode of transportation in the present world. A single airplane crash leads to tremendous loss of human life. Safety is of prime importance since a huge number of people travel across the borders and within them. Abstracting data from a large database is always a difficult task. Data mining is a robust technology in order to extract the knowledge from raw data. Aviation systems take care of the minute precautions in order to prevent aircraft crashes. Factors causing and contributing to crashes needs to be understood studied and prevented in order to further minimize any kind of mishap. It is immensely difficult to find and extract the patterns of the factors due to very less amount of accident rates. In this research work, crash analysis and prediction is done. We have conducted the analysis of airplane crash data while, co-relating it with accidental information. To carry out this we have employed machine learning techniques. Machine learning helps in extracting the relationships between the various factors either affecting or non-affecting the crash to the general information of the airplane and as a result, patterns are formed. Many researchers, in recent times, have been using several machine learning techniques to help the aviation industry and the professionals in determining the hurdles. Supervised machine learning algorithms like SVM, K-NN, ADABoost and XGBoost are used for the purpose of prediction.The work has helped in improving the accuracy to a great extent.

Index Terms – Air crash, safety, prediction, classification, KNN, SVM, ADABoost, XGBoost

Contents

CHAPTER 1. INTRODUCTION

The increase in technology has resulted in advancement in the systems that are used for the purpose of predicting and analyzing the existing records. Machine learning is a technology used for automatically making the system learn without explicitly performing the instructions. Large amount of data with respect to the past records are used for training the model. The attributes that contribute to the crash of a particular airplane are taken into the consideration.

Classifying or predicting the number of individual data that is collectively associated is called as classification. The dataset is filtered and normalized. To predict the exact classifier class for each record in the dataset, it is the prime objective of classification. A classifier is able to tolerate noise and this is its essential quality. Classifier can handle quantitative data but it is very difficult to carry out this process.

Safety is of prime concern for the applications in aviation industries. Companies carry out numerous investigations to create reports and collect information to justify the crash records and hence, this information can be in either conceptual form or structured/non-structured form. Feature selection is one of the prime stages in machine learning. Correct and relatable features should be selected with reference to the output in order to achieve highest accurate results. Data contains a lot of redundant values and these values should be filtered out in order to remove the irrelevant features. Irrelevant values only reduce the valuable assets from the output. The initial number of features is reduced.

The new dataset has features like the ‘aircraft category’, ’purpose of the flight’, ’total injured’, ‘built of the plane’ ,’model type’, ‘manufacturing company’ that are highly appropriate for predicting the safety aspect of the aircraft.

Fours algorithms namely Support Vector Machines (SVM), K-Nearest Neighbours (KNN), ADABoost, XGBoost are used for the purpose of classification. Precision, recall and f1 score are the performance factors used to improve the accuracy of the classification. These performance metrics are essential because they determine the efficiency of the analysis. It is very overwhelming for the humans to manage the datasets.

These algorithms, as a result can efficiently explore them. KNN is an algorithm in which the

similarity in the features of ‘built’, ‘incident type’ is used for creating values of new data points. SVM is used in text, image classification and its main task is to sort the data into two categories and the sorted data is separated with margins away from one another. XGBoost focuses on the ‘schedule’, ‘make’, ‘injury severity’, ‘amateur type’ in order to boost the predictability of the airplane crash.

Scope:

This project consists of a prediction system that predicts the safety of an airplane based the algorithms implemented on the dataset. The analytical scope of the projects causes the fulfillment of the objectives set for the application. The functional scope is constituted to all the records that are included in the database of the investigation of the airplane data.

Motivation:

A major challenge faced by airplane companies is the provision of quality services at affordable costs. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most airline companies today employ some sort of aviation information systems to manage their healthcare or patient data . These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support the crash decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: “How can we turn data into useful information that can enable aviation companies to make intelligent decisions ?” This is the main motivation for this project.

CHAPTER 2. PROBLEM STATEMENT

In today's world there are various type of predicting application used to analyze and provide solution to the future records. Airline industry is advancing day by day. Safety measures are taken at every provided situation by the companies. Also, the risk factors are examined for prevention of human loss. A single airplane crash can lead to a great loss of human life and property. There are numerous factors that leads to the airplane crash which are the airplane type, built of the model, weather conditions, make of the airplane, engine type, phase of the flight etc. Hence, taking all these factors into consideration, based on the details of a particular aircraft the analysis of the airplane crash is carried out. In order to predict whether the airplane is safe or at a risk the application is built where these functionalities are processed and the safety is predicted.

CHAPTER 3. AIM and OBJECTIVES

- The aim of the our project is to provide a comprehensive solution for dealing with the flight crashes**.**
- The project comprises integrating different services facilitating foresight, prediction, communication and eventually decision making in crisis situation.
- The high-quality of the developed solutions is assured by sufficient number of testing segments.
- Our project starts from a deep analysis of recent cases accompanied with scenarios of potential crisis.
- Our project will deliver two core components:
    1. Foresight and Prediction Analysis
    2. Decision-Support Analysis

CHAPTER 4. LITERATURE SURVEY

Many researches have been done on the prediction of aircrafts accidents. We start the section by presenting some related works in prediction in general.

"Hybrid safety and analysis method based on SVM and RST", [1] *Ying Dai, Jin Tian, HaoRong, Tingdi Zhao,* research paper focused on providing a safe landing without an accident.

"Data mining approaches for aircraft accidents prediction", [2] by *A.B Arockia Christopher and Dr. S.Appavu,* used decision tree method to predict the warning level. Dataset used are pilot details, delay details, accident related details, maintenance details and flight details.

A.B Arockia Christopher and Dr. S.Appavu[3] analyzed various data preprocessing techniques to find best techniques which suits for airline data. Classification algorithms and many clustering techniques are used in for comparison. The data mining tool weka was used in this process. As a result of this analysis they have proved that Principal Components Attributes (PCA) Transformer would perform better than other attribute evaluators on airline data to reduce the dataset. On an empirical study on Turkey airline, decision tree technique is used to generate model. This model in turn is used to predict the warning level.

A Short Introduction to Boosting Yoav Freund Robert E. Schapire AT&T Labs Research Shannon Laboratory 180 Park Avenue Florham Park, NJ 07932 USA www.research.att.com/ yoav, schapire yoav, schapire @research.att.co

Boosting is a general method for improving the accuracy of any given learning algorithm. This short overview paper introduces the boosting algorithm AdaBoost, and explains the underlying theory of boosting, including an explanation of why boosting often does not suffer from overfitting as well as boosting's relationship to support-vector machines. Some examples of recent applications of boosting are also described.

XGBoost: A Scalable Tree Boosting System Tianqi Chen University of Washington tqchen@cs.washington.edu Carlos Guestrin University of Washington

guestrin@cs.washington.edu

Tree boosting is a highly effective and widely used machine learning method. In this paper, a scalable endto-end tree boosting system called XGBoost, which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. A novel sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning. More importantly, insights on cache access patterns, data compression and sharding to build a scalable tree boosting system is provided. By combining these insights, XGBoost scales beyond billions of examples using far fewer resources than existing systems.

From tutorialspoint.com:

K-nearest neighbours (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well −

- **Lazy learning algorithm** − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm** − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

CHAPTER 4. SYSTEM DESIGN

PROPOSED SYSTEM

The proposed system provides the person using the system to enter the specifications of the flight in order to know whether the flight is safe or has changes of a crash. Based on the past records of various airline companies the analysis and prediction of the given input is carried out.

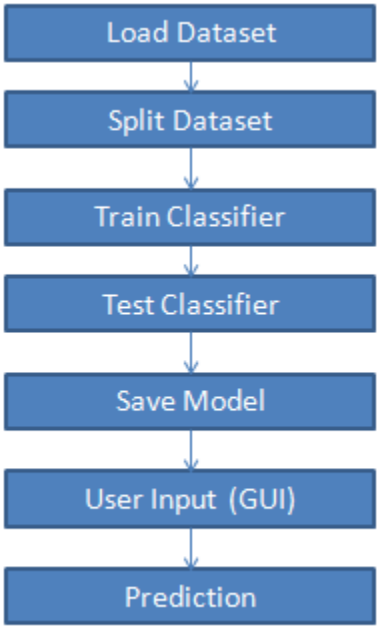

Fig1: flow chart of project

Machine learning is a strong and dependable technology in order to predict the values. Four algorithms are used and based on every dataset the best algorithm would be used in order to predict the value. Every dataset varies majorly and hence the algorithms used for the classification may also differ. To overcome this problem our system works accordingly.

1. **Feature Selection**:

Redundancy, irrelevant data, noise etc. are removed or none the less reduced to a great extent from a huge dataset having multiple attributes. It comes under the pre-processing step in machine learning. The attributes like ‘weather’, ‘model type’, ‘engine type’, ‘injured severity’, ‘manufacturing company’ etc. are the prime ones that add value to the desired output are selected based on the specification of the aviation industry.

Every attribute is taken into consideration and its importance is measured by relating it to the output required. The attributes that does not contribute to the result like the ‘latitude and longitude’ or are of least importance are deleted.

The final dataset with the selected features are evaluated to check whether the subset is most relevant for prediction. Also, these attributes are sorted in a specific order from highest to lowest based on its importance on the prediction. As a result only useful and relevant features are added hence increasing the accuracy of the prediction.

2. Processing on the Dataset:

Once the pre-processing is carried out the specific dataset is loaded into the system. In order to carry out the processing the dataset must be structured. The given dataset is cleaned that is all the missing values are removed by the process of using attribute mean for all samples belonging to the same class also known as aggregation. The dataset is now ready to be loaded. For the purpose of performing machine learning algorithms on this dataset the data has to be split in training data and testing data. The appropriate split ratio for the dataset is 70:30, 70% for training data and 30% for testing data. Now the dataset is ready for the algorithms to carry out its processing.

3. Train and Test the Classifier:

The model views and learns from the training phase. However, the testing phase is used to evaluate the model based on its performance. The records from the dataset which has been converted into a structured format, the training model should not be included in the testing otherwise it won’t produce correct results. Also if the dataset is unbalanced it would create the problem of over-fitting and under-fitting. Hence these phases essentially produce the output which is the actual prediction that is “safe” or ”crash”. All the four algorithms are used for training and testing and the best algorithm with highest accuracy is used for the purpose of prediction.

4. User Input(GUI):

The user enters all the parameters which are the specifications of the airplane that they want to know about. The model takes into consideration all these features and carry out the prediction for these instances. The output of the prediction is either “safe” or “crash” based on the details of the airplane.

Methodology:

In order to explain the concept of time saved in our proposed system, we decided to build a prototype. The prototype is specifically designed to predict the status of the airplane time based on the previous scenarios. The dataset is converted into a structured format.

This forms our database required for processing. The database consists of past records with numerous attributes of airplane crashes. The records from this database constitute the input to our system. Furthermore algorithms are applied on the database. They undergo the process of analysis and noise removal in order to make the process of prediction more precise. The output displayed shows whether the airplane is 'Safe' or is going to 'Crash'.

Analysis: (Process Model):

Waterfall Model:

We will be using waterfall model in order to develop our project. The reasons for using waterfall model are as follows:

- It allows for departmentalization and control. A schedule can be set with deadlines for each stage of development and a product can proceed through the development process model phases one by one.

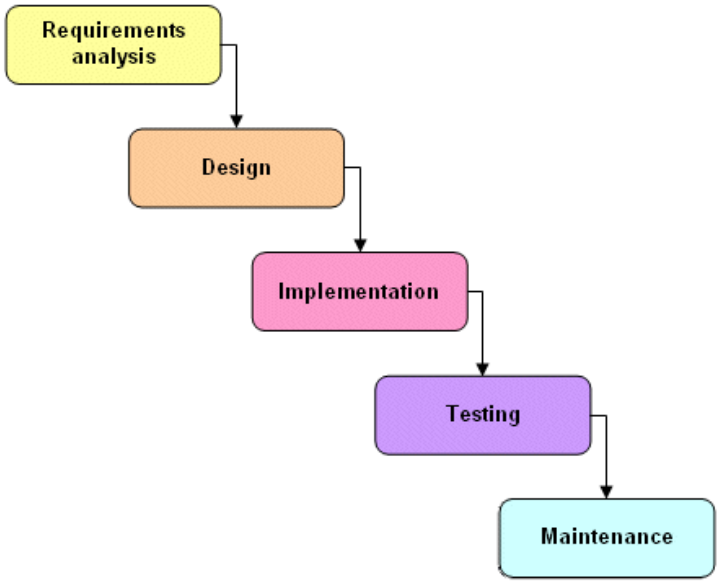

Fig2: Waterfall model

- All the requirements are documented beforehand.
- The waterfall model progresses through easily understandable and explainable phases and thus it is easy to use.
- It is easy to manage due to the rigidity of the model – each phase has specific deliverables and a review process.
- In this model, phases are processed and completed one at a time and they do not overlap.

Feasibility Analysis:

1. **Technical feasibility**: Technical feasibility focuses on the technical resources (software and hardware) available to the organization and helps to determine whether the technical team is capable of converting the ideas into working systems. The software required for our project, Python IDE, is already ready with us. The project involves solvable technical issues.

2. **Economic feasibility**: This assessment typically involves a cost/ benefits analysis of the project. This project developed in minimal amount will emerge as a huge market valued project that will solve the safety issues.

3. **Legal feasibility**: This assessment investigates whether any aspect of the proposed project conflicts with legal requirements like zoning laws, data protection acts, or social media laws. The project does not involve any legal concerns.

4. **Operational feasibility**: This assessment involves undertaking a study to analyze and determine whether—and how well—the organization's needs can be met by completing the project. The main objective of the project is to predict accurately the safety status of the airplane. The time saving factor is achieved by using image processing technique.

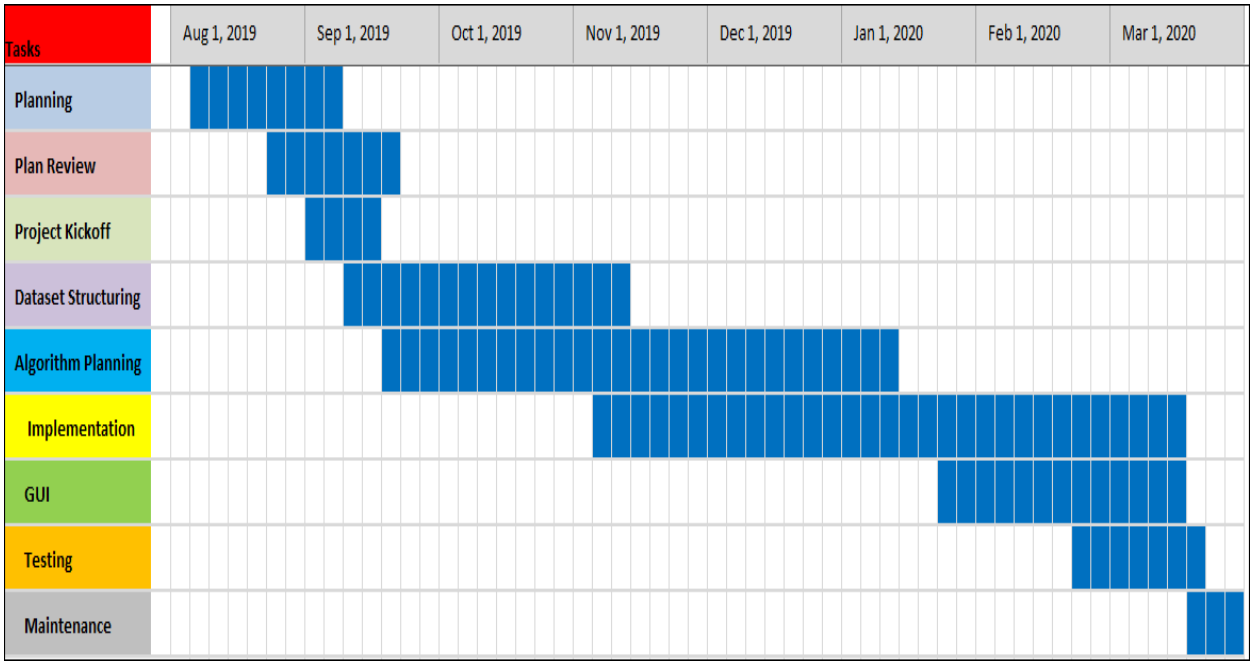

Fig3: Gantt Chart

CHAPTER 5. CLASSIFICATION AND ALGORITHMS

5.1. KNN (K Nearest Neighbour):

KNN algorithm predicts the most comparable training value by assigning the class mark or continuous target value on the labeled attribute. This can be used for classification and regression related problems. KNN Algorithm is an algorithm of supervised machine learning that is fast and simple to implement. The classification method is non- parametric. It is also a learning algorithm based on an instance, where the function is locallyapproximated.

KNN classification is implemented as follows –

i. Determine the value of the distance between the test data point and all the marked data points having value ‘safe’’ and ‘crash’.
ii. Order the named data points of the selected attributes in the increasing order of certain distance metrics which are closest to k.
iii. Pick the data points labelled with the top k and look at the class labels.
iv. Find the class label most of these data points called k have and assign it to the test data point which can either be ‘safe’ or ‘crash’.

Following are some things one should consider:

i. Parameter selection
ii. Presence of noise
iii. Feature selection and scaling
iv. Curse of dimensionality

- **Parameter selection:**

It is the data upon which k's best choice depends. High k values reduce the effect of noise on classification but make the boundaries of the decisions less distinct between groups. Smaller k values tend to be influenced by the noise, with strong class separation.

- **Feature selection and scaling:**

It is important to remove unnecessary traits. If the number of characteristics become too high and the feature gets assumed it requires highly redundant extraction. Hence features that leave a great impact on the result like the ‘weather’, ‘mode type’, ‘engine type’ etc. are selected as a result

the accuracy is improved and the redundancy is reduced. When features are carefully selected, the classification will always give better results.

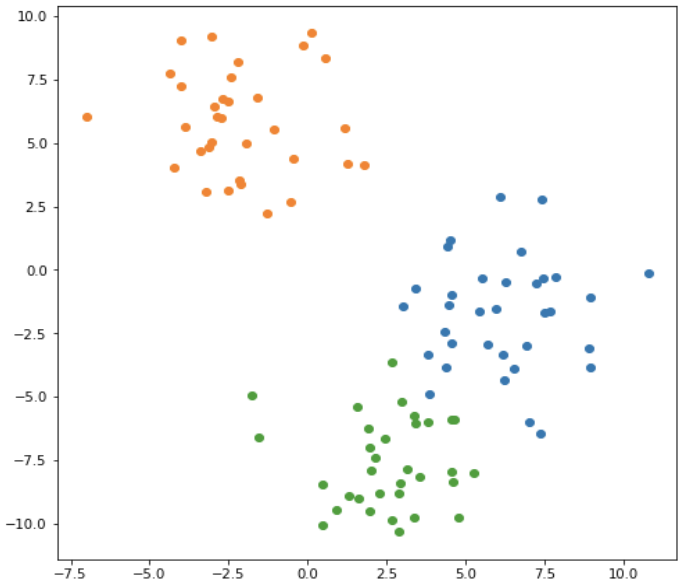

Fig4: Neighborhood of KNN algorithm

Here the neighbors are formed in such a manner for every labeled attribute that classification for the data points is performed accordingly.

5.2. SVM (Support Vector Machine):

Every data item is plotted with its associated value of the labeled attribute. Classification is carried out in such a manner that it separates the given classes, here each separated area is known as the hyper-plane. It is very essential to group the data items in the dataset to their right hyper-plane. This process is known as the identification of the hyper-plane. The data points with ‘safe’ and ‘crash’ values are being separated using this algorithm. These separated values hence generate two hyper-planes for the labeled attributes.

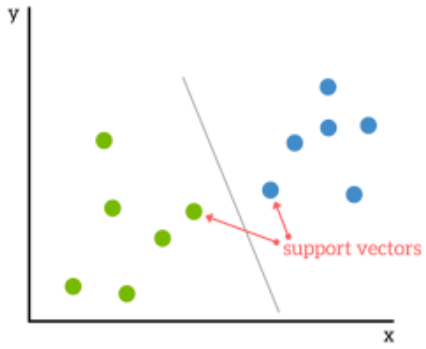

Fig5: Hyper-planes of SVM

The support vector machines in scikit-learn support both dense (numpy.ndarray and convertible to that by numpy.asarray) and sparse (any scipy.sparse) sample vectors as input. SVM to make predictions for sparse data, it must have been fit on such data.

Here the above figure represents how the data points are separated into two hyper-planes for each labeled attribute.

The advantages of support vector machines are:

i. Effective in high dimensional spaces.
ii. Still effective in cases where number of dimensions is greater than the number of samples.
iii. Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
iv. Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

5.3. ADABoost:

Weak classifiers should be converted into strong classifiers, hence boosting came into picture in machine learning. Weak classifiers are always beneficial than random guesses. As a result, these classifiers prove to be robust and solve the problem of over-fitting when applied on large dataset. Hence, the weak ones provide efficient results than random values. A single feature is focused upon which has any random kind of threshold applied on it. If the feature is above the threshold than predicted, it belongs to positive otherwise belongs to negative. AdaBoost stands for "Adaptive Boosting" which transforms weak learners or predictors to strong predictors in order

to solve the problem of classification.

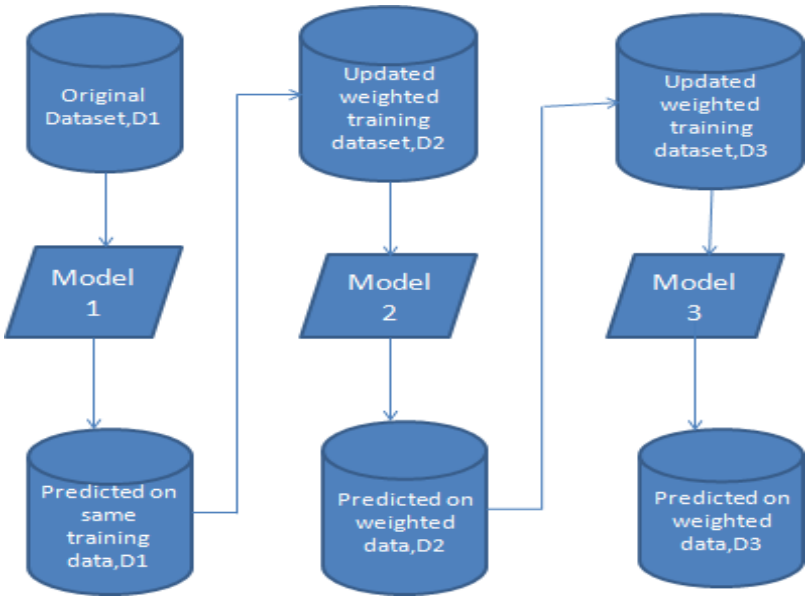

Fig6 : ADABoost Classifier and Flowchart

Here the above figure shows the classification flowchart of ADABoost algorithm.

A weak classifier in this case called as the 'safe' is prepared on the training data using the weighted samples. Only binary classification problems are supported. So each decision stump makes one decision on one input variable. And outputs are +1.0 or -1.0 value for the first or second class value. The misclassification rate is calculated for the trained model. Traditionally, this is calculated as:

$$\text{error} = (\text{correct} - N) / N$$

Here error is the misclassification rate. While correct is the number of training instance predicted by the model. And N is the total number of training instances.

For classification, below is the final equation:

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)),$$

Here fm designates the classifier mth weak and m represents its corresponding weight.

5.4.XGBoost:

XGBoost is a machine learning algorithm that is used for the implementation of gradient boosting decision trees. It is used for classifying non-structured or semi-structured data. This algorithm boosts the speed and improves the performance of the models. The tree model as well as the linear model is included in XGBoost. Since these two models work in a single algorithm parallel computing is enabled on every individual machine. The difference between the target and the predicted outputs are minimized. New trees are iteratively added that are used to predict the errors from the previous trees. Hence, it is known as gradient boosting. New models are added to minimize the loss.

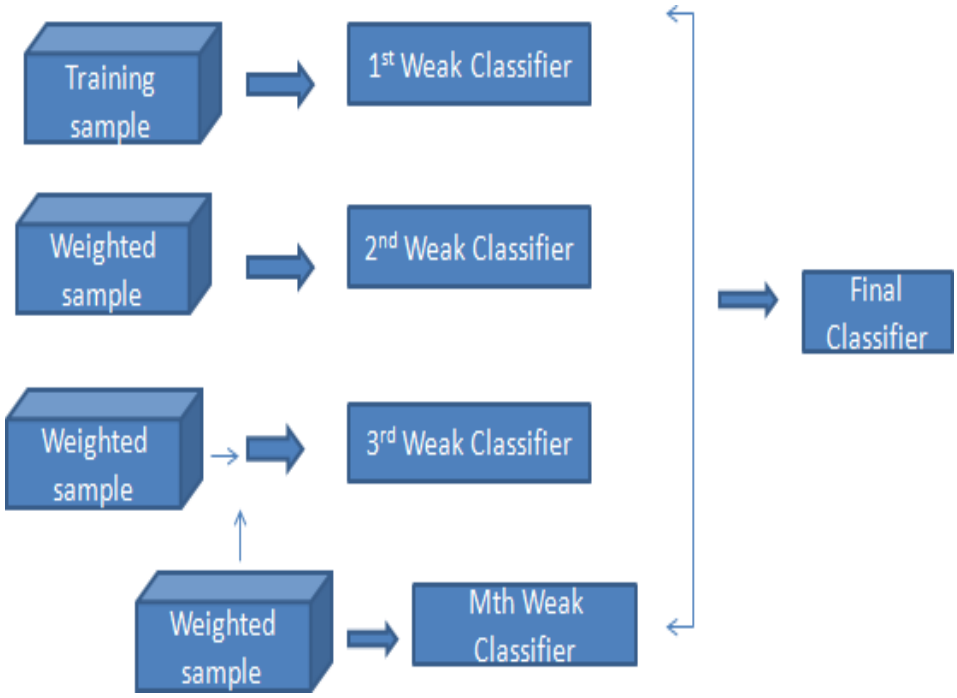

Fig7 : XGBoost Classifier and Flowchart

XGBoost is used for twopurposes in the project:

- Execution Speed.
- Model Performance.

This is an ensemble method that seeks to create a strong classifier (model) based on “weak” classifiers. Weak and strong refer to a measure of how correlated are the learners to the actual target variable. By adding models on top of each other iteratively, the errors of the previous model are corrected by the next predictor, until the training data is accurately predicted or

reproduced by the model.

Gradient boosting also comprises an ensemble method that sequentially adds predictors and corrects previous models of SVM, KNN. However, instead of assigning different weights to the classifiers after each iteration, this method fits the new model to new residuals of the previous prediction and then minimizes the loss when adding the latest prediction. So, in the end, you are updating your model using gradient descent and hence the name, gradient boosting. This is supported for both regression and classification problems. XGBoost specifically, implements this algorithm for decision tree boosting with an additional custom regularization term in the objective function.

CHAPTER 6. DATA ANALYSIS

Companies produce their records for each airplane and hence, these records are collected to form a dataset wherein details about every airplane module are stored hence, a huge dataset comprising of thousands of records is formed. Such datasets are loaded with a large amount of attributes. Numerous attributes are required in order to justify the airplane incident. Hence, the data in these attributes is unstructured and textual. The dataset has to be brought and cut down in such a manner where the classification algorithms can be performed. The attributes that define the output of the prediction are the target attributes. The target variables depend upon the safety of the airplane.

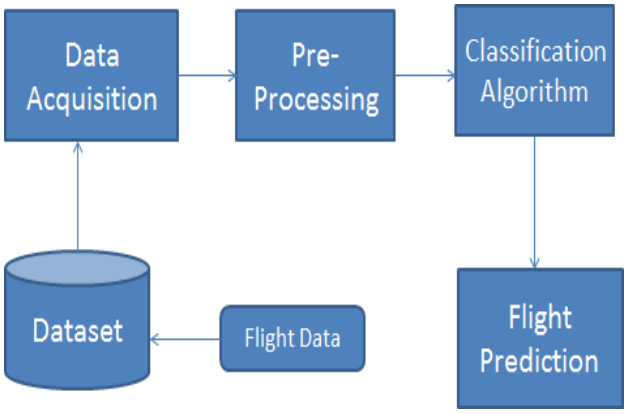

Fig8 : Flight Prediction Flowchart

Classification algorithms carry out the process of data analysis. Since accuracy is the main performance measure, the best classification algorithm is selected is for the purpose of prediction depending on the dataset. The below table depicts the accuracy of each algorithm.

Algorithm used	Accuracy
KNN	79.7%
SVM	80.54%
ADABoost	83.52%
XGBoost	86.12%

Table1:Accuracy of algorithms used

Here the above table defines the accuracy of the different types of algorithms used for the prediction.

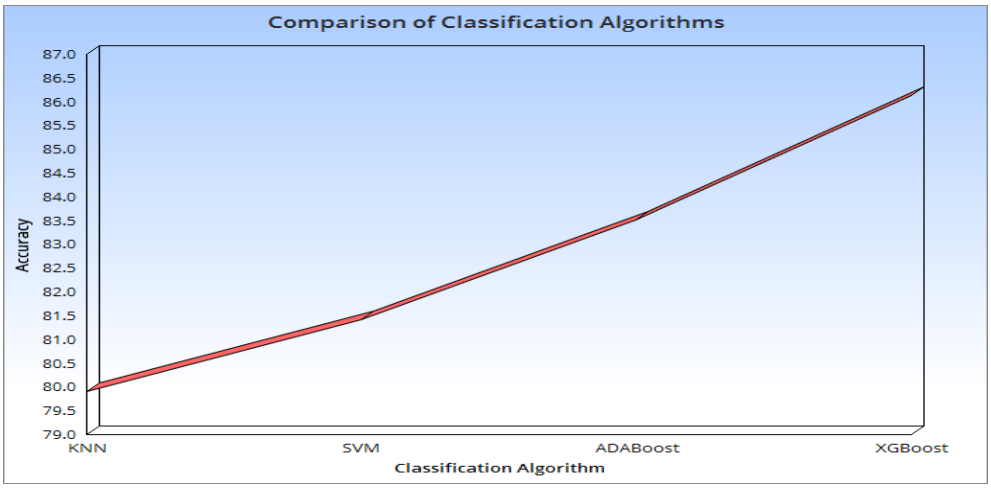

Fig9 : Comparison of Classification Algorithms

Here the above figure shows the comparison of all the four classification algorithms used.

CHAPTER7. PERFORMANCE METRICS

7.1 Precision:

Precision helps when the costs of false positives are high. So let's assume the problem involves the detection of skin cancer. If we have a model that has very low precision, then many data points will be told that they have crashes, and that will include some mis-prediction. Lots of extra tests and stress are at stake. When false positives are too high, those who monitor the results will learn to ignore them after being bombarded with false alarms.

$$Precision = \frac{truepositives}{truepositives + falsepositives}$$

7.2 Recall:

Recall helps when the cost of false negatives is high. A false negative has devastating consequences. Get it wrong and the prediction is wrong. When false negatives are frequent, you get hit by the thing you want to avoid. A false negative is when you decide to ignore the sound of a twig breaking in a dark forest, and you get eaten by a bear. If you had a model that let in crashes by mistake, you would want to throw it out. Then you need to optimize for an evaluation metric that's a combined measure of precision and recall.

$$Recall = \frac{truepositives}{truepositives + falsenegatives}$$

7.3 F1 Score:

F1 is an overall measure of a model's accuracy that combines precision and recall, in that weird way that addition and multiplication just mix two ingredients to make a separate dish altogether. That is, a good F1 score means that you have low false positives and low false negatives, so you're correctly identifying real threats and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}$$

7.4 AUC-ROC Curve:

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

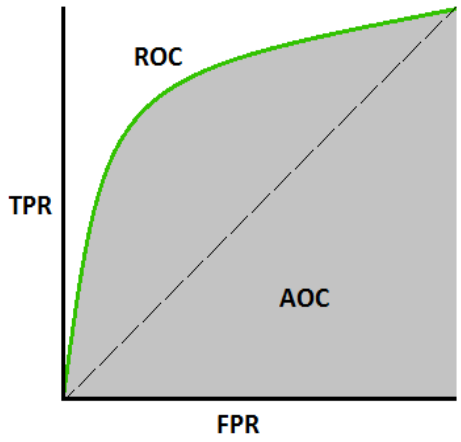

Fig10: AUC ROC curve

Here the above figure represents how an ideal ROC curve should be in order to achieve higher accuracy.

CHAPTER 8. RESULTS

The prediction of the system helps the user in taking the necessary precautions to prevent the mishap or any airplane crashes. The administration department thus become aware of the possible difficulties and hurdles that might come along the way. As a result, the elementary steps taken will help to eradicate any crashes that might occur thus leading to minimize the loss of property and life. Below are all the results of the application.

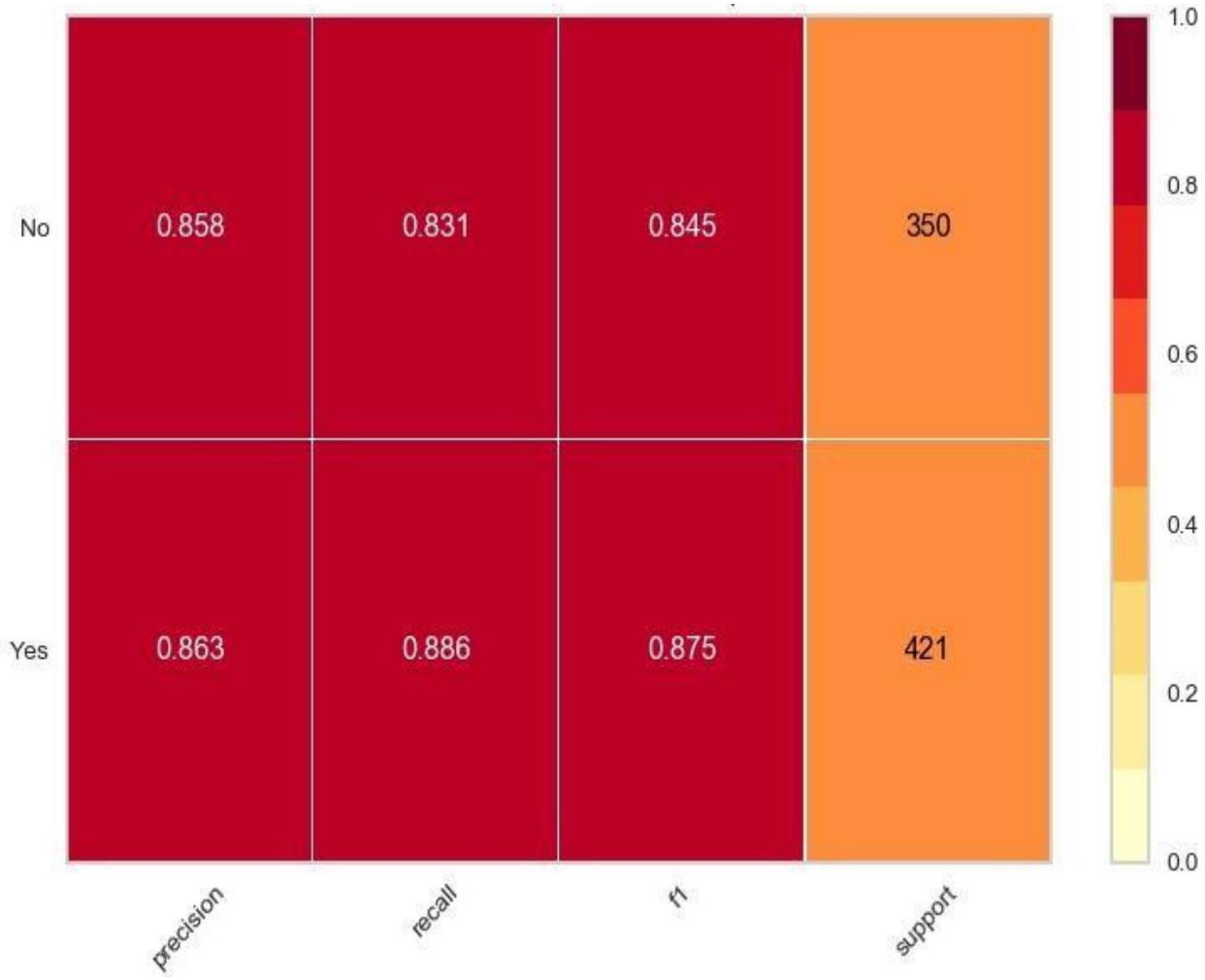

Fig11 : Performance Metrics

Here the above figure represent the performance of the classification algorithm.

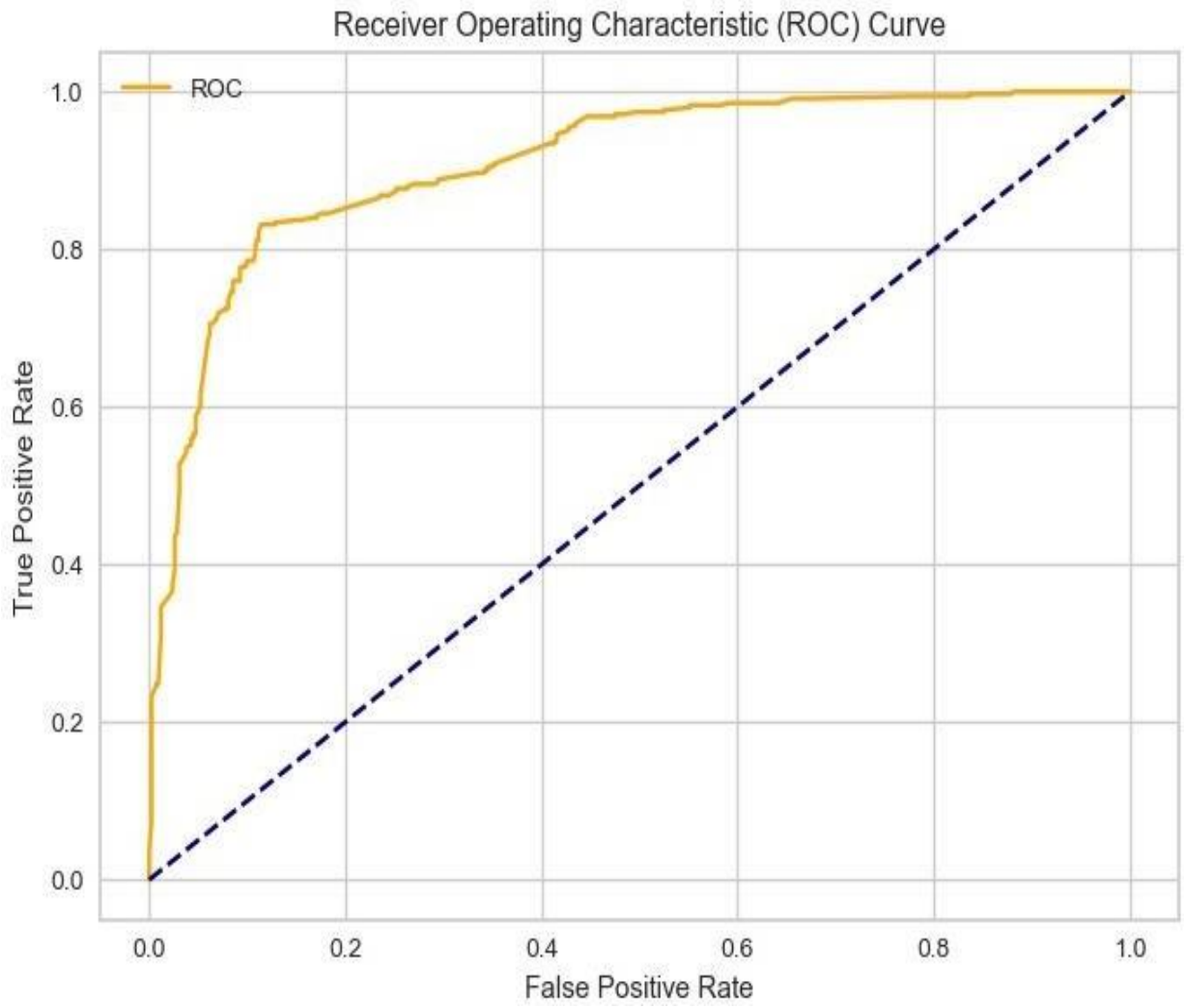

Fig12 : ROC Curve

Here the above figure represents the ROC curve of the model.
The ROC curve is perpendicular to the axis almost forming a right angle hence the accuracy of the model is high. As a result the prediction will be highly accurate.

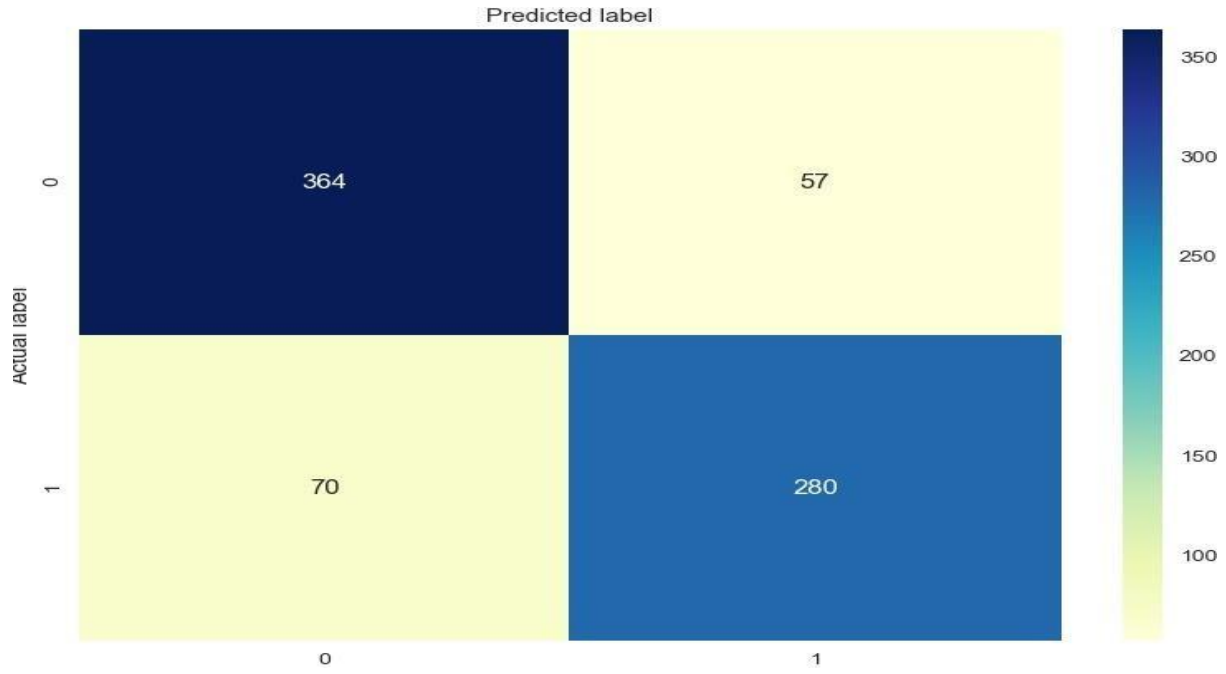

Fig13 : Confusion Matrix

Here the above figure represents the confusion matrix of the dataset.

```
XGBOOST:train set
XGBOOST:Confusion Matrix:  [[820 125]
 [150 702]]
XGBOOST:Accuracy :  84.69671675013912
XGBOOST:Test set
XGBOOST:Confusion Matrix:  [[373  48]
 [ 59 291]]
XGBOOST:Accuracy :  86.12191958495461
AUC: 0.91
              precision    recall  f1-score   support

         Yes       0.86      0.89      0.87       421
          No       0.86      0.83      0.84       350

    accuracy                           0.86       771
   macro avg       0.86      0.86      0.86       771
weighted avg       0.86      0.86      0.86       771
```

Fig14 : XGBoost Algorithm

```
Ada Boost:train set
Ada Boost:Confusion Matrix:  [[815 130]
 [172 680]]
Ada Boost:Accuracy :  83.1942125765164
Ada Boost:Test set
Ada Boost:Confusion Matrix:  [[364  57]
 [ 70 280]]
Ada Boost:Accuracy :  83.52788586251621
AUC: 0.91
              precision    recall  f1-score   support

         Yes       0.84      0.86      0.85       421
          No       0.83      0.80      0.82       350

    accuracy                           0.84       771
   macro avg       0.83      0.83      0.83       771
weighted avg       0.84      0.84      0.83       771
```

Fig15 : ADABoost Algorithm

```
Suppor_Vector_Machine:train set
Suppor_Vector_Machine:Confusion Matrix:  [[800 145]
 [189 663]]
Suppor_Vector_Machine:Accuracy :  81.41346688925988
Suppor_Vector_Machine:Test set
Suppor_Vector_Machine:Confusion Matrix:  [[354  67]
 [ 83 267]]
Suppor_Vector_Machine:Accuracy :  80.54474708171206
AUC: 0.85
              precision    recall  f1-score   support

         Yes       0.81      0.84      0.83       421
          No       0.80      0.76      0.78       350

    accuracy                           0.81       771
   macro avg       0.80      0.80      0.80       771
weighted avg       0.81      0.81      0.80       771
```

Fig16 : SVM Algorithm

```
KNeighborsClassifier:train set
KNeighborsClassifier:Confusion Matrix:  [[810 135]
 [156 696]]
KNeighborsClassifier:Accuracy :  83.80634390651085
KNeighborsClassifier:Test set
KNeighborsClassifier:Confusion Matrix:  [[347  74]
 [ 82 268]]
KNeighborsClassifier:Accuracy :  79.76653696498055
AUC: 0.86
```

	precision	recall	f1-score	support
Yes	0.81	0.82	0.82	421
No	0.78	0.77	0.77	350
accuracy			0.80	771
macro avg	0.80	0.79	0.80	771
weighted avg	0.80	0.80	0.80	771

Fig17 : KNN Algorithm

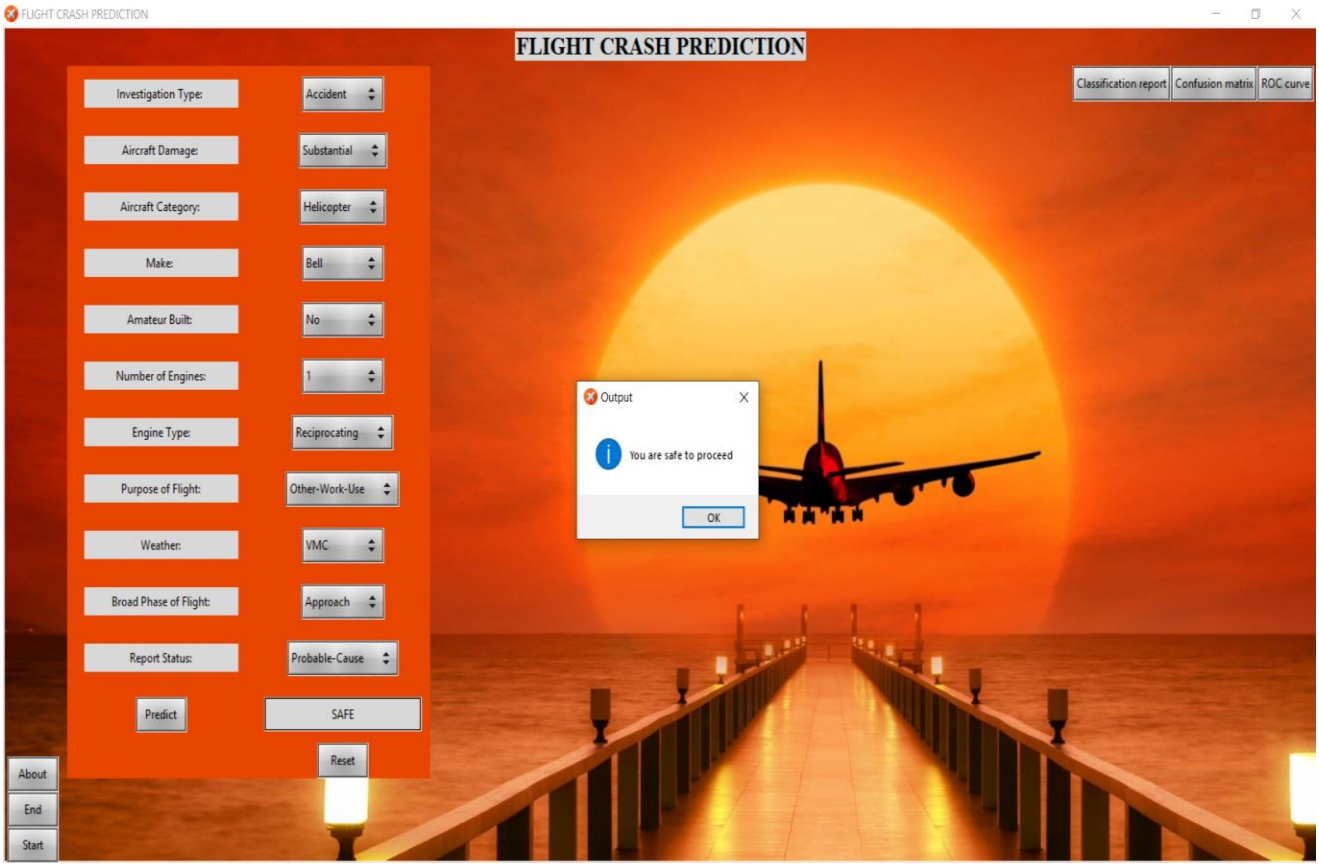

Fig18 : User Interface

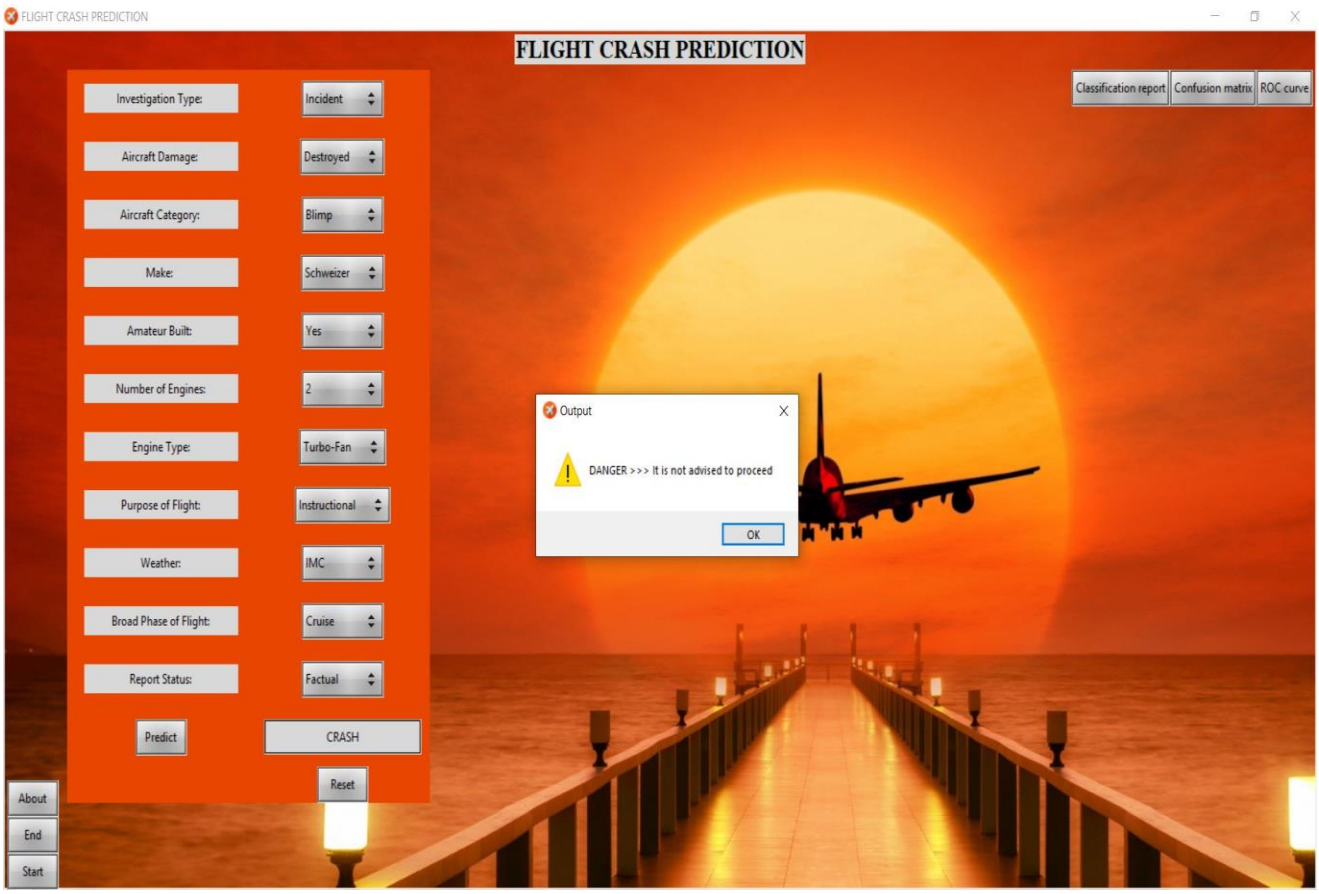

Fig19: User Interface

The user has to select the following values in order to predict whether the airplane is safe or not. The output of SAFE or CRASH will be produced by the system on the basis of the analysis.

CHAPTER 9. HARDWARE AND SOFTWARE REQUIREMENT

Hardware Requirement :

System : Pentium IV2.4

Hard disk: 250 GB

Software Requirement:

Operating system: WindowsXP/7

Application: Python IDE

Coding language: Python

CHAPTER 10. CONCLUSION

In this study, the classification is performed by four different kinds of classification algorithm. The dataset is tested on all the four types of classification algorithms and the patterns of every algorithm are evaluated. The performance of the entire system is essential which mainly depends on the accuracy of results. Accuracy of the XGBoost algorithm is the highest and hence, it is the algorithm selected mainly for the datasets belonging to the aviation industry. The paper has focused on the importance of feature selection and how the relevant features affect the accuracy of the prediction. All the redundant data from the dataset are eliminated. Hence, we extract the key attributes that will highly influence the result of the data and sort them in accordance to their ranking. The prediction is helpful for the company and the pilot to take all the necessary steps to avoid airplane crash. Hence, the classification algorithms have a major role in the data analysis and prediction.

CHAPTER 11.FUTURE SCOPE

The system is able to predict whether the airplane will be be “safe” or not. As a result, the delays of every airplane can also be predicted. The period after which an airplane has to go under the maintenance stage can also be included with the system. Hence, the system will be the one stop destination to check the flight delays, airplane crashes and the period after which the flight should undergo the maintenance phase.

CHAPTER 12. REFRENCES

[1] A.B. Arockia Christopher, S. Appavu, "Data Mining Approaches for Aircraft Accidents Prediction", Emerging Trends in Computing, Communication and Nanotechnology, 2013, Pages 25-26.

[2] https://github.com/AeroPython/flight-safety-analysis

[3] http://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx - dataset

[4] AshA, g.k, mAnJunATh, A.s. and JAyArAm, m.A. A comparative study of attribute selection using gain ratio and correlation based feature selection, Int J of Info Tech and Knowledge Management, July-December 2012, 2, pp 271-277.

[5] https://github.com/Data4Democracy/crash-model

[6] https://nycdatascience.com/blog/student-works/exploring- aviation-accidents-from-1908-through-the-present/

[7] dessureAulT, s., sinuhAJi, A. and colemAn, P. Data mining mine safety data, Mining Engineering Littleton, 2007, 59, (8), p 64. 7
pgs.

[8] D.K.Y Wong, D.E Pitfield, R.E Caves and A.J Appleyard, "The Development of Aircraft Accident Frequency Models", Safety and Reliability for Managing Risk – GuedesSoares, 2006 Pages 83-90.

[9] Michael G.Lenne, Paul M. Salmon, Charles C. Liu, Margaret Trotter, "A System Approach to Accident Causation in Mining", Accident Analysis and Prevention. Vol. 48, Sep 2012,Pages111- 117.

[10] Jin Tian, HaoRong, Tingd Zhao, "Hybrid Safety analysis method based on SVM and RST: An application to carrier landing of aircraft", School of Reliability and Systems Engineering, Vol. 80, Dec. 2015, Pages 56-65.

Paper Published

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

Airplane Crash Analysis and Prediction using Machine Learning

Likita J. Raikar[1], Sayali Pardeshi[2], Pritam Sawale[3]

[1,2,3] *Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, Maharashtra, India*

Abstract - *Airplanes are the most frequent mode of transportation in the present world. A single airplane crash leads to tremendous loss of human life. Safety is of prime importance since a huge number of people travel across the borders and within them. Abstracting data from a large database is always a difficult task. Data mining is a robust technology in order to extract the knowledge from raw data. Aviation systems take care of the minute precautions in order to prevent aircraft crashes. Factors causing and contributing to crashes needs to be understood studied and prevented in order to further minimize any kind of mishap. It is immensely difficult to find and extract the patterns of the factors due to very less amount of accident rates. In this research work, crash analysis and prediction is done. We have conducted the analysis of airplane crash data while, co-relating it with accidental information. To carry out this we have employed machine learning techniques. Machine learning helps in extracting the relationships between the various factors either affecting or non-affecting the crash to the general information of the airplane and as a result, patterns are formed. Many researchers, in recent times, have been using several machine learning techniques to help the aviation industry and the professionals in determining the hurdles. Supervised machine learning algorithms like SVM, K-NN, ADABoost and XGBoost are used for the purpose of prediction. The work has helped in improving the accuracy to a great extent.*

Key Words: **Airplane crash, safety, prediction, classification, KNN, SVM, ADABoost, XGBoost**

1. INTRODUCTION

The increase in technology has resulted in advancement in the systems that are used for the purpose of predicting and analyzing the existing records. Machine learning is a technology used for automatically making the system learn without explicitly performing the instructions. Large amount of data with respect to the past records are used for training the model. The attributes that contribute to the crash of a particular airplane are taken into the consideration. Classifying or predicting the number of individual data that is collectively associated is called as classification. The dataset is filtered and normalized. To predict the exact classifier class for each record in the dataset, it is the prime objective of classification. A classifier is able to tolerate noise and this is its essential quality. Classifier can handle quantitative data but it is very difficult to carry out this process. Safety is of prime concern for the applications in aviation industries. Companies carry out numerous investigations to create reports and collect information to justify the crash records and hence, this information can be in either conceptual form or structured/non-structured form. Feature selection is one of the prime stages in machine learning. Correct and relatable features should be selected with reference to the output in order to achieve highest accurate results. Data contains a lot of redundant values and these values should be filtered out in order to remove the irrelevant features. Irrelevant values only reduce the valuable assets from the output. The initial number of features is reduced. The new dataset has features that are highly appropriate for predicting the safety aspect of the aircraft. Fours algorithms namely Support Vector Machines (SVM), K-Nearest Neighbors (KNN), ADABoost, XGBoost are used for the purpose of classification. Precision, recall and f1 score are the performance factors used to improve the accuracy of the classification. It is very overwhelming for the humans to manage the datasets. Hence, these algorithms can efficiently explore them. KNN is an algorithm in which the similarity in the features is used for creating values of new data points. SVM is used in text, image classification and its main task is to sort the data into two categories and the sorted data is separated with margins away from one another.

2. PROBLEM STATEMENT

In today's world there are various types of predicting applications used to analyze and provide solutions to the future records. Airline industry is advancing day by day. Safety measures are taken at every provided situation by the companies. Also, the risk factors are examined for prevention of human loss. A single airplane crash can lead to a great loss of human life and property. There are numerous factors that leads to the airplane crash which are the airplane type, built of the model, weather conditions, make of the airplane, engine type, phase of the flight etc. Hence, taking all these factors into consideration, based on the details of a particular aircraft the analysis of the airplane crash is carried out. In order to predict whether the airplane is safe or at a risk the application is built where these functionalities are processed and the safety is predicted.

3. PROPOSED SYSTEM

The proposed system provides the person using the system to enter the specifications of the flight in order to know whether the flight is safe or has changes of a crash. Based on the past records of various airline companies the analysis and prediction of the given input is carried out. Machine learning is a strong and dependable technology in order to predict the values. Four algorithms are used and

based on every dataset the best algorithm would be used in order to predict the value. Every dataset varies majorly and hence, the algorithms used for the classification may also differ. To overcome this problem our system works accordingly.

3.1 Feature Selection

Redundancy, irrelevant data, noise etc. are removed or none the less reduced to a great extent from a huge dataset having multiple attributes. It comes under the pre-processing step in machine learning. The attributes that add value to the desired output are selected based on the specification of the aviation industry. Every attribute is taken into consideration and its importance is measured by relating it to the output required. The attributes that does not contribute to the result or are of least importance are deleted. The final dataset with the selected features are evaluated to check whether the subset is most relevant for prediction. Also, these attributes are sorted in a specific order from highest to lowest based on its importance on the prediction. As a result, only useful and relevant features are added hence, increasing the accuracy of the prediction.

3.2 Processing on the Dataset

Once the pre-processing is carried out the specific dataset is loaded into the system. In order to carry out the processing the dataset must be structured. The given dataset is cleaned that is all the missing values are removed by using attribute mean for all samples belonging to the same class also known as aggregation. The dataset is now ready to be loaded. For the purpose of performing machine learning algorithms on this dataset the data has to be split into training data and testing data. The appropriate split ratio for the dataset is 70:30, 70% for training data and 30% for testing data. Now the dataset is ready for the algorithms to carry out its processing.

3.3 Train and Test the Classifier

The model views and learns from the training phase. However, the testing phase is used to evaluate the model based on its performance. The records from the training model should not be included in the testing otherwise it won't produce correct results. Also if the dataset is unbalanced it would create the problem of over-fitting and under-fitting. Hence, these phases essentially produce the output that is the actual prediction - "safe" or "crash". All the four algorithms are used for training and testing and the best algorithm with highest accuracy is used for the purpose of prediction

3.4 User Interface

The user enters all the parameters which are the specifications of the airplane that they want to know about. The model takes into consideration all these features and carry out the prediction for these instances. The output of the prediction is either "safe" or "crash" based on the details of the airplane.

3.5 Flowchart

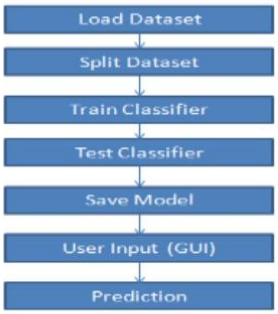

Fig -1: Flowchart of the system

4. CLASSIFICATION ALGORITHMS

4.1 KNN (K-Nearest Neighbors)

KNN algorithm predicts the most comparable training value by assigning the class mark or continuous target value. This can be used for classification and regression related problems. KNN Algorithm is an algorithm of supervised machine learning that is fast and simple to implement. The classification method is non- parametric. It is also a learning algorithm based on an instance, where the function is locally approximated.

KNN classification is implemented as follows –

- Determine the value of the distance between the test data point and all the marked data points.
- Order the named data points in the increasing order of certain distance metrics.
- Pick the data points labeled with the top k and look at the class labels.
- Find the class label most of these data points called k have and assign it to the test data point.

Following are some things one should consider:

- Parameter selection
- Presence of noise
- Feature selection and scaling
- Curse of dimensionality

4.1.1 Parameter Selection

It is the data upon which k's best choice depends. High k values reduce the effect of noise on classification but make the boundaries of the decisions less distinct between groups. Smaller k values tend to be influenced by the noise, with strong class separation.

4.1.2 Feature Selection and Scaling

It is important to remove unnecessary traits. If the number of characteristics become too high and the feature

gets assumed it requires highly redundant extraction. When features are carefully selected, the classification will always give better results.

4.2 SVM (Support Vector Machine)

Every data item is plotted with its associated value. Classification is carried out in such a manner that it separates the given classes, here each separated area is known as the hyper-plane It is very essential to group the data items in the dataset to their right hyper-plane. This process is known as the identification of the hyper-plane.

4.3 ADABoost

Weak classifiers should be converted into strong classifiers, hence boosting came into picture in machine learning. Weak classifiers are always beneficial than random guesses. As a result, these classifiers prove to be robust and solve the problem of over-fitting when applied on large dataset. Hence, the weak ones provide efficient results than random values. A single feature is focused upon which has any random kind of threshold applied on it. If the feature is above the threshold than predicted, it belongs to positive otherwise belongs to negative. AdaBoost stands for 'Adaptive Boosting' which transforms weak learners or predictors to strong predictors in order to solve the problem of classification. For classification, below is the final equation:

$$F(x) = sign(\sum_{m=1}^{M} \theta_m f_m(x)).$$

Here fm designates the classifier mth weak and m represents its corresponding weight.

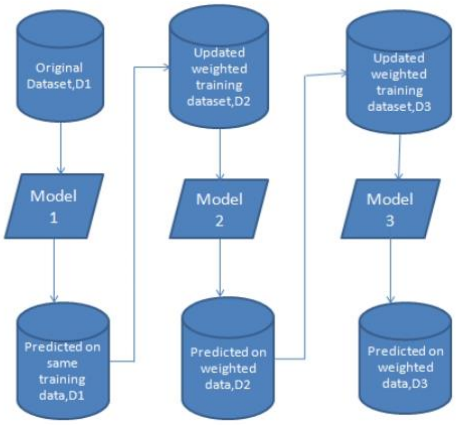

Fig -2: ADABoost Classifier and Flowchart

4.4 XGBoost

XGBoost is a machine learning algorithm that is used for the implementation of gradient boosting decision trees. It is used for classifying non-structured or semi-structured data. This algorithm boosts the speed and improves the performance of the models. The tree model as well as the linear model is included in XGBoost. Since these two models work in a single algorithm parallel computing is enabled on every individual machine. The difference between the target and the predicted outputs are minimized. New trees are iteratively added that are used to predict the errors from the previous trees. Hence, it is known as gradient boosting. New models are added to minimize the loss.

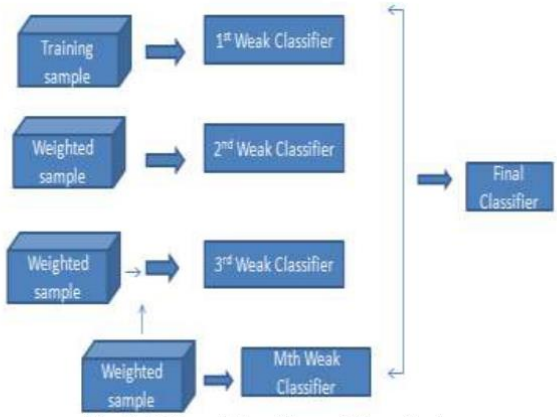

Fig -3: XGBoost Classifier and Flowchart

5. DATA ANALYSIS

Companies produce their records for each airplane and hence, these records are collected to form a dataset wherein details about every airplane module are stored hence, a huge dataset comprising of thousands of records is formed. Such datasets are loaded with a large amount of attributes. Numerous attributes are required in order to justify the airplane incident. Hence, the data in these attributes is unstructured and textual. The dataset has to be brought and cut down in such a manner where the classification algorithms can be performed. The attributes that define the output of the prediction are the target attributes. The target variables depend upon the safety of the airplane.

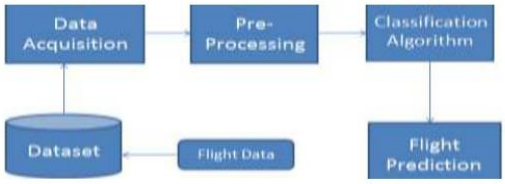

Fig -4: Flight Prediction Flowchart

Classification algorithms carry out the process of data analysis. Since accuracy is the main performance measure, the best classification algorithm is selected is for the purpose of prediction depending on the dataset. The below table depicts the accuracy of each algorithm.

Algorithms used	Accuracy
KNN	79.7%
SVM	80.54%
ADABoost	83.53%
XGBoost	86.12%

Table -1: Accuracy of algorithms

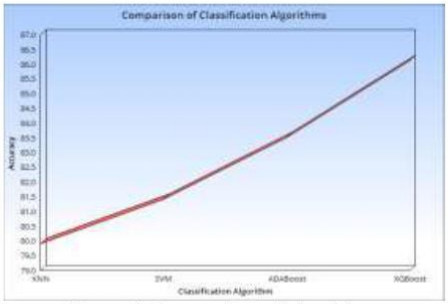

Chart -1: Comparison of algorithms

6. RESULTS

The prediction of the system helps the user in taking the necessary precautions to prevent the mishap or any airplane crashes. The administration department thus becomes aware of the possible difficulties and hurdles that might come along the way. As a result, the elementary steps taken will help to eradicate any crashes that might occur thus leading to minimize the loss of property and life. Below are all the results of the application.

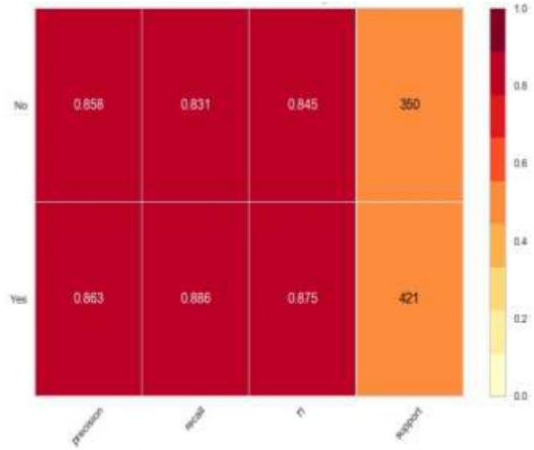

Chart -2: Performance metrics

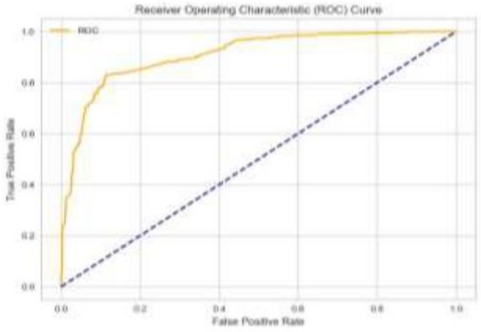

Chart -3: ROC Curve

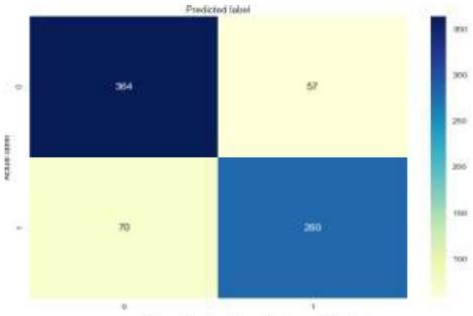

Chart -4: Confusion Matrix

```
XGBOOST:train set
XGBOOST:Confusion Matrix:  [[820 125]
 [150 702]]
XGBOOST:Accuracy :  84.69671675013912
XGBOOST:Test set
XGBOOST:Confusion Matrix:  [[373  48]
 [ 59 291]]
XGBOOST:Accuracy :  86.12191958495461
AUC: 0.91
              precision    recall  f1-score   support

         Yes       0.86      0.89      0.87       421
          No       0.86      0.83      0.84       350

    accuracy                           0.86       771
   macro avg       0.86      0.86      0.86       771
weighted avg       0.86      0.86      0.86       771
```

Fig -5: XGBoost Accuracy

```
Ada Boost:train set
Ada Boost:Confusion Matrix:  [[815 130]
 [172 680]]
Ada Boost:Accuracy :  83.1942125765164
Ada Boost:Test set
Ada Boost:Confusion Matrix:  [[364  57]
 [ 70 280]]
Ada Boost:Accuracy :  83.52788586251621
AUC: 0.91
              precision    recall  f1-score   support

         Yes       0.84      0.86      0.85       421
          No       0.83      0.80      0.82       350

    accuracy                           0.84       771
   macro avg       0.83      0.83      0.83       771
weighted avg       0.84      0.84      0.83       771
```

Fig -6: ADABoost Accuracy

```
Suppor_Vector_Machine:train set
Suppor_Vector_Machine:Confusion Matrix:  [[800 145]
 [189 663]]
Suppor_Vector_Machine:Accuracy :  81.41346688925988
Suppor_Vector_Machine:Test set
Suppor_Vector_Machine:Confusion Matrix:  [[354  67]
 [ 83 267]]
Suppor_Vector_Machine:Accuracy :  80.54474708171206
AUC: 0.85
              precision    recall  f1-score   support

         Yes       0.81      0.84      0.83       421
          No       0.80      0.76      0.78       350

    accuracy                           0.81       771
   macro avg       0.80      0.80      0.80       771
weighted avg       0.81      0.81      0.80       771
```

Fig -7: SVM Accuracy

```
KNeighborsClassifier:train set
KNeighborsClassifier:Confusion Matrix:  [[810 135]
 [156 696]]
KNeighborsClassifier:Accuracy :  83.80634390651085
KNeighborsClassifier:Test set
KNeighborsClassifier:Confusion Matrix:  [[347  74]
 [ 82 268]]
KNeighborsClassifier:Accuracy :  79.76653696498055
AUC: 0.86
              precision    recall  f1-score   support

         Yes       0.81      0.82      0.82       421
          No       0.78      0.77      0.77       350

    accuracy                           0.80       771
   macro avg       0.80      0.79      0.80       771
weighted avg       0.80      0.80      0.80       771
```

Fig -8: KNN Accuracy

Fig -9: User Interface

The user has to select the following values in order to predict whether the airplane is safe or not. The output of SAFE or CRASH will be produced by the system on the basis of the analysis.

7. CONCLUSIONS

In this study, the classification is performed by four different kinds of classification algorithm. The dataset is tested on all the four types of classification algorithms and the patterns of every algorithm are evaluated. The performance of the entire system is essential which mainly depends on the accuracy of results. Accuracy of the XGBoost algorithm is the highest and hence, it is the algorithm selected mainly for the datasets belonging to the aviation industry. The paper has focused on the importance of feature selection and how the relevant features affect the accuracy of the prediction. All the redundant data from the dataset are eliminated. Hence, we extract the key attributes that will highly influence the result of the data and sort them in accordance to their ranking. The prediction is helpful for the company and the pilot to take all the necessary steps to avoid airplane crash. Hence, the classification algorithms have a major role in the data analysis and prediction.

8. FUTURE SCOPE

The system is able to predict whether the airplane will be be "safe" or not. As a result, the delays of every airplane can also be predicted. The period after which an airplane has to go under the maintenance stage can also be included with the system. Hence, the system will be the one stop destination to check the flight delays, airplane crashes and the period after which the flight should undergo the maintenance phase.

ACKNOWLEDGEMENT

We are thankful to our guide Prof. Neha Kudu who supported and guided us in every phase of the project.

REFERENCES

[1] A.B. Arockia Christopher, S. Appavu, "Data Mining Approaches for Aircraft Accidents Prediction", Emerging Trends in Computing, Communication and Nanotechnology, 2013, Pages 25-26.

[2] https://github.com/AeroPython/flight-safety-analysis.

[3] http://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx-dataset

[4] AshA, g.k, mAnJunATh, A.s. and JAyArAm, m.A. A comparative study of attribute selection using gain ratio and correlation based feature selection, Int J of Info Tech and Knowledge Management, July-December 2012, 2, pp271-277.

[5] https://github.com/Data4Democracy/crash-model

[6] https://nycdatascience.com/blog/student-works/exploring-aviation-accidents-from-1908-through-the-present/

[7] dessureAulT, s., sinuhAJi, A. and colemAn, P. Data mining mine safety data, Mining Engineering Littleton, 2007, 59, (8), p 64. 7 pgs.

[8] D.K.Y Wong, D.E Pitfield, R.E Caves and A.J Appleyard, "The Development of Aircraft Accident Frequency Models", Safety and Reliability for Managing Risk – GuedesSoares, 2006 Pages 83-90.

[9] Michael G.Lenne, Paul M. Salmon, Charles C. Liu, Margaret Trotter, "A System Approach to Accident Causation in Mining", Accident Analysis and Prevention. Vol. 48, Sep 2012,Pages111-117.

[10] Jin Tian, HaoRong, Tingd Zhao, "Hybrid Safety analysis method based on SVM and RST: An application to carrier landing of aircraft", School of Reliability and Systems Engineering, Vol. 80, Dec. 2015, Pages 56-65.

Certificates


e-ISSN: 2395-0056 p-ISSN: 2395-0072
International Research Journal of Engineering and Technology (IRJET)
(An ISO 9001 : 2008 Certified Journal)
Is hereby awarding this certificate to
Likita J. Raikar
In recognition the publication of the manuscript entitled
Airplane Crash Analysis and Prediction using Machine Learning
published in our Journal Volume 7 Issue 3 March 2020
Editor in Chief
E-mail : editor@irjet.net
Impact Factor : 7.34
www.irjet.net


e-ISSN: 2395-0056 p-ISSN: 2395-0072
International Research Journal of Engineering and Technology (IRJET)
(An ISO 9001 : 2008 Certified Journal)
Is hereby awarding this certificate to
Sayali Pardeshi
In recognition the publication of the manuscript entitled
Airplane Crash Analysis and Prediction using Machine Learning
published in our Journal Volume 7 Issue 3 March 2020
Editor in Chief
E-mail : editor@irjet.net
Impact Factor : 7.34
www.irjet.net

e-ISSN: 2395-0056 p-ISSN: 2395-0072

International Research Journal of Engineering and Technology (IRJET)

(An ISO 9001 : 2008 Certified Journal)

Is hereby awarding this certificate to

Pritam Sawale

In recognition the publication of the manuscript entitled

Airplane Crash Analysis and Prediction using Machine Learning

published in our Journal Volume 7 Issue 3 March 2020

Editor in Chief

E-mail : editor@irjet.net

Impact Factor : 7.34

www.irjet.net