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

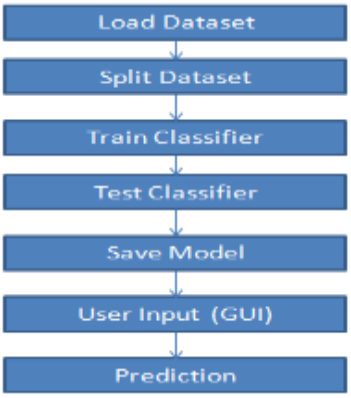

Fig -1: Flowchart of the system

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.

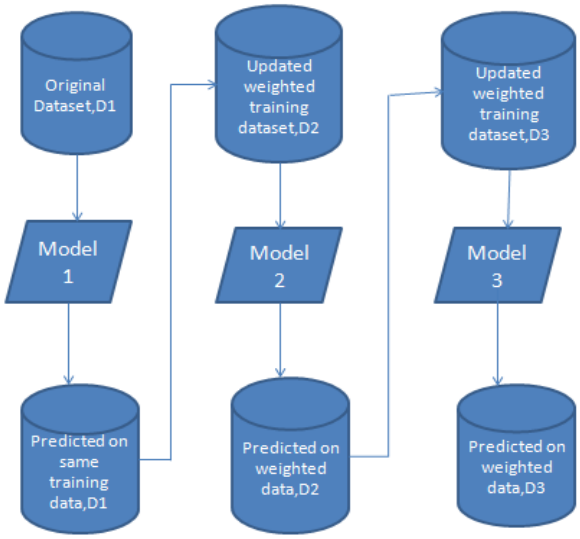

Fig -2: ADABoost Classifier and Flowchart

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.

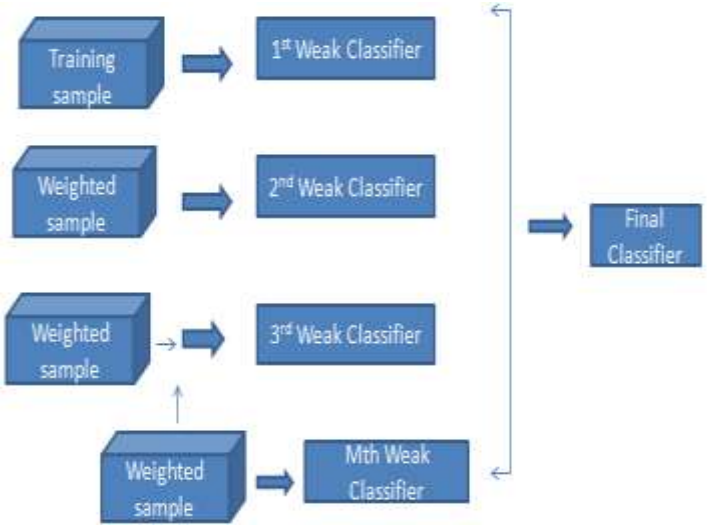

Fig -3: XGBoost Classifier and Flowchart

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.

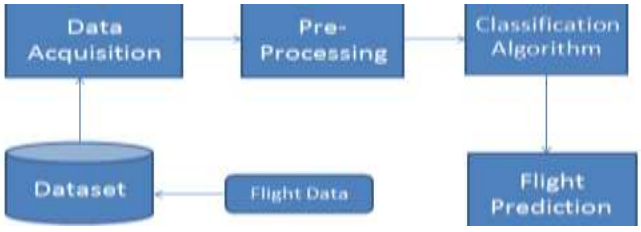

Fig -4: Flight Prediction Flowchart

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

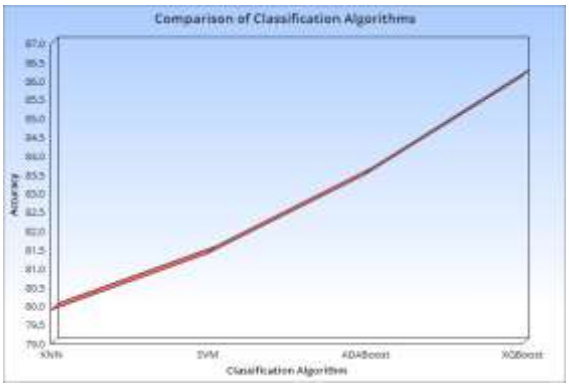

Chart -1: Comparison of algorithms

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.

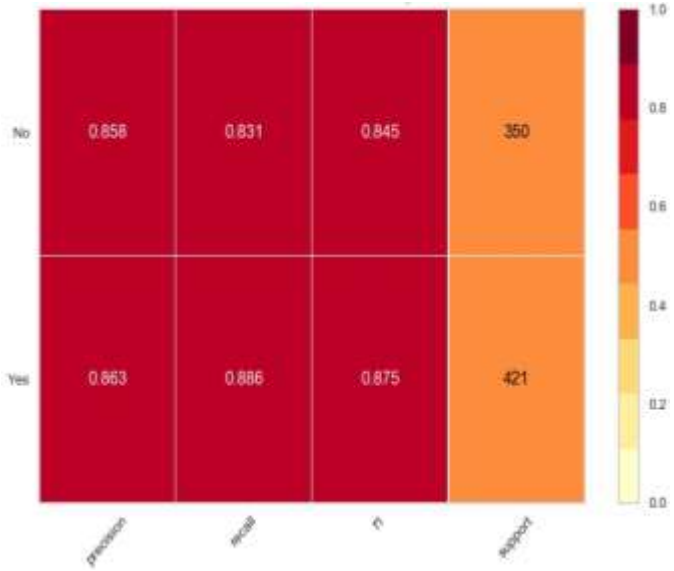

Chart -2: Performance metrics

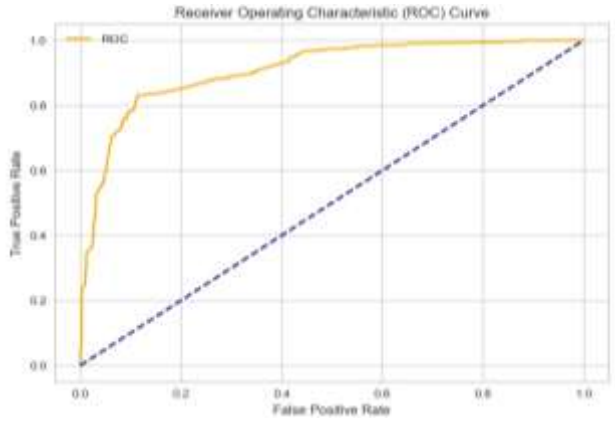

Chart -3: ROC Curve

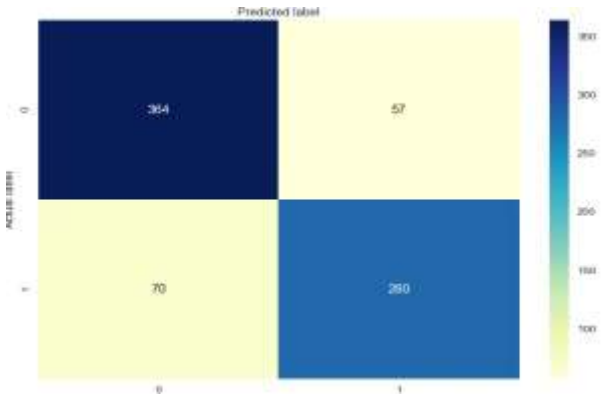

Chart -4: Confusion Matrix

```
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, 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.