PIA Accidents Analysis Using Naïve Bayes Classifier

Jamsher Bhanbhro
Computer Systems Engineering
Mehran University of Engineering and Technology, Jamshoro
Jamshoro, Pakistan
jamsherbhanbhro@gmail.com

Faiez Yousuf
Computer Systems Engineering
Mehran University of Engineering and Technology, Jamshoro
Jamshoro, Pakistan
faiezjaliawala@gmail.com

Sanam Narejo
Computer Systems Engineering
Mehran University of Engineering and Technology, Jamshoro
Jamshoro, Pakistan
sanam.narejo@faculty.muet.edu.pk

Mursal Furqan
Computer Systems Engineering
Mehran University of Engineering and Technology, Jamshoro
Jamshoro, Pakistan
mursalfurqan@gmail.com

***Abstract* — There have been many airline accidents especially in Pakistan, the nation has lost many precious lives and had a bad impact on the economy. This research study is about finding reasons and statistical analysis of airline accidents. There is not recent study available in this area from Pakistan Region, although the research studies are done on accident analysis of the international airline. The data science algorithms and the statistical analysis techniques are used to find out the reasons and data abnormalities. Mostly the model used in this study is based upon a supervised learning algorithm, i.e., Naïve based classifier (because it is very efficient and produces better predictions when there is fewer data in the dataset and in this case, dataset contains only 22 data entries) since it is used to classify analysis. Histogram analysis is used for graphical representation. The dataset used in this research is data of all airline accidents in Pakistan since 1947. The study gives the best results and helps to find out causes and suggests how accidents can be reduced.**

Keywords — Supervised Learning, Histograms, Statistics, Naïve Bayes Classifier, Datamining, K-mean Clustering's

I. Introduction

Airlines are the major source of international transports. They represent the countries on the international levels and play a vital role to develop economies of the countries. However, just one airline crash creates bad effects on the environment, population, economy, and nature. Many crashes occurred in Pakistan, even in the modern era despite having state of the art technical tools. The main reason to conduct this research is to find out tentative reasons because of which the country is facing this issue.

Flying misfortune is caused by many reasons like a pilot's mistake, bad weather, or engine malfunction. Air Crashes happen lesser compared to other modes of transportation However, air travel is the safest and fastest among all other modes of transportation but why do they always cause a bad impact socially and economically if crashes.

Air crash analysis is a major and advanced research area, much research has been done in this field. Many researchers used different techniques to find out the causes of the airline crashes. The major techniques used for these investigations are simple statistics, digital image processing, cloud computing, data mining. These research help to prevent future accident. Our research is on analysis of Pakistan Airline (PIA) crashes. There are almost 20 major air crashes that have happened since 1955. It causes a big loss of PIA financially and socially.

Our focus is to analyze the causes of these crashes and find the similarities among these crashes. Naive Bayes technique is used to find out similarities among these crashes. Naive Bayes technique is a supervised machine learning technique in which the label data set is used to train the model. These label data sets help the model to predict the output. also used a histogram to visualize the causes of air crashes. A histogram is a powerful statistical technique that organizes the numerical data and displays the data graphically.

PIA has lost more than 20 aircraft and almost 1,000+ human lives, and a horrific crash happened recently this year. The verbal examination by the authorities provides explanations such as operators' errors or pilot errors. But there is no improvement, even in this technical era, crashes are taking place. There is not any research done about the PIA accident analysis in history, that can help to understand the actual causes of the crashes This study therefore focuses primarily on the scientific analysis of the causes and finding the results. We have used histogram techniques in the first step to detect the anomalies in the dataset. And then we have again used a histogram to represent the causes of air accidents in the visualization form. The line histograms, circle histogram is used to visualize the dataset. And the final step the Naïve Bayes classifier is used.

II. Literature Review

There is a lot of research being done at the international level, as mentioned above. Several studies are available in which various methods are used to analyze the crash causes. But all these are research are analysis of the accidents of a specific

ISBN-978-969-23372-1-2

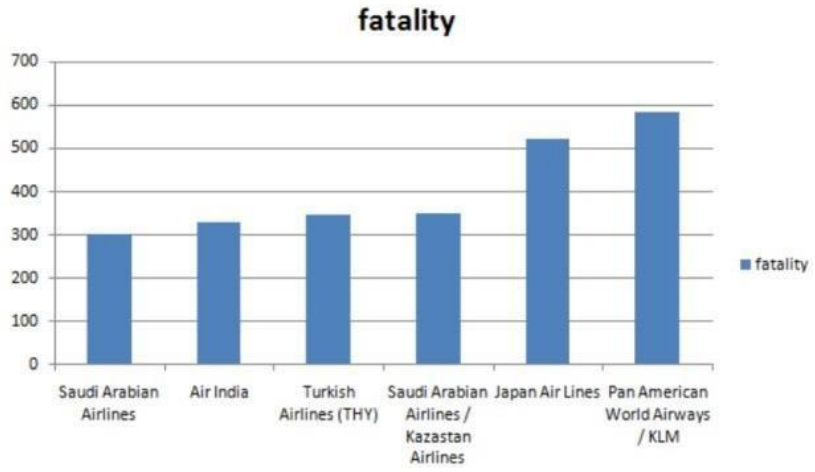

Fig. 1 Airline Crash research using data mining techniques [1].

country. These research studies analyze the causes using Machine learning algorithms or by using simple statistical comparison. The studies find out fatalities and causes by specific airlines and specific reasons. Mostly researches uses both scientific and statistical analysis of the crashes using machine learning algorithms, EDA techniques, and data mining techniques are included below.

2.1. Air Crash Analysis Using Data Mining Algorithms

Indian researchers wrote this paper and used cosine similarity and k-mean clustering to figure out unexplained trends in the datasets. The dataset includes details of worldwide airline accidents from 1908-2008. This research uses data mining and image processing algorithms [1]. Data mining was used for data cleaning and K-mean clustering was used for finding out similar clusters or causes. The result is shown in figure1.

In the above figures, the RapidMiner algorithm of the K-mean clustering is used and the value of k or clusters is considered as 5. This study mainly discusses the fatalities rate concerning different crashes reasons and finds fatalities ratio for the countries. The above figure states the airlines which have faced unfortunate or represents the histogram of the world's well-known airlines and the number of fatalities due to crashes. This figure (1) states that most fatalities are occurred because of the Pan American World Airways airplane crashes.

2.2. Air crash causes prediction using machine learning

This is one of the most advanced and recent works done in march of 2020, in this study the Kaggle dataset is used and that is a collection of the accidents from 1981 to 2019 across the globe. In seven categories, the crash causes are specified, and multiple machine learning algorithms are used. The datasets contain 17 columns or fields; the training was done by using the IBM SPSS modeler. those are: (i) CHAID (ii) Neural Network (iii) XGBoost Tree(iv) Tree-AS(v) XGBoost Linear 75% partitioning was used for training and 25% for the testing. In this study, the result was compared to three different models in the terms of accuracy. The study was analyzed by using Neural network, Quest 1 and C&R Tree 1 models whose comparative result is also mentioned in a subpart of Fig.2. The Fig.2 represents the result of this research study [2]. The Figure- 2 is the comparison of the model's result and that concludes the neural net 1 model is the best. This research study has used histograms and the bars of histograms do the comparisons between the number of accidents and reasons.

Fig.2 Airline crash prediction using Machine Learning [2]

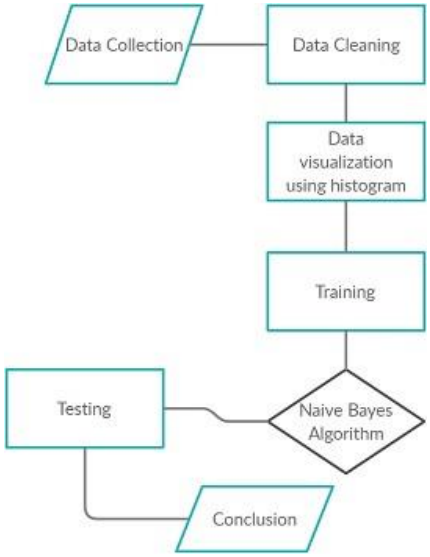

Fig. 3 Process Flow Diagram

2.3. Airline crash prediction due to a bird attack using machine learning

The main focus of this research is on airline accidents caused by bird strikes. A small bird strike may cause a crash due to the relative speed of the aircraft. The National Transportation Safety Board (NTSB) data is used as a training dataset. For the development of a prediction model, machine learning methods decision trees, and Bayesian classifications. The outputs of the prediction vary from 80% to 90% [6].

2.4. An Empirical Analysis of Reasons of Air Crashes using transport

This paper was published in January 2014. The research has found out the region, country, or area in which the most air crashes occur. This study is a little bit simply because, in this, the data sets are taken from different years, and manual or mathematical comparison is done. The conclusion says that North America is the region in which most accidents are occurred [3].

2.5. Textual Indicator based aviation extraction from air crash records

This paper is based on text mining techniques, uses data from the database of the National Transportation Safety Board (NTSB). First creates a classification model based on the NTSB database and different causes (taken from text). the comparative study states that Support Vector machine and deep neural network algorithms are better than all other classification algorithms. After executing results from the above algorithm study further considers specific keywords and finds textual indicators on those keywords to find deeper reasons for accidents [7].

III. Proposed Methodology

Figure-3 is the Flow chart or the methodology of this study. Mainly flow chart specifies the steps that are sequentially taken to develop and demonstrate the model that can train the data in the right way, and the method or the classifier that is to be implemented must be fully capable of producing the 100% accurate prediction.

The step taken of data cleaning is to remove abnormalities from the datasets, because that produce problems in testing, as well as training, i.e. the wrong training, will always produce an incorrect prediction. Visualization is just representing the drawbacks, features, and abnormalities in the datasets. Naïve Bayes classifier is used for the classification of the dataset and it is the method that is implemented to produce our expected results.

The first and important step is to collect data and construct a full dataset so that effective prediction can be produced by

Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/ln	Aboard	Fatalities	Ground
2/25/1956	17:18	unkown	PIA		Demonstration	PESSENGER		1	3	3	
7/1/1957	6:30	bangal	PIA	PIA-12	Test flight	PESSENGER			24	24	
8/14/1959		karachi	PIA	-		Training			1	1	
5/20/1965		Cairo	PAF	Boing 707		PESSENGER			124	124	
9/9/1965	18:30	lowarpass	PIA	PIA		PESSENGER			26	22	
10/17/1970	10:30	Islamabad	PIA	PIA f27		PESSENGER			31	31	
10/17/1972	10:30	Islamabad	PIA	PIA f27		PESSENGER			31	31	
3/5/1986	1:00	Pishawer	PIA	PIA f28		PESSENGER			54	13	
9/3/1988	15:20	Bahalawarpur	US	C130		military			30	30	
7/28/2003		Kohat	PAF	foker F27		military			17	17	
9/24/2003	1:00	Arabian sea	PIA	402B		PESSENGER			8	8	
10/1/2006	23:45	Multan	PIA	f27		PESSENGER			43	43	
11/21/2010		Islamabad	AirBlue	A321231		PESSENGER			152	152	
11/28/2010	23:45	karachi	JSAIR	beechart1900c		cargo			21	21	
3/4/2010		karachi	sunway	IL76		PESSENGER			6	12	
3/30/2012		Islamabad	Bhoja Air	Boing 737		PESSENGER			127	127	
7/7/2015		lahore	PIA	sh737		PESSENGER			40	40	
7/8/2016		Islamabad	PIA	ATR42500		PESSENGER			47	47	

Fig.4 A sample of dataset used in this study [4].

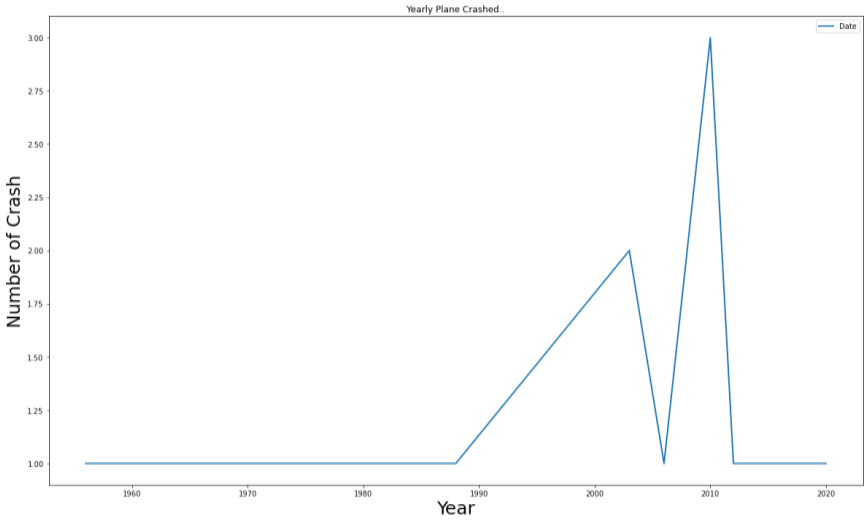

Fig. 5 Visualization of Dataset

algorithms. The algorithms operate sequentially. In the first step, the data is visualized by using histograms (circle and linear). Then naïve Bayes classifier is used for classification of classes in the dataset

3.1. Dataset

I manually collected data from various websites and newspaper articles [4]. There are twelve columns in datasets and almost 22 data entries are in this paper. The data is collected from an investigation report that was submitted at the national assembly and some of it was collected from the nation.pk essay.

3.2. Histogram Visualization

Histograms are also used to represent statistical data graphically. Linear histograms, Line histograms and Circle histograms, are used for this approach. Matplot Library of python is providing the histograms functionality. Next section is about data visualization.

3.2.1. Visualize the Dataset

Figure-5 visualizes the Pakistan airline crashes per year (in numbers). Here, a simple histogram is used on the y-axis number of crashes is defined and, on the x-axis, years are labeled. The data visualization states that most crashes occurred between 2006- 2012.

3.2.2. Defines Year Wise Crashes and Fatalities

Figure-6 is visual representation of the survival and fatalities of the peoples in crashes and specific years. Here is also histogram analysis is used and tells the analysis with respect to survival and fatalities in years. As the first histogram figure states that more crashes occurred in 2006-2012 hence fatalities in this duration has a high value.

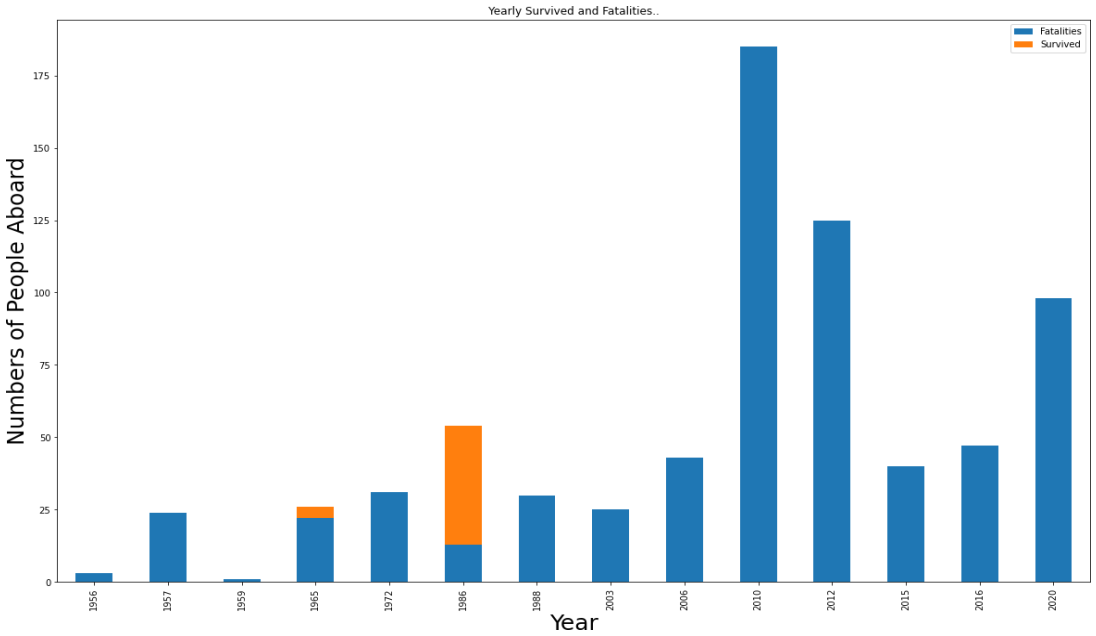

Fig.6 Survival and Fatalities in Years

3.2.3. Shows the percentage of Military and Non-Military planes crashes

The Figure-7 describes the percentage of military and non-military crashes according to the dataset.

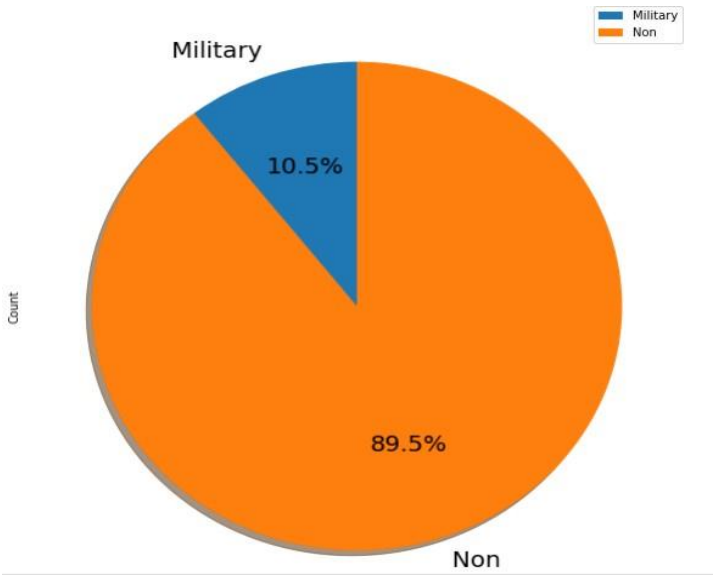

Fig. 7 Military and Non-Military Airline crashes

3.2.4. Shows Crashes Because of Operators

This Figure-8 simply tells the airlines that are mostly crashed because of operator.

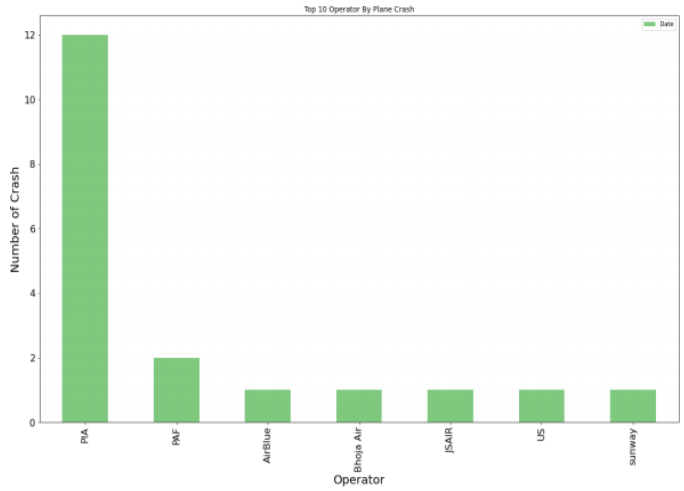

Fig.8 Fatalities with respect to Operators

IV. Algorithms

This paper uses a supervised algorithm called naïve Bayes, used for the classification of the causes and by training the causes the algorithm will analyze result from the dataset.

4.1. Naïve Bayes Classifier

In simple words, a Naive Bayes gives best result if we have more unrelated data. The dataset is trained and particular terms (Causes names) or triggers are selected and classified in classes. The Following equation(equation-1) is used for the classification of the dataset and that uses the training dataset which is discussed in section 4.1.1.

Naïve Bayes Classifier is used for the classification of the dataset. The classification is done by using the training data values.

NaiveBayesClassifier(trainData) → equation-1

Equation-1 states that a classifier which classifies the datasets according to the training data values. The dataset is classified according to the reasons that are used in the training data array called trainData that is discussed in the upcoming section 4.1.1.

4.1.1. Training

First, a summary text file that consists of causes is built and then an extra column of classification is added in the dataset and then the complete dataset is examined or tested upon that. The results of the matching of the words from a summary column of the dataset with the text that is used for training.

The Naïve Bayes classifier compares the testing dataset with a training data set and then tells the percentage in the output. The training of the dataset is done in the following way. The next paragraph is about the training model and the logical way to implement the training and taking special data values from the datasets to predict the actual causes of the crashes.

An array consider trainData is used to store the training data values from the dataset, this array will store the training values by using labels and those labels are the reasons that are stored in the last column of the dataset. The program logic depends on the Summary(last) column of the dataset if the labels that are mentioned, are not available in the summary column than the training model (naïve Bayes) will auto consider it as unknown reasons. If the reasons are found from the dataset's Summary column, then those will be stored in the traidData array. Here the training also includes the actual cause (the thing behind the reason) because of which the accident occurred, let say an accident occurred because of weather that is a reason, but behind this reason, there can be bad, thunderstorm, this is also taken consideration in the training of the datasets and these are called the expressions. These expressions are very important for the model to predict the accurate and exact output.

When the model predicts a reason for the crash it gives the direction and the model predicts according to the expression then it gives the exact point in the direction told by the model

i.e. Plane crashed due to caught on fire, the prediction will also tell why the fire was caught on the plane it may be because of terrorists, it can be shot down, or it can be by a bird strike. The expressions can be called subtype of the reasons. The difference between the reasons and the expressions are discussed below. The reasons for the crashes

of the airlines are stated in the array that is used for the training purpose.t

reasons *= ['weather', 'fire', 'shot down', 'stall/runway', 'pilot/crew error', 'systems failure'].*

The expressions are the causes of the reasons, it means if an airplane caught the fire now why it caught fire that is the expression and the array of expressions used in the above example is:

expresion *= ['((poor | bad). * (weather | visibility) | thunderstorm | fog)' , '(caught fire) | (caught on fire)', '(shot down) | (terrorist) | (terrorism)', '(stall) | (runway)', '(pilot | crew) (error | fatigue)', '(engine. * (fire | fail)) | (structural fail) | (fuel leak) | (langing gear) | (turbulence) | (electrical) | (out of fuel) | (fuel. * exhaust)']*

4.1.2. Testing

Testing is done on the complete dataset of the Pakistan Airlines accidents and the result of the testing is discussed in the Result section of this paper. The testing is mostly done by finding out the locations and causes and by comparing the output with the dataset taken, Results sections truly demonstrate the testing.;

V. Results

After using the Naïve Bayes classifier, the result of this study is found out that most accidents are occurred either because of the weather or because of system failure. The 18.2% of crashes occurred because of landing or runway. It concludes that for safety purposes it is better not to fly in bad weather. The following Figure-9 tells the percentage of the accidents caused. The analysis also tells that most crash incidents occurred in ISLAMABAD.

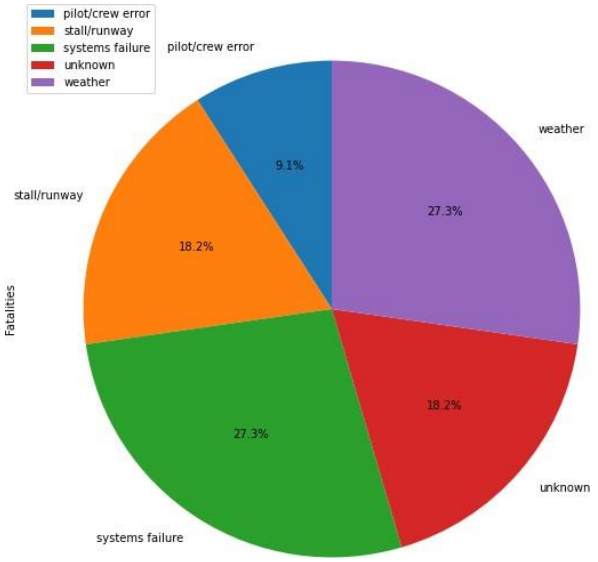

Fig. 9 Percentage of Crashes with respect to labels

VI. Conclusion

Through this analysis, the outcome states that airline accidents occurred mainly because of the operator errors and because of landing or take-off. The study also notes that the high altitude weather in Islamabad can also be dangerous. The accidents can be avoided by making the runways better and by using the advanced technology and tools for the manufacturing of the airplanes, and the weather cause can be avoided by using the modern advanced technology to predict the harmful weather. This study also concludes that untrained pilot or staff is also a sign for these incidents. The study describes well trained military pilots and their planes are safest in Pakistan. This research can be used for all the other specific airlines and also can be used on whole world airlines. This model specifically works on the datasets and the data that is required discussed in section 3.1, if it is available then this model can be used for any airline crash analysis.

VII. Future Work

The future work for this study will be to find out the percentage of untrained pilots because of which the accidents have occurred, the future work will also be research upon the system failure and the find out the actual causes of the system failures. In future work, this also can be calculated the hardware / the engine manufacturing companies whose most engines failed.

REFERENCES

[1] Shagun sharma and A Sai Sabitha(2016). Flight crash analysis using data mining techniques, 1st India International Conference on Information Processing.

[2] Ved Parkash, Sajid Mansoori, Jitendra shreemali, payal paliwal (March 2020). Predicting causes of Airplane crashes using machine learning algorithm, Volume 8 issue 6s IJRTE.

[3] Mobalji Stephens and Wilfred Isiomia . (Jan-2014). An Empirical analysis of crashes of airline crashes from transport management. Mediterranen Journel of Science.

[4] Express Tribune, 2016, Timeline of major air crashes in Pakistan

[5] Kunimitsu Iwadare, Tatsuo Oyama (2015), Statistical Data Analyses on Aircraft Accidents in Japan: Occurrences, Causes and Countermeasures

[6] S. Nimmagadda, S. Sivakumar, N. Kumar and D. Haritha, "Predicting Airline Crash due to Birds Strike Using Machine Learning," 2020 7th International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2020, pp. 1-4, doi: 10.1109/ICSSS49621.2020.9202137.

[7] Hu, Xue & Wu, Jun & He, Jingrui. (2019). Textual Indicator Extraction from Aviation Accident Reports. 10.2514/6.2019-2939.