International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

Air Crash Investigation and Safety Promotion using Data Mining Technique

Vasudeva Rao P V [1], Jyothi Mishel Veigas [2], Kajal B Surve [3], Anvitha P [4], Dhanyashree [5]

[1]*Assistant Professor, Dept. of Information Science Engineering, Sahyadri College of Engineering and Management, Karnataka, India*
[2,3,4,5]*Student, Dept. of Information Science Engineering, Sahyadri College of Engineering and Management, Karnataka, India*

Abstract - *Air crashes are the most uncertain mishaps that take place in modern era. Once the mishap happens, a detailed case study has to be done for determining its cause, considering various factors like mechanical failure, crew error, weather conditions, sabotage etc. The proposed project mines the available air crash records using data mining techniques such as clustering and hierarchical techniques and checks for the possibilities that usually cause the malfunctioning of an aircraft. It will determine the factors which led to the aircraft mishap and list the safety measures to be taken to prevent such incidents from occurring again.*

Key Words: **Air crash, failure, safety measure, K-means, mishap, clustering techniques, tokenization, aircraft, prediction, worst operators.**

1. INTRODUCTION

1.1 Data Mining

Data Mining is the process of automatically identifying the useful information which are extracted from large data repositories. The primary aim of data mining being, extraction of information from large datasets and their transformation into useful information. This technique provides ways for the prediction of outcome of the observations which might be carried out in the future.

The data mining core tasks are considered to be Cluster Analysis, Predictive modeling, Anomaly detection and Association analysis. The aim of Cluster Analysis is to find the set of related observations, which may be of the same cluster and also similar to one another when compared to observations of other clusters. For the grouping of related objects Clustering Techniques may be used. In Predictive modeling a model is built for the target variable which is considered a function of explanatory variable. In Anomaly Detection the primary task is to identify observations that are different from the characteristics of rest of the data. These observations are the anomalies. The main aim is to detect real anomalies. Association analysis is used for discovering patterns that define features that are strongly associated with data. The objective is efficient extraction of interesting patterns.

The data mining tasks are divided mainly into Predictive tasks and Descriptive tasks. The aim of Predictive task is to predict a value of an attribute based on attribute values. The predicted attribute is called the target variable. The Descriptive task aims on deriving patterns which usually summarizes data relationships. These are often explanatory and they require post processing techniques for validation. The Data mining applications include Outliers Identification and Detecting Fraud, Customer Profiling, Prediction and Description, Relationship Marketing, Customer Segmentation, Website Design and Promotion. It is used to answer queries the customer or user asks for. It is also used for the analysis of customer profiles and improvement of marketing plans. The unusual expenses that are claimed by the staff are identified, also the fraud in credit cards are determined. Using Web mining the user behaviour is tracked based on navigation while using the web site.

1.1 K-means

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the *k*-means algorithm; it is also referred to as Lloyd's algorithm, particularly in the computer science community.

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

The algorithm has a loose relationship to the *k*-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with *k*-means because of the *k* in the name. One can apply the 1-nearest neighbor classifier on the cluster centers obtained by *k*-means to classify new data into the existing clusters. This is known as nearest centroid classifier.

2. ARCHITECTURE OF THE PROPOSED MODEL

An architecture diagram as shown in **Fig-1** is a description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. The architecture used in our project gives air

crash investigation and safety measures for a particular air crash. A particular input of an air crash is given and then that particular data is tokenized into parts. These parts are then clustered according to LDA algorithm and stopwords are removed by the NLTK. The system is already provided with a set of pre-existing data records called as dataset. Given data is compared with the already present data in the system. Finally, the crash investigation report and safety measure report is given to the user else a default conclusion is provided.

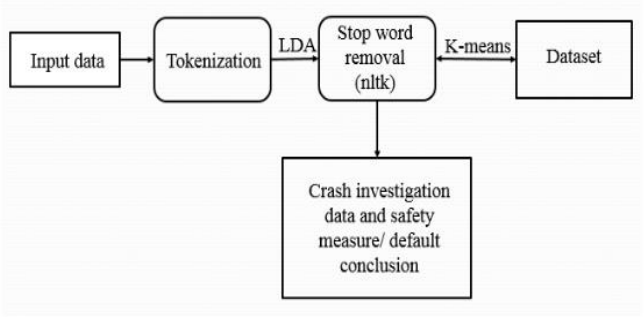

Fig -1.: Architecture diagram

Data flow diagram as shown in Fig-2, is a pictorial representation of the flow of data within a system right from the start till the end. Data flow diagram is often used as a preliminary step to create an overview of the system. Data flow mainly gives the interaction between the system. First, the user logs into the system by giving the username and password. Then the login authentication will be checked by the system and the next process will take place by the system. The data input will be viewed by the system and compared with the already present dataset using tokenization. The details generated will be sent to the user with the safety measures from probability based questions given to the user. Then the user logs out of the system.

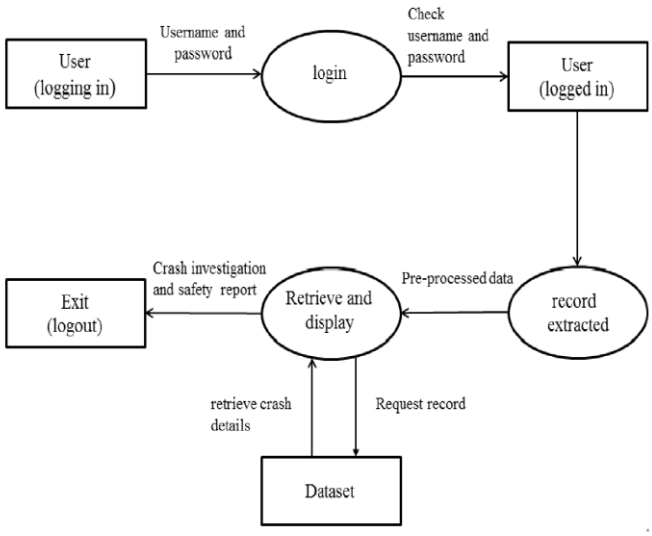

Fig -2.: Data flow diagram

4. RESULTS

The proposed algorithm is known as k-means algorithm. K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. K-means clustering aims to partition 'n' observations into k clusters in which each observation belongs to the cluster with the nearest mean. In the proposed system first the user logs into the system through the authentication of the admin. After the user logs in he gives the input as shown in Fig-3, for which the prediction takes place as shown in fig-4. Prediction starts based on the proposed algorithm which is given in the dataset. Data set is a large file which contains all the necessary data which is required for cluster mining.

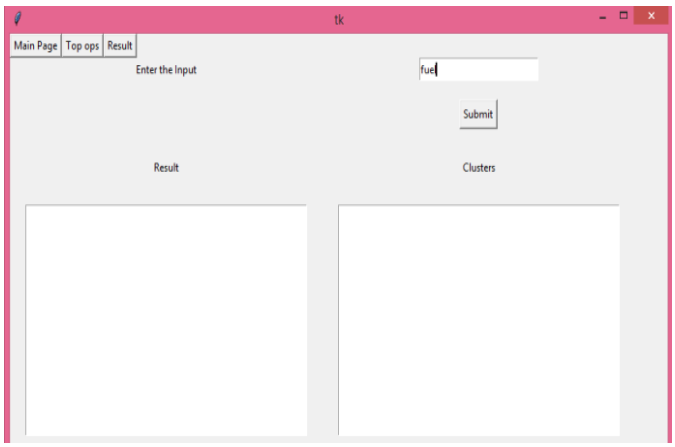

Fig -3.: User input

In fig-4 shows that when the user gives an input in which keyword is not present in the existing cluster by proposed algorithm then, the default output will be produced.

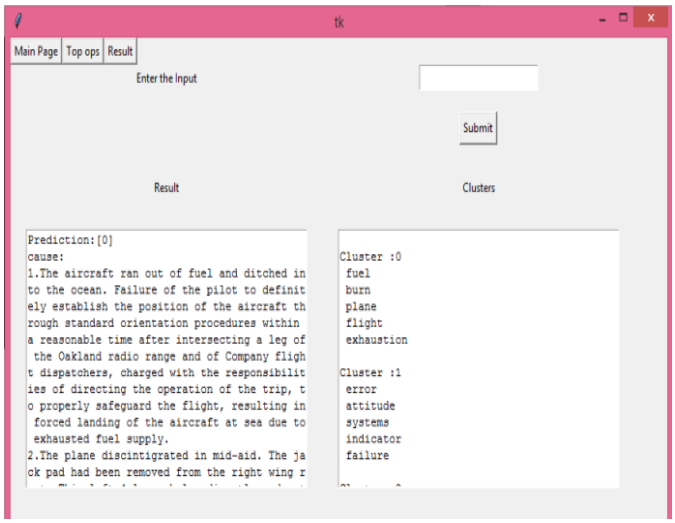

Fig -4.: Prediction of causes and safety measures

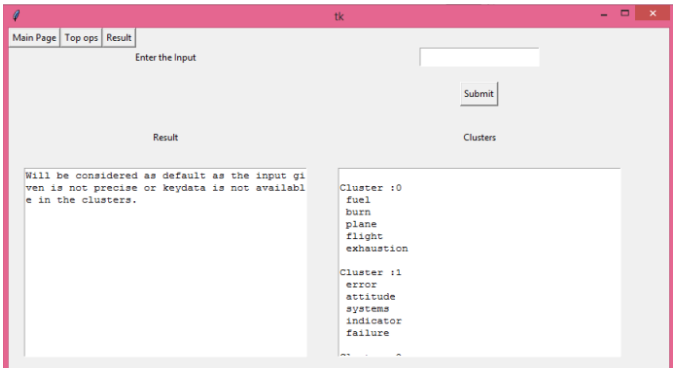

Fig -5.: Default prediction

After the analysis for input data is done the user will also be provided with the top 10 worst operator and dangerous locations as top ops shown in Fig-6. Different aircraft with their worst disasters are discussed to the user through which the user will have an idea about the recent mishaps. The user will also be provided with the worst or dangerous locations from all the recent disasters occurred.

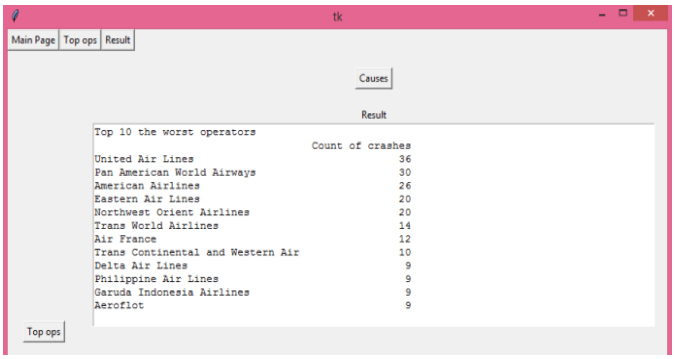

Fig -6.: Top op prediction for top 10 worst operators

Another algorithm called LDA algorithm is used in the proposed model. Latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. All the datasets provided will be further divided with the removal of the stopwords present by the NLTK and given to the user as shown in fig-7.

Fig -7.: Removal of stop-words based on NLTK

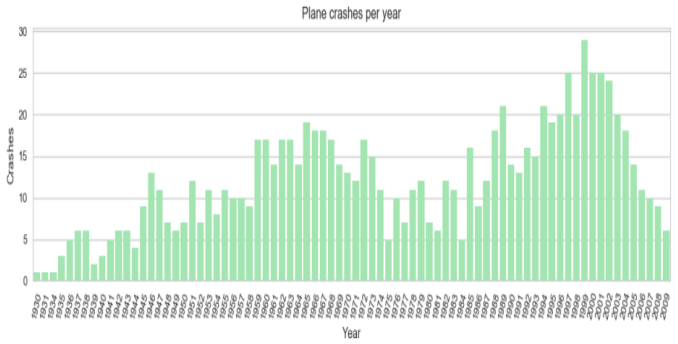

Fig -8.: Graph representing plane crashes per year

Various graphs are generated from the dataset indicating the details about crashes graphically.

Fig -9.: Graph representing plane crashes per day

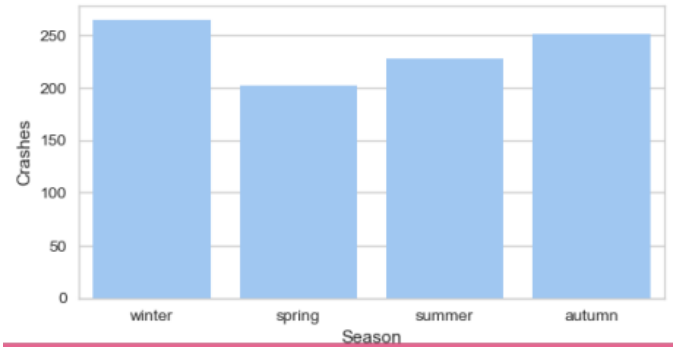

Fig -10.: Graph representing plane crashes per season

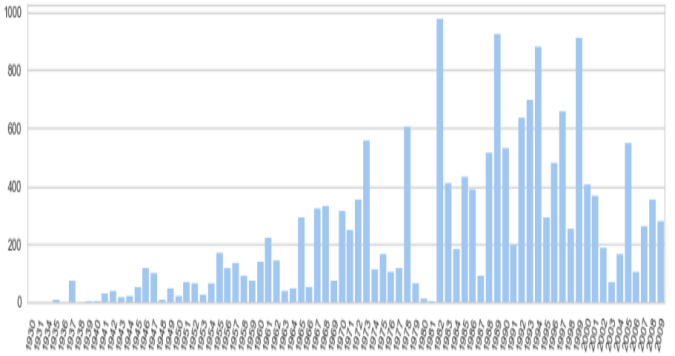

Fig -11.: Graph representing passengers survived per year

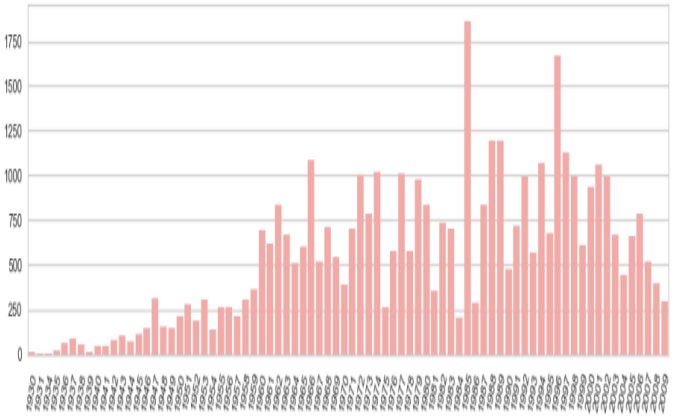

Fig -12.: Graph representing passengers dead per year

3. CONCLUSIONS

This work has investigated the user's request and has given the particular prediction and safety measure for that input. The K-Means clustering technique was used to find the clusters and fatality for the flight crash investigation. This method will also consider other factors like efficiency, weather impact and schedules of other aircraft. Results showed that the proposed algorithm obtained prediction based on user's input with efficiency. The proposed algorithm also provided improved search results for the query given by the user. Possible future work is to improve the efficiency and also increase the count of clusters used.

ACKNOWLEDGEMENT

It is with great satisfaction and euphoria that we are submitting the paper on "Air Crash Investigation and Safety Promotion using Data Mining Technique". We are profoundly indebted to our guide, Mr. Vasudeva Rao P V, Assistant Professor, Department of Information Science & Engineering, for innumerable acts of timely advice, encouragement and we sincerely express our gratitude. We also thank him for his constant encouragement and support extended throughout. Finally, yet importantly, we express our heartfelt thanks to our family & friends for their wishes and encouragement throughout our work.

REFERENCES

[1] Shagun Sharma, Ms. A. Sai Sabitha , "Flight Crash Investigation Using Data Mining Techniques", 1st India International Conference on Information Processing(IICIP)- Delhi, India, 2016, pp. 1-7.

[2] Australian Transport Safety Bureau, Avoidable Accidents No. 5 "Starved and exhausted: Fuel management aviation accidents", Australia, 2003.

[3] Yichen Zhang, Rong Su, Qing Li, Christos G. Cassandras and Lihua Xie, "Distributed Flight Routing and Scheduling for Air Tra_c Flow Management", IEEE Transactions on Intelligent Transportation Systems, 2017, pp. 2681-2692.

[4] Babatope S. Ayo, Yim Fun Hu, Jian-ping Li, "Adverse weather avoidance considering flight level changes", 7th International Conference on Innovative Computing Technology(INTECH)- Luton, UK ,2017 , pp. 144-148.

[5] Ray Haythornthwaite, Adrian Earle, Abdullah Rahal and Dick James, "Case History: Novel FA Techniques Used to Recover EEPROM Data from the Swissair 111 Crash", 2001 IEEE International Conference -Orlando, USA, 2001, pp. 283-288.

[6] Timothy P. Waldron, Saab Sensis Corporation, East Syracuse NY, "Mining Airport Surveillance For Operational Insights", Digital Avionics Systems Conference(DASC)- Seattle , USA, 2011, pp. 2C1-1 - 2C1-14.

[7] Yang Yue, LiWeimin, Li Xueming, "Research into Prevention of Errors of Airline Transport Pilots: A Rough Set Approach" - IEEE 11th International Conference-Harbin, China, 2013, pp. 154- 158.

BIOGRAPHIES:

Vasudeva Rao P V
Assistant Professor,
Areas of Interest: Data mining, Image processing.
Sahyadri College of Engineering and Management

Jyothi Mishel Veigas
Student
Sahyadri College of Engineering and Management

Kajal B. Surve
Student
Sahyadri College of Engineering and Management

Anvitha P. Shetty
Student
Sahyadri College of Engineering and Management

Dhanyashree
Student
Sahyadri College of Engineering and Management