Collective PredictiveAnalysis for Flight Disasters

Ms.Annapurna.K.A.
Department of Computer Science and Engineering,
Channabasaveshwara Institute of Technology
Gubbi, Karnataka, India.

Mr.Harish.T.A.
Assistant Professor,
Department of Computer Science and Engineering,
Channabasaveshwara Institute of Technology
Gubbi, Karnataka, India.

Abstract-**In airline community, identifying patterns of factors which is associated with aircraft accidents is of high interest to the aviation safety. The data collected from disasters, human errors, weather and social media i.e., twitter (live tweets) are used to derive the performance of airlines towards safety. Twitter provides huge amount of data. Analysis is made on four data set i.e., disaster data, human error data, weather data, and twitter data. After performing analysis, from each dataset we will be getting different patterns. As we are using four dataset so four patterns will be obtained. Finally, the four different patterns are merged using contrast set mining algorithm to give predicted outcome.**

This paper focuses on the performance analysis of aviation safety community using contrast set mining algorithm, which establishes the relationship between the four dataset i.e., disaster data, human error data, weather data, twitter data. Finally, predictive results will be obtained in data visual form.

Keywords-**data mining, contrast set mining, predictive analysis, aviation disaster, human involvement, weather, social media dynamics.**

I. INTRODUCTION

Now a day, fatal air crashes are the worst type of crises that airlines face. A disaster could be seen as a natural or man-made (or technological) menace resulting in an incident or substantial extent causing significant physical injury or destruction, loss of life, or radical change to the environment. A disaster can be ostensibly defined as any catastrophic event stemming from events such as earthquakes, floods, catastrophic accidents, fires, or explosions [4]. It is a phenomenon that can cause destruction to life and property and destroy the economic, social and cultural life of people.

In contemporary academia, catastrophes are seen as the consequence of inaptly managed menace. These menaces are the product of combination of both hazards and vulnerability. Hazards that strike in areas with low vulnerabilitywill never become disasters as in the case of uninhabited regions [4].

Two types of catastrophe has been identified which are natural catastrophe and man-made catastrophe. A natural catastrophe is a consequence when a natural menace affects human and/or the built environment. Human susceptibility and lack of appropriate emergency management leads to financial, environmental, or human impact. The resulting Loss depends on the capacity of the population to support or resist the catastrophe; their resilience. This understanding is concentrated in the formulation: “catastrophe occurs when menaces meet susceptibility”, examples are earthquakes, landslides, volcanic eruption, floods and cyclones; these natural and Menaces kill thousands of people and destroy billions of dollars of habitat and property each year.

Man-made catastrophes are the consequence of technological or human menaces. Examples include fires, transport accidents, industrial accidents, oil spills and nuclear explosions/radiation.

Air catastrophe can be seen as an air event resulting in physical harm, loss of Lives and property and even environmental destruction. Handling or avoiding such catastrophes necessitatesthorough analysis to predict the cause for catastrophe. Aviation mishap analysis is carried out in order to determine the cause or causes of amishap or series of mishaps so as to avert further incidents of aanalogous kind. Predictive analysis is the practice of mining information from existing data sets in order to determine patterns and predict future outcomes and trends.

Mishap analysis is performed in four steps:

1. Fact gathering: after a mishap happened a forensic process starts to gather all possibly relevant facts that may contribute to understanding the mishap.
2. Fact analysis: after the forensic process has been completed or at least delivered some results, the facts are put together to give a “big picture”. The history of the mishap is reconstructed and checked for consistency and plausibility.
3. Conclusion drawing: if the mishap history is sufficiently informative, conclusions can be drawn about causation and contributing factors.
4. Counter measures: in some cases the development of counter measures is desired or recommendations have to be issued to prevent further mishaps of the analogous kind.

In thispaper four data sets are considered i.e., disaster data, human error data, weather data and twitter data.

Disaster data describes accident and incident event description, when disaster happened and what are the factors affecting disaster, when and which airline got collapsed. Human error data contains the details of staff. Disasters are greatly affected by weather.

From thunderstorms and snow storms, to wind and fog as well as temperature and pressure extremes, every phase of flight has the potential to be impacted by weather.Twitter contains a very large number of very short messages about the aircraft disaster.In twitter, the tweets may come from common people and from the experts in that field. Here, onlyexpert's tweets are taken for analysis. Patterns are determined from each data set by performing analysis on the data set. We are using contrast set mining algorithm for performing collective feature analysis of four data sets, which establishes the relationship between four data sets. Finally, the predicted solution will be obtained in the data visual form.

II. PROBLEM STATEMENT

Aviation industry has experienced a series of air catastrophes in the recent past. On 4th February, 2015,The ATR-72-600 operated by TransAsia took off from Taipei, Taiwan, for a passenger flight to Kinmen, Taiwan. 53 passengers and 5 crewmembers were on board. The plane crashed shortly after take-off in the water of the Keelung River near the Nankang Software Park, 2.9 NM from the end of runway. 15 occupants were injured, and 43 occupants were killed [9].

On 28th December, 2014,The AirAsia Indonesia Airbus A320-200 took off from Surabaya, Indonesia, for a passenger flight to Singapore, Singapore. 155 passengers and 7 crewmembers were on board [9].

There are at least two to three accidents or disasters per year by large aircrafts, causing many deaths and economic losses there is no single reason for the mishap. Therefore we aim at performing collective predictive analysis that analyse current and historical facts to make predictions about future.

III. RELATED WORK

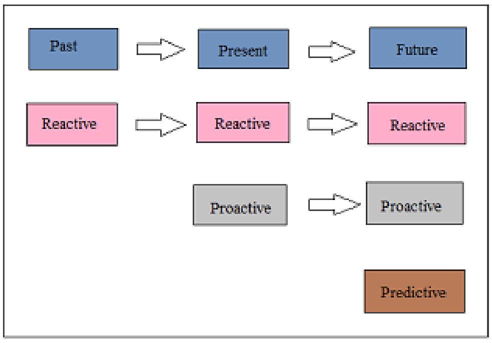
A. Predictive Analysis

Predictive analysis is a branch of data mining concerned with the prediction of future probabilities and trends. So that we can plan and carry out strategies that improve outcomes.

The central element of predictive analysis is the predictor, a variable that can be measured for an individual or other entity to predict future behaviour. Multiple predictors are combined into a predictive model, which, when subjected to analysis, can be used to forecast future probabilities with an acceptable level of reliability.

In predictive modelling, data is collected a statistical model is formulated, predictions are made and the model is validated as the additional data becomes available.Predictive analysis is applied to many research areas, including meteorology, security, disaster avoidance, genetics, economics and marketing [5].

Figure 1: Safety strategies overview



ICAO also recommends predictive analysis, because *"...it deals with hazards when they are at infancy and therefore have no opportunity to start developing their damaging potential. It also allows a high level of intervention, which is a highly efficient one."*

Predictive analysis is based on two steps: Firstly, identifying the factors that contribute to these events and compiling statistics for these factors during normal flight operation. Secondly, using this information to calculate the probability of the incident itself in a statistically valid way. In other words, predictive analysis means looking at statistics and variations that occur during the whole flight operation for a given airline in order to quantify incident probabilities. However, having a numerical value for a certain incident probability is meaningless without accounting for uncertainties.

B. Dataminig

Data Mining is defined as extracting the information from the huge set of data. In other words we can say that data mining is mining the knowledge from data. This information can be used for any of the following applications:

- Market Analysis
- Fraud Detection
- Customer Retention
- Production Control
- Science Exploration

Need of Data Mining:

Here are the reasons listed below:

□ In field of Information technology we have huge amount of data available that need to be turned into useful information.

□ This information further can be used for various applications such as market analysis, fraud detection, customer retention, production control, science exploration etc.

What can data mining do?

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics.

And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history.

By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

C. Existing system:

Previous research on aircraft mishaps has focused on studying mishaps that determine casual factors involved in mishaps. Log-linear modelling technique is used to analyse casual factors involved in loss of separation incidents. Log linear modelling technique is a technique used in analysis which scrutinizes the relationship between the two categorical variables. It is used in hypothesis testing and model building.

Hansen and Zhang tested the hypothesis that adverse operating conditions lead to higher incident rates in air traffic control. In-flight LOC (Loss of Control) is a serious aviation problem. Well over half of the LOC mishaps included at least one fatality.

In about 30 per cent of LOC mishaps, the LOC was secondary to system/component failure/malfunction. 23 per cent of LOC accidents were secondary to aircraft damage. Other frequently cited causes for LOC are derisorypre-flight, improper planning or decisions and flying in obscuration or at night.Dimukes studied 19 airline accidents focusing on pilot errors.

Pilot errors are one of the causes of aviation disaster. Many of all the mishaps are due to pilot error. Pilot must navigate through bad weather. If there are any mechanical issues he must respond and execute safe landing or take-off. Sometimes even mishaps are caused when pilots misread equipment, misjudge weather or if he fails to recognize mechanical errors until it's too late.

Disadvantages:

- Lacks in aim to reveal the casual reasoning behind the events and circumstances leading to amishaps.
- Does not identify the relationship between the incident factor and accidents.

- These studies helps in understanding individual mishaps and their casual factors, the low rate of mishaps however, makes it difficult to discover repeating patterns of these factors.

IV. IMPLEMENTATION

A. Proposed system:

Traditional methods of discovering safety menaces and determining future jeopardy will not take the aviation community to the next level of mishap prevention needed to ensure the success of the next generation of air transportation.

In the proposed system we are performing predictive analysis of different or various factors related flight/airlines catastrophe on the basis of both historical and live data. We also shown the establishment of relationships among different data's to perform a predictive analysis.

Advantages:

- Collection of all Aviation data available to mine for future risks
- Strive to get beyond the limits of looking at Mishaps
- Reveal dangers that lie ahead, and their removal
- Simulating the "Incident Investigations", without the Incidents actually happening.
- Weighing of the importance of individual data elements towards Jeopardy Calculation
- Panel of subject matter experts to evaluate the importance of data in predictive Analysis
- Reveal and forecast the cause of Mishap
- Simulate and Train
- Make Data available to all
- Measures or Controls are put in place to preclude the cause from reoccurring.

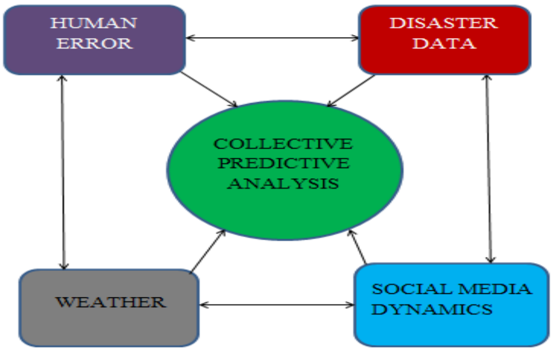


Figure 2: System Architecture

Above system architecture has four modules human error, disaster data, weather and social media dynamics.

Human error: is having a crucial cause for the disaster caused by pilots, co-pilots, flight engineers, fright carrier pilot. All of them will be having their own responsibilities even some small mistake may cause serious disaster[7].

Asst. Airport Manager/Deputy Director of Aviation:

Assistant airport managers, or deputy directors of aviation, help the airport manager fulfil the tasks and responsibilities of maintaining an airport. This may include the purchase, maintenance, and staffing of airport vehicles, and equipment.

Director:

Airport directors are ultimately responsible for what happens at an airport: short- and long-term planning, project management, staffing, operations in general. It's a high-stress job that is also very rewarding.

Airport Engineer / Planner:

Airport engineers or airport planner, work together with architects and contractors to design airportsand runways. They supervise all phases of design and construction in an effort to keep the work on schedule and according to approved plans.

Disaster data is having information about how like tail strike, runway and when like landing, take-off, approaches disaster happened.Flight disaster contain main fields, they are:

Phase of light: current position of the flight when accident happened, it may be landing, take-off, and approach
Event description: why disaster happened? Cause for the disaster
Damage category: damage may be major or minor, substantial or damaged
Injury category: what kind of injury happened fatal or serious?
Major accidents:accidents may be major or minor

Weather: will be the main cause for flight delay, sudden changes in weather may cause air crashes.In passenger-carrying aircraft, turbulence is a major concern, while thunderstorms can close air routes for hundreds of miles. Volcanic ash, especially hazardous to aircraft engines, forces costly re-routes[8]. Fields considered in weather are:

Id: assign a number in chronological order
Flight id: flight identification number
Event_type_desc: description of event type whether accident, incident or delay
Phase of light: current position of the flight
Result**:** what are the results caused by weather disaster?
Effect: what are the effects for the cause?

Twitter: Social media is very popular and it is useful for prediction based on opinion in their tweets.We do three activities before classifying the tweets into positive, negative and neutral[10].

- **Display the tweets:** the live tweets such as previously stored tweets and currently coming tweets are displayed. The tweets from people increases until we stop it.
- **Save the tweets:** In the first activity, it displays only the tweets but it is not saved. Here it saves the past and present tweets and it increases until we stop it.
- **Clean the tweets:** the tweets saved in the second activity are used for cleaning the tweets. It means clean the tweets in order.

We collected huge amount of text posts from twitter. We are taking only expert tweets for the analysis. We classify the tweets into three sets of texts:

- Texts containing positive opinion on performance, such as good, well, great, etc.,
- Texts containing negative opinion on performance, such as bad, poor, very bad, etc.,

Objective texts that only state a fact or do not express any emotions

Since all modules are collectively combined by establishing mutual relationship, by this combined analysis we can easily predict the future problem issues.

B. Methodology:

Identifying patterns of factors associated with aircraft accidents is of high interest to the aviation safety community. We applied the STUCCO algorithm to analyse aircraft accident data in contrast to the aircraft incident data in major aviation safety databases and identified factors that are significantly associated with the accidents. The data pertains to accidents and incidents involving commercial flights. Accident database are analysed against incident data-bases and the results were compared.

A. Phase 1: An aircraft accident

Is an occurrence associated with the operation of an aircraft in which people suffer death or injury and/or in which aircraft receives substantial damage.

B. Phase 2: An aircraft incident

Is an occurrence which is not an accident but is a safety hazard and with addition of one or more factors could have resulted in injury or fatality, and/or substantial damage to the aircraft. Previous research on aircraft accidents has focused on studying accident data to determine factors leading to accidents, causal reasoning behind the events and circumstances leading to an accident.

While these studies help understanding individual accidents and their causal factors, the low rate of accidents however, makes it difficult to discover repeating patterns of these factors. Air Traffic Management related accidents worldwide and showed flight crew is a more important factor in ATM-related accidents than air traffic control.

Already reported no systematic trends were found in the accident causal factors, it does not identify the relationship between the incident factors and accidents.Dataset when performing a trend analysis. While these studies help understanding individual accidents and their causal factors, the low rate of accidents however, makes it difficult to discover repeating patterns of these factors, while studying incident data is helpful to understand incident.

Data:
The data used in the study consists of delay, disaster analysis, human error, social media combined to commercial flights.[1][3].

- National Transportation Safety Board (NTSB) database containing reports of all
Accidents.
- Federal Aviation Administration Accident and Incident Database System
(FAA/AIDS), containing reports of incidents investigated and/or documented by the FAA
- Statistical summary of commercial jet airplane accidents(SASTSUM) containing reports of disasters of airlines
- Real time twitter contains data of social media
- FAA Operational Errors and Deviations (OED), containing mandatory reports of
Air Traffic Control errors

Each report in these databases consists of structured fields plus an unstructured
Narrative explaining the event.

Data Constraints:
Some constraints imposed by the data need to be considered , All accidents in the United States involving civil aircraft are investigated by the National Transportation Safety Board (NTSB)[1][3].

The historical data on incidents is large enough to represent these factors qualitatively. we consider all factors that have been present in an event, regardless of their primary or contributory role in leading to the event.

Data Selection: Since the purpose of the analysis is to identify operational factors under normal conditions,accidents, incidents, delay and disaster due to the following causes were filtered out from the data.[1][3].

- bird/animal strike, such as aircraft encountering a deer on the runway
- oversight/due diligence such as human resource
- events during the phases of operation when the aircraft is not operating

C. Phase 3: Problem Definition

In association rules, we typically deal with airline data where the database D is a set of transactions with each transaction T _I = {I1, I2, I3, I4,,,,,,Im}. Each member of I is a literal called an item, and any set of these literals is called an item set. We generalize the data model to grouped categorical data. The data is a set of k-dimensional vectors where each component can take on a finite number of discrete values.The vectors are organized into n mutually exclusive groups.[6]

The concept of an item set can be extended to a contrast-set as follows:-

Definition 1. Let A1,, A2,,,,,,,,,Ak be a set of k variables called attributes. Each Ai can take on values from the set {Vi1, Vi2, and Vim}. Then a contrast-set is a conjunction of attribute-value pairs defined on groups G1, G2,,,,,,,,,,,,Gn.

Example: (flight disaster = landing) ^ (weather= thunderstorm).

We define the support of a contrast-set with respect to a group ‘G ‘as follows

Our goal is to find all contrast-sets whose support differs meaningfully across groups. Formally, we want to find those contrast-sets

For all ijP (cset = True | Gi)! =P (set = True | Gj)…………………………………… (1)

We call contrast-sets where Equation 1 is statistically valid significant

STUCCO: a mining algorithm:

We treat the problem of mining contrast-sets as a tree search problem. The root node is an empty contrast-set, and we generate children of a node by specializing set by adding one more term. We use a canonical ordering of attributes to avoid visiting the same node twice. Children are formed by appending terms that follow all existing terms in a given ordering[6].

We search this tree in a breadth-first, level wise manner. Given all nodes at a level, we scan the database and count their support for each group and then examine each node to determine if it is significant.

Steps:

- Plot a different graph by considering different attributes for each module
- Extract the high item set individually from each module
- Establish the connection between any two item sets
- Merging all item sets collectively and get required pattern
- Prune out remaining unmatched patterns and save it for future use

- Continue these steps until pattern represents in visualization

Given all nodes at a level, we scan the database and count their support for each group and then examine each node to determine if it is significant and large, if it should be pruned, and if children should be generated.

After finding all significant contrast-sets in the data, we then process the results and select a subset to show to the user. We display the low order results first, which are simpler, and then show only the higher order results that are surprising and significantly different

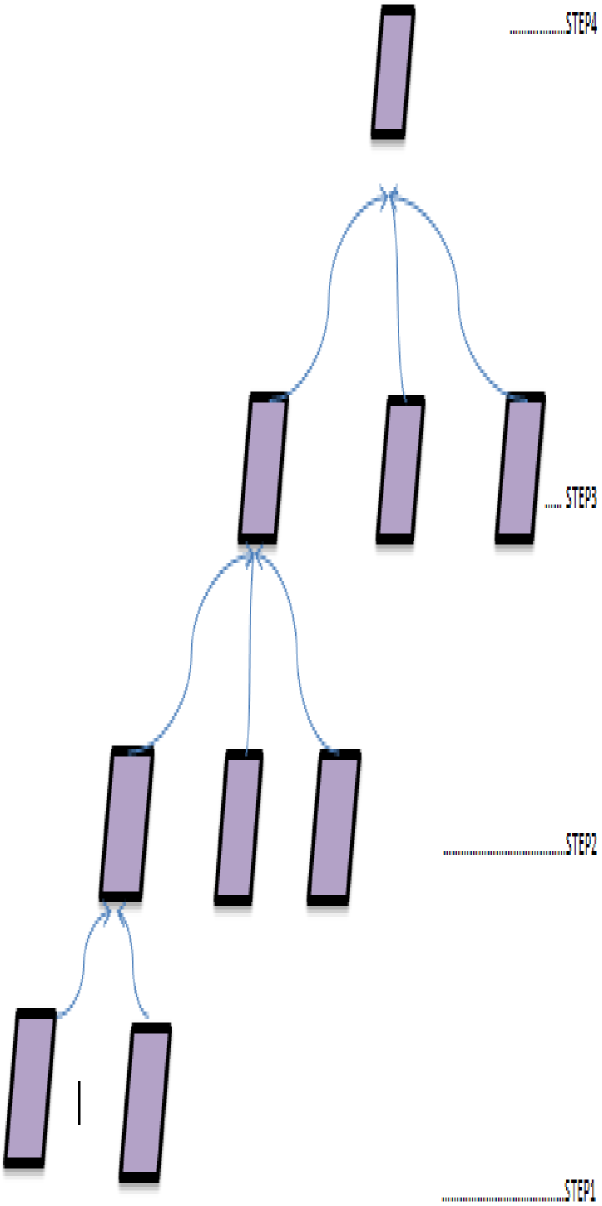


Figure 3: Flow of contra-set mining to obtain surprising data.

D.Phase 4: Finding Significant Contrast Sets

We can check if a contrast-set is significant by testing the null hypothesis that contrast-set support is equal across all groups or, alternatively, contrast-set support is independent of group membership.

The support counts from each group are a form of frequency data which can be analysed in contingency tables and pictorially represent by plotting the graph.

We form a 2*c contingency table where the row variable represents the truth of the contrast-set and the column variable indicates the group membership

Algorithm:-

STUCCO: Search and Testing for Understandable Consistent Contrasts.

```
 Algorithm STUCCO
  Input: data D
  Output: Dsurprising
    Begin
Set of Candidates C {}
    Set of Deviations D {}
      Set of Pruned Candidates P {}
 Let prune(c) return true if c should be pruned

        while C is not empty
        scan data and count support  ¥ c € C
        for each c € C
        if significant(c) ^ large(c) then
      D ← D U  c
        if prune(c) is true then
        P  ←P U c
        else Cnew  ←Cnew U GenChildren(c,P)
        C  ← Cnew
     Dsurprising ← Find Surprising(D)
```

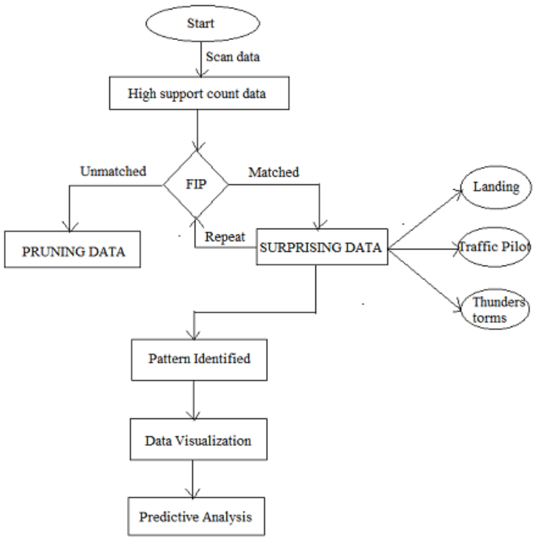


Figure 4: Data flow of stucco algorithm

V. RESULT ANALYSIS

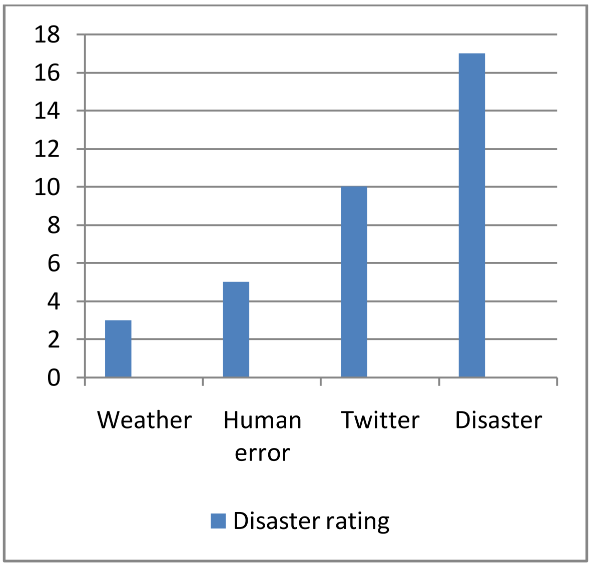


Figure 5: Graph represents major influenced factor for disaster.

The figure above shows major influenced factor that is obtained from the overall collective predictive analysis using contrast set mining algorithm andas per result analysis the highly influenced factor is on disaster.

VI .CONCLUSION AND FUTUREWORK

This paper focus on multiple factors and identifies patterns of factors having major influence for disaster by performing predictive analysis using contrast set mining algorithm, which establishes relationship between factors and give predicted solution in data visual form. This takes aviation to new levels of risk-based decision making and mishap prevention.

In this paper it has shown the way for doing predictive analysis using some existing data or preloaded data. Also, we can do this on real time data by using some streaming API we crawling data. For applying predictive analysis on real time data it helps to predict and cause of further disaster of flights.

VII. ACKNOWLEDGEMENT

We are grateful to express sincere thanks to our faculties who gave support and special thanks to our department for providing facilities that were offered to us for carrying out this project.

REFERENCES

[1] Analyzing relationships between Aircraft accidents and Incidents. A data mining approach – by Zohrehnazeri

[2]Aviation Risk Forecasting – by Futronorporation, www.futron.com • 1.800.807.4927

[3] Contrast-Set mining of Aircraft Accidents and Incidents – by ZOhrehNazeri, Daniel Barbara, Kenneth De Jong, and Lance Sherry http://link.springer.com/chapter/10.1007%2F978-3-540-70720-2_24

[4]http://www.academia.edu/6241117/AIR_DISASTER_AND_ITS_IMPLICATIONS_IN_THE_DEVELOPING_COUNTRIES_A_CASE_STUDY_OF_NIGERIA

[5]http://asndata.aviation-safety.net/industry-reports/IATA-safety-report-2013.pdf

[6]http://www.dbis.informatik.hu-berlin.de/dbisold/lehre/WS0405/KDD/paper/BP99.pdf

[7]https://www.nifc.gov/fireInfo/fireInfo_documents/humanfactors_classAnly.pdf

[8]http://climate.dot.gov/documents/workshop1002/kulesa.pdf

[9]http://www.1001crash.com/index-page-crash-lg-2.html

[10]http://lrecconf.org/proceedings/lrec2010/pdf/385_Paper.pdf

First Author: Annapurna K.A was born in Karnataka, India. She received the B.E Degree in Computer Science and Engineering from *Visvesvaraya Technological University* Belgaum Dist., India in 2013 and pursuingM.Tech Degree in also same branch and University. Her research interests are in the area of data analysis and big data.

Second Author: Harish T.A M-tech was born in Karnataka, India. He is working as Asst.Prof. In computer Science Engineering Dept., CIT, Gubbi, Tumkur (district), Karnataka, INDIA. His research interests are in the area of data analysis and big data.