International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 4, July- August 2017 ISSN 2278-6856

NTSB Aviation Accidents Analysis Using R

Nimeelitha Reddy[1], M.Padma[2]

[1]PG Student, Computer Science and Engineering Dept, GPREC, Kurnool (District), Andhra Pradesh-518004, INDIA.

[2] Assistant Professor, Computer Science and Engineering Dept, GPREC, Kurnool (District), Andhra Pradesh-518004, INDIA.

Abstract

With the ever increasing population, the need of commutation is also increasing resulting in the loss of valuable resources. For instance, when a major accident is occurring on a route with heavy traffic, there will be human loss, public property damage and so on, causing inconvenience to other commuters. Research is going on in this area. In this paper, analysis of the dataset consisting the data of civil aviation accidents and selected incidents with the US and its territories, possessions and in international waters with R programming is being done so that the risk of any further commute incidents or accidents deprecate andcan provide an opportunity for better decision making. For instance, determining weather conditions so that incidents are prevented or when an accident occurs, providing the medical assistance to the victims. This paper is based on the data from the National Transportation Safety Board (NTSB). It involves classifying a set of aircraft accident/incident data covering the years 1948 to 2017 in the United States and its territories, possessions and in international waters. The NTSB provides one of the most extensive online aircraft accident and incident databases. It includes dates, airport names, Injury severity, aircraft and engine types, scheduled and non-scheduled certificated air carrier, and the name of the air carrier. Filters here enable the business aircraft operators to identify risks associated with the aircraft they operate, the types of operations and procedures they typically fly, and the airports they frequent.

Keywords: Aviation Accidents, R programming, data analytics, aviation accidents prediction.

1. INTRODUCTION

Whether it is fine monitoring shop floor operations, assessing consumer sentiment or large-scale analytic challenges, big data is having an exceptional impact on the aviation enterprise. The amount of data that is generated in every flight has risen over the years and diverse information is being stored in digital formats. However, it is not just the access to new data sources, but access to patterns and interrelationships among these elements is of interest. Collecting such changeable types of data does not quickly create any value; it is the analysis of this data to unwrap the insights that will help the organization which is important. This paper is based on data from the National Transportation Safety Board (NTSB) and it consists of79,412 records from 1948 to 2017. The NTSB not only handles data from the USA but from around the world and it currently manages the data from the 1940s to the present. This process generates tremendous amounts of log data which makes the storing and analysis a concern that needs to be solved. Storing vehicle monitoring data is very important for the NTSB to give information support for departments such as public security, criminal investigation, and economic investigation and front-line police.

R is widely known to solve all types of problems with data analytics, and in this paper, usage of some R statistical methods like ggplot, geo_point, summary, dplyr, randomForest is done to analyze the data. Many questions about aviation accidents in the USA have been answered. However, the big question still remains: how can it be safer? Particularly in the context of ever increasing amount of the flights. Consequently and definitively more questions can be solved and analyzed in order to get a more complete analysis of this data.

After analyzing this data, the most-common occurrences in these mishaps is the ostensible failure to stabilize the aircraft during the landing that canbe due to factors such as the probablereadiness of the flight crew to venture the landing of the aircraft in unsafe conditions. Analyzing this data uncovered that the most basic notion ‘bad metrological conditions’ is not at all the main reason for aviation accidents. A detailed analysis was done to the Illinois State, revealing that most aviation accidents in Illinois happen during landing and that Cessna is the aircraft that was involved in more of its accidents in the last 15 years.

Based on the history of aviation accidents in the USA, probable accidents have been forecasted using R for some states. The prediction exposes in what phase of flight Aviation accidents are more probable to occur. In this paper, analysis of the data with R is done to show high degree of sustainability and robustness in analysis and storage of large data.

2. DATASET PREPARATION AND DESCRIPTION

This paper is based on the data from the National Transportation Safety Board (NTSB). The data contains information from 1940 to the present from over the world, but for the purpose of this research data considered is from1948 to 2017 in the USA and its territories, possessions and in international waters. Some missing data is also present during this period which is in lesser proportion than the years before 1948. Missing data for this research was replaced by “unknown” and, thus calculations had to be based on this information.

1. **Attributes**

The data set has 32 attributes, in which few of them may not be required as those are generic fields and few of them not having proper data.

2. **Data Analysis**

3. RELATED WORK

Modeling studies have employed different statistical techniques to unravel the complexity of interactions between distributions and environmental factors. Those include Generalized Linear Models, especially Logistic Multiple Regression (LMR); Generalized Additive Models (GAM) and Classification and Regression Trees (CART, also known as Regression Tree Analysis, RTA).

A. Statistical methods

Methods used in predictive Modeling consist of two main types: global parametric and local non-parametric. Global parametric models use a strategy of global variable selection. Each variable enters the model as 'a whole' to explain its contribution to the response. This strategy is clearly inappropriate when the hypothesis is that variables interact in a non-homogeneous way across their range of values. However, global techniques are still appropriate for small data sets where the analyst is forced to use parametric Modeling techniques because all points will influence almost every aspect of the model. GLM and LMR have several important disadvantages. Ecologists frequently assume a unimodal and symmetric response to gradients, which real life obstinately tends to refute. Such multi-modal or skewed distributions are sometimes dealt with using high-order polynomial functions, but this strategy heavily increases the risk of over-fitting – finding patterns that only apply to the training data – creating models that work almost perfectly with original data but have poor predictive ability with new data.

Secondly, in GLM the relationships between response and predictors are assumed to be linear, when real-world effects are generally more complex. Our hypothesis is that in Modeling organism/communities distributions, response is related to predictor variables in a non-linear and local fashion. Local nonparametric models are suitable under such a hypothesis as they use a strategy of local variable selection and reduction, and are flexible enough to allow non-linear relationships. From this type we have tested CART and MARS. Classification and Regression Trees(CART)[4] is a rule based method that generates a binary tree through binary recursive partitioning, a process that splits a node based on yes/no answers about the values of the predictors. Each split is based on a single variable. Some variables may be used many times while others may not be used at all. The rule generated at each step maximizes the class purity within each of the two resulting subsets. Each subset is split further based on entirely different relationships. CART builds an overgrown tree based on the node purity criterion that is later pruned back via cross-validation to avoid over-fitting. The main drawback of CART models, when used to predict organism distributions, is that with more than just a handful of predictor variables or cases to classify, the generated models can be extremely complex and difficult to interpret, generating a tree with 510 nodes for just ten predictors. In the present study, the optimal tree obtained for the Fagus data set (103 181 cases) has 1726 terminal nodes! Such complexity makes the tree impossible to interpret, whereas in many studies interpretability is a key issue. Moreover, implementation of such a tree within GIS is unworkable.

Prediction maps are often a required outcome of Modeling, and this shortcoming affects the use of CART when complexity grows beyond a reasonable limit. Multivariate Adaptive Regression Splines (MARS) is a relatively novel technique that combines classical linear regression, mathematical construction of splines and binary recursive partitioning to produce a local model where relationships between response and predictors are either linear or non-linear. To do this, MARS approximates the underlying function through a set of adaptive piecewise linear regressions termed basis functions (BF). A model which clearly over fits the data is produced first. In subsequent steps, knots that contribute least to the efficiency of the model are discarded by backwards pruning steps. The best model is selected via cross-validation, a process that applies a penalty to each term (knot) added to the model to keep low complexity values

Another key improvement of MARS over global parametric models is the way it deals with interactions. In local Modeling, interactions can no longer be treated as global. MARS considers interactions not between the original predictors, but between the sub-regions of every basis function generated. A particular sub-region of a given basis function can interact with a particular sub region of another basis function, but other regions of these basis functions might display none or a different interaction pattern.

Most statistical techniques perform poorly with high dimensionality, a problem that in predictive Modeling literature is usually circumvented by limiting the number of variables employed through a priori selection, which is not always biologically warranted, or by eliminating the interactions from the model.

So, In this paper, to analyze the aviation data[16], we are using liner statistical methods. It will predict aviation accidents during a specific phase of flight. Visualization utilities are used in this paper to visualize the difference between injured and uninjured persons in aviation accidents from 1948 to 2017

This algorithm includes various steps like Data preprocessing, data clustering with K means algorithm and then performing some mining rules and finally association rules. So the rest of paper describes how implementation of the R scripts is being used for analyzing this huge NTSB data.

B. Data Preprocessing

The data set used in this project is obtained through NTSB (National Transportation Safety Board) as shown in Fig.2, it has 33 attributes

The time span of the available aviation accident data sets ranges from the year 1948 up to 2017. The data set is comprised of 79,412 accident records described using 32 attributes.

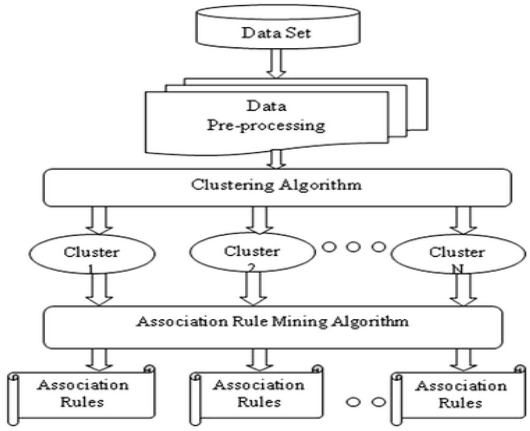
Some attributes of the data set are irrelevant for the purpose of this study and many attributes also presented duplicated data or invalid records, thus a step of data cleansing was required before using the data set.

```
> names(df)
 [1] "Event.Id"              "Investigation.Type"    "Accident.Number"
 [4] "Event.Date"            "Location"              "Country"
 [7] "Latitude"              "Longitude"             "Airport.Code"
[10] "Airport.Name"          "Injury.Severity"       "Aircraft.Damage"
[13] "Aircraft.Category"     "Registration.Number"   "Make"
[16] "Model"                 "Amateur.Built"         "Number.of.Engines"
[19] "Engine.Type"           "FAR.Description"       "Schedule"
[22] "Purpose.of.Flight"     "Air.Carrier"           "Total.Fatal.Injuries"
[25] "Total.Serious.Injuries" "Total.Minor.Injuries" "Total.Uninjured"
[28] "Weather.Condition"     "Broad.Phase.of.Flight" "Report.Status"
[31] "Publication.Date"      "geo_point"
>
```

Figure 12: NTSB data analysis algorithm and flow

```
> summary(df)
         Event.Id    Investigation.Type  Accident.Number      Event.Date
 20001212X19172:    3   Accident:76235    ANC14LA035:    2   5/16/1982:   25
 20001214X45071:    3   Incident: 3177    ANC15LA037:    2   6/30/1984:   25
 20100204X45658:    3                     ANC16CA056:    2   7/8/2000 :   25
 20101022X34140:    3                     ANC16FA001:    2   5/17/1986:   24
 20001204X00086:    2                     CEN14FA230:    2   6/5/1983 :   24
 20001205X00276:    2                     CEN14LA476:    2   8/25/1984:   24
 (Other)       :79396                     (Other)    :79400   (Other)  :79265
          Location             Country        Latitude       Longitude
 ANCHORAGE, AK  :  372   United States :74851   Min.   :-78.02   Min.   :-178.68
 MIAMI, FL      :  185                 :  507   1st Qu.: 33.38   1st Qu.:-115.02
 CHICAGO, IL    :  169   Canada        :  256   Median : 38.19   Median : -94.53
 ALBUQUERQUE, NM:  165   Brazil        :  220   Mean   : 37.70   Mean   : -93.82
 HOUSTON, TX    :  155   United Kingdom:  217   3rd Qu.: 42.57   3rd Qu.: -81.75
 Anchorage, AK  :  141   Mexico        :  210   Max.   : 89.22   Max.   : 177.56
 (Other)        :78225   (Other)       : 3151   NA's   :53543    NA's   :53552
  Airport.Code    Airport.Name    Injury.Severity    Aircraft.Damage
        :34536          :30031   Non-Fatal:60119             : 2414
 NONE   : 1466   N/A    : 1831   Fatal(1) : 7840   Destroyed  :17336
 PVT    :  357   PRIVATE:  216   Fatal(2) : 4625   Minor      : 2514
 ORD    :  146   Private:  173   Incident : 3177   Substantial:57148
 APA    :  142   NONE   :  140   Fatal(3) : 1453
 (Other):42726   (Other):46996   Fatal(4) : 1014
 NA's   :   39   NA's   :   25   (Other)  : 1184
```

Figure 13: Aviation Accident Data set – Total Attributes



The data set also lacks detailed information about aircrafts (i.e. model, Operator), geographic location(i.e. longitude, latitude.) and victims injury severity (i.e. fatal, non fatal).

Event Id	Aircraft Damage	Total Fatal Injuries
Investigation Type	Aircraft Category	Total Serious Injuries
Accident Number	Registration Number	Total Minor Injuries
Event Date	Make	Total Uninjured
Location	Model	Weather Condition
Country	Amateur Built	Broad Phase of Flight
Latitude	Number of Engines	Report Status
Longitude	Engine Type	Publication Date
Airport Code	FAR Description	geo_point
Airport Name	Schedule	
Injury Severity	Purpose of Flight	
	Air Carrier	

The data set is comprised by 79,412 records of Aviation accident events that took place in the world since 1948.

The attributes of the data set can be categorized in different types:

Geospatial Attributes: These attributes, listed in Table I represents where the accident happened in space. They are used in this project to analyze the aviation accidents based on country.

Location, Country, Latitude, Longitude, Airport Code, Airport Name

Relevant attributes: These are attributes that were used to train all the predictive models presented in this project. Except the counting attributes, all attributes were preprocessed using one-hot encoding scheme (aka. one-of-K scheme). Aircraft Damage, Aircraft Category ,Registration Number ,Make ,Model ,Amateur Built ,Number of Engines ,Engine Type ,FAR Description ,Schedule ,Purpose of Flight ,Air Carrier

```
> print(relevant_attributes)
 [1] "Event.Id"              "Investigation.Type"    "Event.Date"
 [4] "Location"              "Country"               "Injury.Severity"
 [7] "Aircraft.Damage"       "Aircraft.Category"     "Make"
[10] "Model"                 "Amateur.Built"         "Number.of.Engines"
[13] "Engine.Type"           "FAR.Description"       "Schedule"
[16] "Purpose.of.Flight"     "Air.Carrier"           "Total.Fatal.Injuries"
[19] "Total.Serious.Injuries" "Total.Minor.Injuries" "Total.Uninjured"
[22] "Weather.Condition"     "Broad.Phase.of.Flight" "Report.Status"
```

International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 6, Issue 4, July- August 2017 ISSN 2278-6856

Irrelevant Attributes: These attributes are irrelevant to the analysis of factors

```
> print(irrelevant_attributes)
[1] "Accident.Number"      "Latitude"            "Longitude"
[4] "Airport.Code"         "Airport.Name"        "Registration.Number"
[7] "Publication.Date"     "geo_point"
```

Figure 14: Irrelevant Attributes

Target attribute: Below attributes are used to predict the number of injured passengers based on aircraft model and weather condition.

```
> print(target_attributes)
[1] "Total.Fatal.Injuries"   "Total.Serious.Injuries" "Total.Minor.Injuries"
[4] "Total.Uninjured"        "Weather.Condition"      "Broad.Phase.of.Flight"
[7] "Report.Status"
```

Figure 5: Target attributes

C..Logic Regression

The Logistic Regression used in this project is the Logistic Regression present in R library, which in turn uses the LIBLINEAR implementation of the Logistic Regression.

The LIBLINEAR implementation solves the following optimization problem:

$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi(w; x_i, y_i)$$

Given a set of instance-label pairs (xi ,yi), i = 1, . . . , l where C is the penalty parameter and

ξ(w; xi , yi) is the loss function,

This for Logistic Regression is:

$$log(1 + e^{-y_i w^T x_i})$$

In this paper, L2 regularized Logistic Regression with the penalty C equal to 1.0 is used.

D.Random Forests

Random Forests from R Library, used as a binary classifier and also to evaluate the feature importance in order to understand which are the most important factors while predicting the Aviation Accident Risk. Random forests are a combination of tree predictors, where each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The implementation combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class. Random Forests were also used to assess injury risk and the importance of factors

A relationship between random forests and the k-nearest neighbor algorithm (k-NN) was pointed out by Lin and Jeon in 2002.[19] It turns out that both can be viewed as so-called weighted neighborhoods schemes. These are models built from a training set $\{(x_i, y_i)\}^n_{i=1}$ that make predictions y^ for new points x' by looking at the "neighborhood" of the point, formalized by a weight function W:

Here, $W(x_i,x')$ is the non-negative weight of the i'th training point relative to the new point x' in the same tree. For any particular x', the weights for points must sum to one.

This shows that the whole forest is again a weighted neighborhood scheme, with weights that average those of the individual trees. The neighbors of x' in this interpretation are the points sharing the same leaf in any tree j. In this way, the neighborhood of x' depends in a complex way on the structure of the trees, and thus on the structure of the training set. Lin and Jeon show that the shape of the neighborhood used by a random forest adapts to the local importance of each feature.

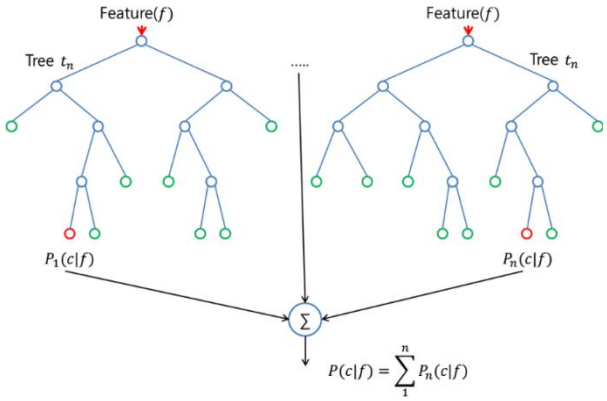


Figure 16: Random forest algorithm

E. K-nearest neighbors

The K-nearest neighbors (kNN) algorithm used in this project is also from R Library, which provides both unsupervised and supervised neighbors-based learning methods.

Despite the simplicity of the algorithm, kNN has been successful in a large number of classification and regression problems.

Attribute scaling is also performed before using kNN, to ensure that the distance measure accords equal weight to each variable.

4. RESULTS

In this paper, we are using below tools to analyze the NTSB data and to predict the aviation accidents.

- ggplot2: To plot heat maps with the geospatial distribution of the accidents, ggplot library is used
- dplyr: To provide data analysis, dplyr, an open source library providing high-performance, data structures and data analysis tools for R programming language.
- randomForest : This project also uses randomForest – an open source Machine Learning R library for implementing Random Forest algorithm.
- summary: to get the aggregate values and summery

information of the accident data, summary function is used in this project

- data.table : An alternative way to organize data sets for very, very fast operations. Useful for big data/large data sets.

As shown in the Figure. The data set has nulls which will impact the data analysis. Removal of the nulls is done as follows

So, using mutate() method, all the nulls are removed as shown the figure.

```
> dfp2 <- dfp %>% mutate(Total.Serious.Injuries = ifelse(is.na(Total.Serious.Injuries),0,Total.Serious.Injuries))
> dfp2 <- dfp2 %>% mutate(Total.Minor.Injuries = ifelse(is.na(Total.Minor.Injuries),0,Total.Minor.Injuries))
> dfp2 <- dfp2 %>% mutate(Total.Uninjured = ifelse(is.na(Total.Uninjured),0,Total.Uninjured))
> dfp2 <- dfp2 %>% mutate(Total.Fatal.Injuries = ifelse(is.na(Total.Fatal.Injuries),0,Total.Fatal.Injuries))
> head(dfp2[18:21])
  Total.Fatal.Injuries Total.Serious.Injuries Total.Minor.Injuries Total.Uninjured
1                    0                      1                    1               0
2                    0                      0                    1               1
3                    0                      1                    0               0
4                    3                      0                    0               0
5                    1                      0                    0               0
6                    1                      0                    0               0
```

Figure 17: Removing the NULLs and NAs

ggplot2: To plot heat maps with the geospatial distribution of the accidents, ggplot library is used

```
> ggplot(subset(dfp2, Country %in% c("Canada", "Brazil", "United Kingdom", "Mexico")),
+        aes(x=Aircraft.Damage,
+            y=Total.Fatal.Injuries,
+            color=Country))+
+   geom_point()
```

Figure 18: Data distribution query

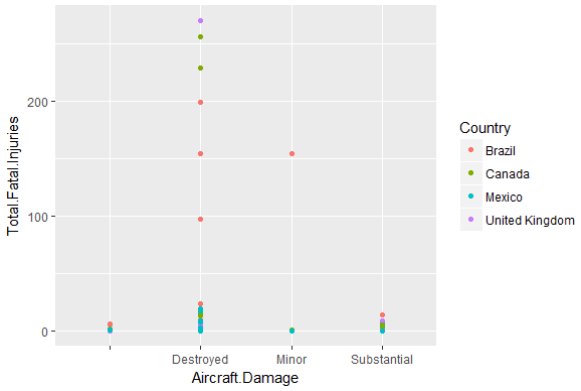


Figure 19: Data distribution

Every flight is different, but in some cases accidents follow well-formed patterns. Whether heedless, hapless, or simply clueless, pilots keep falling into the same traps that have snared others before them and that is the main motto of analysis in this paper. Weather can be a factor to have accidents.

IMC – Instrument Meteorological Condition - this means

$$\hat{y} = \sum_{i=1}^{n} W(x_i, x')\, y_i$$

flying in cloud or bad weather.

VMC - Visual meteorological conditions - It is an aviation flight category in which pilots have sufficient visibility to fly the aircraft.

Injury Severity: Below analysis shows the accidents against the injury severity.3. How many Air accidents happen based on the different types of injury and fatalities?

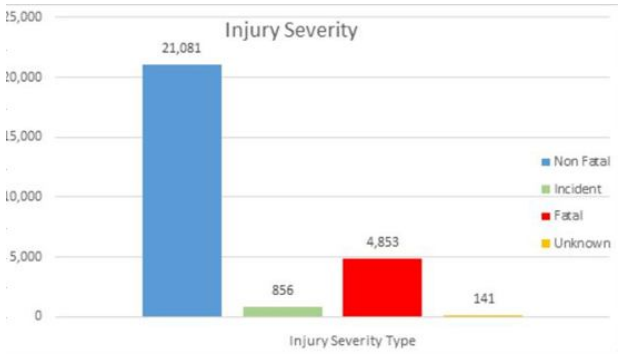


Figure 20: Accidents Analysis by Injury Severity

There were 21,063 Non-Fatal accidents, but 2,671 accidents where it was at least one fatality. See harts below.

Phase of Flight Occurrences: Below analysis shows the accidents against the phase of flight.

Many aviation occurrence reporting systems capture the phase of operation or the phase of flight in which the event that is to be reported occurred. Analyzing this factor, we are obtaining that accidents have mostly occurred when Landing is performed

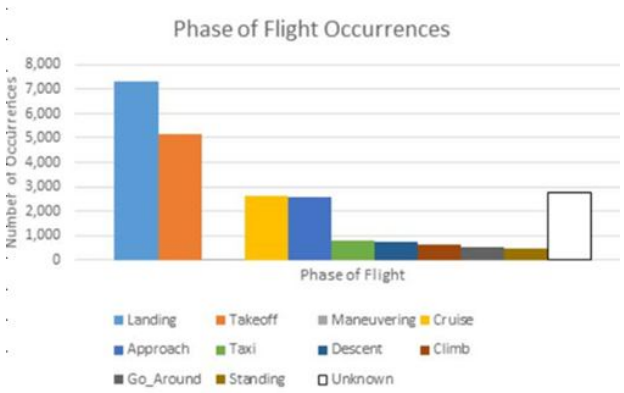


Figure 21: Accidents Analysis by Flight Phase

5. CONCLUSIONS AND FUTUREWORK

In this paper, the proposed analysis method for aviation data gives the various predictions of the data. As the data is to analyze the existing accidents, based on this results, prediction of the accidents severity based on various conditions like weather, flight model and making tear sets.

This analysis is very helpful to reduce the future accidents. In this paper, various analyzations have been provided with respective to weather, flight making model, making year, coordination's so that the impact of an accidents can be

fore seen and take the necessary precautions to rescue the passengers.

With the analysis in R programming, there might be some data misses while analyzing so in future, performing the same analysis in bigdata can get more accuracy

REFERENCES

[1] S. Bobrowski, The Force.com R analytics, Force.com Architects White Paper Series, 2010., pp. 889–896
[2] M. Adrian and N. Heudecker, Hadoop 2017: The Road Ahead, Gartner Webinars, Nov. 2016.
[3] https://wikipedia.com
[4] C. J. Guo, W. Sun , Y. Huang, Z. H. Wang, and B. Gao, "A framework for R analytics for prediction," in Proc. IEEE CEC and EEE, Tokyo, Japan, July 2007, pp. 551–558.
[5] T. Kwok, T. Nguyen, and L. Lam, "aviation accidents data," in Proc. IEEE SCC, Honolulu, Hawaii, Jun 2008, pp. 169–186.
[6] Unkel S, Farrington CP, Garthwaite PH, Robertson C, Andrews N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2012;175(1):49–82.
[7] NARA (National Archives and Records Administration). 2015. "Airworthiness Directives; Airbus
[8] "Airworthiness Standards; Crash Resistant Fuel Systems in Normal and Transport
[9] NASA (National Aeronautics and Space Administration). 2003. Columbia Accident Investigation
[10] Board, Report Volume 1. Washington, DC: NASA. Accessed January 30, 2017.
[11] http://www.nasa.gov/columbia/home/CAIB_Vol1.html.
[12] NTSB (National Transportation Safety Board). 2016. Crash-Resistant Fuel Systems on Airbus
[13] Helicopters. ASR-16/02. Washington, DC: NTSB.
[14] Crash Following Loss of Engine Power Due to Fuel Exhaustion, Air Methods
[15] Corporation, Eurocopter AS350 B2, N352LN, Near Mosby, Missouri, August 26, 2011. AAR-13/02. Washington, DC: NTSB.

Author

T.Nimeelitha Reddy received the bachelor's. Degree in Information Technology from S K University and she is currently pursuing Master's in Computer Science from G.Pulla Reddy Engineering College, Kurnool, A.P. Her research interests include multi-tenant system, big data analysis.

Smt.M.Padmareceived bachelor's Degree in CSIT from JNTU-H and Master's degree in Computer Science& Engineering from Sathyabama University, Chennai. She is currently an Assistant Professor in Dept.of Computer Science& Engineering in G.Pulla Reddy Engineering College, Kurnool, A.P and pursuing her Ph.D from Rayalaseema University, Kurnool.. Her research interests include big data and IOT.