|  |  |
| --- | --- |
| https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcQWwf11bMZR-Abi8gWubzDusZ_UBMOnvbf-QG1i_5GfshW_aQUn  gROUP 4 – Final REPORT  Data Mining For Business Intelligence | “Flying towards a safer tomorrow…”  Team  Alok Nayak, Chirag Shetty, Karan Nandwani, Prag Chinnappan |

Table of Contents

[1. Problem description and current state of affairs 2](#_Toc406029083)

[2. Data description and descriptive statistics 2](#_Toc406029084)

[a. Origin: 2](#_Toc406029085)

[b. Frequency: 3](#_Toc406029086)

[c. Dataset and statistics: 3](#_Toc406029087)

[3. Data preprocessing results 4](#_Toc406029088)

[4. Algorithms Used 5](#_Toc406029089)

[a. Decision tree - C5.0 5](#_Toc406029090)

[b. Bayes Network in SPSS 6](#_Toc406029091)

[5. Describe your proposed evaluation 7](#_Toc406029092)

[6. Conclusion. 9](#_Toc406029093)

[7. Wildcard 9](#_Toc406029094)

# Problem description

Our project focuses on the goal of predicting the safety aspect of an airplane assuming that it has unfortunately been in a crash. We want to analyze why there are higher fatality in certain types of flights, and provide the outcome of our analysis potentially to NTSB (National Transport Safety Board), Airline companies, Aircraft manufactures and other federal authorities and “Make Flying Safer”.

With the growth in technology, speed and capacity of the aircrafts have increased over the years. At the same time the number of passengers using air as a mode of transportation has increased by 1.3% between 2012 and 2013, from 815.7 million to 826 million passengers. With an increase in the number of people using air transport, the safety of airplanes will be imperative for any carrier to ensure continued engagement with the customer. Through data analysis in this project we have looked at instances where a crash has occurred and evaluated the attributes that constitute a particular crash scenario, and tell us if in an event of a crash the percentage of fatality will be high (100%) or if the flight survived the crash and no fatalities occurred.

Once we have been successful in building an algorithm that will tell us if a crash will result in a 100% fatality rate, or if the plane survived with no casualties, these analyses can be used by airplane carries, airport authorities and plane manufacturers to evaluate the various conditions under which a crash results in a 100% fatality rate. The 3 partners can then collaborate to make conditions conducive to ensure that even if a crash occurs, it will not result in any casualties. This will definitely help the customers trust the airline carrier to use them for their regular commute and recreation.

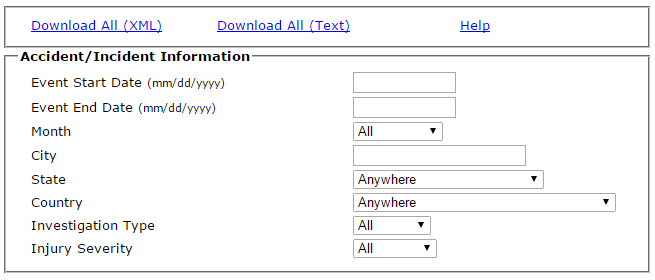
# Data description and descriptive statistics

## Origin:

Every year, while thousands of accidents happen across the world, many go unreported. Our project analyzes the details about the accidents that were reported to NTSB. When the accidents are reported to NTSB, personnel from the board visits the crash site in person and investigate the site, flight and fatality and collect all the details. These details are verified by team of NTSB board personnel again and then formally uploaded to NTSB site.

<http://www.ntsb.gov/aviationquery/>

The data set can be filtered based on start date and end date of the accidents, city or country where the accident happened and injury details.



## Frequency:

NTSB database contains accident data as old as 1965 to the accidents that happened last week. The database is updated with accident details as soon as they are reported and investigation is completed by board members.

## Dataset and statistics:

Initial dataset collected from NTSB had 75K records of reported accidents from across the world. Record for each of the accident contains details about the location of the crash, near by airport, aircraft make and model , no of engines, broad phase of the flight, injury and fatality details. Descriptive statistics on the key columns revealed some interesting information such as average damage for aircraft was substantial, and most of the aircraft were not altered after manufacturing, average injuries in different categories such as fatal, serious, minor and uninjured. Independent variables in our dataset include variables such as location, airport name, aircraft make and model, aircraft damage, whether airplane is amateur build or not and phase of the flight. Dependent variables in the data set include total fatal injuries, serious, minor and uninjured for each accidents.

Below screenshot gives a brief idea about the descriptive stats of the key variables.



As we ran the descriptive statistics, we found that on average aircrafts that met with accidents had more than substantial damage. On average fatalities was higher than average number of serious and minor injuries. This is an alarming and frightening factor as the any aircraft accident has risk of fatality. Main rationale for choosing the dataset was the exponential growth we seen the airline industry and severity of the accidents and impact of accidents on airlines, authorities, safety board and people involved. We want to predict which are the key variables that result in higher rate of fatality in the accidents.

# Data preprocessing results

Dataset that was downloaded from NTSB site was in text format, hence we had to convert it into excel format for preprocessing and doing some calculation. As a part of data cleaning, we removed rows (accident report) that had null values. Also the nominal data with unique values will not help our project to predict the output data , during data preprocessing, the team removed the columns that contained nominal values such as eventid, report number. Also there were transitive dependencies between some of the variables such as latitude and longitude and airport code which were similar to location.

Moreover variables such as aircraft damage, custom built, phase of the flight and few other variables had nominal data and this may not help in calculating the averages or running the algorithm. Also before running the algorithm, we categorized percentage fatality which was the continuous variable into different bins. This bin will help us in running the predictive mining software’s and come up with suggestions.

Before preprocessing we have around 75,000 records in our data with 31 attributes. After preprocessing we had 60,000 rows and 14 attributes.

# Algorithms Used

## Decision tree - C5.0

**Overview of how the algorithm works**

First algorithm that we are going to use in our project is decision tree. Decision tree works based on divide and conquer concept, were in dataset at the top will be divided based on key determinant variable and usefulness of the information is calculated (through entropy which is uncertainty measure). After entropy calculation at the first level, decision tree divides the data sets in the second level and further based on other determinant variable and calculate the entropy at each level. Goal of the algorithm is to have a class of objects that all have the same label. If all the objects belong to the same class, then algorithm stops else data set is further divided into disjoint sets C1,C2 and start over. Based on entropy reduction at each subsequent level, algorithm chooses the variable that gives good results or little confusion. We used the C5.0 decision tree classifier in the SPSS modeler to create our decision tree. We had an option between the conventional C4.5 and C5.0 and chose the C5.0 for the following reasons. They are;

* *Accuracy:* The C5.0 rulesets have noticeably lower error rates when compared to C4.5
* *Speed:* C5.0 is much faster; it uses different algorithms and is highly optimized.
* *Memory:* C5.0 commonly uses an order of magnitude less memory than C4.5 during ruleset construction.

Px = Probability of having label x \* Log of that probability

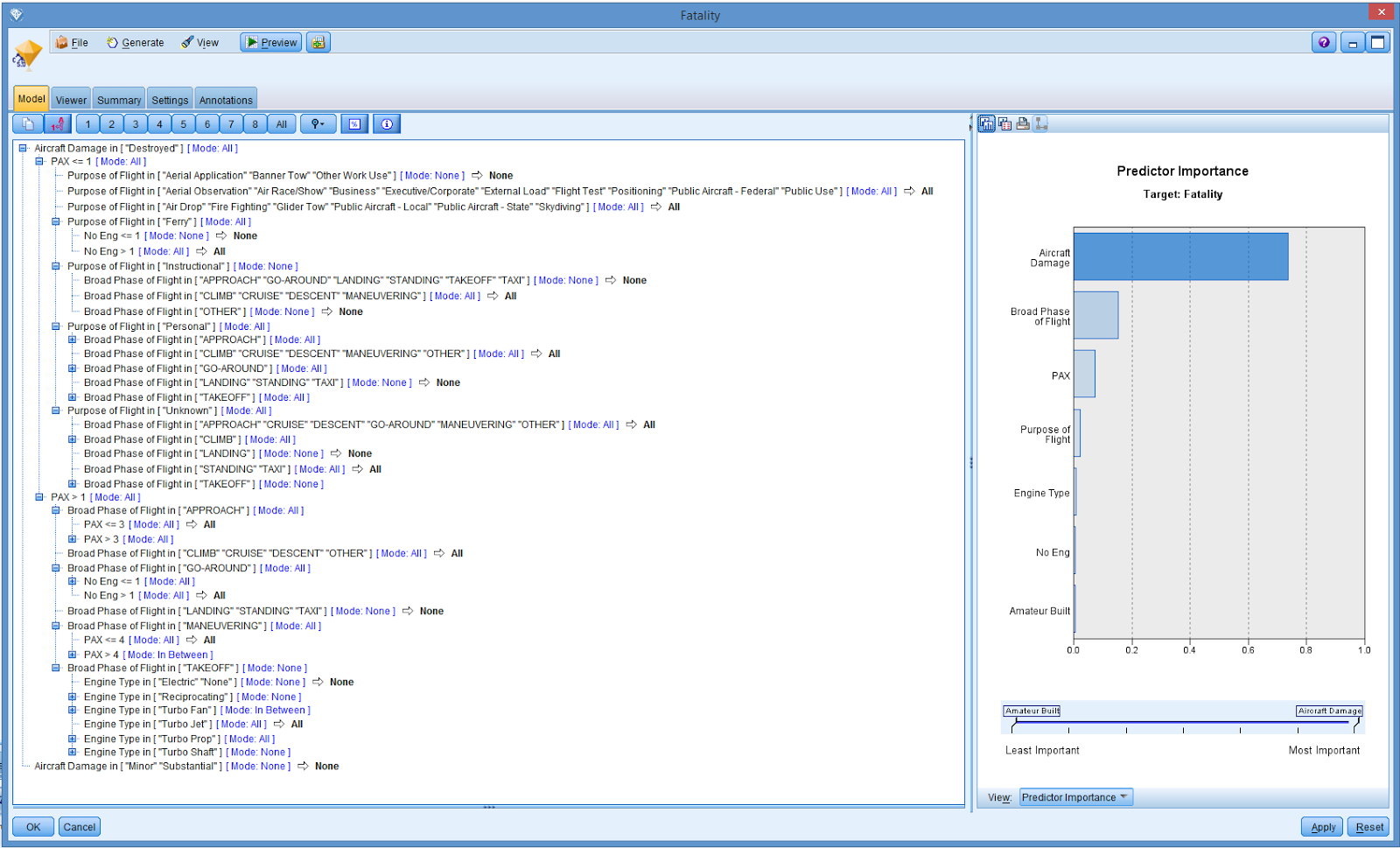
Entropy after splitting into classes C based on attribute = Sum of branch probability \* entropy of the subclass

**Why it’s a good choice for our goal.**

Primary objective of our project is to predict the fatality rates in the aircraft accidents. Team decided to use decision tree algorithm, as the algorithm supports the multi-variable input to predict the output and also predict the discretive values. Fatality rate of the accidents will have 3 bins low, inbetween and high, and input consists of location, airport name, aircraft make and model, aircraft damage, amateur build and phase of the flight. Algorithm will take these inputs values and run the decision tree and predict the output based on the influence of different variables and how much entropy is reduced and how accurate is the prediction. Our predictions are classified into 3 bins; 100% fatality, 0% fatality and in between. When decision analysis in required for predictions, a decision tree is useful to predict outcomes. In the future, if any additional scenarios have to be added (weather data) we can incorporate this data without much difficulty into the decision tress. Another major reason is because the attributes in our data have nonlinear relationships, this may result in algorithm fail making the model invalid. Decision trees do not require any assumptions of linearity in the data. Thus, we can use them in scenarios where we know the parameters are nonlinearly related. Finally, since our business partners are airplane manufacturers and airport authorities, we assume that they will not have much data analysis and tools experience and using decision trees will make it easier for us to explain to them and for them to understand it.

**Results of running the algorithm.**

Majority class in our data set is none (Bin Fatality), and the baseline probability of majority class is 81%. Below is the result of algorithm.



Here we have shown a screenshot of our team running the C5.0 on SPSS modeler and the results of the impact of the attributes on the outcome. At the end of every line of the decision tree, we have a triangle which denotes the predictability of the particular attribute on the dataset.

## Bayes Network in SPSS

**Overview of how the algorithm works**

A Bayesian network or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). Bayesian networks are DAGs whose nodes represent random variables in the Bayesian sense: they may be observable quantities, latent variables, unknown parameters or hypotheses. Each node is associated with a probability function that takes, as input, a particular set of values for the node's parent variables, and gives (as output) the probability (or probability distribution, if applicable) of the variable represented by the node. For example, if m parent nodes represent m Boolean variables then the probability function could be represented by a table of 2^m entries, one entry for each of the 2^m possible combinations of its parents being true or false.

Bayes Rule:

For any two events, A and B,

p(B|A) = p(A|B) x p(B) / p(A)

where you read 'p(A)' as "the probability of A", and

'p(A|B)' as "the probability of A given that B has occurred".

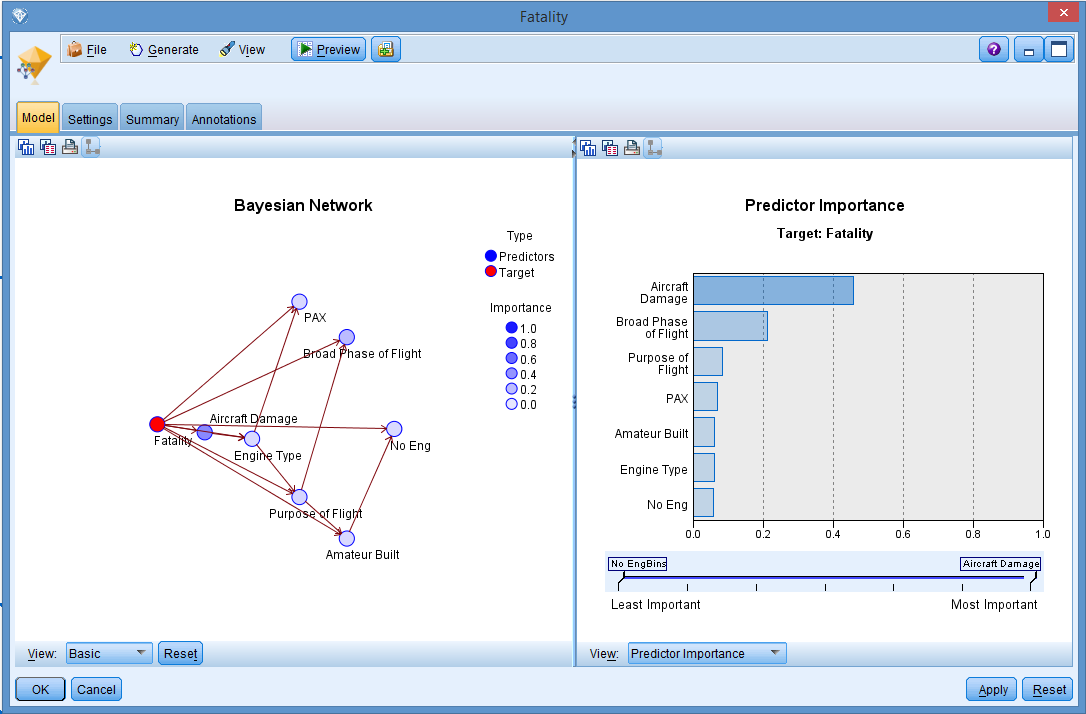
It turns out that Bayes' rule is very powerful and is the basic computation rule that allows us to update all the probabilities in a net, when any one piece of information changes.

**Why it’s a good choice for your goal**

There are several reasons why we decided to use the Bayes Network for our dataset. They are as follows;

It is a solid modeler for decision theory. They can be used for computing the predictive distribution on the outcomes of possible actions. The Bayes Network has good smoothness properties. This means that re-modelling is easier when values of attributes are changed. This will be helpful once our business partners have implemented recommendations based on our analysis and we can evaluate outcomes based on future crashes. Another important reason why we decided to use Bayes network for our data is because of its ability to handle different variable types. Since we have a combination of continuous variables such as number of passengers (PAX) and discrete variables such as aircraft damage or phase of flight, Bayes Network handles these different variables well and prevents unexpected results due to variability of data type.

**Results of Bayes network**



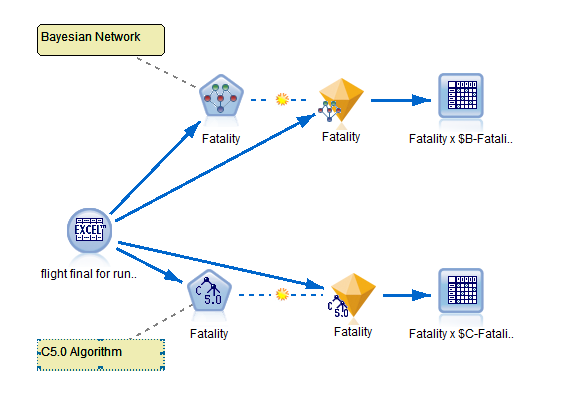
Here we have shown a screenshot of our team running the Bayesian Net on the SPSS modeler which includes the graphical representation of the data set and nodes with our dataset and the relations shared by these nodes.

# Describe your proposed evaluation

After preprocessing our team came up with three categories in fatality: first category is None which means none of the passengers are injured in the accident, All means all the passengers travelled in the aircraft died and in between has range of fatality between 0 and 100%. Majority class in our dataset was “None”, where none of the passengers were fatally injured. These data will be fed to the predictive algorithms using SPSS, where in data from excel will be loaded and then run the algorithm and connect results to display the output.

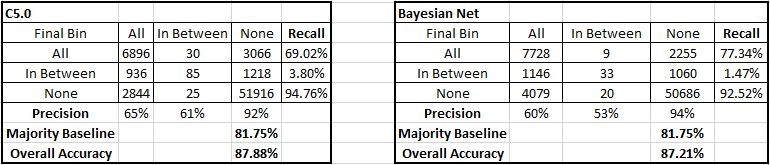
**Execute the evaluation**

Below screenshot describes how the data is processed in SPSS. First step was to load the data from excel into the SPSS and then connect the input to the algorithm (in this case C5.0 and Bayes Net) and then select input variables and run the algorithm. After execution is complete, we connected the results to the output and analyzed final predictions.



**Discuss the evaluation:**

Predictions from both the algorithms are better than our majority class baseline probability. This is due to good data preprocessing and feed the appropriate variables to the algorithms.



# Conclusion.

The majority class for our data set is "None". “None” implies there were no fatalities for that tuple. The majority baseline is 81.75% and the accuracy for predicting the majority class is 94.76% and 92.52% for C5.0 and Bayesian Network respectively. Our overall model accuracy is 87.88% and 87.21% for C5.0 and Bayesian Network respectively. Hence, we can conclude that our predictive models are performing better than majority class prediction. However, for the non-major class i.e. “All” and “In Between”, the accuracy is not as good. We will try to improve the prediction accuracy for these classes by re-training the algorithms, experimenting with different ways of partitioning the data set and tweaking algorithm specific attributes such as “Pruning Severity” & “Minimum Records per Child Branch” for C5.0 and “Likelihood Ratio” for Bayesian Network.

# Wildcard

Over the years, steps and constructive measures taken by manufactures and other transport authorities have made significant improvement in the flight safety. But accidents are still happening and people are losing lives.

As a next step in our project, we want share our analysis and outcome with NTSB (National Transport Safety Board), Airline companies, Aircraft manufactures and other federal authorities. This will help the concerned authorities in making modification to the manufacturing, aircraft designs, also update the install, service and safety procedures.