Analysis of Potential Terrorist and Perpetrators

Khalid bin Huda Siddiqui EP-1449041

Data Mining Techniques

University of Karachi

31-July-2017

Supervisor: Miss Mariam Feroz

CONTENTS

Project Link: https://github.com/binhudakhalid/Analysis-of-Potential-Terrorist-And-Perpetrators-

ABSTRACT

Terrorism is one of gravest problems that society faces today. It is an issue of global concern. The presence of terrorism can be felt all across the globe. It is, today, a much debated issue in all the countries of the world-developing or developed. The perpetrators of acts of terrorism can be individuals and groups. According to definitions state actors or non-state actor may also carry out terrorist acts outside the framework of a state of war. However, the most common image of terrorism is that it is carried out by small and secretive cells, highly motivated to serve a particular cause and many of the most deadly operations in recent times (Sageman, Mark 2004), such as the September 20 attacks, Islamabad Marriott Hotel bombing etc.

In this research we analysis the potential a person or individual can be a Terrorist or a Perpetrators, not group or nation state and to achieved this objective we examine various factor to predict the percentage a person can be a Terrorist or can help them to carry out the attack.

INTRODUCTION

Terrorism is widespread around the world and the most important question is What turns people toward violence or bad activities and to answer this millions of dollars of government-sponsored research had been done but there is still nothing close to a consensus on why someone becomes a terrorist. After Sept. 11, 2001 attack, A researcher Alan B. Krueger out tested widespread assumption that “Poverty was a key factor in the making of a terrorist” but found out that there is no link between economic stress and terrorism. After almost sixteen year the government officials and other organization still believes that money problems as an indicator of radicalization. Some studies suggest that the terrorist are educated or extroverted and other uneducated are at risk. and to solve this issue we should know about the peoples in our society that are likely to become the terrorist or a person who support those terrorist by arranging the

individual in a class of Religion, Sectarian and Race, In order to stop terrorist attack we have to examine the factor and characteristic that a person possessed or the percentage that a person can be a terrorist. The factor include age, religion, educational background , military training or other kind of madrasa tanning. To achieved this objective we apply different type of Data mining Techniques to our Dataset such as Classification and Clustering. Hence, the main Focus of this research is to analyze and interpret the percentage of a person is terrorist from a given set of data by applying classification techniques. Also, we complete the research by classifying of what sorts of people were likely to become Terrorist. In this project, we apply the tools of machine learning. The whole project was implemented and visualized in R programming.

2.1 RESEARCH FOCUS

Our Research is mainly Focus to identify the Terrorist or Perpetrator before they do Terrorist Attack or help them to carry out the Attack. In today world hundred of a people killed by Terrorist Attack, left thousands of People Wounded. As Terrorist Attack led to such loss of life. It is as predicted, that a set of people could become terrorist before they carried out Terrorist Attack. To to-do this, we use an existing data set which contains the case statuses (Military Tanning = Yes or No), and a set of feature parameters where the cases are filed. We then analyze and interpret the factors that affect various cases statuses like, Terrorist or Not Feature selection that is which feature is more significant while classifying innocent person. This will be a major research problem.

1. Complete analysis of what sort of people were likely to Become Terrorist.
2. Tuning of various parameters of each model will be a research problem (Decision Tree regression model with evaluations)
3. Apply Random forest classification Algorithm and other Techniques.

2.2 OTHER IMPORTANT QUESTION

The research question are i) Does All Muslim are Terrorist. ii) Terrorist and Age iii) All student from Madrasa are Terrorist. iv) Educated People are not Terrorist. v)The role of Military Training.

METHOD

SAMPLE

The population of interest for this study is all the people from age having age above 15. The sample are collected from different source such as data-word public safety and sjhaveri. This include people living in different part of the world. This sample size is of 912 persons.

DATA MINING TECHNIQUES

To do research and solve some problem we use various datamining method and techniques.

DATA PREPERATION

Data cleaning

In Data Cleaning Step we analyzed the dataset, cleaned it and remove noise and inconsistent data We remove Column that are not good for analysis such as plot, Status of Case, Number of people Killed, stilled abroad ,weapon involved in plot, initial source, web documented , initial tip and few others. We remove noise by removing record that are not good for are analysis.

Data integration

In Data integration step we combine all are record and data from different source and move them into a single file. We merge the data from tommy Blanchard, Google and Sjhaveri.

We change the name of different column to make merge worked such as religion, Gender and other.

Data transformation (where data are transformed and consolidated into forms):

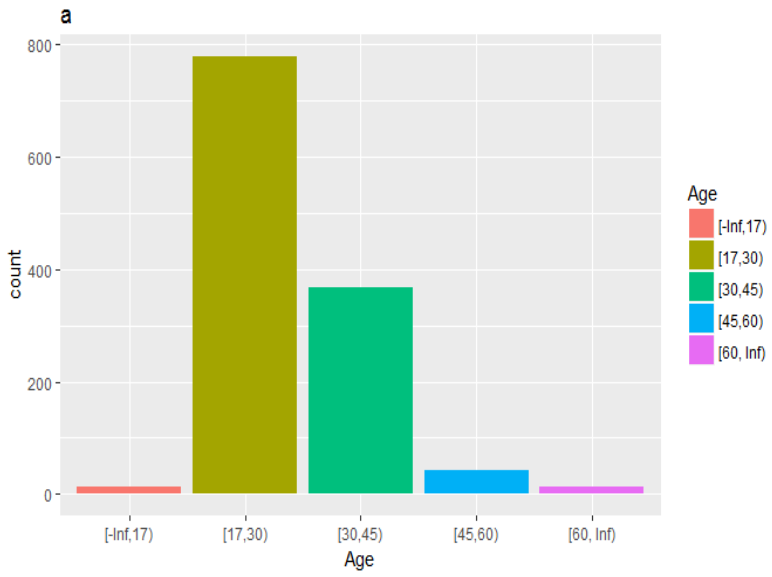
In data transformation we transform are data by merging different column and remaining them appropriately such as first name and last name column merge and renamed to just Name. The Home and nationality column merge and name it to Citizenship and other transformation steps we have done.

Data Description:

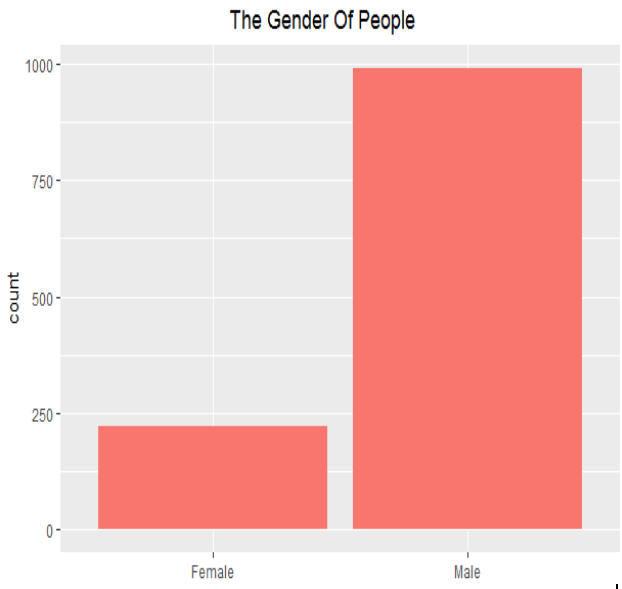
We take the dataset from google form and data world and It has total number of records were about 1299 . The is training set should be used to build your machine learning models. For the training set, an is terrorist was provided which was also known as the ground truth for each person Record. Our model was based on features like Person's gender and its Religion, Military tanning and other classes. The data set was used to check how well model performs on unseen data. In the test set, ground truth for each person was not provided and hence we predicted these outcomes of is Terrorist Class. For each person in the data set, Decision Tree model we trained to predict whether the person is Terrorist or Not. The basic data dictionary and the variables were declared and show in figure 1.

Data Visualization

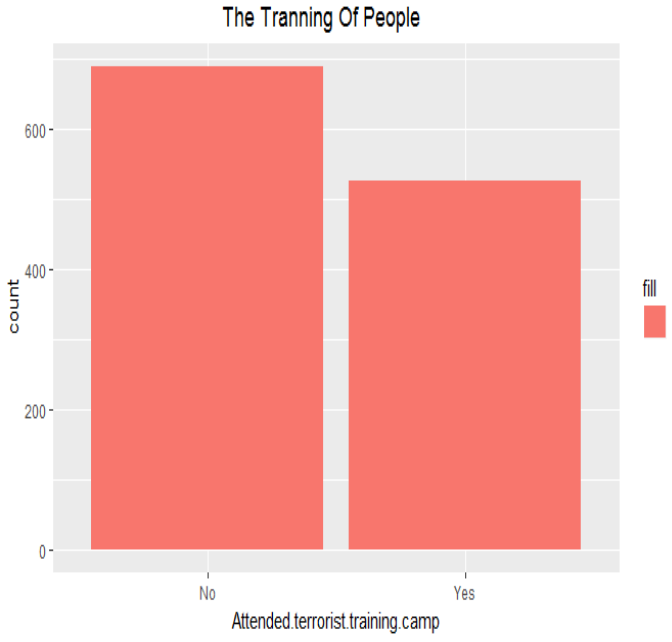
We explore the data and find some interesting fact about it some graph and plot are given blow and some are present at the end in figure 2.



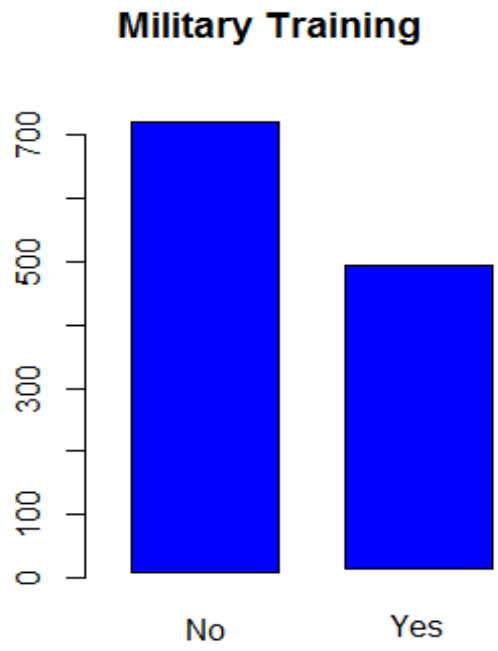
The above plot has x-axis with age category and y-axis has number of people. It shows that most of the person in our record have Age between 17 to 30.



The above plot has x-axis with Gender and y-axis has number of count. It shows that most of the person in our record have are Male.



The above plot has x-axis with categorical variable (yes or No) and y-axis has number of count. It shows that most of the person in our record have not attended Tanning Camp.



The above plot has x-axis with categorical variable (yes or No) and y-axis has number of count. It shows that most of the person in our record have not attended Military Camp.

METHOD USED

We use Classification and Clustering Analysis. In Classification we use Decision tree with rpart library and naïve Bayer classifier and Random forest. In Clustering Analysis we use K-Mode technique to find interesting cluster because it has categorical attributer.

Classification

Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels and our problem is a classification problem that's why we use various type of classification techniques.

Decision Tree

We of the technique we use is Decision tree, A decision tree is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. we model our problem on Decision tree by making are data attributes categorical and used two R libraries such as rpart and party. Our formula is

```
myFormula <- IsTerrorist ~ Marital Status + Educational +  Military + Madrasa training + Mental Illness +  Islam + Sex + Age
```

Model is

```
terro.ct1 <- ctree(myFormula, data = trainData)
```

Decision tree use gini index to slit. The Conditional inference tree with terminal node can found in figure 3.

CART Modeling via rpart

we use rpart`s Classification and regression trees, to grow a tree we use

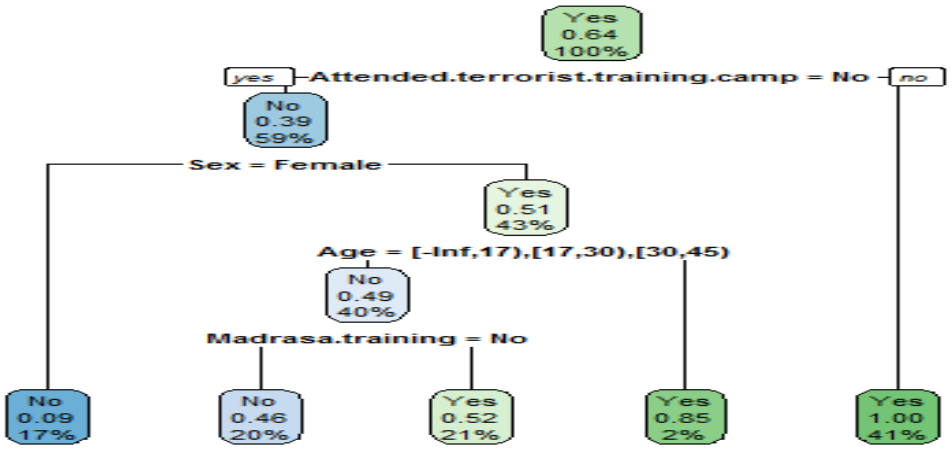
treemodel <- rpart(as.factor(IsTerrorist)~Age + Attended.terrorist.training.camp + Sex +

Madrasa.training + MaritalStatus ,data = train,method="class")

The result of our model is

Predcition	No	Yes
No	101	60
Yes	125	262

If you goes through the confusion matrix you will find out that the result is not that accurate and to make it accurate we have to use other datamining techniques.



From the above diagram we found out that there are more chance that people who attended training camp and gender is male and age is between 17 to 30 are more likely to become terrorist.

Random Forest

The Random Forest algorithm is one of the most widely used machine learning algorithm for classification. This algorithm, is operated by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. So for our problem is good for us as it estimates of what variables are important in the classification. The, advantage by using this algorithm is ,it reduces the chances of overfitting. High Model Performance and Accuracy are observed by using this algorithm.

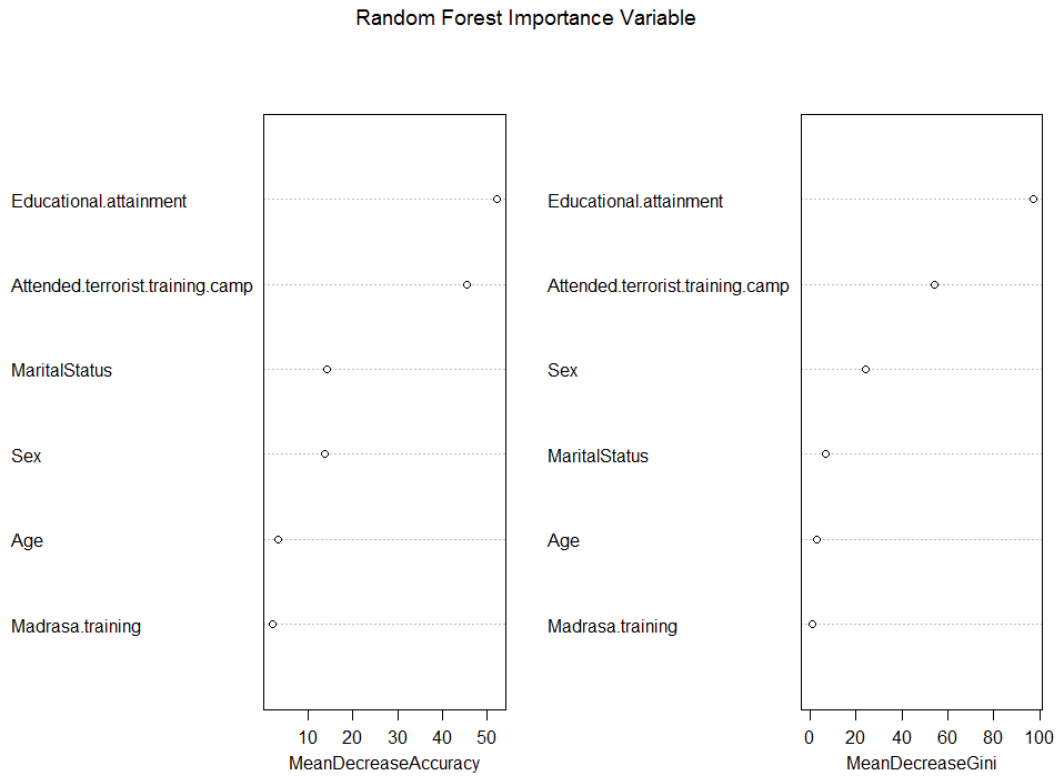
Confusion Matrix and Statistics

Prediction	No	Yes
No	193	68
Yes	0	346

Accuracy : 0.888
95% CI : (0.8601, 0.912)
Sensitivity : 1.0000
Specificity : 0.8357
Pos Pred Value : 0.7395
Neg Pred Value : 1.0000
Prevalence : 0.3180
Detection Rate : 0.3180
Detection Prevalence : 0.4300
Balanced Accuracy : 0.9179
'Positive' Class : No

From the above Table, we can see that the Accuracy of our Random Forest model is 0.888 which is quite good and Sensitivity is 1 and Specificity is 0.8357. Detection Rate is 0.3180 and The Positive Class is NO.

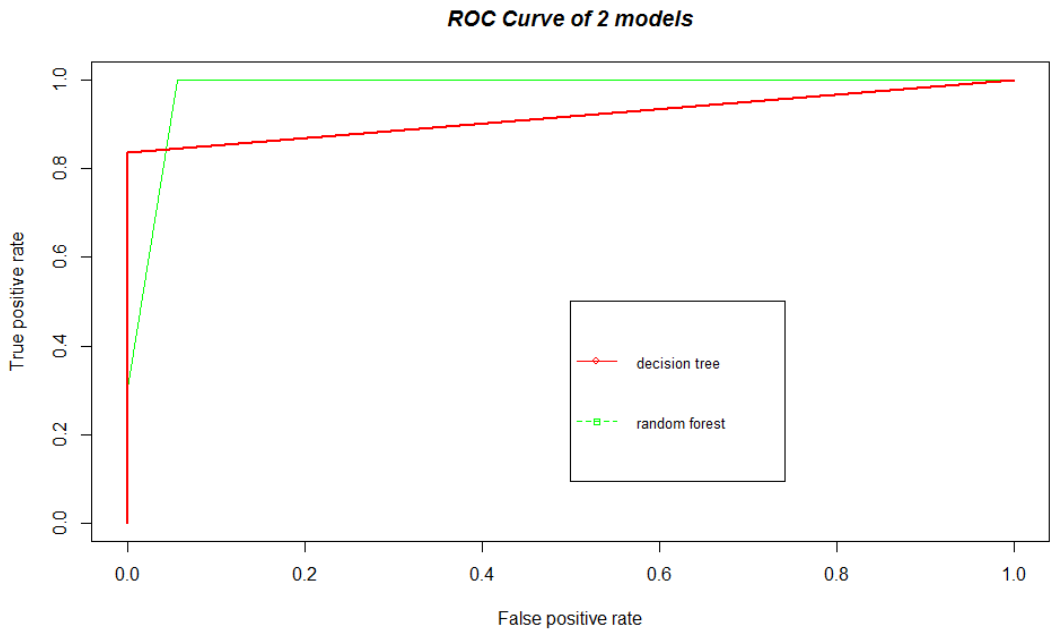
The Feature Importance Plot:



we can infer that the variable ‘Education Attainment has the highest significance on independent variable, then Marital Status, Sex , Age and Madrasa training has significance Respectively.

Evaluation Of Two Model

We use Roc Curve to Evaluate are machine learning models. ROC curves are commonly used to characterize the sensitivity/specificity trade-offs for a binary classifier. Most machine learning classifiers produce real-valued scores that correspond with the strength of the prediction that a given case is positive.



The above figure shows that the Random forest is more Accurate as compared to Decision Tree. Are data has move categorical variable than binary variable this may cause decision tree to have less accurate as compare to random forest.

CLUSTERING

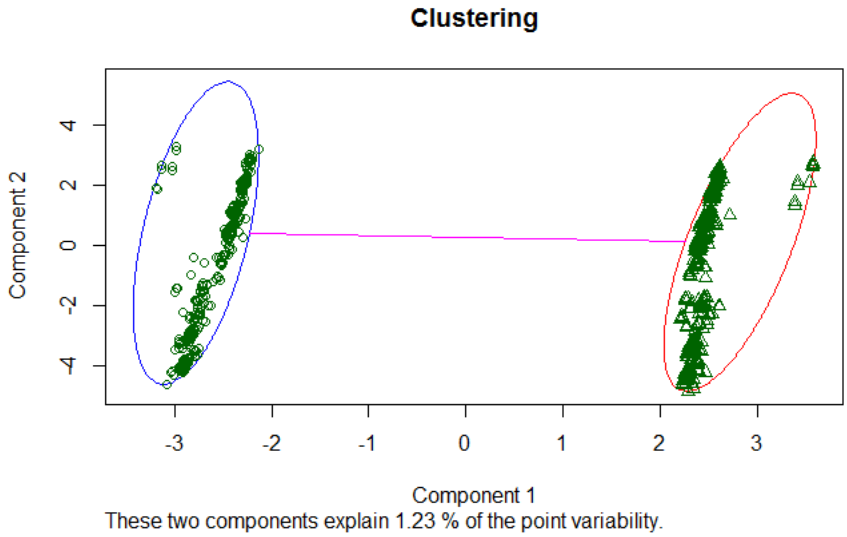
Although our problem is solved by classification`s random forest but we perform clustrin to find out more interesting factor about our dataset.Our data is categorial so we cannot use k-mean method. So we use Partitioning Around Medoids (PAM) technique with cluster library in r. we use first create a matrix

```
daisy.mat <- as.matrix(daisy(dataa, metric="gower"))
```

then we made clusters

```
my.cluster <- pam(daisy.mat, k=2, diss = T)
```

```
The cluster average is 0.4262890 and 0.4180452
```




```
The above plot we can see that there are two cluster and it shows the variabi
lity between component 1 and 2 is 1.23%.
```

RESULT

We have divided the data into training (80%) and testing (20%) by random sampling. And we find out that from naïve Bayer ,Decision tree, Clustering and random Forest the best for prediction is Random forest with accuracy level of 0.888.

OBSERVATIONS AND RESEARCH:

We observe some fact and finding that are given below, are answer to the important questions and graph can be found in figure 4 and 5 in table and Illustration section.

1) **TERRORIST AND AGE**

   If we go through the result of our analysis we find out that people aged from 17 to 30 are more likely to terrorist, as we find out that out of 540 observation of aged between 17 to 30 only 200 are not terrorist. So, The people from that age are more likely to become a terrorist as it is easy to manipulate these people. There is moderate relation between Age and Terrorist.

2) **All student from Madrasa are Terrorist**

   As we found out that the people who only get the education from madras and did not get Any other kind of education like from school are more likely to become terrorist as the Evil people use them, out of 600 people 400 hundred and terrorist.

3) **ALL MUSLIM ARE TERRORIST**

   If we go through the Graph between Islam and terrorist that, out of 600 terrorist in our dataset 400 are the people with other religion and 200 are Muslim and so there is no Direct relation between the Muslim and terrorist.

4) **Gender and Terrorism**

   There is a strong observation from the research is that most of male are terrorist as compare to female, from our data 95% male are more likely to be terrorist and very little change that female will become perpetrator.

5) **Education Level and Potential to become Terrorist**

   There is perception that educated people have less chance to become a terrorist but we found from our study that it is not completely true as people who have done graduation are also found terrorist in large number but it is true people who have no school education are more likely to become terrorist and there is strong relation between the twos.

6) **TRAINING CAMP**

   This is the most obvious that the people who get training from Terrorist camp are terrorist and there is a direct and Strong Relation between them.

CONCLUTION

From all the research and study of previous Research paper we find out that to get better prediction and classification we have to collect data from and all over the world and we have to use different rules to predict from Muslim and different rule from non-Muslim as for non-Muslim we found out that education, mental illness, Terrorist camp training, age and gender are main factor to predict the chance of becoming a terrorist but if you go to analyses for Muslim, you will found out that other factor are important like their firqa, people attack in bombing, relative died in drone attack and attended madrasa are more strong factor in make a prediction.

Limitations

1) Missing values were filled manually and sometime with Mode values. This may cause biasness and make our model less accurate.

2) Our Data set is very small. If our data have many record and have multidimensional and multicultural data sets then are prediction would be more accurate and can apply to other people and it helps us to know more fact about it..

Potential Improvements or Future Work

Accuracy of the models, can be improved by using machine leaning algorithms such as SVM(Support Vector Machine) and other classification algorithms. Also, by increasing the size of the dataset we can further interpret many facts such as “education from which Country” and “People and their Friends Data” and many more like their Facebook Data.

Bibliography

Alan B. Krueger, the Princeton economist ---, John. *Institutes of the Christian Religion*. 2 vols., ed. John T. McNeill, trans. Ford Lewis Battles. The Library of Christian Classics, vols. 20-21. Philadelphia: Westminster Press, 1960. *(Note there is both an editor and translator for this version of the* Institutes*.)*

Han, J. and Kamber, M. Data Mining Concepts and Techniques, 2nd ed . (*Morgan Kaufmann Publisher,2006.)*

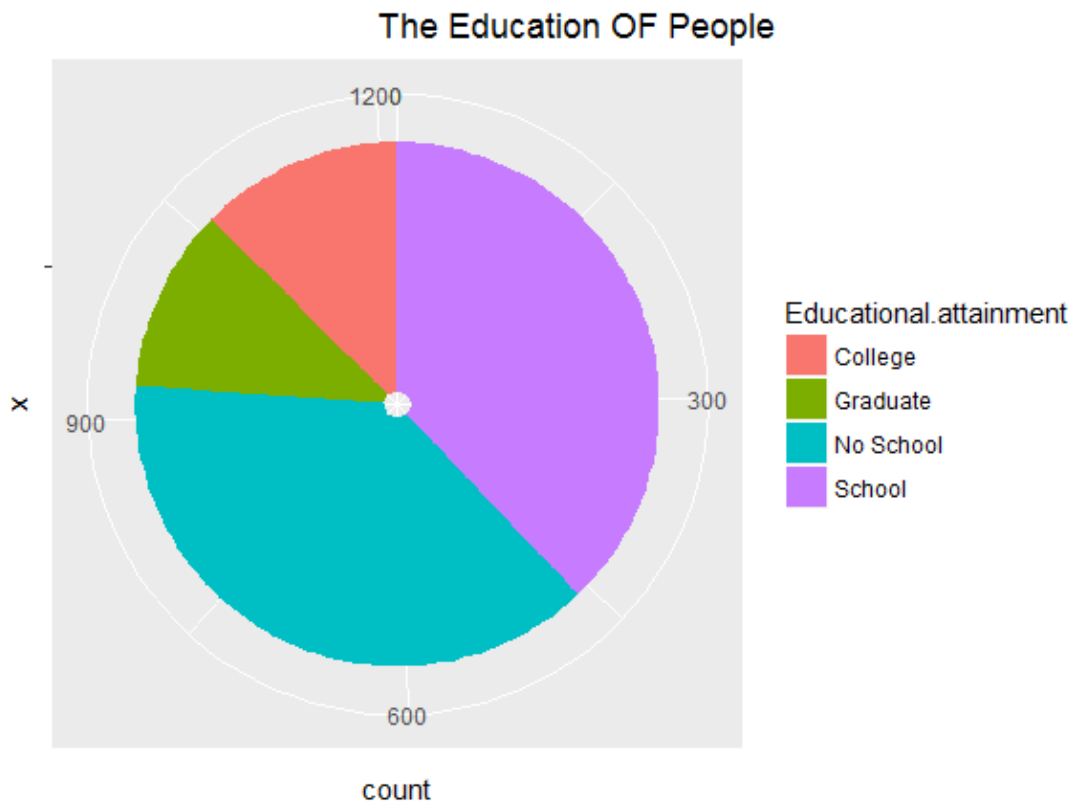
Sageman, Mark (2004). Understanding Terror Networks. Philadelphia, PA: U. of Pennsylvania Press. pp. 166–67. Berkhof, Louis. Systematic Theology. 4th ed. Grand Rapids.

Tables & Illustrations

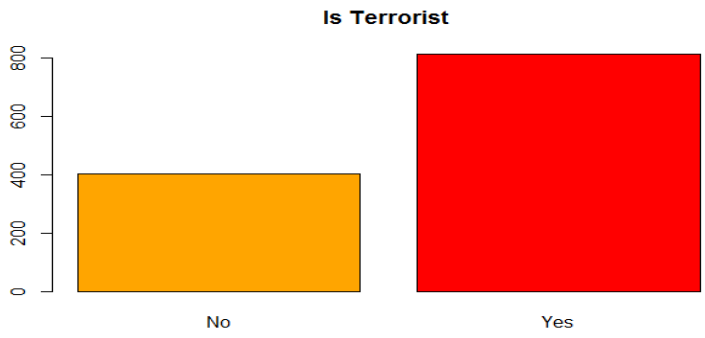
Figure 1, **Data Description**

Attributes	TYPE	DESCRIPTION
X	Integer	Serial Number
Name	String	Name of Person
Age	Integer	Age
Military	Categorical	Military Tanning
Terrorist.training.camp	Categorical	Terrorist Camp
Citizenship	String	Person`s Homeland

***Figure 2*, Data exploration and visualization**



The above Pie chart shows that Collage variable have pink Colour, Graduate have Green Colour, No School have B Color. It shows that in our record people with school and no school are equal and Graduate has least count.



The above plot has x-axis with categorical variable (yes or No) and y-axis has number of count. It shows that most of the person in our record are Terrorist.

***Figure 3*, Decision tree conditional inference**

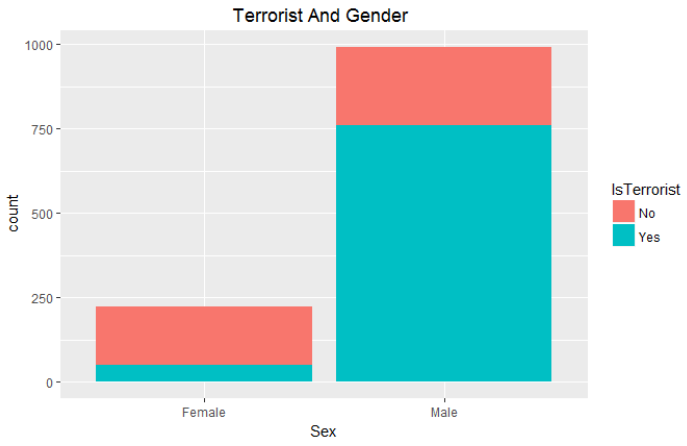
Conditional inference tree with 6 terminal nodes

Response: IsTerrorist
Inputs: MaritalStatus, EducationalAttainment, Military, MadrasAtraining, MentalIllness, Islam, Sex, Age
Number of observations: 637

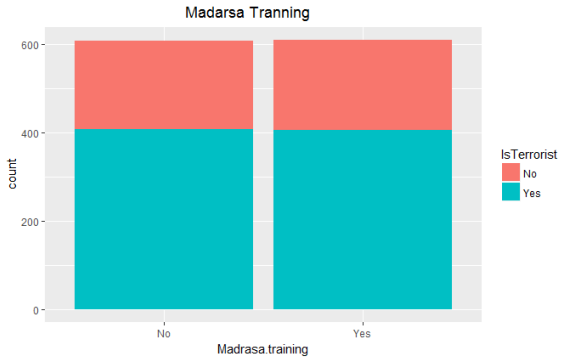
1) EducationalAttainment == {School}; criterion = 1, statistic = 192.077
2) Sex == {Female}; criterion = 1, statistic = 33.622
3)* weights = 29
2) Sex == {Male}
4) Military == {?, Unknown, Yes}; criterion = 0.961, statistic = 12.838
5)* weights = 24
4) Military == {No}
6)* weights = 87
1) EducationalAttainment == {College, Graduate, No School}
7) EducationalAttainment == {College, No School}; criterion = 1, statistic = 86.999
8)* weights = 431
7) EducationalAttainment == {Graduate}
9) Sex == {Female}; criterion = 0.998, statistic = 13.048
10)* weights = 7
9) Sex == {Male}
11)* weights = 59

Figure 4 **Result**

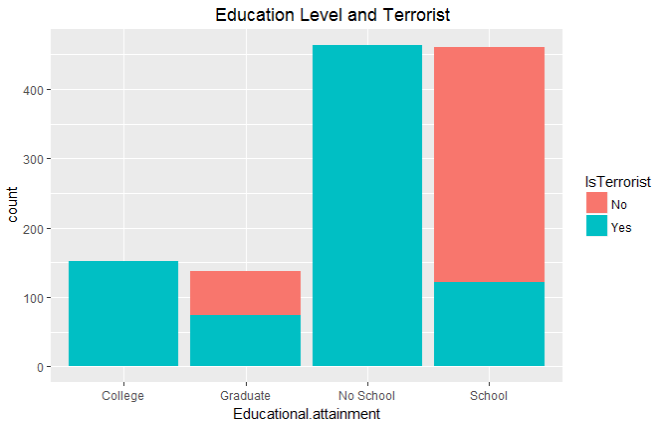
It shows relation between Terrorist camp and those who attended are terrorist.



It show relationship between Terrorist and Gender, male are more likely to become terrorist.



It shows relation between Terrorist camp and Madrasa training and there is moderate relation between them.



It shows relation between Terrorist camp and education level And people with graduation and people with no school are more likely to become a terrorist. who attended are terrorist.