Predictions on US Mass Murders using Data Mining Techniques

Chethana
Department of Computer
Kent State University
Kent, Ohio, USA
cvempara@kent.edu

Davneet Kaur
Department of Computer
Kent State University
Kent, Ohio, USA
dkaur1@kent.edu

Shaista Gulnaar
Department of Computer
Kent State University
Kent, Ohio, USA
shaista@kent.edu

***Abstract*— The number of mass genocides is increasing day by day which results in a lot of loss to life and material. These kinds of happenings should be prevented. The application of Data Mining Techniques would be useful to analyze some underlying patterns, which can help the government and the citizens to take some preventive measures. As part of our project, we use our knowledge of the Data Mining Techniques on the data of US Mass Shootings of the last 50 years. Providing a prediction on mental illness of the shooters based on their age and race; and on which specific dates attacks are more likely to occur and in which states would be very useful for the government and US citizens. We use J48 decision tree and Naïve Bayes classification algorithms on the dataset.**

***Keywords*—Data Mining, Naïve Bayes, J48 Decision Tree, US mass shootings, Mental Illness.**

I. INTRODUCTION

Terror is rising day by day all over the world, and United States is no exception. There have been mass shootings at schools, shopping malls, music concerts and even at a movie theatre. So far, there have been 398 mass shootings recorded only in the past 50 years that resulted in 1996 deaths and 2488 people injured [1]. The average number of genocides is 7 per year for last 50 years which took 39 lives and 48-person injuries per year [1]. These incidents effect the society on a high rate which in turn contributes for such situations to occur again indirectly. There is no denial in the fact that such happenings result in loss of both life and materials such as; personal items and infrastructure. Also, people remain in a state of panic and assume every other race to be a terrorist. This in turn reduces the number of people attending mass gatherings and celebrations, which would otherwise bring closeness amongst the public. There is also a strong psychological effect on children that impact the future generation. Lastly, because of these mass shootings, the reputation of a particular area or state is getting negatively affected. All of these reasons are enough for the society and government to take some preventive measures for this issue.

Generally, it is assumed that severe mental illness increases the risk of violence. However, in an article [2], authors critically addressed the assumption that "Mental illness causes gun violence". They suggested that not only mental illness but other factors such as; social relationships, firearm access during emotional moments etc., also lead to gun violence. However, they failed to cite this with strong facts, numbers etc. Reports suggest that up to 60% of executioners of mass shootings in the United States since 1970 displayed symptoms including acute paranoia, delusions, and depression before committing their crimes [3, 4]. Researchers said that there is a difference between criminals who end up in murders to those who end up in suicide following murder [4]. It is not known about the difference in the type of shooters regarding the difference between workplace, schools, and rampage. The report only presents the difference in shooters who die or live after their attack. The dataset used in this report is from a report named NYPD 2010.There is a clue for scholars and security officers in the criminal pattern, weapons, victims, place of attack etc. According to Van Dorn et al., a history of childhood abuse, binge drinking, and male gender are all predictive risk factors for serious violence [5]. In this study, the authors have re-examined on this issue using dataset of National Epidemiologic Survey on Alcohol and Related Conditions (NESARC), a two-wave study (N = 34,653: Wave 1: 2001–2003; Wave 2: 2004–2005). They found that people with severe illness were relatively more violent than those with minor illness or disorders. Further, they found that historical and the present conditions assisted violence. Finally, they concluded that there is a statistical relationship between severe mental illness and violence. In another article [6], the authors emphasized on the "association between rates of household firearm ownership and homicide across the United States, by age group" with the help of "cross-sectional time-series data (1988–1997)". A proportional relation between the household firearms and murders was found with this cross-sectional time-series data. They found that there was a relation among victims of age group 5-14 and about 35 years and so on. In addition, region wise and state wise there was a connection for all age groups above 5, in spite the issues like unemployment, poverty, urbanization, alcohol consumption. Their study was unable to show the cause for the effect, but

they concluded with the fact that wherever there is a large range of household firearms ownership, there were more numbers of murders in that state. In this analysis, they failed to take gender, mental health and other factors into consideration which helps more to analyze that which type of people (mentally ill) and/ or which gender are doing these cruel activities.

Although there have been many studies and analysis done on various aspects of US mass shootings so far, there have been no predictions done on any kind of dataset using Data Mining and Machine learning techniques. Some exploratory analysis has also been done using the same dataset that is used for our project. They have done analysis on the dataset using Python and R (plotly package) [1]. However, our project is aimed to come up with some useful predictions using this dataset, rather than some analysis. Our project aims to predict; mental illness among shooters based on their age and race; and on which specific date's attacks are more likely to occur at what states. The software WEKA is used for our project [7]. WEKA is a very powerful tool that has 76 classification/regression algorithms. We have used Naïve Bayes and J48 – Decision Tree classification algorithms in WEKA on some of the attributes in the dataset to come up with expected outcomes.

II. Dataset Details

The dataset for this data mining research is obtained from kaggle.com. [1] The dataset includes details of about 300 mass shootings that happened in United States in the last 50 years (1966-2017). The various attributes included in the dataset are title, location, date, summary, fatalities, injured and total victims, mental health issues race, gender, latitude and longitude. Table 1 shows the attributes obtained from kaggle.com along with their description and data type. Not all of these attributes are used in the predictions in this paper. Along with these attribute, another attribute 'Age' is taken from opendatasoft.com [8]. Table 2 shows the description of this attribute.

III. Data Analysis

A. Data Pre-processing

Before any classification algorithms can be run on the dataset, we need to massage the dataset i.e. do some sort of pre-processing on this dataset so that it can be imported in WEKA. Although the downloaded dataset was in ".csv" (Comma Separated Values) format, there were many missing/unknown values in many of the attributes. In addition, not all the attributes in the downloaded dataset were needed as part of our data mining research. We did an extensive refining of the dataset by analyzing the data repeatedly and then removing any unwanted information. For example, in the summary attribute, there were many special characters (like "") that could not be recognized by WEKA. These characters were removed. Also, as part of the pre-processing step, we added some new attributes from existing ones as per our requirement.

Table 1 – Attributes obtained from kaggle.com

Attribute	Description	Type
Title	Mostly same as location	String
Location	City and state where the shooting happened.	String
Date	Date when the shooting happened	Date Time
Summary	A brief description of the incident including some details about the shooter	String
Fatalities	Number of people that died	Numeric
Injured	Number of people wounded	Numeric
Total Victims	Sum of fatalities and injured	Numeric
Mental Health Issues	If the shooter had mental illness or not	String
Race	Race of the shooter	String
Gender	Gender of the shooter	String
Latitude	Latitude of the location	Numeric
Longitude	Longitude of the location	Numeric

Table 2 – Attribute obtained from opendatasoft.com

Attribute	Description	Type
Age	Age of the shooters	Numeric

The initial challenge was to have a refined ".csv" file that could be imported into WEKA to apply the data mining techniques on this file. The very first issue was with the Date attribute. The values in the date attribute had different formats. For example, some dates were in the format MM-DD-YYYY, while others were in the format M/DD/YY or MM/DD/YYYY. We manually checked all these values and changed them into MM/DD/YYYY format. Although, we changed all the values in the Date attribute to same format, if we keep the Date as it is, WEKA will consider it as a multi-nominal attribute. All the date values will be considered as unique classes and so the predictions will not be accurate. Therefore, we split the Date

attribute into two new attributes Month and DOM (day of month). It makes sense to ignore the year in the Date, as we want to do predictions on the future dates.

As mentioned earlier, the Age attribute was taken from opendatasoft.com. Some of the values in Age attribute were missing in this data. For these missing values, they were further derived from the Summary attribute. For the values that could not be taken from the Summary also, they were considered to be missing and then replaced by "?".

The values in the Location consisted of both city and state. Since we wanted to predict only the states where crimes may happen in the coming future, we separated the Location into two new attributes as City and State. There were many missing values in the Location. In case of missing values, we used Latitude and Longitude to find the state. The Google API [9] was used to get the state from longitude and latitude. This API gives a complete address when queried upon latitude and longitude. However, as we were only interested in the state and city, the rest of the address was ignored.

The last challenge in the dataset was with the missing values. After we completed all the above data pre-processing steps, there were many cells in the dataset for which the values were still missing. For all these cells, missing values were replaced by "?". Table 3 provides the final set of attributes along with the percentage of missing values for these attributes. Once all the above data refinements were done, our ".csv" file was ready to be imported in WEKA to start the research.

Table 3 – Correlation Ranking Attribute Evaluator results for Mental Health Issues class

Attribute	% Missing Values
City	0%
State	0%
Month	0%
DOM	0%
Age	14%
Mental Health Issues	34%
Race	14%
Gender	5%

B. Attribute Selection and Data Labeling

We deleted the unimportant attributes in the beginning of the attribute selection process. The below attributes were deleted even before starting the attribute selection process:

- Title: It is mostly same as location.
- Location: Split into city and state.
- Date: Split into month and DOM, and ignored the year.
- Summary: Taken the necessary information (age) and ignored rest of the text.
- Fatalities: Make no sense in our predictions.
- Injured: Make no sense in our predictions.
- Total Victims: Make no sense in our predictions.
- Latitude: Redundant attribute.
- Longitude: Redundant attribute.

Initially, for selecting attributes, we used "Forward selection method" of attribute selection. This is one of the greedy and lazy approaches towards attribute selection. We started from one attribute and tried different combinations of attributes to get satisfactory results. We also tested different attributes for their correlation with the class using CorrelationAttributeEval Attribute Evaluator in WEKA. This evaluator ranks the attributes as per their correlation with the class. We used this evaluator for both predictions once keeping the class value as Mental Health Issues and then keeping the class value as State

Table 4 gives the ranking of different attributes as per the CorrelationAttributeEval for Mental Health Issues class. It can be seen in the table that Race and Age had the highest ranks among all attributes. It was also observed later that doing these predictions by considering Age and Race of the shooters gave the most optimal predictions.

Table 5 gives the ranking of different attributes as per the CorrelationAttributeEva for prediction on State as class. It is seen from the table that DOM, Race and Month attributes have the highest ranks. However, it makes sense to not consider Race while predicting that what states might have such happenings in future, and on what dates. Therefore, we considered only DOM and Month to predict that what State may be prone to such attacks.

Table 4 – Correlation Ranking Attribute Evaluator results for Mental Health Issues class

Attribute	Ranked
Race	0.1646
Age	0.0777
Month	0.0749
Gender	0.0643
City	0.0633
State	0.0576
DOM	0.0423

Table 5 – Correlation Ranking Attribute Evaluator results for State

Attribute	Ranked
DOM	0.0582
Race	0.0481
Month	0.0467
Mental Health Issues	0.045
Age	0.0324

After we decided on the attributes for both the predictions, the next step was to Label the data. For our predictions, we had two classes: Mental Health Issues and State. For Mental Health Issues, we had three labels: Yes, No and Unclear. Yes and No indicated whether a shooter has mental illness or not. For Unclear, the dataset owner was not sure of the mental health status of the shooters. It is important to mention here that we did not considered Unclear as a missing value, and so did not replaced it with a "?'. This is because for a missing value, the value is completely unknown. However, for Unclear, some sources mentioned that the shooter had a mental illness, while others mentioned otherwise. Therefore, these values were decided to be taken as Unclear and not as missing values.

For the attribute State, the Labels are the abbreviations of the 51 US states except for the three fortunate states where mass shootings never happened: New Hampshire, Rhode Island and North Dakota.

IV. Algorithms Used And Results

We have used WEKA Explorer tool to do our data mining research. Among all the 76 classification algorithms in WEKA, Naïve Bayes and J48 – Decision Tree are one of the best classifiers according to our dataset. Therefore, we did our predictions using these algorithms.

A. Prediction on Mental Illness in shooters based on their Age and Race

Initially, for predicting the Mental Health of the shooters, we considered three attributes: Age, Race and State. We wanted to see if there is any relation between State and mental health of shooters. We ran Naïve Bayes first on these attributes and making Mental Health as class. We started with cross-validation as the splitting method for training and test sets, and folds = 5. We changed the number of folds from 5, 6, 7, and so on up to 12. We observed that when we take folds = 9, we get as 59.33% correctly classified instances. This was the best among different values of folds. Next, we ran Naïve Bayes again but with Percentage Split as the splitting method for training and test sets. We started with Percentage split as 66% (i.e. the default) and tested by increasing the percentage to 67%, 68%, 69%, and so on up to 75%. We observed that when we keep the percentage split as 70%, we get the correctly classified instances as 61.29% and this was again the best among different values of percentage split.

Next, we ran J48 Decision tree using the same set of attributes and making Mental Health as class. We started with cross-validation as the splitting method for training and test sets, and folds = 5. We changed the number of folds from 5, 6, 7, and so on up to 12. We observed that when we take folds = 6, we get as 55.98% correctly classified instances. This was the best among different values of folds. Next, we ran J48 again with Percentage Split as the splitting method for training and test sets. We started with Percentage split as 66% (i.e. the default) and tested by increasing the percentage to 67%, 68%, 69%, and so on up to 75%. We observed that when we keep the percentage split as 70%, we get the correctly classified instances as 58.06% and this was the best among different values of percentage split.

Table 6 gives a summary of these results. We can see that by using Naïve Bayes, we have better results than J48 decision tree for both cross-validation and percentage split splitting methods.

Table 6 – Results of Naïve Bayes and J48 Decision Tree using Age, Race and State attributes

Algorithm	Cross Validation	Percentage Split
Naïve Bayes	59.33 % (9 folds)	61.29 % (70 % training set)
J48	55.98 % (6 folds)	58.06 % (70 % training set)

On observation, the results in table 5 are not much convincing. In addition, we observed that in the J48 Decision Tree, there were no rules in which the attribute "states" has been considered. So, we tried to improve our results by doing a next round of predictions. This time, we only took Age and Race attributes to predict the Mental Health of shooters and see if there is any improvement in the results. First, we ran Naïve Bayes on these two attributes and making Mental Health as class. We started with cross-validation as the splitting method for training and test sets, and folds = 5. We changed the number of folds from 5, 6, 7, and so on up to 12. We observed that when we take folds = 9, we get as 62.20% correctly classified instances. This was the best among different values of folds. Next, we ran Naïve Bayes again but with Percentage Split as the splitting method for training and test sets. We started with Percentage split as 66% (i.e. the default) and tested by increasing the percentage to 67%, 68%, 69%, and so on up to 75%. We observed that when we keep the percentage split as 70%, we get the correctly classified instances as 70.97% and this was again the best among different values of percentage split.

Next, we ran J48 Decision tree using the same set of attributes and making Mental Health as class. We started with

cross-validation as the splitting method for training and test sets, and folds = 5. We changed the number of folds from 5, 6, 7, and so on up to 12. We observed that when we take folds = 6, we get as 59.33% correctly classified instances. This was the best among different values of folds. Next, we ran J48 again but with Percentage Split as the splitting method for training and test sets. We started with Percentage split as 66% (i.e. the default) and tested by increasing the percentage to 67%, 68%, 69%, and so on up to 75%. We observed that when we keep the percentage split as 66%, we get the correctly classified instances as 66.67% and this was again the best among different values of percentage split.

Table 7 gives a summary of these results. Again, we can see that by using Naïve Bayes, we have better results than J48 decision tree for both cross-validation and percentage split splitting methods. We saw that doing the predictions only on Age and Race gave us more accurate results than doing predictions by taking State attribute as well into consideration. This makes sense as in the attribute selection step also, we saw that when we ran the CorrelationAttributeEval Evaluator method in WEKA, State had the least rank among Race, Age and State attributes. Table 7 is our final result for this prediction.

Table 7 – Results of Naïve Bayes and J48 Decision Tree using Age and Race attributes

Algorithm	Cross Validation	Percentage Split
Naïve Bayes	62.20 % (9 folds)	70.97 % (70 % training set)
J48	59.33 % (6 folds)	66.67 % (66 % training set)

Figure 1 gives the final pruned J48 tree with the rules derived from the tree. These rules give a clear prediction that shooters of what race and what age have or do not have mental health issues. It also shows that for which race and what age, mental health of shooters is unclear.

B. On which specific date's attacks are more likely to occur at what states

We considered the Month and DOM to see if we can predict that on which specific date's attacks are more likely to occur at what state. We ran Naïve Bayes first on these attributes and making State as class. We started with cross-validation as the splitting method for training and test sets, and folds = 5. We changed the number of folds from 5, 6, 7, and so on up to 12. We observed that when we take folds = 6, we get as 6.27% correctly classified instances. This was the best among different values of folds. Next, we ran Naïve Bayes again but with Percentage Split as the splitting method for training and test sets. We started with Percentage split as 66% (i.e. the default) and tested by increasing the percentage to 67%, 68%, 69%, and so on up to 75%. We observed that for any values of percentage split, we could not get an accuracy of more than 3%. Hence, we did not consider those results.

```
J48 pruned tree
------------------

Race = White
|   Age <= 52: Yes (16.8/6.35)
|   Age > 52: Unclear (2.01/0.01)
Race = Asian: Yes (6.27/0.12)
Race = Black: Yes (7.32/4.14)
Race = Latino
|   Age <= 33: Yes (3.17/0.07)
|   Age > 33: Unclear (2.06/0.03)
Race = Some other race
|   Age <= 35
|   |   Age <= 24
|   |   |   Age <= 18: Yes (3.08/1.01)
|   |   |   Age > 18: No (2.14/0.06)
|   |   Age > 24: Yes (3.21/1.08)
|   Age > 35: No (5.15/0.13)
Race = Black American or African American
|   Age <= 41
|   |   Age <= 25: No (13.79/2.63)
|   |   Age > 25: Yes (21.46/9.25)
|   Age > 41: No (8.64/0.42)
Race = White American or European American: Yes (101.37/37.94)
Race = Asian American: No (7.32/0.21)
Race = Two or more races: No (1.04/0.03)
Race = Native American or Alaska Native: No (3.14/1.09)
Race = Asian American/Some other race: No (1.04/0.03)
```

Fig.1. J48 Tree for Mental Health of shooters based on Age and Race.

Next, we ran J48 Decision tree using the same set of attributes and making State as class. We started with cross-validation as the splitting method for training and test sets, and folds = 5. We changed the number of folds from 5, 6, 7, and so on up to 12. We observed that when we take folds = 7, we get as 8.77% correctly classified instances. This was the best among different values of folds. Here also, we ran J48 again but with Percentage Split as the splitting method for training and test sets. We started with Percentage split as 66% (i.e. the default) and tested by increasing the percentage to 67%, 68%, 69%, and so on up to 75%. In case of J48 also, we observed that for any value of percentage split, we could not get an accuracy of more than 3%. Hence, we did not consider the results with percentage split splitting method.

Table 8 gives a summary of these results. We can see that by using Naïve Bayes, we have better results than J48 decision tree. Although the accuracy of these results is way too less but we cannot improve more on these results. The reason of the results for this prediction being too less is that there are very less number of instances for doing such type of prediction. Table 7 is our result for the second prediction.

Figure 2 gives the final pruned J48 tree with the rules derived from the tree. This J48 tree here is an example of data over fitting. Since there are kind of less number of incidents on same dates, WEKA has come up with rules for almost each

instance. Fortunately, this is the reason for this prediction being so less accurate.

Table 8 – Results of Naïve Bayes and J48 Decision Tree using Month and DOM attributes

Algorithm	Cross Validation
Naïve Bayes	6.27 % (6 folds)
J48	8.77 % (7 folds)

V. Conclusion And Future Works

In our paper, we tried to come up with better results for both the predictions. We have got fairly satisfactory results in case of predicting the mental health of the shooters than doing predictions on state. We could get more accurate and efficient results if we have large datasets. While doing predictions on these kinds of mass shootings, to hope for a large dataset means having more such happenings. This means more loss to lives and material. So, in a way it is better that we have a few numbers of records for such type of dataset. We may have a bit more accurate results if there are no missing values at all even in the present dataset.

Recently, some new attributes are added to this data set. These new attributes included:

- Incident Area
- Open/Close Location
- Target
- Cause
- Policeman Killed
- Employed (Y/N)
- Employed at

Although these new attributes are added to the dataset, the values for these attributes are mostly missing for many records. We may come up with some other predictions if we consider the newly added attributes or may be combining this dataset with some other dataset. For example, if we have a dataset that has the information that how the weapon (if any) used in such shootings was accessible to the shooters (legally or illegally). Then maybe we can come up with a prediction if providing legal license to mentally ill people should be allowed or not.

Acknowledgment

We thank Dr. Kambiz Ghazinour for his expertise, guidance, instructions and concepts of Data Mining and Machine Learning techniques that were needed to make this project successful. We also thank him for his time and supporting throughout our research work.


```
J48 pruned tree
------------------

Month = Oct
|   DOM <= 6
|   |   DOM <= 1: NV (4.0/3.0)
|   |   DOM > 1
|   |   |   DOM <= 4: PA (2.0)
|   |   |   DOM > 4: CA (2.0/1.0)
|   DOM > 6
|   |   DOM <= 22
|   |   |   DOM <= 10: NJ (3.0/2.0)
|   |   |   DOM > 10: OH (7.0/5.0)
|   |   DOM > 22
|   |   |   DOM <= 25: WA (2.0/1.0)
|   |   |   DOM > 25: AZ (3.0/2.0)
Month = June
|   DOM <= 18: FL (13.0/9.0)
|   DOM > 18
|   |   DOM <= 22: WA (2.0/1.0)
|   |   DOM > 22: LA (3.0/2.0)
Month = May
|   DOM <= 16
|   |   DOM <= 7
|   |   |   DOM <= 4: WI (3.0/1.0)
|   |   |   DOM > 4: CA (2.0/1.0)
|   |   DOM <= 25: IL (6.0/4.0)
|   |   DOM > 25
|   |   |   DOM <= 30: WA (2.0/1.0)
|   |   |   DOM > 30: OH (2.0/1.0)
Month = Apr
|   DOM <= 19
|   |   DOM <= 5: TX (8.0/5.0)
|   |   DOM > 5
|   |   |   DOM <= 13
|   |   |   |   DOM <= 8
|   |   |   |   |   DOM <= 6: AL (2.0/1.0)
|   |   |   |   |   DOM > 6: IL (2.0/1.0)
|   |   |   |   DOM > 8: NM (3.0/2.0)
|   |   |   DOM > 13
|   |   |   |   DOM <= 18
|   |   |   |   |   DOM <= 17: VA (2.0/1.0)
|   |   |   |   |   DOM > 17: CA (2.0)
|   |   |   |   DOM > 18: TX (3.0/2.0)
|   DOM > 19: MD (12.0/10.0)
Month = Jan
|   DOM <= 13
|   |   DOM <= 7: FL (4.0/2.0)
|   |   DOM > 7: AZ (3.0/2.0)
|   DOM > 13: CA (21.0/16.0)
Month = Sept
|   DOM <= 16
|   |   DOM <= 10: NV (3.0/2.0)
|   |   DOM > 10: TX (5.0/3.0)
|   DOM > 16
|   |   DOM <= 22: PA (4.0/3.0)
|   |   DOM > 22
|   |   |   DOM <= 26: WA (2.0/1.0)
|   |   |   DOM > 26: FL (3.0/2.0)
Month = July
|   DOM <= 19: CA (10.0/7.0)
|   DOM > 19: GA (6.0/4.0)
Month = Mar
|   DOM <= 16
|   |   DOM <= 9
|   |   |   DOM <= 6
|   |   |   |   DOM <= 5: CA (2.0/1.0)
|   |   |   |   DOM > 5: GA (4.0/2.0)
|   |   |   DOM > 6
|   |   |   |   DOM <= 8: KS (3.0/1.0)
|   |   |   |   DOM > 8: PA (2.0)
|   |   DOM > 9
|   |   |   DOM <= 11: AL (3.0/2.0)
|   |   |   DOM > 11: GA (7.0/5.0)
|   DOM > 16
|   |   DOM <= 22
|   |   |   DOM <= 19: NV (5.0/4.0)
|   |   |   DOM > 19: TX (4.0/3.0)
|   |   DOM > 22
|   |   |   DOM <= 25
|   |   |   |   DOM <= 24: CA (2.0/1.0)
|   |   |   |   DOM > 24: WA (3.0/2.0)
|   |   |   DOM > 25
|   |   |   |   DOM <= 29: NC (5.0/2.0)
|   |   |   |   DOM > 29: VA (2.0/1.0)
Month = Feb
|   DOM <= 20
|   |   DOM <= 15
|   |   |   DOM <= 11: FL (18.0/14.0)
|   |   |   DOM > 11: IL (3.0/2.0)
|   |   DOM > 15
|   |   |   DOM <= 19: CA (6.0/2.0)
|   |   |   DOM > 19: MI (5.0/3.0)
|   DOM > 20
|   |   DOM <= 24: TX (9.0/6.0)
|   |   DOM > 24
|   |   |   DOM <= 26
|   |   |   |   DOM <= 25: KS (3.0/1.0)
|   |   |   |   DOM > 25: WA (2.0/1.0)
|   |   |   DOM > 26
|   |   |   |   DOM <= 27: OH (3.0/2.0)
|   |   |   |   DOM > 27: CA (3.0/2.0)
Month = Dec
|   DOM <= 12
|   |   DOM <= 6
|   |   |   DOM <= 3: CA (3.0/2.0)
|   |   |   DOM > 3: NE (4.0/2.0)
|   |   DOM > 6
|   |   |   DOM <= 8: OH (2.0/1.0)
|   |   |   DOM > 8: NV (3.0/2.0)
|   DOM > 12
|   |   DOM <= 17
|   |   |   DOM <= 14: LA (4.0/3.0)
|   |   |   DOM > 14: PA (2.0/1.0)
|   |   DOM > 17
|   |   |   DOM <= 25: CA (4.0/2.0)
|   |   |   DOM > 25: MA (6.0/4.0)
Month = Nov
|   DOM <= 9
|   |   DOM <= 3: CA (3.0/2.0)
|   |   DOM > 3
|   |   |   DOM <= 6: TX (2.0/1.0)
|   |   |   DOM > 6: OH (2.0/1.0)
|   DOM > 9
|   |   DOM <= 20
|   |   |   DOM <= 12: WA (2.0/1.0)
|   |   |   DOM > 12: FL (5.0/3.0)
|   |   DOM > 20
|   |   |   DOM <= 25: OH (3.0/2.0)
|   |   |   DOM > 25: WA (3.0/1.0)
Month = Aug
|   DOM <= 9: TX (9.0/6.0)
|   DOM > 9
|   |   DOM <= 22
|   |   |   DOM <= 19
|   |   |   |   DOM <= 17: CA (4.0/2.0)
|   |   |   |   DOM > 17: NY (2.0/1.0)
|   |   |   DOM > 19: FL (2.0/1.0)
|   |   DOM > 22
|   |   |   DOM <= 26: VA (2.0/1.0)
|   |   |   DOM > 26: IL (3.0/2.0)
```

Fig.2. J48 Tree for on which specific date's attacks are more likely to occur at what states

REFERENCES

[1] Z. Usmani, "US Mass Shootings Last 50 Years (1966-2017)." 2017. [Online]. Available: https://www.kaggle.com/zusmani/us-mass-shootings-last-50-years. [Accessed Oct. 15, 2017]

[2] J. M. Metzl and K. T. MacLeish, "Mental Illness, Mass Shootings, and the Politics of American Firearm." 12, December 2014. [Online] Available: http://ajph.aphapublications.org/doi/full/10.2105/AJPH.2014.302242. [Accessed Oct. 17, 2017]

[3] M. Follman, "Mass Shootings: Maybe What We Need Is a Better Mental-Health Policy." 9, November 2012. [Online]. Available: http://www.motherjones.com/politics/2012/11/jared-loughner-mass-shootings-mental-illness/. [Accessed Oct. 17, 2017]

[4] A. Lankford, "Mass Shooters in the USA, 1966–2010: Differences Between Attackers Who Live and Die." *Justice Quarterly*, vol. 32, no. 20, June 2013. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/07418825.2013.806675. [Accessed Oct 17, 2017]

[5] R. V. Dorn, J. Volavka and. N. Johnson, "Mental disorder and violence: is there a relationship beyond substance use?" *Social Psychiatry and Psychiatric Epidemiology*, vol. 47 no, Issue 3, p 487–503. March 2012. [Online]. Available: https://doi.org/10.1007/s00127-011-0356-x. [Accessed Oct 18, 2017]

[6] M. Miller, D. Azrael, and D. Hemenway, "Rates of Household Firearm Ownership and Homicide Across US Regions and States, 1988–1997." 10, October 20111. [Online]. Available: http://ajph.aphapublications.org/doi/full/10.2105/AJPH.92.12.1988. [Accessed Oct18, 2017]

[7] E. Frank, M. A. Hall, I. H. Witten and C. J. Pal, "The WEKA Workbench." *Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*, 2016. [E-book] Available: https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.

[8] Dataset for accessing Age, https://public.opendatasoft.com/explore/dataset/mass-shootings-in-america/. [Accessed: Nov 08, 2017]

[9] Google API to find City and State, https://www.findlatitudeandlongitude.com/find-address-from-latitude-and-longitude/#.WiEEQoZKvDd [Accessed: Nov 08, 2017]