Gun Violence and Mass Shootings
Big Data Science Final Project Report - Spring 2018

Daniel Rivera Ruiz
Department of Computer Science
New York University
daniel.rivera@nyu.edu

Herbert Li
Department of Computer Science
New York University
herbert.li@nyu.edu

Ross Abramson
Department of Computer Science
New York University
ross.abramson@nyu.edu

Abstract

With the recent school shooting in Parkland, mass shootings have once again become a divisive and contentious topic across America. The various shootings we've had over the years seem almost random, with every person and political party having their own opinion of the root cause. Assault weapons, violence in mass media, and mental health are among the usual suspects. If we could determine the root causes or common threads between these shootings, it could prove to be an invaluable asset for law enforcement, politicians, and mental health specialists to predict and prevent such events from happening in the future. As a result, we would like to examine mass shootings from a data science and data analysis perspective. In this paper, we will perform clustering on these events in order to identify hidden patterns and similarities between them. We will also examine this in a classification scenario by evaluating the performance of several machine learning models in the context of predicting whether or not a shooting will happen based on features of the surrounding area and population.

1 Business Understanding

1.1 Literature Review and Related Work

Past research has suggested that mental illness, educational background, and the availability of guns are among the factors that can predict whether a certain location or person is likely to experience or engage in a shooting. However, not much data-driven analysis has been conducted in this area. We have listed several relevant papers below:

- **Gun control really works - here's the science to prove it** (Loria, 2018) - How gun control can help prevent gun violence.
- **Firearm regulations in the U.S.: trend of gun violence.** (Bouchet, 2017) - Kaggle dataset.
- US Mass Shootings Analysis 1966-2017 (Smith, 2017) - Kaggle dataset and analysis on how demographics correlate to shooters in mass shootings.
- **A Guide to Mass Shootings in America.** (Follman et al., 2018) - One of the de-facto gun violence datasets.
- **Mental Illness, Mass Shootings, and the Politics of American Firearms.** (Metzl and MacLeish, 2014) - Study regarding the relationship between mental illness and mass shootings.

- **The impact of the Orlando mass shooting on fear of victimization and gun-purchasing intentions.** (Stroebe et al., 2017) - Study regarding the relationship between guns sales and mass shootings.
- **Socioeconomic factors and mass shootings in the US.** (Kwon and Cabrera, 2017) - Study exploring whether population-level measures of income inequality and poverty rates are associated with mass shootings in the United States.
- **Columbine Revisited: Myths and Realities About the Bullying–School Shootings Connection** (Mears et al., 2017) - Study seeking to understand the bullying–school shooting connection.

1.2 Zip Codes

Currently, there is a surprising lack of research in predicting mass shooters and mass shootings. Mass shootings seem to happen sporadically, but we think that there may be recurring patterns. As a result, we want to examine the problem of predicting mass shootings from a data analysis/data mining perspective, answering questions like: What cities/states/zip codes are more prone to such attacks? Is there any correlation between an area's demographic data (earnings, age, education level), guns licenses, guns manufacturing, etc. and whether or not there has been a mass shooting there?

1.3 Stanford Mass Shootings in America (MSA)

Started in 2012 as a reaction to the mass shooting in Sandy Hook, the Stanford Mass Shootings in America (MSA) data set (Center and Libraries, 2018) sets out to create a single repository for mass shooting events. The project attempts to facilitate research on gun violence in the US by making raw data more accessible. We would like to analyze this data set as well, answering questions like: Are there characteristics that make a particular shooting deadlier than others? What role does mental health play in mas shootings? Is there any correlation between shooters and race, gender, income, educational attainment, etc.?

2 Our Data

2.1 Zip Codes

2.1.1 Data Understanding

For this problem, we obtained data from several sources:

(a) Firearms Data - ATF (The Bureau of Alcohol, Tobacco, Firearms, and Explosives)
- Listing of Federal Firearms Licensees
- Listing of Firearms Manufacturers
- Challenge: Lack of data, ATF Records only from 2014 and onward

(b) Census Data - US Census Bureau
- Demographic Data

(c) Gun Violence Data sets, Mother Jones Database, Kaggle repositories and others.

2.1.2 Data Preparation

We noticed that the data is significantly sparser the further back you go. As a result, we will only look at data from recent years (2014-2016). We handled missing values by replacing them with the average value of the feature, and we also identified and removed highly correlated features to simplify our models (see 3 for more details on these processes).
Another feature we wanted to include was some sort of proximity calculation. Not only did we want to calculate gun licenses and manufacturing for a certain zip code, but we also wanted to spread these numbers to all nearby zip codes within a 50 mile radius.
Our final data matrices for three years: 2014, 2015, and 2016, consist of 33120 observations (rows), with each row pertaining to an individual ZIP code. We selected 91 variables / features (columns) in

Table 1: Features of the ZIP Codes Datasets

Type of feature	Count
Income per household	2
Demographic data	55
Educational attainment	31
Guns manufacturing	1
Guns licences	1
Guns purchases	1
Total	**91**

total (to be pruned via feature selection/ranking). The breakdown of these features and their counts is presented in table 1.

The predictive variable (label) we used for these datasets was a binary indicator for violence. A value of 1 means that there is a record of at least one gun-related violent incident in that zip code or within a 50-mile radius in that year, while a value of 0 implies there have been no violent incidents.

2.2 Stanford MSA

2.2.1 Data Understanding

The Stanford MSA Data set (Center and Libraries, 2018) is available via GitHub.

2.2.2 Data Preparation

We reduced the Stanford MSA dataset to the following form: 336 observations (rows), with each row pertaining to a mass shooting event. For each observation we considered only 24 variables / features (columns), which are summarized in table 2:

Table 2: Features of the Stanford MSA Dataset

Type of feature	Count
Date and location	2
Shooter information	7
Type and number of guns	7
Targeted place	3
Targeted victims	1
Possible motive	1
Mental illness	1
Military experience	1
Type of incident	1
Total	**24**

The predicted variable (label) we used was again a binary indicator for how violent the event was, which is either 1, meaning this incident was particularly deadly (the number of deaths was above the median) or 0 otherwise.

3 System Architecture

3.1 RapidMiner

RapidMiner is a data science platform which provides an integrated platform to easily develop models for pre-processing of data, machine learning, and predictive analytics. It offers a template and GUI based interface so that users can easily drag and drop actions for data modeling. We used RapidMiner to retrieve weights of attributes in terms of rule, their correlation to whether a violent gun crime

existed, as well as several models showcased in the Experimental Section. Some of the ranking algorithm were used redundanly (i.e. correlation). We did this to showcase the utility and parallel output of all data tools used throughout our research.

3.1.1 Feature Ranking

Using RapidMiner, we were able to show the most correlated features for the zip code data set as well as their weights in Information Gain and OneRule. The results shown are from 2016 but that is immaterial. Under correlation, as shown in Figure 1, manufacturing and licensing

attribute	wei... ↓
91) Licenses	0.502
92) Manufacturing	0.280
14) Percentage of 10) some college	0.135
41) Population percentage: not Hispan...	0.129
03) Household median income in dollars	0.128
30) Median earnings in dollars, popula...	0.128
17) Percentage of 10) graduate degree	0.126
35) Median earnings in dollars, popula...	0.113
26) Percentage of 24) bachelor's or hi...	0.112
20) Percentage of 18) bachelor's or hi...	0.108
93) Gun purchases approximate	0.103

Figure 1: Highly Correlated Features using RapidMiner for the 2016 Zip Code Dataset.

In figure 2, the list of weights were obtained using Rapid Miner's "Weigh by Rule" (Rule). Rule is a simple classification algorithm which generates one rule for each predictor in the data. The smallest total error of the rules is selected to be the main rule of use. A frequency table is selected for each predictor against the target. In a way, it's a simple summation of which factors are present for the label attribute. This weighting system indicates that Manufacturing and Licenses carrying the most weight in terms of gun violence occurring within an area. Additionally the percentage of bachelor based on age group, highschool dropouts, and approximate gun purchase carry the next highest values of weight importance. It shows that lack of education and availability of guns were the main indicators for a violent gun crime to occur.

3.2 RStudio

R is a programming language and a free software environment for statistical computing and graphics. Along with its IDE *RStudio* it is one of the top technologies now a days to perform data analytics. In the context of our project, we used *RStudio* to perform the steps described in the following sections for all our datasets: the three ZIP Code datasets (2014, 2015 and 2016) and the Stanford dataset.

3.2.1 Data preprocessing

In this step of the process we performed two main tasks:

(a) Missing values substitution. Since we collected our data from several sources, the resulting datasets had missing values (or values with heterogeneous formats) for some of the features. To solve this problem, we decided to substitute all missing values with the mean value of the corresponding feature. Listing 1 shows the code used to achieve this task.

attribute	weight ↓
92) Manufacturing	0.824
91) Licenses	0.764
26) Percentage of 24) bachelor's or higher	0.747
20) Percentage of 18) bachelor's or higher	0.738
23) Percentage of 21) bachelor's or higher	0.731
12) Percentage of 10) 9th to 12th no diploma	0.718
93) Gun purchases approximate	0.695
57) Population total	0.685
29) Percentage of 27) bachelor's or higher	0.664
21) Total population 35 to 44 years	0.663
27) Total population 65 and over	0.660
09) Percentage of 05) bachelor's	0.652
19) Percentage of 18) high school or higher	0.652
08) Percentage of 05) some college	0.651
15) Percentage of 10) associate's degree	0.651
07) Percentage of 05) high school	0.650
30) Median earnings in dollars, population 25 and over	0.650

Figure 2: Weight by Rule Ranking Features for the 2016 Zip Code Dataset.

(b) Highly correlated features removal. In particular for the ZIP Code datasets, we noticed that the number of features we were considering was very high (above 90), which would result in heavy models that would take a long time to train. Thus we decided to use a correlation matrix with a cutoff value of 0.75 to remove highly correlated features from the datasets before training our models. Listing 2 shows the code used to achieve this task.

Listing 1: Missing Values Substitution

```
# Handle missing values: replace with mean
fill_in_values <- function(data, strategy){
    numericCols <- sapply(data[1,], is.numeric)
    data[, numericCols] <- apply(data[ , numericCols], 2, function(x){
        is_na <- is.na(x)
        x[is_na] <- strategy(x[!is_na])
        x
    })
    data
}
datasetMean <- fill_in_values(dataset, mean)
```

Listing 2: Highly Correlated Features Removal

```
# Find correlation matrix
correlationMatrix <- cor(datasetMean[-ncol(datasetMean)])
highlyCorrelated <- findCorrelation(correlationMatrix,
                                    cutoff = 0.75,
                                    exact = TRUE)
datasetClean <- datasetMean[, -highlyCorrelated]
```

3.2.2 Feature Ranking

For the feature ranking step of our process we fitted a *Learning Vector Quantization* (LVQ) model to our data. An LVQ is a special kind of artificial neural network were the neurons are competing with each other to yield the highest similarity score given an input vector from the dataset. The advantage

of this model is that it provides both a ranking of the features in the dataset as well as a classification mechanism given a labeled dataset.
The feature ranking part of the algorithm is performed within the built-in function provided by *R* and it is based in the area under the *ROC* curve to measure the importance of a variable. Regarding the classification part, we trained the model using 10-fold cross validation and generated the corresponding confusion matrix.
Listing 3 shows the code used to achieve this task.

Listing 3: Feature Ranking and LVQ Model

```
# Fit Learning Vector Quantization (LVQ) model using repeated cross validation
control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
model <- train(LABEL ~ .,
               data = datasetClean,
               method = "lvq",
               preProcess = "scale",
               trControl = control)
importance <- varImp(model, scale = FALSE)

# Generate test set and confusion matrix
get_golden <- function(x) x$LABEL
testIndices <- createDataPartition(dataset$LABEL, p = 0.3)[[1]]
testData <- datasetClean[testIndices, ]
cm <- confusionMatrix(predict(model, newdata=testData), get_golden(testData))
```

3.2.3 Feature Selection

Feature ranking helps us get an idea of the variables that are highly related to our classification label and therefore have a higher prediction power. However, this technique usually considers features by themselves without accounting for the interactions among them.
Feature selection algorithms overcome this issue by trying different combinations of features and selecting the ones that (together, not alone) yield the best performance. For our problem we used the *Recursive Feature Elimination* algorithm built-in within *R*, using a 10-fold cross validation technique and a random forest selection function. Listing 4 shows the code used for this task.

Listing 4: Recursive Feature Elimination

```
controlRFE <- rfeControl(functions = rfFuncs, method = "cv", number = 10)
vf <- dim(datasetClean)[2]
results <- rfe(datasetClean[, 1:vf - 1],
               datasetClean[, vf],
               sizes = c(1:vf - 1),
               rfeControl = controlRFE)
predictors <- predictors(results)
```

3.2.4 Hierarchical Clustering

The final step of our analysis in *R* was hierarchical clustering. From the reduced dataset (only with the selected features from the previous step) we generated a dendrogram using the *Agnes* (agglomerative nesting) algorithm.
The *Agnes* algorithm groups the data one by one on the basis of the nearest distance measure of all the pairwise distances between the data points. One of its advantages is that it yields the agglomerative coefficient, which measures the amount of clustering structure found.
In addition to the dendrogram we also generated a new version of our datasets that includes an additional column for the cluster assigned to each data point. The cluster assignment was achieved by cutting the agnes dendrogram into a given number of clusters k.
Finally, we generated a comparison matrix to visualize the relation between cluster and label i.e., how many data points of each label where assigned to the different clusters.
Listing 5 shows the code used for this task.

Listing 5: Hierarchical Clustering

```
# Run Agnes (Agglomerative Nesting) hierarchical clustering algorithm
```

```
datasetPredictors <- scale(datasetClean[, predictors])
agnesClustering <- agnes(datasetPredictors, method = "complete")

# Cut Agnes tree into optimal number of clusters
kClustering <- 4
clusters <- cutree(as.hclust(agnesClustering), k = kClustering)

# Append Cluster column to original data
datasetClusters <- datasetClean
datasetClusters$Cluster <- as.data.frame(clusters)$clusters

# Comparison cluster vs. label
clusterLabel = matrix(nrow = kClustering,
                      ncol = 2,
                      dimnames = list(NULL, c("0", "1")))
for (k in 1:kClustering){
  clusterLabel[k, 1] <- length(which(
            datasetClusters[, ncol(datasetClusters) - 1] == 0 &
            datasetClusters[, ncol(datasetClusters)] == k))
  clusterLabel[k, 2] <- length(which(
            datasetClusters[, ncol(datasetClusters) - 1] == 1 &
            datasetClusters[, ncol(datasetClusters)] == k))
}
clusterLabel <- as.data.frame(clusterLabel)
```

3.3 Weka

Weka is a machine learning library written in Java that is maintained by the University of Waikato. It implements many popular machine learning algorithms like KMeans, SVM, Neural Nets, and also provides utilities for feature selection and feature ranking. In this project, we primarily used the GUI version of Weka, and found that it was very easy to use, and useful in rapidly prototyping models.

3.3.1 Feature Ranking

Using Weka, we performed Feature Ranking based on Information Gain Ratio as well as Correlation with the target label. reduced. Our results for the year 2014 are shown in 4 and 3. Across all years, the highest correlated and ranking features were gun licenses and manufacturing. Other high-scoring attributes were consistently: Median earnings in dollars, population 25 and over, Household mean/median income in dollars.

We also performed feature ranking using the Stanford MSA data set. Our results are shown in 5. Perhaps unsurprisingly, the number of guns and whether or not the shooter had a history of mental illness were highly correlated with the deadliness of the shooting.

4 Experimental Results

4.1 RapidMiner

Through RapidMiner, we utilized several predictive models to evaluate the Stanford MSA dataset. We used 'total number of victims in terms of injured and dead' as the predictive attribute. Given that all incidents have some type of injures, we used this approach to understand what factors may lead to a deadlier shooting. We used five models for evaluation: Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, and Gradient Boosted Trees. Of the models, Random Forest performed the best judged by the least Root Mean Squared Error as shown in 6.

4.1.1 Process Overview

The process shown in 7 first pre-processes the data, second splitting into a training and test set, then 10 fold Cross Validation, and finally application of one of the five models.

(1) **Pre-processing**. The pre-processing step defines our predictive column as well as remove columns that were found to have importance or helpfulness to our predictive model. Those

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 92 VIOLENT?):
        Correlation Ranking Filter
Ranked attributes:
 0.50235    89 91) Licenses
 0.2802     90 92) Manufacturing
 0.16735    28 "30) Median earnings in dollars, population 25 and over"
 0.15501     1 03) Household median income in dollars
 0.15193    32 "34) Median earnings in dollars, population 25 and over bachelor's"
 0.14835     2 04) Household mean income in dollars
 0.13479    12 14) Percentage of 10) some college
 0.13179    33 "35) Median earnings in dollars, population 25 and over graduate degree"
 0.12941    39 41) Population percentage: not Hispanic Asian
 0.12856    31 "33) Median earnings in dollars, population 25 and over some college"
 0.12642    15 17) Percentage of 10) graduate degree
 0.11196    24 26) Percentage of 24) bachelor's or higher
 0.10803    18 20) Percentage of 18) bachelor's or higher
 0.10292    91 93) Gun purchases approximate
 0.10127    21 23) Percentage of 21) bachelor's or higher
 0.09644    22 24) Total population 45 to 64 years
 0.09364    25 27) Total population 65 and over
 0.0917      8 10) Total population 25 years and over
 0.0915     30 "32) Median earnings in dollars, population 25 and over high school"
```

Figure 3: Highly Correlated Features using Weka for the 2014 Zip Code Dataset.

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 92 VIOLENT?):
        Information Gain Ranking Filter

Ranked attributes:
 0.460891    90 92) Manufacturing
 0.233628    89 91) Licenses
 0.140561    91 93) Gun purchases approximate
 0.021006    12 14) Percentage of 10) some college
 0.020755    28 "30) Median earnings in dollars, population 25 and over"
 0.019041     1 03) Household median income in dollars
 0.01901      2 04) Household mean income in dollars
 0.017089    39 41) Population percentage: not Hispanic Asian
 0.015193    15 17) Percentage of 10) graduate degree
 0.014503    32 "34) Median earnings in dollars, population 25 and over bachelor's"
 0.013599    38 40) Population percentage: not Hispanic American Indian or native
 0.012346    31 "33) Median earnings in dollars, population 25 and over some college"
 0.012342    24 26) Percentage of 24) bachelor's or higher
 0.011418    37 39) Population percentage: not Hispanic black
 0.010221    18 20) Percentage of 18) bachelor's or higher
```

Figure 4: Info Gain Ranking Features for the 2014 Zip Code Dataset.

attributes were “Number of Shooters” and “Number of Automatic Guns.” These attributes were removed because they were over the stability threshold, a threshold judges as being constant for 90 percent of the data inputs (the majority of shootings did not include multiple shooters and very few had automatic guns). The data is then filtered for any missing values and is sent to the splitting step.

(2) **Splitting and Testing**. The split operator splits the data into a train:test ratio of 80:20. The data is then sent to the cross-validation operation which uses a 10-fold cross validation. The data is then sent to the main model’s operator. Performance is measured in Root Mean Squared Error.

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 27 LABEL):
        Correlation Ranking Filter
Ranked attributes:
 0.25888   14 Number of Automatic Guns
 0.18509   24 History of Mental Illness - General
 0.17093   15 Number of Semi-Automatic Guns
 0.14701   10 Number of Shotguns
 0.13299   13 Total Number of Guns
 0.12254    9 Type of Gun - General
 0.11677   21 Relationship to Incident Location
 0.11086   11 Number of Rifles
 0.10597   23 Possible Motive - General
 0.10137   20 Place Type
 0.09932   22 Targeted Victim/s - General
 0.09702   18 Shooter's Cause of Death
 0.08815   12 Number of Handguns
 0.07265   26 Class
 0.06847    8 Shooter Race
 0.05899    2 State
 0.05434    5 Number of shooters
 0.05339    7 Shooter Sex
 0.0519     1 City
 0.0514     3 Date
```

Figure 5: Correlation Ranking Features for the Stanford MSA Dataset.

Model	Root Mean Squared Error ↑	Run Time
Random Forest	7.805	28 s
Gradient Boosted Trees	7.856	1 min 28 s
Generalized Linear Model	7.922	321 ms
Deep Learning	8.071	4 s
Decision Tree	9.060	312 ms

Figure 6: Comparison of RapidMiner Model accuracy and time taken to develop.

4.1.2 Models Used

(1) **Deep Learning**. Deep Learning is based on a multi-layer feed-forward artificial neural network that is trained with stochastic gradient descent using back-propagation. Two hidden layers were used with 50 neurons each. Training consisted of 10 epochs. The activation function used by the neurons was a rectifier linear unit.

(2) **Random Forest**. Random forest is an ensemble method which generates several decision trees and outputs the class that is the mode of the classes or mean prediction. The model used 10 randomly generated trees and a subset of attributes are selected via bootstrapping. The criterion for splitting attributes was based on those which minimized the square distance between the average of values in the node. The maximum depth was set to 20.

(3) **Gradient Boosted Trees (GBT)**. GBT works by generating several decision trees. The trees which have little accuracy are considered weak, while those that have higher accuracy are boosted and given more weight. The Number of trees used was 20 with a maximum depth of 5. The gradient boosted model is another ensemble method to obtain predictive results.

(4) **Generalized Linear Regression (GLR)**. GLR works by generalizing linear models to the data by maximizing log likelihood.

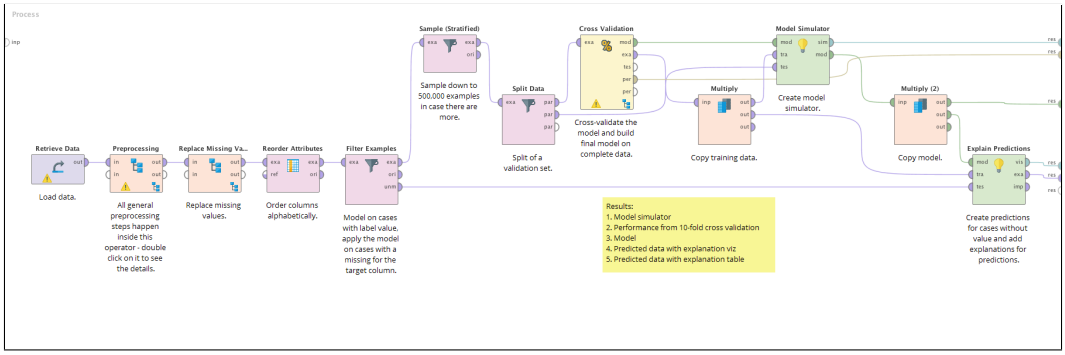

Figure 7: RapidMiner Process overview. All models incorporate the same template aspects with the exception of the cross validation step which only differs based on the Model used.

(5) **Decision Tree (DT)**. DT is a tree consisting of nodes where each node dictates a path decision based on inputted values. The criterion used for splitting data for decision was based on least square where the the attribute is selected for splitting based on whether it minimizes the square distance between the average of values in the node. The maximal depth of the trees was 40.

4.1.3 Results

The models were optimized to select the attributes which would maximize the predicted number of deaths. The models indicated which attributes were supportive and which would contradictory to get that maximal value. Of the models, there was some widespread agreement and some contradictions. Mental illness was considered bother supportive and contradictory. The type of place was considered very supportive for some models. This would make sense as school shootings and crowded in door locations tend to have a large population with a narrow means of escape. All but GBT placed "Number of Semi-Automatic Guns" as the most important attribute in predicting the number of victims in a mass shooting. GBT on the other hand ranked Total number of guns as more important.

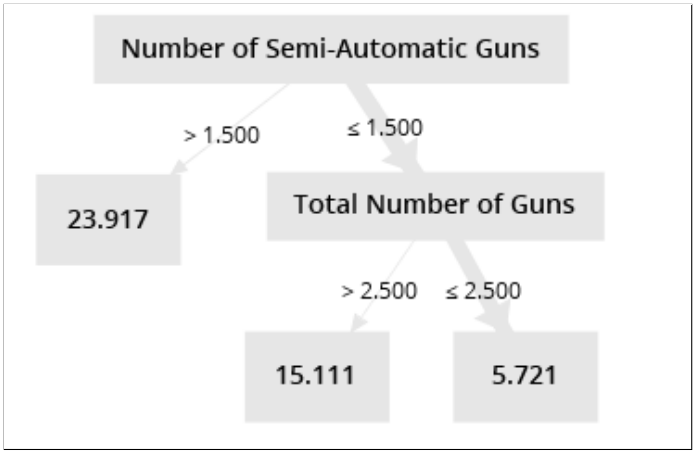

Figure 8: Decision Tree found that the only two aspects necessary for properly predicting a deadly shooting were the use of a semi-automatic gun and the total number of guns.

4.2 RStudio

In this section we present some of the results obtained after performing the analysis of our data on *RStudio*. While not comprehensive, they are a good example of the insights we could extract from the data and provide valuable evidence to support the conclusions presented in section 5.
Figure 11 shows the list of features that were removed from the 2014 ZIP Code dataset after performing the high correlation analysis described in section 3.2.1. The resulting dataset contains the

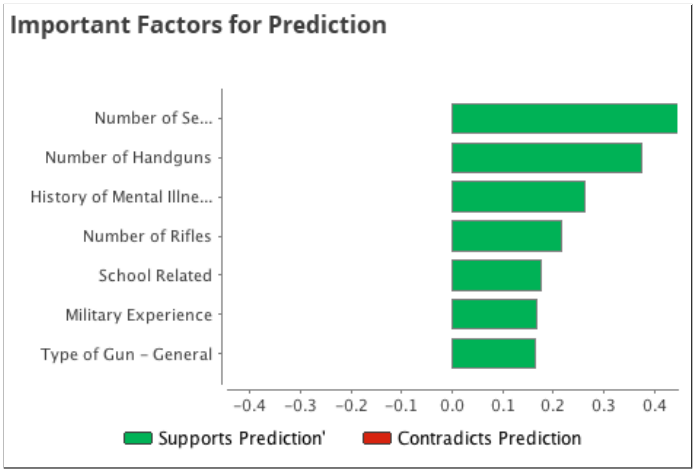

Figure 9: Deep Learning example of most important predictive factors.

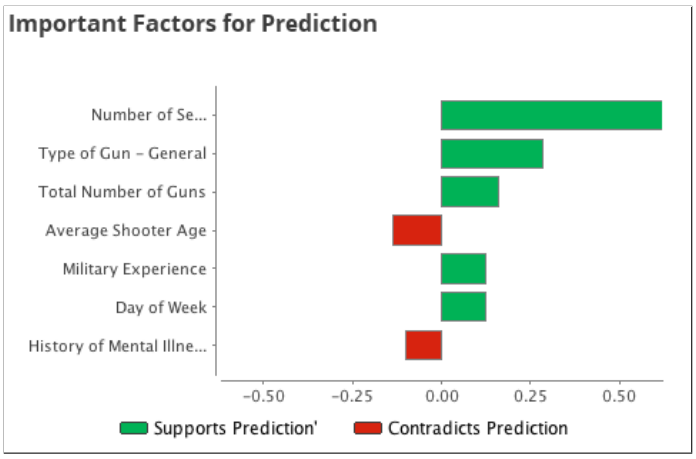

Figure 10: Random Forest example of most predictive features. Notice Mental illness is a negative indicator of a high victim count.

same information as the original one but help us generate and train smaller models since the amount of data is considerably reduced.

Figure 12 shows the confusion matrix and performance metrics of the LVQ Model described in section 3.2.2 for the 2014 ZIP Code dataset. As we can observe, the accuracy achieved in this case was around 82%.

Figure 13 shows a graph of the top 20 features according to their importance after performing the feature ranking algorithm described in section 3.2.2. As we can observe, the two top features are *Guns Manufacturing* and *Guns Licences*.

Figure 14 shows a graph of the accuracy for the 2014 ZIP Code dataset using different classification models depending on the number of features selected. In this case the graph reveals that the optimal number of features equals 4. The actual list of features is stored in the predictors field of the RFE model described in section 3.2.3: *Guns Manufacturing*, *Guns Licences*, *Percentage of non-hispanic white population* and *Percentage of non-hispanic Asian population*.
If we compare the four features selected to the feature ranking from figure 13 we notice that *Percentage of non-hispanic white population* by itself does not rank very high. However, as we explained in section 3.2.3, the RFE algorithm takes into consideration not only the individual rankings but also the interactions among the selected features to optimize the accuracy of the model.

Figure 15 shows the agnes dendrogram for the 2014 ZIP Code dataset with a cut value at $k = 4$. Given the huge number of records in this dataset (over 30,000) generating a good visualization of

```
> # Print highly correlated attributes
> print(colnames(data2014_mean)[highlyC2014])
 [1] "X86..Median.age"
 [2] "X82..Population.percentage..60.years.and.over"
 [3] "X21..Total.population.35.to.44.years"
 [4] "X36..Population.total"
 [5] "X57..Population.total"
 [6] "X83..Population.percentage..62.years.and.over"
 [7] "X18..Total.population.25.to.34.years"
 [8] "X10..Total.population.25.years.and.over"
 [9] "X24..Total.population.45.to.64.years"
[10] "X03..Household.median.income.in.dollars"
[11] "X04..Household.mean.income.in.dollars"
[12] "X93..Gun.purchases.approximate"
[13] "X16..Percentage.of.10..bachelor.s"
[14] "X81..Population.percentage..18.years.and.over"
[15] "X26..Percentage.of.24..bachelor.s.or.higher"
[16] "X47..Population.percentage..Hispanic"
[17] "X80..Population.percentage..16.years.and.over"
[18] "X37..Population.percentage..not.Hispanic"
[19] "X76..Population.percentage..5.to.14.years"
[20] "X78..Population.percentage..18.to.24.years"
[21] "X89..Old.age.dependency.ratio"
[22] "X56..Population.percentage..Hispanic.two.or.more.races.excluding.some.other.race"
[23] "X46..Population.percentage..not.Hispanic.two.or.more.races.excluding.some.other.race"
[24] "X58..Population.percentage..under.5.years"
>
```

Figure 11: Highly Correlated Features for the 2014 Dataset.

```
> print(cm2014)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7580 1420
         1  343  594

               Accuracy : 0.8226
                 95% CI : (0.8149, 0.83)
    No Information Rate : 0.7973
    P-Value [Acc > NIR] : 1.067e-10

                  Kappa : 0.3143
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9567
            Specificity : 0.2949
         Pos Pred Value : 0.8422
         Neg Pred Value : 0.6339
             Prevalence : 0.7973
         Detection Rate : 0.7628
   Detection Prevalence : 0.9057
      Balanced Accuracy : 0.6258

       'Positive' Class : 0
```

Figure 12: Confusion Matrix for the 2014 Dataset.

the clusters is quite challenging. However, in figure 16 we present a summary of the results for the clustering task:

(a) Agglomerative coefficient. The values for this metric range between 0 and 1, so a value of 0.998 is evidence of the strong clustering structure obtained by the algorithm.

(b) Number of elements in each cluster.

(c) Number of elements in each cluster distributed by label. This result is very interesting because during the clustering task the label column of the dataset is completely ignored (clustering is an unsupervised problem). Under this circumstances, we observe that the similarities found by the clustering algorithm do not necessarily coincide with the labels of the data, and therefore we can find data points from both categories in the same cluster.

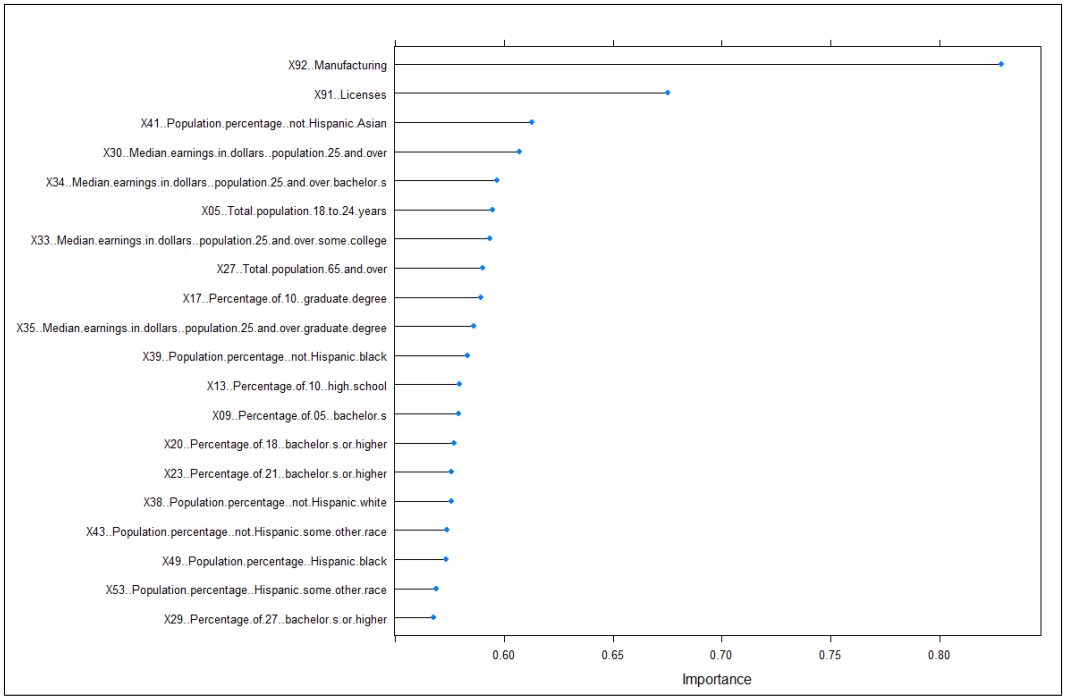

Figure 13: Top 20 Ranked Features for the 2014 Dataset.

Similar results to the ones presented in figures 11 through 16 were obtained for all our datasets. Table 3 summarizes all of these results.

4.3 Weka

4.3.1 Classification

Using Weka, we performed experiments looking at this problem in a classification scenario. We decided to use a couple of well known model such as REP Trees (decision trees), Adaboost, and Multilayer Perceptron (Neural Networks). For the first two, we ran experiments using all features, and then using only the high-ranked features found in the previous sections. However, we noticed no significant increase in performance when we used highly ranked features. Our results for both the Zip Code dataset and the Stanford MSA dataset are summarized in the following table 4. (Note that the Stanford MSA did not produce meaningful results in the binary classification scenario so we elected not to try every single model available).
For REP-Trees, we performed a Grid-Search on the maximum depth of the tree, using ten-fold cross-validation and testing the returned model on a hold-out test containing 20% of our test data. For Adaboost, we used a similar setup, we performed a Grid-Search on the number of weak learners (decision stumps), using five-fold cross-validation and testing the returned model on a hold-out test containing 20% of our test data. Finally, for Neural Networks we performed a Grid-Search on the number of hidden layers, again using five-fold cross-validation and testing our model on a hold-out set containing 20% of our data. We have provided our results in table 4, as well as some images of sample runs below in figures 17 18 19 20. Furthermore, Weka provides a neat visualization of the decision tree, which we have provided in figure 21. The first two attributes to split on are consistently Licenses or Manufacturing, followed by some demographic measure to income measure. These results are also consistent with our findings in feature ranking, as highly-ranked features are usually the first to be split on.

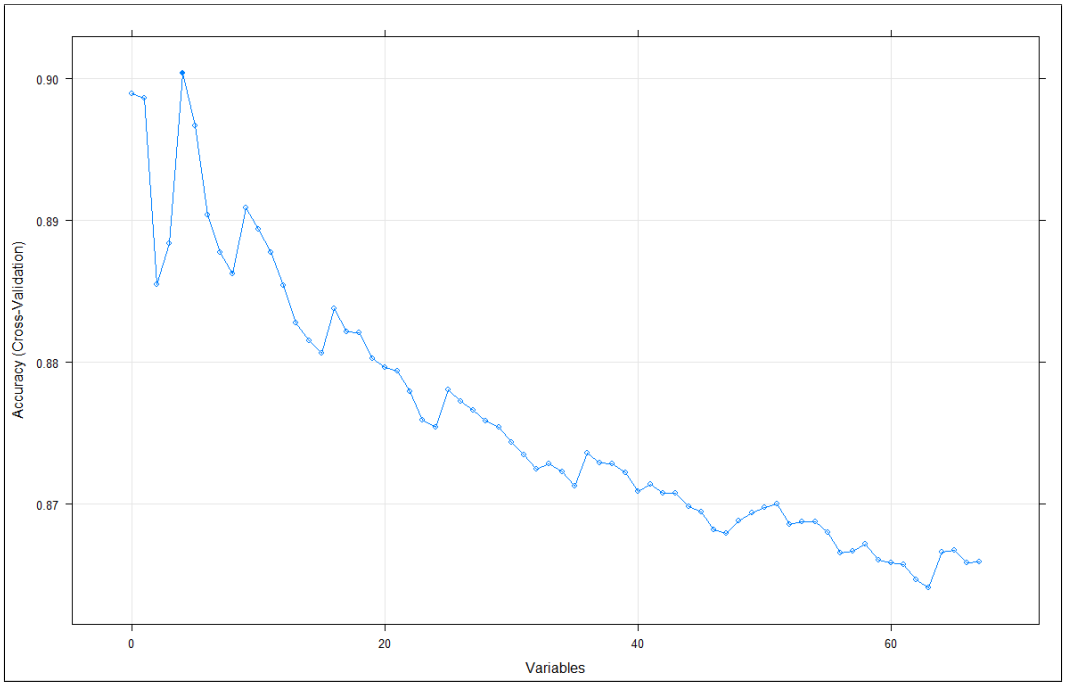

Figure 14: Accuracy vs. Features Selected for the 2014 Dataset.

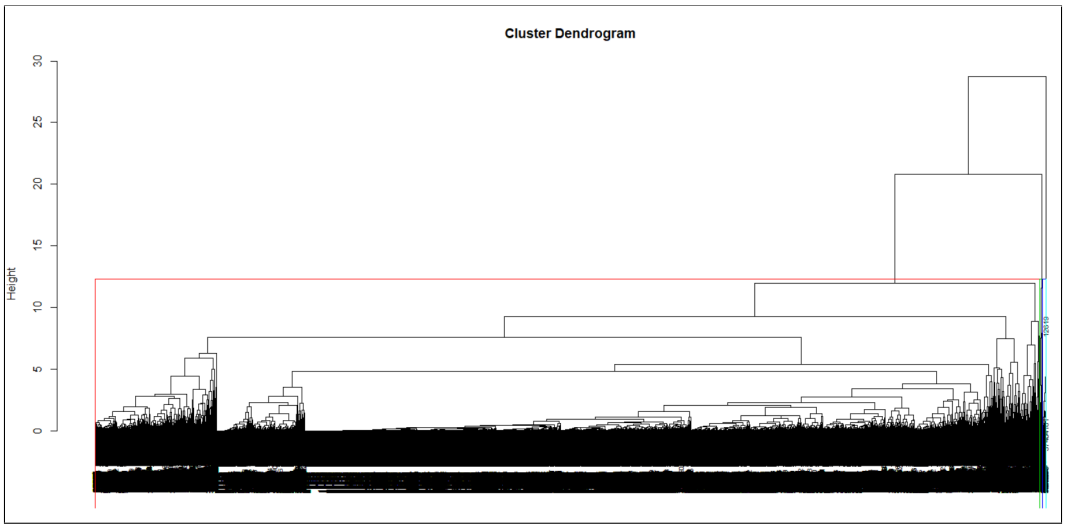

Figure 15: Dendrogram for the 2014 Dataset.

```
>
> agnes2014$ac
[1] 0.9986162
>
```

```
>
> table(clusters2014)
clusters2014
    1     2     3     4
32891   129    99     1
>
```

```
> cluster_label_2014
      0    1
1 26229 6662
2   111   18
3    68   31
4     1    0
```

Figure 16: Summary of Hierarchical Clustering for the 2014 Dataset.

Table 3: Summary of the Data Analysis in *R*

Results	Dataset			
	2014	2015	2016	Stanford
Features removed with CM	24	21	22	6
LVQ Model Accuracy	81.92%	80.06%	74.80%	75.25%
LVQ Model F-Score	89.41%	87.88%	84.90%	71.26%
Features selected with RFE	4	4	2	2
Agnes Agglomerative Coefficient	0.9986	0.9982	0.9998	0.9909

Table 4: Classification Results in Weka

Dataset	Model	Parameters	Accuracy	Precision	Recall
Zip Codes	Adaboost	Weak learners = 500	85.6%	84.4%	85.6%
	Neural Networks	Hidden layers = 4	82.7%	80.6%	82.7%
	REP-Trees	Max depth = 20	87.1%	86.4%	87.1%
Stanford MSA	Adaboost	Weak learners = 500	62.7%	63.3%	62.7%

```
Cross-validated Parameter selection.
Classifier: weka.classifiers.meta.AdaBoostM1
Cross-validation Parameter: '-I' ranged from 100.0 to 500.0 with 5.0 steps
Classifier Options: -I 500 -P 100 -S 1 -W weka.classifiers.trees.DecisionStump

Correctly Classified Instances        5667               85.5525 %
Incorrectly Classified Instances       957               14.4475 %
Kappa statistic                          0.4625
Mean absolute error                      0.2101
Root mean squared error                  0.3193
Relative absolute error                 65.3828 %
Root relative squared error             80.0666 %
Total Number of Instances             6624

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.961    0.570    0.872      0.961    0.914      0.485    0.880     0.962     0
                 0.430    0.039    0.731      0.430    0.541      0.485    0.880     0.668     1
Weighted Avg.    0.856    0.465    0.844      0.856    0.840      0.485    0.880     0.904

=== Confusion Matrix ===

    a    b   <-- classified as
 5102  208 |    a = 0
  749  565 |    b = 1
```

Figure 17: Results from running Adaboost on the 2014 Zip Code Dataset.

```
Cross-validated Parameter selection.
Classifier: weka.classifiers.trees.REPTree
Cross-validation Parameter: '-L' ranged from 10.0 to 100.0 with 10.0 steps
Classifier Options: -L 20 -M 2 -V 0.001 -N 3 -S 1 -I 0.0

Correctly Classified Instances        5772               87.1377 %
Incorrectly Classified Instances       852               12.8623 %
Kappa statistic                          0.5592
Mean absolute error                      0.1685
Root mean squared error                  0.3061
Relative absolute error                 52.4613 %
Root relative squared error             76.7598 %
Total Number of Instances             6624

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.947    0.434    0.898      0.947    0.922      0.566    0.904     0.970     0
                 0.566    0.053    0.725      0.566    0.636      0.566    0.904     0.713     1
Weighted Avg.    0.871    0.358    0.864      0.871    0.865      0.566    0.904     0.919

=== Confusion Matrix ===

    a    b   <-- classified as
 5028  282 |    a = 0
  570  744 |    b = 1
```

Figure 18: Results from running REPTrees on the 2014 Zip Code Dataset.

```
Cross-validated Parameter selection.
Classifier: weka.classifiers.functions.MultilayerPerceptron
Cross-validation Parameter: '-H' ranged from 1.0 to 4.0 with 4.0 steps
Classifier Options: -H 4 -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20

Correctly Classified Instances        5481               82.7446 %
Incorrectly Classified Instances      1143               17.2554 %
Kappa statistic                          0.3226
Mean absolute error                      0.2418
Root mean squared error                  0.3561
Relative absolute error                 75.2786 %
Root relative squared error             89.2903 %
Total Number of Instances             6624

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.958    0.699    0.847      0.958    0.899      0.354    0.790     0.920     0
                 0.301    0.042    0.638      0.301    0.409      0.354    0.790     0.533     1
Weighted Avg.    0.827    0.569    0.806      0.827    0.802      0.354    0.790     0.844

=== Confusion Matrix ===

    a    b   <-- classified as
 5086  224 |    a = 0
  919  395 |    b = 1
```

Figure 19: Results from running Neural Nets on the 2014 Zip Code Dataset.

```
Cross-validated Parameter selection.
Classifier: weka.classifiers.meta.AdaBoostM1
Cross-validation Parameter: '-I' ranged from 100.0 to 500.0 with 51.0 steps
Classifier Options: -I 444 -P 100 -S 1 -W weka.classifiers.trees.DecisionStump

Correctly Classified Instances          42               62.6866 %
Incorrectly Classified Instances        25               37.3134 %
Kappa statistic                          0.2445
Mean absolute error                      0.4394
Root mean squared error                  0.4825
Relative absolute error                 89.2915 %
Root relative squared error             97.5972 %
Total Number of Instances               67

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.607    0.359    0.548      0.607    0.576      0.245    0.663     0.615     0
               0.641    0.393    0.694      0.641    0.667      0.245    0.663     0.729     1
Weighted Avg.  0.627    0.379    0.633      0.627    0.629      0.245    0.663     0.682

=== Confusion Matrix ===

  a  b   <-- classified as
 17 11 |  a = 0
 14 25 |  b = 1
```

Figure 20: Results from running Adaboost on the Stanford MSA Dataset.

```
92) Manufacturing < 0.5
|   91) Licenses < 54.5 : 0 (11141/159) [5631/77]
|   91) Licenses >= 54.5
|   |   93) Gun purchases approximate < 281011.2
|   |   |   39) Population percentage: not Hispanic black < 4.34
|   |   |   |   76) Population percentage: 5 to 14 years < 14.35
|   |   |   |   |   91) Licenses < 108.5
|   |   |   |   |   |   "33) Median earnings in dollars, population 25 and over some colleg
|   |   |   |   |   |   "33) Median earnings in dollars, population 25 and over some colleg
|   |   |   |   |   91) Licenses >= 108.5 : 0 (48/10) [27/3]
|   |   |   |   76) Population percentage: 5 to 14 years >= 14.35 : 0 (66.27/0) [31/4]
|   |   |   39) Population percentage: not Hispanic black >= 4.34
|   |   |   |   93) Gun purchases approximate < 159189.45 : 0 (23/0) [17/1]
|   |   |   |   93) Gun purchases approximate >= 159189.45
|   |   |   |   |   55) Population percentage: Hispanic two or more races including some ot
|   |   |   |   |   |   28) Percentage of 27) high school or higher < 69.75
|   |   |   |   |   |   |   91) Licenses < 82 : 0 (8/0) [4/1]
|   |   |   |   |   |   |   91) Licenses >= 82 : 1 (12/4) [3/1]
|   |   |   |   |   |   28) Percentage of 27) high school or higher >= 69.75 : 1 (10/0) [7/
|   |   |   |   |   55) Population percentage: Hispanic two or more races including some ot
|   |   93) Gun purchases approximate >= 281011.2
|   |   |   93) Gun purchases approximate < 432082.6
|   |   |   |   93) Gun purchases approximate < 409078.95
|   |   |   |   |   73) Population percentage: 75 to 79 years < 2.05
|   |   |   |   |   |   72) Population percentage: 70 to 74 years < 2.5
|   |   |   |   |   |   |   13) Percentage of 10) high school < 40.8 : 1 (10/3) [5/1]
|   |   |   |   |   |   |   13) Percentage of 10) high school >= 40.8 : 0 (5/0) [1/0]
|   |   |   |   |   |   72) Population percentage: 70 to 74 years >= 2.5 : 1 (12/0) [13/6]
|   |   |   |   |   73) Population percentage: 75 to 79 years >= 2.05 : 0 (96/30) [51/15]
|   |   |   |   93) Gun purchases approximate >= 409078.95 : 1 (22/2) [4/0]
|   |   |   93) Gun purchases approximate >= 432082.6 : 0 (67/8) [41/3]
92) Manufacturing >= 0.5
|   41) Population percentage: not Hispanic Asian < 1.71
|   |   91) Licenses < 132.5
```

Figure 21: REP Tree Visualization for the 2014 Zip Code Dataset.

5 Conclusions

5.1 Zip Codes

In this experiment we set out to answer questions like: What cities/states/zip codes are more prone to such attacks? Is there any correlation with demographic data, (earnings, age, education level), guns licenses, guns manufactured, etc. ? Our results show that places have more distributors and manufacturers of guns are much more prone to violent attacks. Furthermore, our results through hierarchical clustering show that there are definitely some similarities between certain clusters of zip codes that have experienced violent attacks. In the classification scenario, using decision trees produced the best results, with 88% accuracy.

5.2 Stanford MSA

Our experiments with the MSA data set produced less fruitful results, perhaps pointing to the fact that there might not be that many similarities between the deadly events. We wanted to see whether there are certain characteristics that make a shooting deadlier than others, for example, the shooter's race, gender, income, educational attainment, etc. We also wanted to see what role does mental health play?. However, the only definitive thing we can say is that deadlier weapons + type of place = a deadlier shooting. Through clustering, we saw that there are some events that share similar features, but it's inconclusive if mental illness plays a role (one experiment said it did, another said it didn't, furthermore the accuracy of each experiment was about the same).

References

Jonathan Bouchet. Firearm regulations in the U.S., 2017. URL https://www.kaggle.com/jonathanbouchet/firearm-regulations-in-the-u-s.

Jason Brownlee. Feature Selection with the Caret R Package, 2014. URL https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/.

United States Census Bureau. American Fact Finder, 2018. URL https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

Stanford Geospatial Center and Stanford Libraries. Stanford mass shootings in america, courtesy of the stanford geospatial center and stanford libraries, 2018. URL https://github.com/StanfordGeospatialCenter/MSA.

Tableau Community. What is Tableau?, 2018. URL https://community.tableau.com/thread/238491.

Mark Follman, Gavin Aronsen, and Deanna Pan. A Guide to Mass Shootings in America, 2018. URL https://www.motherjones.com/politics/2012/07/mass-shootings-map/2/.

UC Business Analytics R Programming Guide. Hierarchical Cluster Analysis, 2018. URL https://uc-r.github.io/hc_clustering.

Roy Kwon and Joseph F. Cabrera. Socioeconomic factors and mass shootings in the United States, 2017. URL https://www.tandfonline.com/doi/abs/10.1080/09581596.2017.1383599.

Kevin Loria. Gun control really works — here's the science to prove it, 2018. URL http://www.businessinsider.com/science-of-gun-control-what-works-2018-2.

Daniel P. Mears, Melissa M. Moon, and Angela J. Thielo. Columbine Revisited: Myths and Realities About the Bullying–School Shootings Connection, 2017. URL https://www.tandfonline.com/doi/abs/10.1080/15564886.2017.1307295.

Jonathan M. Metzl and Kenneth T. MacLeish. Mental Illness, Mass Shootings, and the Politics of American Firearms, 2014. URL https://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2014.302242.

Zachary Smith. US Mass Shooting Analysis, 2017. URL https://www.kaggle.com/zacksmith32/us-mass-shooting-analysis/notebook.

Wolfgang Stroebe, N. Pontus Leander, and Arie W. Kruglanski. The impact of the Orlando mass shooting on fear of victimization and gun-purchasing intentions, 2017. URL http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0182408.