Analysis and Prediction of Youth Suicide Ideation

Group#1
Bala Sai Yashvanth
Chaitanya Tanga
Rahul Kumar Chevvuri
Rohit Kalva

UNIVERSITY AT BUFFALO

December 2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 PROJECT ABSTRACT

Out of the many contributory factors for deaths in USA, Suicide is one of the crucial reasons for deaths in the country. Suicides can possibly follow various socio-cultural norms or constraints of the societies of their occurrence. However that being said there can be a possibility of common set of factors being the contributing reasons that may address the increasing trend of suicides worldwide and specifically in the USA , a trend that is being observed irrespective of age groups, language, ethnicity and gender. The main purpose of this research project is to develop a model-based framework in order to ascertain, analyze and correlate the common and possibly impact factors that may impact suicide drive amongst victims. To keep the scope of the project manageable the focus would be on the suicide trends in the United States with a special focus on the major contributing subgroup of youngsters as the primary input census for developing the model. The project will primarily utilize data set generated from Youth Risk Behavioral system (YBRS).The data set requirement would for example generally consist of various age groups, genders, various age related peer pressure factors such as smoking, drinking, online abuse, financial debts, Medication abuse, depression etc. all of these which can provide significant insight into impact factors influencing suicide drive. In the end the research will provide an in-depth understanding and knowledge about the correlation between various factors for an in-depth casual analysis that can possibly contribute in addressing significant spike in suicides in the recent times especially amongst the youngsters. This predictive analysis will not only help in significant understanding of the various factors that have been contributing to suicides but also will help in developing various prevention strategies and rehabilitative measures for both potential victims and victims of failed suicide attempts.

2 LITERATURE REVIEW

It is seen that among the deaths of youth in the age group of 10-19, suicide stands 3rd in the list of many other factors. A lot of research has been conducted to analyze the suicides among youth. Many factors were observed that initiate the thought of suicide among the youths such as sexual harassment, bullying, career development, academic progress etc.

It has been observed that there are some ill effects of alcohol on the increasing youth suicides. The sample set was sorted according to the gender and age groups where the suicide rates across many years were tabulated. Empirical link between alcohol use and the suicides has been studied and the results were presented. The link is hypothesized using 2 statements, first, the correlation between alcohol consumption and the suicide incidents and the second, negative relation between price, taxes on alcohol and the consumption[1]. The observed results were different across the 2 genders. Male suicides were negatively affected by the laws and the price & taxes on alcohol. However, these factors have no effect on the female suicides. The paper further suggests incorporating drug uses and relevant policies for better understanding.

It was also analyzed that the suicide pattern difference from Urban and Rural areas. County level was gathered from 3141 counties and these were sorted accordingly based on urban and rural areas, population and various others. It was observed that rural area youth suicides are double that of urban suicides, which is an alarming issue. These results can be used to implement various development plans and schemes which can specifically focus towards rural areas and thereby rural areas can be developed and the suicide rates can also be reduced significantly[2].

Another study was performed based on the data obtained after the autopsy. It was seen that 41 individuals below the age of 18 just in NYC committed suicide in 2002 alone. Out of which one death was seen related to antidepressants[3]. Although these had little to no link between anti-depressants and suicide, we cannot ignore the suicide thoughts that a youth taking anti-depressants has.

It was also studied that the youth are getting affected through internet and by learning from others. The paper [4] has studied to determine if these news and other internet articles which focus on suicides and expose people who have attempted/committed suicides persuade young people to suicide or implant any suicidal ideas. Survey was conducted where people were asked if they knew any person who has committed suicide or heard any stories from other friends/relatives. After a span of time the same set were surveyed on the effects of internet and other social media news and articles towards suicide. The results showed that the internet has a significant affect towards inculcating suicidal thoughts on youth. It concluded that help sites are to be kept on internet which speak directly to the suicidal youth.

According to [5] almost half of youth suicides have firearms usage involved. He examined the association between the federal laws which included age specific restriction and suicides among the youth. It was seen that that introducing CAP laws had a fair play in reduction of suicides among the youth aged between 14-17 years of age. It was also noted that with current age restriction on firearms don't have a significant reduction of suicide among the youth.

There is a study performed by [6] on the suicides in Tiwi people of Bathurst islands in the northern territory which showed some new corner reasons for youth suicides. It was seen

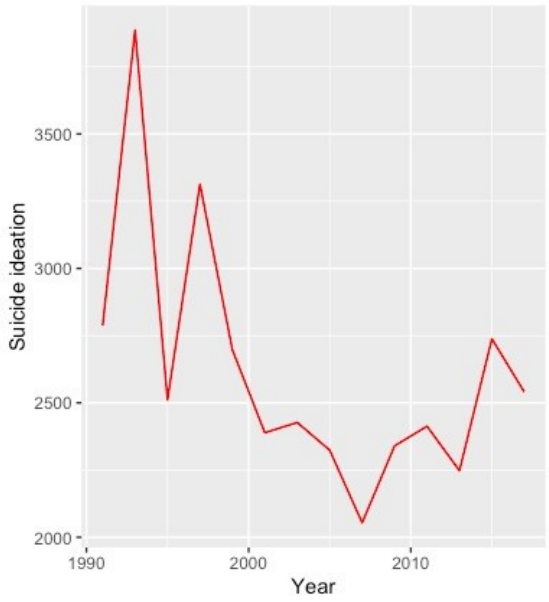
that suicide among aboriginal was observed among the people who has previous history of self-harm behavior. Social disruption was also seen as a major cause among the community.

It is also observed in [7] that, Individual, family, peer and school level risk factors also played a role in suicide among sexual minority youth. He focused on the causes of suicide in sexual minority status people. It is seen that self-identified gay male youth are among those most at risk. Youth engage with same gender attraction or same sex orientations are at risk for committing suicide. It is also noted by the author that most of the suicides are occurred when the respondents were of age 25 or younger. Furthermore, these suicides of prevalence of parasuicide during gay and bisexual males are increasing day by day.

[8] quoted that many reports have considered that gay, lesbian and bisexual youth we are part of vulnerable outcomes, including suicide. Hence it suggests applying 4 protective factors which are family connectedness, teacher caring, other adult caring, and school safety that can help to reduce the suicides among them. Logistic regression was used to study the effects of these protective factors. It is also noticed that GLB youth are at higher risk than the non GLB youth for commuting suicides.

After a thorough literature review, it is found that there are many factors that are affecting and increasing the suicide rates among youth. From the collected data sets and through the literature review, many such factors as listed above were identified. These would be used to predict the risk factors leading to youth suicide and can be used to get a better understanding on these factors. Further, our study can be used to fill out the gaps in literature that the above articles have missed to point in their papers.

3 DATA INTERPRETATION



(a) Suicide Ideations

data2.jpg

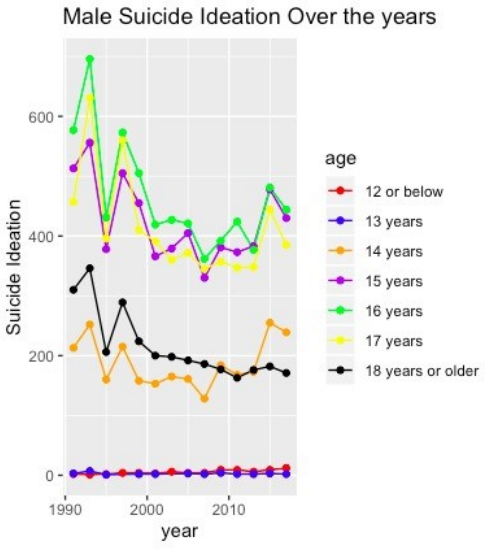
(b) Suicide Ideation-Male and Female

Figure 3.1: Suicide Ideation

There are some unusual trends in the suicidal ideation rates through the year. Further, it can be observed that the trend usually tries to decrease apart from few instances of high increments which can be attributed to various reasons/scenarios happening around the world during those years. From the second graph, it can be clearly interpreted that male suicide rate is considerably high than the female suicide rate. Even the gender wise suicide rate follows the general suicide ideation trends of unusual variation during some years.

In Male Suicide Ideation over the years graphical data it can be inferred that there is no specific overlap in male suicidal thoughts at any given age group, Additionally, there is a consistently low male suicidal thoughts for age groups of 13, 12 and below. There may possibly be multiple factors in play that might have caused this trend, Additionally there has been a unusual spike in suicidal thoughts in all major age groups except 18 years and above during early 2008 to early 2010's which cannot as expected be attributed to the local and various global socio economic factors in play during that period.

The Female Suicide Ideation over the years graph indicates female suicide ideation rates from 1990 to 2017, from the above data it can be inferred there is a unusually high rate of suicidal thoughts among 16 and 17 year old females, although it can also be inferred from the data there has been a perceivable decrease in the these thoughts in the above mentioned target groups due to possible concerted counselling efforts on the specific age groups.



(a) Agewise Suicide Ideation-male

data4.jpg

(b) Agewise Suicide Ideation-female

Figure 3.2: Agewise Suicide Ideation

The below link show the year wise trends in the suicide ideation rates for every state through many years. Almost every state maintains its levels throughout the range of years. However, New York, New Hampshire, and Maine do not follow the levels but shows an immediate increment in the ideation rates abruptly.

https://plot.ly/ chaitu2595/45//

4 DATA PREPROCESSING

The data set was originally a survey questionnaire which had various levels of answers and around 150,000 data points. However, among these values, there were many NA values. Because of large number of NA values, the data was cleaned and preprocessed, then the columns which had more than 70% NA were removed and finally the data points were brought down to a count of 23,000. Further, as the questionnaire had levels of answers, they were then converted to categorical variables. Because these variables were categorical, it became difficult to correlate with Pearson correlation test. Hence, we have used Goodman and Krushal tau's correlation test. These data points were then partitioned into training and test data where training set has 80% of the data points and test set has around 20% of them.

4.1 CORRELATION PLOT

The predictive correspondence of the variables can be found using the 'Goodman and Kruskal' tau's correlation test.

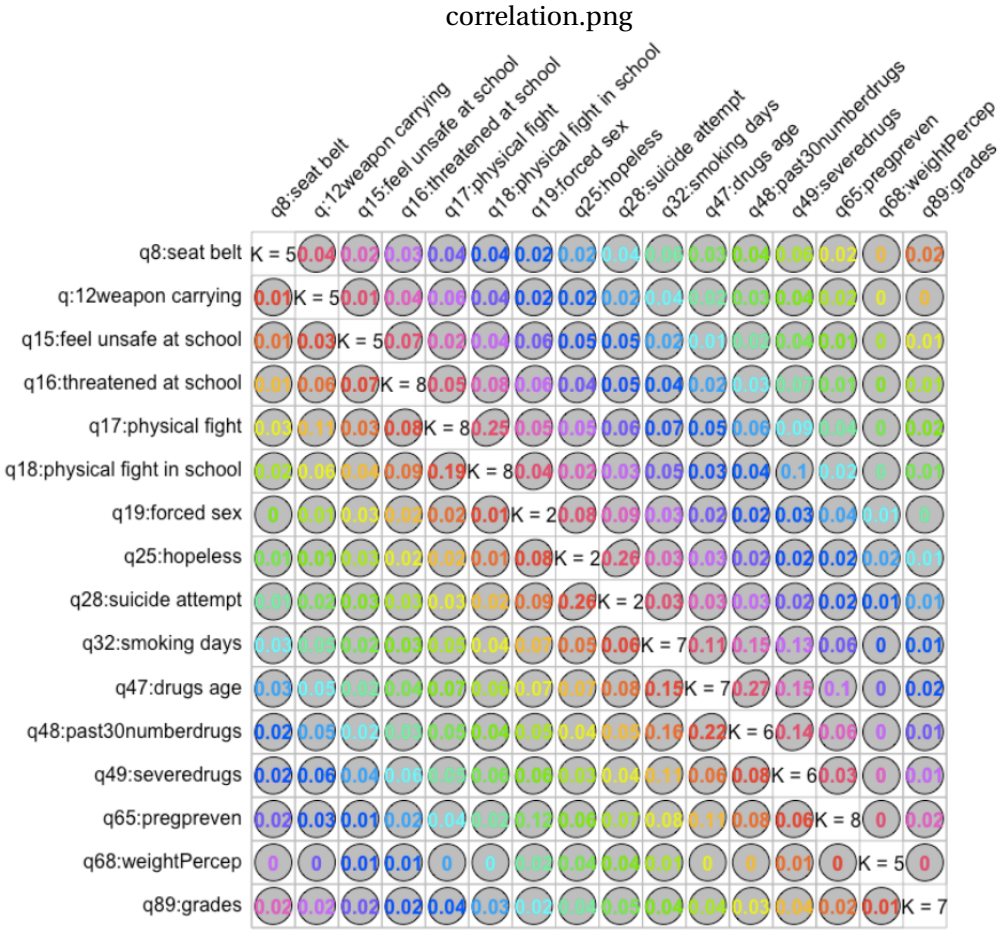


Figure 4.1: Correlation Plot

For example, consider tau(suicide attempt, seat belt)[forward association]=0.04,tau(seat belt, suicide attempt)[Backward association]=0.02 indicating that suicide attempt can be predicted better from seatbelt but the other way (seat belt indicator from suicide attempt).

The variables seem to be very predictable but not on a higher degree, the best predictor of suicide attempt would be hopelessness.

4.2 Dimensionality Reduction

As all these variables are categorical, we can check if we can reduce the dimensionality. This can be done using Multiple Correspondence Analysis (which is similar to Principal Component analysis (PCA), PCA is done for continuous data whereas MCA is done for categorical data).

We can observe that the Dimensions fail to explain the variability in the data, i.e. we need a lot of Dimensions to have high variability, this may be due the presence of many levels in the predictors.

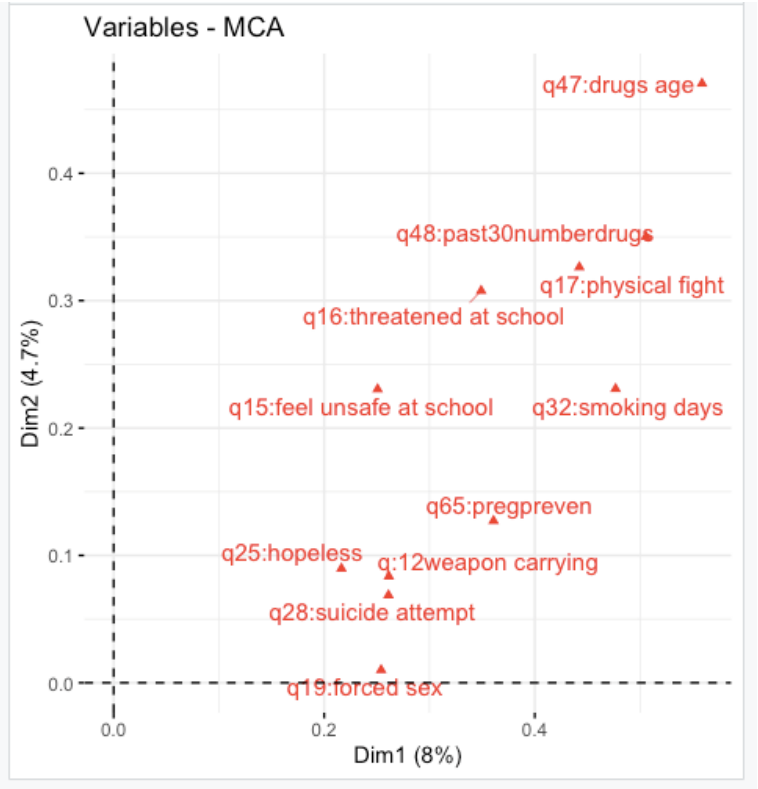


Figure 4.2: correspondence of the predictors with Dimensions 1 and 2

The above figure displays the correspondence of the predictors with Dimensions 1 and 2, we can similarly plot the correspondence with all the Dimensions.
The above plot shows the distribution of levels in Suicide ideation with respect to the Dimensions. We may be able to increase the variability explained by the dimensions by reducing the levels in predictors. Which can be done by grouping the levels in to just 2 categories. [This has been left out for end submission]
As Dimensional Reduction doesn't seem to be a viable option right now, we decided to use all the available predictors.

5 METHODOLOGY

5.1 GENERALIZED LINEAR MODEL (GLM)

General linear models are the models extend the conventional linear modelling framework to variables that are not normally distributed. The form of GLM is yi N(xiT β, σ2), where xi has covariates which are known and β has coefficients that can be estimated. GLMs are mostly used to either model binary or count data. In this model the word general means the dependence on more than one variable. And the errors εi are independent and identically distributed i.e.,

$$E[\varepsilon_i] = 0$$

$$var[\varepsilon_i] = \sigma^2$$

Generalized linear models are mainly made of linear predictors and two functions called link and variance.

$$n_i = \beta_0 + \beta_1 + \beta_2 + ... + \beta_n$$

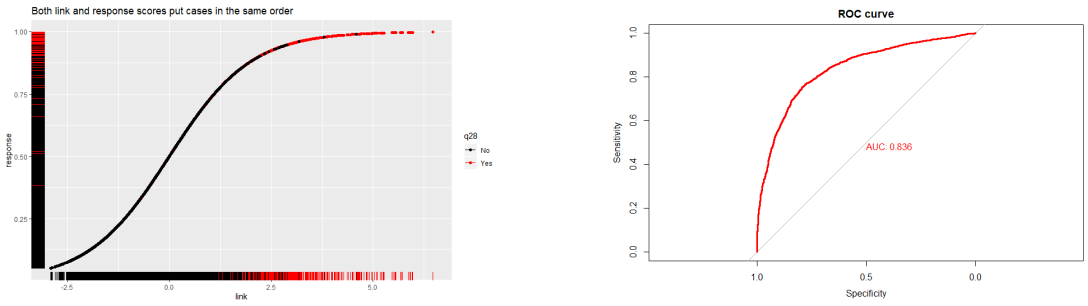
They have a broad class of models like linear regression, ANOVA, ANCOVA, log linear etc. The table below shows various models in GLM .

Model	Random	Link	Systematic
Linear Regression	Normal	Identity	Continuous
Logistics Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical

Table 5.1: Components in various models

Random Component: It tells about the probability distribution of the response variable Y.
Systematic component: It refers the explanatory variables in model, specially their linear combination in creating linear predictor.
Link Function: It specifies what type of link between systematic and random component

Of these, we have used the logistics regression to model the datasets.



(a) Response Curve

(b) ROC Curve

Figure 5.1: Response Curve & ROC Curve

5.2 Recursive Partitioning

Recursive partitioning is a statistical method in multi-variable analysis. It creates a decision tree that classifies member of population by splitting into sub populations based on several independent variables. It is called as recursive as each sub population is again split into many other indefinite number of times until the splitting process is stopped for some stopping criteria is reached.

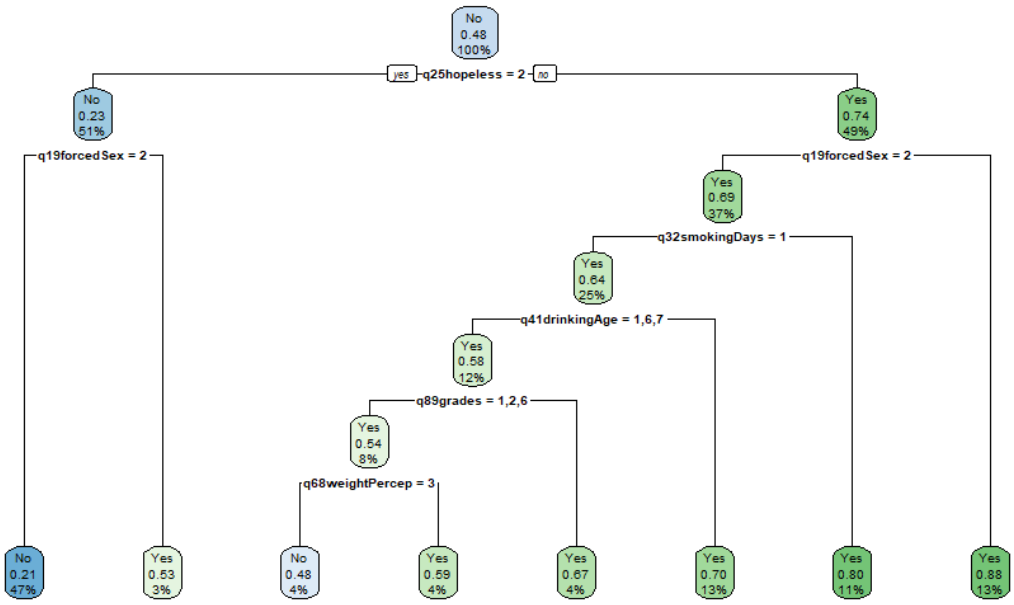


Figure 5.2: Recurssive Partitioning

5.3 BAGGING

Bagging models, also known as Bootstrap aggregating prediction models which is used to fit multiple versions of prediction models and then combine them into an aggregated prediction. Bagging model works in a way that duplicates the original data into b bootstrap copies, which are referred as base learner and applies its algorithm to each of the copies that were created and averages all of the predictions from the duplicated copies. The above algorithm can be represented as the following equation

$$F_{bag} = f_1(X) + f_2(X) + ... + f_b(X)$$

Due to the averages that are being calculated across the duplicates, bagging reduces the variance. Bagging generally works for unstable and high variance base learners, where the outputs gets subjected to major changes by a minimal change in the training data. However, bagging does not offer a greater improvement on the individual base learners.

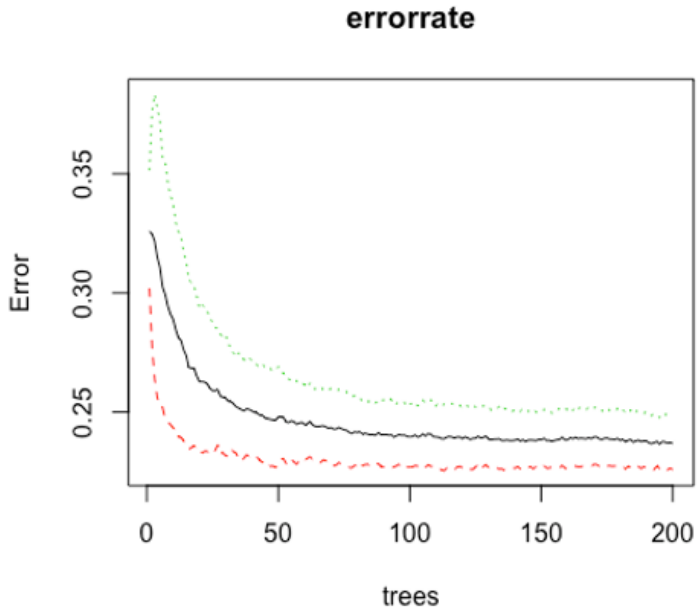


Figure 5.3: Error Rate

5.4 RANDOM FOREST

Random Forests are an ensemble learning method for classification, regression and other tasks that operate 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. Random forests are an improved technique over bagged trees technique in which a decorrelation of the trees takes place as a consequence of which there is a reduction

in the variance when averaging the trees. The random tree technique when building these decision trees takes the condition that when a tree splitting is considered each time, a random selection of "n" predictors is chosen as split candidates from the full set of m predictors. The split is allowed to use only one of those "n" random predictors.

The random forest technique works in the following way in which a fresh selection of n predictors is taken at each split, and typically we choose n ≈ $\sqrt{m}$ which means the number of predictors considered at each split is approximately equal to the square root of the total number of predictors. This method helps in reducing the variance of a single tree by forcing each split to consider only a subset of predictors. Number of trees were around 200 and the model accuracy was 76.39%.

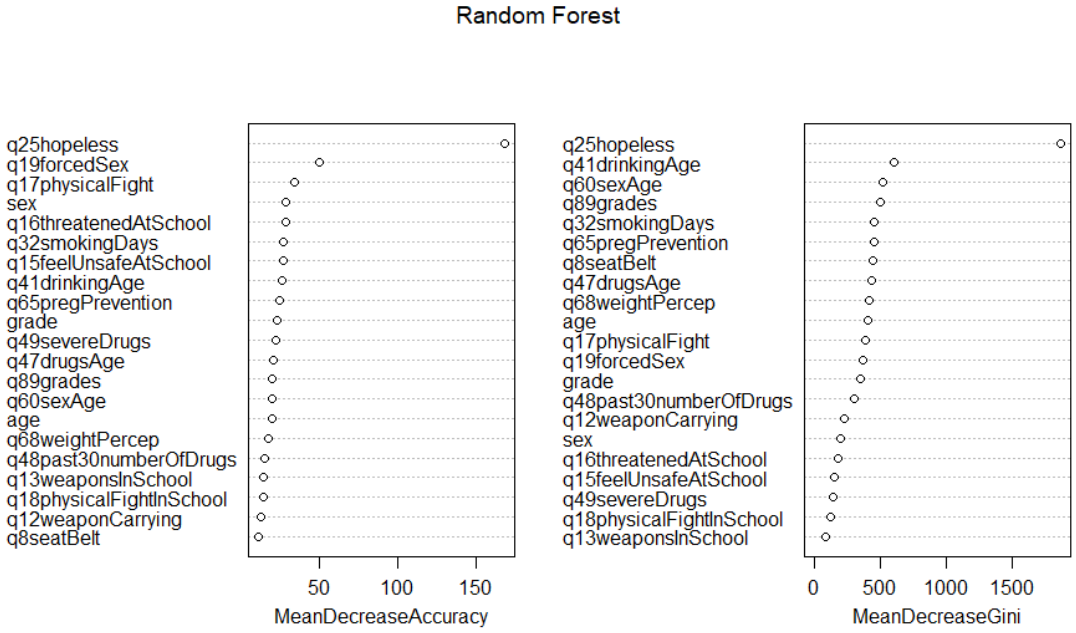


Figure 5.4: Random Forest: Predictor Relevancy

5.5 BOOSTING

Boosting is similar to bagging in terms of creating multiple copies of the original training set. However, unlike fitting a seperate decision tree on a bootstrap data set independent of other trees, boosting works in such a way that the trees grow sequentially, i.e. each tree is grown using the information from the previously grown trees. Hence, we can say that boosting does not involve bootstrap sampling. For the model, 5000 trees were used and the optimal value of shrinkage is 0.01. Accuracy of the model was around 77.16% which is highest among other models.

$$\hat{f}(x) = \sum_{n=1}^{B} \lambda \hat{f}(x)$$

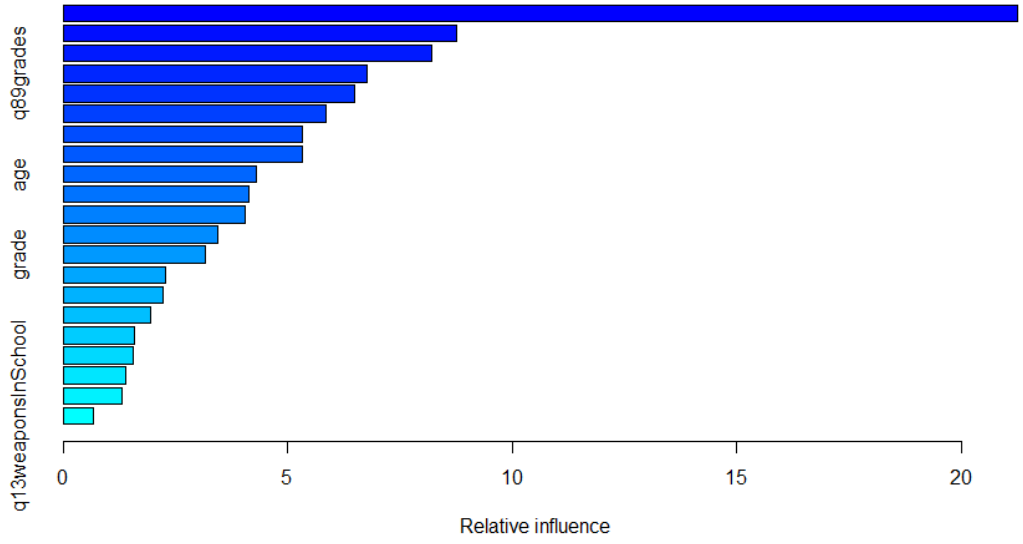


Figure 5.5: Relative Influence

5.6 Support Vector Machine

A Support Vector Machine(SVM) is a discriminative classifier defined by using a separating hyper-plane. In this the given labelled training data is used on the algorithm. It outputs an optimal hyper-plane which categorizes new examples. In 2 Dimensional space this hyper-plane is a line dividing a plane in two parts where in each class lay in either side. SVM performs in many kinds of settings perfectly, it is considered as one of the best “out of the box” classifiers.

Linear and Radial kernels were used and among these, radial was found to be more accurate with almost 10,000 support vectors with a cost parameter of 0.01 and gamma of 0.001. The accuracy was found to be 75.77%.

5.7 K-Nearest Neighbors

KNN is a non-parametric learning algorithm. It uses a database in which the data points are separated into several classes to predict the classification of a new sample point. It also stores all the available cases and classifies the new data or case based on a similarity index. It mostly used to classify a data point based on how its neighbours are classified. Also knownas lazy algorithm as it does not have any training step. And all data points will be used only at the time of prediction. It is used for both classification and regression.

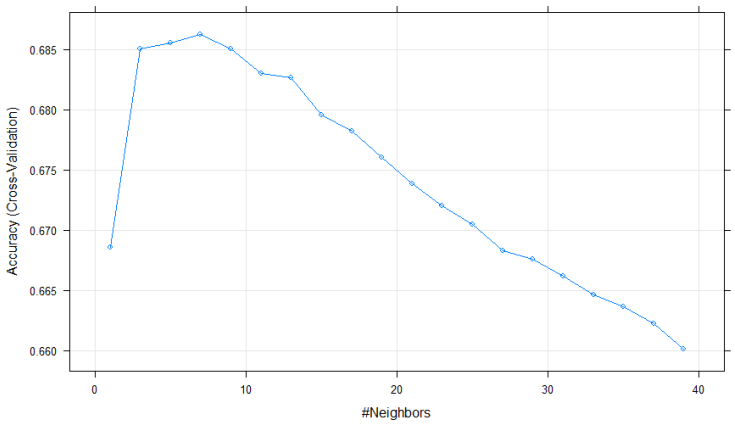


Figure 5.6: Accuracy with various K Values

6 RESULTS

6.1 Discussion

The following table describes the comparison of model accuracies.

Model	Accuracy
Logistics Regression	74.8346%
K-Nearest Neighbors	68%
Recursive Partitioning	75.77%
Bagging	75.15%
Random Forest	76.39%
Boosting	77.16%
Support Vector Machine	75.77%

Table 6.1: Model Accuracies

The model accuracies are all almost around or more than 75% which shows that all the models are prominent enough. However, K-Nearest Neighbors (KNN) has a small accuracy because there were many predictors and it is not easy for the KNN model to map these large number of predictors while considering the nearby data points for the results.

Further, the highest accuracy was obtained for the boosting model with an accuracy of 77.16%.

6.2 Conclusion

The results obtained from all the model methodologies gives out the prominent predictors that influence youth to have suicidal thoughts and ideations. Of all the results we have found,

hopelessness is the most prominent factor. The other findings are forced sex, use of guns etc. Below are some other results that the models have found.

1. GRADE (DIVISION) - The grade in which the youngster is studying.
2. FEELING UNSAFE AT SCHOOL - Which is linked to the external environment of the institution the youngster is studying.
3. THREATENED AT SCHOOL - Pertains to the possibility of bullying in the Institution.
4. PHYSICAL FIGHT - Getting into Physical altercations with Peers, Seniors at the institution.
5. HOPELESSNESS - A feeling that possibly has links to the youngster's perception of his educational performance, Bullying, Mental Harassment, blow to their perception of thinking.
6. DRINKING AGE - Points to the Age at which youngster might have started drinking.
7. SMOKING DAYS - The term refers to the frequency of smoking of the youngster.
8. FORCED SEX - Possibly points out to instances of Sexual Intercourse under Coercion of peers, Partner and its effects on the psych of the youngster.
9. AGE DURING FIRST SEXUAL INTERCOURSE - The age at which the youngster might have engaged in sexual Intercourse.
10. GENDER - The gender of the candidate (Indirect factors to this include Societal, Ethnic, Religious Expectations based on gender).
11. SEVERE DRUGS - Usage of drugs which affect the mental condition of youth and has a major influence in the ideation of suicidal thoughts.
12. WEIGHT PERCEPTION - Perception of their weight by the youngster which is correlated to Bullying, Gender perception, Peer Expectations.
13. CLASS GRADES - Points to the Personal expectations which is correlated to the Societal Expectation on the Academic Performance of the youngster

7 Future Research

The Questionnaire that we found had only limited variables which hasn't considered some other important factors that would initiate the suicidal thoughts. For example, there can be various psychological behaviours among the youth which could arise because of racial differences, income status etc. Also, we had a lot of NA values in the survey. These large number of NA values could have made a difference in interpreting the results. Further, the answers were in categorical variables. If they would have been in ordinal variables, the obtained results would be different.

REFERENCES

[1] Markowitz, S., Chatterji, P., & Kaestner, R. (2003). Estimating the impact of alcohol policies on youth suicides. *Journal of Mental Health Policy and Economics*, 6(1), 37-46.

[2] Fontanella, C. A., Hiance-Steelesmith, D. L., Phillips, G. S., Bridge, J. A., Lester, N., Sweeney, H. A., & Campo, J. V. (2015). Widening rural-urban disparities in youth suicides, United States, 1996-2010. *JAMA pediatrics, 169*(5), 466-473.

[3] Leon, A. C., Marzuk, P. M., Tardiff, K., Bucciarelli, A., PIPER, K. M., & Galea, S. (2006). Antidepressants and youth suicide in New York City, 1999–2002. *Journal of the American Academy of Child & Adolescent Psychiatry, 45*(9), 1054-1058.

[4] Dunlop, S. M., More, E., & Romer, D. (2011). Where do youth learn about suicides on the Internet, and what influence does this have on suicidal ideation?. *Journal of child psychology and psychiatry, 52*(10), 1073-1080

[5] Webster, D. W., Vernick, J. S., Zeoli, A. M., & Manganello, J. A. (2004). Association between youth-focused firearm laws and youth suicides. *Jama, 292*(5), 594-601.

[6] Parker, R., & Ben-Tovim, D. I. (2002). A study of factors affecting suicide in Aboriginal and 'other'populations in the Top End of the Northern Territory through an audit of coronial records. *Australian and New Zealand Journal of Psychiatry, 36*(3), 404-410.

[7] Russell, S. T. (2003). Sexual minority youth and suicide risk. *American Behavioral Scientist, 46*(9), 1241-1257.

[8] Eisenberg, M. E., Resnick, M. D. (2006). Suicidality among gay, lesbian and bisexual youth: The role of protective factors. *Journal of adolescent health, 39*(5), 662-668.