ISyE 6740 – Summer 2020
Final Project

Using Machine Learning Models to Predict Suicide Risk among High School Students in California

Anh L Tran

1. Problem Statement

According to the Youth Risk Behavior Survey (YRBS) – Data Summary and Trend Report 2007-2017, the rates of suicide ideation and attempt among high school students increased significantly from 2007 through 2017 [8]. Moreover, the prevalence of suicide thoughts and attempts vary among different demographic groups and strongly associate with certain health-related risky behaviors [1,4,6]. For example, injection drug use, carrying a weapon on school property, and methamphetamine use were strongly associated with suicide attempts among female students while male students' high suicide risk included injection drug use, using vomiting or laxatives for weight control, and ever having been forced to have sex [6]. Plus, previous non-fatal suicide attempts are likely to increase the chance of future and more severe suicide attempts. Therefore, being able to identify and predict the patterns and trends of students' health-risk behaviors in association with suicide thoughts and attempts are critical in order for providing appropriate early intervention and prevention services.

Machine learning techniques have been used to examine large and complex data to identify hidden patterns and intricate relationships among factors as well as to efficiently produce statistical prediction of a given outcome [2,3,5,10]. Some studies have applied machine learning models such as logistic regression, classification tree, natural language processing, etc. to predict suicide risk and determine health-related behaviors using a variety of data sets [2,3,5]. However, different types of data set (e.g. survey data, clinical notes, administration data, etc.) and populations (e.g. region, age, gender, etc.) might provide different results and exhibit different trends in relation to suicide risk and associated health-risk behaviors. For this final project, I focus on predicting the suicide risk among high school students in California. Specifically, I used the YRBS data and three machine learning classification models: binary logistic regression, k-nearest neighbor, and decision tree to predict low and high suicide risk based on a variety of health-risk behaviors.

2. Methodology

2.1 Data source

The national school-based Youth Risk Behavior Survey (YRBS) was developed and conducted biennially by the Centers for Disease Control and Prevention (CDC). The survey uses three-stage cluster sample design and contains approximately 98 items that ask U.S. grades 9 through 12 students many health-related behaviors such as sexual behavior, substance use, violence victimization, mental health and suicide, unhealthy dietary behaviors, and inadequate physical activity [8]. In this project, I used the YRBS data limited to the state of California. Currently, California YRBS data only include the 2015 and 2017 survey data (3721 surveys total).

Four suicide-related items were chosen as my target variables. Two items ask about suicidal thoughts: "During the past 12 months, did you ever seriously consider attempting suicide?" and "During the past 12 months, did you make a plan about how you would attempt suicide?" Two items ask about suicide attempts: "During the past 12 months, how many times did you actually attempt suicide?" and "If you attempted suicide during the past 12 months, did any attempt result in an injury, poisoning, or overdose that had to be treated by a doctor or nurse?" To ensure that I have a full response data, I excluded rows that have missing data on all those four suicide-related questions. The final sample for analyses is 3707 surveys. Participants' demographic characteristics are shown in Table 1.

2.2. Suicide Risk

I combined the aforementioned four suicide-related items to create one binary (0-1) suicide risk variable: low risk, defined as having no suicidal thoughts or attempts during the past 12 months; and high risk, defined as answering "yes" to at least one suicidal thought question and/or attempting suicide at least once. As shown in Table 1, 22.8% students were in high risk for suicide.

2.3 Health-related Behaviors

Twenty-three health-risk behaviors from YRBS were selected to determine their associations with suicide risks. These behavior items can be classified into eight main categories: community-related violence, school-related violence, mental health issue, substance use, sexual health, weight-related issue, sedentary activities, and low academic performance. Table 2 lists all the details and analytic coding for these items.

Differences in the prevalence of demographic characteristics as well as health-risk behaviors at different level of suicide risk were assessed by chi-square test using R version 3.6. Because the YRBS data were obtained from a three-stage cluster sample design, analyses must account for the sampling design (stratification, clustering, and unequal selection probabilities) [9]. I used the "survey" package, which includes the sampling

design information (i.e. weight, stratum, and primary sampling unit or PSU) with the data into a 'survey design object' for calculating prevalence and chi-square test.

2.4 Data imputation

Missing data might cause some issues and affect many machine learning algorithms, and survey data is notoriously known to have lots of missing data. YRBS data is not an exception. In fact, for the selected variables in the current data, 990 observations and 27 variables (4 demographic variables + 23 health-risk behavior variables) have at least one missing data. Among them, qn62 "currently sexually active" and qn83 "did not play in a sport team" have the highest percent of missing data (about 14%). The Multivariate Imputation by Chained Equations (MICE) package was used to replace the missing data with plausibly predicted values computed by logistic regression (for binary health-risk behavior variables) and Bayesian polytomous regression (for demographic variables). Because some variables are highly correlated (e.g. qn42, qn48, qn62) as shown in figure 1, I applied the "quickpred" package to adjust the values in the predictor matrix based on the bivariate correlations of the variables in the data with 10% default cutoff [7].

The procedure produced five new imputed data sets. The missing values from the original data set then were replaced by the mode of the imputed data for each column per record. This new data set was used to build the models.

2.5 Training and Testing Dataset

Because the level of suicide risk is unbalanced (77% low vs. 23% high), I separated all the low-risk data and all the high-risk data into two data sets. For each of them, I randomly sampled 80% as training data set and the remaining 20% as testing data set; then I combined the corresponding low and high data sets together.

2.6 Machine Learning Models

A. Logistic Regression Model

Logistic regression, also known as a logit model, is a type of generalized linear regression model that predicts a dichotomous outcome variable (e.g. $y_i = 1$ for high-risk suicide and $y_i = 0$ for low-risk suicide; i is a random number in n data points) with one or more predictors (e.g. $x_i = (x_{i1},\ldots,x_{ip})$; p is a selected predictive variable). By applying logistic regression formula, I computed the probability of high-risk suicide Y using the logit link function:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \; ; \; \beta \; is \; coefficient$$

After fitting the glm() function on the training data and predicting the log(odds) of the Y variable, I converted it to prediction probability scores. I then determined the optimal prediction probability cutoff to tune

the model so that it can improve the accuracy. I used "InformationValue::optimalCutoff" function to achieve this. The optimal cutoff value was 0.4791532. If the predicted probability is greater than this cutoff value, it was coded as 1 and 0 otherwise.

B. K-Nearest Neighbor Model

K-nearest neighbor (KNN) model is a pattern-recognition model that classifies the unknown class label (e.g. unknown y_i of a new data point v) based on its distance to the known class labels (e.g. $\{(x_1,y_1),\ldots,(x_n,y_n)$. For each new data point v, the KNN algorithm determines the k nearest neighbors and their labels from that data point and assign its class label based on majority vote. I used the default Euclidean distance [formula = $(\sqrt{}((X_1-X_2)^2+(Y_1-Y_2)^2)]$. Plus, to select the best number of k neighbors and kernel for the model, I applied the leave-one-out cross validation with different kernels (i.e. optimal, rectangular, inv, gaussian, and triangular) technique by using the "knn.cv()" function.

C. Decision Tree

Decision Tree, also known as Classification and Regression Tree, is a robust algorithm that repeatedly partitions the data into a number of smaller groups with similar response values based on a set of decision rules. Because my response variable is categorical (i.e. level of suicide risk), the model predicts the class label Y_i from the class that has majority presentation in the subgroup. The model tries to minimize cross-entropy or Gini index, a measure of purity to ensure the subset is as pure or homogeneous as possible. Plus, I tuned the parameters of the model with 10-fold cross validation and tried different cutoff values to find the best model with good values of accuracy, sensitivity and specificity.

3. Results and Evaluation

3.1 Observation

According to table 3, many categories of demographics and health-risk behaviors varied significantly by level of suicide risk. For example, during the 12 months before the survey, 45.2% of female students and 54.8% of male students reported having no suicidal thoughts or attempts and were considered to have low suicide risk; 61.7% of female students and 38.3% of male students reported at least one incident of suicidal thoughts or attempts and were considered at high risk for suicide. Sexual identity was also statistically significant. Compared to heterosexual-identified peers, sexual-minority and unsure students reported more incidents of suicidal thoughts and attempts (21.6% and 7.0%, respectively) than the ones in low risk (4.6% and 3.3%). There was no significant difference in the level of suicide risk among different races as well as among different grade level.

Moreover, the prevalence of violence-related behaviors (both community and school), feeling hopelessness, substance use, risky sexual behavior, being obese or perceiving overweight, inactive lifestyle and low academic performance (p values range from <0.05 to <0.001) were all significantly greater among students at high risk for suicide compared with students at low risk for suicide.

3.2 Performance Evaluation

Accuracy, sensitivity and specificity were used to evaluate the performance of each model after it fitted the test data set to predict level of suicide risk. Accuracy measures how often the model classifies correctly; its formula is *(True Positive + True Negative)/Total*. The sensitivity measures how accurately the model classifies actual event, which is the high suicide risk level in the current project; its formula is *True Positive/ (True Positive + False Negative)*. The specificity measures how accurately the model classifies non-actual event, which is the low suicide risk level; its formula is *True Negative/ (True Negative + False Positive)*. A good model should try to maximize these performance indicators. Table 4 displays a summary of these performance indicators for my machine learning models.

For the logistic regression model, the accuracy rate was 80.73% with 84.24% sensitivity and 61.40% specificity. This model also showed that sexual identity, bullied electronically, forced to have sex, in a physically fight at school, bullied at school, threatened at school, feel sad/hopelessness, perceived to be overweight, and sleep less than 8 hours/ day are significant factors (p value < 0.05) that increase risk of suicide.

For the KNN algorithm, the model had k=66 and a triangular kernel (see figure 2 for the different kernel performances). However, it performed slightly worse than the logistic regression with 78.44% accuracy, 79.46% sensitivity and 60.98% specificity.

The decision tree model, before adjusting for cutoff value (default 0.5 cutoff), had an accuracy rate of 79.1% with very low 31.4% sensitivity and very high 93.2% specificity. Because predicting correctly a high level of suicide risk is more critical than predicting correctly a low level of suicide risk, I needed to increase the sensitivity of the decision model. Hence, I applied different cutoff values (see table 5). The desirable results were at cutoff 0.4 with 77.80% accuracy, 68% sensitivity and 80.6% specificity. Furthermore, the binary tree of my model (see figure 3) showed that feeling hopelessness and not heterosexual identified (i.e. sexual minority and unsure) group was more likely to have high suicide risk. For heterosexual group, being bullied electronically was also considered at high-risk level for suicide.

4. Conclusion

My current project attempted to examine adolescent suicide risk in California and apply three different types of machine learning models to predict the level of suicide risk. Although the performances of the models

are about 70%-80% range for the indicators, they can be improved and applied to other year(s) or state(s) of the YRBS data. Some of the ideas I have for improvement of predicting the suicide risk include variable selection, carefully select and tune certain parameters, and potentially build other machine learning models (e.g. SVM, neural network).

Reference

[1] Annor, F. B., Clayton, H. B., Gilbert, L. K., Ivey-Stephenson, A. Z., Irving, S. M., David-Ferdon, C., & Kann, L. K. (2018). Sexual orientation discordance and nonfatal suicidal behaviors in US high school students. American journal of preventive medicine, 54(4), 530-538.

[2] Carson, N. J., Mullin, B., Sanchez, M. J., Lu, F., Yang, K., Menezes, M., & Cook, B. L. (2019). Identification of suicidal behavior among psychiatrically hospitalized adolescents using natural language processing and machine learning of electronic health records. PloS one, 14(2), e0211116.

[3] Choi, S., Lee, W., Yoon, J., Won, J., & Kim, D. (2018). Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. Journal of Affective Disorders, 231, 8-14.

[4] Eaton, D. K., Foti, K., Brener, N. D., Crosby, A. E., Flores, G., & Kann, L. (2011). Associations between risk behaviors and suicidal ideation and suicide attempts: do racial/ethnic variations in associations account for increased risk of suicidal behaviors among Hispanic/Latina 9th-to 12th-grade female students?. Archives of suicide research, 15(2), 113-126.

[5] Hill, R., Oosterhoff, B., & Do, C. (2019). Using Machine Learning to Identify Suicide Risk: A Classification Tree Approach to Prospectively Identify Adolescent Suicide Attempters. Archives of Suicide Research : Official Journal of the International Academy for Suicide Research, 1-33.

[6] Lowry, R., Crosby, A., Brener, N., & Kann, L. (2014). Suicidal Thoughts and Attempts Among U.S. High School Students: Trends and Associated Health-Risk Behaviors, 1991–2011. Journal of Adolescent Health, 54(1), 100-108.

[7] van Buuren, S., Boshuizen, H.C., Knook, D.L. (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18, 681--694.

[8] Youth Risk Behavior Survey—Data Summary & Trends Report: 2007–2017. Centers for Disease Control, 2018. https://www.cdc.gov/healthyyouth/data/yrbs/pdf/trendsreport.pdf

[9] Youth Risk Behavior Survey—Software for Analysis of YRBS Data. Centers for Disease Control, 2017.https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2017/2017_YRBS_analysis_software.pdf

[10] Zheng, Z., & Ruggiero, K. (2017, November). Using machine learning to predict obesity in high school students. In 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2132-2138). IEEE.

Appendix

Table 1. Participants' demographic characteristics

Characteristic (N=3707)		n*	%
Gender	Female	1861	50.2%
	Male	1820	49.1%
Race	White	687	18.5%
	Black	188	5.1%
	Hispanic	1890	51.0%
	All other races	849	22.9%
Grade	9th	1108	29.9%
	10th	947	25.5%
	11th	885	23.9%
	12th	739	19.9%
Sexual identify	Heterosexual	3179	85.8%
	Sexual minority †	317	8.6%
	Unsure	166	4.5%
Suicide risk	High	845	22.8%
	Low	2862	77.2%

**Some categories include missing data*

†Sexual minority includes students who responded that they were gay, lesbian, or bisexual

Table 2. Health-risk behaviors

Health-risk behaviors	Item	Question	Analytic coding
Community-related violence			
In a physical fight	# qn17	During the past 12 months, how many times were you in a physical fight?	coded 1 if >= 1 time; 0 otherwise
Bullied electronically	# qn24	During the past 12 months, have you ever been electronically bullied?	coded 1 if yes; 0 otherwise
Forced to have sex	# qn19	Have you ever been physically forced to have sexual intercourse when you did not want to?	coded 1 if yes; 0 otherwise
School-related violence			
Carried weapon at school	# qn13	During the past 30 days, on how many days did you carry a weapon such as a gun, knife, or club on school property?	coded 1 if >= 1 day; 0 otherwise
In a physical fight at school	# qn18	During the past 12 months, how many times were you in a physical fight on school property?	coded 1 if >= 1 time; 0 otherwise
Bullied at school	# qn23	During the past 12 months, have you ever been bullied on school property?	coded 1 if yes; 0 otherwise
Missed school because unsafe	# qn15	During the past 30 days, on how many days did you not go to school because you felt you would be unsafe at school or on your way to or from school?	coded 1 if >= 1 day; 0 otherwise

Threatened at school	# qn16	During the past 12 months, how many times has someone threatened or injured you with a weapon such as a gun, knife, or club on school property?	coded 1 if >= 1 time; 0 otherwise
Mental health issue			
Feel sad/hopeless	# qn25	During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities?	coded 1 if yes; 0 otherwise
Substance use			
Current alcohol use	# qn42	During the past 30 days, on how many days did you have at least one drink of alcohol?	coded 1 if >= 1 day; 0 otherwise
Current marijuana use	# qn48	During the past 30 days, how many times did you use marijuana?	coded 1 if >= 1 time; 0 otherwise
Ever used methamphetamine	# qn52	During your life, how many times have you used methamphetamines (also called speed, crystal, crank, or ice)?	coded 1 if >= 1 time; 0 otherwise
Ever injected drugs	# qn57	During your life, how many times have you used a needle to inject any illegal drug into your body?	coded 1 if >= 1 time; 0 otherwise
Sexual health			
Had sex before age 13	# qn60	How old were you when you had sexual intercourse for the first time?	coded 1 if had sex before 13 years old; 0 otherwise
Currently sexually active	# qn62	During the past 3 months, with how many people did you have sexual intercourse?	coded 1 if >= 1 person; 0 otherwise
Weight-related issue			
Obese	# qnobese	Based on BMI, whethere the student is obese or not	coded 1 if BMI>=95th percentile; 0 otherwise
Percived to be overweight	# qn68	How do you describe your weight?	coded 1 if describing overweight; 0 otherwise
Sedentary activities			
Watched TV >= 3 hours/day	# qn80	On an average school day, how many hours do you watch TV? (1 if more than 3 hours/day)	coded 1 if >= 3 hours/day; 0 otherwise
Computer/video game >= 3 hours/day	# qn81	On an average school day, how many hours do you play video or computer games or use a computer for something that is not school work? (1 if more than 3 hours/day)	coded 1 if >= 3 hours/day; 0 otherwise
Not physically active	# qn79	During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day?	reverse coded 1 if < 5 days)
Did not play in sport team	# qn83	During the past 12 months, on how many sports teams did you play?	reverse coded 1 if 0 team; 0 otherwise)
Slept < 8 hours/night	# qn88	On an average school night, how many hours of sleep do you get?	reverse coded 1 if < 8 hours/night; 0 otherwise)
Academic performance			
Low grade	# qn89	During the past 12 months, how would you describe your grades in school?	reverse coded 1 if less than B grade; 0 otherwise

Table 3. Prevalence of demographics and health-risk behaviors by level of suicide risk

		Low suicide risk		High suicide risk				
		%	95% CI	%	95% CI	F	df	p value*
Gender								
	Female	45.2	(41-49)	61.7	(57-66)	42.464	16	<0.001
	Male	54.8	(50.1-59)	38.3	(33.9-43)			
Race								
	White	25.8	(19-34)	27.1	(20.4-35)	0.283	44	0.823
	Black/ African American	4.4	(2.8-7)	4.4	(2.8-7)			
	Latino/ Hispanic	51.1	(42-60)	51.2	(42.6-60)			
	All other races	18.7	(14.6-24)	17.3	(12.5-23)			
Grade								
	9th	25.9	(19.1-34)	28.5	(21.3-37)	0.808	40	0.476
	10th	26.1	(18.5-35)	25.9	(19.6-33)			
	11th	24.2	(19-30)	24.0	(18.5-30)			
	12th	23.8	(18.6-30)	21.6	(15.6-29)			
Sexual identify								
	Heterosexual	92.1	(90.1-93)	71.4	(66.4-76)	75.397	31	<0.001
	Sexual minority	4.6	(3.8-6)	21.6	(18-26)			
	Unsure	3.3	(2.5-4)	7.0	(4-12)			
Community-related violence								
	In a physical fight	14.5	(12.5-17)	24.8	(20.5-30)	24.576	16	<0.001
	Bullied electronically	9.0	(7.6-10)	29.2	(23.9-35)	186.180	16	<0.001
	Forced to have sex	4.4	(3.6-5)	13.7	(10.4-18)	66.935	16	<0.001
School-related violence								
	Carried weapon at school	2.6	(2-3)	7.4	(5.1-10)	40.767	16	<0.001
	In a physical fight at school	5.6	(4.5-7)	8.1	(5.6-11)	5.288	16	0.035
	Bullied at school	13.2	(11.2-15)	35.2	(31.1-39)	239.180	16	<0.001
	Missed school because unsafe	3.8	(2.9-5)	13.5	(9.9-18)	73.834	16	<0.001
	Threatened at school	3.1	(2.3-4)	11.5	(9-15)	60.909	16	<0.001
Mental health issue								
	Feel sad/hopeless	19.7	(17.7-22)	69.8	(65.6-74)	407.980	16	<0.001
Substance use								
	Current alcohol use	27.0	(23-31)	38.3	(34.1-43)	25.713	16	<0.001
	Current marijuana use	19.9	(17.1-23)	30.6	(26.8-35)	24.742	16	<0.001
	Ever used methamphetamine	1.7	(1.4-2)	5.7	(4-8)	41.204	16	<0.001
	Ever injected drugs	1	(0.6-2)	3.5	(2.3-5)	25.059	16	<0.001
Sexual health								
	Had sex before age 13	2	(1.4-3)	5	(2.9-8)	9.654	16	0.007
	Currently sexually active	2.2	(18.8-25)	28.4	(25-32)	9.007	16	0.008
Weight-related issue								
	Obese	13.1	(10.5-16)	17	(14-20)	7.630	16	0.014
	Percived to be overweight	29.5	(26.6-33)	43	(39.1-47)	54.700	16	<0.001

Sedentary activities							
Watched TV >= 3 hours/day	20.1	(17.6-23)	22.9	(19.9-26)	3.202	16	0.093
Computer/video game >= 3 hours/day	41.2	(36.3-46)	52.9	(48.9-57)	35.011	16	<0.001
Not physically active	46.7	(43-50)	61.7	(58-65)	44.176	16	<0.001
Did not play in sport team	38.5	(34.3-43)	53.3	(48-59)	53.922	16	<0.001
Slept < 8 hours/night	69	(65.8-72)	81.6	(78.6-84)	74.077	16	<0.001
Academic performance							
Low grade	26.9	(23-31)	38.5	(32.4-45)	30.332	16	<0.001

df = degrees of freedom.

**p<0.05 is considered statistically significant*

Table 4. Performance indicators of each machine learning models

Model	Accuracy	Sensitivity	Specificity
Logistic regression	80.73%	84.24%	61.40%
KNN	78.44%	79.46%	60.98%
Decision tree	77.80%	68.00%	80.60%

Table 5. Performance indicators of the decision tree by different cutoff values

cutoff	accuracy	sensitivity	specificity
0.228	1.000	0.000	0.000
0.050	0.228	1.000	0.000
0.100	0.228	1.000	0.000
0.150	0.778	0.680	0.806
0.200	0.778	0.680	0.806
0.250	0.778	0.680	0.806
0.300	0.778	0.680	0.806
0.350	0.778	0.680	0.806
0.400	0.778	0.680	0.806
0.450	0.791	0.314	0.932
0.500	0.791	0.314	0.932
0.550	0.791	0.314	0.932
0.600	0.791	0.314	0.932
0.650	0.790	0.178	0.970
0.700	0.790	0.178	0.970
0.750	0.772	0.000	1.000
0.800	0.772	0.000	1.000
0.850	0.772	0.000	1.000
0.900	0.772	0.000	1.000
0.950	0.772	0.000	1.000
1.000	0.772	0.000	1.000

Figure 1. Correlation matrix

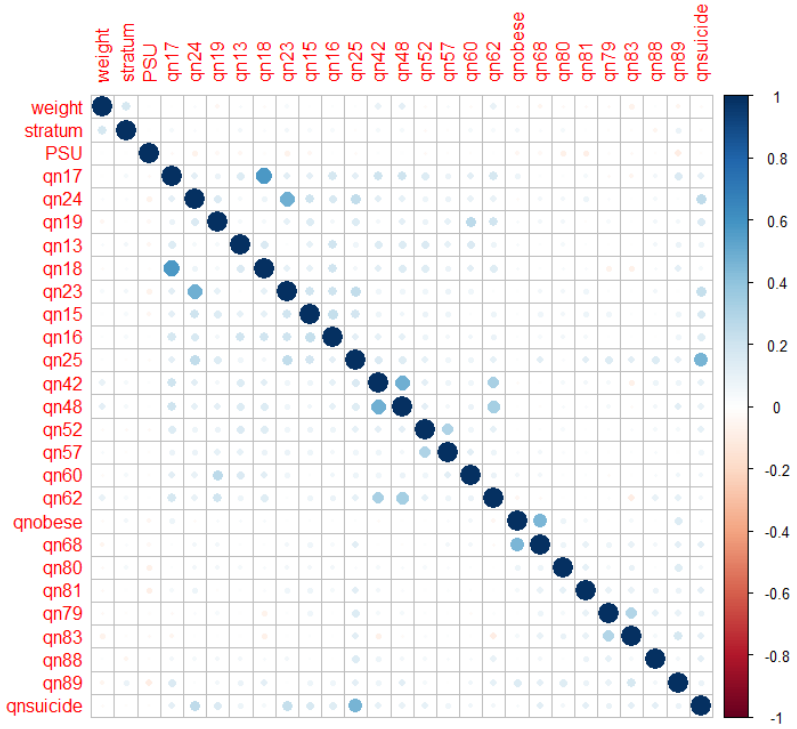

Figure 2. Mean square error and k value by different kernel types.

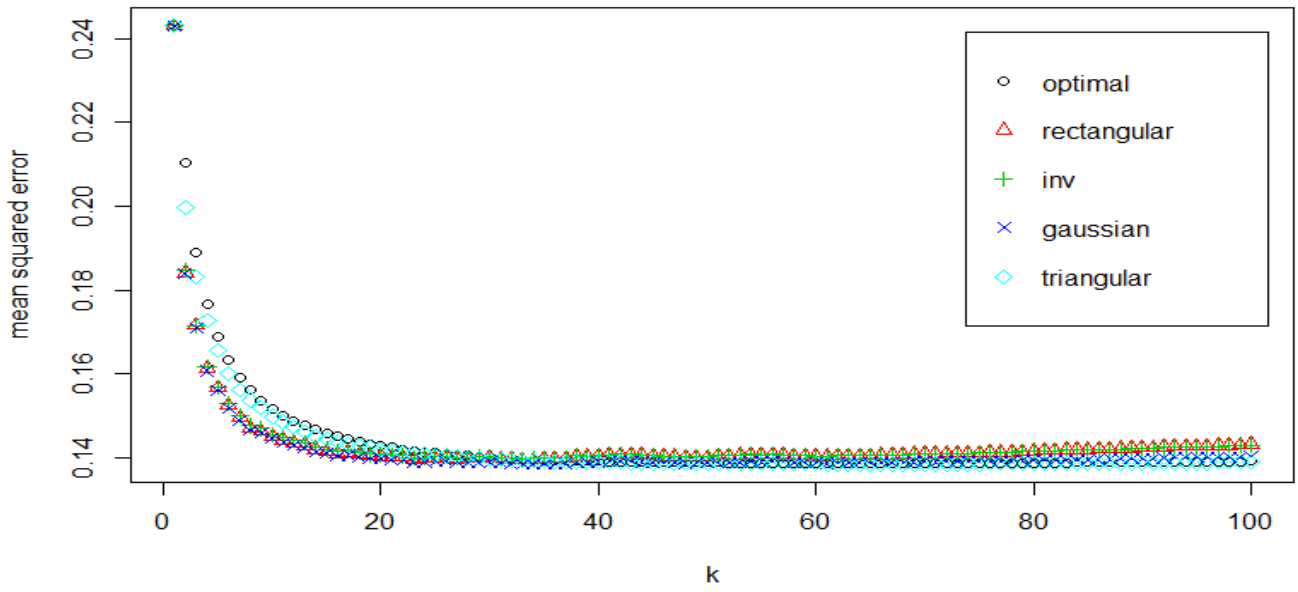

Figure 3. Binary tree