

Classification and Prediction of Suicidal Tendencies of the Youth in the Philippines: An Empirical Study

Stephen Jun Villejo

School of Statistics

University of the Philippines Diliman

This paper investigates suicidal tendencies of youth in the Philippines based on the Young Adult Fertility and Sexuality Study (YAFS) 2002. The main goal of the paper is the classification and prediction of suicidal tendencies using classification algorithms. The different classification algorithms such as Classification and Regression Trees, random forests and conditional inference trees; and the logistic regression have consistent findings on the significant variables affecting suicidal tendencies. Due to the severely unbalanced classes of the response variable, the classification models have very poor predictive ability for the minority class although the overall classification rate is high. A classification algorithm is proposed which improves the predictive ability in terms of balancing out the correct classification in the two classes of the response variable.

Keywords: classification, suicide, prediction, logistic regression

1. Introduction

According to the World Health Organization (2014), an estimated 804,000 suicide deaths occurred worldwide in 2013. This figure could even be possibly underreported since suicide is a sensitive issue and illegal in some countries. Other notable findings are as follows: In richer countries, three times as many men die of suicide than women do. Around 75% of suicides occur in low-and middle-income countries. In almost all regions of the world, suicide rates are highest in persons aged 70 years or over for both men and women. For some countries, suicide rates are highest among the young, and is the second leading cause of death in 15-29 years old. Suicide in the South-East Asian region is the highest as compared to other regions where 39% of those in low- and middle-income countries in the region comprises the total suicides in the world. Suicide has its peak among the young and elderly. Globally, males have higher age-standardized suicide rate per

100,000 compared to females. It applies for all high-income countries, and low- and middle-income countries. It's only in the Western Pacific that the suicide rate is higher for females than males.

Suicidal behavior is very complicated. A lot of factors affect suicidal behavior which are grouped into four key main areas: health systems, society, community, relationships, and individual (World Health Organization, 2014). The societal factors are access to means, inappropriate media reporting and stigma associated with help-seeking behavior. Community factors are disasters, wars and conflicts, stresses of acculturation and dislocation, discriminations, and trauma or abuse. Relational factors are sense of isolation and lack of social support. Individual factors are previous suicide attempt, mental disorders, and harmful use of alcohol, job or financial loss, hopelessness, chronic pain, family history of suicide and genetic and biological factors. (World Health Organization, 2014)

The World Health Organization stresses the fact that suicide is a global phenomenon affecting all countries but it is also preventable. The effort to mitigate suicidal risks is multi-sectoral and demands cooperation and collaboration of the government, the communities, the media, and many more institutions.

Objectives

The main objective of this study is the identification of the variables which will explain suicidal tendencies of the youth in the Philippines. In particular, the response variable to be used is whether or not a person has ever thought of committing suicide. The next step is to find a classification model with an acceptable predictive power.

2. Review of Related Literature

It has been noted by Cho et al. (2013) that physical illnesses and disorders are correlated with suicide risk thus, the importance of access to health care in reducing the risk. In particular, mental disorders is present in up to 90% in high-income countries (Cavanagh et al., 2003). People with more than one mental disorder have higher risks of suicide.

Relationships with other people is a significant factor affecting suicide behavior. Young people who have experienced childhood and family abuse like physical violence, sexual or emotional abuse, and maltreatment among others have high risks of committing suicide (Johnson et al., 2002). Traumas arising from disciplinary or legal crises, financial problems, academic or work-related problems, and bullying are also significant factors affecting suicide (Foster, 2011).

People who lack relational and social support compounded with other factors increases the risk of suicide (Casiano et al., 2013). Conflicts like separation and losses like death of a partner or any loved one can cause psychological stress and grief which causes an increase of suicide risk. (Kposowa, 2003). Those who have family history of suicide have increased risk of suicide or mental disorder (Qin

et al., 2002). Also, chronic suicide thoughts is a major factor affecting suicide. Previous suicide attempts is a strong indicator for future suicides (Beautrais, 2003). Moreover, consumption of alcohol increases risk of suicide. Of all deaths from suicide, 22% was attributed to the use of alcohol according to the World Health Organization's report on alcohol and health (2014). Job loss and financial uncertainty is associated with depression, anxiety, violence and the harmful use of alcohol which increases the risk of suicide (Chang et al., 2013).

Hopelessness which is associated with thoughts such as "things will never get better" and "I do not see things improving" is associated with depression (Beck et al., 1990) which when coupled with mental disorders and prior suicide attempts could lead to increase of suicidal risk. (O'Connor and Nock, 2014).

According to the World Health Organization (2014), the following factors have been proven to increase resilience against suicidal behavior: strong personal relationships, religious or spiritual beliefs, and lifestyle practice of positive coping strategies and well-being. For the religious beliefs, it should be noted that while it can lessen suicidal tendencies, it is not universal and it is dependent on the specific cultural and contextual practices and interpretations. Resilience and the ability to cope up protects against suicide, good self-esteem, self-efficacy and effective problem-solving skills can lessen the impact of emotional stressors. (Luther, 2003). Subjective psychological well-being also helps mitigate suicide risk (Sisasketal, 2008).

Attempted suicide is a dangerous event, a risk factor for completed suicide, and an indicator of other health problems such as substance abuse, depression, or adjustment and stress reactions. Garland and Zigler (1993) have estimated that the number of suicide attempts is 50 to 200 times the number of completed suicides.

Studies have shown that females are more likely to attempt suicide than males. According to Garland and Zigler (1993) and Gelmar (1994), females attempt suicide three to four times as often as males. Also, it is estimated that males complete one out of three attempts while females complete one out of every 25 suicide attempts. (Rosenberg et al., 1990).

Cantor (1989) and Deykin and Buka (1994) cited the following factors related to the rising rate of adolescent suicide: a high level of social and academic competition and pressure, violence that children and adolescents are exposed to, lack of socially ways for youngsters to express anger, lack of connection to religion, abuse of drugs and alcohols, social isolation, occurrence of depressive illness, and greater availability of handguns among others.

Cebula and Zelenskaya (2006) studied the determinants of youth (age 15-24) suicide over the 1960-1998 period. It was found out that the rate of youth suicide was positively influenced by the divorce rate, the violent crime rate, and the Vietnam war.

Tonooka (1999) examined a person's attitude towards suicide when faced with an incurable disease, bankruptcy, or family dishonor, or when he/she is

tired of living. It was found out that religion, race, and education had the highest levels of statistical significance. In particular, religion and suicide approval has a strong negative relationship. In terms of income, the richer people tend to support suicide. Males have slightly higher tendency to support suicide. Religious services attendance has a negative effect on the attitude toward suicide. People who live in the rural areas tend to oppose suicide more than people in the city. People who have higher degree tends to support suicide.

Redaniel et al. (2011) has pointed out that even if suicide rates are low in the Philippines, suicide incidence among adolescents and young adults increases which implies that there should be a focused suicide prevention programs. Moreover, more women than men attempt suicide in the Philippines which is higher than China or India but is comparable to Thailand, Japan, and New Zealand.

3. Data

The data used in the study was the Young Adult Fertility and Sexuality Study (YAFS) conducted by the University of the Philippines Population Institute (UPPI) and the Demographic Research and Development Foundation, Inc. (DRDF). The goal of the national study is to understand the knowledge, attitudes, and behavior of young Filipinos in a wide range of sex and sex-related issues together with some significant events and institutions affecting their behaviors.

The response variable is whether the respondent has ever thought of committing suicide or not. The independent variables with their codes are listed below:

Table 1. Table on Variable Description

Label	Variable
e2	Believes in God or a Supreme being
e3	Believes that doing good to others is necessary for salvation
e4	Believes in eternal life
e5	Believes in heaven and hell
e6a	Feeling of being of worth
e6b	Feeling of having a number of good qualities
e6c	Feeling of proud for one's self
e6e	Satisfaction with self on the whole
e6f	Taking a positive attitude towards self
e6g	Feeling that many things he/she does are not so meaningless
e6i	Feeling of being no good at all
g53	Any friends/acquaintance attempted to commit suicide
g54a	During the past 3 months, respondent has felt sad or helpless that made him/her stop doing usual activities
g54b	During the past 3 months, respondent has felt really hopeless about the future

g54c	During the past 3 months, respondent felt depressed about life in general.
g55	Ever felt unsafe or threatened during the past 3 months
g56	Every experienced being threatened by someone
g57	Every experienced being physically injured by someone during the past 3 months
g59	Ever carried a weapon
h1	Ever had a barkada
h3	Have any barkada at present
j10	Ever had crush
sex	Sex of respondent

Variables e2-e5 are binary variables with values either “Yes” or “No.” The series of e6 variables is measured in a 5-point Likert scale from “Strongly Disagree” to “Strongly Agree”. The variables g49, g53, g55 to g59, and h1, h3, and j10 are also binary variables of either “Yes” or “No.” The series of g54 variables takes values of “Never,” “Sometimes,” “Most of the time,” and “Always.”

4. Methodology

4.1 Classification and regression trees

Classification and regression trees (CART) is a nonparametric statistical method and a supervised learning technique done through an algorithm known as recursive partitioning where a decision tree is constructed by splitting on each node into two daughter nodes. (Izenman, 2008). The binary split is determined by a Boolean condition on the value of a single variable. Observations satisfying the condition for the variable drop down to one of the two daughter nodes while the remaining observations not satisfying the condition drop down to the other daughter node.

A measure of goodness of a split measured by *anode impurity function* will be the basis in choosing the best split over all variables. If we let $\Pi_1, \Pi_2, \dots, \Pi_k$ be the $K \geq 2$ classes and let $p(k|\tau)$ an estimate of $P(X \in \Pi_k | \tau)$, the conditional probability that an observation X is in Π_k given that it falls into node τ . One such function is the *entropy function* defined by

$$i(\tau) = \sum_{k=1}^K p(k | \tau) \log[p(k | \tau)], \quad (1)$$

which reduces to a function of the following form when there are only two classes:

$$i(\tau) = -p \log(p) - (1-p) \log(1-p) \quad (2)$$

Another function which is also the usual default in statistical software is the *Gini diversity index*,

$$i(\tau) = \sum_{k \neq k'} p(k | \tau) p(k' | \tau) = 1 - \sum_k \{p(k | \tau)\}^2 \quad (3)$$

which reduces to $i(\tau) = 2p(1 - p)$ if there are only two classes. The two measures do not significantly differ for practical purposes.

The class associated with a terminal node corresponds to the class where the majority of the observations belongs. This is called the *plurality rule*.

Let T be the tree classifier and let $\tilde{T} = \{\tau_1, \tau_2, \dots, \tau_L\}$ the set of all terminal nodes. The true misclassification rate is estimated by,

$$R(T) = \sum_{\tau \in \tilde{T}} R(\tau) P(\tau) \quad (4)$$

where $p(\tau)$ is the probability that an observation falls into node τ . If we estimate $p(\tau_l)$ by the proportion of all observations that fall into node τ_l , then the resubstitution estimate or $R(T)$ is

$$R^{re}(T) = \sum_{l=1}^L r(\tau_l) p(\tau_l). \quad (5)$$

The resubstitution estimate of the error has its drawbacks for the following reasons: bigger trees have smaller values of the error estimate, and using it as the basis in coming up with the final tree will generate very large trees. In choosing the decision tree, the approach recommended by Breiman et al. (1984) was to let the tree grow to saturation and then prune it until the tree is the right size. The pruning algorithm is to grow a saturated tree, then compute an estimate of $R(\tau)$ at each node, then prune the tree upwards toward its root node such that $R(T)$ is minimized. A regularization approach will be adopted instead of using $R^{re}(T)$ as the estimate or $R(T)$. A complexity approach will be introduced such that

$$R_\alpha(T) = \sum_{l=1}^L R^{re}(T) + \alpha L. \quad (6)$$

The term αL is a penalty term for tree size so that generating a very large tree will have a higher penalty than for smaller trees. An optimally-pruned sub tree is the tree whose $R_\alpha(T)$ is minimized.

According to Breiman et al. (1984), use of an independent test set is preferred when the data set is large while cross-validation is preferred for smaller data sets.

4.2 Random forests

The random forests technique examines a large ensemble of decision trees, by first generating a random sample of the original data with replacement. A decision tree will then be constructed based on the subset of the original data where in which arandomsample of variables will be generated to determine node splitting.

The resampling will be performed with a predetermined number of times. From the multiple subsets of trees that will be built, the support for the role of each variable in each decision is noted. Izenman (2008) noted that as the number of trees is increased, the generalization error converges to a limit.

Random forests gives the importance of each variable in the classification process based on the mean decrease in Gini and mean decrease in accuracy. A variable is deemed more important if the accuracy of the random forest decreases in the addition of the variable. Therefore, variables with a large mean decrease in accuracy are more important. The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes. Also, the Gini coefficient is a measure on node impurity. Hence, variables that results in nodes with higher purity have a higher decrease in Gini coefficient and are therefore of greater importance.

4.3 Conditional inference tree

Conditional inference trees estimate a regression relationship by binary recursive partitioning similar to how classification and regression trees are constructed but in a conditional inference framework. The algorithm works as follows. (1) The first step is to test the global null hypothesis that the response variable and the input variables are independent. (2) If the null hypothesis is rejected, the input variable with the strongest association to the response variable will be selected. Next, a binary split will be implemented in the selected input variables. Steps (1) and (2) will be recursively repeated.

The implementation of the algorithm was developed by Strasser and Weber (1999). The stopping criterion in step (1) is either based on multiplicity adjusted p -values or on the univariate p -values. A split is implemented using some criterion which assures that the right sized tree is grown and no form of pruning or cross-validation is needed. Moreover, the selection of the input variable to split in is based on univariate p -values avoiding a variable selection bias towards input variables with many possible cutpoints.

4.4 Binary Logistic Regression

Binary logistic regression estimates the probability that a characteristic is present given the values of explanatory variables. The response variable is a binary variable, either a characteristic is present or absent. The explanatory variables can be a combination of discrete and continuous variables. The model is given below:

$$\text{Given that } \pi_i = P(Y_i = 1 | X_{i1} = x_{i1}, X_{i2} = x_{i2}, \dots, X_{ip} = x_{ip}), \quad (7)$$

$$\pi_i = \frac{\exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}\}} \quad (8)$$

or,

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} \quad (9)$$

The model does not assume linear relationship between the dependent variable and the independent variables, but it does assume linear relationship between the logit of the response and the explanatory variables. The estimation of the parameters is done using maximum likelihood estimation rather than ordinary least squares.

The interpretation of the parameters is as follows. The β_j is the average increase in the logit(π_i) for every unit increase in x_i . The term $\exp(\beta_j)$, $j = 1, \dots, p$ is the multiplicative change in the odds for every unit increase in x_{ij} . The term $\exp(\beta_0)$ is the odds that the characteristic is present when all $x_{ij} = 0$, $j = 1, \dots, p$. The odds is define as the probability of success divided by the probability of failure

or $\frac{\pi_i}{1-\pi_i}$.

5. Results and Discussion

5.1 Logistic regression

The binary response variable is whether a person has ever thought of committing suicide or not. None of the variables related with religious conviction or spirituality was significant in the logistic model. It should be noted that 99.7% of all the respondents believe in God or a Supreme Being. Only 6.8% does not believe that doing good to others is not necessary for salvation. Also, 12% does not believe in eternal life. Lastly, 94.1% believe in heaven and hell. Only the 4 variables previously mentioned which are related with religiosity were not significant.

The categories of variables measured in a 5-point Likert-scale was collapsed into three categories: Disagree, Agree, and Neither. In the logistic model, the baseline for the binary variables is "No." For the variables measuring agreeability, the baseline category is "Agree." For the variables measuring frequency (Never, Sometimes, Most of the Time, Always), the baseline is "Always."

The Hosmer and Lemeshow goodness of fit test p-value is 0.926. Thus, the logistic model is of good-fit. Table 1 gives the multiplicative change in the odds and their significance.

There is no significant change in the odds of committing suicide between those who agree and disagree that they have a number of good qualities. However, if a person will neither agree nor disagree, the odds of thinking about committing suicide will be 1.137 times the odds if a person will agree. Those who agree that they have a number of good qualities have a lower chance of thinking about committing suicide. Also, those who think of positive attitude towards self will have a lower

Table 2. Table of Odds Ratios and Their Significance

Variables	Sig.	Exp(B)	Variables	Sig.	Exp(B)
e6b(Agree)	0.107		g54a(Sometimes)	0.0000	0.657
e6b(Disagree)	0.690	1.042	g54a(Most of the time)	0.5667	1.066
e6b(Neither)	0.035	1.137	g54b	0.0001	
e6c	0.036		g54b(Never)	0.0002	0.630
e6c(Disagree)	0.496	1.0367	g54b(Sometimes)	0.0291	0.772
e6c(Neither)	0.012	1.164	g54b(Most of the time)	0.1337	0.818
e6e	0.197		g54c	0.0000	
e6e(Disagree)	0.837	0.981	g54c(Never)	0.0000	0.590
e6e(Neither)	0.081	1.114	g54c(Sometimes)	0.0849	0.816
e6f	0.000		g54c(Most of the time)	0.4200	1.112
e6f(Disagree)	0.000	1.240	g55(Yes)	0.0001	1.262
e6f(Neither)	0.000	1.364	g56(Yes)	0.0000	1.450
e6i	0.000		g57(Yes)	0.0006	1.248
e6i(Disagree)	0.000	0.679	g59(Yes)	0.0000	1.399
e6i(Neither)	0.000	0.700	h3(Yes)	0.0004	0.792
g53(Yes)	0.000	1.929	j10(Yes)	0.0006	1.894
g54a	0.000		sex(Female)	0.0000	2.672
g54a(Never)	0.000	0.544	Constant	0.0000	0.119

chance of thinking about suicidal thoughts. The difference is significant against those who neither agree nor disagree of thinking positive thoughts with self where the increase in odds of having suicidal thoughts is 11.4% the original odds. Those who feel proud of themselves and are satisfied with themselves on the whole also have lower chances of thinking about committing suicide than those who neither agree nor disagree that they are proud of their selves and those who neither agree nor disagree that they are satisfied with their selves. The increase in the odds of suicidal thoughts for the two variables when they neither agree nor disagree is 11.4% and 36.4% the original odds, respectively. If a person is not satisfied with one's self, the odds of thinking about committing suicide will increase by 24% the original odds than if the person is satisfied. If a person disagrees that sometimes he/she thinks that he/she is no good at all, the odds of having suicidal thoughts will decline by 32.1% than the original odds if the person agrees that he thinks he is no good at all. If the person neither agrees nor disagrees, the odds will also decline but by 30% of the original odds. This implies that those who think that they are no good at all have higher chances of thinking about committing suicide.

If a person has friends or acquaintances who committed suicide, the odds that he/she will think about committing suicide as well will increase by 92.9% of the original odds that if a person has no friends or acquaintances who committed suicide.

Feelings of depression, sadness or helplessness, and hopelessness about the future are also significant indicators of having thoughts of committing suicide. If a person has never felt being sad or helpless that made him/her stop doing usual activities during the past 3 months, the odds of thinking about committing suicide will be 0.544 times the odds if the person was sad or helpless during the past 3 months. The increase in the odds of suicidal thoughts is 83.82% if the person is sad or helpless than otherwise. If the person was sad or helpless only for sometimes, then the chance of having suicidal thoughts is still lower than if the person was always sad or helpless. The decline in the odds of committing suicide for those who felt sad or helpless sometimes is 34.3% that for those who are always sad or helpless. There is no significant difference between those who are always sad or helpless and those who are sad or helpless most of the time. Furthermore, if a person has never felt being hopeless about the future, the odds of having suicidal thoughts will 0.630 times the odds if a person has always felt such. In other words, those who felt being hopeless about the future will have higher chances of thinking about committing suicide. Lastly, those who felt being depressed about life in general have higher chances of thinking of committing suicide. The odds of thinking about committing suicide for those we never felt depressed is 0.59 times the odds for those who always felt being depressed about life in general.

Those who felt unsafe or threatened during the past 3 months have higher chances of thinking about committing suicide than those who do not. The odds of thinking about committing suicide for those who felt unsafe or threatened is 1.26 times the odds for those who do not. Those who experienced being threatened by someone and being physically injured by someone also have higher chances of thinking about committing suicide than those who do not. In particular, the odds of committing suicide is 45% and 24.8% higher than original odds than for those who do not experience being threatened or being physically injured, respectively.

Those who have carried a weapon (balisong, gun, etc.) has higher chances of thinking about committing suicide than those who have never carried a weapon. The increase in the odds for those who have carried a weapon is 39.9% the odds for those have not carried a weapon.

If a person has a barkada at present, the odds of having thoughts of committing suicide will be 0.792 times the odds of those who do not have barkada at present. In other words, those who have barkada has lower chances of having thoughts of committing suicide than those who have no barkada. Moreover, if a person ever had a crush, the odds of thinking about committing suicide is higher than those who have not had a crush. However, it should be noted that only 2.86% of the respondents in the sample have not had a crush. Lastly, females have higher chances of thinking of committing suicide than males. The odds for females to think of committing suicide is 2.67 times the odds for males. Table 2 gives the classification table from the model.

It can be seen from the table that the total correct classification is 86.2%. The model has a high correct classification (99.2%) for those who did not think of committing suicide. However, the percentage of correct classification for those who thought of committing suicide is only 6.2% which is very small. A strong reason for this is that in the original data, only 13.9% have thought of committing suicide while 86.1% have never thought of such. Moreover, there are a lot of other factors affecting suicidal thoughts not accounted by the model. Hence, the model is biased towards classification on the non-suicidal category.

Table 3. Classification Table from the Logistic Model

		Predicted		
		No	Yes	Correct Classification
Observed	No (without suicidal thoughts)	15949	127	99.2%
	Yes (with suicidal thoughts)	2442	162	6.2%
				86.2%

5.2 Classification and regression tree

The dataset is partitioned into two. One will serve as the training set and the other as the test set. According to Breimanetal (1984) an independent test set is preferred over cross-validation for large datasets. The root node error from the maximal tree using the training set is 13.96%. Figure 1 shows the obtained maximal tree.

Figure 2 shows the pruned tree. The root node error is still 13.96% which is the same as the maximal tree. At the root node, there are 86% observations who have never thought of committing suicide while 14% have thought of committing suicide. There are 9 splits and 10 terminal nodes in this tree. There are 3 paths or sequences of splits for a young person to be declared having suicidal thoughts, and 7 paths for a young person to be declared as having no suicidal thoughts. The first variables used to classify is sex. If male, then there are no suicidal thoughts. The daughter node corresponding to males is already a terminal node which by the plurality rule is a path for nonsuicidal. If a young person is female and during the past 3 months did not feel depressed about life in general, then the person is classified as nonsuicidal which is another terminal node. If a female young person has felt depressed in the past month but do not feel sad or helpless always or even most of the time, the person is also classified as nonsuicidal. If a female young person has felt depressed, and sad or helpless but did not experience being threatened by someone, then the person is classified as nonsuicidal. However, if the same person has experienced being threatened by someone and has carried a weapon (i.e. gun, knife, club, etc.), then the person is classified as suicidal. Otherwise, if the same person did not experience carrying a weapon but did not

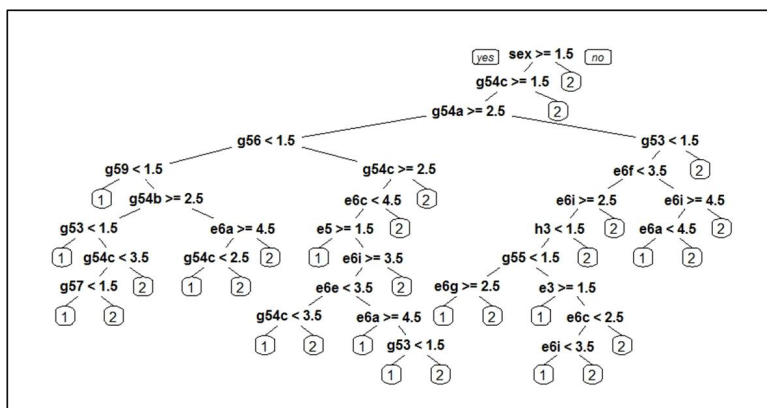


Figure 1. Maximal Tree using the Training Set

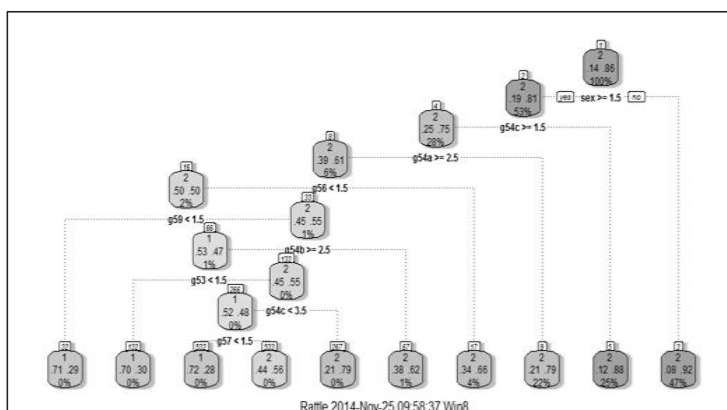


Figure 2. Pruned Tree from the Training Set

feel hopeless about the future, then the person is classified as nonsuicidal. If the person felt hopeless about the future in the past 3 months and has friends or acquaintances who committed suicide, then the person is classified as suicidal. An anomalous finding in the tree is that if the person has no friends or acquaintances who committed suicide and always felt depressed, then the person is classified as non-suicidal. If the female person who felt hopeless about the future, has no friends or acquaintances who committed suicide but experienced being physically injured by someone, then the person is qualified as suicidal. Otherwise, if there is no experience of being physically injured, then the person is qualified as non-suicidal.

Applying the model on the test set, the total percentage of correct classification is 86.4%. For those who have not thought of committing suicide, the percentage

of correct classification is 99.75%. However, for those who committed suicide, the correct classification is only 2.72%. When applied to the training set, the percentage of correct classification for those who thought of committing suicide is around 10%. As expected, the predictive power of the classification model on the minority class is very small while the model is biased towards classifying the observations on the majority class.

5.3 Conditional inference tree

Figure 3 shows the conditional inference tree based on the training set. The variable which has the most significant association with the binary response variable is about the feeling of depression about life in general during the past 3 months. The next variable is sex. The next splits after sex are the variables pertaining to experiences of being threatened by someone, experiences of being physically injured by someone, sadness and helplessness which causes the person to stop doing the usual activities, and the feeling of hopelessness about the future. Then in the latter splits, the factors of having barkada at present, satisfaction with one's self, and thoughts of being no good at all have been used.

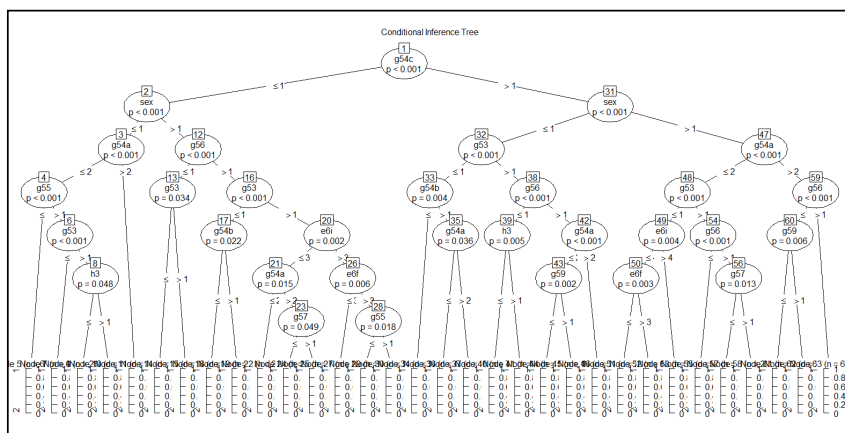


Figure 3. Conditional Inference Tree of the Training Set

After applying the model to the test set, the overall percentage of correct classification is 85.23%. The percentage of correct classification for those who have never thought of committing suicide is 99.19%. However, only 3.66% of those who have thought of committing suicide have been correctly classified by the model. This is the same problem encounter from the logistic model and CART.

5.4 Random forest

A random forest of 2,000 trees was generated. The out-of-bag estimate of error rate is 13.9%. Based on the confusion matrix, the classification error for the category of haven't thought of committing suicide is only 0.91%. However, the classification error for the category of thinking about committing suicide is very high which is 94.1%. This is the same problem encountered from the logistic regression and the conditional inference tree.

Table 3 gives the importance of each variable from the random forest.

Table 4. Importance of Each Variable from the Random Forest

Variable	1	2	Mean Decrease Accuracy	Mean Decrease Gini
e2	-2.818	7.755	6.496	5.808
e3	-4.189	9.238	6.724	64.513
e4	-6.086	18.634	14.498	81.472
e5	-4.127	18.878	15.829	50.694
e6a	-14.696	46.197	36.590	221.602
e6b	-3.871	34.514	30.503	211.628
e6c	-4.123	22.508	19.280	266.646
e6f	4.189	34.707	33.249	242.132
e6g	-1.752	37.367	33.965	259.865
e6i	20.087	24.722	31.123	242.716
g53	61.494	19.169	46.839	109.501
g54a	41.047	34.040	51.968	193.574
g54b	23.640	39.062	49.292	172.366
g54c	47.070	48.954	68.995	184.925
g55	7.332	41.620	42.606	83.806
g56	10.724	65.068	67.786	88.611
g57	14.593	38.017	42.279	79.952
g59	-2.968	45.941	44.118	69.156
h1	0.000	0.000	0.000	0.000
h3	11.613	21.369	24.475	76.561
j10	-1.073	-0.594	-0.959	19.973
sex	67.363	70.165	92.747	130.607

Based on the mean decrease in the accuracy, the most important variable is sex. The other variables considered as the top 5 most important are feeling of depression about life in general, experiences of being threatened by someone, feeling of sadness and helplessness, and feeling of hopelessness about the future. Based on the mean decrease in Gini, the most important variable is the feeling of being proud of self and is followed closely by the feeling that many things a

person does is not meaningful. The other variables are satisfaction with self and the feeling of being no good at all. Also with high importance are self-worth and the feeling of having a number of good qualities.

Figure 4 and Figure 5 shows the Type I and Type II plots.

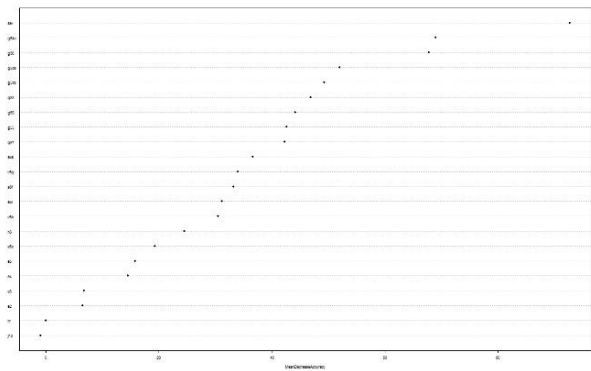


Figure 4. Type I Plot

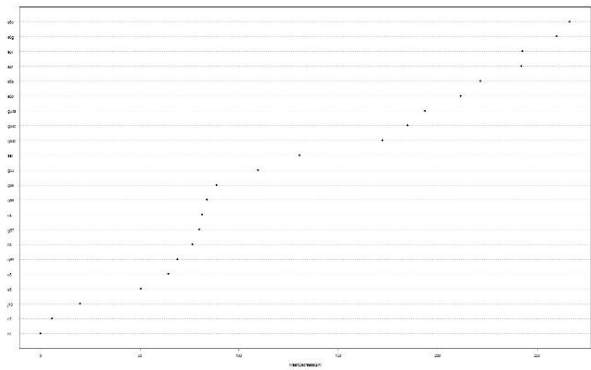


Figure 5. Type II Plot

Balanced dataset

To make the dataset balanced, a subset of the observations in the majority class was generated using simple random sampling without replacement while all the observations in the minority class were retained. The root node error was 50%. From the random forest, the OOB estimate of error rate is 34.93%. Below is the confusion matrix:

Table 5. Confusion Matrix

		1	2	Classification Error
Actual	No (without suicidal thoughts)	1677	927	35.6%
	Yes (with suicidal thoughts)	892	1712	34.25%

It can be seen that the classification errors are balanced out. The classification error for the category of having suicidal thoughts is 35.6% while the classification error for the category of having no suicidal thoughts is 34.25%. Table 4 gives the importance of the variables.

Table 6. Importance of the Variables from the Balanced Data Set

Variables	Mean Decrease Accuracy	Mean Decrease Gini	Variables	Mean Decrease Accuracy	Mean Decrease Gini
e2	-4.08384	2.380507	g54a	25.21504	143.9696
e3	-1.12349	40.29299	g54b	22.94331	127.2711
e4	4.893257	51.53841	g54c	30.32165	144.0239
e5	3.235539	30.81164	g55	16.09948	64.92915
e6a	2.612227	149.7699	g56	14.51834	66.18226
e6b	-0.21282	139.4804	g57	6.784102	54.10658
e6c	4.831406	194.3236	g59	2.960144	43.57345
e6e	7.807873	151.7184	h3	6.666195	51.77953
e6f	5.966944	178.1967	j10	4.648297	18.91633
e6g	7.80058	193.4471	sex	41.52687	106.8387
e6i	13.91409	181.2053	rtype	10.88722	87.46838
g53	30.46371	87.6386			

The identified most important variables based on the mean decrease in accuracy are the following: sex, if any friends/ acquaintance attempted to commit suicide, if the respondent has felt really depressed about life in general during the past 3 months, if the respondent has felt sad or helpless that made him/her stop doing usual activities during the past 3 months, and if the respondent has felt really hopeless about the future during the past 3 months. Based on the mean decrease in Gini, the top 5 most important variables are the following: the feeling of being proud of one's self, satisfaction with self on the whole, taking a positive attitude towards self, feeling that many things he/she does are not so meaningless, and feeling of being no good at all.

After running cart on the balanced data set, below is the classification table on the test set:

Table 7. Classification Table from Cart in the Test Set using the Balanced Data Set

		Actual	
		No	Yes
Predicted	No (without suicidal thoughts)	700	1700
	Yes (with suicidal thoughts)	1904	904
Correct classification		73.12%	65.28%

Table 8. Classification Table from Cart in the Full Balanced Data Set

		Actual	
		No	Yes
Predicted	No	11086	904
	Yes	4990	1700
Correct classification		68.96%	65.28%

Using the test set, the correct classification for those who have not thought of committing suicide is 65.28% while the correct classification for those who have thought of committing suicide is 73.12%. Applying the classification model on the full data set, the percentage correct classification for the two categories are 65.28% and 68.96%, respectively

5.5 Proposed algorithm with resampling

A proposed algorithm to get an estimate of the percentage correct classification is given below:

1. Get a random sample from the observations who have not thought of committing suicide.
2. Merge the random sample from (a) and all observations who have thought of committing suicide. This will be the training set to be used, which is balanced.
3. Run the CART on the training set.
4. Get the classification table using the classification model from (c) on the training set and the full dataset.
5. Do the steps (a) to (b) many times.
6. The final estimate for the percentage correct classification will be the mean of the estimates from the different resamples. The standard error of the estimates will also be computed.

Table 7 and Table 8 gives the mean of the percentage correct classification for each category of the response variable from the 200 resamples.

Table 9. Mean and Standard Deviation of the Percentage Correct Classification from the 200 Resamples

	Mean		Standard Deviation	
	Yes Suicide	No Suicide	Yes Suicide	No Suicide
Training Set	0.6813	0.6762	0.0283	0.0367
Full Dataset	0.6813	0.6517	0.0283	0.0270

Table 10. Minimum and Maximum of the Percentage Correct Classification from the 200 Resamples

	Min		Max	
	Yes Suicide	No Suicide	Yes Suicide	No Suicide
Training Set	0.5956	0.5929	0.7358	0.7596
Full Dataset	0.5956	0.5902	0.7358	0.7145

From using the training set, the mean of the correct percentage classification for the category of having thought of committing suicide and have not thought is 68.13% and 67.62%, respectively. The standard deviations are 2.83% and 3.67%, respectively. From using the full data set as the test set, the mean of the correct percentage classification of having thought of committing suicide and have not thought of committing suicide is 68.13% and 65.17%, respectively. The standard deviations are 2.83% and 2.7%, respectively.

The minimum of the correct percentage classification for the category of having thought of committing suicide is 59.56% and 73.58%, respectively. Using the training set, the minimum and maximum of the correct percentage classification for the category of have not thought of committing suicide is 59.29% and 75.98%, respectively. On the other hand, in using the full data set as the test set, the minimum and the maximum are 59.02% and 71.45%, respectively.

The percentage of correct classification for the minority class improved although there's a decline in the percentage or correct classification for the majority class. Nonetheless, the algorithm, which uses resampling strategy and a balanced training set which is a subset of the full data set, balanced out the predictive ability in the two classes of the response variable.

Figure 6 shows a pruned tree from a balanced training set which is a subset of the full data set.

From this tree, the observations are more distributed in the terminal nodes. There are 13 terminal nodes in which 8 of them are terminal nodes for those who have thought of committing suicide and 5 nodes for those who have never thought of committing suicide. The following paths are for the suicidal:

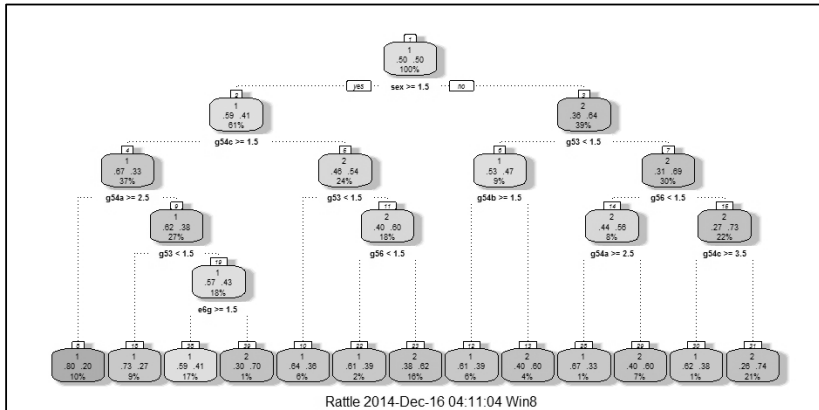


Figure 6. Pruned Tree from one of the Balanced Subsets of the Full Data Set

- Males, those with friends who committed suicide, and felt really hopeless about life;
- Males, those with no friends who committed suicide, experienced being threatened by someone, and felt really sad or helpless most of the time or always;
- Males, those with no friends who committed suicide, did not experience being threatened by someone, but always felt being depressed about life in general;
- Females, has felt depressed about life in general, felt sad or helpless in the past 3 months;
- Females, has felt depressed about life in general, did not feel sad or helpless in the past 3 months, with friends or acquaintances who committed suicide;
- Females, has felt depressed about life in general, did not feel sad or helpless in the past 3 months, with no friends or acquaintances who committed suicide, but feels that many things are not meaningful to him/her;
- Females, did not feel depressed about life in general, but with friends or acquaintances who committed suicide; and
- Females, did not feel depressed about life in general, with no friends or acquaintances who committed suicide, but ever experienced being threatened by someone.

6. Conclusions

In the practical side, the results of the study have identified variables which can significantly predict tendencies of having suicidal thoughts. The different statistical methods employed in the study have equivalent practical significance.

The binary logistic regression have identified significant predictors of the odds of thinking about committing suicide. If a person has friends or acquaintances who committed suicide, the odds that he/she will think about committing suicide as well will increase by 92.9% of the original odds that if a person has no friends or acquaintances who committed suicide. Feelings of depression, sadness or helplessness, and hopelessness about the future are also significant indicators of having thoughts of committing suicide. Those who felt being sad or helpless that made him/her stop doing usual activities during the past 3 months have high odds of thinking about committing suicide. Moreover, those who felt being hopeless about the future and felt depressed about life in general have higher odds of thinking about committing suicide. Furthermore those who experienced being threatened by someone and being physically injured by someone and those who felt unsafe or threatened have higher chances of thinking about committing suicide. Variables about self-esteem, satisfaction with one's self, being proud with self, and other related variables are also significant predictors. The results are logical since those who feel dissatisfied with self or those with low self-esteem have higher chances of thinking about committing suicide. Variables related to relationships with other people can significantly predict suicidal tendencies. Those who have friends or acquaintances who committed suicide have higher odds of thinking about committing suicide. Furthermore, having a barkada will help decrease the odds of suicidal tendencies. Consistent with the literature, females have higher odds of thinking of committing suicide than males. An interesting result in the paper is that those who had a crush have higher chances of thinking about committing suicide compared to those who have not had crush. The results of the logistic regression are consistent with the literature.

The variables used in the classification algorithm of CART and the Conditional Inference Tree are basically the same variables. These variables are the variables found to be significant from the logistic regression. However, there is a problem with the predictive ability of the classification models. The node error of the maximal tree and pruned tree is almost the same as the percentage of observations of the minority class which is the class of those who have thought of committing suicide. This is a common problem with unbalanced data sets. The predictions are biased towards predicting the majority class. In fact, in some if not in most of the cases, the error classification rate is the same as the percentage of the minority class in the sample. No wonder why such problem exists in the results of the models. The random forest also gave bad predictive ability although the important variables identified based on the mean decrease in accuracy and mean decrease in Gini are logical.

The algorithm which uses a balanced subset of the dataset improves the predictive ability in terms of balancing out the correct classification in the two classes of the response variable. There is a significant increase in the percentage

of correct classification for the minority class although there is a decrease in the predictive ability for the majority class. Nonetheless, the resampling strategy and the generation of a balanced data set is still very helpful in getting an acceptable predictive ability at least in the class of more interest which is the class with the suicidal thoughts. Another advantage is that a particular pruned tree from a sample of subset also gave a better distribution of the elements in the terminal nodes.

REFERENCES

- BEAUTRAIS, A., 2003, Subsequent mortality in medically serious suicide attempts: A 5-year follow-up, *Aust NZ J Psychiatry* 37(5): 595-9.
- BECK, A., BROWN, G., BERCHICK, R., STEWART, B., STEER, R., 1990, Relationship between hopelessness and ultimate suicide: A replication with psychiatric outpatients, *Am J Psychiatry* 147(2): 1990-5.
- BREIMAN, L., FRIEDMAN, J., 1997, Predicting multivariate responses in multiple linear regression (with discussion), *Journal of the Royal Statistical Society Series B*. 59:3-54.
- CANTOR, P., 1989, Intervention strategies: Environmental risk reduction for youth suicide. Report of the Secretary's Task Force on Youth suicide: Volume 4. (DHHS Publication No. ADM 89-1623). Washington, D.C: U.S. Government Printing Office.
- CASIANO, H., KATZ, L., GLOBERMAN, D., SAREEN, J., 2013, Suicide and deliberate self-injurious behavior in juvenile correctional facilities: A review, *J Can AcadChilAdolesc Psychiatry* 22(2): 118-24.
- CAVANAGH, J., CARSON A., SHARPE, M., LAWRIE, S., 2003, Psychological autopsy studies of suicide: A systematic review, *Psychol med* 33(3): 395-405.
- CEBULA, R., ZELENSKAYA, T., 2006, Determinants of youth suicide: A friendly comment with suggestions, *American Journal of Economics and Sociology* 65(4):991-996.
- Centers for Disease Control, 1991, Attempted suicide among high school students – United States, 1990, *Morbidity and Mortality Weekly Report* 40:633-635.
- CHANG, S., STUCKLER D., YIP, P., GUNNEL, D., 2013, Impact of 2008 global economic crisis on suicide: Time trend study in 54 countries. *BMJ* 347: f5239.
- CHO, J., LEE, W., MOON, K., SUH, M., SOHN J., HA, K., 2013, Medical care utilization during 1 year prior to death in suicides motivated by physical illnesses, *J Prev Med Public Health* 46(3): 147-53.
- DEYKIN, E., AND BUKA, S., 1994, Suicidal ideation and attempts among chemically dependent adolescents, *American Journal of Public Health* 84:634-639.
- FOSTER, T., 2011, Adverse life events proximal to adult suicide: A synthesis of findings from psychological autopsy studies, 15(1): 1-15.
- GARLAND, A., AND ZIGLER, E., 1993, Adolescent suicide prevention: Current research and social policy implications, *American Psychologist* 48: 169-182.
- Global status report on alcohol and health 2014. World Health Organization; 2014.
- IZENMAN, A., 2008, *Modern Multivariate Statistical Techniques* New York, USA. Springer Science+Business Media, LLC.

- JOHNSON, J., COHEN, P., GOULD, M., KASEN, S., BROWN, J., BROOK, J., 2002, Childhood adversities, interpersonal difficulties, and risk for suicide attempts during late adolescence and early adulthood, *Arch Gen Psychiatry* 59(8): 741-9.
- KPOSOWA, A., 2003, Divorce and suicide risk, *J Epidemiol Community Health* 57(12):993.
- LUTHAR, S., (editor), 2003. *Resilience and Vulnerability: Adaptation in the Context of Childhood Adversities*, Cambridge: Cambridge University Press.
- O'CONNOR R., NOCK, M., 2014, The psychological of suicidal behavior, *Lancet Psychiatry* 1(1): 73-85.
- QIN, P., AGERBO, E., MORTENSEN, P., (2002, Suicide risk in relation to family history of completed suicide and psychiatric disorders: A nested case-control study based on longitudinal registers, *Lancet* 360(9340): 1126-30.
- ROSENBERG, M., RODRIGUEZ, J., AND CHORBA, T., 1990, Childhood injuries: Where we are, *Pediatrics* 86:1084
- SISASK, M., VANI, K., KOLVES, K., KONSTABEL, K., WASSERMAN, D., 2008, Subjective psychological well-being (WHO-5) in assessment of the severity of suicide attempt, *Nord J Psychiatry* 62(6): 431-5.
- STRASSER, H., WEBER, C., 1999, On the asymptotic theory of permutation statistics, *Mathematical Methods of Statistics* 8:220-250.
- TONOOKA, F., 1999, Determinants of Social Attitudes Toward Suicide, *Michigan Sociological Review* 13:80-104.