A Comparison of Random Forest and Decision Tree for Suicide Ideation Classification

Mayam Mohammadian-Khoshnoud
Hamadan University of Medical Sciences Medical School
Tahereh Omidi
Hamadan University of Medical Sciences Medical School
Javad Faradmal (javad.faradmal@umsha.ac.ir)
Hamadan University of Medical Sciences Medical School
Jalal Poorolajal
Hamadan University of Medical Sciences Medical School

Research article

Keywords: Suicide Ideation, Random forest, Decision tree, Logistic regression

Posted Date: September 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-66839/v1

Abstract

Background: Suicide resulted from complex interaction factors. Most classical statistical methods were not efficiently enough to cover this complexity. With the new branch of statistics as statistical/machine learning, complex relationships between risk factors and responses can be modeled.

Methods: We aimed to identify the high-risk groups for suicide using different classification methods including logistic regression(LR), decision tree(DT), and random forest(RF). Also, the prediction accuracy of the models is compared. This study used data obtained from a cross-sectional study conducted in in the Hamadan University of Medical Sciences from 2015-2016 to investigate the prevalence of suicidal ideation and related risk factors among university students. The LR, DT, and RF models were used to evaluate the high-risk group for suicide. Finally, the applied all three models were compared using sensitivity(SE), specificity(SP), and the area under receiver operating characteristics (ROC) curves.

Results: In the training sample, the area under the ROC curve of the DT was greater than the LR and RF. But in the validation sample, the RF model has the best performance and the DT has the worst performance among these methods.

Discussion: In this study, the risk factors for suicide were different for men and women. According to the results of the DT, substance abuse, average, general health score, faculty of education, depression were the risk factors on suicidal ideation in both genders. But despair about the future, residence (parents' house/dormitory) were among the factors contributing to the suicidal ideation of men. On the other hand, parents' education, interested in the discipline and anxiety influence factors on suicidal ideation in women. The results of RF indicated that depression, general health score, average, anxiety and substance abuse were important risk factors for suicidal ideation in both genders. Also, the faculty of education and age are risk factors for suicide in women.

Conclusions: In the training sample, the DT had better performance but in the validation sample, the RF model provided better results. The LR was the best model for diagnosis of the patient and the DT and RF are considered the best models to diagnosis a healthy individual.

Introduction

The term suicide was first defined in 1642, with a combination of Latin terms Sui and CADER meaning self and killing, respectively(1). If someone commits suicide but remains alive, this behavior will be defined as attempted suicides(2) and suicide ideation is defined as the thought of injuring or killing yourself (3). In the last decade, the suicide rate among university students in worldwide has increased significantly(4). Suicide is the third cause of death in people aged 15 to 24 years old(5), and the second leading cause of death among college students(6). However, suicide is observed in all age groups, but it is of greater importance in young adults because of their loss of life expectancy(7, 8). University years are the most vulnerable period for these people due to pressure to succeed, financial burden and responsibility for the transition to adulthood(9). Suicide is a multidimensional phenomenon which is resulted from a complex interaction of biological, genetic, psychological and environmental factors(10). The suicide ideations are closely related to suicide, therefore identifying risk factors of suicide ideation can be very important in reducing suicide rates(11–15). Due to the complex relationships of various factors in the creation of behavior, the most classical statistical methods that are based on the functional form of a predetermined and simple relationships between factors are not sufficiently accurate to cover this complexity. But today, with the new

branch of statistics as statistical/machine learning, complex relationships between factors and responses can be modeled. This method leads to the partitioning of people based on different risk factors. This will, therefore, help identifying high-risk groups. This study aimed to identify the high-risk group for suicide using different classification methods including logistic regression (LR), decision tree (DT) and random forest (RF). Also, we evaluate prediction performance of these models to identification high-risk person.

Methods

Data and settings:

This study used data obtained from a cross-sectional study conducted in the Hamadan University of Medical Sciences from 2015–2016 to investigate the prevalence of suicidal ideation, suicide attempt and related risk factors among university students(16). We enrolled students who had passed at least one semester of their education at the university. The associated response to this study is suicide ideation which is recorded as a binary variable. Three methods of LR, DT, and RF were fitted to the data to gender segregation. The required data needed to be divided into two subsamples of training and validation. The training sample finds the model and the validation sample tests the performance of the trained model. Because the causes of suicide differ between men and women, the risk factors of suicidal ideation for men and women were analyzed separately. Predictors of age, marital status, residence (parents' house/dormitory), mother's education, father's education, Educational level, faculty of education, average, interested in the discipline, despair about the future, substance abuse, general health score, depression, anxiety, boyfriends/girlfriends, emotional breakdown, illegitimate heterosexual and/or homosexual intercourse, cigarette smoking, city, birth order were analyzed.

Statistical models:

Logistic Regression

LR is a modeling mechanism that can be used to describe the relationship between multiple predictive variables with a binary dependent variable. LR is one of the most commonly used methods in the study of Epidemiological data when the dependent variable is dichotomous(17). One of the causes of the popularity of LR is related to the regression estimators presented in the range of zero to one and also describes the s-shape of the combined effect of several risk factors for disease(17). The logit (log odds) has the following linear relation(18):

logit(π(x)) = log (π(x))/(1-π(x)

Decision Tree

In this analysis, the response variable is a dichotomy. This model is designed for quantitative variables but applicable to any form of variable. DT are the most powerful classification algorithms that have gained more popularity through the growth of data mining in various fields. Popular DT algorithms include Quinlan ID3, C4. 5 (19, 20) C5 and CART(21). This technique recursively separates the observations in the branches to construct a tree to improve the predictive accuracy. To do this, mathematical algorithms such as Information Gain, Gini index and Chi-Square test are used to identify a variable and associated threshold for the variable which divides the input observations into two or more subgroups. This process on each leaf node is repeated until the complete tree is constructed. The splitting algorithm aims are to find the variable and threshold pair that maximizes the homogeneity (order) of two or more subgroups of sample. The mathematical algorithms such as Information Gain

in C4.5, C.5, ID3 trees, Gini index in CART and Chi-Square test in CHAID are used(22). The CART algorithm can be used as one of the best-known diagnostic and predictive classification methods in the medical sciences(23). Also, in the CART model, the classification tree pruning is based on the complexity cost. The easy perception of DT, using both nominal and categorical data and the absence of hypothesis about the nature of data are appropriate properties of DT(24).

Random forest algorithm

Breiman introduced the RF classification algorithm in 2001(25). The RF is an ensemble of unpruned regression and classification trees(26). The Bootstrap sample is extracted to construct a RF. Afterward, the recursive partitioning is used to the Bootstrap sample. The q predictors are randomly selected of the p predictors at each node. The recursive partitioning is run to the end and a tree is formed. The above steps are repeated until a forest is formed. The forest-based classification of the majority vote from all trees is formed(27). Generally, RF demonstrates significant performance to single tree classifiers such as CART and C4.5(24).A RF, unlike trees, is extremely large for interpretation. One way to summarize or quantify information is to identify the forest's important predictors.

Evaluation models’ performance:

The results derived from the training and validation samples were evaluated by utilizing the sensitivity (SE), specificity (SP) and the area under curve (AUC).

The accuracy of diagnostic tests is assessed with two conditional probabilities: sensitivity, specificity(28). The ability of a test to identify correctly those who have the disease (or characteristic) of interest is called sensitivity(29). The ability of a test to identify correctly those who do not have the disease (or characteristic) of interest is called specificity(29). The receiver operating characteristic (ROC) curve is the plot that displays the complete picture of the compromise between sensitivity and 1- specificity over a series of cutoff points. Area under the ROC curve is as a measure of the intrinsic validity of the diagnostic test(30).

Software:

The data were analyzed through 'tree', 'RF', 'ggplot2' packages of R software 3.5.1(31–33). The control argument in the tree package was used to determine the minimum number of observations per node, the smallest size of each node and the within-node deviance, were set at 10, 20, and 0.01, respectively. Also, the cut-off point 0 .3 was used in the DT for allocating the response to each node. In other words, if the response value is less than 0.3, then no is allocated to the final node.

Results

The statistical population consisted of 1259 students in this study who after refining the data 1247 individuals were included in the final analysis. Of the 1247 participants, 491(39.4%) were male and 756(60.6%) were women. The mean age of students was 18 to 49 years with a mean of 22.52 (3.33). Of participants,1044 (83.8%) were single, 160 married (12.8%) and 43 (3.4%) divorced. Of the university students who participated in the study, 146 (around 12%) had suicidal ideation during the past year and 63 (5%) students, had attempted suicide at least once in the past year. Suicide ideation was higher among men than in women. For analyzes LR, DT, and RF were fitted for suicidal ideation to gender segregation.

Women’s suicide

The LR fit results indicated that in women, father's education, depression, substance abuse has a significant relationship to suicidal ideation. The DT for women's suicidal ideation has 16 terminal nodes. Of the 21 selected features for entering in the tree decision, variables of depression, age, average, substance abuse, mother 's education, general health score, interested in the discipline, faculty of education, father 's education, anxiety are important attributes in the DT. The misclassification error rate is 0.58. A total of 500 trees were used to construct the RF of suicide ideation in women.

According to the Gini importance index, depression, general health score, average, anxiety, faculty of education, age and substance abuse were the most important predictors, respectively.

Men’s suicide

Regression fitting results showed a significant relationship between age, marital status, substance abuse, depression and suicidal ideation in men. The DT for suicidal ideation in men has 14 terminal nodes. Of 21 entered features, depression, substance abuse, average, faculty of education, despair about the future, residence (parents' house/dormitory) and general health score were selected by the DT. The misclassification error rate is 0.079. Important predictors based on the Gini importance index in the RF were depression, general health score, anxiety, substance abuse.

Comparing the results of logistic regression, the variables of depression, substance abuse in both genders have a significant relationship with the response. By comparing DTs, four features of faculty of education, general health score, average, substance abuse, depression are the effective factors in both genders. The results of RF indicated that depression, general health score, anxiety, substance abuse, average were identified as important variables.

The comparison of the area under the ROC curve, the sensitivity and specificity of the three methods of LR, DT, and the RF to gender segregation are shown in Table 1. In both genders, the area under the ROC curve in the training sample and DT is better than the LR and RF. In fact, in the existing samples, the DT has good performance and the RF in the training sample has the lowest performance among the three methods compared. The RF model has the best performance in the validation model, and the DT has the weakest performance among the three methods. In a new sample, RF performs better than the other two methods. The sensitivity of LR in training and validation samples is higher than the DT and RF. Among the three methods compared, RF and DT have the highest specificity, respectively.

Table 1
The comparison of classification methods for suicidal ideation

Female	model	suicidal ideation					
		train sample			test sample		
		LR(SE)	DT(SE)	RF(SE)	LR(SE)	DT(SE)	RF(SE)
	ROC	0.90	0.96	0.86	0.84	0.72	0.85
	sensitivity	0.85(0.011)	0.65(0.015)	0.36(0.015)	0.77(0.027)	0.41(0.31)	0.20(0.025)
	specificity	0.82(0.012)	0.97(0.005)	0.98(0.004)	0.89(0.020)	0.94(0.015)	0.99(0.006)
Male	ROC	0.94	0.95	0.87	0.77	0.73	0.77
	sensitivity	0.82(0.012)	0.54(0.016)	0.37(0.015)	0.68(0.030)	0.37(0.031)	0.32(0.030)
	specificity	0.87(0.011)	0.98(0.004)	0.99(0.003)	0.78(0.026)	0.92(0.017)	0.97(0.011)

Discussion

The purpose of this study was to identify significant risk factors associated with gender-specific suicidal ideation. In this study, the risk factors for suicide were different for men and women. According to the results of the DT, substance abuse, average, general health score, faculty of education, depression were the risk factors on suicidal ideation in both genders. But despair about the future, residence (parents' house/dormitory) were among the factors contributing to the suicidal ideation of men. On the other hand, parents' education, interested in the discipline and anxiety influence factors on suicidal ideation in women. The results of RF indicated that depression, general health score, average, anxiety and substance abuse were important risk factors for suicidal ideation in both genders. Also, the faculty of education and age are risk factors for suicide in women. The relationship between depression and suicide among students has been reported in other studies(34). There is a well-known relation between depression and suicidal ideation(35, 36). The results of this study were consistent with the study results which have reported the relationship between anxiety and suicide(37–39). The relationship between substance abuse and suicidal ideation in medical students has also been confirmed in previous studies(40, 41). The next purpose of this study was to compare the performance of LR, DT and RF based on sensitivity, specificity, and area under the ROC curve. The predictive power of the classifiers was investigated by the area under the ROC curve (AUC), in women's suicidal ideation, the area under ROC curve of the DT in the training sample is higher than LR and RF, in a similar study, the area under ROC curve for the DT is better than LR and discriminant analysis(42). But in the validation sample, the area under the ROC curve of RF is higher than the LR and DT, in Tian study, the area under the ROC curve is better for RF than DT and LR(43). The sensitivity range between 0.2 (the validation sample of the RF for women's suicidal ideation) was 0.85 (the training sample of the LR for women's suicidal ideation). The minimum specificity was 0.78 (the validation sample of the LR for male's suicidal ideations) and the maximum specificity was 0.99 (the validation sample of the RF for female suicidal ideations and the training sample of the RF for male's suicidal ideations). The area under the ROC curve of the DT in the training sample has better than LR and RF, The area under the ROC curve of the RF in the validation sample and specificity of RF for validation and training samples better than the other two methods. In women's suicidal ideations, the sensitivity of the LR model in the training and validation samples was better than the other two methods. These results were also confirmed in men's suicidal ideations.

In the training sample, although the area under the curve of the DT was higher than the LR, the sensitivity of the DT and RF in suicidal ideations of men and women in both training and validation samples was lower than the LR. Power of detection of individuals exposed to suicidal ideations using LR better than RF and DT. But, in the case of specificity in both training and validation samples, the DT and the RF has better performance than LR. In identifying people who do not have suicidal ideations, the DT and RF perform better than LR.

Conclusion

In binary classification, the DT has better performance in the training sample, but in the validation sample, the RF provides better results. Among the three classifiers, to identify the patient, LR is the best model and the DT and the RF are the best models to identify the healthy person. By comparing the results of the three methods of LR, DT, RF in suicidal ideation of men and women, depression, substance abuse were identified as important risk factors.

Abbreviations

RF: Random forest

DT: Decision tree

LR: Logistic regression

SE: Sensitivity

SP: Specificity

AUC: Area under curve

ROC: Receiver operating characteristic

Declarations

Ethics declarations

Ethics approval and consent to participate

This study was approved by the ethics committee, Hamadan University of Medical Sciences. We took the suicide data from the registry and we had permission from the ethics committee to access this data, so we did not have any direct contact with the patient to ask for consent (No. IR.UMSHA.REC.1396.778)

Consent for publication

Not Applicable.

Availability of data and materials

Readers who wish to gain access to the data can send email to poorolajal@umsha.ac.ir.

Competing interests

The authors declare that they have no competing interests.

Funding

The study was founded by Vice-Chancellor for Research and Technology, Hamadan University of Medical Sciences(No.9611107202).

Authors ' contributions

Conceptualization: Maryam Mohammadian-Khoshnoud, Javad Faradmal, Tahereh Omidi.

Data curation: Jalal Poorolajal

Formal analysis: Maryam Mohammadian-Khoshnoud, Javad Faradmal, Tahereh Omidi.

Funding acquisition: Maryam Mohammadian-Khoshnoud, Javad Faradmal, Jalal Poorolajal.

Methodology: Maryam Mohammadian-Khoshnoud, Javad Faradmal.

Writing – original draft: Maryam Mohammadian-Khoshnoud, Javad Faradmal, Tahereh Omidi, Jalal Poorolajal.

Writing – review & editing: Maryam Mohammadian-Khoshnoud, Javad Faradmal, Tahereh Omidi, Jalal Poorolajal.

Acknowledgments

We would like to appreciate the Vice-Chancellor for Research and Technology of the Hamadan University of Medical Sciences for supporting this work.

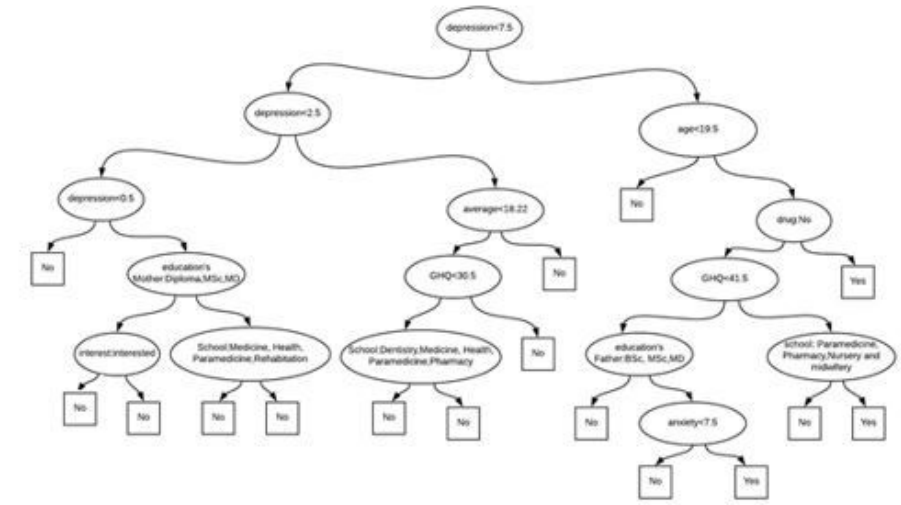
References

1. Minois G. History of suicide: voluntary death in Western culture: Baltimore: Hopkins Univers ity; 2009.
2. Haghighat S. Semnan university of medical sciences and health services 2017/12/03 [Available from: semums.ac.ir/?siteid=1&pageid=659.
3. Medicine Io. Reducing Suicide: A National Imperative. Washtington: DC: The National Academies Press; 2002.
4. Collins I, Paykel E. Suicide amongst Cambridge University students 1970–1996. Social Psychiatry and Psychiatric Epidemiology. 2000;35(3):128-32.
5. Anderson R, Smith B. Deaths: leading causes for 2002. Natl Vital Stat Rep. 2005;53(17):1-89.
6. Schwartz A. College student suicide in the United States: 1990–1991 through 2003–2004. Journal of American College Health. 2006;54(6):341-52.
7. Aliverdinia A, Rezaee A, Peyrou F. Sociological analysis of students' tendency to suicide. Applied Sociology. 2011;44(4):1-18.
8. Rezaeian M. Suicide epidemiology: Nevisande; 2010.
9. Farabaugh A, Bitran S, Nyer M, Holt D, Pedrelli P, Shyu I. Depression and suicidal ideation in college students. Psychopathology. 2012;45(5):228-34.
10. World Health Organization. PREVENTING SUICIDE A RESOURCE FOR GENERAL PHYSICIANS. . Geneva: WHO 2000.

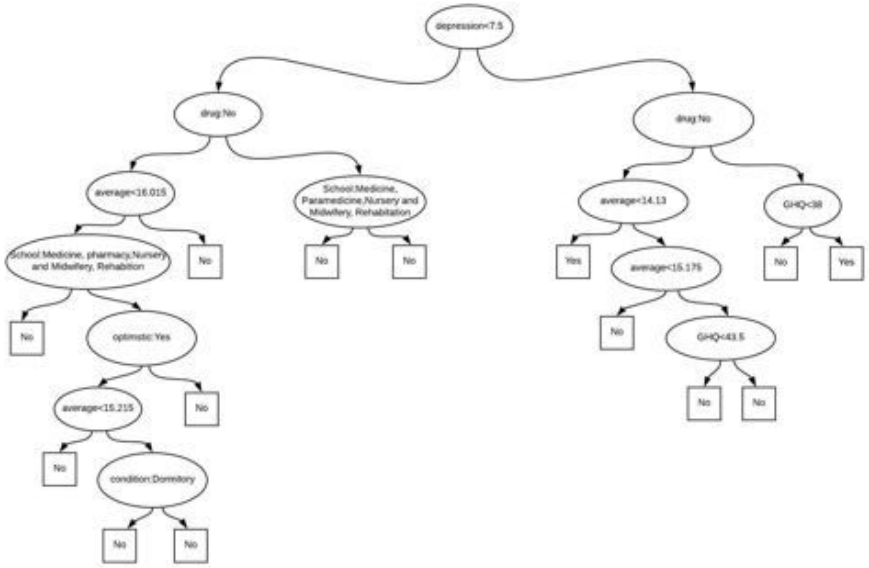
11. Casey P, Dunn G, Kelly B. Factors associated with suicidal ideation in the general population: five-centre analysis from the ODIN study. Br J Psychiatry. 2006;189:410-5.
12. De Leo D, Cerin E, Spathonis K. Lifetime risk of suicide ideation and attempts in an Australian community: prevalence, suicidal process, and help-seeking behaviour. J Affect Disord. 2005;86(2-3):215-24.
13. Kessler R, Borges G, Walters E. Prevalence of and risk factors for lifetime suicide attempts in the National Comorbidity Survey. Arch Gen Psychiatry. 1999;56:617-26.
14. Neeleman J, DeGraaf R, Vollebergh W. The suicide process: prospective comparison between early and later stages. J Affect Disord. 2004;82:43-52.
15. Suominen K, Isometa E, Suokas J. Completed suicide after a suicide attempt: a 37-year follow-up study. Am J Psychiatry. 2004;161:562-3.
16. Poorolajal J, Panahi S, Ghaleiha A, Jalili E, Darvishi N. Suicide and Associated Risk Factors Among College Students. International Journal of Epidemiologic Research. 2017;4(4):245-50.
17. Kleinbaum DG, Klein M. Logistic Regression, A Self-Learning Text. 3rd ed. New York: Springer; 2010.
18. Agresti A. Categorical Data Analysis. 3rd ed. New Jersey: John Wiley & Sons; 2013.
19. Quinlan JR. Induction of decision trees. Machine Learning. 1986;1(1):81-106.
20. Quinlan JR. C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning). 1st ed: Morgan Kaufmann; 1992.
21. Breiman L, Friedman J, Olshen R. Classification and Regression Trees: Monterey, Calif: Wadsworth & Brooks; 1984.
22. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005;34(2):113-27.
23. Behnampour N, Hajizadeh E, Semnani S, Zayeri F. The Introduction and Application of Classification Tree Model for Determination of Risk Factor for Esophageal Cancer in Golestan Province. The research Quarterty of student research commitee. 2012;1(2):38-46.
24. Zhao Y, Zhang Y. Comparison of decision tree methods for finding active objects. Advances in Space Research. 2008;**41**(12):1955-59.
25. Breiman L. Random Forests. Machine Learning. 2001;45:5-32.
26. Smith A. Image segmentation scale parameter optimization and land cover classification using the Random Forest algorithm. Journal of Spatial Science 2010;55(1):69-79.
27. Zhang H, Singer B. Recursive Partitioning and Applications. New York: Springer Series in Statistics; 2010.
28. Agresti A. An Introduction to Categorical Data Analysis Edition. 2007.
29. Szklo M, Nieto FJ. EpidEmiology Beyond the Basics. 3th ed: Jones & Bartlett Publishers; 2014.
30. KUMAR R, INDRAYAN A. Receiver Operating Characteristic (ROC) Curve for Medical Researchers. INDIAN PEDIATRICS. 2011;48:277-87.
31. Breiman L, Cutler A. Breiman and Cutler's Random Forests for Classification and Regression. 2018.
32. Ripley B. Classification and Regression Trees. 2018.
33. Wickham H, Chang W, Henry L. Create Elegant Data Visualisations Using the Grammar of Graphics. 2018.
34. Wilcox H, Arria A, Caldeira K, Vincent K, Pinchevsky G, O'Grady K. Prevalence and predictors of persistent suicide ideation, plans, and attempts during college. J Affect Disord. 2010;127(1-3):287-94.

35. Hirsch J, Visser P, Chang E, Jeglic E. Race and ethnic differences in hope and hopelessness as moderators of the association between depressive symptoms and suicidal behavior. J Am Coll Health. 2012;62(2):115-25.
36. Schulenberg JE, Zarrett NR. Mental Health During Emerging Adulthood: Continuity and Discontinuity in Courses, Causes, and Functions. In J. J. Arnett & J. L. Tanner (Eds.), Emerging adults in America: Coming of age in the 21st century. 2006:135-72.
37. Appleby L. Panic and Suicidal Behaviour: Risk of self-harm in patients who complain of panic. Br J Psychiatry. 1994;164(6):719-21.
38. Sareen J, Cox B, Afifi T, de Graaf R, Asmundson G, ten Have M, et al. Anxiety disorders and risk for suicidal ideation and suicide attempts: a population-based longitudinal study of adults. Arch Gen Psychiatry. 2005;62(11):1249-57.
39. Weissman M, Klerman G, Markowitz J, Ouellette R. Suicidal ideation and suicide attempts in panic disorder and attacks. N Engl J Med. 1989;321(18):1209-14.
40. Coentre R, Góis C. Suicidal ideation in medical students: recent insights. Advances in Medical Education and Practice. 2018;9:873–80.
41. Toprak S, Cetin I, Guven T, Can G, Demircan C. Self-harm, suicidal ideation and suicide attempts among college students. Psychiatry Research.187(1-2):140-4.
42. Sui M, Huang X, Li Y, Ma X, Zhang C, Li X, et al. Application and Comparison of Laboratory Parameters for Forecasting Severe Hand-Foot-Mouth Disease Using Logistic Regression, Discriminant Analysis and Decision Tree. Clin Lab. 2016;62(6):1023-31.
43. Tian X, Chong Y, Huang Y, Guo P, Li M, Zhang W, et al. Using Machine Learning Algorithms to Predict Hepatitis B Surface Antigen Seroclearance. Computational and Mathematical Methods in Medicine. 2019;ID 6915850:1-7.

Figures



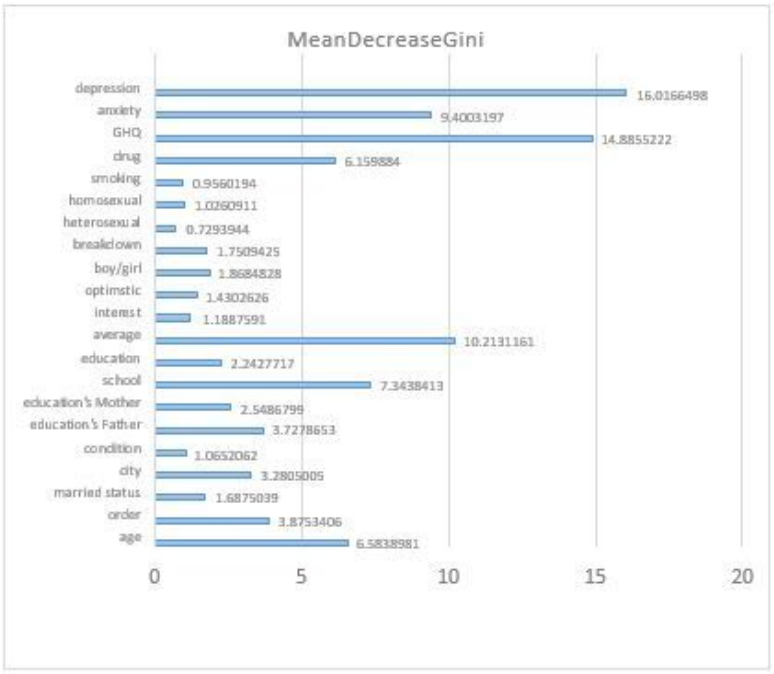
A



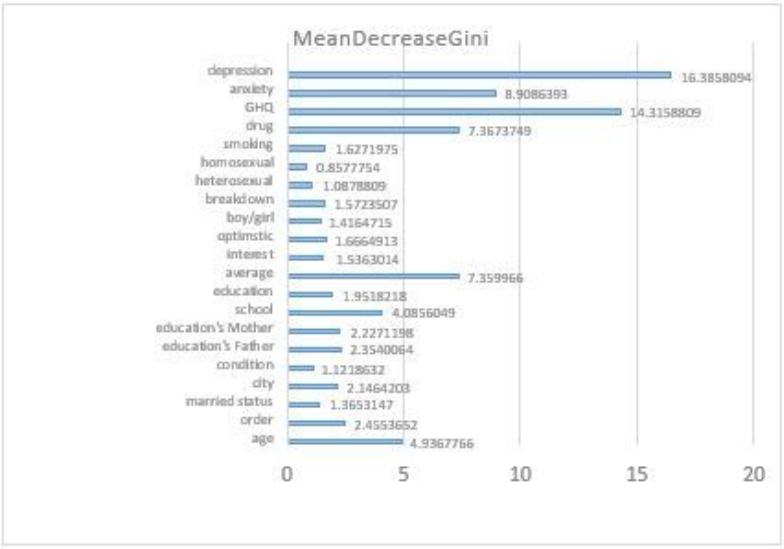
B

Figure 1

DT with allocating the response to each node A) women B) men



A



B

Figure 2

Important predictors based on the Gini importance index in RF A) women B) men

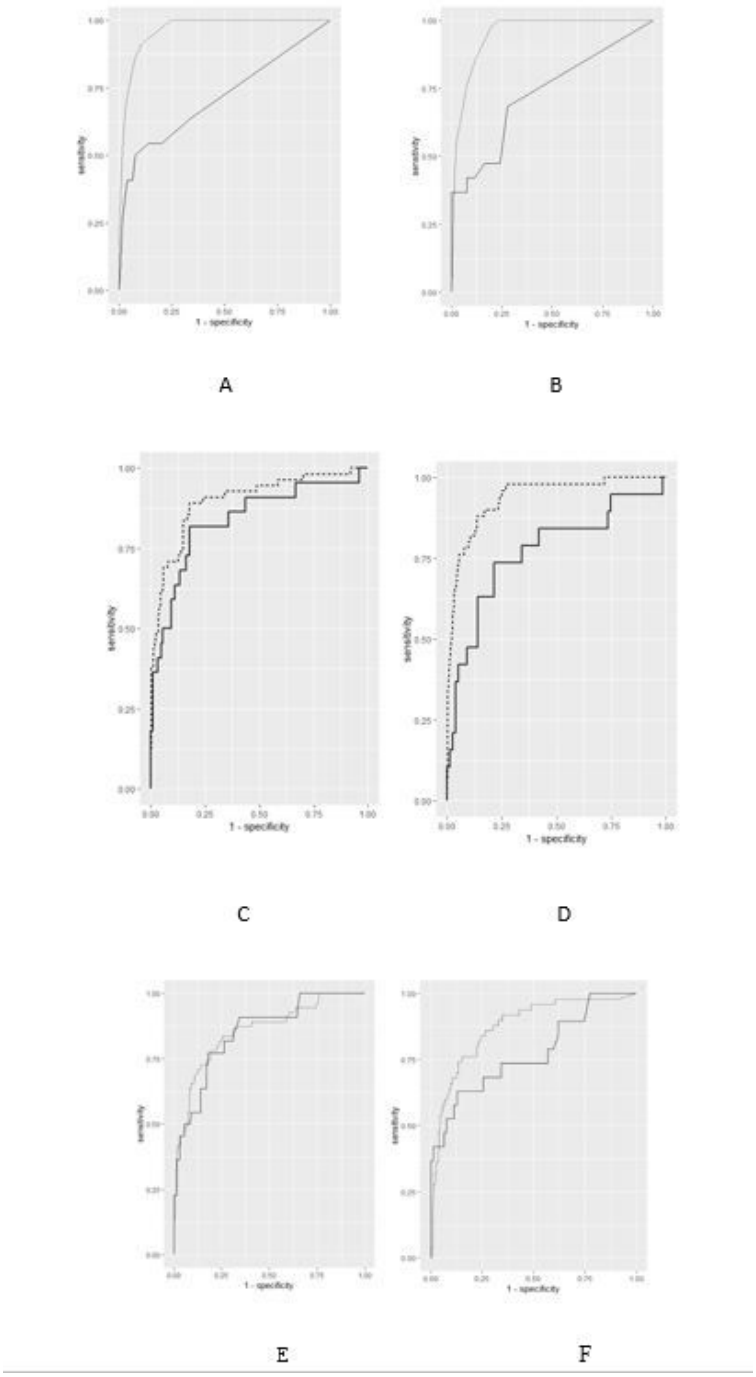


Figure 3

. The area under the ROC curve A) DT in women B) DT in men C) LR in women D) LR in men E) RF in women F) RF in men (training sample (- - -), validation sample (_____))