Iran J Public Health, Vol. 45, No.9, Sep 2016, pp.1179-1187

Original Article

Evaluating the High Risk Groups for Suicide: A Comparison of Logistic Regression, Support Vector Machine, Decision Tree and Artificial Neural Network

Payam AMINI [1], ***Hasan AHMADINIA*** [2], ***Jalal POOROLAJAL*** [3], ****Mohammad MOQADDASI AMIRI*** [2]

1. *Dept. of Epidemiology & Reproductive Health, Reproductive Epidemiology Research Center, Royan Institute for Reproductive Biomedicine, ACECR, Tehran, Iran*
2. *Dept. of Biostatistics & Epidemiology, Hamadan University of Medical Sciences, Hamadan, Iran*
3. *Research Center for Health Sciences and Dept. of Biostatistics & Epidemiology, Hamadan University of Medical Sciences, Hamadan, Iran*

***Corresponding Author:** Email: mohammad_moqaddasi@yahoo.com

(Received 12 Oct 2015; accepted 20 Feb 2016)

Abstract

Background: We aimed to assess the high-risk group for suicide using different classification methods includinglogistic regression (LR), decision tree (DT), artificial neural network (ANN), and support vector machine (SVM).

Methods: We used the dataset of a study conducted to predict risk factors of completed suicide in Hamadan Province, the west of Iran, in 2010. To evaluate the high-risk groups for suicide, LR, SVM, DT and ANNwere performed. The applied methods were compared using sensitivity, specificity, positive predicted value, negative predicted value, accuracy and the area under curve. Cochran-Q test was implied to check differences in proportion among methods. To assess the association between the observed and predicted values, Ø coefficient, contingency coefficient, and Kendall tau-b were calculated.

Results: Gender, age, and job were the most important risk factors for fatal suicide attempts in common for four methods. SVM method showed the highest accuracy 0.68 and 0.67 for training and testing sample, respectively. However, this method resulted in the highest specificity (0.67 for training and 0.68 for testing sample) and the highest sensitivity for training sample (0.85), but the lowest sensitivity for the testing sample (0.53). Cochran-Q test resulted in differences between proportions in different methods ($P<0.001$). The association of SVM predictions and observed values, Ø coefficient, contingency coefficient, and Kendall tau-b were 0.239, 0.232 and 0.239, respectively.

Conclusion: SVM had the best performance to classify fatal suicide attempts comparing to DT, LR and ANN.

Keywords: Suicide, Support vector machine, Neuralnetworks, Logistic regression, Decision tree, Classification

Introduction

Suicide is an important health challengeand one of the main leading causes of premature death worldwide (1, 2). Suicide rate is among the top three causes of death in people aged 15-44 yr old (3). The rate of suicide related death is one million, annually. Suicide causes 1.53 million deaths by 2020 (4). The rate of suicide ideation is reported to be highest among elderly (1). Completed suicide is 5% to 10% of suicide attempts including one attempt in every 3 and one death from suicide in every 40 sec (5).

According to the report of Iranian Ministry of Health and Medical Education in 2004, suicide is the 13th cause of death (6) and the second cause of death from external causes of morbidity(7), the rate of suicide attempt is 3 times in women,

while, the rate of completed suicide is 4 times more common in men (8). Among Iranian population at 2012, 3216 suicide attempts occurred including 6 suicide attempts per 100,000 populations with the highest rate in the second decade of life (9, 10).

Classification is defined as an approach to determine a class for a new object applied using different methods such as data mining (machine learning) techniques (11). Decision tree (DT), k-nearest neighbor, logistic regression (LR), naive Bayes, C4.5, support vector machine (SVM) and linear classifier are among conventional classification methods (12-14).

Classification methodsinclude two main steps: First, a training sample of the dataset is determined randomly to find the model and the second step tests the resulted model (12). According to the kind of dataset, different methods result in different accuracy of prediction. The comparison among the methods can be applied using different criteria such as area under curve, which measures the accuracy of the prediction (13).

Among different classification methods, LR is the most popular predicting the presence or absence of an attribute using covariates. However, DT is preferable when there are predetermined set of attributes, the response is discrete and disjunctive and graphical results are required (15). Artificial neural network (ANN)as a non-linear, flexible, and general tool is capable of dealing with any sort of arbitrary function. Support vector machine (SVM)is a kind of generalized linear models with a classification decision according to the value of the linear combination of features(16, 17).

This study aimed to determine factors putting people at a higher risk of completed suicide using different classification methods including LR, DT, ANN and SVM.

Methods

We used the dataset of a study conducted to predict risk factors of completed suicide in Hamadan Province, the west of Iran, in 2010 (18). The dataset was based on a large population survey conducted in 2010 where all cases of suicide occurring in Hamadan Province from Apr 2008 to Mar 2010 were enrolled. Of 5414 people who attempted suicide, 457 died of suicide.

The presence of missing values was 17% in the dataset; therefore, expectation maximization (EM) algorithm was used for imputation. For this purpose, parameters in the equations imputed the missing values (expectation), then, parameters were updated using all observations including the imputed ones (maximization). This procedure ended at the convergence (19). To assess the fatal suicide, which is a binary variable; using several risk factors, several classification methods were performed. Factors affecting a fatal/non-fatal suicide were included in classification methods. Affective factors were then determined in each method. The data needed to be divided into two sub-sets where training sample of the dataset finds the model and the testing sample tests the resulted model. The test and training samples were composed randomly among cases. The result derived from the learning sample (70% of cases) was then evaluated by utilizing the test sample (30% of cases). The applied methods were compared using sensitivity (SE), specificity (SP), positive predicted value (PPV), negative predicted value (NPV), accuracy (ACC), and the area under curve (AUC).

Logistic Regression: LR is one of the most common applied classification methods in medical data analysis. The model can be written as:

$$log(\frac{\pi}{1-\pi}) = \alpha + \sum_{i=1}^{k} \beta_i x_i$$

In this model,thex_i's are the covariates to classify the response and the β_i's are the regression coefficients. The logit $log(\frac{\pi}{1-\pi})$ indicates the odds ratio of classifying the response in category one than zero.

Artificial Neural Network: This method is an information-processing tool based on human brain performance. Among different ANN models, multilayer perceptron (MLP) is the most com-

mon used method, which includes layers as input, output, and hidden with nodes in each layer. An activation function transforms the data in each layer to the latter one by introducing a degree of non-linearity. Input layer consist of all risk factors affecting the result of suicide, here including 6 variables. The response variable is shown in the output layer with two nodes as the possible outcomes for suicide attempts. To find the best performance of the network, a complicated non-linear mapping between input and output layers is found using the number of nodes determined empirically in the hidden layer (20).

Support Vector Machine: A mapping function whether a classification or regression function is used in SVMs. To classify the result of suicide, a non-linear kernel function is used in order to transform the input data to a high-dimensional space where the input data can be separated as well. Radial basis function (RBF) kernel consists of two parameters trading off misclassification of training sample against simplicity of the decision surface (cost parameter) and to evaluate the influence degree of a training sample. Choosing the kernel function as well as the parameters, acclaims SVM as a flexible method, which the ability of the user can make the results more appealing. Using maximum-margin hyper planes, the classes will be best separated in the data. By contrasting two parallel hyper planes on each side of the separating hyper plane, the minimum generalization error will be achieved when the distance between the hyperplanes takes place (21).

Decision Tree

The DT can be applied when the aim of the research is to identify or discriminate high-risk subjects. Three components are included in DT: decision nodes, branches, and leaves. The direction begins at the node and extends to the leaf, which connects the features. The tree is a disjunction of these connections and these disjunctions separate the branch population into sets with the same likelihood of events. At each stage, the disjunctions cause the highest possible predictive power. The graphical feature presentation makes ease of interpretation and allowing to different alternatives (15).

To check the adequacy of the models, indices such as sensitivity, specificity, diagnostic accuracy (DA), positive predictive value (PPV), negative predictive value (NPV), and the area under curve (AUC) were calculated using the observed data as the gold standard. The Cochran-Q test was used to check differences in proportion among methods. To assess the association between the observed and predicted values several statistics were measures such as Ø coefficient, contingency coefficient, and Kendall tau-b.

Results

Of 5414 people who attempted suicide 50.8% were male, 53.7% were married, 92.8% had no history of suicide, 47.3% and aged between 20 to 29 yr, 8.4% (457 subjects) died of suicide. The mean age of subjects was 26.3 yr (25.3 yr in females and 27.3 yr in males) ranged from 10 to 90 yr. To identify the risk factors affecting completed suicide, LR, SVM, DT and ANN were performed to the data.

The test and train samples were composed of 1626 (30%) and 3788 (70%) cases, respectively. The test sample evaluated the results from training sample. The factors such as gender, job, age, education, marital status, and history of suicide attempt were considered as the explanatory variables for the performed methods.

Completed suicide was significantly associated with gender ($P<0.0001$) and age ($P<0.05$) in the LR model. Accordingly, males were 8.55 times more kill themselves by suicide than women. Those aged between 20-29 yr old was 3.14 times more likely to die from suicide than those aged 10-19 yr (Table 1).

Among several ANN models, the best model included one hidden layer and six hidden nodes. Hyperbolic tangent and softmax were the activation functions for hidden and output layers, respectively. The importance of the variables is shown in Fig. 1 presented by scores using sensitivity analysis.

Table 1: Logistic regression model results

Variables	Adjusted Odds Ratio(95% CI)	*P* value
Gender		
Female	1.00	
Male	8.55 (3.90, 18.78)	<0.0001
Age group (yr)		
10-19	1.00	
20-29	1.68 (1.04, 2.72)	0.033
30-39	3.14 (1.80, 5.50)	0.001
40-49	3.09 (1.60, 5.98)	0.001
50-59	5.72 (2.77, 11.83)	0.001
60-69	6.50 (2.51, 16.87)	0.001
70-79	4.90 (1.67, 14.43)	0.004
80-90	6.93 (1.22, 39.51)	0.029

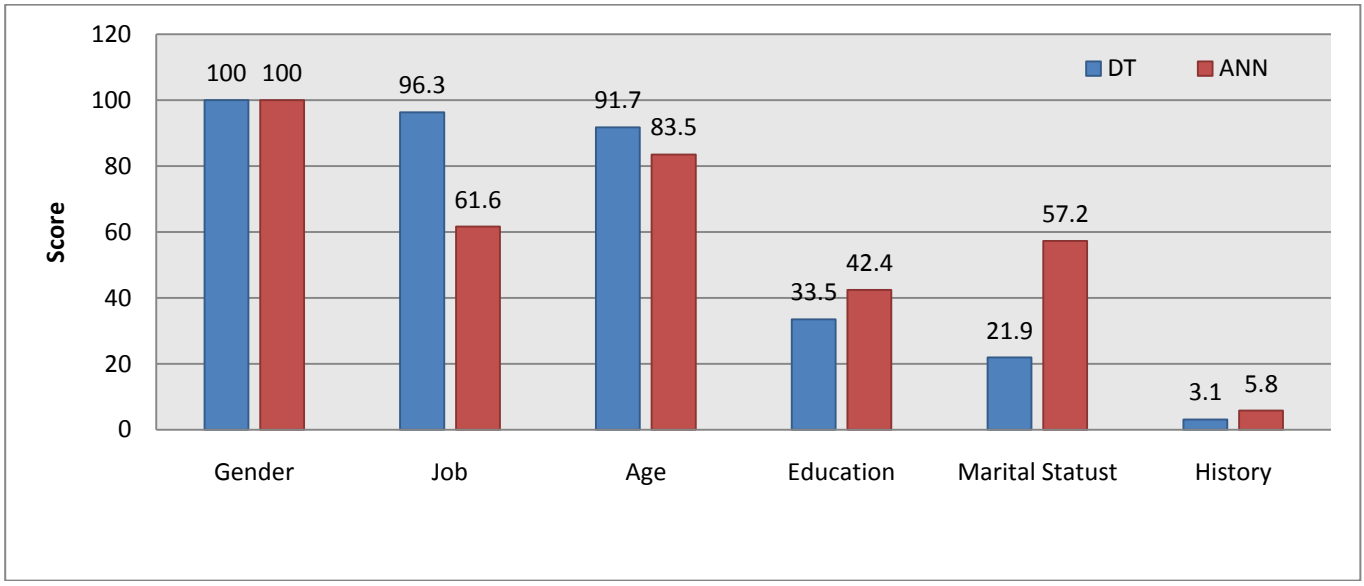

Fig. 1: The normalized importance of the variables in decision tree and artificial neural network

To perform the SVM model, Gaussian radial basis function was used as the best non-linear kernel function for classifying the successful attempts. This method showed a kernel parameter (sigma) of 0.24, a cost parameter of 5, and 2178 support vectors as the estimated parameters of the kernel function. In training sample, the weight assigned to the SVM method was 11 for completed suicide and one for suicide attempt.

The decision tree analysis resulted in 8 rules. In each node, the probability of completed suicide is presented according to the condition mentioned in its corresponding branch (Fig. 2). Moreover, the sensitivity analysis was performed (Fig. 2).

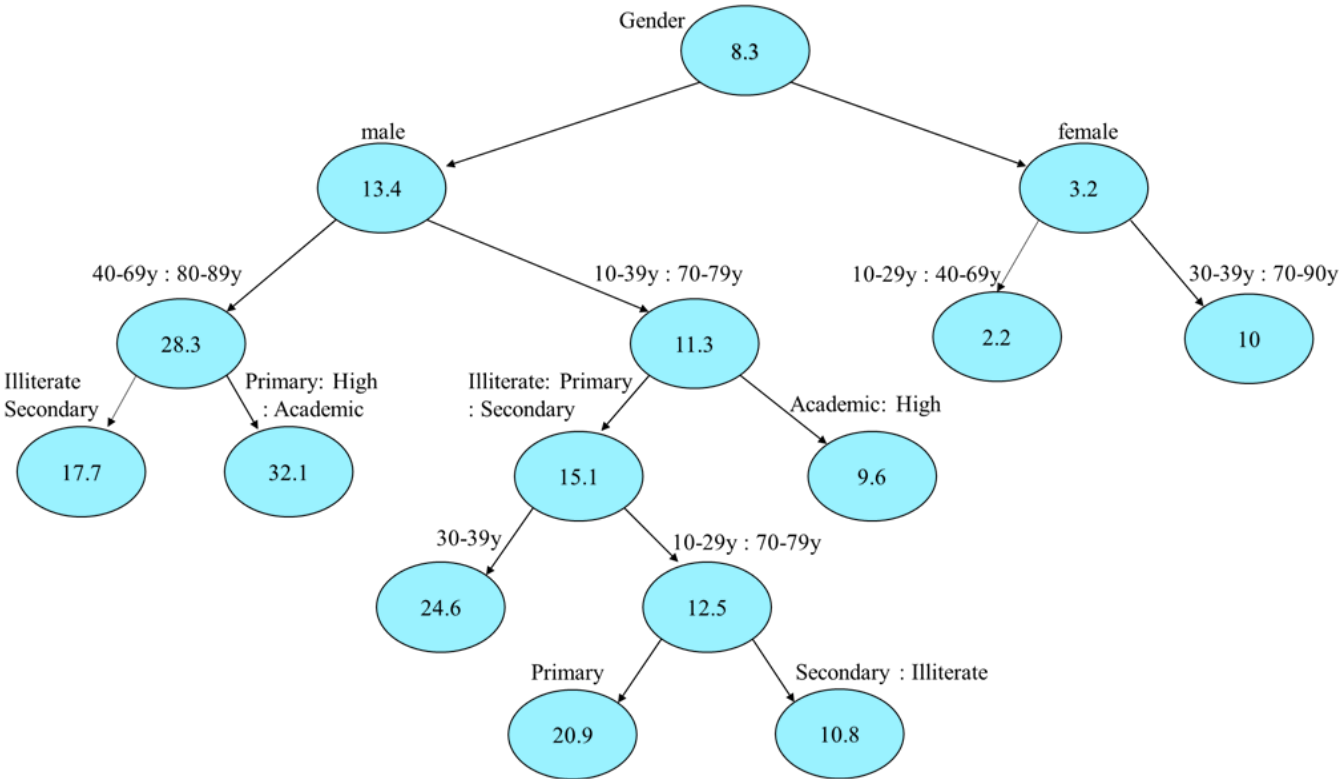

Fig. 2: The classification tree with the probabilities of success for suicide attempts in each node

A comparison of sensitivity, specificity, positive probability value, negative probability value, accuracy and the area under curve for training and testing sets of classification methods are shown in Table 2 and Fig. 3. Cochran-Q test resulted in differences between proportions in different methods (P<0.001). Multiple comparison adjusted for significance level was performed using McNemar test which showed a significant difference in proportions of any two methods (P<0.001). To evaluate the association of the method predictions and observed value of suicide attempts, Ø coefficient, contingency coefficient, and Kendall tau-b were performed which resulted in the best performance of SVM in comparison to others (Table 3).

Table 2: Comparison of classification techniques

Model	Training Sample				Testing Sample			
	LR	DT	ANN	SVM	LR	DT	ANN	SVM
Sensitivity	0.72	0.88	0.74	0.85	0.73	0.85	0.75	0.53
Specificity	0.63	0.46	0.60	0.67	0.65	0.46	0.60	0.68
Positive predictive value	0.15	0.13	0.14	0.19	0.16	0.13	0.15	0.14
Negative predictive value	0.96	0.97	0.96	0.98	0.96	0.97	0.96	0.94
Accuracy	0.64	0.50	0.62	0.68	0.65	0.49	0.62	0.67

LR: logistic regression, DT: decision tree, ANN: artificial neural network, SVM: support vector machine

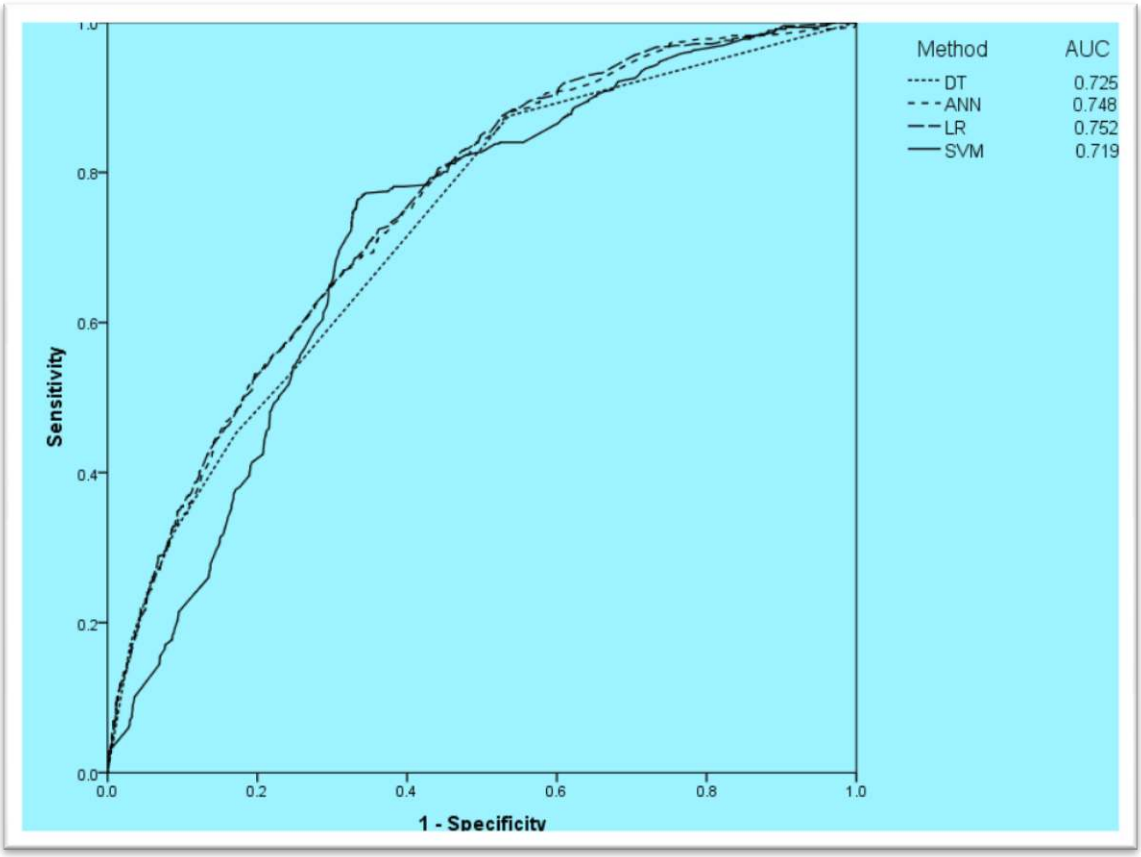

Fig. 3: The area under curve for the performed methods

Table 3: The association of performed methods with observed values

Association Coefficient	Method			
	LR	DT	ANN	SVM
Ø coefficient	0.206	0.190	0.197	0.239
Contingency coefficient	0.202	0.187	0.193	0.232
Kendall tau-b	0.206	0.190	0.197	0.239

LR: logistic regression, DT: decision tree, ANN: artificial neural network,SVM: support vector machine

Discussion

In this study, gender was recognized as a significant risk factor for predicting completed suicide so did for age, and educational level in different applied methods. Despite not being significant, marital status and history of suicide were the less important variables predicting completed attempts in DT and ANN.

In a study, the risk ratio of completed suicide was reported7.1 for males comparing to females. Furthermore, age of 21-30 yr was associated with the highest rate of completed suicide. Classifying educational level into three categories (low-intermediate-high), the intermediate educated cases associated with the highest ratio of completed suicide. Moreover, they showed that married cases were more prone to die comparing to the single people (22). In other study, men selected high-risk methods of suicide and suicide related mortality rate was higher in men (23). The rate of completed suicide was 2.5 in males compared to females. Moreover, age, occupation, ma-

rital status, and educational levelwere reported as significant risk factor for completed suicide (24). Age was affecting significant variable for suicide with the highest age-specific suicide after 45 yr old in both Japan and South Korea (25). Marital status was a significant risk factor resulted from the odds of completed suicide 2.77 for married cases (26).

This study showed that among four statistically different classification methods including SVM, LR, DT and ANN for this data, SVM had the best performance in classifying the risk factors associated with completed suicide. In spite of the least sensitivity in the testing sample and presence of unbalanced data (8.3% fatality in training sample), SVM had the outperformance among mentioned methods and indicated the highest association between SVM predicted and the observed values as well as the highest accuracy. Although, the assigned weight for the training sample was the best choice among all other assignments, the testing sample did not result in the same shape as the training sample because of different rates for fatal to non-fatal suicide attempts.

To compare seven classification methods based on sample size and type of attributes, a sufficient number of records DT, SVM, k-nearest neighborhood and C4.5 obtained a higher area under curve than LR, naive Bayes, and linear classifier (27). In another study, several classification methods including linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests were used to predict the dementia. Despite of the highest specificity and lowest sensitivity of SVM, this method had the highest accuracy among all different methods (28).

The functioning of SVM based methods against ANN assessed in a study of analytical chemistry. They recommended that the SVM-based approach for practical application according to the robustness (29). Conducting an empirical comparison between SVM and ANN, for classifying document-level sentiment, ANN showed a better statistically significant prediction comparing to SVM, even on the context of unbalanced data (30).

Finding predictive models for pre-operative diagnosis of rotator cuff tear, ANN and LR were compared. The study resulted in a higher predictive accuracy of ANN than LR (31). In a study, ANN and DT were applied to predict hospital charge for gastric cancer patients. An outperformance for ANN was found compared to DT where the mean absolute errors for the former were less than the latter one (32). To classify the magnetic resonance imaging data in Alzheimer's disease, different classification methods DT, ANN, SVM and orthogonal projections of latent structures (OPLS) were compared. Although there was no statistical difference among several methods, SVM and OPLS outperformed slightly than DT and ANN (33).

In a study that assessed differences between SVM and LR, concluded that SVM achieves a better performance in comparison to LR when fewer variables are included (34). To determine statistically the sex from craniometrists, three different methods LR, SVM and linear discriminant analysis were compared. The study showed a better reliability existed for males than females using all the methods while the results for SVM had to be developed. Moreover, they found that LR was much more feasible than SVM according to the choice about the kernel function and the parameters (35). Predicting the hospital mortality in critically ill patients with hematological malignancies, SVM and LR were applied. The comparing results were not statistically significant even though LR was resulted in a better predictive accuracy comparing to SVM. Moreover, to predict the model using SVM, only 4 variables were needed, whereas this number was 7 and 8 for LR (36).

Conclusion

Despite its limitations such as missingness in the data imputed, this study compared four different methods suggesting SVM as the best classifier model, which may help the policymakers determining suicide risk factors. This may reduce the amount of suicide attempts and its social consequences. SVM had a better performance in classi-

fying risk factors of completed suicide than other classification methods including DT, k-nearest neighbor, LR, naive Bayes, C4.5, SVM, and linear classifier. The flexibility of this method according to several choices for parameters and kernel function can make it as the first choice method for classification of such data.

Ethical considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgments

We appreciated Vice-Chancellor of Research and Technology, Hamadan University of Medical Sciences, for financial support of this work. The authors declare that there is no conflict of interests.

References

1. Simon M, Chang E, Zeng P, Dong X (2013). Prevalence of suicidal ideation, attempts, and completed suicide rate in Chinese aging populations: a systematic review. *Arch Gerontol Geriatr*, 57(3):250-6.
2. Organization WHO (2009). Preventing suicide: a resource for police, firefighters, and other first line responders. *www.who.int/mental_health/resources/preventingsuicide/en/*
3. Reddy M (2010). Suicide incidence and epidemiology. *Indian J Psychol Med*, 32(2):77.
4. Bertolote JM, Fleischmann A (2009). *Oxford Textbook of Suicidology and Suicide Prevention.In: A global perspective on the magnitude of suicide mortality*. Oxford University Press, UK. pp.: 91-8.
5. Organization WHO (2014) . Preventing suicide: a global imperative. Available from: *www.who.int/mental_health/suicide-prevention/world_report_2014/en/*
6. Naghavi M, Jafari N (2007). *Mortality of 29 provinces-2004*. Tehran: Ministry of Health and Medical Education:(Persian).
7. Poorolajal J, Esmailnasab N, Ahmadzadeh J, Motlagh TA (2012). The Burden of Premature Mortality in Hamadan Province in 2006 and 2010 Using Standard Expected Years of Potential Life Lost: A Population-based Study. *Epidemiol Health,* 34:e2012005.
8. Memari AM, Ramim T, Amirmoradi F, Khosravi K, Godarzi Z (2006). Causes of suicide in married women. *Hayat*. [Research], 12(1):47-53.
9. Mohammadkhani P (2004). Epidemiology of Suicide thoughts and attempts in young girls from highrisk regions in Iran. *Social Wellfare*, 13(4):(persian).
10. Hajebi A, Ahmadzad Asl M, Zaman M, Naserbakht M, Mohammadi N, Davoudi F, et al (2011). Designing a Registration System for Suicide in Iran. *Ira J Psychiatr Clin Psychol,* 17(2):106-9.
11. Han J, Kamber M (2006). *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann.
12. Pang-Ning T, Steinbach M, Kumar V (2006). *Introduction to data mining*. Pearson.
13. Kantardzic M (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.
14. Witten IH, Frank E (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
15. Jenhani I, Amor NB, Elouedi Z (2008). Decision trees as possibilistic classifiers. *Int J Approx Reason,* 48(3):784-807.
16. Moreno JJM, Pol AP, Gracia PM (2011). Artificial neural networks applied to forecasting time series. *Psicothema*, 23(2):322-9.
17. Delen D, Oztekin A, Tomak L (2012). An analytic approach to better understanding and management of coronary surgeries. *Decis Support Syst*, 52(3):698-705.
18. Amiri B, Pourreza A, Rahimi Foroushani A, Hosseini SM, Poorolajal J (2012). Suicide and associated risk factors in Hamadan Province, West of Iran, in 2008 and 2009. *J Res Health Sci*, 12(2):88-92.
19. Lin TH (2010). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Qual Quant*, 44(2):277-87.

20. Ayat S, Farahani HA, Aghamohamadi M, Alian M, Aghamohamadi S, Kazemi Z (2013). A comparison of artificial neural networks learning algorithms in predicting tendency for suicide. *Neural Comput Appl*, 23(5):1381-6.
21. Auria L, Moro RA (2008). Support vector machines (SVM) as a technique for solvency analysis. DIW Berlin Discussion Paper No. 811. Available at SSRN: http://ssrn.com/abstract=1424949.
22. Ghaleiha A, Khazaee M, Afzali S, Matinnia N, Karimi B (2009). An annual survey of successful suicide incidence in Hamadan, western Iran. *J Res Health Sci*, 9(1):13-6.
23. Cibis A, Mergl R, Bramesfeld A, Althaus D, Niklewski G, Schmidtke A, et al (2012). Preference of lethal methods is not the only cause for higher suicide rates in males. *J Affect Disord*, 136(1):9-16.
24. Poorolajal J, Rostami M, Mahjub H, Esmailnasab N (2015). Completed suicide and associated risk factors: a six-year population based survey. *Arch Iran Med*, 18(1):39-43.
25. Kim SY, Kim M-H, Kawachi I, Cho Y (2011). Comparative epidemiology of suicide in South Korea and Japan: effects of age, gender and suicide methods. *Crisis*, 32(1):5-14.
26. Zhang J (2010). Marriage and suicide among Chinese rural young women. *Soc Forces*, 89(1):311-26.
27. Entezari-Maleki R, Rezaei A, Minaei-Bidgoli B (2009). Comparison of classification methods based on the type of attributes and sample size. *J Converg Inform Technol*, 4(3):94-102.
28. Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Res Notes*, 4(1):299.
29. Balabin RM, Lomakina EI (2011). Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. *Analyst*, 136(8):1703-12.
30. Moraes R, Valiati JF, Neto WPG (2013). Document-level sentiment classification: An empirical comparison between SVM and ANN. *Expert Syst Appl*, 40(2):621-33.
31. Lin CC, Chan HH, Huang CY, Yang NS (2014). Predictive Models for Pre-Operative Diagnosis of Rotator Cuff Tear: A Comparison Study of Two Methods between Logistic Regression and Artificial Neural Network. *Appl Mech Mater*,595:263-8 .
32. Wang J, Li M, Hu Y-t, Zhu Y (2009). Comparison of hospital charge prediction models for gastric cancer patients: neural network vs. decision tree models. *BMC Health Serv Res*, 9(1):161.
33. Aguilar C, Westman E, Muehlboeck J-S, Mecocci P, Vellas B, Tsolaki M, et al. (2013). Different multivariate techniques for automated classification of MRI data in Alzheimer's disease and mild cognitive impairment. *Psychiatry Res*, 212(2):89-98.
34. Salazar DA, Vélez JI, Salazar JC (2012). Comparison between SVM and logistic regression: Which one is better to discriminate? *Revista Colombiana de Estadística*, 35(2):223-37.
35. Santos F, Guyomarc'h P, Bruzek J (2014). Statistical sex determination from craniometrics: Comparison of linear discriminant analysis, logistic regression, and support vector machines. *Forensic Sci Int*, 245:204. e1-. e8.
36. Verplancke T, Van Looy S, Benoit D, Vansteelandt S, Depuydt P, De Turck F, et al (2008). Support vector machine versus logistic regression modeling for prediction of hospital mortality in critically ill patients with haematological malignancies. *BMC Med Inform Decis Mak*, 8(1):56.