International Journal of Mechanical Engineering and Technology (IJMET)
Volume 9, Issue 4, April 2018, pp. 293–301, Article ID: IJMET_09_04_034
Available online at http://www.iaeme.com/ijmet/issues.asp?JType=IJMET&VType=9&IType=4
ISSN Print: 0976-6340 and ISSN Online: 0976-6359

SUICIDAL BEHAVIOR PREDICTION USING DATA MINING TECHNIQUES

Alina Joseph
Department of computer science, Christ University, Bangalore, Karnataka, India

Ramamurthy B
Department of computer science, Christ University, Bangalore, Karnataka, India

ABSTRACT

Background: *Suicide is one of the most serious public health problem that has affected many people. After being recognized as a public health priority by the WHO (World Health Organization) various studies have been going out for its prevention. It is one of a serious health problem and it is preventable and can be controlled by proper interventions and study in the field. The objective of the study is to create a prediction model for individuals who are at higher risk of suicide by studying the different predictors of suicide such as depression, anxiety, hopelessness, stress etc. by using data mining techniques for the prediction.*

Study Design*: Systematic review and predictive analysis for suicidal behavior.*

Methods: *The research applies data mining process to analyze the data and on the basis of analysis create the model to predict suicidal behaviors present in the individual. Prediction is done on the basis of analysis of risk factors which are Depression, anxiety, hopelessness, stress, or substance misuse which is calculated by using various psychological measures such as Beck hopelessness scale,suicidal ideation subscale,hospital anxiety and depression scale.Various data mining algorithms for classification are compared for the purpose of prediction.*

Results: *Six different data mining classification algorithms which are namely Classification Via Regression, Logistic Regression. Random Forest, Decision Table, SMO are compared and Classification Via Regression was found to the highest accuracy in prediction.*

Conclusions: *Data required for the development of such a model requires continuous monitoring and needs to be updated on a periodic basis to increase the accuracy of prediction.*

Keywords: Data Mining, Classification, Prediction, Suicide, Depression, Risk Factors.

Cite this Article: Alina Joseph and Ramamurthy B, Suicidal Behavior Prediction Using Data Mining Techniques, International Journal of Mechanical Engineering and Technology, 9(4), 2018, pp. 293–301.
http://www.iaeme.com/IJMET/issues.asp?JType=IJMET&VType=9&IType=4

1. INTRODUCTION

Suicide has been defined as the act of deliberately imposing one's own death. According to WHO around 800,000 people die due to suicide each year and even more number of them attempt suicide. It is ranked among the top causes of death worldwide [1]. It is the second largest cause of suicide among the age group 15-29. It is the 10^{th} leading cause in the US. Risk factors embrace mental disorders, Depression, manic-depressive illness, dementia praecox, temperament disorders, alcohol dependency, or drug misuse. Other areas include impulsive acts due to stress like from monetary issue, relationship troubles, or due to bullying. Antecedent tries of suicide have a higher risk for future tries. More than one lakh lives are lost each year because of suicide in India. Inside the most recent couple of years there has been immense increment in the suicide rates. The rates were the same in 1975 and 1985 around; from 1985 to 1995 there is an ascent of 35% and from 1995 to 2005, the expansion was 5%. Nevertheless, the male-ladylike proportion has been steady at around 1.4 to no less than one 1. There is a wide variety in suicide rates inside the national nation. Kerala, Karnataka, Andhra Tamil and Pradesh Nadu possess a rate of suicide greater than 15 whereas Punjab, Uttar Pradesh, Jammu and Bihar and Kashmir, the rate of suicide is less than 3. This adjustable pattern has been steady for the last twenty years. Higher literacy, a much better reporting program, lower exterior aggression, higher socioeconomic position and higher expectations will be the feasible explanations for the bigger suicide rates in the southern states [2]. 37.8% out of the aggregate suicides that happen in India are endeavoured by those that are underneath the age of thirty years. The near equivalent suicides rates of young people and ladies and consistently limit man: female proportion indicates that considerably more Indian ladies die by suicide than their Western partners. Harming (34.8%), hanging (31.7%) and self-immolation (8.5%) had been the basic techniques used to commit suicide [3]. Data mining has been proven to be useful in the field of medicine. It has been used to in psychiatry to estimate risks associated with suicide using the information available in the electronic medical records [4]. Frequent warning signs of their distress are given by youth who are contemplating suicides. The people who are at a key position to pick up these signs are Parents, teachers, and friends and get help. The major concerns is how the warning signs can be observed and predict the students at high risk. For accurate prediction, the data needs to be collected on a periodic basis [5]. In an examination they have built up a linguistic driven expectation models which were utilized to evaluate the danger of suicide. These models were created from unstructured clinical notes taken from a national specimen of U.S. Veterans Administration (VA) restorative records. They made three gatherings first the veterans who submitted suicide, second the veterans who utilized psychological wellness benefits and did not confer suicide, and third the veterans who did not utilize emotional wellness benefits and did not confer suicide amid the perception time frame. Each gathering containing 70 instances. Single keywords and multi-word phrases were generated from the clinical notes, and prediction models were constructed based on a genetic programming framework using a machine learning algorithm. The resulting inference accuracy was consistently 65% or more. The data therefore suggests that computerized text analytics can be applied to unstructured medical records to estimate the risk of suicide. The subsequent framework could enable clinicians to conceivably screen apparently vulnerable patients at the essential care level, and to persistently assess the suicide chance among mental patients. [6] In the paper, they portray a strategy of self-organizing maps (SOM) for finding the most pertinent factors notwithstanding when their connection to self-destructive conduct is firmly nonlinear. The investigation of the factors engaged with suicidal behaviour is imperative from a social, therapeutic, and practical perspective. Given the high number of potential factors of intrigue, a huge populace of subjects was broken down with a specific end goal to get definitive outcomes. They have connected the technique to a dataset of more than 8,000 subjects and 600 factors and found

four gatherings of factors engaged with suicidal behaviour. Based on the outcome the suicide attempters are divided into four main categories of risk factors namely mental disorders, liquor addiction, impulsivity, and childhood abuse [7].

2. METHODS

The data is obtained from UK data archive. The data consist the details of 267 Para suicidal patients (aged 16 years or older) who had been seen by the Liaison Psychiatry benefit, Edinburgh. Participants completed the various psychological measures and the given information were recorded. Table 1 contains the list of attributes. Suicidal Ideation was assessed using the suicidal ideation subscale of the Suicide Probability Level it consists of suicidal cognitions, negative affect, and presence of a suicide plan e.g., "Personally i think that people would be better off if I were dead". The Beck Hopelessness Scale assesses pessimism for future years e.g., "I anticipate the future with hope and enthusiasm". The Hospital Anxiety and Depression Scale were used to assess anxiety and depression e.g. "Worrying thoughts {go through my mind". The values hence recorded is then measured on a numerical scale and kept in the database.

Table 1 List of attributes in the dataset

Variable	Variable Label
Age	Age
Sex	Sex
Intention	Intention to kill oneself?
Marital	Marital Status
suicid	Suicide Ideation
Bhs	Hopelessness
poswk	Positive Week
posyr	Positive Year
negwk	Negative Week
negyr	Negative Year
postot	Positive Total
Negtot	Negative Total
Anx	anx1
Dep	Depression
selfm	Self-oriented perfectionism
otherm	Other-oriented perfectionism
socialm	Socially prescribed perfectionism
Diseng	Goal Disengagement
Reengag	Goal Reengagement
BIS	BIS
BASDrive	Bas Drive
BASFUN	Bas Fun seeking
BASReward	Bas reward
Attempts	Suicide attempts

The proposed system applies the process of data mining to be able to analyze the data and on the basis of analysis provide methods to predict suicidal behaviors present. Finding the right data mining technique for prediction by evaluating the different learning methods in WEKA. Depression and anxiety have already been found to be the most effective predictors

of suicidal behavior. Proposed model uses these predictors along with other psychological measures in order to predict the suicidal behavior.

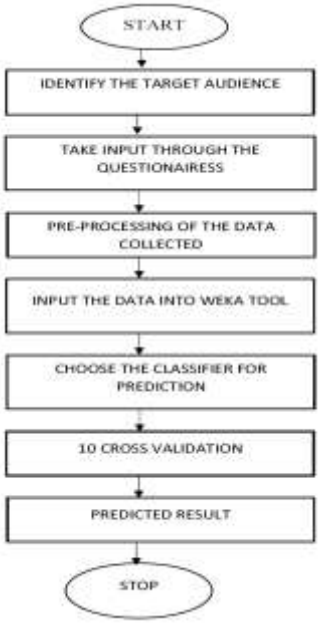

Figure 1 Proposed model flowchart

3. RESULTS

The tool used for implementation is WEKA which is a machine learning tool written in java. The dataset consists of different psychological measures converted into nominal scale using SPSS measurement. Attempts are chosen to be the class attribute for prediction. Different classifiers in WEKA were applied to the dataset and their performance was analyzed using 10-fold cross-validation. The classifiers used were:

ClassificationViaRegression

Regression approaches are requested classification under this classifier. Single regression model is built for every single instance of the class. Utilized when the dependent variable can be binary. Table 2 shows the detailed accuracy in WEKA [8].

Detailed Accuracy by Class:

Table 2 ClassificationViaRegression Accuracy

			Weighted Average
TP Rate	0.916	0.382	0.764
FP Rate	0.618	0.084	0.466
Precision	0.788	0.644	0.747
Recall	0.916	0.382	0.764
F-Measure	0.847	0.479	0.743
MOC	0.359	0.359	0.359
ROC Area	0.765	0.765	0.765
PRC Area	0.869	0.561	0.781
Class	Yes	No	

Logistic Regression

Like all regression analyses, the logistic regression can be used for prediction. Logistic regression is used to explain the data and explain the relationship between one dependent binary adjustable and a number of nominal, ordinal, ratio-level or interval independent variables. Table 3 shows the detailed accuracy in WEKA.

Detailed Accuracy by Class:

Table 3 Logistic Regression Accuracy

			Weighted Average
TP Rate	0.869	0.382	0.730
FP Rate	0.618	0.131	0.480
Precision	0.779	0.537	0.710
Recall	0.869	0.382	0.730
F-Measure	0.822	0.446	0.715
MOC	0.282	0.282	0.282
ROC Area	0.716	0.716	0.716
PRC Area	0.845	0.507	0.749
Class	Yes	No	

DecisionTable

Used to construct in form of all table almost all possible situations which a decision might encounter and also to specify which thing to do in each one of these circumstances. A matrix representation of the logic of a decision. Specifies the feasible circumstances and the resulting actions. Table 4 shows the detailed accuracy in WEKA.

Detailed Accuracy by Class:

Table 4 DecisionTable Accuracy

			Weighted Average
TP Rate	0.880	0.368	0.734
FP Rate	0.632	0.120	0.486
Precision	0.778	0.549	0.713
Recall	0.880	0.368	0.734
F-Measure	0.826	0.441	0.716
MOC	0.285	0.285	0.285
ROC Area	0.661	0.661	0.661
PRC Area	0.813	0.454	0.711
Class	Yes	No	

RandomForest

Collaborative learning way of classification, regression and alternative activities that functions by creating a gathering of decision trees at training time and generating the class this is the approach to the classes known as classification or mean prediction referred to as regression for each single tree. Table 5 shows the detailed accuracy in WEKA [9]

Detailed Accuracy by Class:

Table 5 RandomForest Accuracy

			Weighted Average
TP Rate	0.911	0.276	0.730
FP Rate	0.724	0.089	0.543
Precision	0.760	0.553	0.701
Recall	0.911	0.276	0.730
F-Measure	0.829	0.368	0.698
MOC	0.242	0.242	0.242
ROC Area	0.679	0.679	0.679
PRC Area	0.804	0.476	0.710
Class	Yes	No	

Sequential Minimal Optimization (SMO)

Implementation of SVM in WEKA is done with the help of SMO. Extremely popular machine learning technique Controls complexity and over fitting issues, so that it works well about an array of practical problems. Due to this, it can deal with high dimensional vector areas, making feature selection less essential. Table 6 shows the detailed accuracy in WEKA.

ClassificationViaRegression was found to have the highest performance compared to others

Detailed Accuracy by Class:

Table 6 SMO Accuracy

			Weighted Average
TP Rate	0.911	0.342	0.749
FP Rate	0.658	0.089	0.496
Precision	0.777	0.605	0.728
Recall	0.911	0.342	0.749
F-Measure	0.829	0.437	0.724
MOC	0.311	0.311	0.311
ROC Area	0.627	0.627	0.627
PRC Area	0.771	0.394	0.664
Class	Yes	No	

Model performance chart depicts the performance of all the 5 models. It can be seen through the chart that ClassificationViaRegression has better performance compared to the other 4 models.

X axis: False Positive Rate

Y axis: True Positive Rate

× Logistic (CLASS: YES)

+ ClassificationViaRegression (Class: Yes)

° SMO (Class: Yes)

▵ DecisionTable (Class: Yes)

▿ RandomForest (Class: Yes)

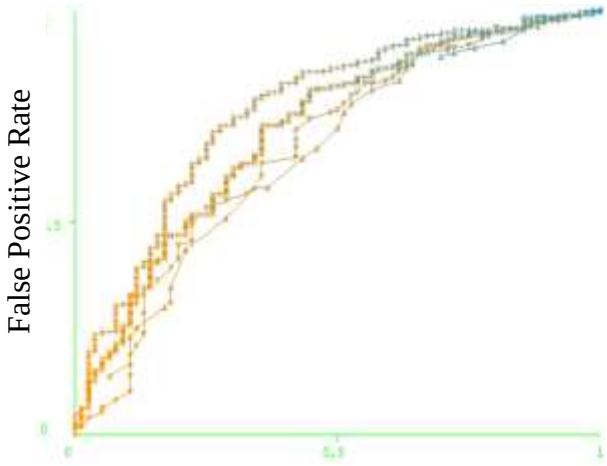

Figure 2 Model Performance chart

4. DISCUSSION

The paper, describes the entire process to automatically gather the suspect tweets according to a vocabulary of topics a suicidal person uses to convey his feelings. The methodology consists of first defining a vocabulary by identifying what people who have suicidal tendency usually talk about like depression, feeling of loneliness, bullying etc. The tweets which indicate risky suicidal behaviour are classified and captured by simple classification methods such as JRIP, IBK, IB1, J48, Naive Bayes, SMO using 10 cross validation and loo (leave one out) validation. They have observed Naïve Bayes to have the better performance compared to other classifiers. After the suspected tweets are captured they have been given to a web interface implemented for psychiatrists enabling them to analyse the suspect tweets and then consult the profiles associated with the tweets. [10]

The study conducted a decision tree analysis of data mining using the Answer Tree 3.0 program. In order to test the fitness of the model Gain charts which is the ratio of the target category in a particular node and Risk chart which is the probability of the prediction model through testing data. Training data can be assumed as generalized if there is no difference in risk estimates of the model between training data and testing data. They came up with a conclusion that school performance record and depression were significant variables to predict suicide attempts [11].

They have built a predictive model using the data mining analysis methods. The study is conducted on 707 Chilean mental health patients by analysing three hundred and forty-three variables from five questionnaires. The used six data mining techniques using the R statistical tool and found that support vector machines (SVM) gave better results compared to the others. Out of the three hundred forty-three variables only twenty-two variables are selected by the model and those variables are then used to build a clinical tool which can be used to predict the suicide risk in a person. They found that the risk factors are related to individual satisfaction, belief in one's capabilities and the reason to live [12].

The goal of this article was to decide if longitudinal historical information, generally reachable in electronic health record (EHR) frameworks, can be used to foresee patients' future risk of self-destructive conduct. They have made Bayesian model. EHR information from a sizable social insurance information source spreading over 15 years (1998-2012) of inpatient and outpatient arrangements were used to foresee long haul archived self-destructive conduct. Display effectiveness was assessed reflectively utilizing a fair testing set. The most

grounded indicators found by the model included both surely understood and less regular hazard factors, demonstrating information driven methodology can yield more broad risk profiles. EHR, electronic risk screening strategies may improve prediction beyond what is be possible by an individual clinician [13].

The paper consists of an implementation of a counselling system in order to predict the presence of suicidal tendencies and depression among the students. They analysed the different observable and non-observable warning signs which includes classroom behaviour, interpersonal communication etc. and then data mining algorithms are applied to generate the result. Data for the study is collected from different students and then on the basis of the results of the data mining algorithm they have designed a gradation system. The gradation system thus designed was used to identify the students who are at high risk. [14].

5. CONCLUSION

The paper analyses the various data mining classification algorithms for the prediction of suicidal behaviour present in an individual. ClassificationViaRegression was found to have higher efficiency compared to the other algorithms used. Suicide prevention is of higher priority on the global general public health agenda, the quality and option of data available on suicide and suicide attempts is poor. Improved surveillance and monitoring of suicide and suicide attempts is necessary for effective suicide avoidance strategies. This includes essential registration of suicide, hospital-based registries of suicide attempts and representative surveys collecting details about self-reported suicide attempts nationally. With proper availability of data prediction accuracy also will increase and thus more and effective suicide prevention strategies can be made.

REFERENCES

[1] Nock MK, Borges G, Bromet EJ, ChaCB, Kessler RC, Lee S(2008), Suicide and suicidal behavior. Epidemiologic Reviews (2008), 30(1), pp.133– 154.

[2] Lakshmi Vijayakumar, Indian Research on suicide. Indian J Psychiatry (2010), 52(7), pp. 291-296.

[3] Accidental Deaths and suicides in India. National Crime Records Bureau. Ministry of home affairs. Government of India; 2007.

[4] Tran T, Phung D, Luo W, Harvey R, Berk M, Venkatesh S. An integrated framework for suicide risk prediction. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD (2013), pp1410-18.

[5] Luan, J., Data Mining and Its Applications in Higher Education. New

[6] Directions for Institutional Research (2002), pp. 17–36.

[7] Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, et al. Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes. PLoS .9(1), 2014.

[8] Leiva-Murillo, López-Castromán, Baca-Garcia and Consortium, Characterization of Suicidal Behaviour with Self-Organizing Maps. Computational and Mathematical Methods in Medicine (2013), Volume 2013, pp.1-9.

[9] R. Bal and S. Sharma (2016), Review on Meta Classification Algorithms using WEKA, International Journal of Computer Trends and Technology, vol. 35(1), pp. 38-47.

[10] S. Venkata Lakshmi1 and T. Edwin Prabhakaran, Performance Analysis of Multiple Classifiers on KDD Cup Dataset using WEKA Tool Indian Journal of Science and Technology (2015), 8(17), pp. 1-10.

[11] Abboute A., Boudjeriou Y., Entringer G., Azé J., Bringay S., Poncelet P. Mining Twitter for Suicide Prevention. In: Métais E., Roche M., Teisseire M. (eds) Natural Language Processing and Information Systems. NLDB 2014. Lecture Notes in Computer Science, 8455. Springer, pp. 250-253

[12] Sung Man Bae, Seung A Lee, Seung-Hwan Lee, Prediction by data mining, of suicide attempts in Korean adolescents, national study, Neuropsychiatric Disease and Treatment (2015), 11, pp. 2367-2375.

[13] J. Barros, S. Morales, O. Echávarri, A. García, J. Ortega, T. Asahi, C. Moya, R. Fischman, M. P. Maino, and C. Núñez, Suicide detection in Chile: proposing a predictive model for suicide risk in a clinical sample of patients with mood disorders, Revista Brasileira de Psiquiatria (2017), 39(1), pp. 1–11, 2017.

[14] Y. Barak-Corren, V. M. Castro, S. Javitt, A. G. Hoffnagle, Y. Dai, R. H. Perlis, M. K. Nock, J. W. Smoller, and B. Y. Reis, Predicting Suicidal Behavior from Longitudinal Electronic Health Records, American Journal of Psychiatry (2017),174(2), pp. 154–162.

[15] Omprakash L. Mandge, A Data Mining Tool for Prediction of Suicides among Students. Proceedings of National Conference on New Horizons in IT (2013), pp 178-181.

[16] Kavitha G and Dr. Elango N.M, An Overview of Data Mining Techniques and its Applications. International Journal of Civil Engineering and Technology, 8(12), 2017, pp. 1013-1020.

[17] N.K. Senthil Kumar, M. Uvaneshwari, M. Viswanathan and K. Amsavalli, One-Tier Cache System Applied to Data Mining Techniques to Enhance the Information Security in the Cloud. International Journal of Civil Engineering and Technology, 8(10), 2017, pp. 1709–1717.

[18] R. Lakshman Naik, D. Ramesh, B. Manjula, Instances Selection Using Advance Data Mining Techniques, International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 2, July- September (2012), pp. 47-53

[19] Parag Deoskar, Dr. Divakar Singh, Dr. Anju Singh, Mining Lung Cancer Data and Other Diseases Data Using Data Mining Techniques: A Survey, International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 2, March – April (2013), pp. 508-516

[20] S V Subrahmanyam, M. M. M. Sarcar, Wedm Process Modeling with Data Mining Techniques, International Journal of Advanced Research in Engineering and Technology (IJARET), Volume 4, Issue 7, November - December 2013, pp. 161-169