International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-10 Issue-1, October 2020

Machine Learning Based Prediction of Suicide Probability

Avhishek Biswas, Ananya Talukder, Deep Bhattacharjee, Arijit Chowdhury, Judhajit Sanyal

Abstract*: Many factors have led to the increase of suicide-proneness in the present era. As a consequence, many novel methods have been proposed in recent times for prediction of the probability of suicides, using different metrics. The current work reviews a number of models and techniques proposed recently, and offers a novel Bayesian machine learning (ML) model for prediction of suicides, involving classification of the data into separate categories. The proposed model is contrasted against similar computationally-inexpensive techniques such as spline regression. The model is found to generate appreciably accurate results for the dataset considered in this work. The application of Bayesian estimation allows the prediction of causation to a greater degree than the standard spline regression models, which is reflected by the comparatively low root mean square error (RMSE) for all estimates obtained by the proposed model.

Keywords*: Bayesian model, classification, machine learning, spline regression, suicide prediction.

I. INTRODUCTION

The modern era has witnessed an increase in mental health problems among the general populace, largely due to stress from different socio-politico-economic factors. As a consequence, there has been an extremely significant increase in the number of suicides and attempted suicides in the past few years. Researchers have therefore been trying to understand, identify and predict suicidal behaviour patterns and identify potentially vulnerable people through different techniques. Some researchers have applied text mining techniques to identify suicidal tendencies among war veterans and predict the probability of suicide attempts using their model [3]. Machine learning based models have also been used to determine suicidal probability in military personnel [14]. Authors have extensively surveyed different data mining techniques related to suicide prediction from Twitter feeds [13]. Suicide ideation and the analysis of depression in Twitter users has been carried out through questionnaire based analysis reinforcing the trend analysis obtained from Twitter data, in [8]. Neural network based machine learning models have also been used by the authors in [16] to identify suicidal probabilities from Twitter data. A similar approach involving scanning of embedded documentation is presented in [10]. The technique proposed in [10], based on the Latent Dirichlet Allocation algorithm, is found to be quite effective compared to other state-of-the-art methods, considering a limited set of ways to describe depressive behaviour, which is its fundamental weakness.

Most researchers agree that depression is linked to suicidal tendencies, and, as a result, numerous studies have focussed on identification of depression in individuals as a way to evaluate how likely such people are to commit suicide. The researchers in [9] used an apriori algorithm to predict human depression, while ML based identification of behaviour patterns correlated to depression have been investigated in [5]. Others have endeavoured to identify biomarkers associated with suicide and depression [4], and establish the correlation between psychological vulnerability and depressive trends through Bayesian methods [2]. It is relevant to note that the application of Bayesian models in the estimation of groups most likely to attempt suicide is seen to be highly accurate in [2], considering the availability of a requisite number of psychological parameters. The analysis of family history using ML techniques is employed effectively in [7]. The application of ML techniques to manage suicide ideation has also been explored in recent years [6]. Extensive analysis of suicide ideation has been used to predict suicide attempts in [1]. Big data analytics have been employed by some researchers for suicide prediction and prevention [11]. The development of many suicide prediction methodologies have relied on complex models based on machine learning and neural networks for accurate prediction [12]. Other approaches include sensor-based models for suicide prediction [18]. Another approach to suicide prediction involves identification of groups of individuals highly at risk in terms of being probable suicides [15]. The effectiveness of this approach lies in the fact that statistical significance of test results allow for greater variance to a significant extent of input data with no appreciable change in error probability. Similar approaches have been proposed by other researchers, based on the Generalized Additive Model, which allows for socio-economic factors linked to suicides to be incorporated into the estimation model [17].

Revised Manuscript Received on September 15, 2020.
* Correspondence Author
Avhishek Biswas, Computer Science and Electrical Engineering, University of North Dakota, Grand Forks, North Dakota, United States of America. Email: avhishekofficial@gmail.com
Ananya Talukder, Electronics and Communication Engineering, MAKAUT, Kolkata, India. Email: talukderananya@gmail.com
Deep Bhattacharjee, Electronics and Communication Engineering, MAKAUT, Kolkata, India. Email: deepbhattacharjee486@gmail.com
Arijit Chowdhury, Electronics and Communication Engineering, MAKAUT, Kolkata, India. Email: arijitcontact96@gmail.com
Judhajit Sanyal*, Electronics and Communication Engineering, MAKAUT, Kolkata, India. Email: judhajit.sanyal.2019@gmail.com

The approach presented in this work uses adaptive Bayesian estimation to generate suicide probabilities and estimates of the number of suicides, with respect to different professional, educational and social categories.

The paper presents the theoretical model used in the work and analysis of experimental data, in section II. The results achieved using the proposed model are presented and discussed in section III. Section IV concludes the paper with a discussion on possible future enhancements to the proposed model.

II. THEORETICAL BACKGROUND AND EXPERIMENTAL DATA ANALYSIS

From the brief review of recent literature in the domain of suicide prediction, the work done by researchers suffers from any one or both of two drawbacks. First, some of the models require a great deal of psychological details, which entails that the person being tested can be accurately identified from among a large group of individuals. Second, some of the models effective in identifying individuals correctly in terms of being potential suicides are extremely complex in nature. As a consequence, the present work relies on a much simpler model, based on adaptive Bayesian inference, to predict the number of deaths by suicide among individuals. Different levels of education, employment and types of marital statuses are used to generate the estimates.The model uses linear spline regression to initially generate estimates. Equation 1 shows the general spline formulation used in the work.

$$D(P_F) = k_1 P_F + k_2 \quad (1)$$

Here $D(P_F)$ signifies the estimated deaths by suicide dependent on the probability PF based on the factor F (professional, educational or social, or a combination). The constants k_1 and k_2 are adjusted according to the variation of the Bayesian probability. In a manner similar to equation 1, the final model is dependent on the Bayesian probability P. This is a function of the cascaded probability functions for professional, educational and social factors. If the factors are represented as Pr, Ed and S, the following functional relationship holds.

$$P = f(P_{Pr} * P_{Ed} * P_S) \quad (2)$$

The three classes are shown in table I, with corresponding categories.

Table- I: List of factors with categorical segmentation

Type of Factor	Category
Educational	No Education
	Primary
	Middle
	Matriculate/Secondary
	Higher Secondary
	Diploma
	Graduate
	Post-Graduate or Above
Professional	Unemployed
	Farming/Agriculture
	Government Service
	Private Sector
	Self-Employed or Other Activity
Social	Never Married
	Married
	Separated
	Divorcee
	Widow/Widower

The training simulations considering the individual factors are shown in figures 1 to 3, which follow.

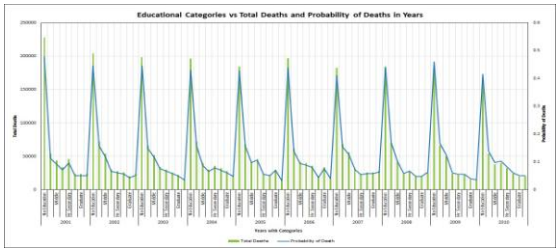

Fig. 1.Training dataset for educational categories.

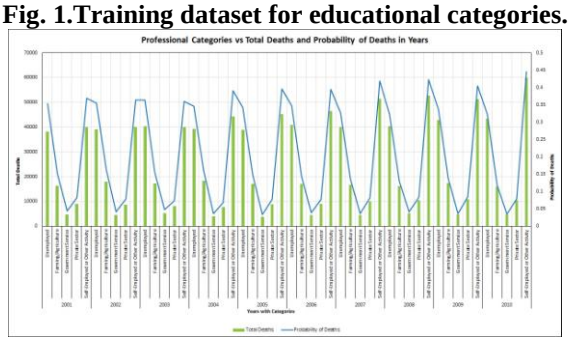

Fig. 2.Training dataset for professional categories.

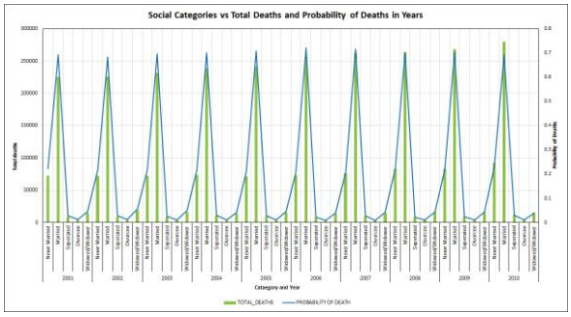

Fig. 3.Training dataset for social categories.

In each of the three figures, the adaptive linear spline estimate line is observed to follow the trend with an acceptable degree of accuracy. From the training datasets, it is also observed that the proposed model never significantly underestimates the probability of suicide for any of the categories, which makes this model safer to adopt in terms of correct identification of individuals at risk of committing suicide. The model is then tested through comparison with actual categorical suicide data, for the years 2011 and 2012. The factor specific test results are outlined in figures 4 to 6, which follow.

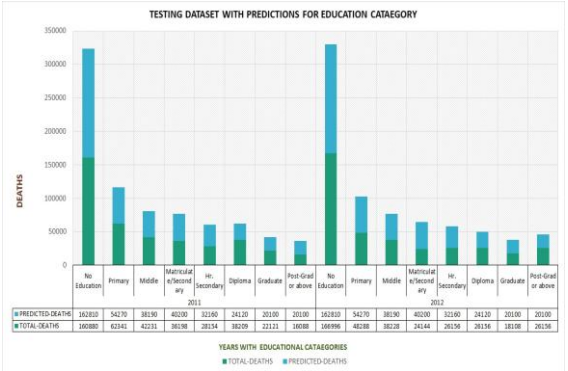

Fig. 4.Testing dataset for educational categories.

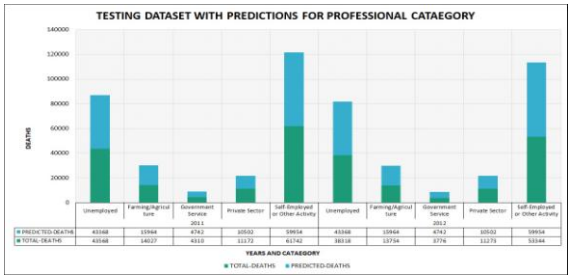

Fig. 5. Testing dataset for professional categories.

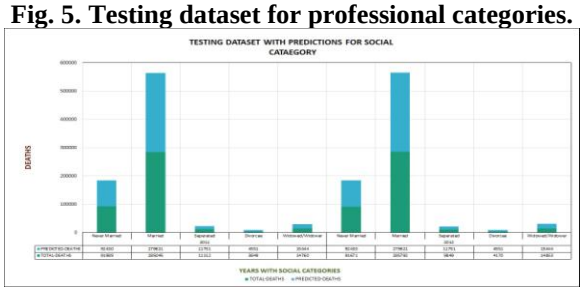

Fig. 6.Testing dataset for social categories.

A number of microtrends become evident by examination of the test results. For example, for educational classification, in case of individuals with no education, the risk of suicide is greatest, and least for individuals with graduate or post-graduate degrees, decreasing approximately monotonically with increase in the level of education. Similarly, with respect to professions, government servants are least prone to committing suicide, while self-employed individuals have the greatest suicide probability among the classes. For social category, married persons represent the group with highest suicide probability while divorcees have lowest probability of suicide. The proposed model is adapted to predict suicide probability for a combination of factors, for increased accuracy. A comparative study of accuracy in terms of RMSE (root mean square error) for the individual and combined factors is presented in table II.

Table- II: Accuracy by Bayesian Estimation Model for different combinations of factors

Factor	RMSE (%)
Educational	0.75
Professional	8.3
Social	0.65
Educational and Professional	1.6
Professional and Social	0.8
Social and Educational	0.13
Educational, Professional and Social	0.6

The variation of RMSE with the chosen factors clearly indicates that the combination of all three factors generates an appreciably low RMSE. In this case the RMSE is less close to zero than most of the other cases, while being significantly low, indicating that this model ensures accuracy without the drawback of data overfitting. The RMSE results obtained by the proposed model, for the combination of three factors, is compared to the basic linear spline and non-linear spline regression estimates. The results are presented in table III.

Table- III: Comparison of accuracy for Bayesian model, linear spline regression and non-linear spline regression

Type of model	RMSE (%)
Linear Spline Regression	4.1
Non-Linear Spline Regression	1.6
Proposed Bayesian Model (Adaptive Linear Spline)	0.6

From the data presented in table III, it is clear that the proposed model has best accuracy among the models compared.

III. RESULTS AND DISCUSSIONS

This The present work outlines an endeavour to build an adaptive spline regression model based on Bayesian estimation, for the prediction of suicide probability and estimation of the number of suicides per year, in the categorized data set shown here. The model proposed in this work is found to have several advantages. First, since it is essentially an adaptive linear model, the computational complexity associated with the model is much lower than other non-linear models, with no appreciable loss in accuracy. Second, in situations where the original data set is prohibitively large, the model allows for identification of subcategories most prone to suicidal attempts, which can help in the initial sampling of data from the dataset. It is also relevant to mention that this sampling is important, since carrying out of extensive psychological testing on a large number of people may not always be feasible, in which case the proposed model allows for more effective sampling through probabilistic gradation of suicide risk in individuals.

It is seen from the RMSE values in table II that the model is most effective in prediction of suicide when examining a combination of social and educational factors. Assuming this is not due to overfitting, this result may indicate that marital status and level of education may be greater determinants of the likelihood of a suicide attempt, compared to type of employment. Also, a careful analysis of the data clearly shows that people who are self-employed and married are most likely to attempt suicide, among the categories examined in this work. This fact may be linked to stress and depression arising out of the nature and uncertainty of their type of employment (self-employed), coupled with familial pressure (married individuals), which can lead to suicidal tendencies when either or both of these domains create excessive stress on the individuals, which corresponds to the results obtained in [7] and [12]. The proposed model is therefore found to be quite accurate in estimating the number of suicides, with added features such as the scope of identification of microtrends and probabilistic risk classification of individuals according to categories.

IV. CONCLUSION

The model proposed here has been found to be appreciably accurate for the examined dataset. Further, the inferences from the results obtained in this work have been found to agree with the findings of other researchers. Categorization of data can however be further increased to make the model more robust with equal or greater accuracy achieved. The model can also be applied to perform analyses using psychological parameters related to stress and suicidal tendencies as the categorical variables. In each of the cases, the nature of Bayesian estimation allows for a greater understanding of causality leading to better model performance, especially as the dimensionality of the dataset increases.

For more complex datasets, the model can be augmented by replacing the adaptive linear spline based approach with support vector machine (SVM) based algorithm in conjunction with the Bayesian estimation scheme. Another possible approach to the problem would be the use of Kohonen maps to reduce dimensionality in larger datasets with a prohibitively large number of factors. The authors intend to pursue further research in the present domain from these aspects, with an aim to propose a technique that can handle high dimensionality in datasets with comparatively low time-complexity, through the utilization of a combination of the approaches discussed above.

ACKNOWLEDGMENT

The authors wish to acknowledge the Department of Electronics and Communication Engineering, Techno International New Town, for providing the infrastructure and encouragement to carry out this research.

REFERENCES

1. G. T. Agarwal, A. Dhawan, A. Jain, A. Jain and S. Gupta, "Analysis and Prediction of Suicide Attempts," 2019 International Conference on Computing, Power and Communication Technologies (GUCON), NCR New Delhi, India, 2019, pp. 650-665.
2. Jorge Barros, Susana Morales, Arnol García, Orietta Echávarri, Ronit Fischman, Marta Szmulewicz, Claudia Moya, Catalina Núñez & Alemka Tomicic. Recognizing states of psychological vulnerability to suicidal behavior: a Bayesian network of artificial intelligence applied to a clinical sample. BMC Psychiatry 20, 138 (2020).
3. A. Ben-Ari and K. Hammond, "Text Mining the EMR for Modeling and Predicting Suicidal Behavior among US Veterans of the 1991 Persian Gulf War," 2015 48th Hawaii International Conference on System Sciences, Kauai, HI, 2015, pp. 3168-3175.
4. Hilario Blasco-Fontecilla, Maria A. Oquendo (2016) Biomarkers of Suicide: Predicting the Predictable?. In: Courtet P. (eds) Understanding Suicide. Springer, Cham.
5. A. A. Choudhury, M. R. H. Khan, N. Z. Nahim, S. R. Tulon, S. Islam and A. Chakrabarty, "Predicting Depression in Bangladeshi Undergraduates using Machine Learning," 2019 IEEE Region 10 Symposium (TENSYMP), Kolkata, India, 2019, pp. 789-794.
6. S. Colic, J. D. Richardson, J. P. Reilly and G. M. Hasey, "Using Machine Learning Algorithms to Enhance the Management of Suicide Ideation," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, 2018, pp. 4936-4939.
7. Theodoros Iliou, Georgia Konstantopoulou, Christina Lymperopoulou, Konstantinos Anastasopoulos, George Anastassopoulos, Dimitrios Margounakis, Dimitrios Lymberopoulos (2019) Iliou Machine Learning Data Preprocessing Method for Suicide Prediction from Family History. In: MacIntyre J., Maglogiannis I., Iliadis L., Pimenidis E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2019. IFIP Advances in Information and Communication Technology, vol 559. Springer, Cham.
8. S. Jain, S. P. Narayan, R. K. Dewang, U. Bhartiya, N. Meena and V. Kumar, "A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter," 2019 IEEE Students Conference on Engineering and Systems (SCES), Allahabad, India, 2019, pp. 1-6.
9. L. Jena and N. K. Kamila, "A Model for Prediction of Human Depression Using Apriori Algorithm," 2014 International Conference on Information Technology, Bhubaneswar, 2014, pp. 240-244.
10. N. Jones, N. Jaques, P. Pataranutaporn, A. Ghandeharioun and R. Picard, "Analysis of Online Suicide Risk with Document Embeddings and Latent Dirichlet Allocation," 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), Cambridge, United Kingdom, 2019, pp. 1-5.
11. Ronald C. Kessler, Robert M. Bossarte, Alex Luedtke, Alan M. Zaslavsky & Jose R. Zubizarreta (2019) The Role of Big Data Analytics in Predicting Suicide. In: Passos I., Mwangi B., Kapczinski F. (eds) Personalized Psychiatry. Springer, Cham.
12. Ronald C. Kessler, Samantha L. Bernecker, Robert M. Bossarte, Alex R. Luedtke, John F. McCarthy, Matthew K. Nock, Wilfred R. Pigeon, Maria V. Petukhova, Ekaterina Sadikova, Tyler J. VanderWeele, Kelly L. Zuromski, Alan M. Zaslavsky. Suicide prediction models: a critical review of recent research with recommendations for the way forward. Mol Psychiatry 25, 168–179 (2020).
13. E. R. Kumar and A. K. V. S. N. R. Rao, "Suicide Prediction in Twitter Data using Mining Techniques: A Survey," 2019 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, Tamilnadu, India, 2019, pp. 122-131.
14. G. Lin, M. Nagamine, S. Yang, Y. Tai, C. Lin and H. Sato, "Machine Learning Based Suicide Ideation Prediction for Military Personnel," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 1907-1916, July 2020.
15. S. S. Priyanka, S. Galgali, S. S. Priya, B. R. Shashank and K. G. Srinivasa, "Analysis of suicide victim data for the prediction of number of suicides in India," 2016 International Conference on Circuits, Controls, Communications and Computing (I4C), Bangalore, 2016, pp. 1-5.
16. N. Shahreen, M. Subhani and M. Mahfuzur Rahman, "Suicidal Trend Analysis of Twitter Using Machine Learning and Neural Network," 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), Sylhet, 2018, pp. 1-5.
17. J. Shen, S. Zhao and M. Ye, "Suicide Prediction Analysis with Generalized Addictive Model," 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York City, NY, USA, 2019, pp. 1069-1073.
18. John Torous, Mark E. Larsen, Colin Depp, Theodore D. Cosco, Ian Barnett, Matthew K. Nock & Joe Firth. Smartphones, Sensors, and Machine Learning to Advance Real-Time Prediction and Interventions for Suicide Prevention: a Review of Current Progress and Next Steps. Curr Psychiatry Rep 20, 51 (2018).

AUTHORS PROFILE

Avhishek Biswas is working as a research scholar in the School of Computer Science and Electrical Engineering at the University of North Dakota (UND), USA, at the Data, Energy, Cyber, and Systems (DECS) Laboratory. He completed his B.Tech in Electronics and Communication Engineering from Techno International Newtown, Kolkata, India. His research interests include computer vision, Machine Intelligence, Deep Learning, and Pattern Recognition.

Ananya Talukder is a graduate engineer trainee at Capgemini India. She completed her B. Tech. degree in Electronics and Communication Engineering from Techno International Newtown, Kolkata, India. Her research interests include cyber security and database management systems.

Deep Bhattacharjee is a software engineer trainee at Cloudkaptan Consultancy Pvt. Ltd. He completed his B.Tech degree in Electronics and Communication Engineering from Techno International Newtown, Kolkata, India. His research interests are in CRM software and Salesforce.

Arijit Chowdhury completed his B.Tech degree in Electronics and Communication Engineering from Techno International Newtown, Kolkata, India and was offered a position at Cognizant. His research interests are in the domains of robotics and artificial intelligence.

Judhajit Sanyal completed his B. Tech. in Electronics and Communication Engineering and M. Tech. in Radiophysics and Electronics. He currently teaches Electronics and Communication Engineering at Techno International New Town. His research interests are in the domains of microwave communication, artificial intelligence and optimization of 5G networks.