Towards Developing a Machine Learning Model For Suicidal Attempt Prediction

Tanvir Hasan
Department of CSE
East West University
Dhaka, Bangladesh
tanvirfuad00@gmail.com

Md. Serajus Salekin khan
Department of CSE
East West University
Dhaka, Bangladesh
khannahid64@gmail.com

Md. Rajib Hawlader
Department of CSE
East West University
Dhaka, Bangladesh
rajib104.ewubd@gmail.com

***Abstract*—Nowadays, suicidal tendency among the people has been increased. Therefore, we were motivated to study regarding this area. In this study [N=469], we have classified the samples who have attempted for suicide and who did not. For classification purpose, we have used random forest classifier and we got 83.7% accuracy regarding classification of these two groups of people. In addition to our classification analysis, we have also shown which factors can be important for suicidal attempt. Our findings show that age and strategy used to improve oneself (e.g. join clubs) are the most important factors.**

***Index Terms*—Suicide, Random Forest, Prediction, Features**

I. Introduction

Suicide causes immeasurable pain, suffering, and loss to individuals, families, and communities nationwide [2]. According to the World Health Organization, every year close to 800 000 people take their own life and there are many more people who attempt suicide [3]. Every suicide is a tragedy that affects families, communities and entire countries and has long-lasting effects on the people left behind [3]. In 2019, an average of 32 deaths can be attributed to suicide in Bangladesh every day [4].

There are many previous studies regarding people's suicidal attempt. However, to our best knowledge, there is no authentic work who has worked on this data set [5]. Investigating this dataset may give us some unique insightful information which motivated us to work on this data set.

We did classification and feature selection analysis. Important feature selection analysis showed that person's age and how one improves oneself (e.g. join clubs, gym) are the most important features. On the other hand, in case of classification analysis we found that our model can 83.7% correctly classify the persons who have attempted to commit suicide and who did not. Thus, our findings will help the researchers in this area to investigate more which factors are the most important for suicidal attempt and also will help to develop a better model for predicting suicidal attempt.

II. Related Work

A. Factors which Influence to Commit Suicide

Using subjective data, Liu [6] examined the relation between sleep patterns and sleep problems and adolescent suicidal behavior. Findings of that study shows the association between short sleep duration and nightmares and suicidal behavior. Liu [6] also highlight the potential role of sleep intervention in the prevention of adolescent suicide. On the other hand, using machine learning and data mining techniques, Amin & Sayed [10] conducted a study to identify the root causes behind the suicide. They found that ratio of suicidal cases is higher for men than women. Beside this, their findings show that there is age difference between men and women who attempt or commit suicide. Age group for most of the men was 30 to 44 years, however, age group for most of the women was 15 to 29 years.

B. Prediction of Suicidal Tendency

Cassells et al. [7] remarks that the prediction of suicide remains a major challenge for health care professionals in inpatient settings. Walsh et al. [8] worked in this are with the aim to use machine learning to enable clinical screening using routine real world data to predict non-fatal suicide attempts among adolescents in a major academic medical center. Their computational models performed well and did not require face-to-face screening. Walsh et al. [11] conducted a study to evaluate the accuracy and temporal variation of a potentially scalable suicide attempt risk detection strategy: machine learning applied to electronic health records (EHRs). They developed machine learning algorithms that accurately predicted future suicide attempts. Accuracy of their findings improved from 720 days to 7 days before the suicide attempt, and predictor importance shifted across time.

III. Methodology

A. Data Set Description

We used a data set [5], available in Kaggle. In that data set, there were 469 participants and 84% of the participants were male. This dataset consists of participants' subjective data and the survey took place from May, 2016 to September 2016. Most of the participants (81.02%) were young whose age were 17 to 30 years. Meanwhile, 34% participants had income of 0$ and 21% participants had income up to 10,000$.

B. Data Preprocessing

Though there were in total 19 features, we did not use every feature in our machine learning model. From our perception,

survey conducting date and time cannot contribute much as an important feature in the predictive model. Therefore, we have removed that feature.

C. Feature Selection And Classification Analysis

For feature selection analysis, we used Pandas and Scikit-Learn python library. To build up the model, we have used ExtraTreesClassifier which is an ensemble learning method basically based on decision trees.

In a previous study [8], random forests significantly outperformed logistic regression. This motivated us to use Random Forest Decision Tree classifier instead of other classifiers. In this case also, we used scikitlearn library where random forest and other required built-in functions (e.g. accuracy_score) were available.

For training and testing purpose, we kept 426 participants (90.83%) in the train group and remaining 43 participants were kept in the test group.

D. Pseudocode to Analyze the Data

1 Select the features from dataset
2 For each item in features
2 (a). Calculate information gain
2 (b). Select node which has highest information gain
2 (c). Split node into sub-nodes
2 (d). Repeat step a, b and c to construct the tree until reaching minimum number of samples required to split.
3 Repeat steps 1,2 for N times to build forest of N trees
4 Compare all of the trees and predict output

IV. Findings

A. Feature Analysis

Figure 1 shows the visual representation regarding the 10 most important features scores. Brådvik [9] remarks that depression strongly related to both suicidal ideation and attempt. Thus, we expected that depression will be the most important feature. However, it was surprising to see that participants age and different strategies used to improve oneself (e.g. join clubs, gym) were found as the most important features. On the other hand, depression got least score among the most important 10 features.

B. Classification Analysis

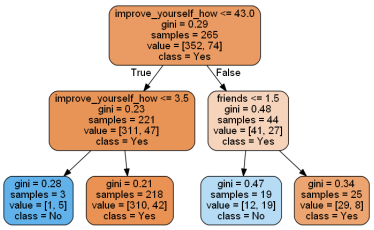

Fig. 1. Decision Tree 1

Based on the features, our model correctly predicted 83.7% times to classify the people who have attempted to commit suicide and who did not. Figure 2, Figure 3, Figure 4 show the decision trees based on our participants' data.

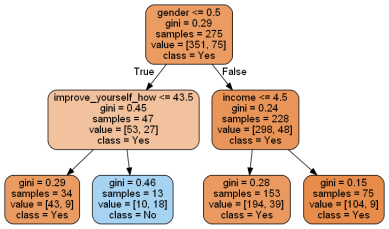

Fig. 2. Decision Tree 2

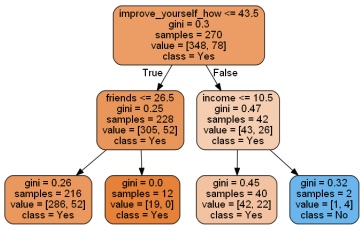

Fig. 3. Decision Tree 3

V. Conclusion

We conducted a study with the aim to develop a model which can classify suicidal attempt tendency of the people. We have also presented feature analysis regarding this area. Our findings show that age and strategy to develop oneself is the most important factor in suicidal attempt. Our developed model which is based on random forest predicted peoples' suicidal attempt tendency 83.7% accurately. The presented work is a step towards developing a better model in this area.

VI. Limitations

Though we have tried to minimize the limitations, there are still few limitations in our study which should be explored by future studies. In this study, we have used subjective data. Thus subjective data may cause to have insufficient evidence to uncover the actual factors which can be important for suicidal attempt prediction. Moreover, we did not analyze whether our developed model is over fitted or not.

Acknowledgment

We like to thank the data set owner for his effort to construct this data set. We also thank to our supervisor for his insightful comments on the submitted draft paper.

References

[1] Kimberly A. Van Orden, Tracy K. Witte, Kathryn H. Gordon, Theodore W. Bender, and Thomas E. Joiner Jr., "Suicidal Desire and the Capability for Suicide: Tests of the Interpersonal–Psychological Theory of Suicidal Behavior Among Adults," Journal of Consulting and Clinical Psychology, vol. 76, no. 1, pp. 72-83, 2008
[2] U.S. Department of Health & Human Services, "Suicidal Behavior," https://www.mentalhealth.gov/what-to-look-for/suicidal-behavior (Accessed on 9th May 2020)
[3] WHO, "Suicide," https://www.who.int/news-room/fact-sheets/detail/suicide (Accessed on 9th May, 2020)

[4] Kamrul Hasan, "World Suicide Prevention Day What is Bangladesh doing to reduce the risk of suicide?" https://www.dhakatribune.com/bangladesh/2019/09/10/world-suicide-prevention-day-what-is-bangladesh-doing-to-reduce-the-risk-of-suicide (Accessed on 9th May 2020)
[5] LiamLarsen, "The Demographic /r/ForeverAlone Dataset," https://www.kaggle.com/kingburrito666/the-demographic-rforeveralone-dataset (Accessed on 9th May 2020)
[6] Xianchen Liu, "Sleep and Adolescent Suicidal Behavior," SLEEP, Vol. 27, No. 7, 2004.
[7] Clare Cassells, Brodie Paterson, Dawn Dowding, and Rhona Morrison, "Long- and Short-Term RiskFactors in the Predictionof Inpatient Suicide:A Review of the Literature," Crisis, vol. 26, no.2, pp. 53–63, 2005
[8] Colin G. Walsh, Jessica D. Ribeiro, and Joseph C. Franklin, "Predicting suicide attempts in adolescents withlongitudinal clinical data and machine learning," Journal of Child Psychology and Psychiatry, vol. 59, pp 1261–1270, 2018
[9] L. Brådvik, "Suicide Risk and Mental Disorders," Int J Environ Res Public Health vol. 15, no. 1, 2018
[10] I. Amin, S. Syed, "Prediction of Suicide Causes in India using Machine Learning, " Journal of Independent Studies and Research – Computing, vol. 15, no. 2, December 2017
[11] C. G. Walsh, J. D. Ribeiro, and J. C. Franklin, "Predicting Risk of Suicide Attempts Over Time Through Machine Learning," Clinical Psychological Science, 2017.