

Suicidal Ideation Prediction Using Machine Learning

A AYUSH GOWDA¹, ABHIRAM P S², HEMASHREE D B³, KEERTHANA R⁴, S J SHAHEENA BEGUM⁵

^{1, 2, 3, 4}Dept of CSE, Ghousia College of Engineering Ramanagara, Karnataka, India

⁵Assistant Professor, Dept of Computer Science and Engineering

Abstract- *Suicide is become a major cause of death worldwide. With each year that goes by, the suicide rate rises dramatically. It is crucial to remember that suicidal thoughts is a complicated problem that cannot be exclusively linked to one mental illness or circumstance. According to reports, sadness, anxiety, high levels of stress, unsatisfying relationships, sleep difficulties, and physical or mental impairments are the main factors that contribute to suicidal ideation. Therefore, when forecasting suicidal ideation, a comprehensive approach that takes into account a number of indicators is crucial. Using a large sample of review data, our proposed work employs machine learning approaches to predict suicidal ideation. We preprocessed and trained the data using a variety of machine learning techniques after gathering it using a tagged dataset. Models like Decision Trees (DT), Support Vector Machines (SVM), Random Forest (RF), and Logistic Regression (LR). According to the results, LR has the highest accuracy of all the computed models (76%), followed by SVM (75%), MLP (74%), RF (74%), and DT (74%). The clinical industry and, consequently, the welfare of our society depend on the suggested framework. Index Terms: depression, anxiety, machine learning, suicidal thoughts, and drowsiness*

concerns including developing their own identities, self-worth, independence, & responsibility, as well as forming new close relationships. They are currently going through ongoing, evolving mental and bodily alterations over this period. Additionally, their close friends and family may have unreasonably high expectations of those. Feelings of helplessness, uncertainty, anxiety, and relinquishing control are unavoidable outcomes in such circumstances. To effectively manage these challenges and emotions, most people need to have access to a safe living environment, close friendships, a structure, and financial resources. The home context in which young people live or have grown up is one of the most important sources of support when it comes to resolving the various problems that adolescents face. Numerous studies have connected suicide behaviour to a number of risk factors, including family dynamics & procedures. Anxiety disorders have also been connected to suicide. The consequences of substance misuse and temperamental issues, which & more common under these circumstances, present in these circumstances, are challenging to assess. As a result, we can comprehend that there are numerous variables and causes related to suicidal ideation. In order to respond appropriately, it is crucial to comprehend suicidal thoughts in individuals at the appropriate moment.

I. INTRODUCTION

Suicidal thoughts or desires to end one's life are referred to as suicidal ideation. This can include considering suicide as a way to end one's life or as a way to get away from issues. It's crucial to remember that suicidal thoughts are a serious issue that needs to be addressed. The elements that contribute to suicidal ideation—gender, age, relationship status, insomnia, despair, anxiety, and physical or mental disabilities—are the main focus of this essay. Females were more likely than males to meet the criteria for depression in 50–65% of suicides. Young people need to deal with

II. BACKGROUND AND RELATED WORKS

According to studies, college students have a higher suicide incidence than the overall population. Numerous factors, including as despair, stress, loneliness, strained relationships, and greater academic obligations, can lead to suicidal thoughts and behaviours. Even though the number of persons experiencing emotional difficulties is rising, less than half of those who have contemplated suicide have sought professional assistance. According to Namik Kirlic et al. in order to guide suicide prevention

initiatives, it is necessary to enhance the early detection and evaluation of students with STBs as well as uncover modifiable factors linked to STBs. This study used machine learning techniques to investigate the elements that contribute to resilience or suicide risk in college students. According to the study, measures meant to implies that increasing these parameters through treatments could enhance wellbeing and lower the risk of suicide in this demographic.

Gen-Min Lin et al.'s study [1] examined the application of machine learning techniques to forecast suicide thoughts among military personnel, who are known to have elevated stress levels. The study showed how well machine learning techniques, such as support vector machines and multilayer perceptrons, can predict suicidal thoughts based on psychological stress categories. An accuracy rate of about 100 produced the best outcomes. A machine learning method was employed by Jeongyoon Lee et al. [3] to forecast suicide outcomes. He employed four machine learning classifiers, including LR, RF, SVM, and XGBoost. Three algorithms (SVM, RF, and XGBoost) had their hyperparameters modified using grid search and cross-validation. The algorithm was assessed using performance metrics, such as sensitivity, specificity, and accuracy. utilised to assess which algorithm performs the best. The findings demonstrated that suicidal thoughts may be accurately predicted with high sensitivity (0.808 to 0.853) and accuracy (0.84 to 0.86). In order to predict near-term suicidal behaviour, a tool for evaluating the suicide crisis syndrome, Neelang Parghi et al. [4] used machine learning (ML) to predict Suicide Crisis Inventory (SCI) analysis. Twenty of the 591 high-risk psychiatric inpatients who provided the data attempted suicide, whereas the remaining 571 did not. Three sampling strategies and three prediction approaches (logistic regression, random forest, and gradient boosting) were used to examine the data. The improved bootstrap approach produced the best results, with gradient boosting and random forest showing the highest accuracy of 98% and 33.9%, respectively, in predicting near-term suicidal behaviour. recollection, 71 percent AUPRC, and 87.8 percent AUROC. Yanmei Shena et al. [5] conducted research to create a machine learning system that might predict medical college students' likelihood of

attempting suicide. The study used a random forest model, 37 predictors of suicide attempt, and online self-report data from 4,882 medical students. According to the findings, the random forest model predicted suicide attempts with 90.1% accuracy. Imran Amin's study [6] employed machine learning and data mining methods to forecast the reasons behind suicides in India. The goal was to identify the reasons behind suicide so that authorities could address the likely causes and increase public awareness. The research reasons. The study employed a predictive strategy to predict future causes and a descriptive and statistical approach to identify the suicide pattern. According to the report, most female suicide victims are between the ages of 15 and 29, and most male suicide victims are between the ages of 30 and 44. Christiane Arrivillaga et al.'s study [7] of 2196 Spanish teenagers sought to ascertain whether emotional intelligence serves as a preventative measure in the relationship between self-murder creativity and inadequate internet and smartphone usage. The findings indicated that self-murder inventiveness was positively correlated with inappropriate internet and mobile device use, however emotional intelligence was adversely correlated with both. The negative correlation between smartphones, the internet, and Emotional intelligence is a moderator of creativity. Using data from the Millennium Cohort Study of 7,347 17-year-olds, Kristin Jankowskyl et al. [8] found that the logistic regression and elastic net regression algorithms outperformed the gradient boosting machine method in predicting teenage suicide attempts. The most important predictor was past self-harm; victimisation, mental health, emotion and motivation, drug use, sexuality, and demography were next. These studies show how machine learning techniques including logistic regression, decision trees, random forests, gradient boosting regression trees, support vector machines, and multilayer perceptrons can be used to predict suicidal ideation and behaviour. Suicide prediction and prevention are investigated through the use of psychometric measurements, demographic, socioeconomic, and psychosocial variables. suicide. The results indicate that machine learning algorithms may be helpful in anticipating suicide behaviour in the near future, but proper data pre-processing methods must be used.

III. METHODOLOGY

There are five steps in the process shown in Figure 1 for creating predictive models for suicidal behaviours. Preprocessing is the first step in gathering and preparing data for analysis. Next, statistical methods are used to identify significant predictors. Next, a variety of algorithms are used to train machine learning models. Lastly, the models' effectiveness is assessed using suitable measures. Every stage is essential to guaranteeing the final model's precision and efficacy. It's crucial to use representative and trustworthy data, appropriately prepare it for analysis, choose significant predictors, employ a variety of algorithms, and assess model performance using the right metrics. All things considered, this methodology offers a methodical way to create precise and useful predictive models for suicide tendencies.

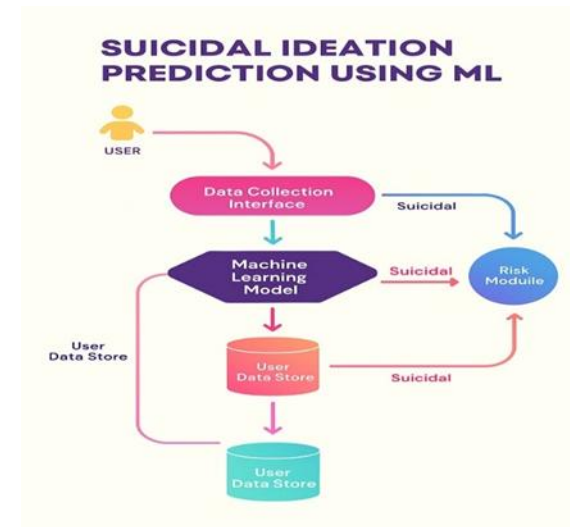


Fig 1. Use case diagram

A. Information Gathering the dataset for this study came from an open science framework that was filled with anonymous online survey responses from Bangladeshi university students using convenience sampling techniques. [9]. A systematic questionnaire was used in the data collection procedure. to compile data on a range of topics, including suicidal behaviour patterns, sociodemographic data, academic data, psychological disorders, and behaviours. Two distinct scales were used in the study to assess different facets of mental health. The ISI [11] scale was used to gauge insomnia, and the DASS-21 [10] tool, which consists

of 42 items divided into three subsections of seven questions each, was used to assess depression, anxiety, and stress levels.

B. Feature Selection and Data Preprocessing data preparation, which includes operations like standardising data, filling in missing values, and eliminating outliers, is the initial stage in constructing a predictive model. Our research focusses on a number of variables that have previously been found to be potential risk factors for suicidal ideation. These elements include anxiety, sleeplessness, sadness, stress, and strained relationships. bad relationships, worry, sleeplessness, and physical or mental impairments. As a result, we have decided to incorporate these characteristics into our research. A crucial stage in the model-building process is feature selection, which tries to enhance the dimensionality of the dataset to boost the accuracy and efficacy of the model. and eliminating unnecessary or superfluous features. Feature significance scores, which quantify each feature's contribution to the model's prediction accuracy, are a popular method for identifying significant features. The features can be ranked according to these scores, and the most pertinent ones can be chosen for model training. shows how several elements are related to suicidal thoughts, according to the study. According to the data, suicidal thoughts were most strongly correlated with depression. The study also found a correlation between suicidal thoughts and relationship status, stress, and anxiety. The study highlights how important it is to address these characteristics in programs aimed at preventing suicide. Furthermore, it implies that treatments for depression and sleeplessness may be especially successful in lowering the risk of suicide. Figure 3. Variable Significance the dataset includes statistics on how pupils behave and how likely they are to think about suicide. Only the target variable "suicidal" and pertinent features remain after the superfluous columns are eliminated in order to train the classifiers. The formula is used to standardise the feature values. are standardised using the z-score normalisation formula:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where x , μ , and σ represent the characteristic's mean and standard deviation, respectively.

C. Model training

1) LR Classifier : A linear model called the Logistic Regression (LR) classifier forecasts the target variable's likelihood. The logistic function,

$$p(y = 1|x) = \frac{1}{1 + e^{-\theta^T x}} \quad (2)$$

where y is the target variable, x is the feature vector, θ is the weight vector, and e is Euler's number, is used to train it on the preprocessed dataset. The linear model's result is converted into a probability value using the logistic function.

2) SVM Classifier: The Support Vector Machine (SVM) classifier is a linear model that determines which hyperplane best separates the data into discrete groups. The optimisation problem on the preprocessed dataset is used to train it.

$$f(x) = w^T x + b \quad (3)$$

3) DT Classifier: Based on feature values, the tree-based Decision Tree classifier (DT) separates the data into subgroups. It is trained on the preprocessed dataset using a recursive method that selects the best feature to split the data into distinct nodes. Two hyper-parameters that need to be adjusted in the decision tree classifier are the maximum tree depth and the minimum number of instances required to split a node.

4) RF Classifier: To lessen overfitting, the Random Forest classifier (RT) is an ensemble model that incorporates many Decision Trees. Several Decision Trees, each trained on a random subset of the features and samples, are used to train it on the preprocessed dataset. Once more, we utilise the preprocessed data as input to train the RF classifier. A random forest model is fitted to the data using scikit-learn, and the hyper-parameters—such as the total number of trees in the forest, the maximum depth of a tree, and the bare minimum of samples required to split an internal node—are adjusted.

5) MLP Classifier: A neural network with multiple layers of nodes is called a Multilayer Perceptron

classifier (MLP). Backpropagation is used to train it. It is trained on the preprocessed dataset using the backpropagation method, which adjusts the node weights to reduce the error between the expected and actual output. The MLP classifier has hyperparameters that need to be adjusted, including the number of hidden layers, the number of nodes in each layer, and the learning rate.

D. Metrics for Evaluation the following formula is used to determine the accuracy score:

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}} \quad (4)$$

The following formulas are used to determine the precision and F1-score:

$$\text{precision} = \frac{\text{True Positive}}{\text{False Positive} + \text{True Positive}} \quad (5)$$

$$\text{F1 - score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

where

$$\text{recall} = \frac{\text{True Positive}}{\text{False Negative} + \text{True Positive}} \quad (7)$$

E. Selecting a Model it is crucial to choose a model that performs well on each of these evaluation metrics. However, the specific problem at hand and the expenses associated with false positives and false negatives will determine how important each indicator is. As a result, it's critical to weigh the trade-offs between recall, accuracy, and precision while selecting a model.

IV. RESULT

Model	Accuracy	Precision	F1-score
Logistic Regression	0.76 ± 0.06	0.60	0.54
SVM	0.75 ± 0.10	0.56	0.59
Decision Tree	0.74 ± 0.06	0.62	0.61
Random Forest	0.74 ± 0.06	0.62	0.61

MLP	0.76 ± 0.07	0.60	0.59
-----	-----------------	------	------

TABLE 1 RESULTS FOR DIFFERENT MODELS

After taking into account a variety of variables, including the number of features, the size and complexity of the dataset, and the required degree of interpretability, we have established Because since it is more likely to produce the most accurate predictions for the specific problem, the Logistic Regression model is the best-performing model for the dataset. V. SUMMARY Suicidal ideation is frequently linked to mental illnesses and traumatic experiences, and suicide is an increasing global issue. Suicidal ideation in big datasets can be effectively predicted using machine learning models. Gender, family income, marital status, age, insomnia, despair, anxiety, and physical or mental handicap are among the risk variables for suicide ideation that were emphasised in the study. The findings demonstrate that in a sizable dataset of student survey responses, the Logistic Regression model offered the highest accuracy for predicting suicidal thoughts. All things considered, our study offers the clinical and research community helpful information and resources to enhance communities to enhance efforts to prevent suicide. VI. ACCEPTANCE the Department of Science and Technology, New Delhi (DST-FIST Order No: SR/FST/COLLEGE-346/2018 dated December 20, 2018) provided financial support, for which the authors are grateful.

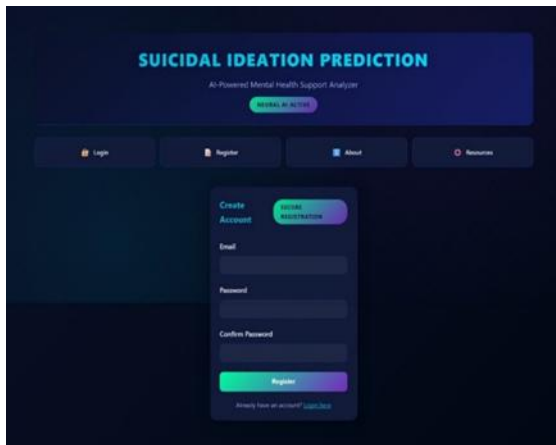


Fig 2. Prediction Result Screen

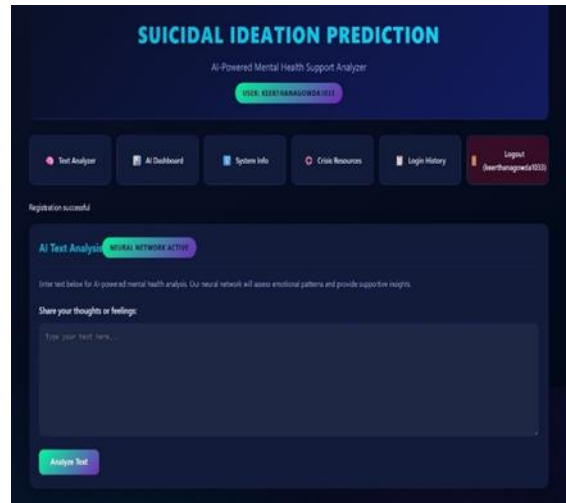


Fig 3. AI Text Analyzer Interface

CONCLUSION

This project is the proposed Voice based Email system for visually impaired people, which is developed as an application which helps the blind and handicapped people to access mails easily and efficiently. It provides a voice based mailing service where the visually impaired person could read and send mail by their own without the help of others. It requires basic information about keyboard shortcuts. System has eliminated all these concepts and overcome all difficulties faced by the visually impaired. It uses a speech recognition application which provides an efficient voice input method for mailing devices for blind. It is also useful for handicapped and illiterate people.

REFERENCES

- [1] K. C. Deshpande, R. D. Kharat, and A. S. Deshpande, "Prediction of Suicidal Ideation Using Machine Learning Algorithms," *International Journal of Engineering Research & Technology (IJERT)*, vol. 10, no. 05, pp. 1–6, 2021.
- [2] D. Burnap, M. L. Williams, O. Rana, W. Housley, A. Edwards, J. Morgan, and L. Sloan, "Detecting tension in online communities with machine learning and lexical analysis," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 3, pp. 1–23, 2017.

- [3] J. C. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. H. Cohen, and J. Hurdle, "Sentiment analysis of suicide notes: A shared task," *Biomedical Informatics Insights*, vol. 5, pp. 3–16, 2012.
- [4] M. Matero, K. Idnani, A. Son, S. Giorgi, H. Vu, D. Zamani, L. Schwartz, H. Eichstaedt, and L. Ungar, "Suicide risk assessment with multi-level dual-context language and BERT," *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pp. 39–44, 2019.
- [5] D. Benton, M. Mitchell, and D. Hovy, "Multi-task learning for mental health using social media text," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, vol. 1, pp. 152–162, 2017.
- [6] K. Sawhney, S. Joshi, P. Jha, and R. Shah, "Analyzing the language of suicide ideation on social media," *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 942–945, 2018.
- [7] A. K. Roy, S. Poria, and E. Cambria, "Deep learning-based mental health analysis from social media text," *Information Fusion*, vol. 87, pp. 1–13, 2022.