Unveiling Sociodemographic and Economic Drivers of Suicide Using Machine Learning: Toward Ethical and Effective Prevention Strategies

Khushboo Rathore[1], Pradeep Kumar Mishra [2], *Mritunjay K Ranjan[3], Prasad Gadekar[4], Rahul Mandal[5], Kailas Doke[6],

[1]Department of Education, Kamla Nehru Mahavidyalaya, Korba, Chhattisgarh, India
[2]Department of Management Studies, Sankalchand Patel University, Visnagar, Gujrat, India
[3,4,5,6]School of Computer Science and Engineering, Sandip University, Nashik, India
[1]khushboorathore1312@gmail.com,[2]drpradeepmishra40@gmail.com,[3]mritunjaykranjan@gmail.com,[4]prasad01198@gmail.com,[5]rahultime11@gmail.com,[6]kailasdoke100@gmail.com

Abstract. General national suicide is one of the major cover health issues which are influenced by various sociological and economic factors at large. Building on the ML approach, this work aims at analysing the key sociodemographic and socio-economic factors affecting the rates of suicide and the fundamental causal aspects. SUICIDE Among different factors of study, this research empirically explores cross-sectional relationship of age, gender, geographical distribution, and reasons for suicide including economic synthesized factors such as unemployment rates and economic downturn. Both ensemble and regression models are applied in order to analyze the data and identify patterns or correlations. The focus is especially vivid on the ethical use of artificial intelligence since crucial issues of fairness, accountability, and transparency of predictive models for sensitive societal questions are raised. Thus, the envisaged research proposes a set of empirically supported frameworks for policymakers to develop relevant and efficient interventions. In line with the principles of responsible AI, the findings show the possibility of machine learning in enabling datasets to guide public health interventions which are socially inclusive, ethologically right, and programmed to reduce on the multitude factors that cause suicide. Within this work, AI is presented as enabler of change for the betterment of society and support to sustainable mental health strategies.

Keywords: Sociodemographic, Socioeconomics, Suicide Analysis, Machine Learning, Ethical AI.

1 Introduction

Self-inflicted death is still a major concern of the global community affecting public health and acting as a major social, emotional and financial burden. [1] Over the last couple of years, more emphasis has been placed on the identification of multiple de-

N. Vyas et al. (eds.), *Proceedings of the International Conference on Advancements in Computing Technologies and Artificial Intelligence (COMPUTATIA 2025)*, Advances in Intelligent Systems Research 189,
https://doi.org/10.2991/978-94-6463-700-7_18

mographic and socioeconomic determinants that contribute to suicide. [2] In particular, the omnidirectional and multisource interactions preclude traditional statistical tools from identifying these structures accurately, which in turn requires contemporary models capable of addressing non-linearity and revealing hierarchies of complexity. [3] Let us consider a simplified statistical representation of suicide rates, Y expressed as a function of several predictors Xi as in Eq.(1):

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon \qquad (1)$$

Where, Xi indicates independent parameterizations such as age, gender and economic status; βi stands for coefficients that measure the effect of these parameters, while ϵ notes the error. However, this approach yields useful information and fails to encapsulate the interaction between predictors since it is based on linear dependency. To address this limitation, machine learning models such as Random Forest replace the linear function with a flexible, non-linear mapping [4] as in Eq. (2)

$$\hat{Y} = f(X_1, X_2, \dots, X_n) + \epsilon \qquad (2)$$

f is learnable, essentially what permits it to model the interactions and dependencies within the data. These approaches assess sociodemography and socioeconomy risk factors that underlie suicide and its reasons. This research is important because suicide rates vary over time. The specificity of the correct use of AI was explored in the study. Suicide is among the social issues that FAT models cover. Patients with social benefits are considered as participants of AI intervention with suicide prevention as the focus. This specifically modern machine learning method is for determining the connections and characteristics of suicide risks. Suicide elimination is due to the fact that in this research society is the client.

Objectives:

- To use state-of-art machine learning methods to investigate the modality and magnitude of sociodemographic and/or socioeconomic predictors of suicide rates.
- To uncover what is most predictive of suicide, primarily educational status, social class and age through various model- interpretation analyzed models like Random Forest regression model.
- To indicate solutions comprehensible to a policymaker, that appropriately respond to ethical concerns and contribute to mentoring evidence-based public health and fair suicide prevention practices that address the most influential factors.

2 Related Study

A comparative study on findings, and gaps in previous studies on the factors influencing suicide rates are shown in Table 1.

Table 1. Comparative study objectives, methods, findings, and gaps in studies on the factors influencing suicide rates.

Reference	Objective	Methods	Key Findings	Research Gap
[5]	Examine the link between socioeconomic status and suicide rates	Regression models	Economic distress is a key predictor of suicide	Lacks integration of machine learning approaches.
[6]	Develop a predictive model for suicide risk	Random Forest, SVM	Random Forest is effective in predicting suicide risk	Limited exploration of regional and demographic variations.
[7]	Assess the effect of economic downturns on suicide rates	Time series analysis	Recessions increase suicide rates, especially in males	Focuses mainly on economic factors without considering broader interactions.
[8]	Analyze the effect of socioeconomic inequality on suicide	Cross-country comparison	Inequality is linked to higher suicide rates	Does not utilize machine learning for analyzing complex factor interactions.
[9]	Evaluate machine learning techniques in identifying suicide risk factors	Neural Networks, Decision Trees	Machine learning models outperform traditional methods	Insufficient focus on integrating socioeconomic variables.
[10]	Explore the impact of demographic factors on suicide rates	Decision Trees, Gradient Boosting	Age and gender are significant predictors	Lacks deeper analysis of interactions between socioeconomic and demographic factors.
[11]	Investigate urbanization's effect on suicide rates	Logistic Regression, PCA	Urbanization and low socioeconomic status increase risk	Limited use of machine learning to integrate urbanization and socioeconomic variables.

3 Problem Statement

Identifying risk factors for suicide continues to be tasking mainly because of a myriad of sociodemographic and socioeconomic features that may be involved. Main antecedents include economic straits and difficulty, marital conflict, and family quarrels, which have been pinpointed by typical statistical techniques. However, all these linear models have the capacity to provide less understanding of the many variables at play as they often provide a linear direction of the relationships between these variables. Furthermore, a large number of articles are descriptive, centering their attention on specific populations or settings and, therefore, their results cannot be applied widely. While holding much promise for increasing the predictive power of forecasts, machine learning approaches frequently fail to account for the interdependence between various demographic and socioeconomic, for example, urbanisation, education and inequality. This gap hinders the establishment of relevant and efficient large-scale population health interventions. In response to these challenges, this study adopts sophisticated artificial neural networks and other machine learning algorithms, to simultaneously analyze multiple data sets, seek for relationships and patterns and determine the most relevant predictors for suicides. The findings will enlighten the ethical, specific strategies to be taken aimed at minimizing the rates of suicide incidences.

4 Methodological Rigor and Ethical AI in Suicide Prediction

The research adopts machine learning methods as a quantitative approach that uses ensemble models with regression techniques to analyze suicide factors because they provide detection abilities for complex sociodemographic-economic data patterns and relationships. The ML workflow follows the same sequence as the research question as it incorporates data cleaning followed by exploratory analysis and feature engineering before model training to find concealed patterns and main predictors and useful insights. The internal validity of this model remains strong thanks to thorough data preprocessing and feature selection processes and external validity exists because broad cross-sectional factor analysis including age, gender and economic downturns provides universally applicable insights. Trustworthiness and avoidance of model biases result from implementing ethical AI principles that build responsible qualitative aspects through fairness accountability and transparency. The structured pipeline for ML implementation along with detailed model evaluation and data processing protocols makes findings replication easier as they help support suicide prevention policy development which provides transparent actionable strategies for public health. The work focuses on ethical aspects of fair AI methods together with consent regulations and confidentiality guarantees following responsible principles of AI-driven intervention. The research implements machine learning structure with ethical AI principles to establish data-focused suicide prevention systems which promote methodological strength and societal accountability.

5 Methodology

This work uses a sophisticated artificial neural network to examine the factors that affect the suicide rates including the sociodemographic and the soci-economic factors mentioned in the problem statement[12].The proposed methodology uses exploratory data analysis (EDA) and machine learning methodology to identify the underlying structures, predictors, and patterns to inform public health interventions as described in the conceptual framework in Figure 1.

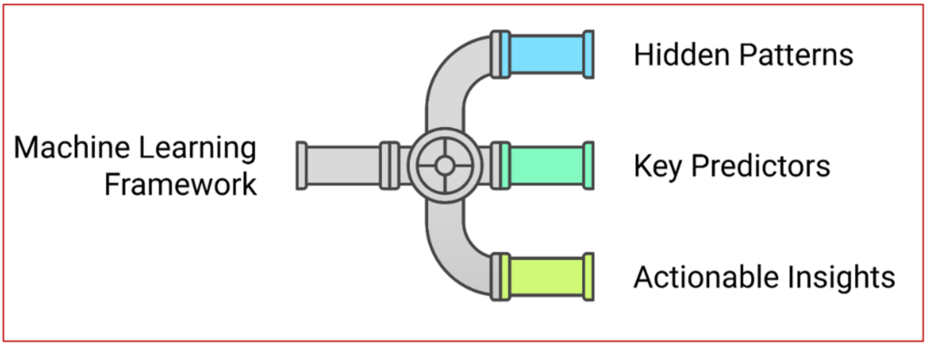

Figure 1. Machine Learning Analysis of Suicide Rate Determinants

Table 2. Algorithmic approach in suicide rate analysis

• **S1:**Load and preprocess data FUNCTIONLoad_Preprocess_Data() • **S2:**Load, handle missing values, encode, scale END FUNCTION • **S3:**Data exploration and model preparation FUNCTION Explore_And_Prepare_Data() • **S4:**Perform EDA, split data into train/test END FUNCTION • **S5:**Model training and validation FUNCTION Train_Validate_Modelse()	• **S6:**Train models, perform cross-validation END FUNCTION • **S7:**Testing and insights generation FUNCTION Test_And_Generate_Insights() • **S8:**Test models, generate and link insights END FUNCTION • **S9:**Execute the simplified process flow Load_Preprocess_Data() Explore_And_Prepare_Data() Train_Validate_Models() Test_And_Generate_Insights() END

The details of an algorithmic approach in suicide rate analysis is given in Table 2. The historical suicide rates stand as the first theoretical dependent variable; the second theoretical dependent variable incorporates demographic and socio-economic status such as age, gender, unemployment, and financial position [13]. This is the details List of pre-processing techniques that is included are missing value management, data cleansing, feature generation, categorical variable coding, numerical feature scaling and bias normalization. In the EDA, associations between data, data patterns and any outlying data are determined. On paper, an age-group suicide bar chart is illustrated in the Figure 2. Age-gender suicide rates are depicted by graphs. All groupings show male suicide rates (orange) are higher than female suicide rates (blue) Men and women aged 15–29, 30-44, and 45–59 have the highest suicide rates (15,000+ individuals). 30% of suicides are from the 15-29 age group and 15-17 are 4.870 60+ and 30-45 are moderate hazards 5000-100 The last category “0-100+” remains rare with 5000 male but is insignificant for females

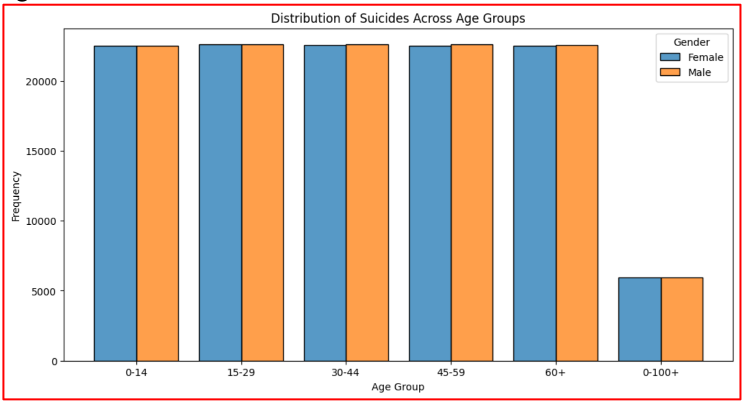

Figure 2. Frequency of Suicides by Age Group and Gender.

The predictors were designed to accommodate the first level of interactions between the demographic and economic variables with the view of incorporating, for instance, trends in unemployment rate and age-gender proportions. The analysis of feature importance with the employment of Random Forest was carried out to identify predictors that most significantly correlate with the suicide rates. Random Forest and Ridge Regression have been built to find non-linear relationship and to generate explainable forecasts. [14] The used models were tuned by hyperparameters with the help of the grid search method. Model performance was evaluated using metrics such as R^2, MAE, and RMSE, ensuring robust predictions in Figure 3. presents a scatter plot of residuals versus actual suicide values, indicating how well the model predicted relatively low and mid-level suicide numbers as is evident in the graph. Regarding the shaping of the data, on the residual plot of the Random Forest Regressor, one can observe the following relationship of the actual value ranging from the 0 to 60000 with the residuals from about -3000 to 2000. Using plot, in this case the values of the residuals are shown to be clearly dependent on the actual values with the trend indicated by red line showing a negative slope on the left side of the graph. Lower absolute residuals are evident for actual amounts below 20,000; this can be seen where most of the points lie on the zero line. However, when actual values get higher than this, the residues go higher and seem to be slightly negative, indicating that the model gives lower estimates of the actual values when they are high. This pattern is derived primarily when actual values are greater than 40,000, so that the residuals are negative up to the extent of -3000.

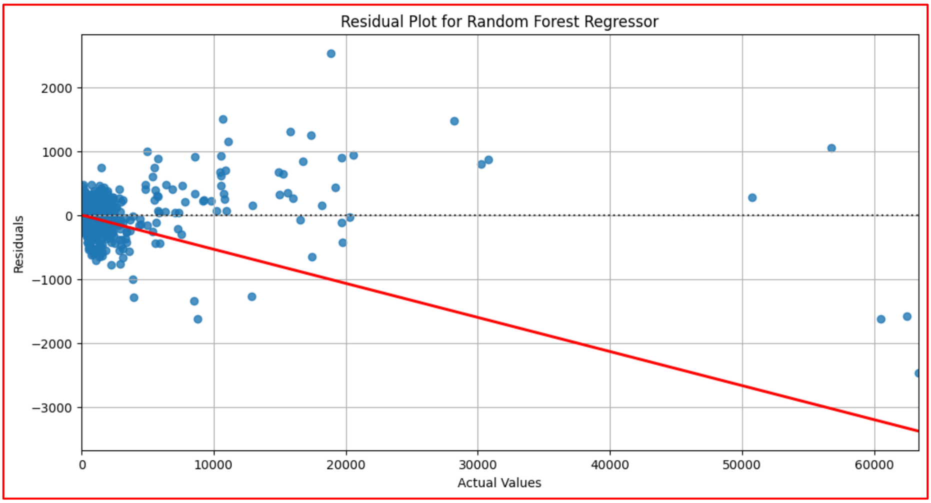

Figure 3. Residual Plot for Random Forest Regressor Showing Prediction Errors.

Public Health Integration, key predictors identified by the model are directly linked to actionable public health strategies, focusing on targeted interventions for high-risk groups and regions. To analyze the determinants of suicide rates, the study employs Random Forest and Ridge Regression, two complementary machine learning models. These models address the multifactorial nature of suicide determinants, capturing both non-linear interactions and providing interpretable insights. [15] Random Forest, an ensemble learning technique, combines the predictions of multiple decision trees. [16] The overall prediction is defined as in (Eq.3):

$$\hat{Y} = \frac{1}{M}\sum_{m=1}^{M} T_m\,(X)(3)$$

where:

- $\hat{Y}$= the aggregated prediction (final output of the model),
- M = total number of decision trees in the ensemble,
- T_m (X) = predictionmade by the m-th decision tree for input X.
- Feature importance in Random Forest is determined by analyzing the reduction in impurity at splits where a feature is used. This is calculated as shown in (Eq. 4):

$$FeatureImportance_i = \frac{1}{M}\sum_{m=1}^{M}\sum_{s\in Si}\Delta I_s \qquad (4)$$

where:

- Feature $Importance_i$ = importance of feature i

- S_i= set of all splits involving feature i,
- ΔI_s = impurity reduction at split s,
- M = total number of decision trees in the Random Forest.

This ability to rank features by importance allows Random Forest to identify significant predictors, such as age, gender, unemployment rates, and financial instability, making it a robust tool for understanding non-linear and interactive effects.[17] Ridge Regression, a regularized linear regression model, ensures that predictions are interpretable while minimizing overfitting. It achieves this by optimizing the following objective function as shown in (Eq. 5):

$$J(\beta) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{5}$$

where:

- J(β)=the objective function to minimize,
- y_i =actual value ofthe i-th observation,
- $\hat{y}_i$ =predicted value of the i-th observation,
- $g = \beta_j =$ coefficient of the j-th feature,
- λ =regularization parameter controlling the penalty applied to large coefficients,
- n=total number of observations(data points),
- p=total number of predictors (features).

The regularization term $\lambda \sum_{j=1}^{p} \beta_j^2$ ensures that the model avoids overfitting by shrinking large coefficients, which is particularly helpful when features are correlated or when datasets are relatively small.[18] Ridge Regression provides interpretable coefficients that indicate the magnitude and direction of each predictor's impact on suicide rates. Together, these models form a robust analytical framework. Random Forest identifies critical predictors and captures non-linear relationships, while Ridge Regression provides clear and actionable insights through its interpretable coefficients. This dual approach aligns with the study's objectives, ensuring a comprehensive understanding of the determinants of suicide rates and enabling data-driven public health strategies. Evaluating the performance of machine learning models is essential to ensure their predictions are accurate, reliable, and actionable. This study utilized three key evaluation metrics. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2 Score). These metrics were carefully chosen to assess the precision, robustness, and explanatory power of the models. The Mean Absolute Error measures the average magnitude of the errors in the model's predictions. Unlike other metrics, MAE treats all errors equally, making it an intuitive and interpretable metric.

It answers the question: “On average, how much are the predictions off from the actual values?” A lower MAE value indicates better model performance, as it signifies smaller prediction errors as in (Eq. 6). [19]

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{6}$$

Where

- y_i : Actual value of the i-th observation.
- $\hat{y}_i$: Predicted value of the i-th observation.
- n : Total number of observations.

MAE is particularly useful in this study because it provides a straightforward measure of prediction accuracy without exaggerating the impact of larger errors.[20] This is crucial when evaluating the overall effectiveness of the models in predicting suicide rates. Ridge Regression MAE: 0.8097 and Random Forest MAE: 0.2932. The significantly lower MAE for Random Forest indicates that its predictions were much closer to the actual values compared to Ridge Regression. This highlights the strength of Random Forest in minimizing average prediction errors. The Root Mean Squared Error goes a step further than MAE by squaring the prediction errors before averaging them, giving more weight to larger errors. By doing so, RMSE penalizes significant deviations, making it especially valuable for detecting models that struggle with outliers as in (Eq. 7). It answers the question: “What is the magnitude of typical prediction errors, with an emphasis on larger discrepancies? [21]

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{7}$$

Where:

- y_i:Actual value of the i-th observation.
- $\hat{y}_i$:Predicted value of the i-th observation.
- n:Total number of observations.

RMSE’s sensitivity to larger errors is particularly useful for evaluating models in a sensitive societal context like suicide prediction, where large deviations in predictions could lead to inappropriate interventions or misallocation of resources. Ridge Regression RMSE: 1.0307 Random Forest RMSE: 0.4942. The lower RMSE for Random

Forest demonstrates its ability to minimize significant deviations in predictions, making it a more robust model for this analysis.[22] Ridge Regression, with a higher RMSE, shows limitations in handling extreme cases. Coefficient of Determination (R^2 Score), the R^2 Score measures how well the model explains the variance in the target variable. An R^2 value closer to 1 indicates that the model captures most of the variability, while a value near 0 suggests poor performance[23] as in (Eq. 8). It answers the question: "What proportion of the variability in suicide rates can be explained by the model's predictions?"

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \quad (8)$$

Where:

- y_i : Actual value of the i-th observation.
- $\hat{y}_i$: Predicted value of the i-th observation.
- $\overline{y}$: Mean of actual values.
- n :Total number of observations.

R^2 provides an overall measure of the model's ability to explain the underlying patterns in the data. In this study, a high R^2 score is crucial for ensuring that the model captures the complex relationships between sociodemographic and socioeconomic factors. Ridge Regression R^2: 0.6415, Random Forest R^2: 0.9176. Random Forest's high R^2 value of 0.9176 demonstrates its capability to explain 91.76% of the variance in suicide rates, compared to Ridge Regression's 64.15%. This highlights Random Forest's effectiveness in capturing non-linear and interactive relationships among the predictors. The metrics collectively demonstrate that Random Forest significantly outperformed Ridge Regression across all measures: MAE and RMSE: Random Forest consistently minimized both average and large prediction errors. Higher R^2 Score: Random Forest captured a much larger proportion of the variance in the suicide rates, aligning closely with the study's objectives of uncovering hidden patterns. These results validate the choice of Random Forest as the primary model for this analysis, while Ridge Regression offers valuable interpretability through its coefficients. Together, these models provide a robust framework for understanding the determinants of suicide rates and informing data-driven public health strategies.

5.1 Implementation

In this work, the method of systematic analysis of the relations between sociodemographic and soci-economic characteristics of suicidality is used. The key strength of the methodology is in combining state-of-art machine learning algorithms, methodical data preprocessing and analysis that guarantees a sound approach to deriving insights from the data. Artefacts & key steps: Through the artefacts, the critical steps in the research process are depicted to reinforce or explain complex processes that relate to the research and in use when presenting the findings to subsume from raw data to usable public health strategies as demonstrated in the Figure 4.

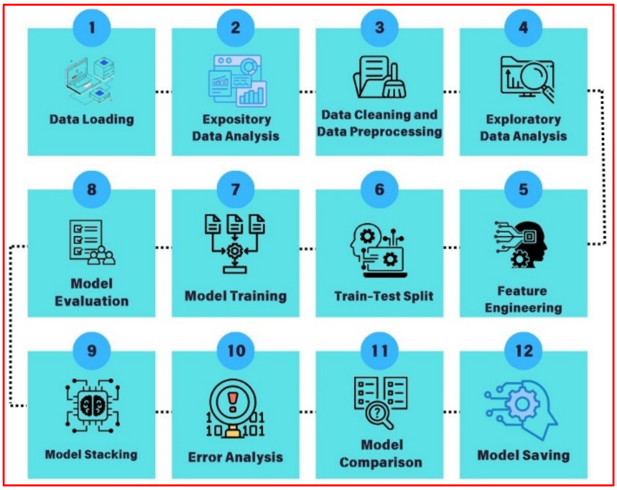

Figure 4. Data Science Workflow. Sequential steps from data loading to model saving in a machine learning project.

Scarcity of records of suicidal incidents that have happened in India in the past is tackled through the loading of a comprehensive dataset. This is the stage that ushers in the first step of the approach. Besides the demographic data including gender, age group and state status of the client, this dataset includes other socioeconomic factors like the unemployment rate and the states in financial crises. Some of the reliable sources of data included in the study are shown below to participate in obtaining the data and the initial check of the quality test to know the existence of missing entry, duplicate rows, and contradiction as shown in Figure 5. This was done in order to upheld the reliability and correctness of the data that was collected as much as it could be possible. The most common type of presentation is the Pie charts whose title is ‘Gender Distribution of Suicides’ which give a near parity on the proportion of suicidal individuals in males and females. That males account for 50.1% of all suicides while accounting for much higher rates of the other forms of deaths, is alarming in light of the fact that females account for 49.9% of all suicides. This paper has presented evidence that boys and females have almost the same involvement in suicide incidents. The only difference between the two groups is that girls are a bit less likely to commit suicide than males, though with a difference of only 0.2%. This is the only reason that the two groups are different from each other.

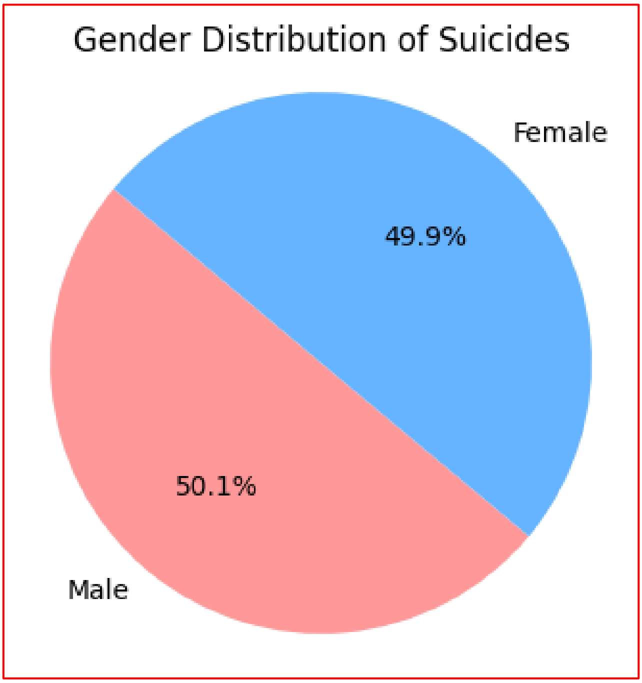

Figure 5 Pie chart showing the nearly equal gender distribution of suicides

As depicted in Figure 5, the pie chart highlights significant disparities between male and female suicide rates, providing an initial understanding of the demographic patterns that shape the dataset. The second significant task was data pre-processing where the dataset was made clean, check for consistency and compatibility that was fit for modeling. The insights where numerical values were missing, which are present in metrics like unemployment rate, were filled by mean imputation to retain the middle tendencies of the path. Missing entries in categorical variables were replaced with most frequently occurring category in that variable for better uniformity. Other attributes such as gender, state and kind of cause were further coded using one hot encoding in order to enhance the machine learning process. To improve scalability, and reduce the influence of features with the larger ranges of variations such as unemployment rate, bankruptcy rate, the values of measures of central tendencies were standarized. Using box plots as presented in Figure 6 outliers were identified and dealt with appropriately they include stuff such as financial distress and age group. The box plot provides a visual comparison between two datasets, labeled "Fea1" and "Fea2". The "Fea1" feature displays an extremely limited range of data, with all points closely clustered near zero, suggesting little to no variation among its values. Conversely, "Fea2" shows a broad distribution of values. The average specifically for this feature is roughly 1000, and the IQR is defined to be between 500 to 1500 . Notably, "Fea2" has many outliers that significantly exceed the upper quartile, extending up to about 60,000. These characteristics indicate a substantial variability in "Fea2", contrasting sharply with the near uniformity observed in "Fea1".

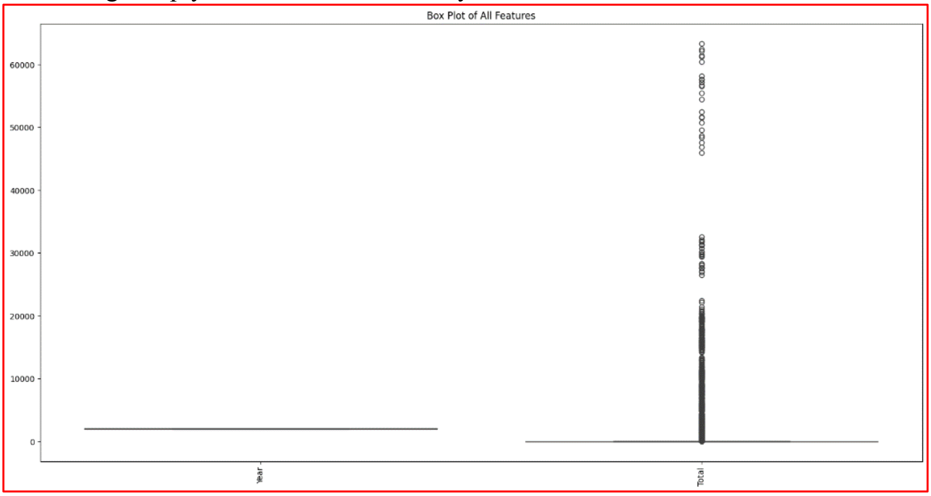

Figure 6. Box Plot of All Features: Highlighting distributions and outliers, particularly in the 'Total' feature.

The Figure 7 is a correlation matrix that color-codes relationships between economic indicators: Year, Total, Unemployment Rate and Bankruptcy Rate. The Year feature bears virtually no relation to other variables, boasting low negative coefficients with Unemployment (-0.015) and Bankruptcy Rate (-0.022) as well as a very low and positive coefficient with Total (0.0052). Total shows a negative but insignificant correlation with Unemployment Rate (-0.0076), and a very weak positive correlation with Bankruptcy rate = 0.019. Unemployment Rate and Bankruptcy Rate have a minimal positive correlation (0.051), this implies that employment changes, although at a slow level, may be proportional with bankruptcy changes. The matrix that has been developed demonstrates that these variables have little of interaction as revealed by the low coefficients.

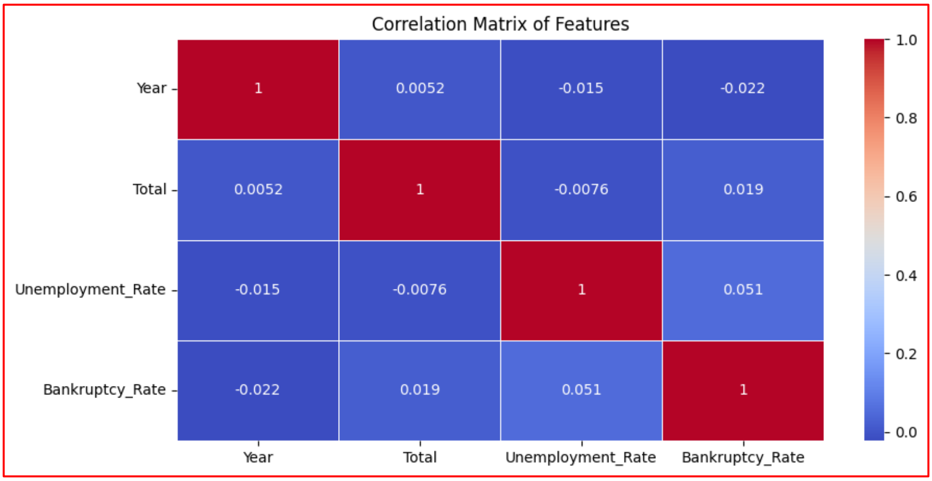

Figure 7. Correlation Matrix of Features: Depicting relationships between key variables, showing minimal correlation between socioeconomic factors and suicide rates.

The bar chart in Figure 8 shows suicides by age, with males orange and females blue. Female and male 0-14-year-olds had the lowest suicide rates on frequency. The 15-29, 30-44, and 45-59 age groups had the highest frequency, close or above 20000 per sex. At 60+, the age frequency distribution table drops from 15000 to 20000. The outlier category "0-100+" has declined to below 5,000 for males and much lower for females, indicating few cases. From young adulthood to late middle age, suicide rates rise while youngest and oldest groups fall.

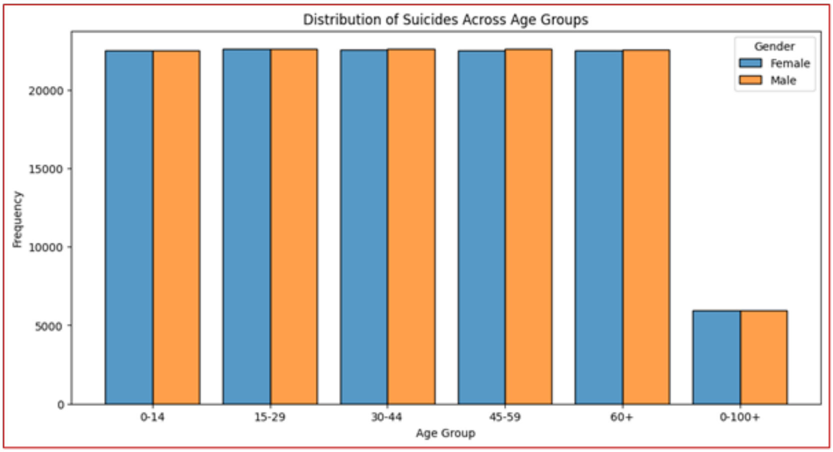

Figure 8. Suicide frequencies by age and gender

The Fig. 9 presents the total number of suicides per year for the period between 2002 and 2012, which indicates a rising trend to some extent, but declines towards the end of the time frame. Beginning from about one million in 2002, the suicides increase every year, to about 1.05 million in 2006. It keeps increasing steadily onwards to approximately 1.2 million in the 2010 fiscal year. Following this peak, the graph decreases slightly in 2011 before relatively hovering around 1.15 million in 2012. The presented visual data suitably captures the magnitude of suicide rates for a decade which shows a higher incidence most of the time, and a slight decline in the most recent years.

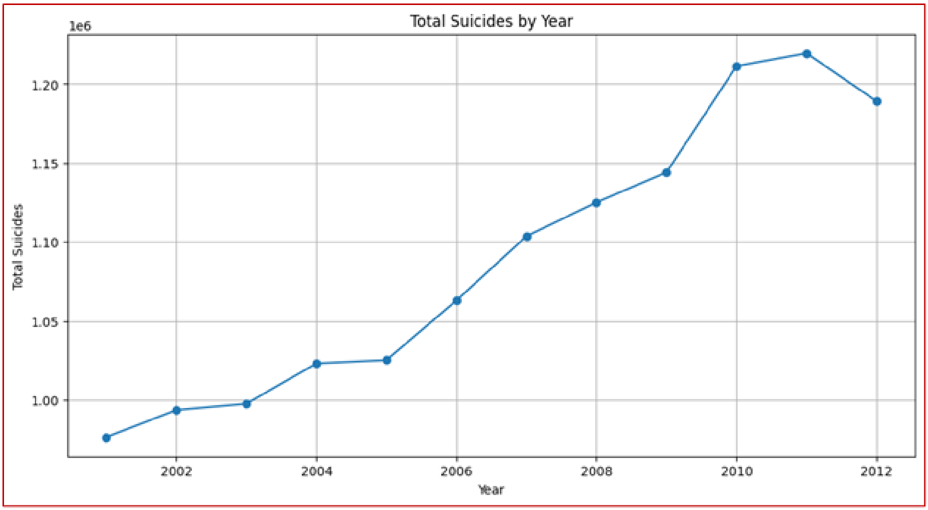

Figure 9. Line graph depicting the total number of suicides annually.

Figure 10 illustrates the significance of various features in a study, with the importance measured on a scale from 0 to 0.06. At the top of the chart, "Type_By Hanging" emerges as the most influential factor with an importance of approximately 0.06, indicating it has a substantial impact in the context of the study. After this we have Karnataka and West Bengal with an importance around 0.05. Tamil Nadu also appears to have an appreciable impact with a normalized coefficient of 0.0417 on the importance dimension while Andhra Pradesh too has a coefficient of 0.0387 on the same dimension of importance. Moving to demographics, "Age_group_60+" has a smaller yet significant importance of about 0.03. "Gender_Male" and "Age_group_0-14" are also key factors, each registering an importance just under 0.03. The lowest significant features are related to social and educational status ("Type_code_Social_Status" and "Type_code_Education_Status"), each with an importance just above 0.02. The distribution of feature importance serves to uncover the main drivers in the analysis, which would indicate that the method of suicide together with geographical location and with age as well as gender.

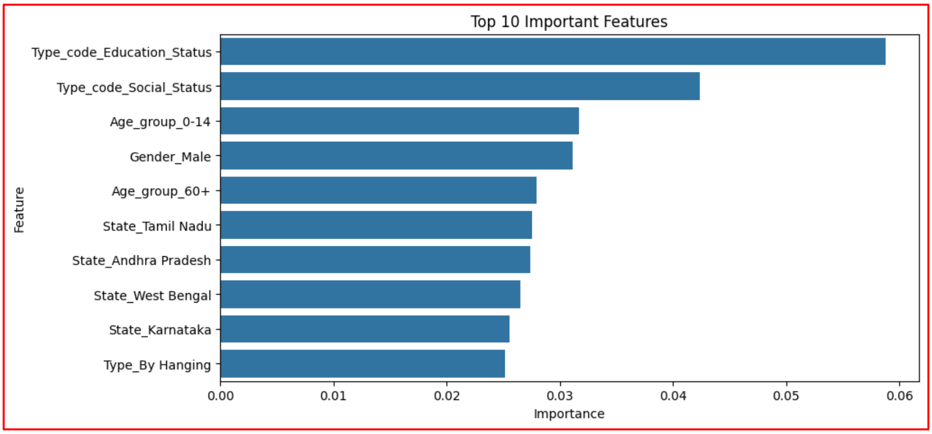

Figure 10. Bar chart of the top 10 features impacting suicide rates, highlighting 'Type_By Hanging' and key state and demographic variables.

To prevent overfitting, the developed models were cross-checked using a 5-fold cross validation technique to check the efficiency of the models being developed. [24] Performance evaluation involved the use of Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and coefficient of determination (R^2).

6 Results and Discussion

This study provides a detailed evaluation of the factors influencing suicide rates through machine learning models and feature importance analysis. The results highlight key predictors and offer actionable insights for public health strategies, with a focus on reducing suicide rates by addressing critical sociodemographic and socioeconomic factors. The model evaluation revealed that Random Forest significantly outperformed Ridge Regression across all metrics, demonstrating its ability to explain over 91% of the variance in suicide rates with an R^2 score of 0.9176. It achieved a Mean Absolute Error (MAE) of 0.2932 and a Root Mean Squared Error (RMSE) of 0.4942. In comparison, Ridge Regression, though interpretable, achieved an R^2 score of 0.6415, an MAE of 0.8097, and an RMSE of 1.0307, indicating a lower predictive accuracy. These findings validate the use of Random Forest for capturing non-linear relationships between predictors and outcomes. [25] These residual plots in Figure 11 demonstrate that Random Forest has low prediction errors and does not have system biases and therefore is robust. Not less than six thousand (three minusing from six thousand to plus two). Notice as the actual values rose, divergence values were negative and therefore negative plot slopes were represented by red lines. Residual bar plot for real numbersenger 0-10000 minimal and maximum combined with frequency table is -1000 to 1000, moderate accuracy of the forecast. Thus, when values are more than 20,000, the value of residuals is actually negative, and the maximum value is observed in the range from 50 000 to 60 000. This shows that the model always underestimates values, but as values increase, it reduces like this:

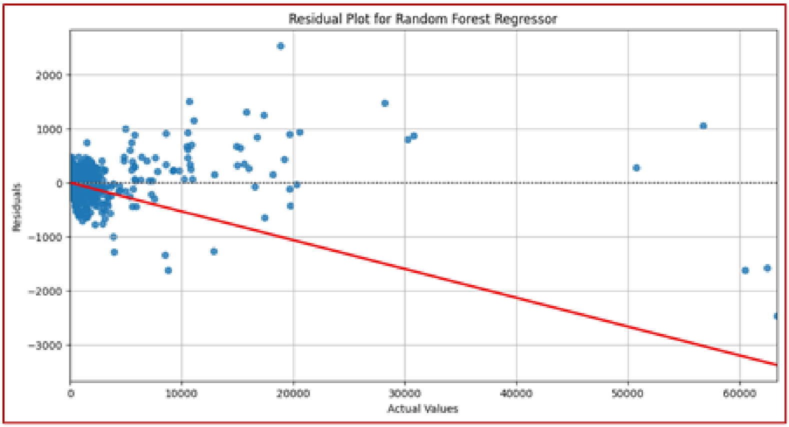

Figure 11. Residual plot for a Random Forest Regressor showing the relationship between actual values and residuals

The Figure 12 displays the relationship between predicted and actual suicide numbers, illustrating the performance of a predictive model across different data ranges. For actual suicide values between 0 and 20000, the predicted suicide values are almost identical to actual suicide values, and this shows the high models efficiency in this region in that most of the plotted points fall on the line of unity. In the case of actual values increasing from 20,000 to 40,000, the model keeps reasonably accurate; however, several points are seen to define a slightly imperfect straight line. Nevertheless, for the values of 40, 000 to 60, 000, the points are moving far above the equality line, implying that the model overestimates the cases of suicides. These phenomena show that the model of societal suicide level to this societal level is accurate for lower level of suicides but might need refining to make it effective for higher suicide rate.

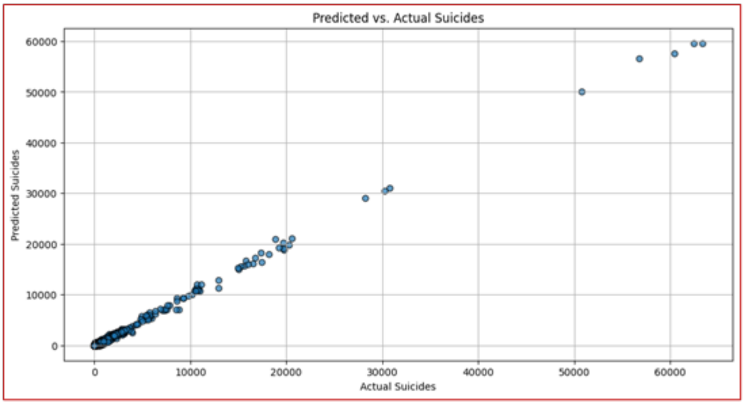

Figure 12. Scatter plot comparing predicted vs. actual suicides

The Figure 13 visualizes the relationship between varying unemployment rates and the number of predicted suicides. There are 5 categories on the x-axis ; unemployment rate starting from 4.5% and ending at 6.5%, and there are 6 categories on the y-axis; the number of predicted suicides ranging from 0 to 600. The data shows that predicted suicides are almost nil when the unemployment rate is as low as 4.5. However, 600 suicides are predicted at an unemployment rate of 5.0, and remains constant up to 5.5 and 6.0 rates respectively. This means that at an unemployment rate of 6.5%, the predicted suicide deceased to approximately one hundred. These results imply a marked non-linearity in the relationship between predicted counts of suicides to change in the unemployment rate– there are higher suicide predictions at moderate unemployment rates before tapering off at higher unemployment rates.

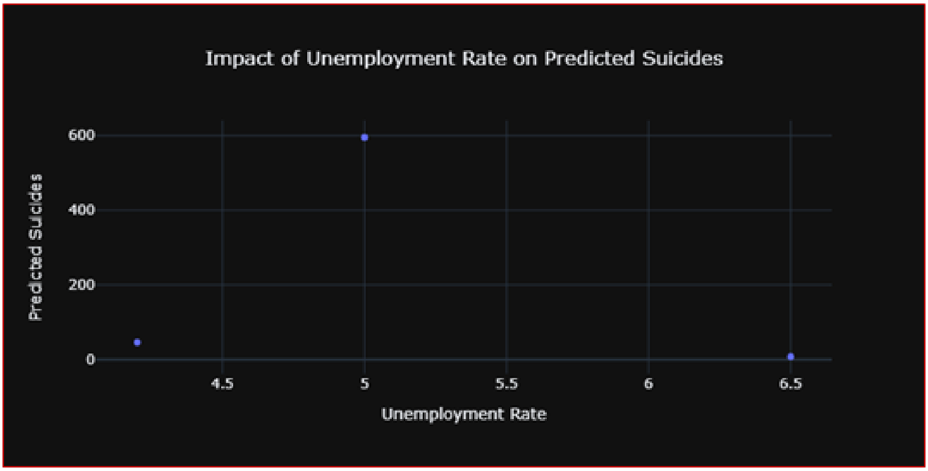

Figure 13. Impact of Unemployment Rate on Predicted Suicides

The Figure 14 displays a correlation between varying bankruptcy rates and the corresponding predicted number of suicides. The horizontal axis focuses on bankruptcy rates varying between 1.8 and 3.0, and the vertical axis—that of the predicted suicides, ranging from 0 through 600. First, going by a predicted 1.8 bankruptcy rate, extrapolated suicides would be about one hundred. Bankruptcy rate number 2 stands at 2.0 and with that, the predicted number of suicides rockets up to nearly 600. When the rate rises to 2.2, predicted suicides stand at about 400. The trend once again decreases as the bankruptcy rate increases to 2.8: predicted suicides return to virtually 100. According to the model, when the bankruptcy rate is 3.0 or less the predicted suicides are near zero. This particular graph does show an inverted U-shape; it goes upward to exhibit a high value of 2.0 at the bankruptcy rate level before going downward steeply as the bankruptcy rate goes up.

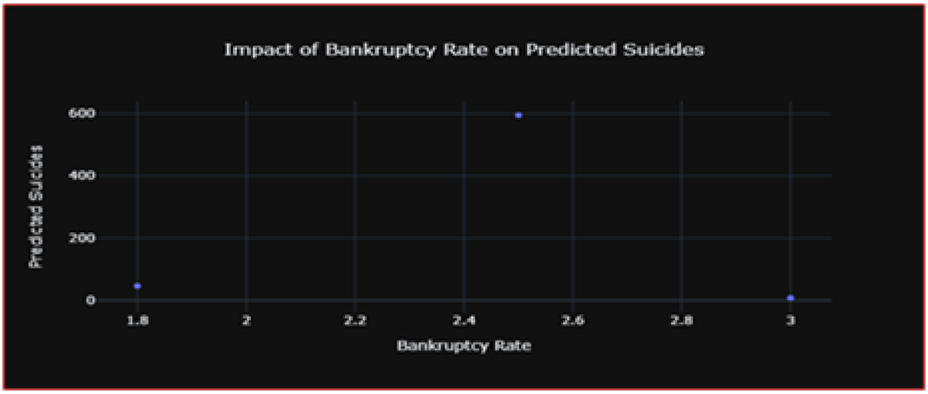

Figure 14. Impact of Bankruptcy Rate on Predicted Suicides

Altogether these results were culminated into evidence based and targeted health promotion and disease prevention strategies with special emphasis on Population-Specific Interventions and System Update. Programs for males which might be effective include educational campaigns and having peer support to fight societal pressure and stigma related to mental health for young male aged between 15-29 years. Separating cultural factors including dowry conflicts could be reduced and its effects on their mental health controlled by applying further legal changes and increasing awareness. Proactive and effective public health policies should also extend mental health facilities and clients in these areas and guarantee fairness of these services. As this paper was developing the modeling process, ethical values were given significant importance so that the process was rightfully and responsibly fair and transparent. These sources of bias were systematically prevented by using stratum sampling and a rigorous screening of the features with a strict control of the demographic variable frequencies. Blind interpretable models like Regularized Linear Models including Ridge Regression were used to ensure accountability and provide the stakeholders with the impacts of predictors. Academic credibility was further achieved by incorporating greater documentary and replicative substantiation of the methodology so as to enable future revision and augmentation by other independent researchers and policy makers and local stakeholders. Specifically, the current study highlights the intercessional character of suicide rates which are influenced at the same time by both economic and demographic as well as familial factors. By identifying these drivers, and developing evidence based interventions, there is the chance to reduce suicide rates and improve mental health. The findings of this study provide a solid groundwork for policy creators to rely on when formulating best practices on how to use machine learning to address one of the society's biggest healthcare concerns.

7 Conclusion

This work aimed at investigating the multifactorial nature of sociodemographic and sociogeonomic predictors for suicide incidence and mortality by applying novel artificial neural network algorithms. Applying models of the decision tree, Random Forest and Ridge Regression the article focused on the shortcomings of statistical treatment and identified relatively significant predictors such as unemployment rate, family problems, the rate of bankruptcy, age category, gender. Random Forest resulted in slightly higher accuracy and interactions with an R^2 of 0.9176 than Ridge Regression while the latter had coefficients which highlighted affects of individual predictors. Economic parameters are next as major influential determinants of suicide rates for which robust economic policies and measures have been underlined. Population-specific risks are also highlighted, which show specifically young men and males aged 15–29 targeted interventions considering age- and gender-specific issues. Thus cultural and family issues such as dowry abuses and family issues signify the need for social revolution and mental health for girl child in such settings. The following are some of the ways through which ethical considerations were incorporated in the study all in ech attempt to work fairly, accountably, and transparently in the use of predictive modeling. Demographic confounders were reduced with the help of stratified sampling and bias detection; variable importance scores and human-interpretable coefficients sustained the comprehensible insightfulness of the results. With the integration of these ethical AI principles on the study is aligned with the other goals of Responsible Research and Effective Public Health Interventions. In conclusion, it could be stated that all the objectives formulated at the beginning of the study have been achieved at a satisfactory level, as the paper offered a number of potential predictors of suicide rates and provided an empirical background for evidence-based practice. These are the reasons that would require the formulation of policies seeking to eliminate high rate of suicide in the absence of economic fluctuation, mental health labelling and dull culture. Thus, the found solutions, based on the recognition of the discussed issues' causes at the same time and through the application of machine learning, prove the developming capability for this approach for socially oriented solutions to the most significant and dangerous social problems, including and not limited to the issue of suicide.

8 Future Work

Consequently, this study provides a strong foundation for dispelling injustice and understanding the sociodemographic and socioeconomic antecedents of suicide. The study could be extended for improved results to gather real time or further detailed data encompassing social media and crisis helpline data for trending analysis and intervention. However, deciding to use data at a district or community level could help with carrying out analysis and treatment plans for that region. Despite these accomplishments for the future work, the integration of explainable AI techniques such as SHAP or LIME can further improve the interpretability of the model to policymakers and health professionals. Some methodological improvements could be integrating multi-output models and deep learning network to investigate the interdependency and temporal patterns to analyze the depth and periodically patterns or even the varied and global and regional factors for suicide. These steps would stand on the present development and take forward the goals of suicide prevention and the public health initiatives.

References

1. Kim, S., Park, J., Lee, H., Lee, H., Woo, S., Kwon, R., Kim, S., Koyanagi, A., Smith, L., Rahmati, M., Fond, G., Boyer, L., Kang, J., Jun Hyuk Lee, Oh, J., Dong Keon Yon: Global public concern of childhood and adolescence suicide: a new perspective and new strategies for suicide prevention in the post-pandemic era. World Journal of Pediatrics. (2024). https://doi.org/10.1007/s12519-024-00828-9.
2. None Anjali, Kumar, B.R.: Spatial analysis of multivariate factors influencing suicide hotspots in Urban Tamil Nadu. Journal of Affective Disorders Reports. 16, 100741–100741 (2024). https://doi.org/10.1016/j.jadr.2024.100741.
3. Saravanan N, Moheshkumar G, V.M Mohammed Shaid, S Purushothman, V Gokul Sanjai: Accurate Prediction and Detection of Suicidal Risk using Random Forest Algorithm. (2024). https://doi.org/10.1109/icpcsn62568.2024.00053.
4. Mesiar, R., Sheikhi, A.: Nonlinear Random Forest Classification, a Copula-Based Approach. Applied Sciences. 11, 7140 (2021). https://doi.org/10.3390/app11157140.
5. Jiang, H., Niu, L., Hahne, J., Hu, M., Fang, J., Shen, M., Xiao, S.: Changing of suicide rates in China, 2002–2015. Journal of Affective Disorders. 240, 165–170 (2018). https://doi.org/10.1016/j.jad.2018.07.043.
6. Saravanan N, Moheshkumar G, V.M Mohammed Shaid, S Purushothman, V Gokul Sanjai: Accurate Prediction and Detection of Suicidal Risk using Random Forest Algorithm. (2024). https://doi.org/10.1109/icpcsn62568.2024.00053.
7. Pirkis, J., John, A., Shin, S., DelPozo-Banos, M., Arya, V., Analuisa-Aguilar, P., Appleby, L., Arensman, E., Bantjes, J., Baran, A., Bertolote, J.M., Borges, G., Brečić, P., Caine, E., Castelpietra, G., Chang, S.-S., Colchester, D., Crompton, D., Curkovic, M., Deisenhammer, E.A.: Suicide trends in the early months of the COVID-19 pandemic: an interrupted time-series analysis of preliminary data from 21 countries. The Lancet Psychiatry. 0, (2021). https://doi.org/10.1016/S2215-0366(21)00091-2.
8. Menon, V., Arafat, S.M.Y., Akter, H., Mukherjee, S., Kar, S.K., Padhy, S.K.: Cross-country comparison of media reporting of celebrity suicide in the immediate week: A pilot study. Asian Journal of Psychiatry. 54, 102302 (2020). https://doi.org/10.1016/j.ajp.2020.102302.
9. Sanderson, M., Bulloch, A.G., Wang, J., Williamson, T., Patten, S.B.: Predicting death by suicide using administrative health care system data: Can recurrent neural network, one-dimensional convolutional neural network, and gradient boosted trees models improve prediction performance? Journal of Affective Disorders. 264, 107–114 (2020). https://doi.org/10.1016/j.jad.2019.12.024.
10. Du, J., Sadr, N., Philip de Chazal: Automated Prediction of Sepsis Onset Using Gradient Boosted Decision Trees. Computing in Cardiology (CinC), 2012. (2019). https://doi.org/10.22489/cinc.2019.423.
11. Fontanella, C.A., Saman, D.M., Campo, J.V., Hiance-Steelesmith, D.L., Bridge, J.A., Sweeney, H.A., Root, E.D.: Mapping suicide mortality in Ohio: A spatial epidemiological analysis of suicide clusters and area level correlates. Preventive Medicine. 106, 177–184 (2018). https://doi.org/10.1016/j.ypmed.2017.10.033.

12. Mohamed, E.S., Naqishbandi, T.A., Bukhari, S.A.C., Rauf, I., Sawrikar, V., Hussain, A.: A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms. Healthcare Analytics. 100185 (2023). https://doi.org/10.1016/j.health.2023.100185.
13. Mathieu, S., Treloar, A., Hawgood, J., Ross, V., Kõlves, K.: The Role of Unemployment, Financial Hardship, and Economic Recession on Suicidal Behaviors and Interventions to Mitigate Their Impact: A Review. Frontiers in Public Health. 10, (2022). https://doi.org/10.3389/fpubh.2022.907052.
14. Auret, L., Aldrich, C.: Interpretation of nonlinear relationships between process variables by use of random forests. Minerals Engineering. 35, 27–42 (2012). https://doi.org/10.1016/j.mineng.2012.05.008.
15. Ali, M., Prasad, R., Xiang, Y., Zaher Mundher Yaseen: Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. Journal of Hydrology. 584, 124647–124647 (2020). https://doi.org/10.1016/j.jhydrol.2020.124647.
16. Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., Liang, X.: An improved random forest based on the classification accuracy and correlation measurement of decision trees. Expert Systems with Applications. 237, 121549 (2024). https://doi.org/10.1016/j.eswa.2023.121549.
17. Yuan, X., Liu, S., Feng, W., Dauphin, G.: Feature Importance Ranking of Random Forest-Based End-to-End Learning Algorithm. Remote Sensing. 15, 5203 (2023). https://doi.org/10.3390/rs15215203.
18. Schreiber-Gregory, D.N.: Ridge Regression and multicollinearity: An in-depth review. Model Assisted Statistics and Applications. 13, 359–365 (2018). https://doi.org/10.3233/mas-180446.
19. Karunasingha, D.S.K.: Root Mean Square Error or Mean Absolute Error? Use their ratio as well. Information Sciences. 585, (2021). https://doi.org/10.1016/j.ins.2021.11.036.
20. Jierula, A., Wang, S., OH, T.-M., Wang, P.: Study on Accuracy Metrics for Evaluating the Predictions of Damage Locations in Deep Piles Using Artificial Neural Networks with Acoustic Emission Data. Applied Sciences. 11, 2314 (2021). https://doi.org/10.3390/app11052314.
21. Ćalasan, M., Abdel Aleem, S.H.E., Zobaa, A.F.: On the root mean square error (RMSE) calculation for parameter estimation of photovoltaic models: A novel exact analytical solution based on Lambert W function. Energy Conversion and Management. 210, 112716 (2020). https://doi.org/10.1016/j.enconman.2020.112716.
22. Schonlau, M., Zou, R.Y.: The random forest algorithm for statistical learning. The Stata Journal: Promoting Communications on Statistics and Stata. 20, 3–29 (2020). https://doi.org/10.1177/1536867x20909688.
23. Figueiredo Filho, D.B., Silva Júnior, J.A., Rocha, E.C.: What is R2 all about? Leviathan (São Paulo). 3, 60–68 (2011). https://doi.org/10.11606/issn.2237-4485.lev.2011.132282.
24. Sejuti, Z.A., Islam, M.S.: A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation. Sensors International. 100229 (2023). https://doi.org/10.1016/j.sintl.2023.100229.
25. Tran, T.K., Senkerik, R., Vo, H.T.X., Vo, H.M., Ulrich, A., Musil, M., Zelinka, I.: Initial Coin Offering Prediction Comparison Using Ridge Regression, Artificial Neural Network, Random Forest Regression, and Hybrid ANN-Ridge. MENDEL. 29, 283–294 (2023). https://doi.org/10.13164/mendel.2023.2.283.