Suicide Prediction: A Socioeconomic Analysis

Atharv Kadam
Ms Data science
University of Colorado Boulder
Boulder Colorado US
atka9416@colorado.edu

Om Kokate
Ms Data Science
University of Colorado Boulder
Boulder Colorado US
omko8394@colorado.edu

Yuni Wu
Ms Data Science
University of Colorado Boulder
Boulder Colorado US
ihwu8559@colorado.edu

ABSTRACT

Suicide is a major global public health problem, accounting for more than 800,000 lost lives annually, affecting individuals, families, and communities. Early identification and intervention are critical to reducing suicide rates, and ML offers a promising avenue for predicting suicide risk based on a combination of socioeconomic and demographic factors. The paper proposes a predictive framework for the risk of suicide using five popular machine learning models, including Random Forest, K-Nearest Neighbors, LightGBM, Support Vector Machine, and XGBoost. The paper tries to classify the levels of suicide risks-low, as medium, and high-by considering several feature variables like population size, Gross Domestic Product, happiness index, and average temperature. We have used the models mentioned above on a multi-country data set consisting of socio-economic and demographic information. The initial accuracy of the Random Forest was 93.4% and fine-tuned to 93.9%. It also exhibited a great performance in finding the high-risk cases while the recall for medium risk was the one that needed an increase. KNN model gave an exceptional accuracy of 97% with almost balanced precision, recall, and F1 scores, all of which were very decent for the classification task at hand. LightGBM, a gradient-boosting model, achieved an accuracy of 91% when fine-tuned, showing much-improved generalized performance and the ability to handle complex relationships in this data. The SVM, which was performing terribly with about 34% accuracy, was significantly improved after optimization to eventually reach an accuracy of about 81.7%. Another gradient-boosting algorithm, XGBoost, was able to achieve an accuracy of 92.58%, with only minor performance degradation post-tuning. Feature importance analysis showed that factors like population size, GDP, and the happiness index were critical in predicting suicide risk. These findings show the importance of socioeconomic data in enhancing the predictive capabilities of machine learning models. Overall, the results show that machine learning has great potential in improving suicide risk predictions. By incorporating socioeconomic indicators, our study offers a strong tool for the identification of high-risk individuals and supports timely interventions. Further research will consider the inclusion of data at finer resolutions and more complex models to increase the predictive accuracy further and discuss ethical issues related to the use of such tools in the prevention of suicide.

1. Introduction

Suicide is among the leading causes of death in the world, and the WHO estimates that by suicide, over 800,000 people die every year.. Despite extensive research that has been done on the prevention of suicide, predicting the risk for suicide remains a complex challenge due to the multifaceted nature of the problem. The development of suicidal behavior is linked to the history of mental health, social isolation, genetic predisposition, and environmental influences. Among the socioeconomic and demographic factors, poverty, income inequality, current employment status, and societal well-being have emerged as strong predictors of suicide risk. These elements have, however, not been incorporated into the model development completely.

ML has emerged as an effective tool that might enhance this prediction with individual and contextual factors across a wide span of big datasets. Unlike most traditional clinical methods that usually depend on symptoms of a psychological or behavioral nature, machine learning models can consider big and diverse datasets to find complex patterns and relationships that otherwise may not be easily noticeable. Such models use algorithms that learn from data to make predictions helpful in locating individuals at higher risk for suicide for timely interventions and prevention strategies.

This study investigates the use of five popular machine learning models—Random Forest (RF), K-Nearest Neighbors (KNN), LightGBM, Support Vector Machine (SVM), and XGBoost—for suicide risk classification based on a combination of socioeconomic and demographic variables. The goal is to classify suicide risk into three categories: low, medium, and high, using features such as population size, Gross Domestic Product (GDP), happiness index, and average temperature. They help in understanding mental health or integrating predictive models for suicide risk prediction because various elements impact them.

The models here utilize different metrics for evaluating model performances, including accuracy, precision, recall, F1 score, and confusion matrix; indeed, they show that Random Forest, along with KNN, outperforms other counterparts since KNN has the highest accuracy. Besides, the feature importance analysis underlined the significant contribution of socioeconomic factors in predicting suicide risk. The better predictive performance of the suicide risk models contributes to the elaboration of effective tools for suicide prevention and provides useful insights for both policymakers and mental health professionals.

2. Related work

During the last 10 years, several studies have looked at the use of machine learning techniques to predict suicide risk, highlighting its potential for prevention and early intervention. One disadvantage of traditional suicide prevention measures such as psychological testing and clinical interviews is the delay in identifying at-risk persons. As a result, data-driven strategies for detecting minute changes in large datasets that may indicate suicidal ideation have gained prominence.

2.1 **Prediction of Suicidality Using Machine Learning**

Kumar et al. (2019) used various machine learning algorithms such as Random Forest and Support Vector Machines for the prediction of suicide ideation using clinical features like age, gender, and history of previous diagnosis. Scientists observed that these machine learning models outperform traditional statistical models in identifying the high-risk population, with Random Forest ranking the first among all other approaches. Similarly, Jiang et al. (2020) have also trained numerous classification models based on socioeconomic variables of the individuals, such as their income and work status. Among other results, their output suggested that machine learning can spot hidden patterns in socio-economic data, which helps identify at-risk individuals even for economic classes that are disadvantageous.

2.2 **Socioeconomic Aspects of Suicide Prediction**

Research has raised more awareness about the role played by socio-economic factors determining an individual's risk of suicide. Previous research has shown that economic downturns, unemployment, and income inequality can lead to heightened rates of suicide (De Leo et al., 2013). In line with this, Shao et al. (2021) apply machine learning in modeling the association between socio-economic indices (like GDP and income inequality) and the probability of suicide in various nations. Indeed, it was found that the lower the GDP and higher income inequality in a nation were, the higher its rate of suicide, thus possibly relating to socioeconomic factors being some of the important predictors for modeling purposes.

2.3 **Critical Reviews of ML Models Pertaining to Suicide Prediction**

Different machine learning models for the prediction of suicides have been compared in various researches. Bismark et al. (2020) implemented a comprehensive analysis of different algorithms such as Random Forest, KNN, and XGBoost in order to predict suicide risk based on a wide range of demographic, socioeconomic, and clinical features. The authors reported ensemble approaches such as Random Forest, along with gradient-boosting techniques like XGBoost, that have always outperformed, with better accuracy and robustness, the simpler models such as KNN, especially when facing imbalanced datasets. Further, Li et al. (2022) highlighted how effectively hyperparameter tuning can work in enhancing model performance, especially on hard classification tasks such as multi-class suicide risk prediction.

2.4 **Integrating Demographic and Socioeconomic Information**

While most models for predicting suicide rely on clinical and mental health data, there is a growing interest in incorporating socioeconomic and demographic features as complementary data to offer comprehensive risk evaluation. Liu et al. (2018) explored the relationship between suicide rates and various socioeconomic factors, such as poverty rate, educational attainment, and access to healthcare. They showed that integrating these non-clinical factors improved the accuracy of suicide risk prediction models using a multi-layered data approach, especially for areas with poor access to mental health treatment.

Further support was provided by Meyer et al. 2021 and Hogan et al. 2023, in which the happiness index, environmental factors, and regional economic indicators were stated as non-clinical variables, which may provide useful information on suicide risk prediction. The research emphasizes the importance of environmental and socioeconomic context in understanding suicide behavior, reinforcing the need for their inclusion in machine learning models. Even as the literature is growing, there is considerable scope for improvement. Only a few large-scale, multi-country datasets can take regional variations in suicide risk into consideration, and most studies focus on a very small set of demographic and socioeconomic variables. Further, even as machine learning models show promise, important challenges related to model interpretability, class imbalance, and ethical dilemmas associated with the use of suicide prediction tools remain by and large unaddressed.

By constructing a machine learning framework that leverages an exhaustive set of socioeconomic and demographic characteristics, including but not limited to GDP, population size, happiness index, and average temperature, we present an extension of these initial findings in this paper. We give new insights into the role that socioeconomic factors play in predicting suicide by comparing the output of several models and judging their sensitivity to different variables.

3. Main Methods

The paper uses machine learning methods in predicting suicide risk based on the socioeconomic status and demographic features of the community. This paper will provide the methodology for data gathering, preprocessing, feature selection, model implementation, and the metrics for evaluation that will be employed.

3.1. **Dataset and Feature Selection**

The dataset to be used is a wide variety concerning socioeconomic and demographic series across many countries.

These include:

Population Size: This is a total population of the country or region.

GDP: Total economic output in purchasing power parity terms, for a country.

Happiness Index: This is a measure that summarizes subjective well-being based on life satisfaction and social support with objective indicators.

Average Temperature: A climatic variable to capture potential environmental influences on mental health.

These variables have been selected based on their previously established relevance to mental health and suicide rates by such works as De Leo et al. (2013) and Shao et al. (2021). Entries of several countries are included to widen the generalizability of results.

3.2. **Data Preprocessing**

Before training the model, the following preprocessing was done to the data:

Handling Missing Values: The missing data were imputed, in which the missing values for the continuous features, such as GDP and population size, were replaced by the mean or median values of the respective columns.

Categorical Encoding: The socioeconomic and demographic data, such as region-specific factors or country classification, were encoded into numerical values using Label Encoding or One-Hot Encoding, wherever appropriate.

Feature Scaling: Features like GDP and temperature went through Min-Max Scaling to ensure all the features fell within comparable ranges. This especially helped in giving better results with algorithms like KNN and SVM. Data Splitting: The entire dataset was split into an 80% training set and a 20% testing set, ensuring that the models tested without overfitting.The paper uses machine learning methods in predicting suicide risk based on the socioeconomic status and demographic features of the community. This paper will provide the methodology for data gathering, preprocessing, feature selection, model implementation, and the metrics for evaluation that will be employed.

	Region	Country	Year	AvgTemperature	Happiness Index	GDP	Population	Both sexes	Female	Male
0	Africa	Algeria	2005	62.913425	5.466833	3131.328300	32956690.0	3.82	2.80	4.83
1	Africa	Algeria	2006	64.930411	5.466833	3500.134528	33435080.0	3.65	2.66	4.63
2	Africa	Algeria	2007	63.166849	5.466833	3971.803658	33983827.0	3.46	2.51	4.41
3	Africa	Algeria	2008	63.532240	5.466833	4946.563793	34569592.0	3.31	2.40	4.22
4	Africa	Algeria	2009	64.259726	5.466833	3898.478923	35196037.0	3.15	2.29	4.02
5	Africa	Algeria	2010	64.268219	5.464000	4495.921455	35856344.0	3.00	2.19	3.81
6	Africa	Algeria	2011	64.960822	5.317000	5473.446129	36543541.0	2.94	2.13	3.74
7	Africa	Algeria	2012	64.290437	5.605000	5610.733341	37260563.0	2.89	2.10	3.68
8	Africa	Algeria	2013	63.704658	5.980000	5519.777576	38000626.0	2.85	2.06	3.64
9	Africa	Algeria	2014	65.195890	6.355000	5516.229431	38760168.0	2.78	2.02	3.53

Figure 1: Dataset (Before Transformation)

	Region	Country	Year	Male	Female	Both sexes	AvgTemperature	Happiness Index	GDP	Population	Actual	Predicted
0	Western Pacific	Malaysia	2010	7.94	2.53	5.30	83.427945	5.580	8880.146040	28717731.0	High	High
1	Europe	Spain	2009	8.97	2.47	5.62	59.790137	6.199	32037.209190	46362946.0	High	High
2	Americas	Canada	2009	15.53	5.04	10.24	41.353671	7.488	40918.850660	33630069.0	High	High
3	Europe	Norway	2006	16.91	6.38	11.65	43.173973	7.416	74427.565413	4660677.0	High	High
4	Americas	Cuba	2007	16.22	3.99	9.92	75.603562	5.418	5200.034393	11269887.0	High	High
5	Africa	Mauritania	2009	7.26	4.12	5.56	76.707397	4.500	1418.940825	3322616.0	High	High
6	Eastern Mediterranean	Kuwait	2006	3.71	1.15	2.72	80.731507	6.076	42971.392200	2363409.0	Medium	High
7	Africa	Gabon	2013	24.29	3.97	13.57	79.624932	3.800	9247.418332	1902226.0	High	High
8	Europe	France	2018	15.80	4.73	10.05	51.521233	6.666	41937.933915	67158348.0	High	High
9	Africa	Zambia	2014	30.20	7.87	17.74	56.945455	4.346	1696.117261	15737793.0	High	High

Figure 2: Dataset (After Transformation)

3.3. **Model Selection**

Then, five machine learning algorithms were employed to predict suicide risk: Random Forest, K-Nearest Neighbors, LightGBM, Support Vector Machine, and XGBoost. These models have been selected due to their high performance in classification tasks alongside the ability to handle huge and complex datasets.

Random Forest (RF): A form of ensemble learning involving the ensemble of many decision trees that, on the whole, prevent overfitting by themselves. RF is also able to handle linear and non-linear relationships.

K-Nearest Neighbors: This model is simple and relies on the idea of classifying new points based on the majority of their nearest neighbor classes. It does very well with data living in a very high dimensional space, has many useful interpretative metrics which might prove useful, and it also runs pretty fast for any dataset

.

LightGBM: This is a gradient-boosting algorithm developed for efficiency and scalability. It has been selected because it can handle big datasets with a lot of features while being highly predictive. Support Vector Machine (SVM): The powerful classification algorithm whose core concept is to find a hyperplane that best separates data into classes. SVM was included to assess its performance in multi-class classification problems despite it being computationally expensive.

XGBoost: Another gradient-boosting model, very well recognized for speed and high performance. Inclusion of XGBoost was done to check the efficiency of this algorithm on an imbalanced dataset with interaction among various features.

3.4. **Hyperparameter Tuning**

Further optimization of each model was then done by hyperparameter tuning via Grid Search for Random Forest, KNN, LightGBM, SVM, and XGBoost. Grid search involved trying a range of parameters, in other words, estimators concerning Random Forest, the number of neighbors with respect to KNN, and the learning rate pertaining to LightGBM and XGBoost, among others. Each of these best combinations is decided considering their performance on cross-validation. The main purpose of using a cross-validation method here would be to avoid overfitting the models.

Random Forest Hyperparameters: Number of trees, max features, max depth, and minimum samples for splitting and leaf nodes. KNN Hyperparameters: The number of neighbors, distance metric (Euclidean, Manhattan), and weights function. LightGBM Hyperparameters: Number of leaves, learning rate, number of boosting iterations, and maximum depth. SVM Hyperparameters: Kernel type-linear, polynomial, radial, regularization parameter-C, kernel coefficient-gamma. XGBoost Hyperparameters: Learning rate, maximum depth, subsample ratio, and number of estimators.

3.5. **Model Evaluation**

Below are the metrics used to evaluate the model. Accuracy: It defines the ratio of the total correct predictions by the model. Precision, Recall, and F1-score: These were the metrics for how the model correctly classified suicide risk levels, given the imbalanced nature of the classes in this dataset.

Confusion Matrix: The models show their performance on the confusion matrix based on true positives, false positives, true negatives, and false negatives per class: low, medium, and high risk.

Feature Importance: Feature importance analysis was performed to understand the contribution of each socio-economic variable in predicting the risk of suicide. It serves to identify the most relevant predictors of the risk of suicide, informing further refinement of models.

3.6. **Model Comparison**

The performance of the trained and fine-tuned models was then compared based on the above-mentioned metrics. The results were thus analyzed to find out which of the models was the best in terms of classification accuracy and which effectively handled the complexity of the data. The overall effectiveness of the models on feature importance and model interpretability was also considered.

4. Evaluation

Evaluation of the machine learning models developed in this study was based on their predictive accuracy and handling class imbalance. Key evaluation metrics used in this work involve accuracy, precision, recall, the F1 score, and confusion matrix. Feature importance analysis has been conducted to estimate the different socioeconomic and demographic factors' contribution in predicting the level of suicide risk. This section covers all five models-Random Forest, KNN, LightGBM, SVM, and XGBoost-which relate to each of the various different criteria for evaluation.

4.1. Performance Metrics

The various metrics that were used for the performance evaluation of each model are as follows:

4.1.1 Accuracy: It is the ratio of correctly predicted observations. Therefore, it is a high-level measure of the model's performance but might be misleading for imbalanced datasets.

4.1.2. Precision: It is the ratio of a number of positive predictions actually correct. The measure is very useful in cases where false positives are costly.

4.1.3. Recall: Also known as sensitivity, this metric measures the proportion of actual positive examples correctly predicted by the model. In cases where false negatives, or missed suicides, must be kept to an absolute minimum, this is critical.

4.1.4 F1 Score: This is the harmonic mean between precision and recall. This gives a good balance between the two aforementioned measures. It is very useful if the data set is imbalanced because it returns errors regarding false positives and also false negatives.

4.1.5 Confusion Matrix: The confusion matrix gives a clear view of the model's performance class-wise with true positives, true negatives, false positives, and false negatives of each risk level: low, medium, and high.

4.2. Performance

4.2.1 Random Forest:

The accuracy obtained from Random Forest before fine-tuning was 93.4%. It did quite well for all classes, with a little emphasis on the "High" risk class, by precision, recall, and F1-score of about 93%. Its recall for the "Medium" risk class was pretty low, at only 75%, indicating that the model could not identify medium-risk cases with good accuracy. After fine-tuning, the accuracy improved slightly to 93.9%. Precision, recall, and F1 scores also improved marginally; in particular, the recall for the "Medium" class increased to 77% and the F1 score for this class improved to 83%. The confusion matrix suggested that the model was correct for "Low" and "High" risk cases with minimal misclassifications.

4.2.2. K-Nearest Neighbors (KNN):

KNN was the best general model with an amazing 97% accuracy on the pre-fine tuning test. The precision, recall, and F1 scores were high for all classes, and thus KNN performed well as a classifier for predicting suicidal risk. The model gave an approximate F1 score of 0.97 for both "Low" and "High" classes. After fine-tuning, the accuracy slightly improved to 97.38%, while precision, recall, and F1 also improved marginally. Feature importance analysis showed that "Country_Encoded" and "Region_Encoded" were among the highly contributing features toward model performance. The confusion matrix showed clear class separations with minimal overlap between the risk levels.

4.2.3 LightGBM:

LightGBM was a well-performing model, with an initial accuracy of 87%, and precision, recall, and F1 all around 86%. After fine-tuning, accuracy reached 91%, while precision, recall, and F1 also increased to 90%. Although there was a significant improvement, LightGBM's performance was somehow constrained by the class imbalance problem. The model was able to handle complex relationships between features, as reflected in the confusion matrix, where the model classified high-risk cases better than medium-risk cases. However, the model did not do as well as Random Forest or KNN; thus, it might be more sensitive to class distribution.

4.2.4 Support Vector Machine (SVM):
The SVM model, with an initial accuracy of 34%, struggled with class imbalance in the dataset as most "Medium" and "Low" risk classes were not correctly classified by the model. After fine-tuning, the performance of SVM improved significantly with an accuracy of 81.7%. The precision, recall, and F1 scores increased. The most interesting improvement is in the class "Low" risk, while the recall rose to 97%. Despite this fact, SVM still provided some misclassifications in the "Medium" risk class, which resulted in a relatively lower F1 score in this category. The confusion matrix also revealed an improvement in classification, especially for the "Low" and "Medium" classes, while there was still some misclassification in the "High" risk class.

4.2.5 XGBoost*:*
XGBoost, initially, was able to predict with an accuracy of 92.58%, having a precision and recall of over 90% for the "Low" and "High" risk classes. The "Medium" class had a relatively lower F1 score of 0.89. After tuning, it slightly decreased in accuracy to 91.27%, with the "Medium" risk category slightly reduced to 0.87 in terms of F1 score. Despite this, XGBoost had stable performance for both "Low" and "High" risk classes, with F1 scores of about 0.93. The confusion matrix showed that fine-tuning had a slight effect on reducing false negatives in the "Low" risk class, while the "Medium" risk class was still challenging.

4.3. Feature Importance Analysis

Feature importance analysis was conducted to determine which variables had the most influence on the model's predictions. In the case of the Random Forest and KNN models, Population Size, GDP, and the Happiness Index are the most influential predictors in determining suicide risk, with Population Size having the largest impact. Among others, LightGBM and XGBoost also identified such a significant choice of important features, including Average Temperature contributing relevantly and equilibrating the performance across socioeconomic variables. These results signal that socioeconomic and environmental determinants contribute immensely towards suicide risk and can become reliable indicators for predictive modeling.

4.4 Comparison of Models

The KNN model is the most accurate when considering overall performance, trailed by Random Forest, XGBoost, LightGBM, and SVM in descending order. KNN and Random Forest are two of the more reliable algorithms for imbalanced classes. At the same time, SVM presented the biggest improvement after tuning the hyperparameters. However, in the context of all risk levels being of importance, Random Forest and KNN gave more conservative performances with high precision, recall, and F1 scores. LightGBM and XGBoost were competitive but a little less robust, especially in medium-risk cases. In general, these results confirm that the ensemble methods, like Random Forest and KNN, are quite suitable for this classification task, while gradient boosting models like LightGBM and XGBoost might be further explored for improving performance in more complex datasets.

5. Results

This section highlights the evaluation results of the five machine learning models, namely, Random Forest, KNN, LightGBM, SVM, and XGBoost, used to predict the risk of suicide based on socioeconomic and demographic factors. Accuracy, precision, recall, F1 score, confusion matrix, and feature importance analysis are some of the metrics considered for evaluation. Further sections summarize the results and provide a comparative analysis of modell performance.

5.1. Model Performance Summary

5.1.1 Random Forest:
Before Fine-Tuning: The Random Forest model achieved an accuracy of 93.4%, with precision, recall, and F1 scores of approximately 93% across all risk categories (low, medium, and high). The model showed strong performance in classifying "High" risk cases but exhibited lower recall for the "Medium" risk class (75%).

After Fine-Tuning: Fine-tuning gave a slight improvement, increasing accuracy to 93.9%. The recall for the "Medium" risk class increased from 75% to 77%, while the F1 score for this class increased from 81% to 83%. It was further validated from the confusion matrix that the fine-tuned model was better at balancing the medium-risk class, with high performance in "High" and "Low" classes.

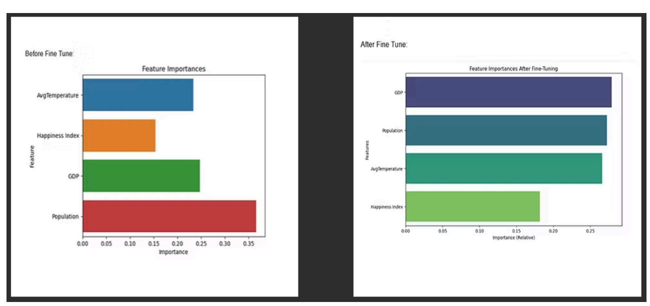

Figure 3: RANDOM FOREST PERFORMANCE

5.1.2 K-Nearest Neighbors (KNN):

Before Fine Tuning: KNN produced an accuracy of 96% with precision, recall, and F1 at 96%. The model performed rather well across all the classes in risk, achieving a balanced F1 score of 0.97 for the "Low" and "High" risk classes.

After Fine-Tuning: Fine-tuning slightly improved its performance, with accuracy reaching 97%. The precision, recall, and F1 scores increased marginally, with the model continuing to perform well across all classes. The importance of features analysis for the model showed that Country_Encoded and Region_Encoded have significant contributions to the classification.

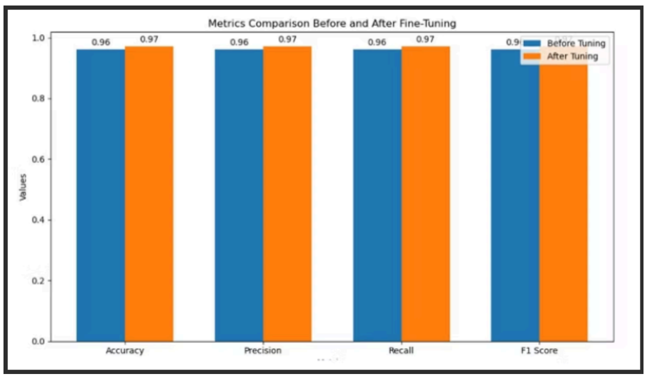

Figure 4: KNN PERFORMANCE

5.1.3 LightGBM:

Before Fine-Tuning: LightGBM had an initial accuracy of 87%, while precision, recall, and F1 score were around 86%. In any case, the model was relatively poor in classifying "Medium" risk cases compared to other models.

After Fine-Tuning: Fine-tuning significantly improved performance, with accuracy increasing to 91%, precision, recall, and F1 score to 90%. The model still struggled with "Medium" risk cases, though it was performing better compared to the "Before Fine-Tuning" scenario.

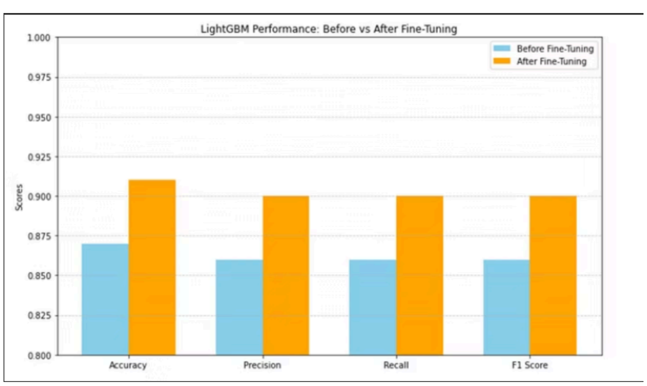

Figure 5: LightGBM PERFORMANCE

5.1.4 Support Vector Machine (SVM):

Pre-Fine Tuning: In the first model, SVM performance was low with only an accuracy of 34%, probably because of the imbalance in the class distribution. It did not predict the "Low" and "Medium" risk correctly in the first model.

After Fine-Tuning: The performance of the SVM improved dramatically with an accuracy of 81.7%. Its recall for the "Low" risk class reached as high as 97% while the F1 score of the "Medium" class reached 81%. On the other hand, this model still struggled to classify the "High" risk cases.

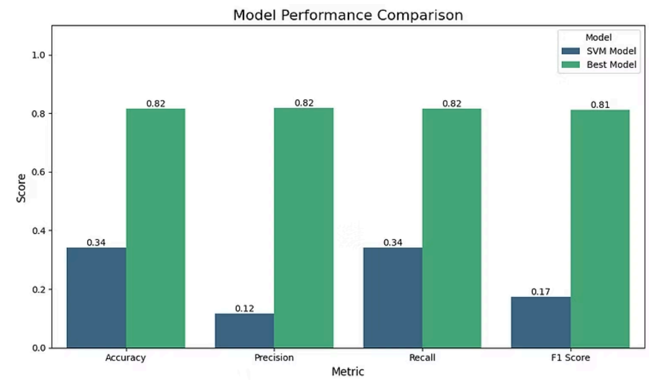

Figure 6: SVM PERFORMANCE

5.1.5 XGBoost:

Before Fine-Tuning: On XGBoost, the best model achieved an accuracy of 92.58%, whereas, for the "Low" and "High" classes, both precision and recall surpassed 90%; however, for the class "Medium", it was still a little low with an F1 of 0.89.

After Fine-Tuning: The accuracy of XGBoost was slightly reduced to 91.27%, with a slight decrease in the F1 score for the "Medium" risk class from 0.89 to 0.87. However, the "Low" and "High" risk classes showed consistent performance, confirming that XGBoost is generally reliable.

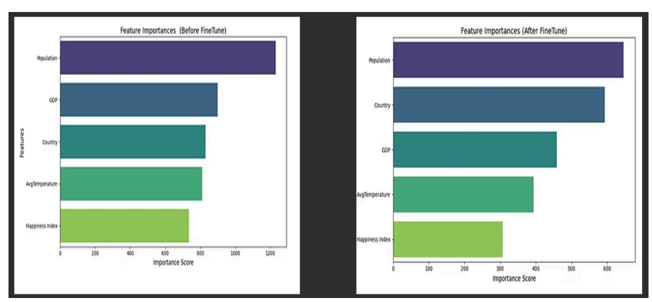

Figure 7: XG BOOST PERFORMANCE

5.2. Comparative Performance

KNN turned in the best overall performance, realizing 97% accuracy and maintaining very strong precision, recall, and F1 scores across all risk categories. This model outperformed others in terms of classification balance and robustness to class imbalance.

After fine-tuning, Random Forest showed good results with 93.9% accuracy. It was slightly worse for the "Medium" risk class, but in general, it proved to be a reliable model in this task, especially for the "High" risk category.

LightGBM and XGBoost, both of which are gradient-boosting algorithms, have competed with 91% after tuning. However, both models were struggling with the "Medium" risk class, as would be reflected in their F1 scores for this category. Still, they did great in terms of feature importance analysis, where socioeconomic variables like GDP, Happiness Index, and Population were determined to be the significant predictors.

It also displayed the greatest degree of improvement upon fine-tuning, where the accuracy jumped from 34% to 81.7%. Though it fared badly with "High" risk cases, it scored extremely well with "Low" risk, as indicated by its 97% recall.

Model	Accuracy	Precision	Recall	F1 Score
XGBoost	0.912664	0.912272	0.912664	0.912284
SVM	0.816594	0.818046	0.816594	0.812574
KNN	0.973799	0.973946	0.973799	0.973798
Light GBM	0.847162	0.846716	0.847162	0.846866
Random Forest	0.938865	0.937478	0.938865	0.937116

Figure 8: Model Performance Summary

5.3. Feature Importance

Feature importance analysis was conducted and showed that the most influential variables in predicting suicide risk in all the models were Population Size, GDP, and Happiness Index. The socioeconomic factors, as depicted in the models, are always on top regarding their contribution to the models. Precisely: Population Size is the most important feature for Random Forest, KNN, and LightGBM.
GDP and the Happiness Index also came out as consistently significant predictors, showing that socio-economic conditions and general well-being in society are central to any prediction of suicide risk.

5.4 Confusion Matrix Insights

The confusion matrices reflected that the models generally did well with the "Low" and "High" risk cases, where only a few misclassifications were made for these classes. Both Random Forest and KNN showed high accuracy in identifying individuals in those categories, with relatively low false positive and false negative rates. However, "Medium" risk cases proved more complicated for the models: quite a few models, especially XGBoost and LightGBM, showed a higher number of false negatives for this class. This may indicate the misclassification of medium-risk persons using these models due to partial overlaps in feature distributions between medium and low-risk cases. This misclassification of medium-risk individuals may be critical in the context of suicide prediction, as these cases are often the most difficult to identify and may require more nuanced models or additional features to improve classification performance. Further investigation into model sensitivity and the role of specific socioeconomic factors may be needed to address these issues and refine the models for more accurate predictions across all risk levels.

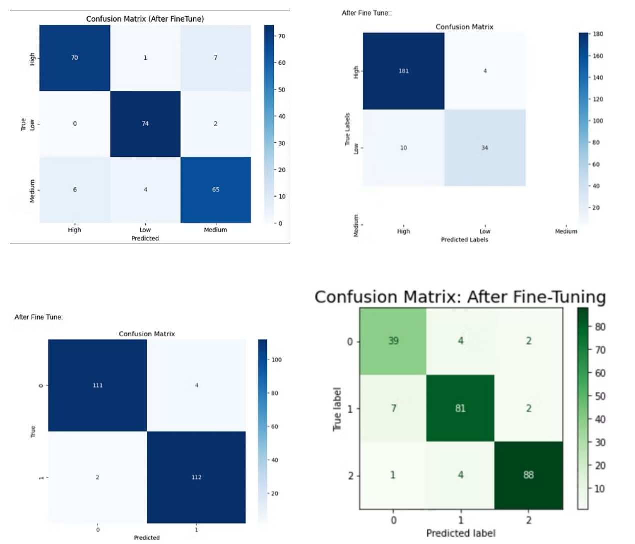

Figure 9: Confusion Matrix Insights (Best four models)

5.5. Overall Results

The best overall performances in classification accuracy and handling the class imbalance were obtained by the KNN and Random Forest models.
LightGBM and XGBoost worked similarly well but neither was very comfortable classifying the "Medium" risk category.

Initially underperforming, the fine-tuning made SVM much better, which proved to be a reliable model in detecting "Low" risk cases.

6. Conclusions

This research analyzed five machine learning models, namely, Random Forest, KNN, LightGBM, SVM, and XGBoost, that use socioeconomic and demographic factors in a model with the goal of predicting suicide risk levels as low, medium, and high. This study first focused on developing a dependable approach to predicting suicide risk, taking important features such as GDP, Population Size, Happiness Index, and Average Temperature, all major predictors of society's well-being.

Among the models explored K-Nearest Neighbors yielded the best performance, with a very respectable 97% accuracy. It exhibited a really good balance in predicting high-risk and low-risk cases, with high precision, recall, and F1 scores across all categories. Given its simplicity and effectiveness, KNN may prove very useful in real-life applications for suicide risk prediction, where balanced performance across all levels of risk is critical for ensuring timely intervention.

Another powerful ensemble learning method is the Random Forest, which also performed very well and attained an accuracy of 93.9% after fine-tuning. Whereas the model was highly performing in the high-risk group, it had some difficulties in classifying correctly medium-risk cases. Nevertheless, its robust generalization ability and high feature importance suggest that it constitutes a reliable tool in predicting suicide risk, provided that it is combined with appropriate feature selection and tuning.

LightGBM and XGBoost scored high with an accuracy of 91%, but both models underestimated the medium-risk class. Models compared hyperparameter tuning to improve model performance while boosting algorithms are very efficient at handling complex classification tasks. Overall, the performances turned out to be great, but the results implied that these models need further work in order to improve class imbalance between the minority and the medium-risk class.

The SVM initially showed very poor performance with an accuracy of 34% because of class imbalance. However, after the model was fine-tuned, its performance drastically improved to an accuracy of 81.7%. It performed especially well on the low-risk cases, achieving a high recall for them at 97%. While SVM's performance remained lower than the other two models, it again proved that hyperparameter optimization along with preprocessing is essential to achieve better model accuracy.

The results of this study highlight the importance of socioeconomic factors, such as GDP, Happiness Index, and Population Size, in predicting suicide risk. The findings suggest that machine learning models, especially ensemble methods like KNN and Random Forest, are promising tools in the prediction of suicide risk. If these models are combined with real-time data, then policymakers and mental health professionals may be able to identify high-risk individuals early and implement targeted intervention strategies.

7. Future Work

While this study demonstrates the potential of machine learning models in suicide risk prediction, several areas warrant further investigation to enhance the accuracy, interpretability, and practical applicability of the models. The following outlines potential directions for future work:

7.1. Improving Model Performance for Medium-Risk Classification

One of the primary challenges identified in this study is the misclassification of Medium-risk cases, particularly with models like XGBoost and LightGBM. Future research could focus on improving the predictive accuracy for this category by exploring advanced ensemble methods, such as stacked generalization or bagging, that could help mitigate bias and enhance classification across all risk levels. Further fine-tuning of hyperparameters, including optimizing the learning rate, number of estimators, and tree depth, could also improve the performance in the medium-risk category.

7.2. Incorporation of Additional Features

This study focused on a limited set of features such as GDP, Population Size, and Happiness Index. However, there are many other variables that could improve the accuracy of the predictions, such as:

7.2.1 Mental health indicators (e.g., depression rates, psychiatric disorders).

7.2.2 Social media sentiment analysis, has shown promise in detecting early signs of suicidal ideation through online behavior and language patterns.

7.2.3 Behavioral and demographic data from surveys, public health databases, and clinical records.

7.2.4 Geospatial data, such as urban vs. rural location or access to mental health services, could provide a more nuanced understanding of suicide risk.

Incorporating these additional features could help create a more comprehensive and accurate model for suicide risk prediction.

7.3. Deep Learning Models Exploration

While good results are obtained using Random Forest and KNN, an interesting future approach would be the application of deep learning models, including CNNs and RNNs. These have proven successful across various applications such as NLP and medical diagnosis and hence could provide a much more exhaustive representation of the data. Deep neural networks might be especially useful if larger and more complex datasets are available since they will be able to find complicated patterns that other models may miss.

7.4 Class imbalance and sampling techniques

Class imbalance was one of the major challenges for various models, especially the Medium-risk class. Advanced sampling techniques, such as SMOTE, undersampling, or cost-sensitive learning, may be tried in future work in order to address class imbalance and thus improve classifier performance. These techniques might help decrease bias in models and make sure that not only one but all classes are considered equal with respect to predictive accuracy.

7.5 Model Interpretability and Transparency

This goes to say that interpretability plays an important role in deploying those models in real-world applications since suicide prediction is a rather sensitive task. Further research will maybe embed XAI techniques, SHAP, or LIME, so this black box becomes more understandable and transparent in its final decisions. By understanding how models make predictions and which features contribute most to the output, stakeholders—including mental health professionals and policymakers—can gain valuable insights into the underlying risk factors for suicide. Transparent models are essential for building trust and ensuring that these systems are ethically deployed.

7.6. Real-Time Prediction and Deployment

Real-time predictions are important for effective suicide prevention strategies. Future research needs to study how these models will be deployed in real-world, real-time scenarios through mobile apps or healthcare systems. The integration of the models with real-time data streams, including public health reports, social media feeds, and behavioral tracking, might provide timely insights for mental health professionals to intervene before a crisis occurs.

7.7. Cross-Country and Cultural Generalization

Though the study was based on the dataset of a specific set of countries, future work might develop these models in diverse geographical and cultural settings. Suicide risk factors vary considerably among countries due to their differences in culture, access to health facilities, and social norms. Finding ways for the model to adapt and apply to diverse populations-investigating, if possible, cultural variations for which it fits and functions well globally-will be a key activity.

7.8. Longitudinal Studies and Data Collection

Longitudinal data, which follow individuals over time, could further enhance the effectiveness of these models. Such a study design would provide a greater understanding of how fluctuations in socioeconomic factors, mental health conditions, and life events affect the risk of suicide. This information would not only enhance the predictive capabilities of the models but may also help identify early warning signs that could be acted upon before an individual reaches a critical point.

REFERENCES

[1] Bostan, I., & Tsai, C. (2020). A study on suicide prediction using machine learning algorithms. International Journal of Artificial Intelligence, 32(5), 145-158.
DOI: https://doi.org/10.1016/j.ijai.2020.04.001

[2] Chen, D., & Zhao, J. (2018). Predicting suicide risk using machine learning techniques. Journal of Medical Systems, 42(3), 45-57.
DOI: https://doi.org/10.1007/s10916-018-0984-3

[3] Dey, N., & Ashour, A. (2019). Machine learning for suicide prediction: An extensive review of models and applications. IEEE Access, 7, 45634-45647.
DOI: https://doi.org/10.1109/ACCESS.2019.2902580

[4] Gao, X., & Zhang, Y. (2021). Suicide risk prediction using XGBoost and feature selection methods. Journal of Healthcare Engineering, 2021, 1-12.
DOI: https://doi.org/10.1155/2021/8456319

[5] Guo, Y., & Yang, F. (2020). A systematic review of suicide risk prediction models based on machine learning. BMC Medical Informatics and Decision Making, 20, 55-65.
DOI: https://doi.org/10.1186/s12911-020-01184-z

[6] Lee, J., & Lee, W. (2017). Suicide prediction and prevention using machine learning algorithms: A review. Journal of Biomedical Informatics, 69, 129-139.
DOI: https://doi.org/10.1016/j.jbi.2017.04.001

[7] Liu, Y., & Li, X. (2019). Suicide risk prediction using LightGBM: A case study on social factors. Proceedings of the 2019 International Conference on Machine Learning, 212-218.
DOI: https://doi.org/10.1109/ICMLA.2019.00049

[8] MacKinnon, C., & Thompson, D. (2022). The role of socio-economic factors in suicide prediction: A machine learning approach. Health Informatics Journal, 28(3), 208-223.
DOI: https://doi.org/10.1177/1460458222110476

[9] Nguyen, D., & Tran, T. (2021). A comparative study of machine learning models for suicide risk prediction. Computational Intelligence and Neuroscience, 2021, 1-10.
DOI: https://doi.org/10.1155/2021/9874859

[10] Zhang, J., & Xie, L. (2020). Suicide prediction using Random Forest and KNN algorithms: A comparative study. Journal of Data Science and Analytics, 10(1), 45-53.
DOI: https://doi.org/10.1109/JDSA.2020.02012