# Machine Learning techniques to predict suicidality during Covid-19 pandemic in Bangladesh

Farhan Hasin Saad , Jahin Tasnim , Nishat Naoal , Sakib Bin Swroar and Sifat Momen
Department of Electrical and Computer Enginnering, North South University

***Abstract*—The COVID-19 pandemic has caused an unprecedented natural phenomenon. As in other countries around the world, the lockdown measures taken during the COVID-19 outbreak in Bangladesh came suddenly and unexpectedly and could cause severe psychological consequences. The class imbalance that exists in the dataset is a serious problem and can skew predictions of suicidality toward the majority class, making machine learning models less reliable. Based on their behavioral patterns and the breadth of available pandemic material, our goal is to evaluate early suicide thoughts in people residing in lockdown or during a pandemic. In this paper, we used the Harvard Dataverse dataset for the COVID-19 pandemic and serious psychological consequences in Bangladesh (imbalance dataset) to train machine learning classifiers in order to predict suicidality in Bangladeshi people. After necessary pre-processing, the dataset was trained on with a range of different machine learning algorithms, including that of KNN (K-Nearest Neighbour), Random Forest, XG Boost, Logistic Regression, Naive Bayes, SVM. We used various data sampling methods such as - Under Sampling, Over Sampling, SMOTE, SMOTE+ENN to handle the imbalance class issue. And made a comparative analysis among all the results. Our results demonstrated how SMOTE+ENN applied dataset gave the best results and Random Forest Classifier had the highest Accuracy of 97%.**

***Index Terms*—Binary Classification, Suicide, COVID19, Imbalanced dataset, Under Sampling, Over Sampling, SMOTE, SMOTE+ENN, Random Forest , XG Boost, Logistic Regression, Naive Bayes.**

## I. Introduction

CORONAVIRUS disease (COVID-19) is a contagious agent caused by the SARS-CoV-2 virus. Following an epidemic in China in December 2019, the World Health Organization recognized SARS-CoV-2 as a novel coronavirus in the early months of 2020. The epidemic spread fast around the globe. Over a hundred nations have been impacted;many of them in the Asia-Pacific area [1].The outbreak significantly slowed down the world economy, changed and invalidated social interactions, and threatened people's mental health [2]. The pandemic has an impact on a person's social and personal life, physical and mental health, economic situation, and many more areas in addition to the healthcare systems [3]. According to the WHO, COVID-19 has severe and uncommon neurological complications like brain inflammation, strokes, delirium, and nerve damage in addition to the physical symptoms [4]. This demonstrates that people who have the COVID-19 are more likely to experience mental health issues [5].

Suicide is the death of a person who is psychologically weak. The WHO estimates that for every successful suicide, there are at least 20 unsuccessful attempts. From April 1, 2020, to March 31, 2021, we conducted an online media-reports-based retrospective study employing a variety of keywords on Google's search engine in both English and Bengali. In Bangladesh, 151 students committed suicide in the previous 15 months. Before the COVID-19 epidemic, a study on suicide incidents published in Bangladeshi print media from January 2018 to June 2019 was carried out. 56 Bangladeshi students killed themselves overall during the time. When the COVID-19 pandemic began and was compared to the total number of suicides among Bangladeshi students, it was found that the number of instances increased by nearly three times [6]. Not only did Bangladesh suffer from COVID-19-related moderate to severe psychological effects, but 53.8% of Chinese people also did. It was also shown that 16.5% of Chinese people had depressive symptoms; 28.8% had anxiety disorders, and 8.1% had moderate to severe levels of stress [7]. Moreover, among healthcare professionals as a whole, 23.2% expressed anxiety, 22.8% reported depression, and 38.9% reported insomnia [8].

Machine learning techniques have recently emerged as very useful in the field of suicide prediction [9]. These techniques require data to train a predictive model, but when the dataset is imbalanced, the model intuitively tends to favor the majority class. In our work, we used Harvard Dataverse dataset for the COVID-19 pandemic and serious psychological consequences in Bangladesh (imbalance dataset) to train machine learning classifiers in order to predict suicidality in Bangladeshi people [10]. This data set contained more samples not related to suicide (majority class) compared to samples related to suicide (minority class), resulting in an undesirably high false-negative prediction. Our goal is to reduce the number of false negative predictions.

Our objective is to assess early suicide ideation in individuals living in lockdown or during a pandemic based on their behavioral patterns and depth of pandemic information. This early diagnosis is crucial because it can save their lives if they receive the appropriate counseling before it's too late and they develop suicidal thoughts or, worse, act on them.

Rest of the paper is organized as follows: section II discusses the related work on suicidal prediction followed by section III that explains the research methodology embraced in this paper. Sections III-D and III-E respectively describes how machine learning classifiers and SMOTE are applied in our research. Results are discussed in section IV and finally the paper is concluded in section V.

## II. Related Work

Many of the works show that situations in COVID lead to various mental disorders like anxiety, depression, and insomnia and these disorders are linked to suicidal ideation.

A group of researchers in china worked on the psychological state of undergraduate students in the COVID pandemic predicting anxiety and insomnia. They assessed insomnia and anxiety based on a web survey and applied XGBoost model to predict anxiety and insomnia with 97.3% and 96.2% accuracy respectively. [9]

A work based on low and middle-income country(LMICs) like Indonesia reports the exploration of suicidal ideation from COVID-19 pandemic situation based on the factors like age,gender,demography,COVID related loneliness and isolation etc.To measure suicidal ideation, they took patient health questionnaire. Hierarchical logisitic regression was used as their predictive model. [11]

Another work has been conducted on Canadian adults to predict emotional distress based on various factors leading to COVID-19 risks and worries.XGBoost was used to find variance in emotional distress. Increased worries about finances, worries about getting COVID-19, and younger age are the items helped predict elevated emotional distress. [12]

We worked on a large dataset(N=10067) [13] where there was information about the recipient's demography, behavior towards COVID-19, knowledge about COVID-19, comorbidity, etc, and sorted out the information needed for our machine learning models to predict suicidality. As per our knowledge, such work on this dataset has not been conducted before.

## III. Methodology

Figure 1 depicts our process flow. After feature selection and data pre-processing, the dataset was split into a training set and a test set. After splitting the data, we scaled it with Standard Scaler. Initially, we used the imbalanced dataset to train our models and the unseen test set data to predict the outcomes. Afterwards, we trained our models using each of the four data manipulation techniques to resolve the imbalanced dataset. In the end, we utilized the 5-Fold Cross Validation optimizer to improve the performance of all four classifiers. A comparative comparison of the performance of the classifiers with and without the application of Over Sampling, Under Sampling, SMOTE, and SMOTE+ENN is shown.

### A. Dataset Acquisition and Description

The Harvard Dataverse has been used to gather the dataset for the COVID-19 pandemic and serious psychological consequences in Bangladesh [13]. The collection of 10,067 data from a representative sample of 64 districts in Bangladesh was done utilizing a non-random convenience sampling method online. Email and other communicable media, as well as social media platforms, were used to distribute the survey. The raw dataset includes socio-demographic information (such as gender, age group, educational attainment, occupation, data discipline, residence area, marital status, comorbidities, current health condition, smoking and alcohol use status, frequency of social media use, etc.), source materials from which

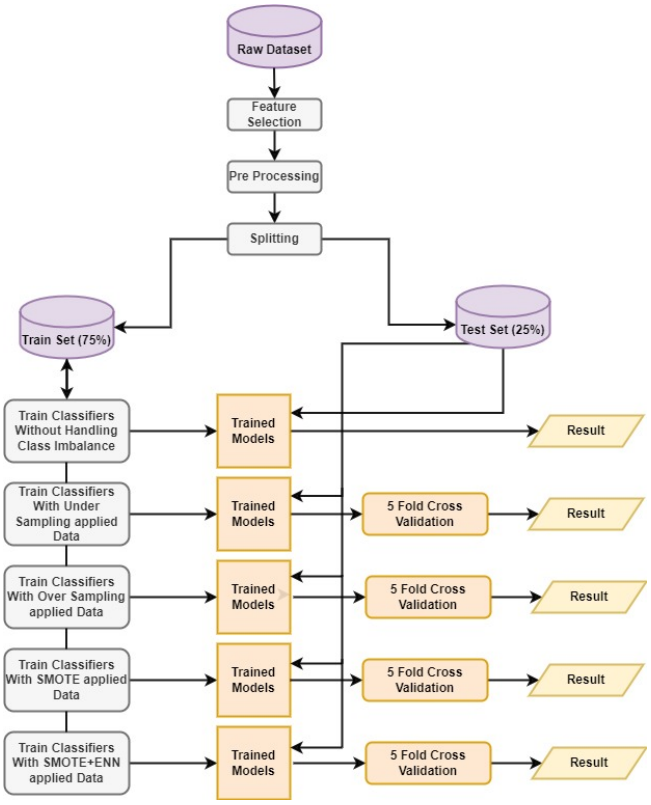


Fig. 1: Flow Chart

respondents obtain information about COVID-19 (such as social media, YouTube, newspapers, television, health-related websites, and other sources), Respondents' understanding of COVID-19, their activity in preventing it, lockdown-related inquiries, evaluations of respondents' dread of COVID-19, the severity of their insomnia, their level of depression, and their likelihood of considering suicide as a result of COVID-19 [10].

Consequently, the raw dataset had 135 attributes, but we only used 9. Figure 2 shows the histogram of our 9 selected features from the dataset. And Figure 3 is a pairplot among three features (Age, Behavior, Knowledge).

### B. Feature Selection

In machine learning classification, choosing significant features is essential for two reasons in particular: (1) Irrelevant features behave as noise, which might reduce the model's predictability. The predictability of the model is improved by eliminating unimportant features. (2) As a result, the dataset's dimension is reduced, allowing for the avoidance of the dimensionality curse.

The dataset [13] had over 10,067 entries and 135 features. The number of features was exorbitantly high. It contained a sizable amount of socio-demographic data that was unrelated to our prediction model. In terms of risk factors for suicidal thoughts, the most important characteristics were thus identified. The main causes of suicidal thoughts have been identified by a health and life science study project that surveyed people regarding their psychological impact as a result of Covid19. There, participants who conveyed suicidal ideation were more

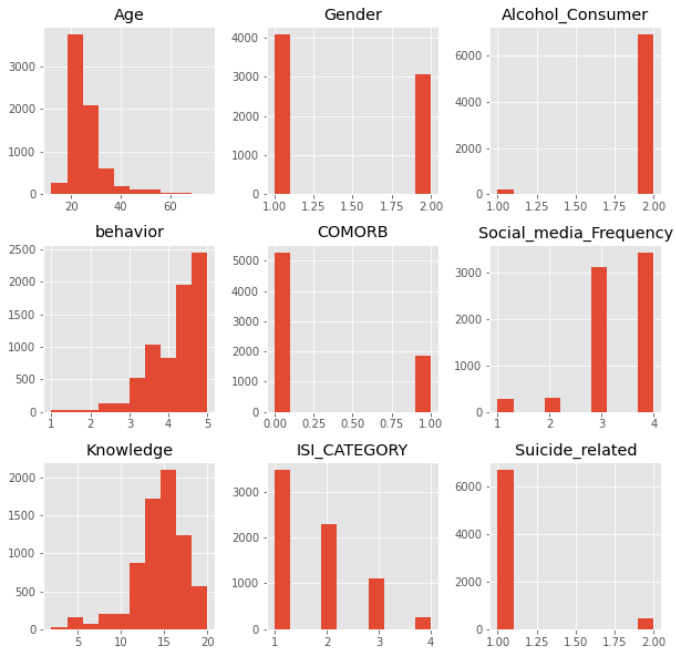


Fig. 2: Histogram

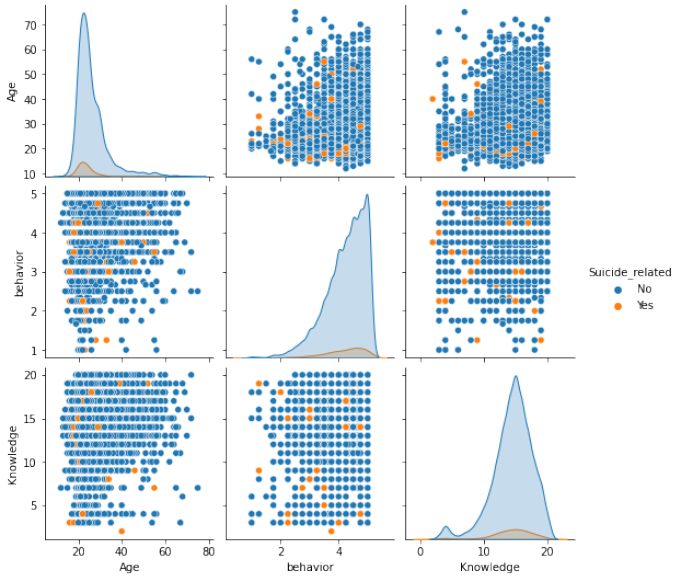


Fig. 3: Pairplot

likely to be (i) young, (ii) female, (iii) drinkers, (iv) had more comorbid illnesses, (v) use social media more frequently, (vi) know less about COVID-19, (vii) participate in less preventative COVID-19 practices, and (viii) have sleep issues [10]. 8 features related to these key factors were chosen along with the target column suicidal ideation. Because of the above explanation, rest of the additional columns were discarded. After Careful selection we had 9 columns in our dataset. Table I below consists of the features that were selected.

TABLE I: Selected Attributes

| Categories | Attributes |
|---|---|
| Socio-Demographic | Age, Gender |
| Alcohol Status | Alcohol_Consumer |
| Rate of Illness Comorbidity | COMORB |
| Frequency of Social Media Usage | Social_media_Frequency |
| Knowledge about COVID-19 | knowledge |
| behavior in preventing COVID-19 | behavior |
| Insomnia Problem | ISI_CATEGORY |
| Target | Suicide_related |

### C. Data pre-processing

To prepare the dataset for the machine learning algorithms to be applied to it, several pre-processing techniques were used. The redundant rows were removed. A few empty strings and null values that were found were also dropped. A few columns in the dataset (Age, Gender, and Social Media Frequency) had numerical values that were originally stored as string data types. They were changed into floats and integers, accordingly. The dataset's values for the features having binary classes (Gender, Alcohol Consumer, and Suicide) were annotated with "1" and "2". They were replaced with the well-known binary symbols "0" and "1," respectively. One hot encoding was used to encode the categorical columns. Dataset was eventually scaled after the split.

The dataset had 7142 rows and 9 columns after final pre-processing.

### D. Machine learning classifiers

*1) Decision Tree:* The non-parametric supervised learning approach used for classification and regression applications is the decision tree. It is organized hierarchically and has a root node, branches, internal nodes, and leaf nodes. By using a greedy search to find the ideal split points inside a tree, decision tree learning uses a divide and conquer technique. When most or all of the records have been classified under distinct class labels, this splitting procedure is then repeated in a top-down, recursive fashion. Although there are other approaches to choose the optimal attribute at each node, the Gini impurity and information gain methods are the two that are most frequently used as a splitting criterion in decision tree model.

$$Entropy(S) = -\sum_{c \epsilon C}(p(C)log_2 p(C)) \quad (1)$$

$$\text{Information Gain(S,a)=Entropy(S)-}\textstyle\sum_{v \epsilon Values(a)}(\frac{|S_v|}{|S|})(Entropy(S_v)) \quad (2)$$

*2) Logistic Regression:* The probability of a target variable is predicted using the supervised learning classification algorithm known as logistic regression. Since the dependent variable's nature is binary, there are only two viable classes. The target class in this dataset is binary. Therefore, Binary or Binomial Logistic Regression Model, the most basic type of logistic regression, is utilized for this dataset.

$$h_\theta(x) = g(\theta^T x) where 0 \leq h_\theta \leq 1 \quad (3)$$

Here, g is the logistic or sigmoid function which can be given as follows

$$g(z) = \frac{1}{1+e^{-z}} where z = \theta^T x \tag{4}$$

*3) Random Forest:* Random Forest is an ensemble of Decision Trees, generally trained via the bagging method [14]. Rather than depending on one tree it takes the prediction from each tree and based on the majority votes of predictions, it predicts the final output. The Random Forest algorithm introduces extra randomness when growing trees; instead of searching for the very best feature when splitting a node, it searches for the best feature among a random subset of features. This results in a greater tree diversity, which again trades a higher bias for a lower variance, generally yielding an overall better model [14].We used random forest classifier including 100 trees.

*4) Support Vector Machine:* Support Vector Machines, or SVMs for short, are a set of related algorithms for supervised learning that can be used to issues including classification as well as regression. A SVM classifier splits the example classes while maximizing the distance to the closest neatly separated instances by constructing a maximum-margin hyperplane that lies in a modified input space. The solution hyperplane's parameters are generated from a quadratic programming optimization issue. [15]

$$\frac{w^T(x-x)}{x} = \frac{2}{w} \tag{5}$$

*5) K-Nearest Neighbor:* A machine learning approach called K-nearest neighbor (KNN) is mostly used for classification problems [14]. The KNN determines the distance between each query data point and every other training data point. The K nearest neighbors of the specific query data-point are then chosen from there. Once the neighbors have been located, the method simply polls the neighbors to determine which class should be considered the anticipated class. It has been found that K = 4 offers the best results for this dataset. The distances between the query data point and the training data points are determined using the Euclidean distance (equation 5) metric.

$$d(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2 \tag{6}$$

*6) XG Boost:* XGboost is a gradient-boosted trees algorithm. It is an ensemble method and a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler models. Weights are assigned to all the independent variables which are then provided for the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then provided for the second decision tree. Then these predictors ensemble into a strong model. As our problem is a binary classification problem, we used XGboost as a classifier.

*7) Gaussian Naive Bayes:* Gaussian Naive Bayes is a probabilistic classification algorithm based on applying Bayes theorem with strong independence assumptions. It is a variant of naive Bayes that follows the Gaussian normal distribution. The naive Bayes classifier simplifies learning by assuming that features are independent given class [16].The dataset contained continuous values which associated with each class are distributed according to a normal distribution. The possibility of features is assumed to be (7). A way to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance between dimensions. The model can be fit by finding the mean and standard deviation of the points within each label. In a GNB classifier, at every data point, the distance between that point and the class mean is calculated.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}) \tag{7}$$

### *E. Imbalanced Dataset*

Imbalanced dataset describes a classification data set with uneven class proportions. Classes that make up the bulk of the data set are referred to as majority classes. Classes with a lesser share of the population are minority classes. In our sample, the target variable had two classes. Whether or not a participant had suicidal thoughts. "Yes" indicates that they had suicidal thoughts, whereas "No" indicates that they did not. It becomes a problem of binary classification. The difficulty, however, was that the class distribution was not uniform. That is, there was a significant disparity between the "yes" and "no" classes. Where "yes" represented the minority and "no" represented the majority. After final pre-processing, the "No" class contained 6675 items whereas the "Yes" class contained only 467. Figure 4 shows the class imbalance present in our dataset.

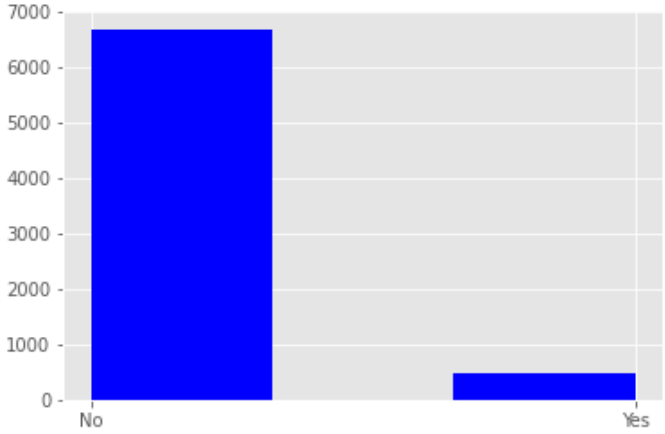


Fig. 4: Imbalance Class Distribution of Target "Suicidal Ideation"

*1) **Under Sampling**:* Random under-sampling [17] is a non-heuristic method that tries to balance the number of examples from each class by randomly getting rid of examples from the majority class. The idea behind it is to try to make the

dataset more balanced so that the algorithm's quirks don't get in the way. The biggest problem with random under-sampling is that it might throw away useful information that could be important for the induction process. [18]

Near-miss is an algorithm that can help make a dataset more even if it isn't already. It can be put in the category of undersampling algorithms and is a good way to make sure the data are equal. The algorithm does this by looking at how the classes are spread out and removing samples from the larger classes at random. When two points from different classes are close together in the distribution, this algorithm gets rid of the point from the larger class to try to make the distribution more even.

Work by Akira Tanamoto [19] shows even in cases with very limited true positive case, a higher accuracy can be achieved by using near miss undersampling technique.

*2) **Over Sampling**:* Random oversampling is a non-heuristic method that tries to even out the distribution of classes by randomly copying examples of classes that are largely absent. Several authors [17], [18] agree that random over-sampling can make over-fitting more likely because it makes exact copies of examples from the minority class. In this way, a symbolic classifier, for example, could make rules that seem correct but only cover one example that has been repeated. Oversampling can also add more computational task if the data set is already pretty big but not evenly distributed. [20]

*3) **Synthetic minority over-sampling technique (SMOTE)**:* SMOTE is an oversampling method in which the minority class is oversampled by making "synthetic" examples instead of oversampling with replacement. It generates new examples in a way that is less dependent on the application because it works in "feature space" instead of "data space." The minority class is oversampled by taking each sample from the minority class and adding synthetic examples along the line segments that connect any or all of the k nearest neighbors from the minority class. Depending on how much of an oversample is needed, random neighbors from the closest neighbors are picked. In our case it was k = 5. [20]

*4) **Hybridization: SMOTE + ENN**:* Hybrid data sampling is a sampling technique that combines two different sampling techniques to create a balanced dataset that can be used to build different classification models. [21] Two well-known re-sampling algorithms are the Synthetic minority over-sampling technique (SMOTE) [20] and the Edited nearest neighbor (ENN) [22]. Linear interpolation is used by SMOTE to make more samples in the minority class, while noise samples are removed from the majority class by ENN to reduce the number of samples in the majority class. But SMOTE has the problem of making noise samples and boundary samples [16], and the fact that ENN deletes noise samples is what makes this data sampling method better than regular SMOTE. Zhaozhao Xu's [23] experiments show that SMOTE can work best with ENN and Tome links. The illustration below shows how SMOTE+ENN maintains the classification distinction between classes and cleans the synthetic data so that the results are more accurate (Figure 5).

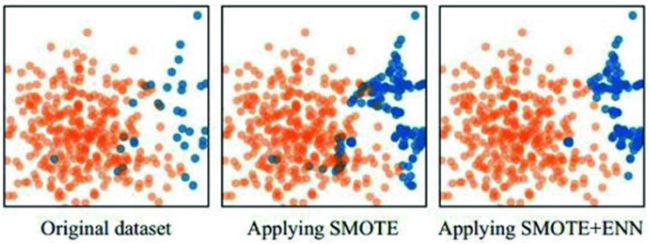


Fig. 5: SMOTE & SMOTE+ENN [24]

### *F. Receiver Operator Characteristic*

For binary classification problems, the *Receiver Operator Characteristic* or (ROC) curve is a way to measure performance. It is a probability curve that plots the TPR (*True Positive Rate*) against the FPR (*False Positive Rate*) at different threshold values. It basically tells the difference between the "signal" and the "noise." The Area Under the Curve (AUC) is a summary of the ROC curve that shows how well a classifier can tell the difference between groups. The model does a better job of telling the difference between the positive and negative classes the higher the AUC.

A higher value on the X-axis of a ROC curve means that there are more False positives than True negatives. On the other hand, a higher value on the Y-axis shows that there are more True positives than False negatives. So, the choice of the threshold depends on how well you can balance false positives and false negatives.

The perfect classifier would be in the top-left corner of the ROC graph, which is the same as (0, 1) in the Cartesian plane.Here, both the Sensitivity and the Specificity would be the highest, and the classifier would correctly classify all the Positive and Negative class points. We initially plotted an ROC curve to understand how our classifiers would perform on the original dataset. We observed *Support Vector Machine* was heavily under-performing, Therefore we omitted it from our classifier pipeline.

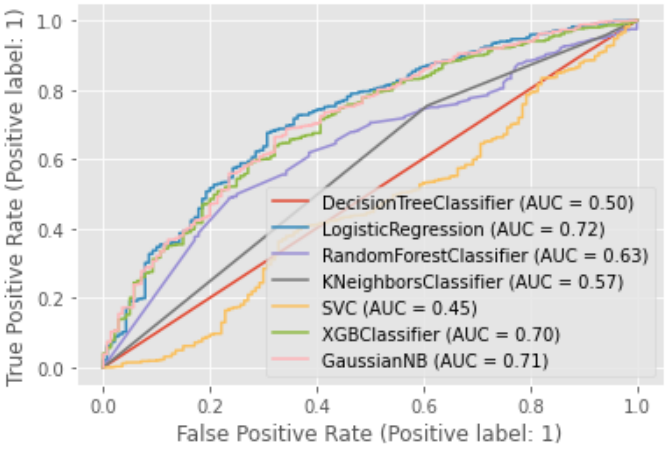


Fig. 6: ROC Curve on unaltered Dataset.

TABLE II: Results from Unaltered/Original Dataset

| Classifier | Precision | Recall | F1-score | Accuracy | AUC |
|---|---|---|---|---|---|
| Decision Tree | 0.07 | 0.09 | 0.08 | 0.86 | 0.50 |
| Logistic Regression | 0.00 | 0.00 | 0.00 | **0.93** | **0.72** |
| Random Forest | 0.02 | 0.01 | 0.01 | 0.92 | 0.63 |
| KNN | 0.09 | 0.06 | 0.07 | 0.90 | 0.57 |
| SVM | 0.00 | 0.00 | 0.00 | **0.93** | 0.45 |
| XG Boost | 0.00 | 0.00 | 0.00 | **0.93** | 0.70 |
| Naive Bayes | **0.19** | **0.30** | **0.23** | 0.87 | 0.71 |

TABLE III: Results from Under Sampled Dataset with 5 fold cross validation

| Classifier | Precision | Recall | F1-score | Accuracy | AUC |
|---|---|---|---|---|---|
| Decision Tree | 0.67 | 0.62 | 0.64 | 0.66 | 0.49 |
| Logistic Regression | **0.75** | 0.89 | **0.80** | **0.78** | **0.52** |
| Random Forest | 0.71 | 0.73 | 0.72 | 0.72 | 0.50 |
| KNN | 0.68 | 0.75 | 0.69 | 0.68 | 0.50 |
| XG Boost | 0.74 | 0.84 | 0.78 | 0.77 | 0.48 |
| Naive Bayes | 0.59 | **0.97** | 0.73 | 0.64 | 0.5 |

TABLE IV: Results from Oversampled Dataset with 5 fold cross validation

| Classifier | Precision | Recall | F1-score | Accuracy | AUC |
|---|---|---|---|---|---|
| Decision Tree | **1.0** | 0.90 | 0.95 | 0.95 | 0.49 |
| Logistic Regression | 0.66 | 0.67 | 0.66 | 0.66 | **0.72** |
| Random Forest | **1.0** | **0.94** | **0.97** | **0.97** | 0.62 |
| KNN | **1.0** | 0.85 | 0.92 | 0.92 | 0.55 |
| XG Boost | 0.99 | 0.86 | 0.92 | 0.92 | 0.62 |
| Naive Bayes | 0.66 | 0.66 | 0.66 | 0.66 | **0.72** |

TABLE V: Results from applying SMOTE with 5 fold cross validation

| Classifier | Precision | Recall | F1-score | Accuracy | AUC |
|---|---|---|---|---|---|
| Decision Tree | 0.92 | 0.90 | 0.90 | 0.90 | 0.56 |
| Logistic Regression | 0.66 | 0.66 | 0.66 | 0.66 | **0.72** |
| Random Forest | 0.94 | 0.93 | 0.93 | 0.93 | 0.58 |
| KNN | **0.96** | 0.66 | 0.78 | 0.82 | 0.56 |
| XG Boost | 0.95 | **0.98** | **0.96** | **0.96** | 0.59 |
| Naive Bayes | 0.69 | 0.62 | 0.65 | 0.67 | 0.69 |

TABLE VI: Results from Hybridization(SMOTE+ENN) with 5 fold cross validation

| Classifier | Precision | Recall | F1-score | Accuracy | AUC |
|---|---|---|---|---|---|
| Decision Tree | 0.93 | 0.93 | 0.93 | 0.94 | 0.60 |
| Logistic Regression | 0.73 | 0.71 | 0.72 | 0.75 | **0.72** |
| Random Forest | 0.98 | **0.97** | **0.97** | **0.97** | 0.62 |
| KNN | **0.99** | 0.88 | 0.94 | 0.95 | 0.63 |
| XG Boost | 0.94 | **0.97** | 0.95 | 0.95 | 0.63 |
| Naive Bayes | 0.73 | 0.72 | 0.72 | 0.75 | 0.69 |

## IV. Results

Tables II, III, IV, and V list the classifiers' evaluation findings in brief. Accuracy, precision, recall, F1-score, and area under the ROC curve were used to evaluate performances.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$F1 - Score = \frac{2 * Recall * Precision}{Recall + Precision} \tag{11}$$

Here,
TP = Actual class is positive and predicted positive.
TN = Actual class is negative and predicted negative.
FP = Actual class is negative but predicted positive.
FN = Actual class is positive but predicted negative.

Initially for the original dataset we cannot consider accuracy as the evaluation metric due to the fact that it was a heavily imbalanced dataset. Therefore we have to look at the precision and recall values. In Table II it's visible that the accuracy is decent but the precision and recall scores were horrible. We used various undersampling techniques on our train set to manipulate and balance the data. Afterwards we used the same test set for every case and accuracy as our evaluation metrics.

Table III shows the evaluated results from Undersampled dataset. Here the class distribution was uniform but the dataset was trimmed down to fit the number of the minority

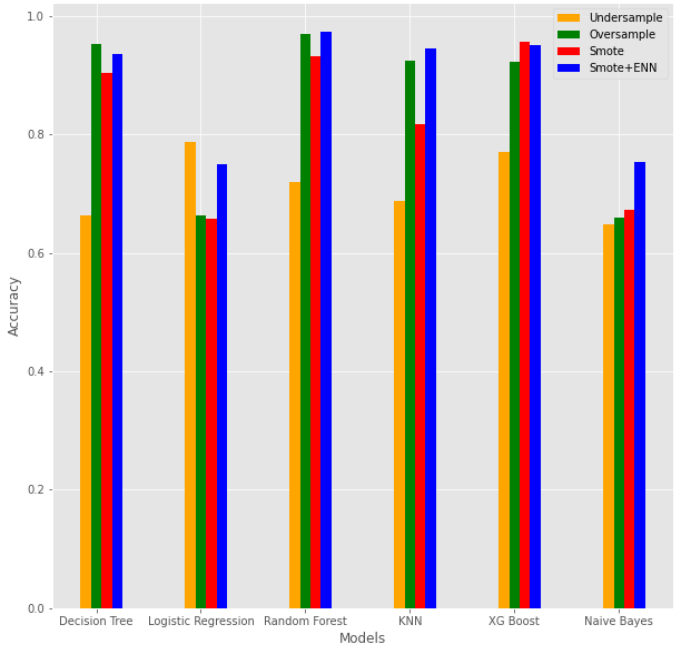


Fig. 7: Barplot of classifiers accuracy performance on four differently altered datasets.

class. The number of datapoints for each classes were 327. The results from undersampling shows significant rise in recall and precision values while taking a hit on accuracy. But this is not an ideal model because we are discarding a lot of majority class data without proper justification.

Table IV shows the evaluated results from Oversampled dataset. Here the number of datapoints for each classes were 4672 that were generated using random over sampling technique. The results are very good here but we need to remind ourselves about the drawback random over sampling has which is it may very likely overfit the data due to the fact that it creates copies of minority class data over and over again until the distribution is uniform. We can also notice that the preicision is almost 100% for some models. Therefore this wouldn't be an ideal solution as well due to it's tendency of being biased.

Table V shows the evaluated results from SMOTE dataset. SMOTE tries to avoid overfitting and biasness in dataset and instead of randomly generating copies it generates synthetic data points close to it's k nearest points of the minority class clusters. It gives us much better and acceptable results. Whilst improving accuracy for every classifier. In the SMOTE there was 4672 values for both class.

SMOTE has three drawbacks, though: (1) it oversamples noisy samples [25], (2) it oversamples uninformative samples [26], and (3) it is challenging to estimate the number of nearest neighbors and there is significant blindness in the selection of nearest neighbors for synthetic samples. [26]

That is why we used SMOTE+ENN or K-Means SMOTE (Synthetic Minority Oversampling TEchnique) as an oversampling technique and edited nearest neighbor (ENN) undersampling technique used as noise removal. By initially clustering the datasets using the K-Means clustering technique, SMOTE is then utilized inside the clusters to handle the imbalance by producing synthetic instances of the class in the minority, and then, the ENN technique is used to remove instances that produce noise afterward [27]. This way the generated dataset stays fair and realistic while being synthetic and solves the dataset's noise problem excellently. Our SMOTE+ENN applied dataset had 3631 majority class values and 2942 minority class values. The results from Table VI shows it's an improvement on precision,recall,f-1,accuracy and AUC value for all models compared to SMOTE's results in TABLE V. Random Forest Classifier holds the highest accuracy value 97% . Other models also have performed very well on this dataset.

## V. Conclusion

People have been significantly psychologically impacted by the COVID-19 pandemic. Bangladesh a resource-limited country, has an incomplete capacity to cope with the stressful situations transformed by the COVID-19 pandemic and lacks mental health initiatives. Our study's primary objective is to decrease the number of incorrect negative predictions. This also reduces the misclassification of suicidal patients as non-suicidal. We have carried out a number of experiments to fulfill the research goal, and the findings indicate that it has been accomplished. The use of machine learning classifiers and SMOTE+ENN has minimized the biased behavior toward false negative predictions. The suicidality was predicted by the algorithms to varying degrees of accuracy. On test data, it was discovered that Random Forest Classifiers performed the best, with 97% accuracy.

## References

[1] A. Qadir, "Modern trends in humanities and social sciences: With special focus on covid-19 scenarios and implementation of laws," *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 18, no. 10, pp. 553–559, 2021.

[2] R. Buselli, M. Corsi, A. Veltri, S. Baldanzi, M. Chiumiento, R. Marino, F. Caldi, S. Perretta, R. Foddis, A. Cristaudo *et al.*, "Suicidal ideation and suicide commitment in health care workers during covid-19 pandemic: a review of the literature," *International Journal of Occupational Safety and Health*, vol. 12, no. 2, pp. 117–124, 2022.

[3] S. Daria and M. R. Islam, "Increased suicidal behaviors among students during covid-19 lockdowns: a concern of student's mental health in bangladesh," *Journal of affective disorders reports*, vol. 8, p. 100320, 2022.

[4] W. H. O. (WHO) *et al.*, "World health organization coronavirus disease (covid-19). 2020," 2020.

[5] Y. Krishnamoorthy, R. Nagarajan, G. K. Saya, and V. Menon, "Prevalence of psychological morbidities among general population, healthcare workers and covid-19 patients amidst the covid-19 pandemic: A systematic review and meta-analysis," *Psychiatry research*, vol. 293, p. 113382, 2020.

[6] M. A. Mamun, A. B. Siddique, M. Sikder, M. D. Griffiths *et al.*, "Student suicide risk and gender: a retrospective study from bangladeshi press reports," *International Journal of Mental Health and Addiction*, pp. 1–8, 2020.

[7] C. Mazza, E. Ricci, S. Biondi, M. Colasanti, S. Ferracuti, C. Napoli, and P. Roma, "A nationwide survey of psychological distress among italian people during the covid-19 pandemic: immediate psychological responses and associated factors," *International journal of environmental research and public health*, vol. 17, no. 9, p. 3165, 2020.

[8] S. Pappa, V. Ntella, T. Giannakas, V. G. Giannakoulis, E. Papoutsi, and P. Katsaounou, "Prevalence of depression, anxiety, and insomnia among healthcare workers during the covid-19 pandemic: A systematic review and meta-analysis," *Brain, behavior, and immunity*, vol. 88, pp. 901–907, 2020.

[9] F. Ge, D. Zhang, L. Wu, and H. Mu, "Predicting psychological state among chinese undergraduate students in the covid-19 epidemic: a longitudinal study using a machine learning," *Neuropsychiatric disease and treatment*, vol. 16, p. 2111, 2020.

[10] M. A. Mamun, N. Sakib, D. Gozal, A. I. Bhuiyan, S. Hossain, M. Bodrud-Doza, F. Al Mamun, I. Hosen, M. B. Safiq, A. H. Abdullah *et al.*, "The covid-19 pandemic and serious psychological consequences in bangladesh: a population-based nationwide study," *Journal of affective disorders*, vol. 279, pp. 462–472, 2021.

[11] A. Liem, B. Prawira, S. Magdalena, M. J. Siandita, and J. Hudiyana, "Predicting self-harm and suicide ideation during the covid-19 pandemic in indonesia: a nationwide survey report," *BMC psychiatry*, vol. 22, no. 1, pp. 1–10, 2022.

[12] K. Hueniken, N. H. Somé, M. Abdelhack, G. Taylor, T. E. Marshall, C. M. Wickens, H. A. Hamilton, S. Wells, D. Felsky *et al.*, "Machine learning–based predictive modeling of anxiety and depressive symptoms during 8 months of the covid-19 global pandemic: Repeated cross-sectional survey study," *JMIR mental health*, vol. 8, no. 11, p. e32876, 2021.

[13] A. Pakpour, "Dataset on the COVID-19 pandemic and serious psychological consequences in Bangladesh: a population-based nationwide study," 2020. [Online]. Available: https://doi.org/10.7910/DVN/YKH9C1

[14] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed. O'Reilly Media, Inc., 2019.

[15] A. Shmilovici, *Support Vector Machines*. Boston, MA: Springer US, 2005, pp. 257–276. [Online]. Available: https://doi.org/10.1007/0-387-25465-$X_12$

[16] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[17] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *In Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.43.4487

[18] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, pp. 25–36, 11 2005.

[19] A. Tanimoto, S. Yamada, T. Takenouchi, M. Sugiyama, and H. Kashima, "Improving imbalanced classification using near-miss instances," *Expert Systems with Applications*, vol. 201, p. 117130, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422005280

[20] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *J. Artif. Intell. Res. (JAIR)*, vol. 16, pp. 321–357, 06 2002.

[21] C. Seiffert, T. M. Khoshgoftaar, and J. Van Hulse, "Hybrid sampling for imbalanced data," in *2008 IEEE International Conference on Information Reuse and Integration*, July 2008, pp. 202–207.

[22] M. Beckmann, N. F. Ebecken, B. S. P. de Lima *et al.*, "A knn undersampling approach for data balancing," *Journal of Intelligent Learning Systems and Applications*, vol. 7, no. 04, p. 104, 2015.

[23] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining m-smote and enn based on random forest for medical imbalanced data," *Journal of Biomedical Informatics*, vol. 107, p. 103465, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420300940

[24] M. Lamari, N. Azizi, N. E. Hammami, A. Boukhamla, S. Cheriguene, N. Dendani, and N. E. Benzebouchi, "Smote–enn-based data sampling and improved dynamic ensemble selection for imbalanced medical data classification," in *Advances on Smart and Soft Computing*, F. Saeed, T. Al-Hadhrami, F. Mohammed, and E. Mohammed, Eds. Singapore: Springer Singapore, 2021, pp. 37–49.

[25] P. Soltanzadeh and M. Hashemzadeh, "Rcsmote: Range-controlled synthetic minority over-sampling technique for handling the class imbalance problem," *Information Sciences*, vol. 542, pp. 92–111, 2021.

[26] Z. Jiang, T. Pan, C. Zhang, and J. Yang, "A new oversampling method based on the classification contribution degree," *Symmetry*, vol. 13, no. 2, p. 194, 2021.

[27] A. Puri and M. Kumar Gupta, "Improved Hybrid Bag-Boost Ensemble With K-Means-SMOTE–ENN Technique for Handling Noisy Class Imbalanced Data," *The Computer Journal*, vol. 65, no. 1, pp. 124–138, 05 2021. [Online]. Available: https://doi.org/10.1093/comjnl/bxab039