Spring 5-9-2023

Predicting Suicide Risk Among Youths Using Machine Learning Methods

Saswati Bhattacharjee

Predicting Suicide Risk Among Youths Using Machine Learning Methods

By

Saswati Bhattacharjee

A Thesis
Submitted to the Graduate School,
the College of Arts and Sciences.
and the School of Computing Sciences and Computer Engineering
at The University of Southern Mississippi
in Fulfillment of the Requirements
for the Degree of Master of Science

Approved by:

Dr. Chaoyang Zhang, Committee Chair
Dr. Sarah Lee
Dr. Ahmed Sherif

Published by the Graduate School

THE UNIVERSITY OF SOUTHERN MISSISSIPPI

ABSTRACT

Suicide is the second leading cause of death among youths in the USA. Although machine learning approaches have provided great potential for predicting suicide risk using survey data, prediction accuracy may not meet the need for clinical diagnosis due to the intrinsic characteristics of datasets. In this study, I perform a comparative study of six classification algorithms including naïve Bayes (NB), logistic regression (LR), multilayer perceptron (MLP), AdaBoost (Ada), random forest (RF), and bagging using YRBSS dataset and investigate the effectiveness of several data handling techniques to improve the overall performance of suicide risk prediction.

The dataset consists of 76 health risk-related questions with 13,437 responses collected from 136 high school students in the USA. Various preprocessing techniques such as missing value imputation, feature selection, and sampling techniques for handling the imbalanced ratio of the class label were applied to the dataset. The data was partitioned into a training dataset (70%) and a test dataset (30%) using a stratified partitioning method. The performance of the classifiers was evaluated using five evaluation metrics including accuracy, precision, recall, F2 score, and area under the receiver operating characteristic curve (AUROC). The result showed that RF classifier with undersampling method achieved the highest recall of 0.84, F2 measure of 0.72, and AUROC of 0.85 followed by LR and Ada classifiers.

Therefore, I can conclude that RF, LR, AdaBoost are powerful tools for predicting suicidal tendencies in youth. Feature selection and undersampling methods are crucial preprocessing steps necessary to identify adolescents who are at high suicide risk.

Acknowledgment

I would like to express my sincere gratitude to my thesis advisor, Dr. Chaoyang (Joe) Zhang, for his unwavering support and guidance throughout this research. His expertise, feedback, and encouragement were invaluable to me and greatly contributed to the completion of this project.

I am also grateful to the members of my thesis committee, Dr. Sarah Lee and Dr. Ahmed Sherif, for their constructive feedback and valuable suggestions that have helped me refine my research and writing skills. I also extend my appreciation to the faculty members of the Computer Science department, who have provided me with a strong academic foundation and have challenged me to think critically and analytically.

My deepest gratitude goes to my parents and husband for their love, encouragement, and unwavering support throughout my academic journey. Without their belief in me, I would not have been able to achieve this milestone.

Finally, I would like to thank my friends and classmates who have offered their support, assistance, and words of encouragement when I needed them the most.

Table of Contents

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATION

NB	Naïve Bayes
LR	Logistic Regression
MLP	Multilayer Perceptron
RF	Random Forest
SVM	Support Vector Machine
SMOTE	Synthetic Minority Over-sampling Technique
Pre	Precision
Rec	Recall
TP	True positive
FP	False positive
FN	False negative
TN	True negative
YRBS	Youth risk behavior survey
CDC	Center for disease control and prevention
MI	Mutual information
RValue	Rank Value
SL_NO	Serial _Number

CHAPTER I – INTRODUCTION

1.1 Significance of suicide risk prediction

In literature, suicide can be defined as self-inflicted harm that one can do to kill himself or herself. This is a major health problem all over the world. The incident of attempting suicide can be found anywhere irrespective of age, gender, and geographical and economical background. According to statistics from the Centers for Disease Control and Prevention (CDC), 45,979 people died by suicide in 2020 (https://www.cdc.gov/suicide/facts/index.html). It is the second leading cause of death among youth in the USA and the fourth leading cause of death in the world[1]. Therefore, early prediction and prevention of suicidal behavior are important. Traditional psychiatric treatments are not very helpful in the prediction of suicidal tendency due to the lack of biomarkers, laboratory tests, or screening reports which can assist in identifying suicidal risk or diagnosis. Some risk assessment tools are available for predicting suicide risk including psychological scales such as the Beck Suicide Intent Scale, and scales derived from statistical models, such as the Repeated Episodes of Self-Harm score[5]. However, studies show that the reliability of those tools is inadequate, costly, time-consuming, and positive predictive value fails to distinguish between low and high-risk patient[6]. Nowadays, the development of artificial intelligence in medical research has shown great potential to help predict and prevent suicide in many cases using survey datasets. However, predicting suicidal tendency using survey datasets can be challenging due to some intrinsic characteristics of survey data such as high dimensionality, multiple data types and missing values, and imbalanced class distribution. Thus, in this study, I perform a comparative study of six representative classification algorithms such as Naïve Bayes (NB), Bagging, Random Forest (RF), Multilayer perceptron (MLP), Logistic Regression (LR), and Adaboost(Ada) using Youth Risk Behavior Survey (YRBS) datasets and investigate the effectiveness of several data handling techniques to improve the overall performance of the predictive models for predicting suicide risk.

1.2 Literature review and related studies

Predicting suicidal tendency is challenging work and researchers engage in finding suicidal tendencies among children[19], adolescents[7,11-15,20], adults[17,18], mentally ill people[9,16], serviceman[21], and aged people. The data mining techniques and machine learning algorithms such as Naive Bayes[13], Logistic Regression[13,14], Support Vector Machine, Random Forest[7,11,13,14], Decision Tree[13], Artificial Neural Network, XGBoost[13], K- Nearest Neighbors[13] algorithm, and Bagging show significant improvement to predicting suicidal thoughts, behavior, and future attempts. The researchers of the existing research found some important factors that influence suicidal tendencies among various groups of people. Those important predictors can be identified as previous self-harm history[20], sex[20], age[20],demography[13], socia-economical status, family history[11],[13], race, relation with peers[13], sexual activity, victimization history, hopelessness[11], bipolar disorder[20], substance abuse[13], unhealthy diet etc. Marcel Miché et-al[14] , used Logistic Regression, and Random Forest to identify risk groups for suicide attempts among adolescents and young adults. The study was developed on the age group of 14-24 years old, and the data collected over 10 years of time frame from 1995 -2005. They achieved similar results for area under the curve (AUC) for four predictive models namely, logistic regression(0.828), lasso(0.826), ridge(0.829), and random forest(0.824). The results show all classifiers are similarly important for predicting suicidal attempts. In this study, prior suicidal attempts, education, and prior help-seeking are important features for the prediction of the suicide attempt. Orion Weller et-al[13], comparing between K-Nearest Neighbors, Naive Bayes, Logistic Regression, DecisonTree, XGBoost, and LightGBM classifiers for predicting suicidal behavior and thoughts among adolescents. They found tree-based classifier LightGBM has strong predictive power and outperformed all other classifiers with 91% accuracy for the test set. The significant predictors for their study are familial life, drug consumption, demographics, and peer acceptance in school. Colin G. Walsh et-al[7], found Random Forest successfully able to predict suicide attempts among adolescents with high AUC (0.8s–0.9s) for different control groups. The most important predictors for his study were body mass and medication to predict the depressed youths from the sample. Melissa Macalli et-al[11], showed in their study that random forest was the best classifier to identify suicidal thoughts and behavior among college students. The data was collected between 2013 and 2019 and 5066 student volunteers participated in the study. They partitioned the data into two groups – students who had suicidal thoughts and behaviors and students who did not have suicidal thoughts and behavior. They found better results for 1st group in terms of AUC (0.83–0.86, respective of girls and boys), sensitivity (0.79-0.81, respective of girls and boys), predictive positive values (0.40 – 0.36, girls-boys) which outperformed the 2nd group with AUC (0.72-0.74), sensitivity (0.63 – 0.64). They also showed that important factors varied gender-wise. For girls, depression, self-esteem, trait -anxiety, academic stress, and perceived stress were the top 5 factors influencing suicidal thoughts and behavior whereas, for boys self-esteem, trait -anxiety, the option of hobbies, perceived health, and, family source of income were the important factors for influencing suicide thoughts and behavior. Frank Iorfino et-al[20], conducted his study on 1962 young people at the Youth Mental Health Clinic of the Brain and Mind Center in Sydney. SMOTE and undersampling were applied to the dataset to make it a balanced dataset. Random Forest, Lasso Regression, Elastic Net Regression, Bayesian Additive Regression Tree, and Logistic Regression were used to predict self-harm tendency in youths. All models showed similar performance on the test dataset. The top features for their study were a history of self-harm, age, social and occupational functioning, sex, bipolar disorder, psychosis-like experiences, treatment with antipsychotics, and a history of suicide ideation. Haque[34] analyzed social media tweets using natural language processing and used both deep learning models and machine learning models to predict suicidal indentation. The study found that random forest performed best with an accuracy of 93% and an F1 score of 0.92 and outperform all other machine learning models whereas all deep learning models perform similarly with an accuracy 93.2% for predicting suicidal indentation. Huang[35] compared three predictive machine learning algorithms namely random forest, support vector machine, and decision tre,e and achieved a high accuracy of 87.3% and AUC of 92.4% for the random forest for predicting suicidal indentation. They achieved a high accuracy of 84% and auc of 90.1% for predicting depression on adolescents. They also find some influencing features for depression and suicidal indentation such as anhedonia, lack of social support, relationship with mother, and emotional neglect during childhood. In the study of predicting suicidal attempts among medical college students, Shen[36] achieved a high accuracy of 90.1%, auc 0.9255, a sensitivity of 73.5% and a specificity of 91.68% for the random forest classifier.

Table 1.1 presents a summary of the existing work done on machine learning in recent years in suicide prevention. The summary also depicts several influencing features for suicide attempt. In Table 1.1 the name of some classifiers is presented in green color. The highlights show that those classifiers achieved the best classification results for predicting suicide risks.

Table 1.1 *A summary of related studies that explore machine learning algorithms for predicting suicide risk*

Researcher	Year	Classifiers	Result	Imp features
Miché M[14]	2020	LR, lasso, ridge, RF	LR auc-0.828 Lasso auc – 0.826 Ridge auc – 0.829 Rf auc – 0.824	Prior suicidal attempt, education, prior help seeking
Weller O[13]	2021	KNN, NB, LR, DT, XGBOOST, LightGBM	Accuracy 91%	Familial life, drug consumption, peer acceptance
Walsh CG[7]	2018	RF, LR	AUC (0.8 – 0.9) depending upon control group	Diagnostic, medication, semieconomic factors
Macalli[11]	2021	RF	AUC (0.83 – 0.86) Sensitivity (0.79 – 0.81) PPV (0.63 – 0.64)	Girl – depression,self esteem, trait anxiety, academic stress Boys – self esteem, trait anxiety, family source of income
Frank[20]	2020	RF, lasso regression, elastic net regression, Bayesian additive regression tree, LR	AUROC (0.744–0.755)	Self harm, age, social and occupational functioning, sex, bipolar disorder, history of suicide indentation
Haque[34]	2022	Deep learning, machine learning (RF)	Acc 87.3% F1 – 0.92	

Table 1.1 continued

Huang[35]	2022	RF, SVM, DT	Suicidal indentation Acc – 87.3% AUC – 92.4% Depression Acc – 84% AUC – 90.1%	Anhedonia, lack of social support, relationship with mother, emotional neglect during childhood
Shen[36]	2020	RF	AUC – 0.9255 Acc – 90.1% Sensitivity – 73.5% Specificity – 91.68%	

In addition to the comparative study of classifications, the objective of my study is to find important characteristics from the dataset that may influence suicide risk. In my research, I found that sadness and depression, and bullying from peers (Ref. Table 4.5, 4.7,4.8) are two important factors influencing suicidal thoughts. Furthermore, an in-depth analysis reveals the fact that youth who are addicted to alcohol and drugs (Ref. Table 4.6) are more likely to attempt suicide.

1.3 Contribution to the work

In this work, data cleaning was done by eliminating noisy data and imputing missing values. The dataset was a high-dimensional, imbalanced dataset. Therefore, the imbalanced ratio was handled using different sampling techniques, and features were selected to reduce the dimension of the dataset and improve the overall performance of the prediction result. Different classification algorithms were applied to the test dataset to measure the performance of different classifiers. A comparative study was conducted to get an overview of the efficacy of the classifiers on a survey dataset.

In chapter II, data collection various data preprocessing techniques such as sampling and feature selection techniques are discussed. In chapter III, an overview of various classification algorithms and evaluation metrics are discussed. In chapter IV, the classification results of sampling and feature selection techniques are presented and discussed. In chapter V, further discussion about prediction results, limitations of this work, and the future scope of the study are depicted. Chapter VI provides a conclusion of the overall study.

Chapter II – DATA COLLECTION AND PROCESSING

2.1 Source of the Dataset

The dataset I used in my study is a publicly available dataset related to health risk problems among youth. The dataset is known as the "National YRBS Datasets of 2019" which is the subset of the "Youth Risk Behavior Survey Dataset" and can be downloaded from the website of CDC (https://www.cdc.gov/healthyyouth/data/yrbs/data.htm). This is a survey dataset, and the survey was conducted on 136 high school students in the United States in 2019.

2.2 The overview of the Dataset

The dataset consisted of 99 survey questions on health-related problems among youth. The questions were answered by 13,678 students. The questions can be categorized into six categories which included a)behaviors that contribute to unintentional injuries and violence, b)sexual behaviors that contribute to unintended pregnancy and sexually transmitted infections, including HIV infection, c)alcohol, and other drug use, d)tobacco use, e)unhealthy dietary behaviors, and f)Inadequate physical activity. The dataset had two types of data values, numerical and categorical. A predefined encoding technique was already applied to the dataset by the data provider. The encoding was based on the Response of Interest (ROI). For example, if a question had four answer choices and among them three answer choices related to the health risk behavior of the youth then those three answer choices are considered as ROI. Answer choices falling under ROI were set as 1 and all other responses for that question were set as 2. Therefore, the categorical columns of the entire dataset were converted to numerical values depending on the ROI. The dataset also contained missing values.

Table 2.1 shows the type of survey questions that were included in the questionnaire. The green-colored answer choices are ROI. Therefore when the dataset was converted to numerical numbers those green-colored answer choices were set as 1 and all other answer choices were set as 2.

Table 2.1: *An overview of the dataset*

Sl_No	Q_No	Questions	Answer choices
1	Q26	During the past 12 months, did you ever seriously consider attempting suicide?	Yes No
2	Q23	During the past 12 months, have you ever been bullied on school property?	Yes No

Table 2.1 continued

3	Q47	During the past 30 days, how many times did you use marijuana?	0 times 1 or 2 times 3 to 9 times 10 to 19 times 20 to 39 times 40 or more times
4	Q19	Have you ever been physically forced to have sexual intercourse when you did not want to?	Yes No
5	Q77	During the past 7 days, on how many days did you eat breakfast?	0 days 1 day 2 days 3 days 4 days 5 days 6 days 7 days

After applying the ROI data preprocessing techniques table 2.1 looks like the following (table 2.2). Q26 is the class attribute. As per the ROI preprocessing the answer “yes” for Q26 was changed to 1 and “no” was changed to 2. However, in this work, I followed the traditional encoding to change the class label. Therefore, I further changed the “yes” label to 1 and the “no’ label to 0 using python code.

Table 2.2 *An overview of preprocessed dataset*

Sl_No	Q_No	Questions	Answer choices
1	Q26	During the past 12 months, did you ever seriously consider attempting suicide?	1 2
2	Q23	During the past 12 months, have you ever been bullied on school property?	1 2
3	Q47	During the past 30 days, how many times did you use marijuana?	2 1 1 1 1 1
4	Q19	Have you ever been physically forced to have sexual intercourse when you did not want to?	1 2

Table 2.2 continued

5	Q77	During the past 7 days, on how many days did you eat breakfast?	1 2 2 2 2 2 2 2

2.3 Data Preprocessing

The dataset contained missing values. Therefore, missing values were required to be deleted. In my study I eliminated the instances and features depending upon three criteria, i) eliminating the features that had more than 50% instances missing, ii) eliminating the instances which did not have any class values iii) eliminating the features which were strongly correlated with the class feature. I have excluded features that have 50% missing values. I considered question26 (table 2.1 & 2.2) was the class label for the study. I observed 240 instances from the class label were missing. I removed those instances and use 13437 instances for our prediction. I also observed Q27, Q28, Q29 had strong correlation with question 26, the class label. Using strongly corelated features for training can give me very high accuracy which can be misleading therefore, I eliminated those three features. Therefore, I eliminated total 16 features to clean the dataset. There were instances where less than 20% of values were missing. I imputed the missing values using mean imputation method. After clean the data I used 76 features and 13,437 instances for our study. The dataset had an imbalanced class distribution (4). The class of interest was the minority class. I used different sampling techniques to handle the data imbalanced problem and different feature selection techniques to find out the important features for predicting suicide tendency.

2.3.3 Feature selection methods and their implementation

The dataset that I used, was a high-dimensional dataset. It is challenging to handle high dimensional data (multiple features) and trained the model using traditional classification algorithms. When the dimensionality increases, the volume of the space increases so fast that the available data becomes sparse. Therefore, the amount of data that need for accurate classification also grows exponentially. This is known as the curse of dimensionality[25]. Therefore, it is important to find out those features which have significant effect on classification. Noisy features can be discarded from the training and test set to improve the performance of the machine learning models. Feature selection is one of the methods of discarding noisy features and selecting important features. I used three feature selection techniques to find out the features which have a significant impact on classification results. I used the sciKit-learn package to different feature selection techniques. The three different feature selection methods that I used for my study is i)

Random forest feature importance, ii) Mutual information feature importance and iii) Recursive feature elimination method.

2.2.3.1 A description of Random Forest feature importance

Each tree of the random forest can calculate the importance of a feature according to its ability to increase the pureness of the leaves. The higher the increment in leaves purity, the higher the importance of the feature. This is done for each tree, then the averaged value of all trees were taken. Finally, the average value normalized to 1. In this study I used random Forest from skLearn ensemble package to train the classifier. The attribute, *feature_importances_* of the random forest classifier gives the importance of each feature in the order in which the features are arranged in the training dataset. I used top 40 attributes for classification.

2.2.3.2 A description of Mutual information feature selection

Mutual information (MI)[2] is a measure between two random variables X and Y, that quantifies the amount of information obtained about one random variable, through the other random variable. The mutual information can be represented as

$$I(X,Y) = \sum log \frac{p(x,y)}{p(x)p(y)} \qquad (1)$$

P(x,y) is the joint probability density functions of X and Y, p(x) and p(y) marginal density function.

The alternative way of write down mutual information is

$$I(X,Y) = H(X) - H(X|Y) \qquad (2)$$

Where H(X) can be defined as marginal entropy, H(X|Y) is the conditional entropy. If H(X) represents the measure of uncertainty about a random variable, then measures what Y does not say about X. This is the amount of uncertainty in X after knowing Y and this substantiates the intuitive meaning of mutual information as the amount of information that knowing either variable provides about the other. In our method, a mutual information measure is used to calculate the information gain among features as well as between feature and class attributes.

The range of the MI score should be in between 0 to ∞. The higher value indicates that there is a close connection between feature and class level, therefore, the feature can be used for prediction. If the MI score is 0 or very low like 0.01. the low score suggests a weak connection between this feature and the class level.

In this study, I used skLearn package for mutual information feature selection method.

2.2.3.3 A description of recursive feature elimination method for feature selection

Recursive feature elimination(RFE) method is a process of eliminating the features from the dataset that have weak connection with class label. It is an iterative process. It can iteratively eliminate the weak features until desired number of features number reached. It is a wrapper-based feature selection technique. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. I used REF method from skLearn feature selection package. The parameters that was used for REF are as follows, I used logistic regression as estimator, multinomial as mutli_class. The number of iteration has been used for the regression model was 2000 and I selected top 40 features using this wrapper method.

2.3.4 Sampling methods and their implementation

In an imbalanced dataset, the class distribution is not equal. Mostly, imbalanced data distribution is skewed to the majority class. The majority class of imbalance class[26] level distribution has a great influence on the classification result. Therefore, we cannot rely on the evaluation metrics. If the classifiers achieve an accuracy of more than 90% for an imbalanced dataset that does not mean the classifiers are able to detect the minority class efficiently. The experimental dataset must have a balanced class distribution to obtain efficient classification results. Different types of data sampling techniques, namely random oversampling, random undersampling, and Synthetic Minority Oversampling Technique (SMOTE) were used in this study to balance class distribution.

2.3.4.1 Random oversampling method

Random oversampling can be defined as the random replication of minority class to make the dataset a balanced dataset. Although there is no possibility of data loss in this method, this method often suffers from overfitting problems. I used randomoversampler method which is available on imblearn over sampling package of python. The minority sampling technique was used as a parameter of the randomoversampler to assign a balanced class distribution.

2.3.4.2 Random undersampling method

Random undersample method was used to obtain a balanced dataset where we randomly discarded some data samples from the majority class. Although we obtained a balanced dataset after applying this method, it can cause some information loss from the dataset. I used randomundersampler method which is available on imblearn under sampling package of python.

The majority sampling strategy has been used as the parameter of randomundersampler to make a balanced class distribution.

2.3.4.3 Synthetic Minority Oversampling Technique (SMOTE)

In this method new synthetically generated samples add to the minority class. The new instances are called synthetic instances because those minority instances are created from existing minority instances of the dataset. This method reduces the problem of overfitting.

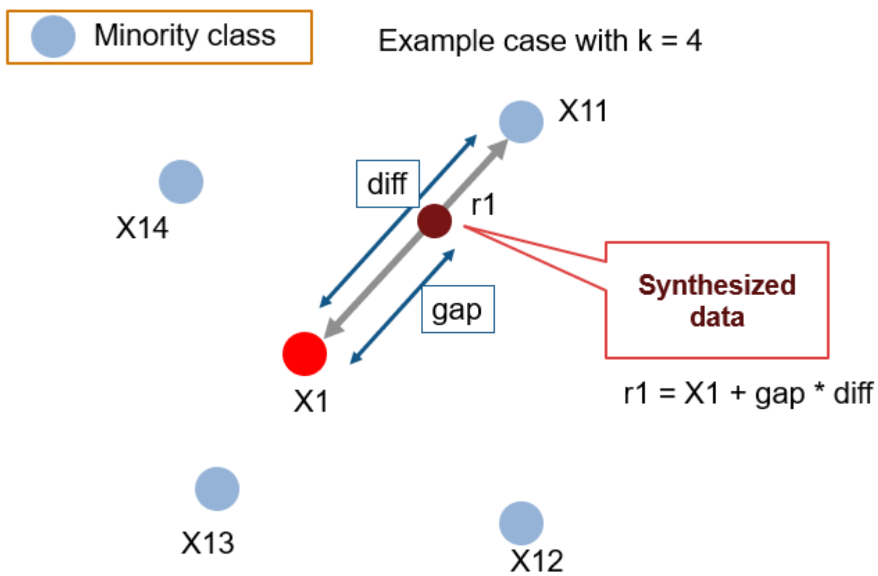

Figure 1 *An overview of SMOTE method* (https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/)

Figure 1 (SMOTE) describe the methodology of SMOTE. In the figure, r1 described as a synthetic data which need to be generated to balance the minority class. All blue circles indicate the minority class. In the above example k =4, k refers the number of nearest neighbors. SMOTE algorithm selects one of the minority class instances randomly, in this case X1. After that it will consider nearest neighbors (in this case no of nearest neighbors is 4) near to the minority class instance X1. As per the figure the algorithm chooses X11 as the nearest neighbor of X1. The r1 can be generated by finding the distance between X1 and X11 which will be multiplied with the a random value between 0 and 1 and adding the chosen minority class instance (X11).

$$r1 = X1 + (X11 - X1) * rand(0,1) \qquad (3)$$

The method is effective because new synthetic instances are generated from the minority class that are plausible, relatively close in feature space to existing instances of the minority class. However, the main drawback of the algorithm is SMOTE only consider the neighboring instances of one class and ignore the neighboring instances of the other class. Therefore, it can increase the overlapping of classes and generate noise. I used SMOTE method which is available on imblearn over sampling package of python.

Table 2.3 the class distribution of the dataset. My problem was a binary classification problem therefore, there were two class values in the dataset one is the “yes” class. Yes class was a minority class and the class of interest. The other one is the majority class and is defined as “no” class. The dataset was divided into the training set and the test set. If no sampling technique was used on the dataset the training set had 1843 yes class instances and 7562 no class instances and the test set had 790 yes class instances and 3242 no class instances. The imbalanced ratio was 4. The oversampling method made the class distribution for the training set a balance class distribution with 7562 yes class instances and 7562 no class instances. Undersampling and SMOTE also made the training set class distribution a balanced class distribution. Sampling methods only applied to the training set, test set remained imbalanced.

Table 2.3 *An overview of class distribution for different sampling techniques*

Sampling techniques	Class distribution on training set		Class distribution on test set	
	Yes	No	Yes	No
No Sampling	1843	7562	790	3242
Oversampling	7562	7562	790	3242
Undersampling	1843	1843	790	3242
SMOTE	7562	7562	790	3242

2.3.5 Data partition method

The preprocessed data were partitioned into two sets training set and a test set using stratified data partition techniques. I used the stratified data partitioning technique to keep the imbalance ratio intact in the training and test sets. The training set consisted of 70% or 9405 instances of the total dataset. The test set contained 30% or 4032 instances of the total dataset.

Table 2.4 shows the class distribution for the entire dataset as well as the training set and test set after using stratified partitioning techniques. The dataset had 2633 instances of suicidal tendencies and 10804 instances of no suicidal tendencies. The data set had an imbalanced ratio of 4. partition method. The training and test sets kept the imbalanced ratio of 4 intact after using the stratified data partition method. I used the train_test_split method from skLearn to split the data into training and test sets.

Table 2.4 *An overview of class distribution on training and test dataset*

Methods	Suicidal tendencies	Non suicidal tendencies	Total number of instances	Imbalance ratio
Dataset	2633	10804	13437	4
Training set (70%)	1843	7562	9405	4
Test (30%)	790	3242	4032	4

Chapter III – Model training methods

This section describes the classification methods that we used for prediction. Six predictive classification algorithms including naïve Bayes (NB), logistic regression (LR), multilayer perceptron (MLP), Ada Boost, random forest (RF), and bagging were used to predict suicide risk among youths. These six algorithms have been widely accepted algorithms to solve binary classification problems. Therefore, we decided to train our data using these six algorithms to predict suicidal tendencies.

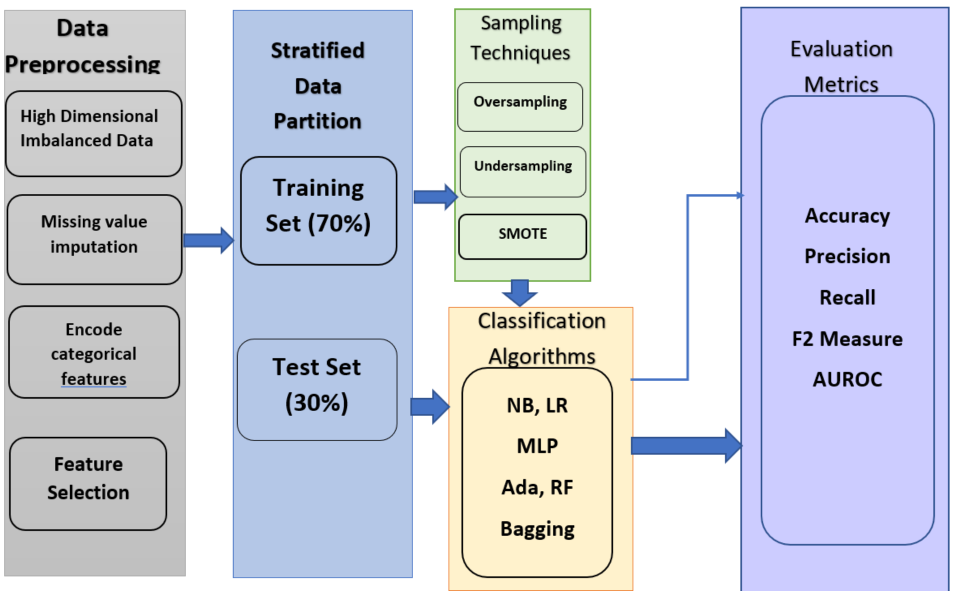

Figure 2 *The workflow diagram*

Figure 2 depicts the workflow of the entire study. The grey-colored column shows the data preprocessing techniques. The blue colored column shows the data partition techniques. The green colored column shows the three sampling strategies used to balance class distribution. The classification algorithms are shown using the yellow-colored column. The purple-colored column shows the evaluation metrics. The arrows show the connection between the two phases. The data partitioned stage has two outgoing arrows connecting the sampling phase from the training set and the classification phase from the test set. This describes the sampling techniques applied only to the training set. There is a downward arrow from the sampling technique phase depicting that the classification algorithms were applied to the training set after applying the sampling techniques to the training set. There are two outgoing arrows from the classification algorithms phase. The thin arrow refers to the results of classification algorithms getting from the training set. The thick arrow refers to the results of classification algorithms getting from the test set.

3.1 Machine Learning Algorithms

This work is a comparative study of six machine learning algorithms. I used three single predictive classifiers such as naïve Bayes (NB), logistic regression (LR), multilayer perceptron (MLP), and three ensembled classifiers such as adaboost (Ada), random forest (RF), and bagging to train the models and predict the suicidal risk.

3.1.1 Naïve Bayes

The naïve bayes, a supervised machine learning algorithm, is a probabilistic approach for classifying a binary class classification problem. This method is based on conditional independence, that is the values of the attributes are assumed not to depend upon each other given the class label. We used categorical naïve bayes model from skLearn naïve bayes package. The model learned class prior probability, the prior probability adjusted according to the data and the additive smoothing parameter set to 1.0. Those were the hyperparameter that we used to train the model.

3.1.2 Logistic regression

Unlike linear regression, logistic regression, a supervised machine learning algorithm, is used for binary classification. The method uses complex sigmoid cost functions to find the probability between 0 to 1. In our study, we used logistic regression classifier from skLearn linear model package. The hyperparameter that we used for this study is as follows, we used Limited-memory Broyden–Fletcher–Goldfarb–Shanno (lbfgs) as the solver of the model, maximum iteration 1000 epochs so that the model can converge, and multinomial loss function which can fit across the entire probability distribution, even when the data is binary, and an L2 regularization to reduce the chance of model overfitting.

3.1.3 Multilayer Perceptron (MLP)

A multilayer feed-forward neural network which is also known as multilayer perceptron (MLP) consists of an input layer that defines the input value, one or more hidden layers with non-linear activation function, and an output layer that defines the final outcome. The classifier uses backpropagation, an iterative learning process, to learn a multi-layer perceptron to classify instances. The MLP classifier from skLearn neural network package had been used to train the model. The following hyperparameter had been used to train the model. We used relu as an activation function, adam as a solver, a learning rate of 0.001, a momentum of 0.2 and the maximum iteration is 200 epochs. There were four hidden layers in our model. First hidden layer contained 256 nodes, second hidden layer contained 128 nodes, third hidden layer contained 64 nodes followed by 32 nodes.

3.1.4 Ada Boost (Ada)

AdaBoost also called Adaptive Boosting is a technique in Machine Learning used as an Ensemble Method. In our study, we used Ada Boost classifier from the sklearn ensemble package. The following hyperparameter had been used to build the model. I used the decision tree classifier with max_depth 1 as a base estimator. The 50 number of estimators were used at which boosting was terminated. I used the learning rate and algorithm 1.0 and SAMME.R respectively.

3.1.5 Bagging

Bagging is an ensemble method. Bagging can train a number of different models on different randomly-selected subsets of the training data. After the training has been completed bagging combined the prediction result depends on the voting scheme. In our study, I used bagging classifier from the sklearn ensemble package. The following hyperparameter had been used for the model. The base estimator was the decision tree classifier. In this study, we combined 10 decision trees models to get the generalized result. Maximum 8 features were required to train the each base estimator.

3.1.6 Random Forest (RF)

The random forest algorithm is an ensemble method, an extension of the bagging method which is based on both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness generates a random subset of features, which ensures low correlation among decision trees. During classification, each tree votes, and selected the most frequent class among all the trees. In our study, we used random forest classifier from the sklearn ensemble package. The following hyperparameter had been used for the model. The number of trees in the forest was 100. We used Gini impurity to measure the quality of the split. The trees can be expanded until all leaves are pure or until all leaves contain less than min_samples_split samples. Therefore, we used max_depth= none parameter. The minimum number of samples required to split an internal node was two. The parameter can be defined by min_samples_split. The number of features to consider when looking for the best split was the square root of all features.

3.2 Evaluation Methods

In this study, I used six evaluation metrics to measure the performance of the classifiers. The evaluation metrics are accuracy, precision, recall, F2 measure, AUROC and confusion matrix. Accuracy can be more reliable for a balanced dataset however in this study an imbalanced dataset was used. Therefore, the classification results of the recall and F2 measure are more important to identify the classification performance. The aim of the study is to minimize false negatives and maximize true positives.

3.2.1 Confusion Matrix

A confusion matrix is a summary of prediction results made by the classifier in a tabular format. Each row of the confusion matrix indicate the actual class and each column of the confusion matrix indicate the predicted class. Table 3.1 depicted an overview of confusion matrix. True positive define correctly predicted students who have suicide risk. True negatives define correctly predicted students who do not have suicide risk. False positives define students who have misclassified as suicidal. False negatives refer students who have misclassified as not suicidal.

Table 3.1 *An overview of confusion matrix*

	Predicted positive	Predicted Negative
Actual positive	True positives (TP)	False negative (FN)
Actual negative	False positives (FP)	True negatives (TN)

3.2.2 Accuracy

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. To find the effectiveness of an algorithm relying only on the accuracy metric could be misleading. In particular, working with an imbalanced dataset can create biases toward the majority class. The dataset that has been used in the research is an imbalanced dataset therefore achieving high accuracy can be misleading for prediction. Other evaluation metrics can be considered to get the overall idea of the performance of an algorithm.

$$Accuracy = \frac{TP + TN}{P + N} \quad (4)$$

3.2.3 Precision (or positive predictive value)

Precision can be thought of as a measure of exactness (i.e., what percentage of tuples labelled as positive are actually such). The class of interest in this research is the positive class. We consider the precision of positive classes in our evaluation. Precision refers here out of all suicide tendencies prediction how many cases have true suicide tendencies. Precision metrics can be misleading for a highly skewed dataset.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

3.2.4 Recall (or sensitivity)

The recall is a measure of completeness (what percentage of positive tuples are labeled as such). Recall is an important metric for our research. In our cases, predicting more false negatives are costly than predicting false positives. Recall refers to out of all actual suicide tendencies class label how many suicide tendencies our classifier correctly identifies.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

3.2.5 F- measure

The F measure is the harmonic mean of precision and recall. It gives equal weight to precision and recall.

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

3.2.6 F2 -Measure

It has the effect of lowering the importance of precision and increase the importance of recall. If maximizing precision minimizes false positives, and maximizing recall minimizes false negatives. Then F_2 measure puts more attention on minimizing false negatives than minimizing false positives.

$$F2 - Measure = \frac{5 \times Precision \times Recall}{4 \times Precision + Recall} \quad (8)$$

3.2.7 Receiver Operator Characteristics (ROC) curve

Receiver Operator Characteristics curve (ROC) curves is the ratio of the number of correctly classified positives examples (True positive rate- TPR) to the number of incorrectly classified negative examples (False Positive Rate -FPR).

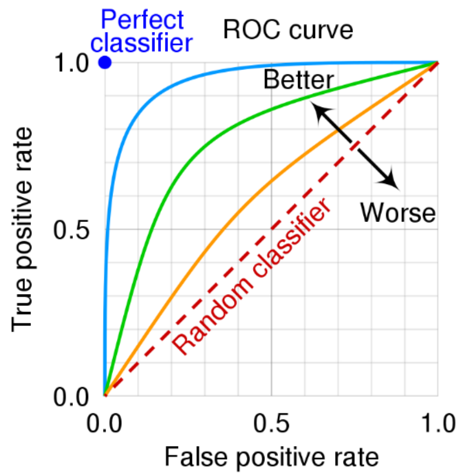

Figure 3 *ROC curve* (https://en.wikipedia.org/wiki/Receiver_operating_characteristic)

Figure 3 depicts the ROC curve. A ratio of 0.5 can be described as random guess of the classifier. Any value over 0.5 can be consider as a good prediction and any value less than 0.5 can be consider as bad prediction. In this scenario we can discard that model.

3.2.8 Area Under Receiver Operator Characteristics (AUROC) Curve

AUROC measures the entire two-dimensional area underneath the entire ROC curve from (0,0) to (1,1). In the figure 3 the area under the blue curve refers as AUROC.

Chapter IV – Result and Discussion

A total of 13437 responses were used to build the predictive models using machine learning algorithms. A class distribution is shown in Table 2.3 and Table 2.4. The dataset was an imbalanced dataset. As per the imbalanced ratio, suicide attempters were a quarter of non-suicide attempters. Therefore, three different data sampling methods were used to balance the dataset. The dataset was partitioned into a training set(70%) and a test set(30%) using a stratified partition method (ref Table 2.3). After data preprocessing, the top 40 features were selected using different data filtering techniques, and the models were trained using the top 40 selected features. Feature selection methods were applied to both the training and test set whereas sampling methods were applied only to the training set. The following tables of the prediction results present the classification reports of the positive class of the test set.

4.1 Prediction results and discussion with all features

Table 4.1 shows the classification results of the different classifiers with all the features. The bar chart of figure 4 is the visual representation of the classification results of Table 4.1. The number of features used for classification was 76. Evaluation metrics used to analyze the results were accuracy, precision, recall, F2 measure, the area under the receiver Operator Characteristics curve (AUROC), and confusion matrix. From table 4.1 and figure 4, we can see that all the classifiers except the NB achieved an accuracy of more than 80%. However, classification accuracy is misleading for skewed class distribution. Further analysis, showed that although NB achieved poor accuracy NB predicted the highest number of true positives (448) instances. Also, NB achieved the highest recall (0.57) ratio. Therefore, accuracy solely could not give a clear representation of classification results. All other classifiers achieved poor recall and F2 values with respect to NB. Further analysis of classification results using different data handling techniques is needed to improve recall and F2 scores.

Table 4.1: Prediction results of different classifiers with all features(wos)

Classifier	**Acc**	**TP**	**TN**	**FP**	**FN**	**Pre**	**Recall**	**F2**	**AUROC**
NB	0.78	448	2686	556	342	0.45	0.57	0.54	0.79
LR	0.83	340	3024	218	450	0.61	0.43	0.46	0.85
MLP	0.83	296	3069	173	495	0.63	0.37	0.40	0.85
Ada	0.84	352	3024	218	438	0.62	0.45	0.48	0.86
Bagging	0.83	305	3035	207	485	0.6	0.39	0.42	0.83
RF	0.83	228	3116	103	576	0.67	0.30	0.34	0.85

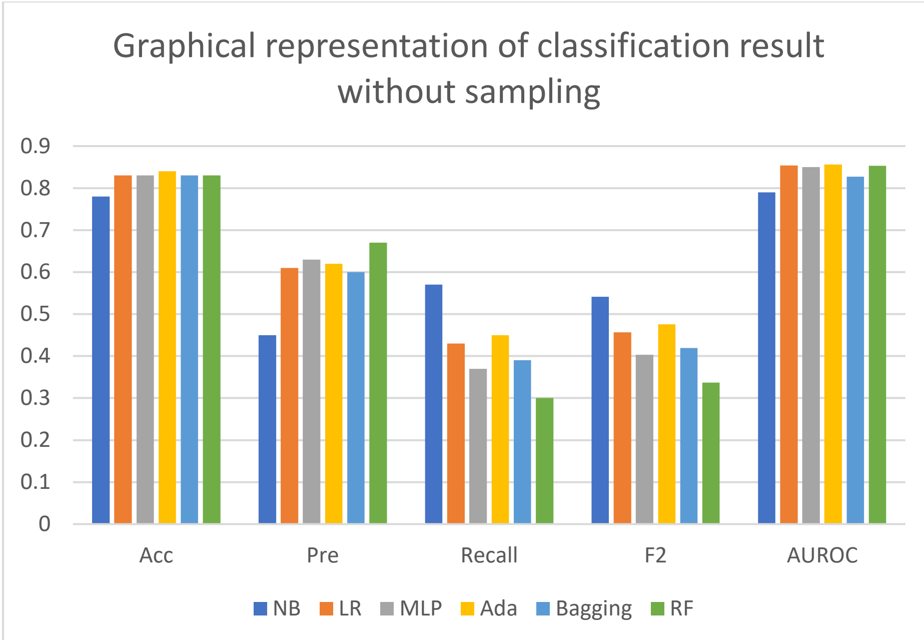

Figure 4 *Graphical representation of classification results with different classifiers*

Table 4.2 shows the classification results of six classifiers with different sampling methods with all features. Figure 5 is a graphical representation of classification reports with sampling methods. Logistic regression and Ada Boost (Table 4.2) with the oversampling method and random forest with the undersampling method achieved the highest recall. As shown in figure 5, random forest with undersampling methods demonstrated the highest recall ratio (0.84) followed by LR with oversampling (0.82), and Ada with oversampling (0.81) method. The recall indicates the ability of the classifiers to correctly predict the students who have suicide risk.

Also, RF with the undersampling method achieved the highest F2 measure (0.72), followed by LR with oversampling (0.71) method and Ada with oversampling (0.70) method. The highest F2 score of RF indicates that RF was able to identify as many positive classes as possible.

As per the results of the confusion matrix of Table 4.2, the RF classifier was able to predict 662 students, the LR classifier was able to predict 647 students, and Ada was able to predict 643 students out of 790 positive class students who have suicidal tendencies. In general, the classification results of SMOTE are the worst with respect to all other sampling methods for each classifier.

Table 4.2: *Prediction results of different classifiers with different sampling methods*

Classifier	**Sampling**	**Acc**	**TP**	**TN**	**FP**	**FN**	**Pre**	**Recall**	**F2**	**AUROC**
NB	Over	0.74	550	2416	826	240	0.40	0.70	0.61	0.79
	Under	0.73	548	2392	850	242	0.39	0.69	0.60	0.78
	SMOTE	0.71	350	2516	726	440	0.33	0.44	0.41	0.69
LR	Over	0.78	647	2479	782	156	0.46	0.82	0.71	0.85
	Under	0.77	645	2446	796	145	0.45	0.82	0.70	0.85
	SMOTE	0.77	645	2464	778	145	0.45	0.82	0.70	0.85

Table 4.2 continued

MLP	Over	0.80	350	2860	382	440	0.48	0.44	0.45	0.78
	Under	0.77	633	2477	765	157	0.45	0.80	0.70	0.85
	SMOTE	0.78	425	2705	537	365	0.44	0.54	0.52	0.76
Ada	Over	0.77	643	2476	766	147	0.46	0.81	0.70	0.85
	Under	0.77	643	2468	774	147	0.45	0.80	0.70	0.85
	SMOTE	0.82	466	2826	416	324	0.53	0.59	0.58	0.84
Bagging	Over	0.82	330	2966	276	460	0.54	0.42	0.44	0.81
	Under	0.77	582	2531	711	208	0.45	0.74	0.65	0.82
	SMOTE	0.81	286	3000	242	504	0.54	0.36	0.39	0.81
RF	Over	0.83	360	2994	248	430	0.59	0.46	0.48	0.85
	Under	**0.77**	**662**	**2440**	**802**	**128**	**0.45**	**0.84**	**0.72**	**0.85**
	SMOTE	0.83	300	3044	198	490	0.60	0.38	0.41	0.85

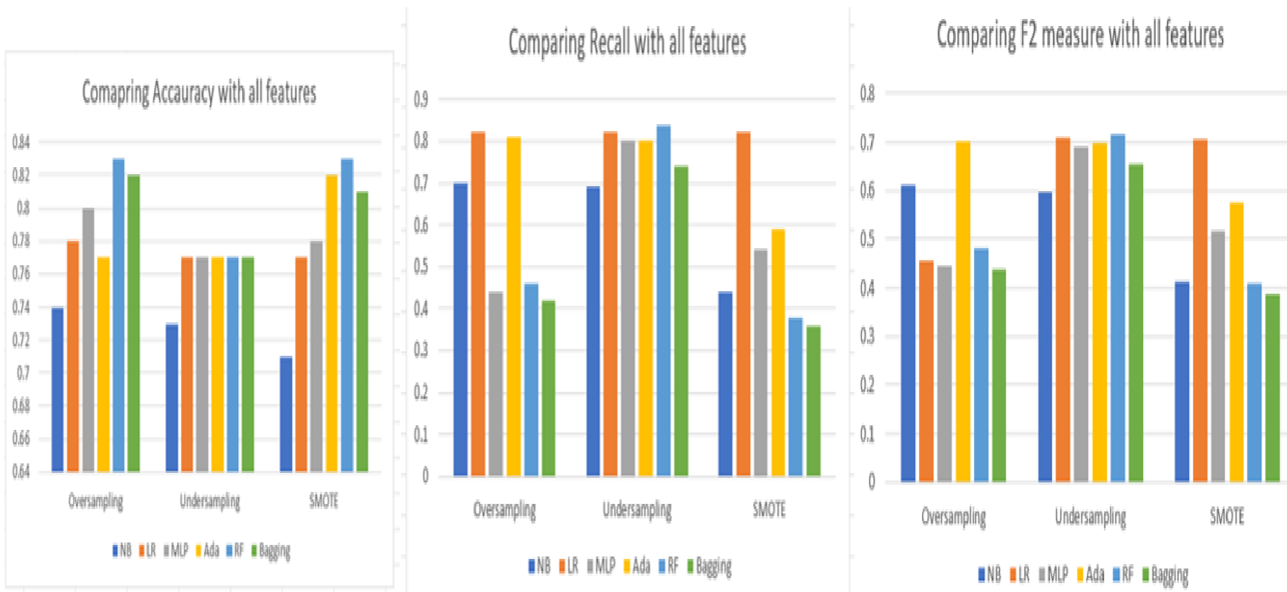

Figure 5 *Comparison results of accuracy, recall and F2 measure*

The aim of my study is to maximize the true positive instances and minimize the false negative instances. The recall is one of the very important metrics for the class imbalanced problem. The recall is a ratio of true positives and false negatives. Therefore, a high recall ratio is certainly an indication of improvement of true positive instances and deterioration of false negative instances. There is a significant improvement in the recall rates using oversampling, undersampling, and SMOTE methods compared to no sampling methods (Table 4.1). The classifiers were able to predict more students who have a suicidal tendency (true positives) when applying sampling methods on the training models compared to no sampling method. However, this improvement is not very significant for SMOTE. One of the reasons for that, the dataset we used was a categorical dataset and also an imbalanced dataset. However, SMOTE works well with a numerical dataset and as it generated a lot of synthetical data points the models can suffer from the overfitting problem.

4.2 Effectiveness of feature selection method

The feature selection method is important for my dataset, as the dataset I used for my study is a high-dimensional dataset and feature selection reduces the dimensionality of the dataset. Redundant data need to be eliminated to improve the models' training time and prediction results. Three feature selection methods were used in this study to reduce the dimensionality of the dataset. Table 4.3 shows the top 40 features among 76 features using the mutual information (MI) feature selection method. I used the mutual information feature selection method from the skLearn package of python.

Table 4.3 *Top 40 selected features using MI feature selection method.*

Serial No	Rank value	Features	F type	Serial No	Rank value	Features	F type
1	0.116021	Qn25	Sad & hopeless-Ness	**21**	0.009263	Qn43	Largest number of drinks
2	0.032856	Qn23	Bullying at school	**22**	0.008662	Qn78	Physical activity >= 5 days
3	0.024720	Qn24	Electronic bully-Ing	**23**	0.008189	Qn46	Initiation of marijuana use
4	0.019272	Qn49	Ever prescription pain medicine use	**24**	0.007658	Qn55	Ever steroid use
5	0.018059	Qn91	Ever used LSD	**25**	0.007515	Q8	Seat belt use
6	0.017364	Q2	What is your sex	**26**	0.007359	Qn40	Initiation of alcohol use
7	0.017201	Qn68	Weight loss	**27**	0.007265	Qn84	HIV testing
8	0.016981	Qn35	Current electronic vapor use	**28**	0.005990	Qn60	Multiple sex partners
9	0.015638	Qn57	Illegal drugs at school	**29**	0.005963	Qn47	Current marijuana use
10	0.015063	Qn37	Current smokeless tobacco use	**30**	0.005512	Qn72	Potato eating
11	0.014933	Q52	Ever heroin use	**31**	0.005104	Qn51	Ever inhalant use

Table 4.3 continued

12	0.014933	Qn15	Safety concerns at school	**32**	0.004805	Q6	How tall are you
13	0.013286	Qn22	Physical dating violence	**33**	0.004603	Q56	Illegal injected drug use
14	0.012057	Qn34	Electronic vapor product use	**34**	0.004519	Qn73	Carrot eating
15	0.011614	Qn13	Weapon carrying at school	**35**	0.004519	Qn45	Ever marijuana use
16	0.011358	Qn30	Ever cigarette use	**36**	0.004307	Qn41	Current alcohol use
17	0.010601	Qn17	Physical fighting	**37**	0.003874	Qn67	Perception of weight
18	0.010492	Qn58	Ever sexual intercourse	**38**	0.003790	Qn61	Current sexual activity
19	0.009273	Qn80	Computer use	**39**	0.003704	Qn88	Sleep
20	0.009272	Qn93	Plain water	**40**	0.003470	Q4	Are you Hispanic/Latino

As shown in Table 4.3, sadness and depression and bullying at school, and bullying using electronic medium are the influencing factors for attempting suicide. I further categorized those top 40 attributes into six categories. Table 4.4 shows the six categories and the features related to those categories. As shown in Table 4.4, the most populated category is alcohol and other drug use with 16 features followed by behavior that contribute to unintentional injuries and violence, sexual behaviors related to unintended pregnancy and sexually transmitted diseases, including HIV, unhealthy dietary behavior, inadequate physical activity, and tobacco use. The names of six categories are pre-determined by the owner of the dataset and can be obtained from the YRBBS website. I determined which features belong to which category depending on the type of features.

Table 4.4 *Categorization of top 40 attributes*

Behaviors that contribute to unintentional injuries and violence	Qn25, Qn23, Qn24, Qn22, Qn13, Qn17, Q8
Sexual behaviors related to unintended pregnancy and sexually transmitted diseases, including HIV infection	Qn58, Qn84, Qn60, Qn61
Alcohol and other drug use	Qn49, Qn91, Qn35, Qn57, Qn37, Q52, Qn34, Qn43, Qn46, Qn55, Qn40, Qn47, Qn51, Q56, Qn45, Qn41
Tobacco use	Qn30
Unhealthy dietary behaviors	Qn93, Qn72, Qn73
Inadequate physical activity	Qn80, Qn78, Qn88

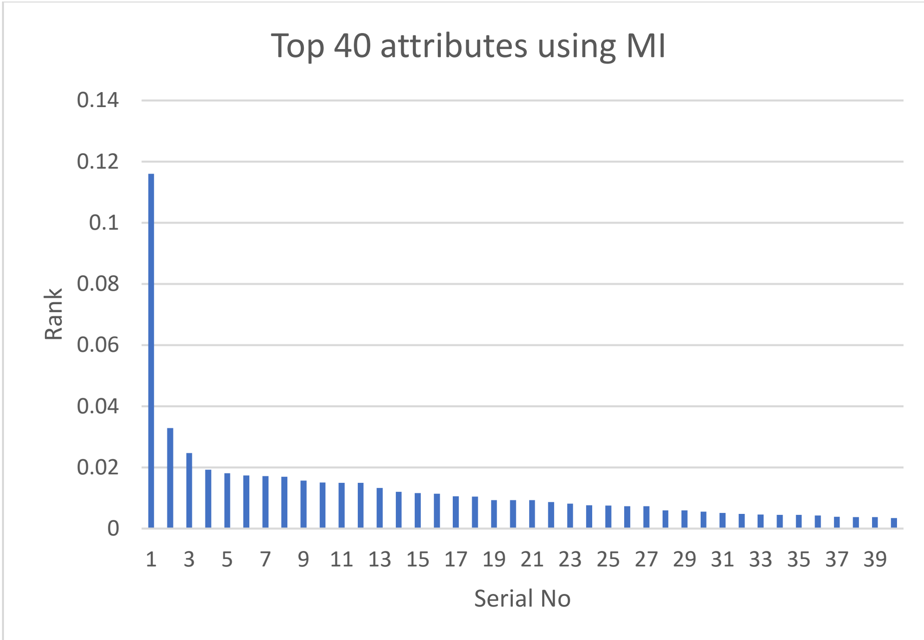

Figure 6 *Top 40 features using mutual information feature selection method*

Figure 6 shows the bar chart of the top 40 features. It can be shown from the bar chart that the ranks of the features started from 0.12 and gradually decreased. The ranks of the last few features are almost zero. We can also find from the bar chart that few features such as Q2(sl_no:6) and Q68(sl_no:7) have an almost similar ranking, Qn55(sl_no:24), Q8(sl_no:25), Qn40(sl_no:26), Qn84(sl_no:27) have similar ranking while there is a certain decrease in ranking between Qn84(sl_no:27) and its next feature Qn60(sl_no:28). Similar ranks have the same effect on prediction results.

4.2.1 Prediction results and discussion with 40 features

The MI method was used for selecting the top 40 features from the survey dataset. Table 4.5 shows the prediction results of the different classifiers using the top 40 features. The classification result of the NB classifier presented in Table 4.5 achieved the highest recall, F2 measure, and AUROC without sampling methods. Ada achieved the highest accuracy (0.84) followed by LR, MLP, RF, Bagging, and NB. The dataset is an imbalanced dataset therefore I considered a recall and F2 measure as two important prediction metrics for predicting suicidal risk. As per Table 4.5, all the classifiers achieve low recall and F2 measures. Therefore, I applied sampling techniques (Table 4.6) to improve the prediction results.

Table 4.5: *Classification results of 40 top features using MI feature selection methods*

Classifiers	Acc	TP	TN	FP	FN	Pre	Recall	F2	AUROC
NB	**0.80**	**457**	**2750**	**492**	**333**	**0.48**	**0.58**	**0.56**	**0.80**
LR	0.83	325	3034	208	465	0.61	0.41	0.44	0.85
MLP	0.83	259	3084	158	531	0.62	0.33	0.37	0.84
Ada	0.84	343	3029	213	447	0.62	0.43	0.46	0.85
Bagging	0.81	285	2996	246	505	0.54	0.36	0.38	0.80
RF	0.83	280	3073	169	510	0.62	0.35	0.38	0.84

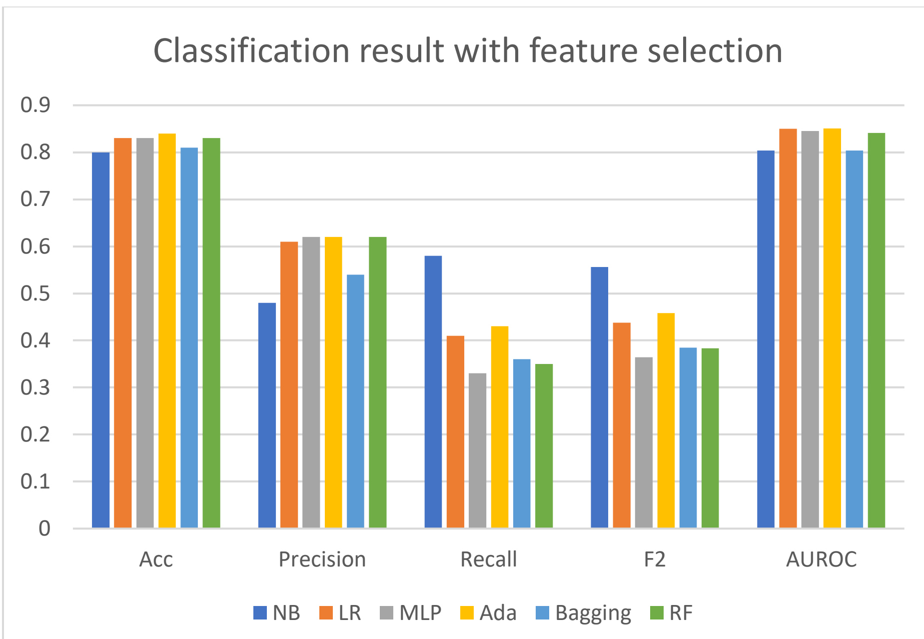

Figure 7 *Graphical representation of classification result with 40 top features(MI)*

As shown in figure 7, all classifiers achieved high accuracy but low recall and F2 score. This analysis shows that the classifiers predicted less number of positive classes. Table 4.5 shows that the TP varies from 457 instances to 259 instances which is slightly more than 50% of all positive instances. Therefore, the classifiers predicted only half or less than half of the total true positive instances from the test dataset.

Table 4.6 shows the classification results of the different classifiers using three sampling methods, oversampling, undersampling, and SMOTE. Comparing Table 4.5 with Table 4.6 shows a significant improvement in recall and the F2 measure can be observed. RF with Undersampling and Ada with oversampling method achieved the highest recall (0.82), followed by LR (0.81). Therefore, it can be shown from the results that the ensembled-based (RF, Ada) machine learning algorithms work well with the top 40 features. Although comparing Table 4.6 with Table 4.2 shows that the recall (reduce 0.02) and F2 measure (reduce 0.012) for RF is slightly reduced with the feature selection method, training time improved with the feature selection method.

Table 4.6 : *Classification results of 40 features (MI) with different sampling methods*

Classifiers	Methods	Acc	TP	TN	FP	FN	Pre	Recall	F2	AUROC
NB	Over	0.74	558	2425	817	232	0.41	0.71	0.62	0.80
	Under	0.74	558	2407	835	232	0.40	0.71	0.61	0.80
	SMOTE	0.74	444	2524	718	346	0.38	0.56	0.51	0.75
LR	Over	0.77	640	2452	790	150	0.45	0.81	0.70	0.85
	Under	0.76	643	2432	810	147	0.44	0.81	0.69	0.85
	SMOTE	0.77	640	2461	781	150	0.45	0.81	0.70	0.85
MLP	Over	0.78	370	2763	479	420	0.44	0.47	0.46	0.75
	Under	0.72	584	2331	911	206	0.39	0.74	0.63	0.79
	SMOTE	0.78	402	2728	514	388	0.44	0.51	0.50	0.75
Ada	**Over**	**0.77**	**644**	**2459**	**783**	**146**	**0.45**	**0.82**	**0.70**	**0.85**
	Under	0.77	643	2467	775	147	0.45	0.81	0.70	0.85
	SMOTE	0.81	512	2739	503	278	0.50	0.65	0.61	0.84
Bagging	Over	0.80	350	2891	351	440	0.50	0.44	0.45	0.79
	Under	0.76	569	2511	731	221	0.44	0.72	0.64	0.81
	SMOTE	0.81	313	2959	283	477	0.53	0.40	0.42	0.80
RF	Over	0.82	387	2922	320	403	0.55	0.49	0.50	0.84
	Under	**0.77**	**646**	**2454**	**788**	**144**	**0.45**	**0.82**	**0.70**	**0.84**
	SMOTE	0.83	347	2983	259	443	0.57	0.44	0.46	0.84

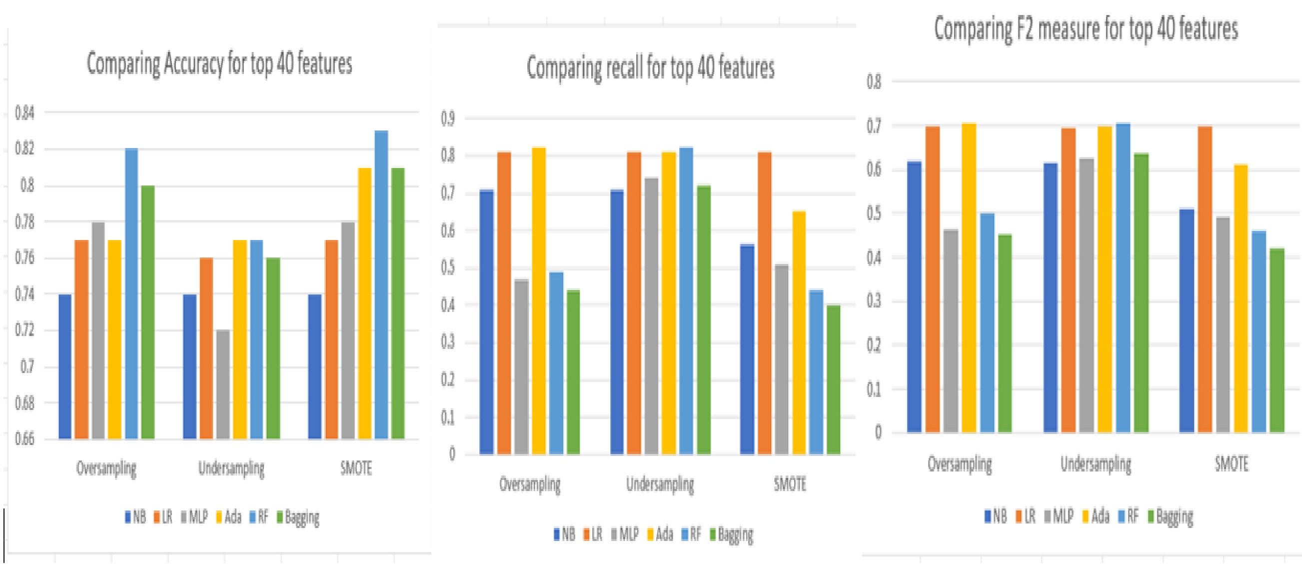

Figure 8 *Comparison result of evaluation metrics for 40 features using MI*

Figure 8 is the graphical representation of various evaluation metrics with different classifiers. The feature selection method is useful to train the classification models as it reduces the dimensionality of the dataset and the training time of the models.

Table 4.7 shows the top 40 features using the random forest feature importance method. Similar to the MI feature selection method sadness and hopelessness is considered the first feature which influences suicidal behaviors. As shown in Table 4.7 electronic bullying and bullying at school ranked fifth and sixth respectively in the random forest feature importance ranking. Therefore, as per the random forest feature importance methods, those three features play important role in influencing suicidal tendencies.

Table 4.7 *Top 40 selected features using random forest feature importance method*

S_NO	Rvalue	Features	F type	S_NO	Rvalue	Features	F type
1	0.140595	Qn25	Sad & hopelessness	21	0.012418	Qn78	Physical activity >= 5 d
2	0.054161	Q7	Weight	22	0.012306	Qn82	Sports participation
3	0.044150	Q6	How tall are you	23	0.012239	Qn73	Carrot eating
4	0.024580	Q1	How old are you	24	0.012025	Q75	No soda drinking
5	0.024461	Qn24	Electronic bullying	25	0.012011	Qn76	No milk drinking
6	0.024206	Qn23	Bullying at school	26	0.011926	Qn68	Weight loss
7	0.022421	Q3	what grade are you	27	0.011860	Qn77	Breakfast eating
8	0.018336	Qn22	dating violence	28	0.011825	Qn58	Ever sexual intercourse
9	0.017732	Qn49	Ever prescription pain medicine use	29	0.011511	Qn87	Asthma
10	0.017355	Qn19	Forced intercourse	30	0.011492	Q72	Potato eating
11	0.015621	Qn20	Sexual violence	31	0.011265	Qn34	Electronic vapor use
12	0.014813	Qn67	Perception of weight	32	0.011221	Qn51	Ever inhalant use
13	0.014375	Q2	What is your sex	33	0.011137	Qn89	Grades in school

Table 4.7 continued

14	0.013361	Qn81	PE attendance	34	0.011096	Qn47	Current marijuana use
15	0.013213	Qn80	Computer use	35	0.011082	Qn41	Current alcohol use
16	0.013143	Qn57	Illegal drugs school	36	0.011052	Qn61	Current sexual activity
17	0.012758	Qn92	Sports drinks	37	0.010951	Qn35	Current e vapor use
18	0.012680	Qn45	Ever marijuana use	38	0.010443	Qn86	Oral health care
19	0.012622	Qn69	Fruit juice drinking	39	0.010438	Qn40	Initiation of alcohol use
20	0.012470	Qn71	Green salad eating	40	0.010311	Qn30	Ever cigarette use

Table 4.8 shows the 40 important features selected by recursive feature elimination methods. Recursive feature elimination (RFE) was used in the study using a logistic regression classifier to select the important features that are more efficient for building the predictive model. It can be shown from Table 4.8 that sad and hopelessness, bullying at school, and electronic bullying are also found within the first fifteen features.

Table 4.8 *Top 40 selected features using recursive feature elimination method*

S_No	**Features**	**F type**	S_No	**Features**	**F type**
1	Q2	What is your sex	21	Qn47	Current marijuana use
2	Q4	Are you Hispanic/Latino	22	Qn48	Ever synthetic marijuana use
3	Qn12	Weapon carrying	23	Qn49	Ever prescription pain medicine use
4	Qn13	Weapon carrying at school	24	Qn51	Ever inhalant use
5	Qn14	Gun carrying past 12 mon	25	Q52	Ever heroin use
6	Qn15	Safety concerns at school	26	Qn54	Ever ecstasy use
7	Qn16	Threatened at school	27	Qn56	Illegal injected drug use

Table 4.8 continued

8	Qn18	Physical fighting at school	28	Qn57	Illegal drugs at school
9	Qn19	Forced sexual intercourse	29	Qn67	Perception of weight
10	Qn22	Physical dating violence	30	Qn69	Fruit juice drinking
11	Qn23	Bullying at school	31	Qn71	Ever inhalant use
12	Qn24	Electronic bullying	32	Q77	How tall are you
13	Qn25	Sad or hopeless	33	Q78	Illegal injected drug use
14	Qn32	Current cigarette use	34	Qn80	Carrot eating
15	Qn34	E vapor product use	35	Qn84	Ever marijuana use
16	Qn35	Current e vapor use	36	Qn85	Current alcohol use
17	Qn37	Current smokeless tobacco use	37	Qn88	Perception of weight
18	Qn38	Current cigar use	38	Qn91	Current sexual activity
19	Qn40	Initiation of alcohol use	39	Qn92	Sleep
20	Qn45	Ever marijuana use	40	Qn93	Are you Hispanic/Latino

Besides the three features mentioned above, several common features can be identified from Table 4.3, Table 4.7 and Table 4.8. Those common features have the same importance to classify suicide attempters from non-suicide attempters. The following Table 4.9 shows the common features obtained from three feature selection methods.

4.2.2 Common features of different feature selection methods

Table 4.9 shows 16 common features that can be obtained using three feature selection methods, mutual information feature selection methods, random forest feature selection methods, and recursive feature elimination method. Although these 16 features were common to the three feature selection methods, their rankings were different for the different feature selection methods. To rank these 16 common features same rank was required for each feature. I used the average ranking method to rank the features from highest to lowest. The process can be described as taking the ranks for particular features for two feature selection methods, MI and random forest features importance, and calculating the average rank for that particular feature. Since ranks are not applicable to the REF method, this method is only used to find the common features and did not participate in the calculation of average rank. As shown in Table 4.9 sad and hopelessness, bullying at school and electronic bullying ranked first, second, and third respectively.

Table 4.9 *An overview of common features*

Serial No	Feature importance	Feature names	Feature_types
1	0.127659	Qn25	Sad and hopelessness
2	0.028006	Qn23	Bullying at school
3	0.025678	Qn24	Electronic bullying
4	0.018905	Qn49	Ever prescription pain medicine use
5	0.015948	Qn22	Physical dating violence
6	0.014933	Q2	What is your sex?
7	0.01457	Qn57	Illegal drugs at school
8	0.014095	Qn35	Current electronic vapor use
9	0.011597	Qn34	Electronic vapor product use
10	0.011128	Qn80	Computer use
11	0.010313	Qn78	Physical activity>= 5 days
12	0.008891	Qn40	Initiation of alcohol use
13	0.008792	Qn45	Ever marijuana use
14	0.008789	Qn67	Perception of weight
15	0.007843	Qn47	Current marijuana use
16	0.0069	Q4	Are you Hispanic or Latino?

Figure 9 is the graphical representation of Table 4.9. As per the figure, the average ranking spread out from 0.12 to nearly 0. The result of common features also shows that almost seven features or 45% of features are related to different types of addictions. Therefore, besides sadness and bullying addiction is one of the important factors influencing suicide tendency. The same influencing features (addiction) can be found by analyzing Table 4.4.

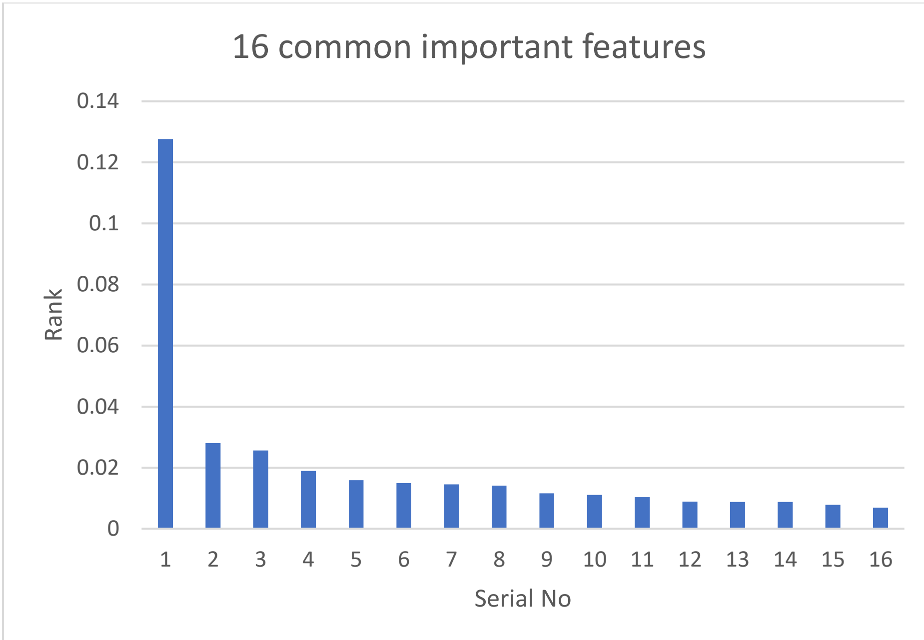

Figure 9 *Common features of three different feature selection methods*

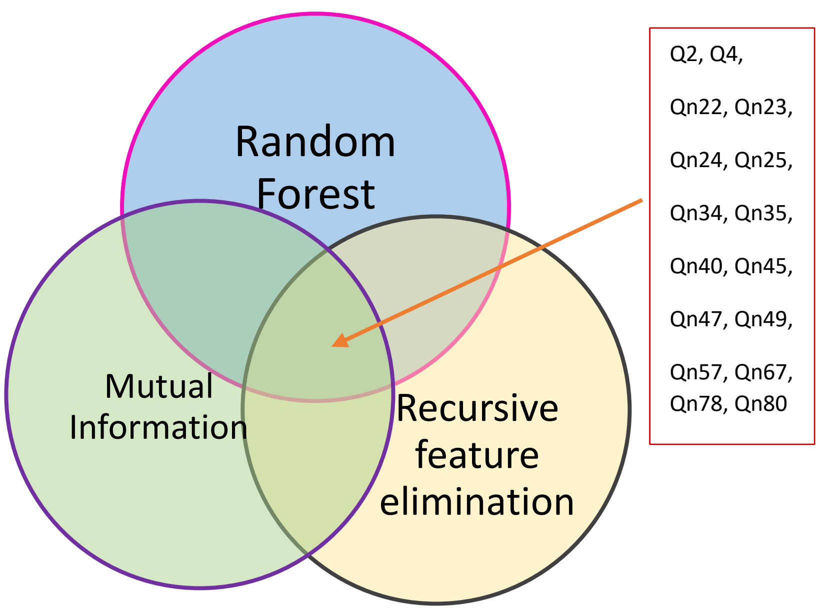

Figure 10 *Venn diagram of 16 common features*

Figure 10 is an extended representation of figure 9. Figure 10 represents a Venn diagram to describe the common features of three feature selection methods. A Venn diagram is used to define the logical relationship between two or more sets. As per figure 10, random forest feature importance, MI, and RFE are three sets of features that are represented as blue, green, and yellow circles respectively. Common features are represented as the common area of the three circles where the arrow is pointed. The text box represents the names of the 16 common features.

Comparing table 4.6 with table 4.1 and table 4.4 we found selecting 16 common features improved the recall ratio of all classifiers other than logistic regression. The improvement range varies from 0.09 to 0.01 which is slightly better than the top 40 feature selection method.

As shown in Table 4.10, LR with undersampling methods and Ada with undersampling methods with 16 common features achieved the highest recall (0.82) and F2 measure (0.704). It can be shown from the confusion matrix that although both classifiers achieved the highest recall and F2 measures LR predicted more true positive instances (305) than Ada (184). On contrary, LR predicted less number of false negative instances (486) with respect to Ada (607). As I discussed earlier the study aims to minimize false negative instances and maximize true positive instances, in which case LR performs better than Ada. Also, comparing LR with oversampling methods and undersampling it can be shown that there is no significant difference between recall and F2 measure for these two sampling methods. The undersampling method of LR achieved a recall ratio of 0.82 and an F2 measure of 0.70 whereas the oversampling method of LR achieved a recall ratio of 0.81 and F2-measures of 0.70. However, analyzing the confusion matrix I found oversampling methods predicted 643 instances as true positives and 147 instances as false negatives whereas undersampling methods predicted 305 instances as true positives and 486 instances as false negatives. These results show that LR with oversampling method is better than LR with undersampling methods with 16 common features.

Table 4.10 *Prediction results of 16 common features with sampling methods*

Classifier	Sampling	Acc	TP	TN	FP	FN	Pre	Recall	F2	AUROC
NB	No sam	0.82	448	2840	405	342	0.53	0.57	0.56	0.83
	Over	0.76	576	2492	750	214	0.43	0.73	0.64	0.83
	Under	0.76	577	2493	749	213	0.44	0.73	0.64	0.83
	SMOTE	0.77	551	2567	675	239	0.45	0.70	0.63	0.82
LR	No sam	0.83	328	3037	205	462	0.62	0.42	0.45	0.85
	Over	0.77	643	2451	791	147	0.45	0.81	0.70	0.85
	Under	**0.77**	**645**	**2441**	**801**	**145**	**0.45**	**0.82**	**0.70**	**0.85**
	SMOTE	0.77	643	2447	795	147	0.45	0.81	0.70	0.85
MLP	No sam	0.83	236	3111	131	554	0.64	0.30	0.33	0.84
	Over	0.77	453	2635	607	337	0.43	0.57	0.53	0.72
	Under	0.71	617	2249	993	173	0.38	0.78	0.64	0.79
	SMOTE	0.78	406	2743	499	384	0.45	0.51	0.50	0.71
Ada	No sam	0.83	330	3033	209	460	0.61	0.42	0.45	0.85
	Over	0.77	641	2459	783	149	0.45	0.81	0.70	0.85
	Under	**0.77**	**645**	**2457**	**785**	**145**	**0.45**	**0.82**	**0.70**	**0.85**
	SMOTE	0.79	598	2596	646	192	0.48	0.76	0.70	0.83

Table 4.10 continued

RF	No sam	0.82	330	2964	278	460	0.54	0.42	0.44	0.81
	Over	0.79	456	2739	503	334	0.48	0.58	0.56	0.79
	Under	0.75	613	2405	837	177	0.42	0.78	0.66	0.82
	SMOTE	0.81	397	2868	374	393	0.51	0.50	0.50	0.81
Bagging	No sam	0.81	309	2945	297	481	0.51	0.39	0.41	0.79
	Over	0.78	409	2752	490	381	0.45	0.52	0.50	0.77
	Under	0.74	570	2432	810	220	0.41	0.72	0.62	0.80
	SMOTE	0.81	362	2891	351	428	0.51	0.46	0.47	0.79

Comparing feature selection with 40 features (Table 4.6) with 16 features it can be shown that LR with oversampling methods is able to predict 3 more instances as true positives and 3 less instances as false negatives. Therefore, 16 features have more potential to predict suicidal tendencies among youths with the LR classifier. Besides, logistic regression all other classifiers with 16 features showed significant improvement in predicting true positive instances with compare to 40 features. Therefore, those 16 common features are important to build predictive models. Figure 11 is the graphical representation of accuracy, recall, and F2 score metrics with different numbers of features for oversampling methods.

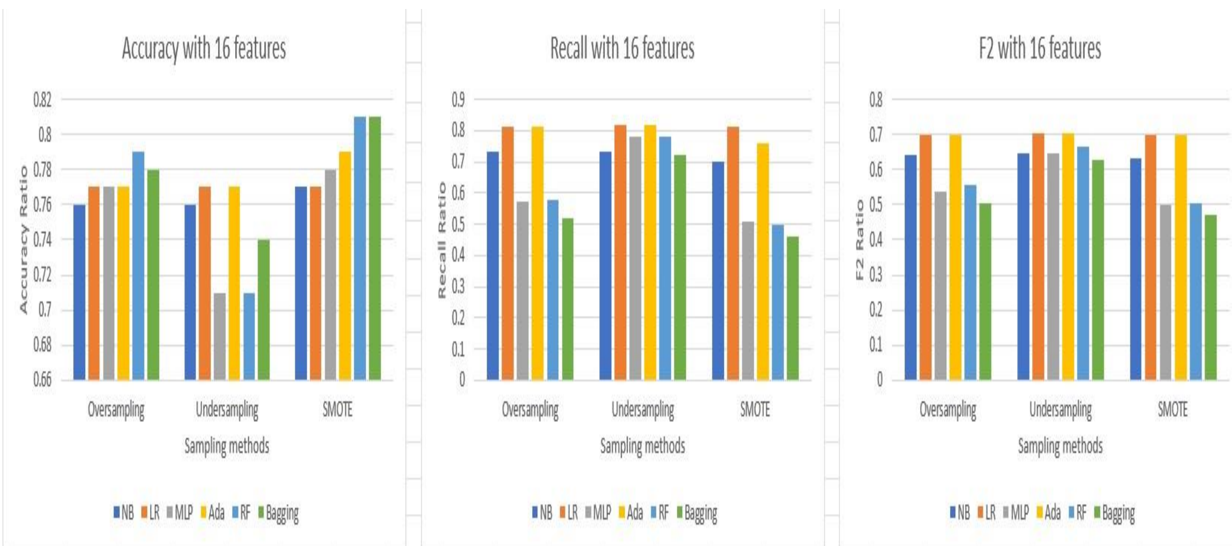

Figure 11 *Comparison of different metrics with 16 features*

As per figure 11, All classifiers improve the accuracy with oversampling and SMOTE methods. However, only the LR and Ada classifiers achieved the highest recall ratio (0.82), F2 measure (0.70), and AUROC (0.85) with the undersampling method. Therefore, the undersampling method works well with 16 features. The recall ratio of the Ada classifier also improved for 16 features with respect to 40 features with the undersampling method. Therefore, the 16 common features I found using three different feature selection methods, are effective for predicting suicidal tendency.

Chapter V – Further Discussion

After analyzing the classification results for various sampling and feature selection methods, I found that LR with 16 features and the undersampling method performs better for predicting suicidal risk than all other classifiers. Overall, all classifiers with 16 features perform well than all features and 40 features with the undersampling method. The dataset that I used for my study is an imbalanced dataset therefore besides accuracy metrics, I need to consider a recall, F2 measure, and confusion matrix to get a better understanding of the classification report. Although LR and Ada classifiers did not achieve the highest accuracy with 16 features and undersampling methods they were able to predict the highest true positives instances (645) and the least number of false negative instances (145) (Table 4.10) with undesampling method. They also achieved a recall ratio of 0.82 which means both classifiers able to predict 82 students out of 100 students who have suicide risk. Therefore, both classifiers are important for predicting suicide risk. RF classifier with 40 features and undersampling methods outperforms all other classifiers by predicting the highest true positive instances (646) and least false negative instances (144) (Table 4.6). The adaBoost classifier also works well with oversampling techniques with 40 features. It was able to predict 644 instances as true positive instances and 146 instances as false negative instances (Table 4.6). According to Table 4.10 and Figure 9, there is no great improvement in prediction results compared to 40 features and 16 features because reducing many features from a dataset sometimes discards some important features as well. Since I found almost the same result with 40 and 16 features, it can be concluded that I am able to reduce the redundant features and the training time to run the models. I also compared my results with some existing works conducted for suicide risk prediction. Although, my dataset is a survey dataset and to my knowledge, no previous work was done using this dataset but some similar suicide risk predictions with other datasets were found while analyzing other works. Table 1.1 presents a summary of some similar suicide risk prediction studies with different datasets. Most of the researchers found significant effects of RF classifiers in predicting suicide risk. A few studies also found LR classifier is a useful tool for predicting suicidal nature. In my work, in addition to LR and RF, I found AdaBoost can predict suicide risk efficiently. As per the summary of Table 1.1 most of the study achieved 90% accuracy. I achieved lower accuracy with sampling methods due to the imbalanced nature of the dataset. Therefore, I focused on other metrics such as recall, F2 measures, and AUROC. The work done by the research group of Frank[20] (Table 1.1) was conducted on the imbalanced dataset and they achieved an AUROC ratio between 0.744–0.755. However, in my study, I achieved an AUROC ratio of more than 0.80. Therefore, my data handling methods are better than contemporary data handling methods. Besides finding the best classifiers in my study I also emphasize finding important features which influence suicidal tendencies. I found besides depression, interactions with peers and addiction to drugs and alcohol are two top features that influence suicidal tendencies. These new findings through my research make my work unique, as contemporary works (Table 1.1) did not find any effect of peer interaction and alcohol, and drug use as potential factors of the suicide attempt. All other studies showed that depression is the main cause of suicide irrespective of age group.

5.1 Limitation of the study

There are some limitations of the current work. First, the survey was conducted on school students and focused on a specific age group. Therefore, the dataset does not have any reference point to compare suicide risk with other age groups. Second, the study was conducted only on the six traditional machine learning algorithms, and the recall and F2-measure could not improve by more than 0.82 and 0.70 respectively. Additional analyses using deep learning methods may be applied to improve the performance of predictive models. Third, in this study, we used three feature selection methods that did not improve the classification result significantly. Others feature selection methods can be used to improve the performance of the classifiers. Although the performance of the machine learning models is satisfactory, I believe the performance will be improved by using state-of-art machine learning algorithms.

5.2 Future Work

In the future, I would like to conduct the study with some other dataset using the same data handling techniques that I used in this work. My plan is to use the datasets that were used for the studies described in Table 1.1. This way, I can compare their results with the results I will achieve and verify that my data handling techniques can be used on any dataset to improve classification results. I would also like to use deep learning to improve the overall classification result. In addition, I would like to explore the clinical variables to understand the importance of feature selection methods as well as to identify possible clinical variables that could be added to my study to classify the suicidal risk more precisely.

Chapter VI – Conclusions

In conclusion, six predictive machine learning algorithms with data handling techniques have been used for predicting suicide risk among high school students. The single model LR and the ensemble-based models Ada Boost and RF achieved the highest recall (0.82-0.84), F2 measure(0.72-0.70), and AUROC(0.85) with different data sampling methods. Therefore, all those classifiers with certain conditions can be used to classify suicide attempters with non-suicide attempters. All other classifiers show significant improvement in classification results with oversampling and undersampling methods compare to no sampling method. The highest recall, F2 measure, and AUROC can be achieved using the undersampling method. Therefore, undersampling methods play a crucial role to improve the prediction results. Feature selection also has some contribution to classification reports. Selecting the top 16 common features using three different feature selection methods gives better results than the top 40 features for recall, F2 measure, and AUROC. Although the improvement is not very significant, feature selection helped to eliminate redundant data from the dataset. Also, fewer features reduce the training time of the models. Therefore, the machine learning algorithms selected in this comparative study, combined with feature selection and undersampling methods, can provide a powerful supplementary tool to identify approximately 82 percent of adolescents with a high suicide tendency based on the survey of their behaviors, which can greatly facilitate the clinical diagnosis and early prevention.

References

1. World Health Organization. Figures and facts about suicide. No. WHO/MNH/MBD/99.1. World Health Organization, 1999.
2. Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. Expert Systems with Applications, 41(14), 6371-6385.
3. K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, S. Pranavanand, Heart disease risk prediction using machine learning classifiers with attribute evaluators, Appl. Sci., 11 (2021), 8352. https://doi.org/10.3390/app11188352
4. Yaakov Ophir, Refael Tikochinski, Christa S. C. Asterhan, Itay Sisso & Roi Reichart," Deep neural networks detect suicide risk from textual facebook posts", October 07, 2020, Nature, Scientific report.
5. Velupillai S, Hadlaczky G, Baca-Garcia E, Gorrell GM, Werbeloff N, Nguyen D, Patel R, Leightley D, Downs J, Hotopf M, Dutta R. Risk Assessment Tools and Data-Driven Approaches for Predicting and Preventing Suicidal Behavior. Front Psychiatry. 2019 Feb 13;10:36. doi: 10.3389/fpsyt.2019.00036. PMID: 30814958; PMCID: PMC6381841.
6. Carter G, Milner A, McGill K, Pirkis J, Kapur N, Spittal MJ. Predicting suicidal behaviours using clinical instruments: systematic review and meta-analysis of positive predictive values for risk scales. Br J Psychiatry (2017) 210:387–95. doi: 10.1192/bjp.bp.116.182717
7. Walsh CG, Ribeiro JD and Franklin JC. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. J Child Psychol Psychiatry 2018; 59(12): 1261–1270.
8. Belsher BE, Smolenski DJ, Pruitt LD, et al. Prediction models for suicide attempts and deaths: a systematic review and simulation. JAMA Psychiatry 2019; 76(6): 642.
9. Hettige NC, Nguyen TB, Yuan C, et al. Classification of suicide attempters in schizophrenia using socio-cultural and clinical features: a machine learning approach. Gen Hosp Psychiatry 2017;
10. David Delgado-Gomez, Hilario Blasco-Fontecilla, Federico Sukno, Maria Socorro Ramos-Plasencia, Enrique Baca-Garcia, Suicide attempters classification: Toward predictive models of suicidal behavior, Neurocomputing,Volume 92, 2012, Pages 3-8, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2011.08.033.
11. Macalli, M., Navarro, M., Orri, M. et al. A machine learning approach for predicting suicidal thoughts and behaviours among college students. Sci Rep 11, 11363 (2021). https://doi.org/10.1038/s41598-021-90728-z
12. Bae S-M, Lee S-H and Lee SA. Prediction by data mining, of suicide attempts in Korean adolescents: a national study. Neuropsychiatr Dis Treat 2015; 11: 2367–2375.
13. Weller O, Sagers L, Hanson C, Barnes M, Snell Q, Tass ES (2021) Predicting suicidal thoughts and behavior among adolescents using the risk and protective factor framework: A large scale machine learning approach. PLoS ONE 16(11): e0258535. https://doi.org/10.1371/journal. pone.0258535
14. Miché M, Studerus E, Meyer AH, Gloster AT, Beesdo-Baum K, Wittchen HU, Lieb R. Prospective prediction of suicide attempts in community adolescents and young adults,

using regression methods and machine learning. J Affect Disord. 2020 Mar 15;265:570-578. doi: 10.1016/j.jad.2019.11.093. Epub 2019 Nov 13. PMID: 31786028.
15. Predicting Suicide Attempts in a Community Sample of Adolescents Using Machine Learning Algorithms
16. Edgcomb, Juliet Beni, Trevor Shaddox, Gerhard Hellemann, and John O. Brooks III. "Predicting suicidal behavior and self-harm after general hospitalization of adults with serious mental illness." Journal of psychiatric research 136 (2021): 515-521.
17. Nordin, Noratikah, Zurinahni Zainol, Mohd Halim Mohd Noor, and Chan Lai Fong. "A comparative study of machine learning techniques for suicide attempts predictive model." Health informatics journal 27, no. 1 (2021): 1460458221989395.
18. Amini, Payam, Hasan Ahmadinia, Jalal Poorolajal, and Mohammad Moqaddasi Amiri. "Evaluating the high risk groups for suicide: a comparison of logistic regression, support vector machine, decision tree and artificial neural network." Iranian journal of public health 45, no. 9 (2016): 1179.
19. Su, Chang, Robert Aseltine, Riddhi Doshi, Kun Chen, Steven C. Rogers, and Fei Wang. "Machine learning for suicide risk prediction in children and adolescents with electronic health records." Translational psychiatry 10, no. 1 (2020): 1-10.
20. Iorfino, Frank, Nicholas Ho, Joanne S. Carpenter, Shane P. Cross, Tracey A. Davenport, Daniel F. Hermens, Hannah Yee et al. "Predicting self-harm within six months after initial presentation to youth mental health services: A machine learning study." PLoS one 15, no. 12 (2020): e0243467.
21. Colic, Sinisa, J. D. Richardson, James P. Reilly, and Gary M. Hasey. "Using machine learning algorithms to enhance the management of suicide ideation." In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4936-4939. IEEE, 2018.
22. YRBSS Data & Documentation (https://www.cdc.gov/healthyyouth/data/yrbs/data.htm)
23. Le Cessie, Saskia, and Johannes C. Van Houwelingen. "Ridge estimators in logistic regression." Journal of the Royal Statistical Society: Series C (Applied Statistics) 41, no. 1 (1992): 191-201.
24. Platt, John. "Sequential minimal optimization: A fast algorithm for training support vector machines." (1998).
25. Verleysen, M., & François, D. (2005). The curse of dimensionality in data mining and time series prediction. In Computational Intelligence and Bioinspired Systems: 8th International Work-Conference on Artificial Neural Networks, IWANN 2005, Vilanova i la Geltrú, Barcelona, Spain, June 8-10, 2005. Proceedings 8 (pp. 758-770). Springer Berlin Heidelberg.
26. Prati, R. C., Batista, G. E., & Monard, M. C. (2009, December). Data mining with imbalanced class distributions: concepts and methods. In IICAI (pp. 359-376).
27. Anxiety, Depression, and Suicide in Youth, https://doi.org/10.1176/appi.ajp.2020.21020186
28. Risk and Resiliency Factors Influencing Suicidality Among Urban African American and Latino Youth, https://doi.org/10.1023/B:AJCP.0000014317.20704.0b

29. Suicide in New Zealand II: a review of risk factors and prevention, Beautrais, Annette. "Suicide in New Zealand II: A review of risk factors and prevention." NZ Med J 116.1175 (2003): U461.
30. School Risk and Protective Factors of Suicide: A Cultural Model of Suicide Risk and Protective Factors in Schools, Marraccini, Marisa E., et al. "School risk and protective factors of suicide: A cultural model of suicide risk and protective factors in schools." School Psychology Review (2020): 1-24.
31. Four Risk Factors Associated with Adolescent Depression: Discrimination, Academic Stress, Migrant Status and Family Bullying, Liu, Yutong, et al. "Four Risk Factors Associated with Adolescent Depression: Discrimination, Academic Stress, Migrant Status and Family Bullying." 2022 5th International Conference on Humanities Education and Social Sciences (ICHESS 2022). Atlantis Press, 2022.'
32. Risk Factors for Suicidal Behaviors in American Indian and Alaska Native Peoples: A Systematic Review, Fetter, Anna Kawennison, et al. "Risk Factors for Suicidal Behaviors in American Indian and Alaska Native Peoples: A Systematic Review." Clinical Psychological Science (2022): 21677026221126732.
33. Kalin, N. H. (2021). Anxiety, Depression, and Suicide in Youth. American Journal of Psychiatry, 178(4), 275–279. https://doi.org/10.1176/appi.ajp.2020.21020186
34. Haque, R., Islam, N., Islam, M., & Ahsan, M. M. (2022). A comparative analysis on suicidal ideation detection using NLP, machine, and deep learning. Technologies, 10(3), 57.
35. Huang, Y., Zhu, C., Feng, Y., Ji, Y., Song, J., Wang, K., & Yu, F. (2022). Comparison of three machine learning models to predict suicidal ideation and depression among Chinese adolescents: A cross-sectional study. Journal of Affective Disorders, 319, 221-228.
36. Shen, Y., Zhang, W., Chan, B. S. M., Zhang, Y., Meng, F., Kennon, E. A., ... & Zhang, X. (2020). Detecting risk of suicide attempts among Chinese medical college students using a machine learning algorithm. Journal of affective disorders, 273, 18-23.