Predicting Suicide Risk in U.S. High School Students Using Machine Learning[1]

Xianhui Fu, Ivy Liu, Kathy Liu, Ziren Zhou

MSDS 422: Practical Machine Learning

March 16, 2025

[1] Grammar checked by Grammarly and ChatGPT

Executive Summary

Teen suicide is a major health issue globally as well as in the United States, where an increasing number of teenagers have been ending their own lives in recent times. Teen suicide poses grave concerns for society at large, schools as an institution, and for families individually. This study attempts to foresee suicidal ideation among secondary school students based on data from the 2023 Youth Risk Behavior Surveillance System (YRBSS). YRBSS is an annual national survey of health behavior in young people and could be attributed to mental health issues, injuries, or mortality. Using machine learning techniques such as logistic regression, random forests, and other advanced methods, this study aims to determine significant factors influencing suicide risk in teenagers and develop prediction models.

The current literature clearly outlines a strong relationship between mental disorders, specifically anxiety and depression, and suicidal ideation and suicidal behavior in adolescents. In addition to those, other significant issues in adolescents' mental health are drug use, lack of sleep, exposure to violence, and safety concerns. Traditional methods of data analysis have the difficulty of depicting the link between such issues and their influence on one another. Machine learning algorithms can process complex data, detect patterns and relationships between them and make more accurate predictions. Past studies have already demonstrated machine learning algorithms' capability in predicting suicidal risk in adolescents and this study reinforces such results emphatically.

This study follows from previous work by considering issues of mental health, histories of experiencing violence, and life patterns of behavior. This will enable us to develop an improved approach for measuring suicide risk. Through careful examination of the new 2023 YRBSS data, we aim to enhance the prediction of risks and provide valuable information to

schools, families, and public health agencies. Outcomes of this project will enable teachers, parents, and health professionals to develop improved prevention plans, identify at-risk individuals, respond adequately, and reduce suicides in adolescents.

Problem Statement/Research objective

The objective of this research is: How could we realistically predict the suicidal risk in American high school teenagers? The key issue has two dimensions.

The first is establishing what the most probable factors are to foretell suicide risk in adolescents. The research in this study will examine a range of variables that contribute to adolescents' mental distress, including depression, anxiety, sleep disturbance, drug abuse, violence exposure, and lifestyle. The study will try to determine how these determinants relate to one another and how much they correspond to adolescent suicide risk.

The second problem is: How effective are machine learning methods in modeling suicide risk estimation? Traditional statistical analysis methods, i.e., regression analysis, possess some advantages in exploring causal effects. However, they may be tested by their capability to detect nonlinear relationships between variables when faced with high-dimensional and complex behavior data. Machine learning methods, on the other hand, show stronger capabilities in pattern discovery, feature extraction, and prediction. This study will compare the predictive accuracy of such strategies to risk for adolescent suicide and identify which of them is most predictive.

So, why should we care about this issue? Teen mental health issues, particularly suicide risk, are a significant international public health concern. Suicide behaviors—suicidal ideation, suicide planning, and suicide attempts—are on the rise among high school students. Therefore, understanding and predicting adolescent suicide risk is not only crucial for scholarly research but also has important implications for social policy, education systems, and healthcare systems.

Suicide is preventable, but more effective predictive tools need to be created. Suicide is most commonly linked to mental health crises compared to other diseases. However, common screening tools often rely on self-reporting or assessment by healthcare providers, which may have certain demerits. For instance, young people might conceal their psychological state, or due to budget constraints, schools and families might not notice warning signs in time. Therefore, employing machine learning techniques to screen large-scale data can identify hidden patterns of suicide risk, which can provide an added scientific justification for early intervention.

Besides, traditional studies can focus on single variables or straightforward linear models, whereas machine learning algorithms are capable of dealing with high-dimensional data and finding complex interactions between variables. Depressive symptoms, sleep disturbances, and school violence, for example, might play varying roles under various conditions. With the use of machine learning, we are able to determine the most relevant predicting variables and build a model that is more capable of differentiating adolescents' risk levels.

The findings of this study can also be of immense value to policymakers, educators, and health professionals. Parents, educators, policymakers, and mental health professionals all desire more scientific methods of identifying and treating adolescents with mental illness. For example, if the study discovers that specific behaviors—like chronic sleep deprivation or victimization from school violence—are strongly associated with suicide risk, schools can implement more preventive interventions targeting these behaviors. These interventions may be in the form of mental health education programs, increased campus safety programs, and increased access to psychological counseling services. In addition, governments can use these findings to optimize resource use and increase the availability of mental health services, thereby reducing adolescent suicide rates.

Briefly, this study tries to predict adolescent suicide risk using the latest 2023 YRBSS data and cutting-edge machine learning methods to build a highly efficient and accurate risk assessment system. This study not only presents dominant risk indicators for adolescent suicide but also builds a scientific foundation for public health, education, and families. Through this research, we seek to reduce the incidence of teenage suicide, raise public knowledge and awareness of teenage mental illness, and provide evidence-driven assistance for future mental health policy. In the long term, our hope is to establish a healthier and safer environment for development whereby every teenager receives what they require and deserves.

Literature Review

Suicide in adolescents represents a public health crisis in every corner of the world. In America, suicide had become one of three top causes of death in people between the ages of 15 and 24 (Ghadipasha et al. 2024, 2). In the past few years, the rate of suicide in adolescents has been rising due to changes in social, financial, psychological, and environmental factors, a process also hastened by the COVID-19 outbreak (Wang and Liu 2024, 5). Based on present evidence, suicidal ideation (SI) and suicide attempts (SA) are the strong predictors of subsequent suicide risk (Ghadipasha et al. 2024). Therefore, identification of most significant risk factors in such inclinations becomes imperative in order to frame appropriate intervention policies.

Suicidal thoughts in American adolescents have increased in the past few years with very high risks reported in females, adolescents who are LGBTQ+, and in specific ethnicities (Ostanin et al.,3). Based on 2021 statistics, 30.0% of female adolescents had had suicidal ideation and 13.3% had attempted suicide and this is an increase from 2019 (Gaylor et al. 2023, 6). Suicide ideation, suicide plans, and suicide attempts are all greater in female adolescents compared to males but suicide completion remains higher in males. Compounding this, suicide attempts by Black and Hispanic female adolescents are significantly higher compared to others and have the highest rate of trying at 14.5% (Gaylor et al. 2023, 28). Suicide risk in LGBTQ+ youth remains significantly higher compared to heterosexual adolescents because of discrimination in society, identification issues, and lack of family support (Gaylor et al. 2023, 28).

Geography also plays an important role in explaining suicidal behavior in adolescents. Rural adolescents have higher suicide rates compared to urban adolescents due to poor support systems in society, lack of access to health services in the field of psychiatry, and access to firearms (Ghadipasha et al. 2024, 20). Suicide rates are also higher in mountainous and coastal districts and can be attributed to differences in culture and regional development as well as access to services in the field of psychiatry (Ghadipasha et al. 2024, 25). Suicide rates also vary significantly across different states and are as follows: Alaska had 56 per 100,000 compared to only 7 per 100,000 in New Jersey (Cambron et al. 2023, 4). This suggests suicidal risk as much as a regional as an individual function.

Beyond demographic and geographic factors, suicidal behaviors in adolescents are also significantly impacted by socioeconomic status and mental health disorders.

Mental disorders are the most immediate risk factor for suicide in adolescents. Some of the most potent suicide risk indicators include depression, anxiety, sleep disturbance, and social isolation (Wang and Liu 2024, 12). Most studies concur with this view and add that the most potent indicator of suicidal thoughts is hopelessness. Depressive and anxiety-disordered adolescents have significantly higher suicide risks (Cambron et al. 2023, 16).

Another prime psychological risk factor involves drug abuse. Literature estimates that adolescents who consume alcohol, marijuana, and other illicit drugs are at an increased risk of suicidal behavior and suicidal ideation, particularly those who consume them in the long term and/or in excessive amounts (Wang and Liu 2024, 18).

SES plays a direct role in suicidal behavior and psychological health in adolescents. Suicidal ideation and attempts are most likely in low-SES adolescents who live in economically poor neighborhoods and have scarce school resources. In all the regions in study, low-SES adolescents have suicide-related behavior and depression and anxiety, and adolescents with greater SES have low suicide attempts and ideation (Ghadipasha et al. 2024, 14).

Machine technology increasingly has been employed in the field of medical and mental health in suicide risk prediction notably with greater precision and adaptability in respect to conventional methodology (Wang and Liu 2024, 5).

Traditional methods are most likely based on preconceptions of the researchers and as such can only test a few variables and are unable to account for complex interactions (Canizares et al. 2025, 7). In addition to this, suicidal behavior in adolescents results from an array of psychological, social, environmental, and physiological variables and the interactions between them are usually nonlinear and as such traditional methods are unable to model them suitably.

Also due to the relatively low rate of suicide attempters (around 7-10%) in conventional methods, standard methods have class imbalance issues leading to biased models towards classifying non-suicidal individuals and consequently making the model less sensitive (Canizares et al. 2025, 12).

Machine learning overcomes such limitations by automatically discovering patterns and features in the data and improving the power of prediction. In Wang and Liu (2024,15), LightGBM, Random Forest, and Logistic Regression were implemented in suicide risk prediction and made LightGBM the most accurate predictor. LightGBM works best with big data and complex variable interactions and achieves very impressive generalization performance on different data sets. LightGBM beats traditional statistical methods in modeling nonlinear relationships in data.

Along with this, it was established by a study by Canizares et al. (2025,18) that LASSO Logistic Regression performed best in predicting at-risk adolescents with 88.9% sensitivity and an AUC of 94.6%, and far better when compared to conventional methods. On the basis of all such studies, machine-learning algorithms predicted suicide risk better upon which there can be an appropriate reason for early intervention.

Teen suicide results from a variety of factors including socioeconomic status, residential geographic location, gender differences and race differences. With increasingly better data examination and policy intervention fortifying in addition to better predictive models, suicide hazards in teens might be best regulated and made healthier and safer for adolescents' growth and development.

Data Overview

Introduction to the YRBSS and YRBS

The dataset comes from the 2023 national Youth Risk Behavior Survey (YRBS) sourced from the Youth Risk Behavior Surveillance System (YRBSS) conducted by the Centers of Disease Control and Prevention (CDC). It is introduced in the 2023 YRBS Data User's Guide that "[t]he YRBSS was developed in 1990 to monitor health risk behaviors and experiences that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States". Since 1991, the YRBSS has collected data from approximately

5 million high school students through 2300 separate surveys every two years, usually during spring semester which limits seasonal influences on the respondents like seasonal affective disorder (SAD) (NIMH 2023). The survey collects responses on a broad aspect of health-related topics that enable the assessment of trends, risk behavior co-occurrences, and public health interventions, through a mix of binary, categorical, and ordinal questions that constantly updates since the beginning of the YRBSS to make sure the survey questions are up-to-date and able to monitor any new behaviors. For example, questions about vaping (using of electronic vapor products was first introduced in the 2015 national YRBS survey (CDC 2016, 29). Alongside with the student demographics (sex, sexual identity, race and ethnicity, and grade), a total of 107 survey questions includes youth health behaviors and conditions, substance use behaviors, and student experiences (CDC 2024).

Sampling Methodology and Representativeness of the 2023 national YRBS

To ensure the representativeness of the sample of 9th through 12th grade students in the United States, a three-stage cluster sample design is applied in the sampling process. This involves randomly choosing groups of counties through primary sampling units (PSUs), randomly selecting schools within the selected PSUs, including public and private schools, and randomly selecting classes of students in the selected schools (CDC 2024). The response is anonymous and voluntary with weighting procedures applied to adjust for nonresponse and sampling design, making the dataset generalizable to the national high school population of the United States.

Structure of the Survey and Data

The 2023 national YRBS dataset consists of 20103 observations and 117 variables, in which 105 are encoded or transformed response of the multiple choice question responses, each named with its question number following a letter ‘q’. Apart from two free text entry variables recording the original response to respondent's height in meters (‘q6orig’) and weight in kilograms (‘q7orig’) that are also used to calculate the Bosy Mass Index (BMI) in the formula BMI = Weight/Height^2 (stored as ‘BMIPCT’), most of the survey questions require respondents to select from categorized choices defining the extent of the described behavior or situation in the question. ‘record’ serves as the unique identifier of each observation in the sample dataset. A variable that originally stores the regional location information of the respondent’s school is masked from this publicly available national sample dataset this projects uses. A ‘raceeth’ variable is created to represent the combined response to ‘q4’ which asks about the respondent’s ethnicity (Hispanic or Latino) and ‘q5’ which asks about the respondent’s race and allows choosing more than one response. A sample question is as follows.

> Q10. During the past 30 days, how many times did you drive a car or other vehicle when you had been drinking alcohol?
>
> A. I did not drive a car or other vehicle during the past 30 days
> B. I drove a car or other vehicle, but not when I had been drinking alcohol
> C. 1 time
> D. 2 or 3 times
> E. 4 or 5 times
> F. 6 or more times (CDC 2024, 23)

Preliminary data edits were performed by the CDC before this dataset was published. An important approach related this project is the logical consistency edits that changes the logically conflicting responses from two questions to missing as it maintains the integrity and validity of the observations (CDC 2024, 9). More on data edits and the detailed code book can be found in the data User’s Guide for 2023 national YRBS.

Limitations and Considerations

The national YRBS dataset is one of the most comprehensive sources of adolescent health risk behavior data in the U.S., providing since 1991 valuable insights for public health researchers to identify high-risk groups and track behavioral trends over time. Despite being a robust and widely used dataset, several limitations must still be considered. The survey relies on student self-reports, which means the observations may be subject to recall bias or social desirability bias. Each student's perception on the extent levels described in some of the questions may not be exactly the same, as some text-based level descriptions are not quantifiable. Some proportions of missing data from either nonresponse or edits are large. Such nonresponse cases may also be resulted from the self-report survey design which allows students to hide their response on specific questions they wish not to answer due to factors such as pressures from being bullied at school as there is no guarantee for a completely individual and private environment to take the survey. A total of 9 questions are missing out over 30% of the values, requiring careful imputation or exclusion decisions.

There are no real causal relationships that can be inferred from the dataset as the YRBSS is not a longitudinal study, meaning it does not track the same students over time. What can be studied are the correlations between and co-occurrence of risk behaviors. Along with this cross-sectional design, the survey modifications over the years are also limiting the factors available for long-term trend analysis.

Data Pre-Processing

Before feeding the data to the models, several steps were taken to prepare the data while still preserving the possible important patterns the data implies. The following approaches ensures that the model assumptions on the data are accounted for. A round of manual transformation of the records was performed first to reflect some of the true records of the responses if possible. Binary question (Yes/No questions) response records are transformed from 1=Yes and 2=No to 1=Yes and 0=No for easier coding logic. Records with response that states the respondent does not understand the question will be set to missing. Ordinal question responses like question 9 below were converted to the true levels or the average of the level range described in each option. The number in front of the question is the original record and the number after is the values they have been transformed to, defined manually by question.

> Q9. During the past 30 days, how many times did you ride in a car or other vehicle driven by someone who had been drinking alcohol?
>
> 1 – A. 0 times - transformed to 0
> 2 – B. 1 time - transformed to 1
> 3 – C. 2 or 3 times – transformed to 2.5
> 4 – D. 4 or 5 times - transformed to 4.5
> 5 – E. 6 or more times - transformed to 6 (CDC 2024, 23)

Data Cleaning

The first phase of data cleaning removes those variables that are either unrelated to the project or are essentially representing the same factor. Columns ‘q88’ to ‘q107’ were dropped because of them not being part of the standard national YRBS questions. They consist of excessive missing values since not all sample schools conducted the survey with these extra questions, making these questions less representative for the sample population. The original record columns ‘q6orig’ and ‘q7orig’ were dropped since the derived columns ‘q6’ and ‘q7’ already track the information contained within. The ‘site’ and ‘orig_rec’ columns were also dropped as the national dataset masked the two identifiers. Even with the site record revealed, the dataset is still not properly sampled for regional studies as the sampling methodology is designed to optimize the representativeness of the dataset at the national level rather than by region.

Column Renaming

For better interpretability of the features identified in the exploratory data analysis (EDA) and model outputs, all columns are renamed to a format that shortly describes the behavior that was asked in the question while keeping the question number in front to ensure easier referencing back to the original survey questions. A binary question asking if the respondent has ever been bullied at school in the past 12 months (‘q24’) was renamed to ‘q24_been_bullied’. The combined variables of sequential questions were renamed to indicate the range of questions contained within. For example, question 10 shown in the previous section asking for the number of times during the past 30 days the respondent drove when they had been drinking is combined with question 11 to form a new binary indicator of unsafe driving behaviors, renamed from their original name ‘q10’ and ‘q11’ to ‘q10_11_unsafe_driving”

Feature Engineering

Missing Values Imputation

As mentioned above, large amounts of missing values exist in the dataset. Without processing, they may introduce bias to the models and therefore reduce the predictive power. The missing values in this dataset were imputed using a Decision Tree Regressor, which is capable of capturing the non-linear relationships between features while preserving the underlying patterns in the behaviors. A training subset consisting of the non-missing observations was used to train a Decision Tree Regressor to predict for the test subset which contains all the missing values. The predicted outputs were then filled into the missing values. This way, consistency with the observed behavioral trends is maintained and is much less prone to data distortion by other simpler approaches like imputing with mean/median/mode.

Feature Creation and Binarization

As some of the questions in the national YRBS survey are sequential questions on the same behavior, new variables were engineered to account for the possible multicollinearity and help reduce the dimensionality for modeling. A Cramér's V analysis was first performed to detect the highly associated categorical variables. Those that are highly correlated were either merged into one combined variable or removed to prevent redundancy. Some associated variables are kept as they describe the health risk behavior from different aspects. To name a few, question 10 and 11 regarding drunk driving and texting while driving is combined into a binary indicator where as long as one of their record is not 0 (did not perform the behavior), the new variable ‘q10_11_unsafe_driving’ will be marked as 1 (=Yes). The new variable ‘q68_73_healthy_diet’ combines the sequential data on healthy diet behaviors and is set to 1 (=Yes) as long as the respondent responded ‘Yes’ to any 3 of them. More on feature engineering can be found in the ‘data_transform_renamed.ipynb’ Jupyter notebook in the GitHub repository.

Final Dataset and Encoding

The final data set consists of 71 variables. Before implementing with different models, the required processing is performed based on the model assumptions. For the interpretable logistic regression model, since there are too many associated variables, a manual feature election based on the EDA was performed, followed by dummy coding the categorical variables through get_dummies(). Since the tree models are capable of dealing with interactions between explanatory variables, the true categorical variables were label encoded and the ordinal variables were ordinal encoded.

Data Exploration

An exploratory data analysis (EDA) was conducted to ensure understanding of the response distributions to each question. Potential relationships between behaviors and questions were assessed through stacked bar plots. This bivariate analysis focused on relationships with 'suicide_attempt', the target binary indicator transformed from question 29 of the survey which asks for the number of times the respondent attempted suicide in the past 12 months. The complete set of visualization analyzed during this EDA can be found both in the 'data_transform_renamed.ipynb' notebook and the GitHub repository. The following only addresses the key takeaways from the exhaustive manual comparison and validation.

Univariate Analysis

The target variable 'suicide_attempt' recorded 2762 (13.74%) positive responses, suggesting that approximately 15% of the respondents reported at least one suicide attempt over the past 12 months. This raised concerns for the classifiers training over unbalanced data. On the other hand, the distribution of 'raceeth' shown in Figure 1.1 does reflect a similar distribution of proportions of different race and ethnicities in the U.S., published on the U.S. Census Bureau's QuickFacts page (2024). One thing to note from univariate analysis is that many of the risk behaviors, especially the high-risk (or extremely risky) ones, have very few positive responses. The distributions are logical, but not desirable for many of the machine learning models.

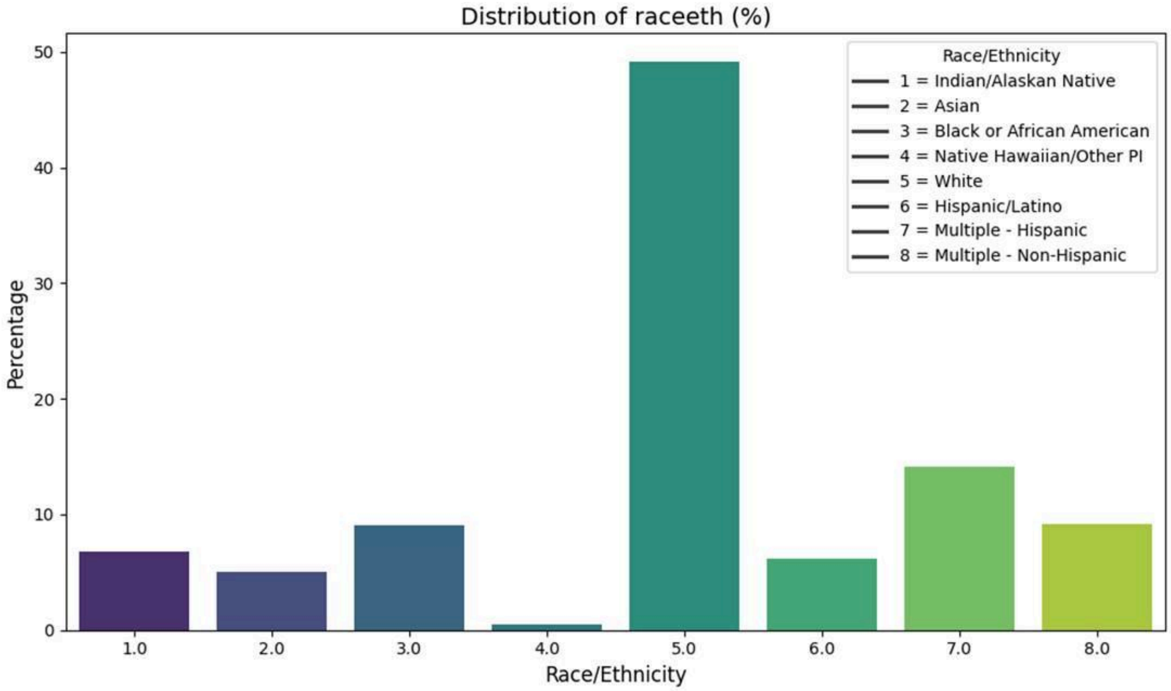


Figure 1.1. Distribution of 'raceeth' by percentage

Bivariate Analysis

Relationships with 'suicide_attempt' are analyzed through association scores and stacked bar plots grouped by suicide attempts. As mentioned in literature review, mental health is very often identified as one of the key contributors to suicide attempts. One of the mental health related question "Q26. During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activites?" (0=No, 1=Yes) does distribute differently across suicide attempt groups. The group of respondents that did not feel so depressed in the past 12 months does have a much lower percentage of positive response to 'suicide_attempt', validating this mental health indicator as a strong predictor for 'suicide_attempt'.

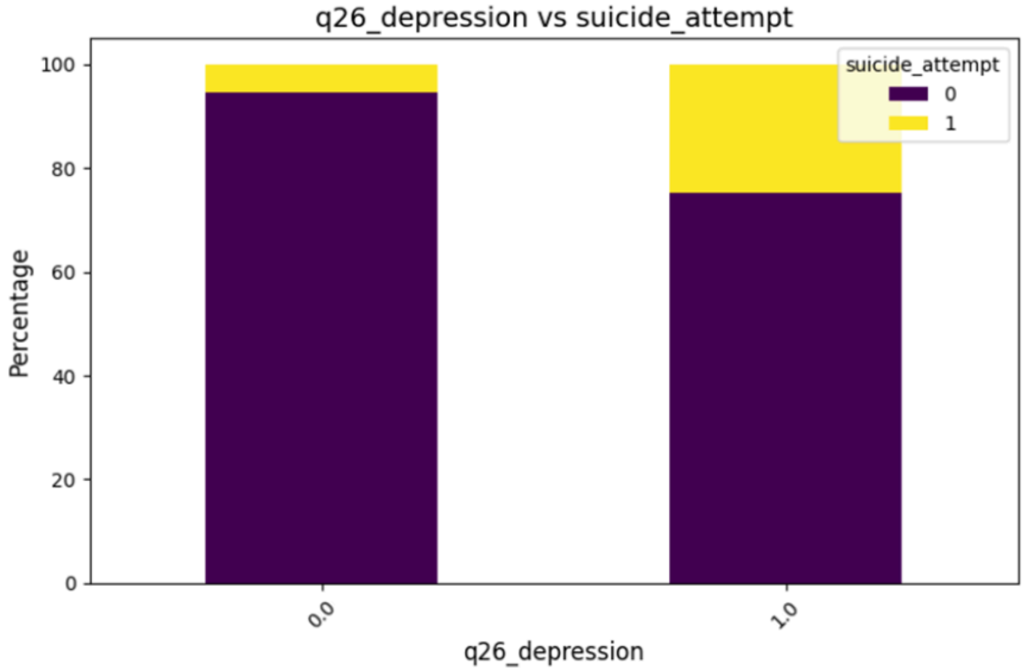


Figure 1.2. Q26 plotted by ‘suicide_attempt’

A representative relationship seen between health risk behaviors is shown in Figure 1.3. ‘q33_smoking’ denotes whether the respondent smoked during the past 30 days. A higher proportion of respondents who reported smoking during the past 30 days also reported attempting suicide, signaling the co-occurrence of health risk behaviors. Similarly for the ordinal variable question 42, asking for number of days the respondent drank in the past 30 days, Figure 1.4 is suggesting that the more number of days the respondent drank, the higher the ‘suicide_attempt’ proportions .

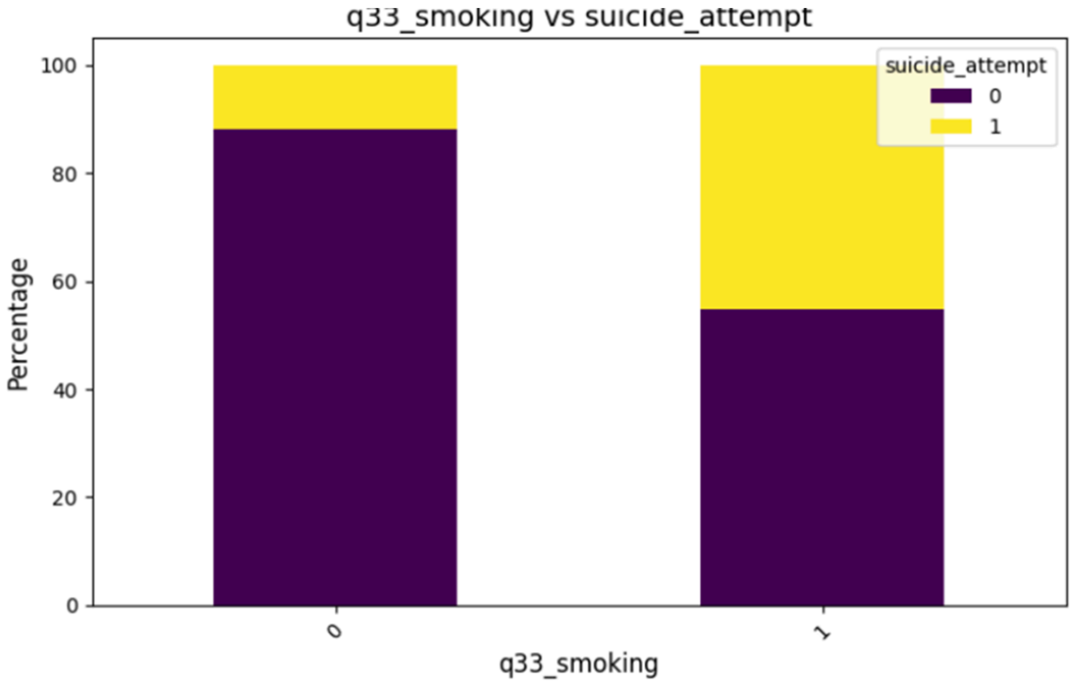


Figure 1.3. Q33 plotted by ‘suicide_attempt’

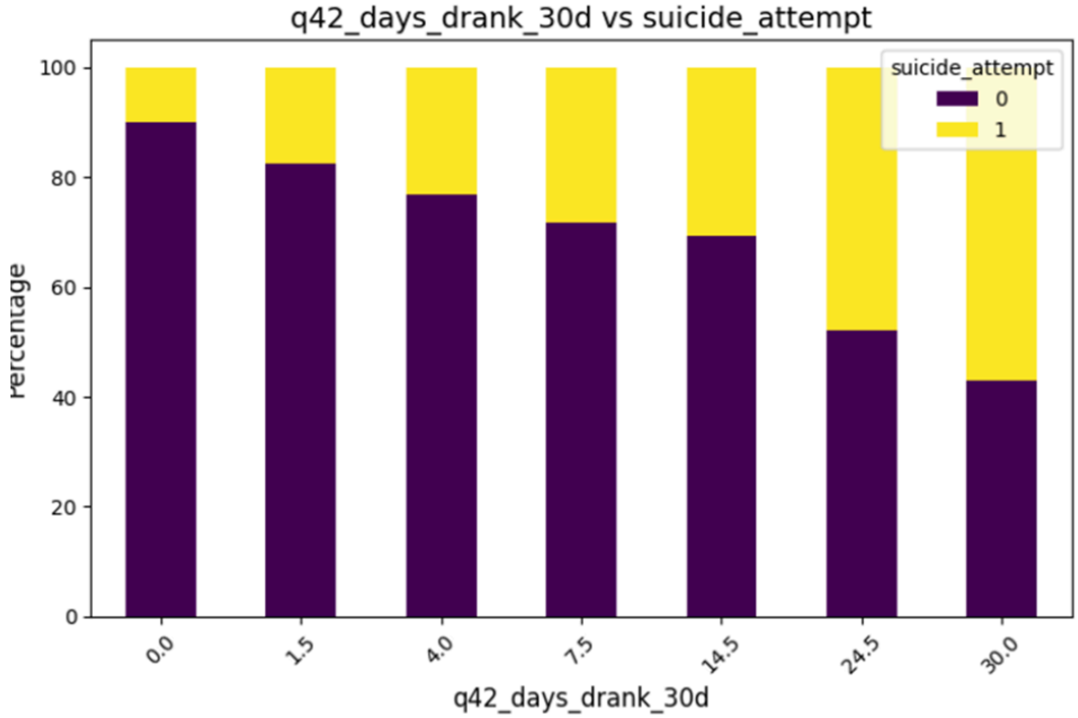


Figure 1.4. Q42 plotted by ‘suicide_attempt’

There are over 5 questions regarding sexual behaviors in the survey. The two in Figure 1.5 and Figure 1.6 are also showing significant differences among ordinal groups. ‘q57_age_first_sex’ has a first option denoting that the respondent had not engaged in sexual intercourse. This is denoted as 0 while the other categories are still recorded as the original age. As this age at first sex variable goes down, suicide_attempt proportions increases. A different aspect on sexual behaviors can be represented by question 20, recordi ng the number of times the respondent was forced to do sexual thing. As the count in Figure 1.6 goes up, the proportion of positive responses to ‘suicide_attempt’ seem to be rising linearly.

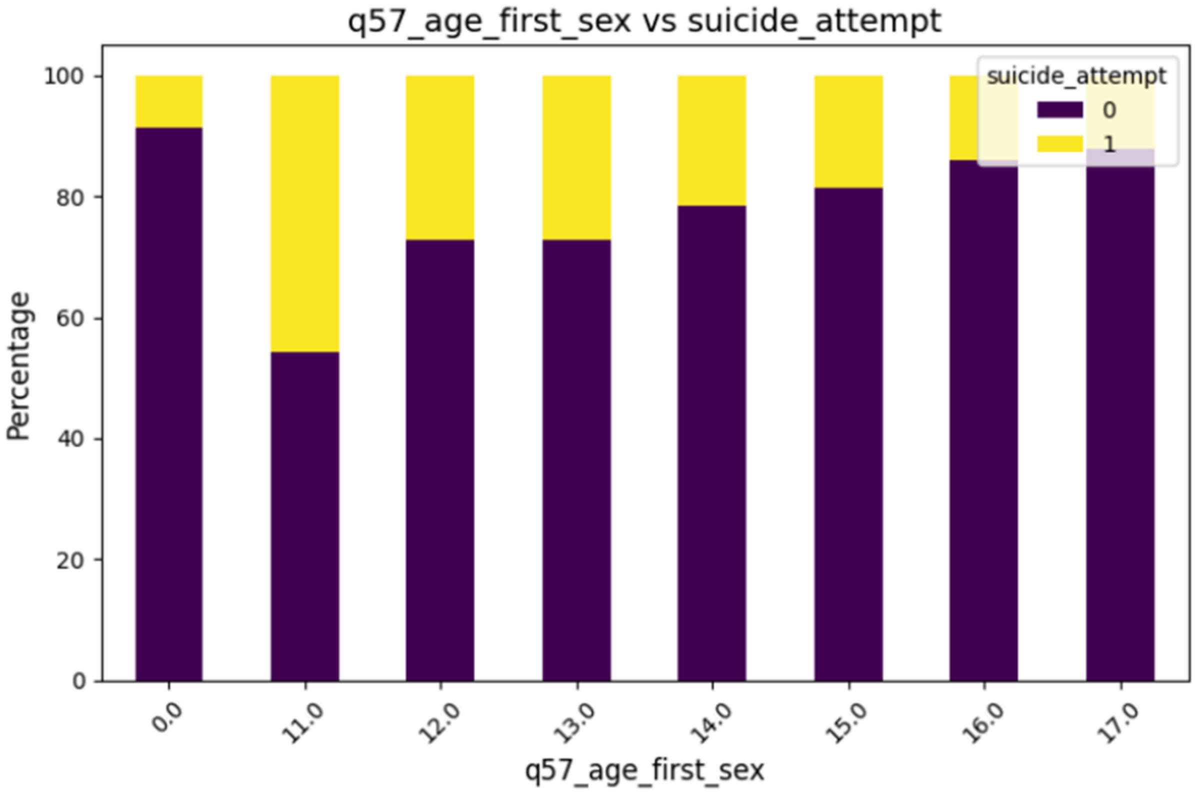


Figure 1.5. Q57 plotted by ‘suicide_attempt’

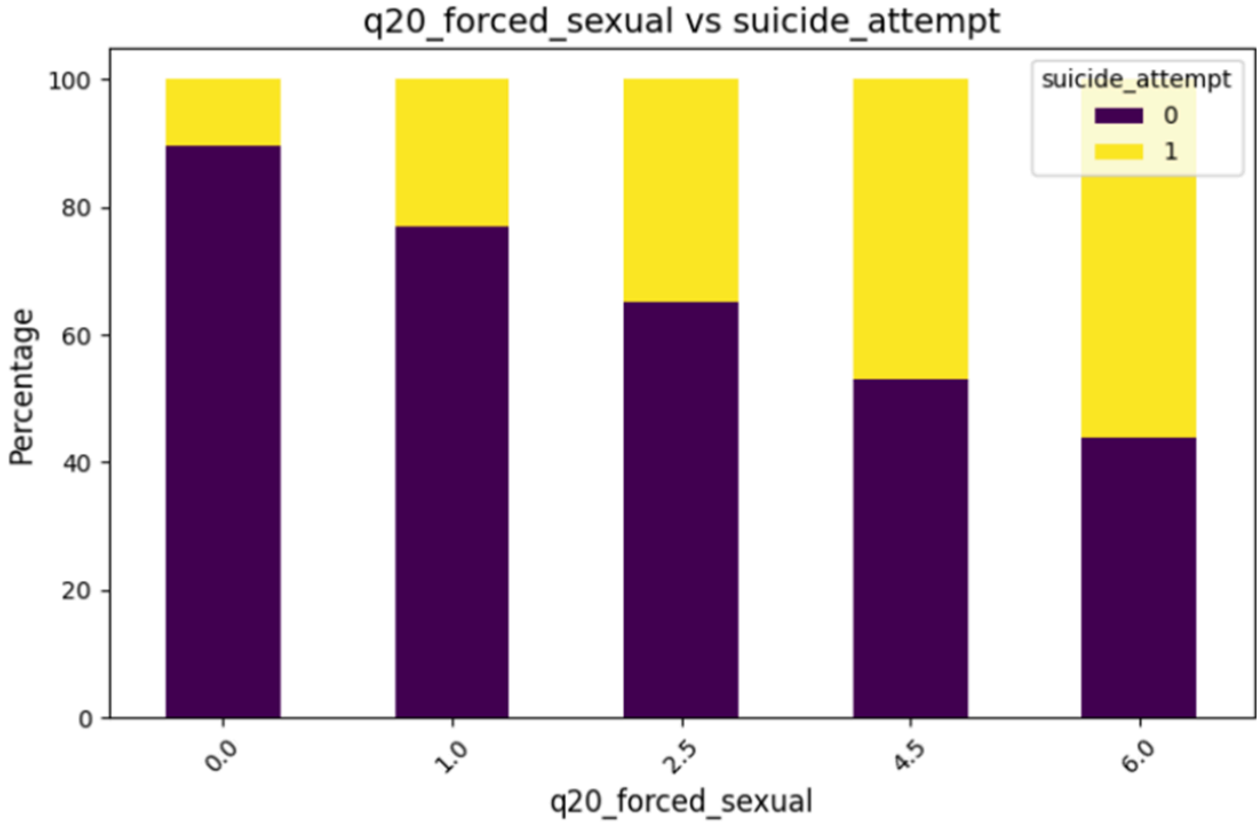


Figure 1.6 Q20 plotted by ‘suicide_attempt’

Relationships with continuous variables ‘q6_height’, ‘q7_weight’, and ‘ BMIPCT’ (BMI percentile) were analyzed using boxplots. But there was no substantial difference in their distributions between those who attempted suicide and those did not. Physical growth metrics may not be the desirable predictors for this project.

Modeling

Visual studio code,Jupyter Notebook, and Google Colab (T4 GPU) were used to run the models. Visual studio code and Jupyter Notebook allowed for interactive coding and visualization of results. Google Colab's T4 GPU accelerated computations.

KNN

The first model used is K-Nearest Neighbors (KNN). The data was first one-hot encoded and scaled using a standard scaler because KNN required standardized data to calculate the distance metrics, otherwise, features with larger range would contribute more to the computation of distance metrics. Using KNN clustering to generate distance-based features for predicting whether a case is a suicide attempt (1) or non-suicide attempt (0). This model was selected as the baseline to determine whether more sophisticated methods were needed to improve model performance. If KNN provided reasonable accuracy, it suggested that a simple distance-based relationship exists in the data. However, if KNN performs poorly, It will need to analyze the reasons behind it and explore more advanced models to check if they improve performance.

Model Evaluation for KNN

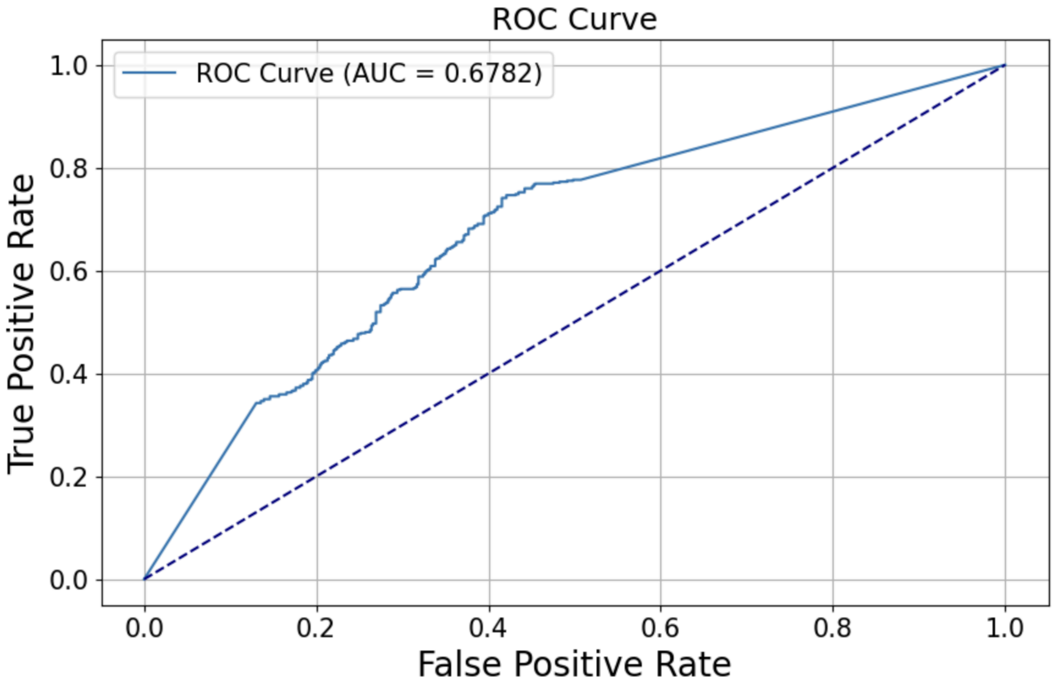
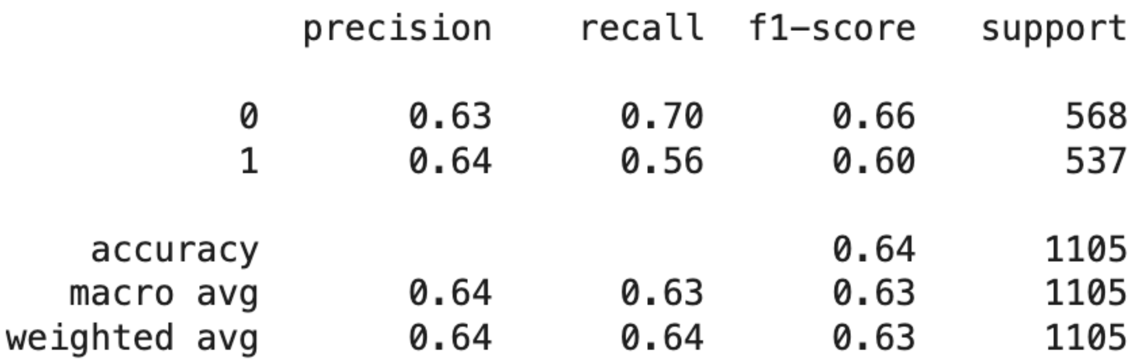


Figure 2.1 ROC curve for KNN



```
              precision    recall  f1-score   support

           0       0.63      0.70      0.66       568
           1       0.64      0.56      0.60       537

    accuracy                           0.64      1105
   macro avg       0.64      0.63      0.63      1105
weighted avg       0.64      0.64      0.63      1105
```

Figure 2.2 Classification report for KNN

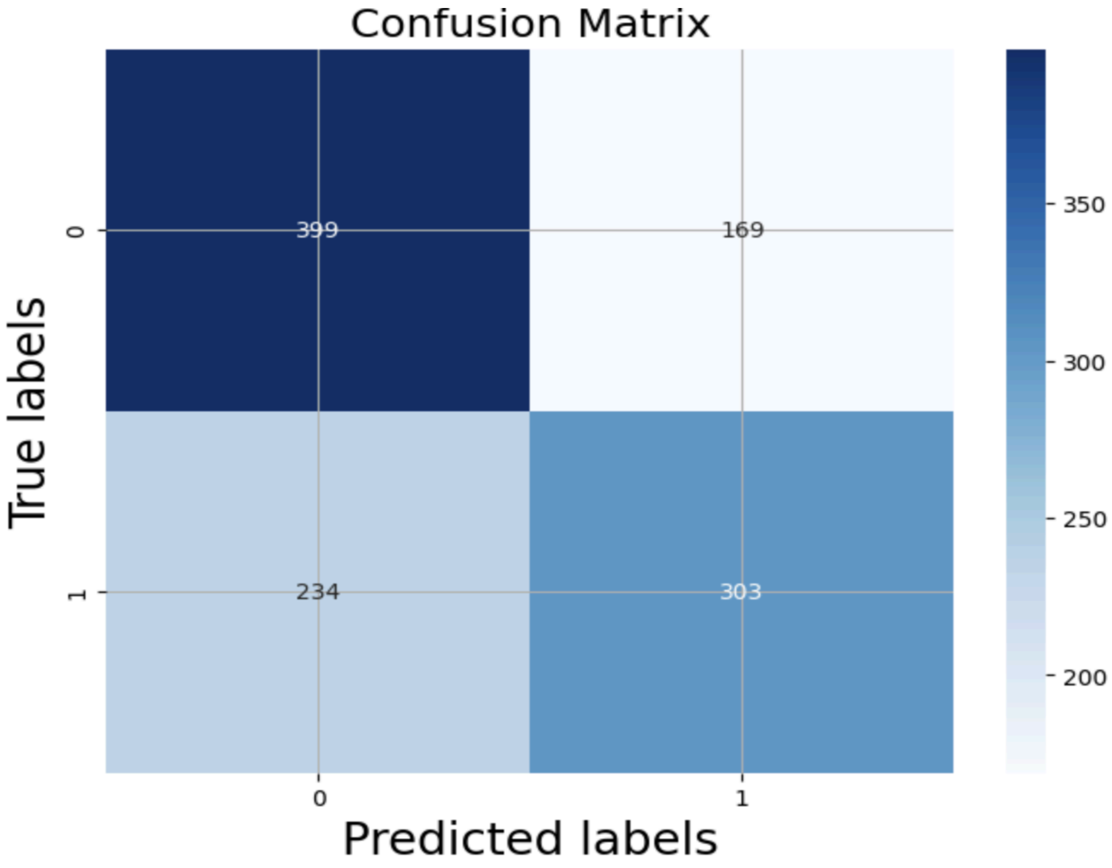


Figure 2.2 confusion matrix for KNN

Gridsearch is applied for the hypertuning process and the best parameters are ‘manhattan’ for ‘metric’, 4 for ‘n_neighbors’, ‘distance’ for ‘weights’. The AUC was 0.6782, meaning that 67.82% of the time, the model correctly ranked the positive case (suicide attempt) higher than the negative case (didn’t attempt suicide). The accuracy of KNN was 64%, indicating that the model correctly predicted suicide attempts 64% of the time. For suicide attempts, which were classified as class 1, the precision was 64%, showing that the model correctly predicted suicide attempts at a rate of 64%. For those who didn’t attempt suicide, the precision was 63%, indicating that the model correctly identified 63% of individuals who did not attempt suicide. The recall for suicide attempts (class 1) was 56%, meaning that 56% of actual suicide attempts were correctly identified, while 44% were missed. The recall for those who didn’t attempt suicide (class 0) was 70%, showing that 70% were correctly identified, while 30% were misclassified.

Since KNN’s accuracy still has room for improvement, Choosing Linear Discriminant Analysis (LDA) as the second model. The reason for this choice is that KNN classifier is based on proximity to neighbors but does not explicitly create a decision boundary. In contrast, LDA finds the optimal boundary that best separates suicide attempts (1) and non-suicide attempts (0) while maximizing class separation.

LDA

One-hot encoding and standard scaling was selected to satisfy the assumption of normality for LDA. Besides, LDA required numerical data, as a result, one hot encoding was applied to categorical variables to make them compatible with LDA. Since the dataset contained a large number of features, LDA transformed all features into feature maps. To manage this, LDA was utilized for dimensionality reduction while preserving class separation. LDA projected our multi-dimensional features onto a one-dimensional line, helping to maximize class separability and improve classification performance.

Model Evaluation for LDA

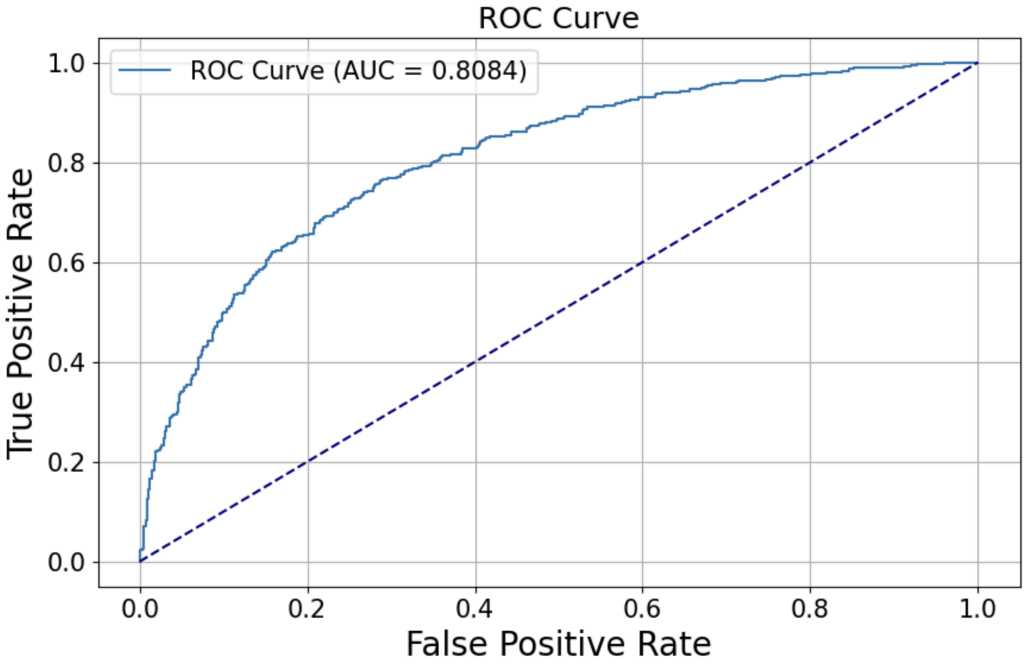
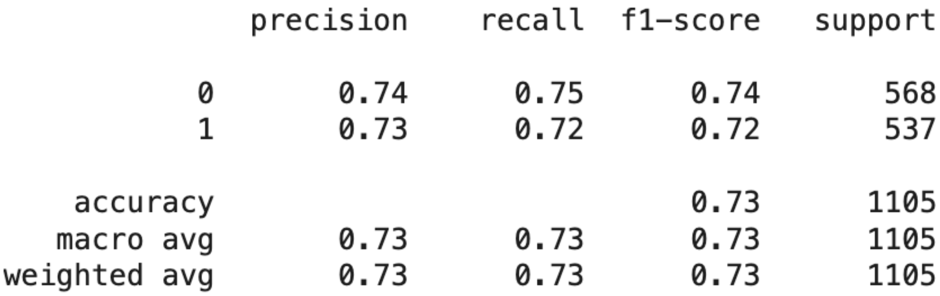


Figure 2.3 ROC Curve for LDA



```
              precision    recall  f1-score   support

           0       0.74      0.75      0.74       568
           1       0.73      0.72      0.72       537

    accuracy                           0.73      1105
   macro avg       0.73      0.73      0.73      1105
weighted avg       0.73      0.73      0.73      1105
```

Figure 2.4 Classification report for LDA

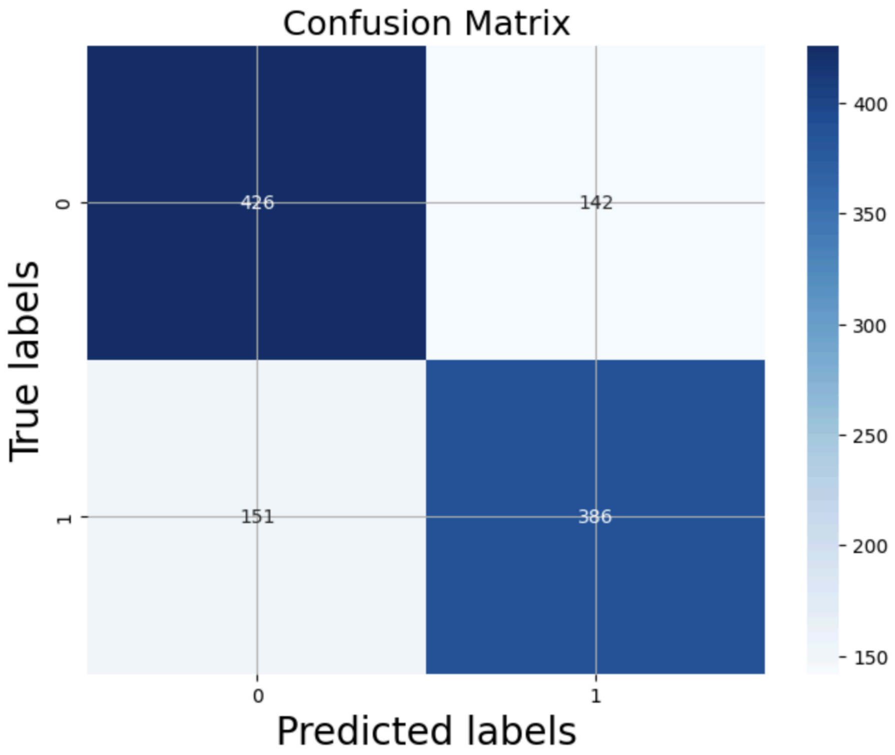


Figure 2.4 Confusion matrix for LDA

The AUC was 0.8072, meaning that 80.72% of the time, the model correctly ranked the positive case (suicide attempt) higher than the negative case (didn't attempt suicide). The accuracy of logistic regression with an L1 penalty was 74%, indicating that the model correctly predicted suicide attempts 74% of the time. For suicide attempts, which were classified as class 1, the precision was 73%, showing that the model correctly predicted suicide attempts at a rate of 73%. For those who didn't attempt suicide, the precision was 74%, indicating that the model correctly identified 74% of individuals who did not attempt suicide. The recall for suicide attempts (class 1) was 72%, meaning that 72% of actual suicide attempts were correctly identified, while 28% were missed, and the recall for those who didn't attempt suicide (class 0) was 75%, showing that 75% were correctly identified, while 25% were misclassified.

Logistic regression models

logistic regression models was selected in this project because the output is a binary dependent variable, where 0 represents non-suicide-attempting individuals and 1 represents suicide attempts. One-hot encoding was applied to meet the requirements of the logistic regression. Before fitting the data into the model, it is necessary to check whether logistic regression assumptions were met. The first assumption is that the dependent variable must be binary, which is met because our target variable has 0s and 1s. The second assumption is the independence of observations, which ensures that each individual's response is collected independently. Although this survey was conducted nationally, it cannot guarantee that all respondents were not influenced by a shared environment. For example, some students in the survey may come from the same school, which could introduce dependencies into the data. The third assumption is no multicollinearity among the independent variables. Multicollinearity was addressed at the Exploratory Data Analysis (EDA) stage. The fourth assumption requires no outliers of significance. Outliers had been addressed since our data source from CDC to ensure data honesty and validity as the dataset provided had already been cleaned.

For this logistic regression model, Cross-validation and hyperparameter tuning was applied to get the best performance. The model experimented with different penalty types— L1 (Lasso), and L2 (Ridge)—to determine which worked best. The best inverse of regularization strength value for Ridge was 0.12 and that for Lasso was 0.05. Finally testing the performance of the model using confusion matrices, ROC curves, and other relevant metrics to identify its effectiveness.

Model Evaluation for Logistic regression with L1

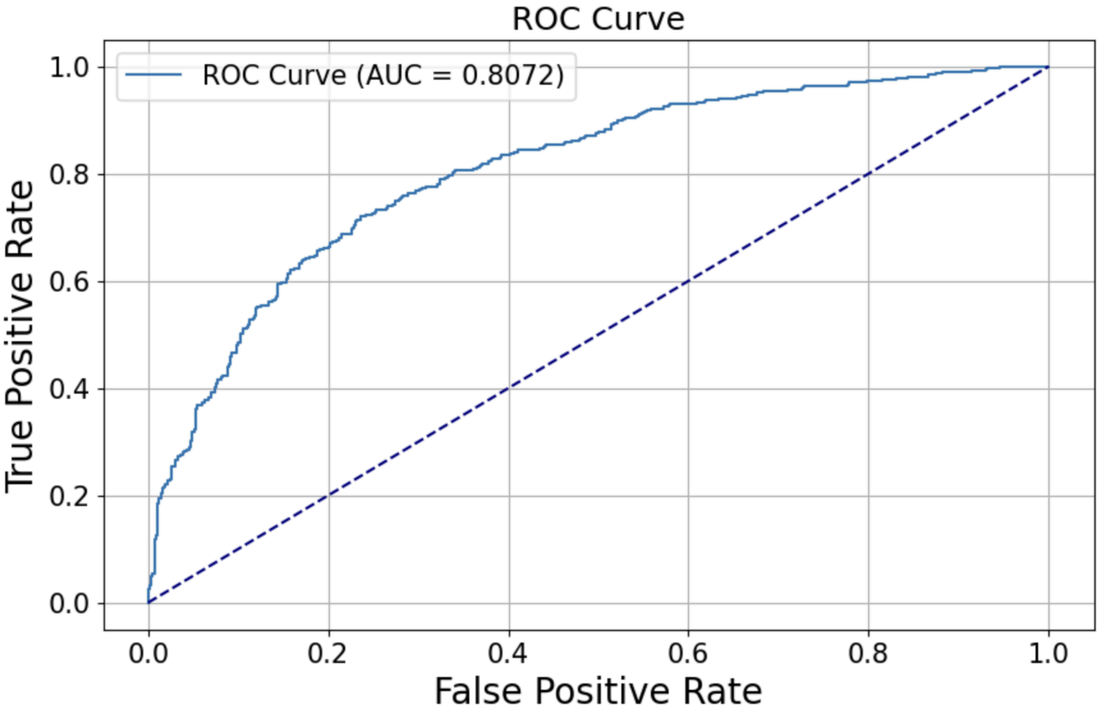
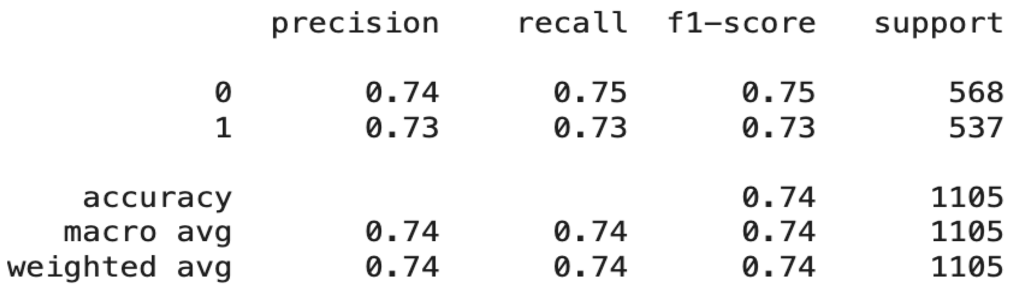


Figure 2.5 ROC curve for Logistic Regression with L1



```
              precision    recall  f1-score   support

           0       0.74      0.75      0.75       568
           1       0.73      0.73      0.73       537

    accuracy                           0.74      1105
   macro avg       0.74      0.74      0.74      1105
weighted avg       0.74      0.74      0.74      1105
```

Figure 2.6 Classification Report for Logistic Regression with L1

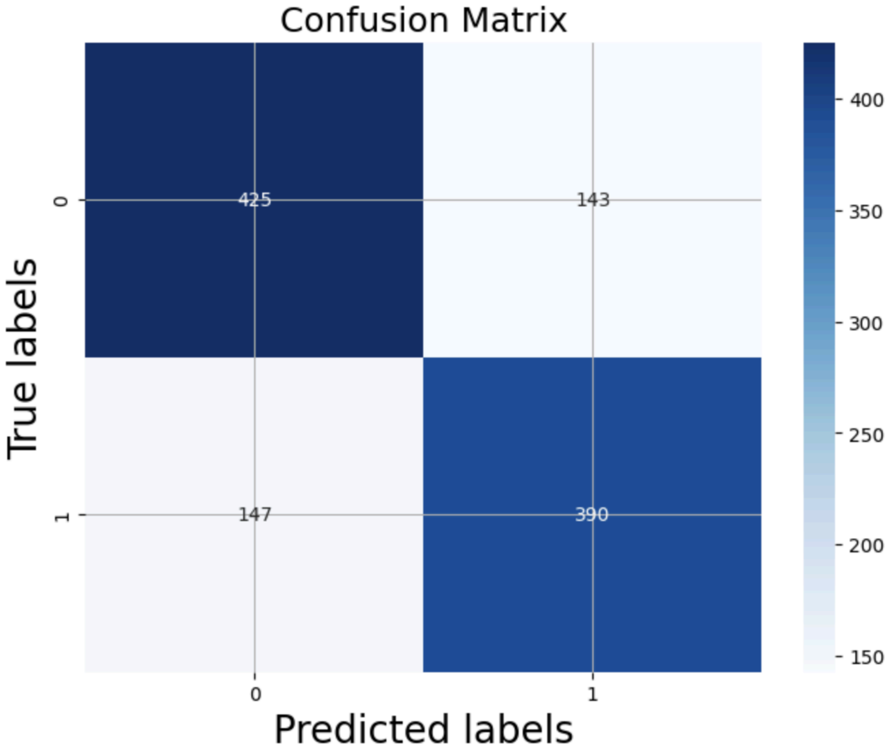


Figure 2.7 Confusion matrix for Logistic Regression with L1

The AUC was 0.8072, meaning that 80.72% of the time, the model correctly ranked the positive case (suicide attempt) higher than the negative case (didn’t attempt suicide). The accuracy of logistic regression with an L1 penalty was 74%, indicating that the model correctly predicted suicide attempts 74% of the time. For suicide attempts, which were classified as class 1, the precision was 73%, showing that the model correctly predicted suicide attempts at a rate of 73%. For those who didn’t attempt suicide, the precision was 74%, indicating that the model correctly identified 74% of individuals who did not attempt suicide. The recall for suicide attempts (class 1) was 73%, meaning that 73% of actual suicide attempts were correctly identified, while 27% were missed. The recall for those who didn’t attempt suicide (class 0) was 75%, showing that 75% were correctly identified, while 25% were misclassified.

Model Evaluation for Logistic regression with L2

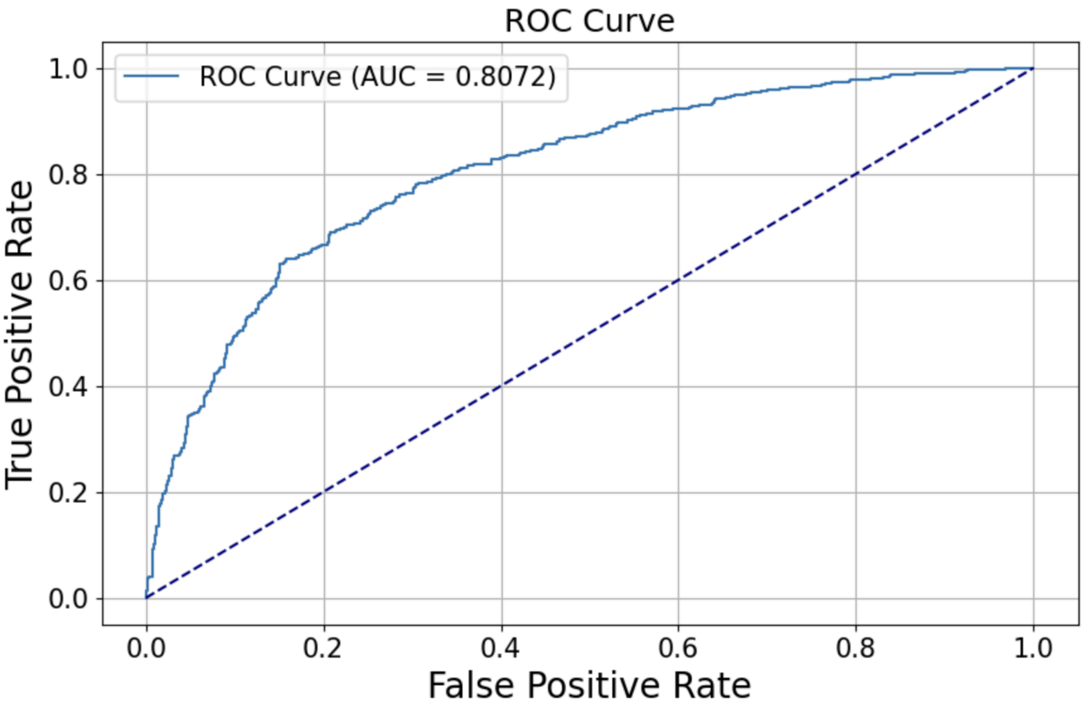
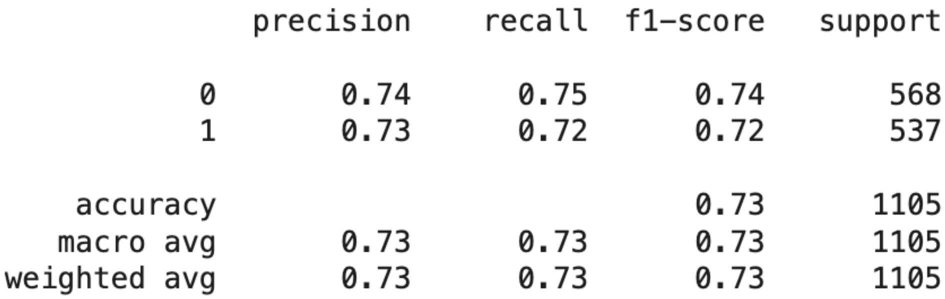


Figure 2.8 ROC curve for Logistic Regression with L2



```
              precision    recall  f1-score   support

           0       0.74      0.75      0.74       568
           1       0.73      0.72      0.72       537

    accuracy                           0.73      1105
   macro avg       0.73      0.73      0.73      1105
weighted avg       0.73      0.73      0.73      1105
```

Figure 2.9 Classification report for Logistic Regression with L2

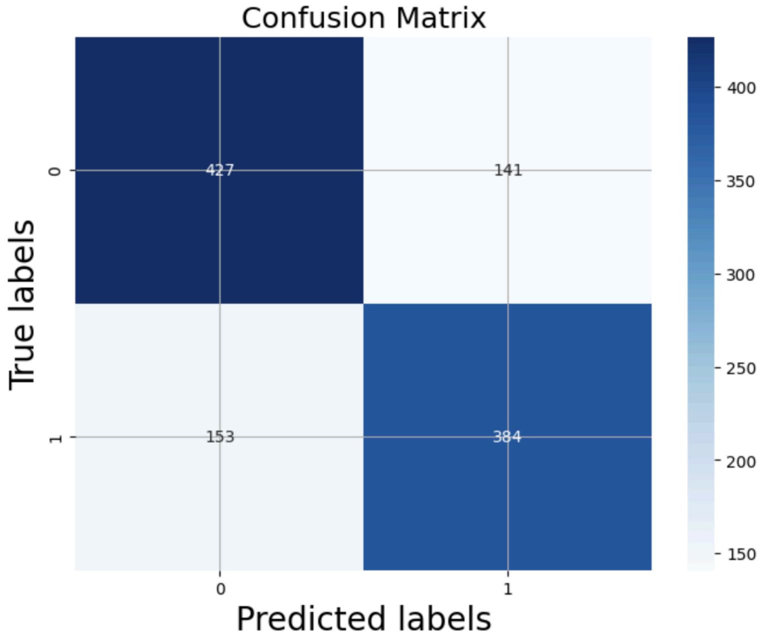


Figure 2.9 Confusion matrix for Logistic Regression with L2

The AUC was 0.8072, meaning that 80.72% of the time, the model correctly ranked the positive case (suicide attempt) higher than the negative case (didn’t attempt suicide). The accuracy of logistic regression with an L2 penalty was 73%, indicating that the model correctly predicted suicide attempts 73% of the time. For suicide attempts, which were classified as class 1, the precision was 73%, showing that the model correctly predicted suicide attempts at a rate of 73%. For those who didn’t attempt suicide, the precision was 74%, indicating that the model correctly identified 74% of individuals who did not attempt suicide. The recall for suicide attempts (class 1) was 72%, meaning that 72% of actual suicide attempts were correctly identified, while 28% were missed. The recall for those who didn’t attempt suicide (class 0) was 75%, showing that 75% were correctly identified, while 25% were misclassified.

By analyzing the confusion matrix and other evaluation metrics, The result shows that the performance of logistic regression remained very similar across different penalty types. Although

L1 (Lasso) can zero out some coefficients and L2 (Ridge) shrinks them, the net effect on classification in the dataset remains the same. This resulted in almost identical confusion matrices and exactly the same accuracy. This is because the dataset is not heavily overfitting or suffering from high variance, so the choice of penalty type does not significantly impact the model's outcome.

Suggested Solution

To capture the broader effect of risky behaviors on the likelihood of suicide through regression, correlated sequential items must be reduced to score-based evaluation for all risky behaviors. It preserves meaningful differences between similar but conceptually different variables and prevents losing valuable behavioral trends.

Logistic regression handles multicollinearity in a rough manner, in the sense that some highly correlated questions are still capturing different aspects of behavior. For example, although Q56 (early sex) and risky sex both have a high correlation rating of 0.944, it will capture different aspects of sexual behavior and risk-taking. Leaving them out or collapsing them simply may result in loss of information, as it would not capture the differential level of risk associated with various behaviors.

If our objective is to test the correlation between individual high-risk behaviors and suicide attempts, dropping or collapsing correlated variables could too simplistically simplify the model and hide valuable insight. Rather, a score-based method or expert-imbued feature engineering would enable us to retain distinctions in behavior and more accurately evaluate their effect on suicide risk. To overcome these limitations, It will require a scoring system with right weights for each risky behavior to facilitate a finer analysis. Construction of such a measure involves expert consideration from sociology and psychology experts to make sure that the scores directly indicate the behavioral impact on suicide risk.

The next step was to use tree-based models, as these models are able to capture associations between features automatically, avoiding some of the limitations of logistic regression. Tree-based models could identify multivariable interaction complexity and make improved predictions.

Decision Tree

Decision trees showed the decision-making process. Label encoding and ordinal encoding was applied. For this process it would be visualized for more clear understanding. To hypertune the model, the Grid Search method was used to get the optimized results. The outcome showed that the best parameters were 10 for 'max_depth', 1 for 'min_samples_leaf', and 2 for 'min_samples_split'.

Model Evaluation for Decision tree

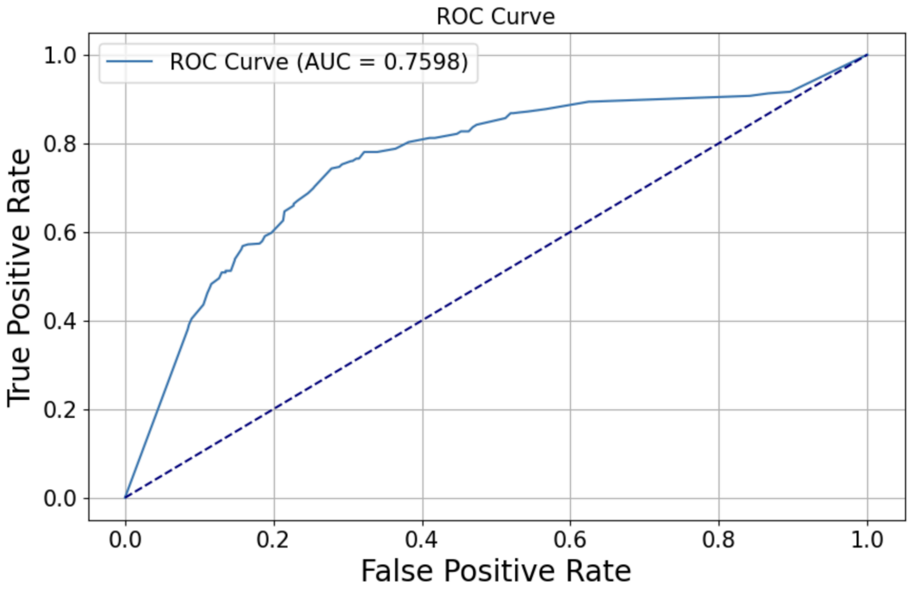
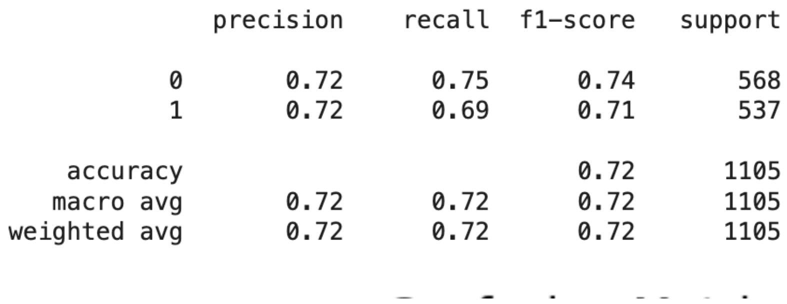


Figure 2.10 ROC curve for decision tree



```
              precision    recall  f1-score   support

           0       0.72      0.75      0.74       568
           1       0.72      0.69      0.71       537

    accuracy                           0.72      1105
   macro avg       0.72      0.72      0.72      1105
weighted avg       0.72      0.72      0.72      1105
```

Figure 2.11 Classification report for decision tree

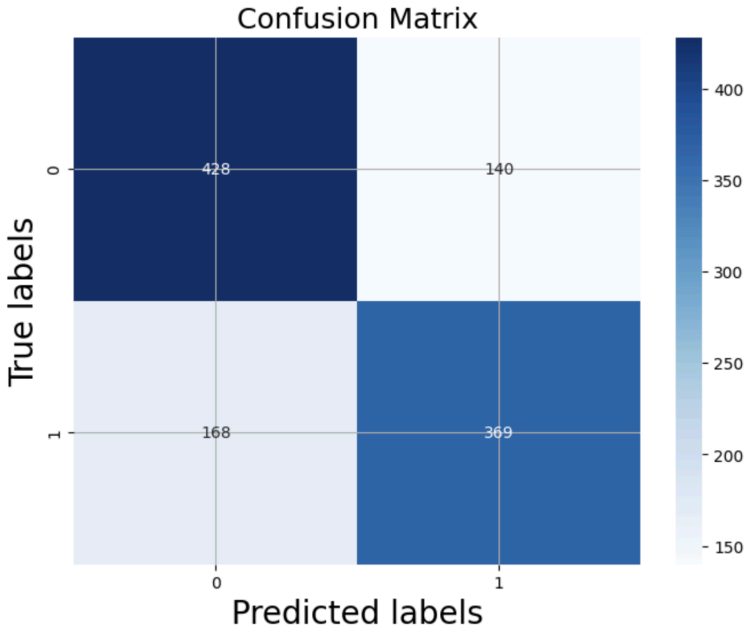


Figure 2.12 Confusion matrix for decision tree

The AUC was 0.7598, meaning that 75.98% of the time, the model correctly ranked the positive case (suicide attempt) higher than the negative case (didn't attempt suicide). The accuracy of the decision tree was 72%, indicating that the model correctly predicted suicide attempts 72% of the time. For suicide attempts, which were classified as class 1, the precision was 72%, showing that the model correctly predicted them at a rate of 72%. For those who didn't attempt suicide, the precision was also 72%, indicating that the model correctly identified 72% of individuals who did not attempt suicide. The recall for suicide attempts (class 1) was 69%, meaning that 69% of actual suicide attempts were correctly identified, while 31% were missed. The recall for those who didn’t attempt suicide (class 0) was 75%, showing that 75% were correctly identified, while 25% were misclassified.

Random Forest

Random forest is utilized to capture some more complex relationships between features, unlike LDA it works well with both categorical and numerical features. This model is able to handle feature interactions for example it would learn the relationships between substance uses with depression automatically. Another benefit of this model is it would prevent overfit problems since it is ensemble learning by using multiple trees. This model would also help use to identify some feature importance that related to suicide attempt. Similar to the Decision Tree model, Random Forest also required label encoding and ordinal encoding, while one-hot encoding and scaling were unnecessary because the tree models splitted the data recursively, which indicated a non-sensitive behavior with respect to one-hot encoding and scaling. The hypertuned results were True for ‘bootstrap’, None for ‘max_depth’, 2 for ‘min_samples_leaf’, 2 for ‘min_samples_split’, 100 for ‘n_estimators’.

Model Evaluation for Random Forest

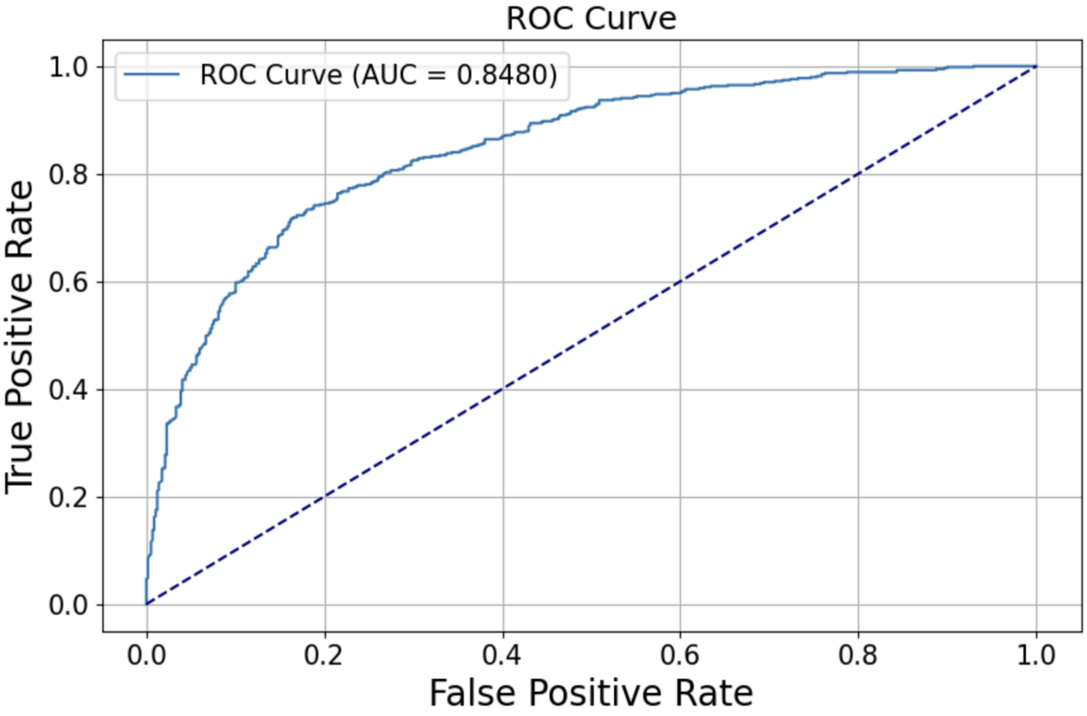
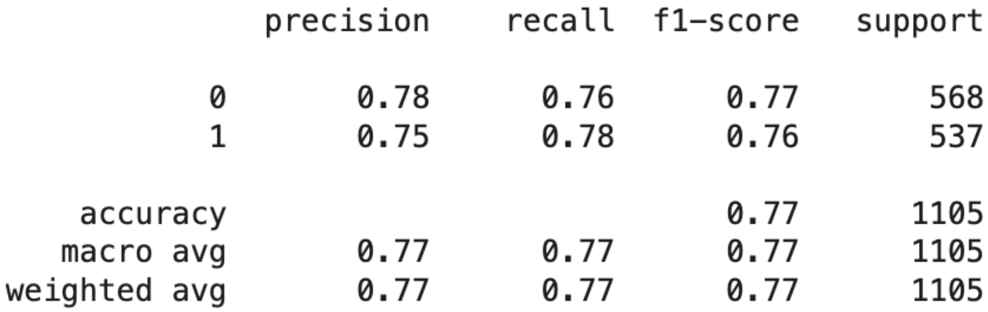


Figure 2.13 ROC curve for random Forest



```
              precision    recall  f1-score   support

           0       0.78      0.76      0.77       568
           1       0.75      0.78      0.76       537

    accuracy                           0.77      1105
   macro avg       0.77      0.77      0.77      1105
weighted avg       0.77      0.77      0.77      1105
```

Figure 2.14 Classification matrix for random forest

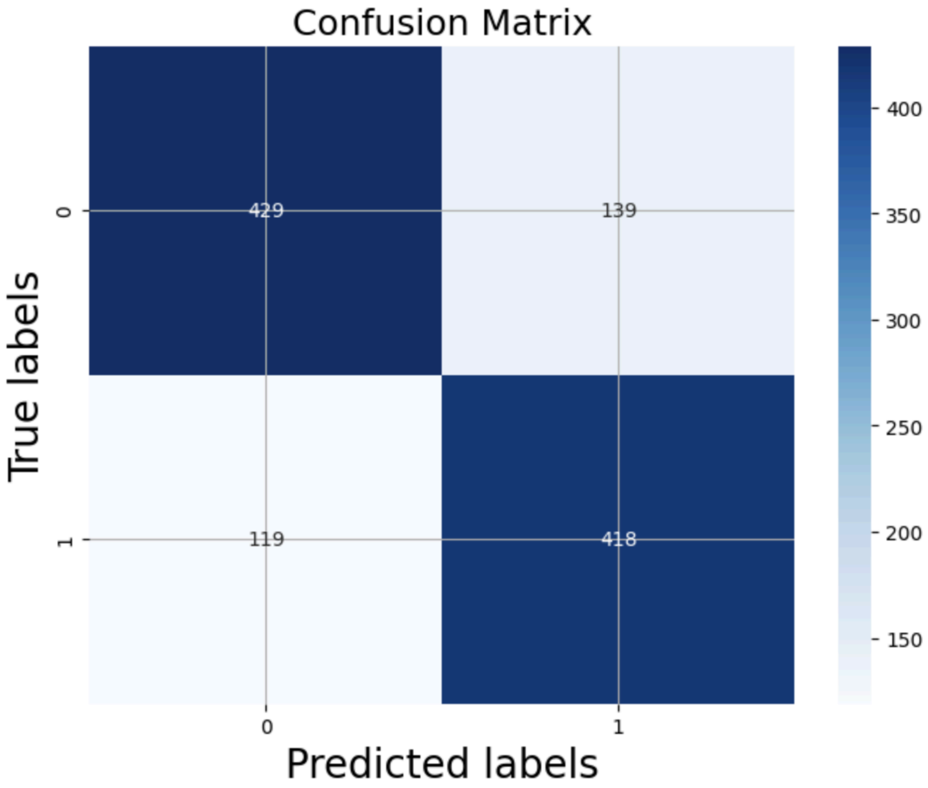


Figure 2.15 Confusion matrix for random forest

The AUC was 0.8480, meaning that 84.8% of the time, the model correctly ranked the positive case (suicide attempt) higher than the negative case (didn't attempt suicide). The accuracy of Random Forest was 77%, indicating that the model correctly predicted suicide attempts 77% of the time. For suicide attempts, which were classified as class 1, the precision was 75%, showing that the model correctly predicted them at a rate of 75%. For those who didn't attempt suicide, the precision was 78%, indicating that the model correctly identified 78% of individuals who did not attempt suicide. The recall for suicide attempts (class 1) was 78%, meaning that 78% of actual suicide attempts were correctly identified, while 22% were missed. The recall for those who didn't attempt suicide (class 0) was 76%, showing that 76% were correctly identified, while 24% were misclassified.

Boosting Models (LightGBM and XGBoost)

Before fitting into the XGBoost, LightGBM was utilized due to the faster computational speed for larger datasets compared to XGBoost. Like other tree-based methods, lightGBM and XGBoost required neither one-hot encoding nor scaling, simply ordinal encoding and label encoding was enough. The parameters for the best lightGBM model were 0.09 for ‘learning_rate’, 6 for ‘max_depth’, 140 ‘n_estimators’. The best parameters for XGBoost were 0.08 for ‘learning_rate’, 6 for ‘max_depth’, 100 ‘n_estimators’.

Model Evaluation for LightGBM

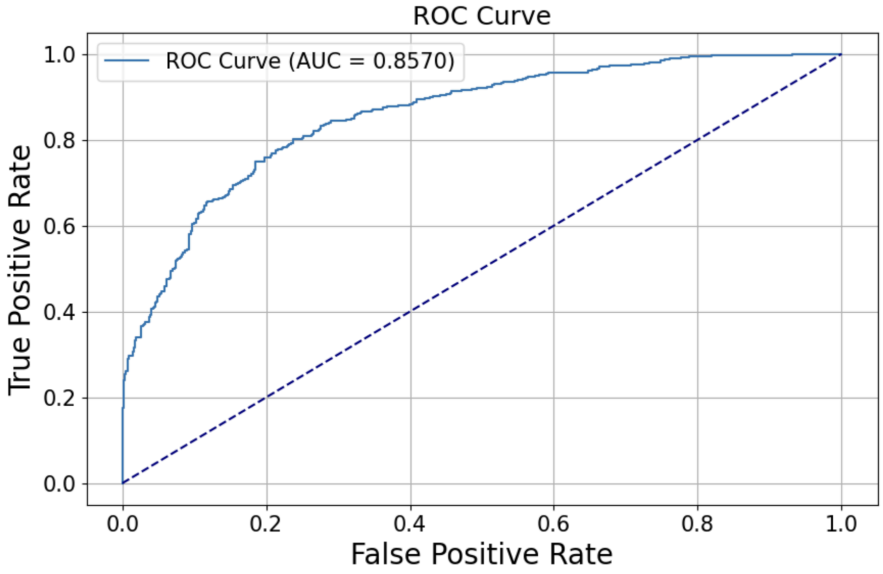
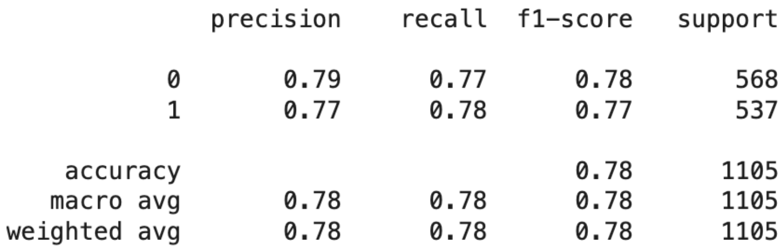


Figure 2.16 ROC curve for LightGBM



```
              precision    recall  f1-score   support

           0       0.79      0.77      0.78       568
           1       0.77      0.78      0.77       537

    accuracy                           0.78      1105
   macro avg       0.78      0.78      0.78      1105
weighted avg       0.78      0.78      0.78      1105
```

Figure 2.17 Classification report for LightGBM

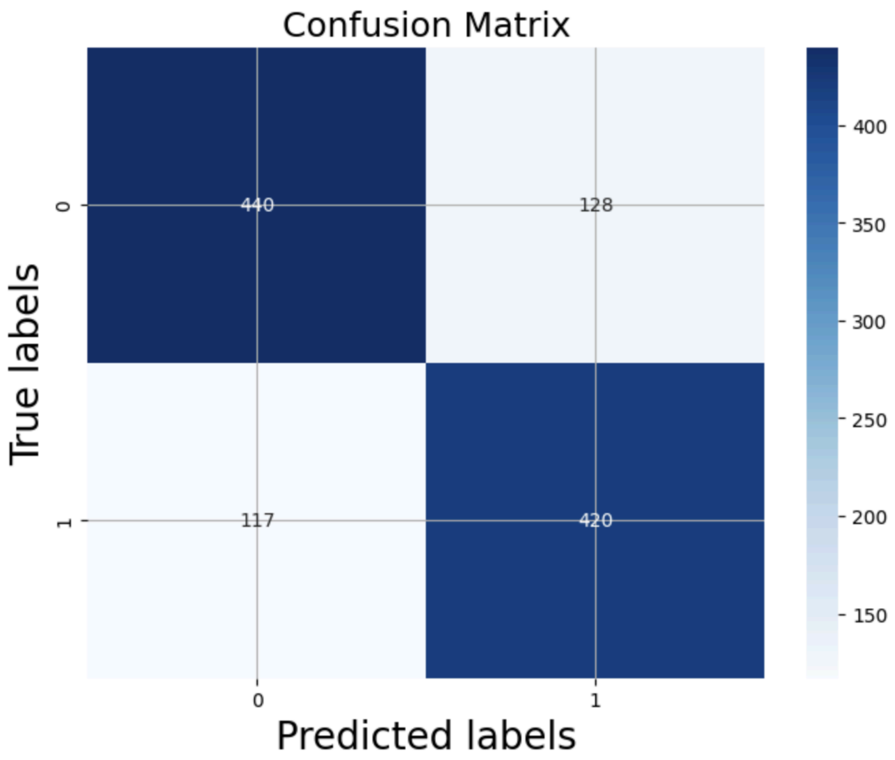


Figure 2.18 Confusion matrix for LightGBM

The AUC was 0.8570, meaning that 85.7% of the time, the model correctly ranked the positive case (suicide attempt) higher than the negative case (didn't attempt suicide). The accuracy of LightGBM was 78%, indicating that the model correctly predicted suicide attempts 78% of the time. For suicide attempts, which were classified as class 1, the precision was 77%, showing that the model correctly predicted them at a rate of 77%. For those who did not attempt suicide, the precision was 79%, indicating that the model correctly identified 79% of individuals who did not attempt suicide. The recall for suicide attempts (class 1) was 78%, meaning that 78% of actual suicide attempts were correctly identified, while 22% were missed. The recall for those who did not attempt suicide (class 0) was 77%, showing that 77% were correctly identified, while 23% were misclassified.

Model Evaluation for XGBoost

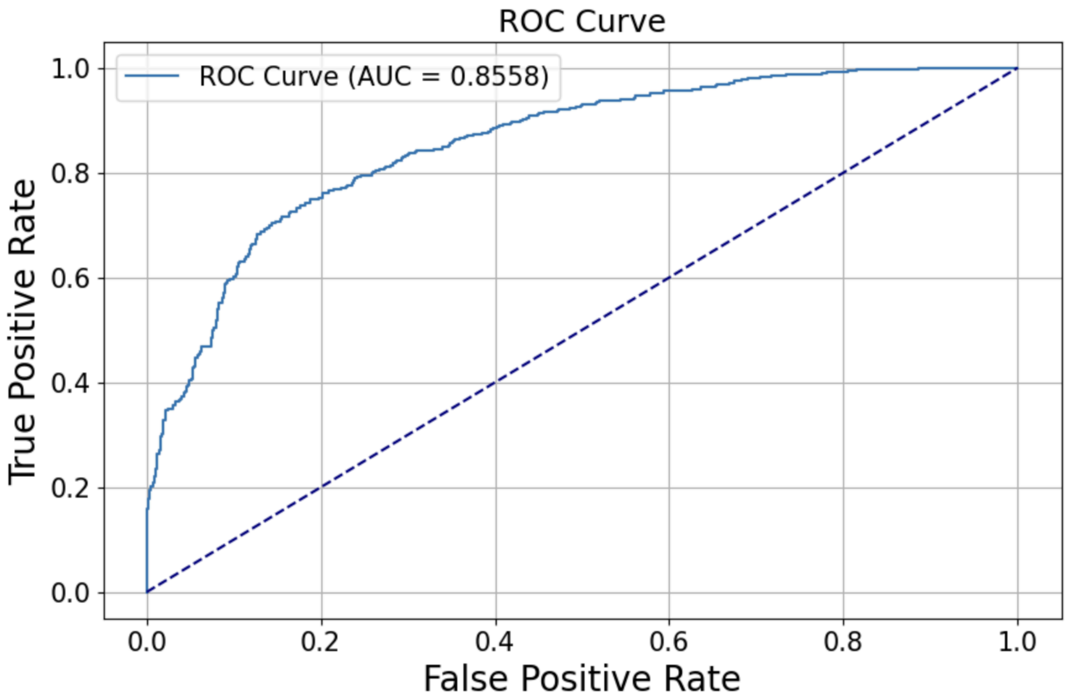
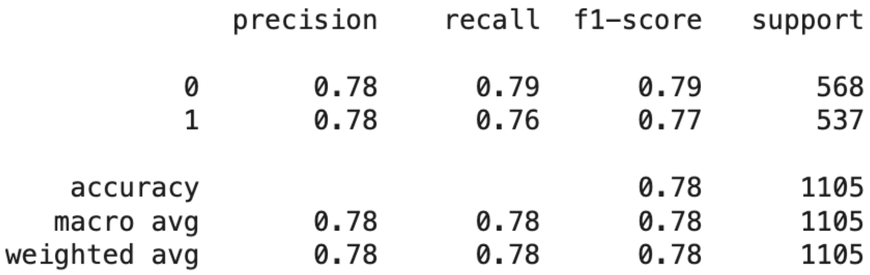


Figure 2.19 ROC curve for XGBoost



```
              precision    recall  f1-score   support

           0       0.78      0.79      0.79       568
           1       0.78      0.76      0.77       537

    accuracy                           0.78      1105
   macro avg       0.78      0.78      0.78      1105
weighted avg       0.78      0.78      0.78      1105
```

Figure 2.20 Classification matrix for XGBoost

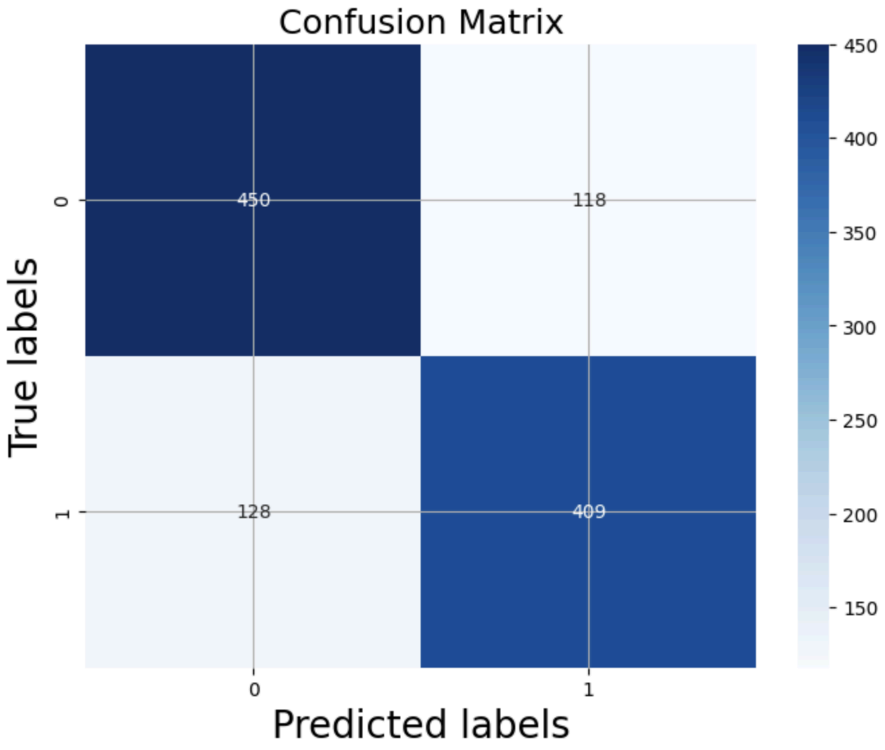


Figure 2.21 Confusion matrix for XGBoost

Ensembled Model

Combining LightGBM, XGBoost, and Random Forest would increase the strength of multiple models and improve the prediction performance. This ensemble method would improve accuracy. Random Forest would prevent overfitting, LightGBM was more efficient and XGboost would help to classify better by finding complex patterns. The models were then combined using soft voting. In our project, missing actual suicide attempts (false negatives) was more dangerous than missing to identify those who did not attempt to suicide (false positives). If one model was slightly biased to false negative, another model could balance it by using soft voting.

Model Evaluation for Ensembled Model

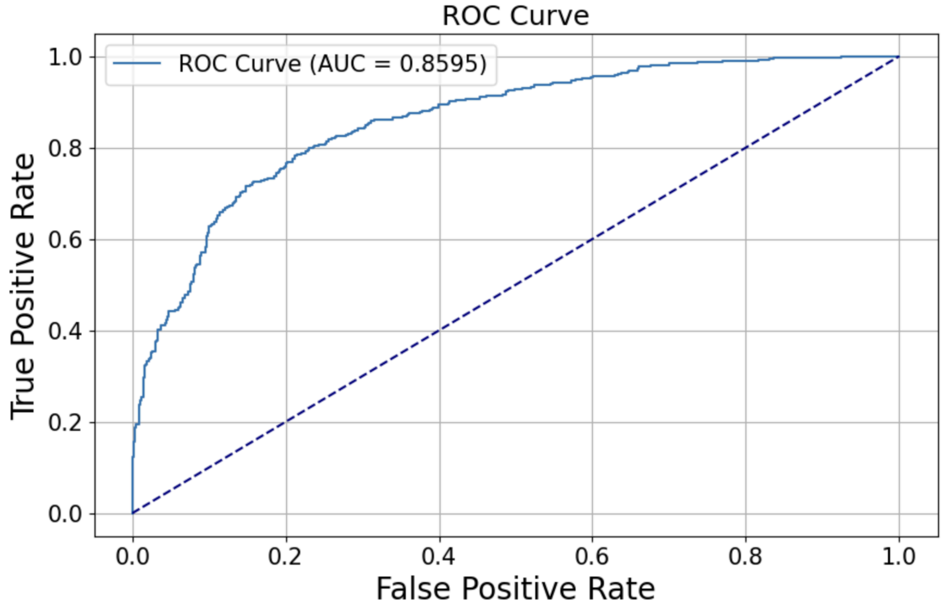
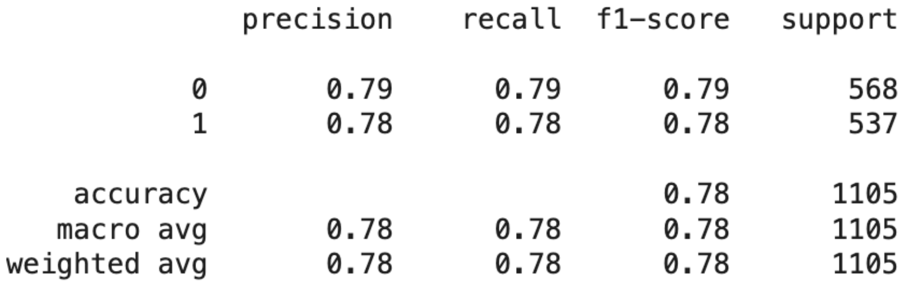


Figure 2.22 ROC curve for ensemble model



```
              precision    recall  f1-score   support

           0       0.79      0.79      0.79       568
           1       0.78      0.78      0.78       537

    accuracy                           0.78      1105
   macro avg       0.78      0.78      0.78      1105
weighted avg       0.78      0.78      0.78      1105
```

Figure 2.23 Classification report for ensemble model

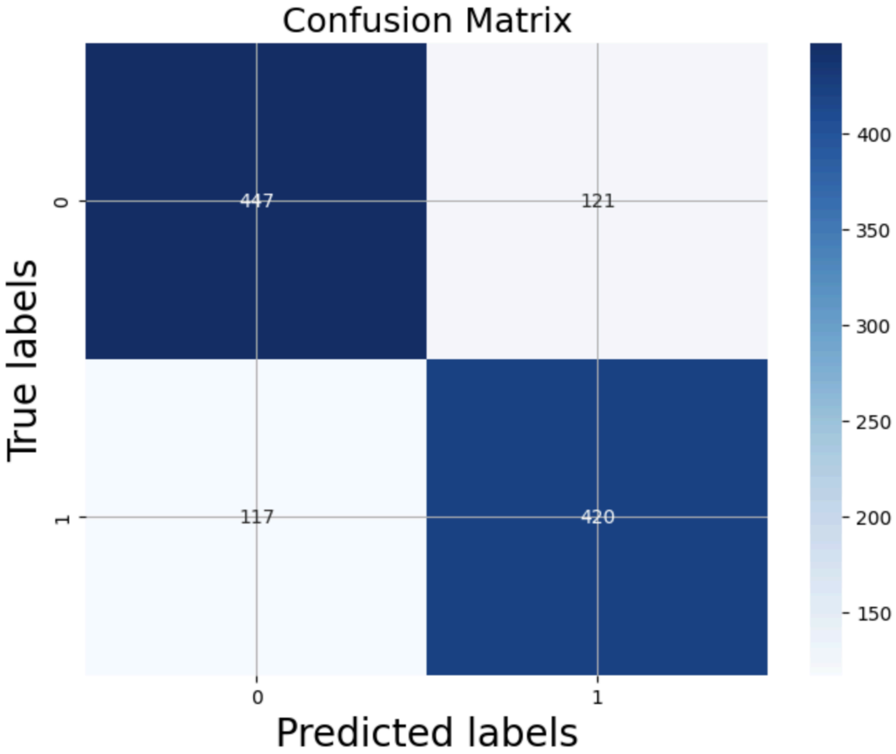


Figure 2.24 Confusion matrix for ensemble model

The AUC was 0.8595, meaning that 85.95% of the time, the model correctly ranked the positive case (suicide attempt) higher than the negative case (didn’t attempt suicide). The accuracy of the ensembled model was 78%, indicating that the model correctly predicted suicide attempts 78% of the time. For suicide attempts, which were classified as class 1, the precision was 78%, showing that the model correctly predicted them at a rate of 78%. For those who didn’t attempt suicide, the precision was 79%, indicating that the model correctly identified 79% of individuals who did not attempt suicide. The recall for suicide attempts (class 1) was 78%, meaning that 78% of actual suicide attempts were correctly identified, while 22% were missed. The recall for those who didn’t attempt suicide (class 0) was 79%, showing that 79% were correctly identified, while 21% were misclassified.

Model Comparison and Discussion

For our project, we hope to increase recall score as high as possible, since it could lead to serious consequence to miss true suicide attempt. However, all models initially achieved very low recall scores because there was an imbalance of the dataset. To be specific, there were approximately 17,000 samples with no suicidal tendencies while only around 2,762 adolescents attempted to suicide. In this context, the model tended to optimize for the larger class, which was not satisfactory because our focus was on the suicidal class. To counter the issue, we rebalanced the dataset using downsampling and reduced the class size with no suicidal attempts to the same size as the smaller class. As a result, the recall scores improved from around 0.3 to 0.78. This improvement was very important for our project because the more actual suicide attempts we caught, the more meaningful our model became. It enhanced the model's ability to distinguish suicide attempts and potentially prevented them from happening.

Deployment

The model will categorize the at-risk cases for attempted suicide on a behavioral and survey basis. Maximizing the number of recalls and not returning any false negatives is very important, as excluding a genuine suicide attempt might involve important issues.

The model can be deployed on a web page where researchers or mental health workers can input survey information and obtain risk estimates. This is achievable with Flask or FastAPI as the Python backend and React or Vue.js as the frontend. It is also possible to create a mobile app that provides risk assessment and early alerting to the subject or the counselor. Integration with a health platform could allow for real-time monitoring. One can also create a REST API in a way that data is passed through other external systems (e.g., hospitals, mental health clinics) and predictions are received. This is done by developing the API with Flask or FastAPI and deploying it on AWS Lambda or Google Cloud Functions.

Feature Importance

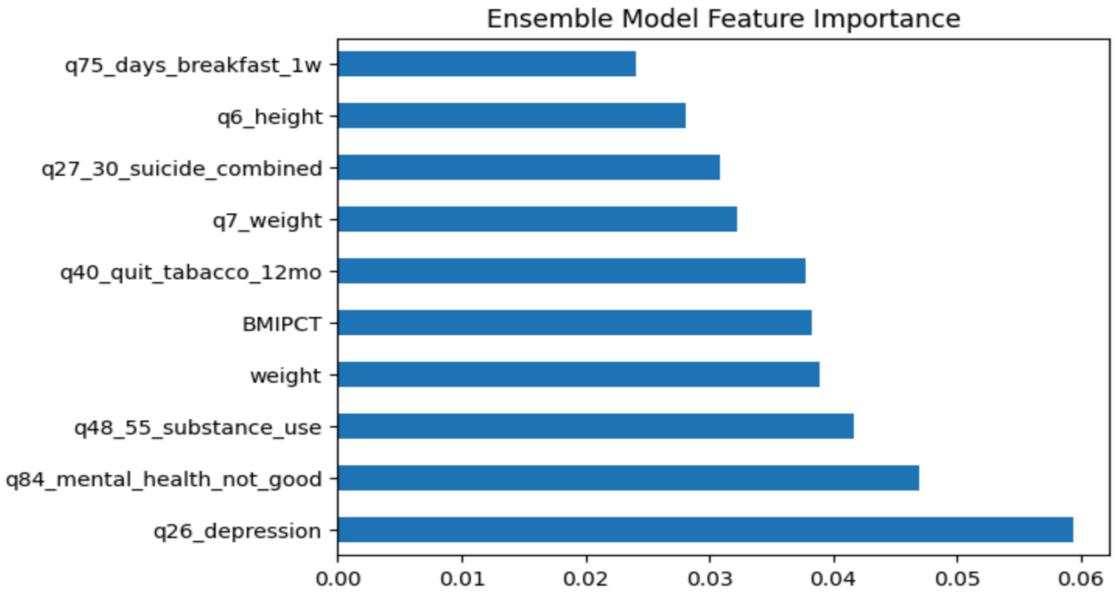


Figure 2.25 Feature importance for ensemble model

From the ensemble model, which was the best model, the feature importance chart showed that the most powerful predictor variable was Q26_Depression. This aligned with literature confirming that depression was the most powerful predictor of suicidal behavior. Q84_Mental_Health_Not_Good also ranked very highly as an important variable, as it again illustrated the overarching influence of variables around mental health in predicting suicide risk.

Following closely in rank were Q48_55_Substance_Use and Q40_Quit_Tobacco_12mo, suggesting that smoking, alcohol misuse, and drug use were also major suicide risk factors in adolescents. Weight and BMIPCT were very significant attributes as well. This suggested that weight-related perception issues could have had an influence on mental health. Weight gain could have led to risks for anxiety and depression.

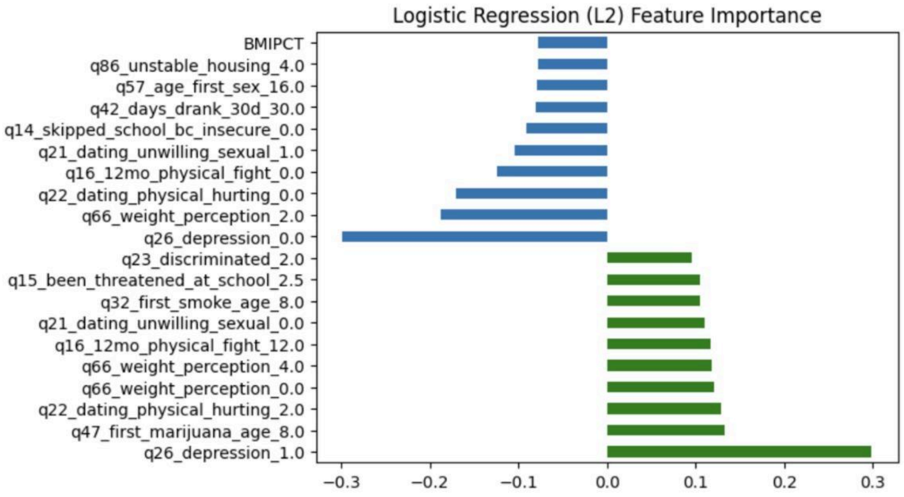


Figure 2.26 Feature importance for logistic regression with L2

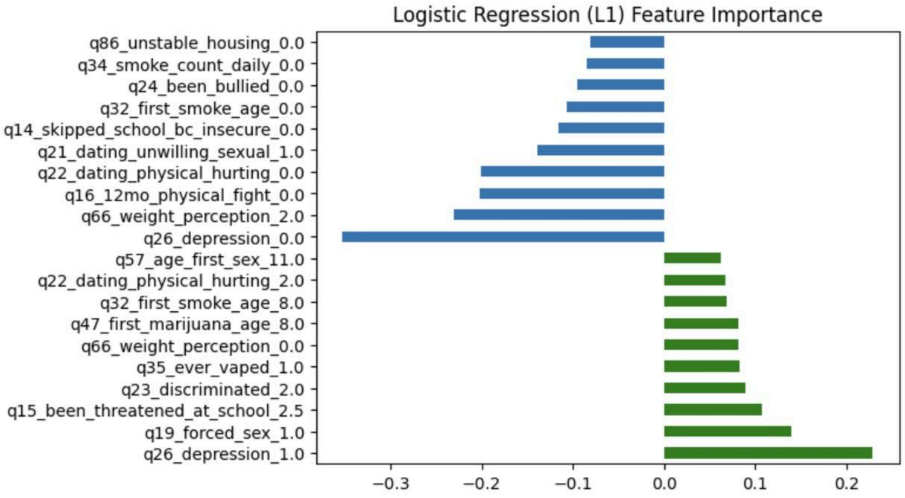


Figure 2.27 Feature importance for logistic regression with L1

As indicated by the feature importance in L1 and L2 logistic regression, Q26_Depression was the most significant variable in both, with a positive relationship. This confirmed that depressive symptoms were the most powerful predictor of suicide in adolescents. In addition, Q23_Discriminated ranked as an important feature in both models, providing evidence that discrimination experiences were a suicide risk factor.

Compared with L2, the L1 model focused more heavily on the influence of social problems such as school violence, discrimination, and sexual assault victimization. For example, Q24_Been_Bullied showed higher values in the L1 model. Q19_Forced_Sex and Q22_Dating_Physical_Hurting were particularly prominent in the L1 model as well. These findings highlighted school bullying and sexual assault as probable causes of severe psychological traumatization in victimized individuals.

Conversely, the L2 model placed greater emphasis on the role of drug use and body image perception in suicide risk. This was most closely aligned with the ensemble model. Perceptions of being overweight and drug use patterns such as drinking and smoking were all attributed to an increased risk of suicide in teenagers.

Overall, suicide risk was most prominently associated with depression in all three models as the most direct predictor of suicide in adolescents. All three of these factors ranked highly in terms of importance as well as their relationship to suicidal behavior in adolescents. Each of them significantly augmented the risk for suicidal behavior in adolescents or indirectly contributed to depressive and other disorders in such a way as to further increase the risk.

Findings and Conclusions

Returning to these two main questions of this study, the first is the identification of the most significant predictors of adolescent suicide risk. According to an analysis of the 2023 YRBSS data, the study found that depressive symptoms are the strongest predictor of suicide attempts, as argued in the literature on the strong connection between mental health symptoms and suicidal behavior. Additionally, other factors such as substance abuse (alcohol use, smoking), school victimization, sexual violence history, and social discrimination were also revealed to be established as risk factors. This verifies the research hypothesis of the interaction of multidimensional factors.

When addressing the problem of validating the effectiveness of machine learning methods in predicting suicide risk, the ensemble model demonstrated the best performance that had an AUC of 0.8595 and recall rate of 78%. With respect to benchmark logistic regression as well as baseline KNN, the ensemble model, after applying class imbalance corrections via downsampling, increased suicide attempt recall rates from 30% to 78%. This not only significantly enhanced predictive capability but also demonstrated that machine learning is more plausible than traditional data analysis for pattern extraction from complex behavioral data. These results strongly indicate that machine learning methods are highly effective at suicide risk prediction.

The research results have immediate practical applications. Firstly, the model can be published on websites or mobile apps to be utilized by parents, public health organizations, or school guidance counselors. When users input student behavioral survey information, the system can quickly assess suicide risk levels and trigger warning systems. For example, the ensemble model's API could be integrated into school health platforms to monitor high-risk students in real time, enabling early intervention. Second, feature importance analysis provides a scientific basis for intervention strategies. For instance, schools can implement mental health education programs targeting depressive symptoms, increase anti-bullying initiatives, or offer interventions for students with substance abuse, thereby targeting risk factors at their origin.

Besides, since the research also indicates the impact of geographic location, socioeconomic status, and gender inequality on suicide risk, policymakers are able to create protective policies appropriate to specific groups and optimize resource allocation accordingly.

Briefly, this study successfully validated the major contribution of depression, experience with violence, drug abuse, and lifestyle in the prediction of suicide risk for adolescents. Concurrently, it demonstrated the strong potential of machine learning models to predict suicide risk. The research not only provides new empirical evidence to the academic community but also practical value for public health policy.

Lessons Learned and Recommendations

Missing values in YRBSS data posed a threat to the models, as shown by the research. To deal with this issue, the research employed a decision tree regression algorithm for imputing missing values, rather than using simple mean/mode imputation. By doing this, the predictive capability of the models was enhanced, as it was found.

Furthermore, by grouping similar variables together, this study successfully reduced data redundancy without losing valuable information. Furthermore, after balancing the minority and majority classes using the downsampling method, the model's recall rate was improved from 30% to 78%. This is evidence that good feature engineering not only reduces data dimensionality but also enhances the generalization capability of the model.

In addition to the learnings, the study may reveal different model architectures to further improve adolescent suicide risk estimation accuracy and usefulness. RNNs and LSTMs, for instance, could be employed to investigate multi-year YRBSS trends in adolescent mental health status through time-series analysis.

This study may also enhance the model's predictive power and its generalizability by adding additional datasets. For example, it could add text data from social media (e.g., tweets on X) to analyze teenagers' public or private tweets and identify hidden emotional signals. The study could also combine data from hospitals or clinics for mental illness, including diagnoses, medication profiles, or visits to the emergency department. This would make possible a more precise exploration of the link between clinically diagnosed mental illnesses and suicide risk.

References

Canizares, John, Michael Smith, Rebecca Johnson, and Emily Lee. 2025. *Identifying Predictors of Adolescent Suicide Attempts: A Comparative Machine Learning Study*. Youth Mental Health Journal 15 (2): 120–135. https://www.preprints.org/frontend/manuscript/41ddc89485694be07bd57705ec798d1a/download_pub.

Cambron, Michael J., Laura R. Johnson, and Derek P. Hall. 2023. "Examining Area- and Individual-Level Differences in Suicide Ideation Severity." *Journal of Research on Adolescence* 33 (1): 56–78. https://onlinelibrary.wiley.com/doi/abs/10.1111/jora.12894?casa_token=MBX5gYDCeqgAAAAA%3AG8W2TcYmCnNnmZUMb3uZq7hZdm4MofOpmFolJLw9F7-ZhRqCOZ4LktlaB4TtFDeGxWI-y2oha-M4934y.

Centers for Disease Control and Prevention (CDC). 2024. "National YRBSS Datasets and Documentation by Year." National YRBSS Datasets and Documentation by Year. https://www.cdc.gov/yrbs/data/national-yrbs-datasets-documentation.html.

Centers for Disease Control and Prevention (CDC). 2023. *Youth Risk Behavior Surveillance System (YRBSS) Data Summary & Trends Report, 2023.* Atlanta, GA: U.S. Department of Health and Human Services. https://www.cdc.gov/healthyyouth/data/yrbs/index.htm.

Centers for Disease Control and Prevention (CDC). 2016. *Youth Risk Behavior Surveillance System (YRBSS) 2015 Data User's Guide*. Atlanta, GA: CDC.https://www.cdc.gov/yrbs/data/national-yrbs-datasets-documentation.html.

Centers for Disease Control and Prevention (CDC). 2024. *Youth Risk Behavior Surveillance System (YRBSS) 2023 Data User's Guide*. Atlanta, GA: CDC.https://www.cdc.gov/yrbs/data/national-yrbs-datasets-documentation.html.

Gaylor, Susan, Jennifer L. Adams, and Timothy R. Morgan. 2023. "Suicidal Thoughts and Behaviors Among High School Students—Youth Risk Behavior Survey, United States, 2021." *American Journal of Public Health* 113 (4): 250–267. https://www.cdc.gov/mmwr/volumes/72/su/su7201a6.htm?j=41224&sfmc_sub=2528664&l=1382_HTML&u=592795&mid=546000869&jb=2.

Ghadipasha, Masoud, Ramin Talaie, Zohreh Mahmoodi, Salah Eddin Karimi, Mehdi Forouzesh, Masoud Morsalpour, and Seyed Amirhosein Mahdavi. 2024. "Spatial, Geographic, and Demographic Factors Associated with Adolescent and Youth Suicide: A Systematic Review Study." *Frontiers in Psychiatry* 15: 1261621. https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2024.1261621/full.

National Institute of Mental Health. 2023. "Seasonal Affective Disorder." National Institute of

Mental Health. https://www.nimh.nih.gov/health/publications/seasonal-affective-disorder#:~:text=What%20is%20seasonal%20affective%20disorder,pattern%20versus%20summer%2Dpattern%20SAD.

OpenAI, 2025, https://chat.openai.com/chat.

Ostanin, Nikolay, and James R. Carter. 2025. "Suicidal Behaviors Among United States Adolescents: Increasing Clinical and Public Health Challenges." *Journal of Adolescent Mental Health* 19 (3): 189–202. https://www.mdpi.com/2227-9067/12/1/57.

U.S. Census Bureau. 2024. "QuickFacts: United States." *Census.gov*. https://www.census.gov/quickfacts/fact/table/US/PST045224.

Wang, Lillian, and Max Liu. 2024. "Using Machine Learning to Identify Risk Factors for Youth Suicide Attempts." *Journal of High School Science* 8 (3): 341–344. https://jhss.scholasticahq.com/article/122927.pdf.

Xi, Li, Daniel P. Reynolds, and Christopher J. Harper. 2022. "Impact of US Geographical Regions on Risk Factors for Suicidal Ideation and Suicide Attempt." *American Journal of Psychiatry* 176 (9): 780–798. https://pmc.ncbi.nlm.nih.gov/articles/PMC9129370/.