Journal of High School Science

Using Machine Learning to identify risk factors for youth suicide attempts

Wang L[1], Liu M[2]

Submitted: July 8, 2024, Revised: version 1, August 18, 2024, version 2, August 20, 2024
Accepted: August 21, 2024

Abstract

Suicide has become the second-leading cause of death among adolescents in the U.S. since the COVID-19 pandemic, with 9% of high schoolers having attempted suicide. This study used machine learning to identify the most predictive risk factors for suicide attempts among American high schoolers from 2021 data, and also to understand if one machine learning model could be better than another. The Center for Disease Control and Prevention (CDC) conducted the 2021 Youth Risk Behavior Survey (YRBS) and received 17,232 survey responses. Using the survey results, this study compared three machine learning models (LightGBM, Logistic Regression and Random Forest) for their accuracy in predicting suicide attempts, and identified the top risk factors. The most accurate predictive model methodology was LightGBM, based on the Area Under the Curve (AUC) value, Area Under the Precision-Recall Curve (AUPRC) value, Accuracy, and the Kolmogorov-Smirnov (K-S) statistic. The leading predictors identified by the LightGBM model included excess body fat, sleep disturbances, Sars-CoV2 infection, consumption of alcohol, and physical inactivity among others. Identifying these risk factors enables the development of targeted prevention strategies, which include increasing physical activity and maintaining routine sleeping habits. Interventions should also incorporate strategies in promoting the mental wellness of adolescents. The results underscore the value of machine learning in predicting suicide attempts based on risk behaviors among high schoolers.

Keywords

Youth suicide, Risk factors, Predictors, Machine Learning, LightGBM, Random Forest, Hyperparameter tuning, Logistic Regression, Adolescents, Youth Risk Behavior Survey

[1]Corresponding author: Lillian Wang, Adlai E. Stevenson High School, 1 Stevenson Dr, Lincolnshire, IL 60069, USA. lwang25@students.d125.org
[2]Max Liu, Adlai E. Stevenson High School, 1 Stevenson Dr, Lincolnshire, IL 60069, USA. mliu25@students.d125.org

Introduction

Mental health issues have become increasingly prevalent worldwide, especially among teenagers. Approximately 5 million adolescents between the ages of 12 and 17 in the United States have experienced at least one major depressive episode, constituting 20.1% of the adolescent population (1). Consequently, suicide has become the second-leading cause of death among adolescents in the U.S., with 9% of high schoolers having attempted suicide (2).

Previous studies of suicide attempts for high schoolers identified many behavioral patterns, along with psychological and environmental factors. The purpose of this paper is to apply machine learning to identify the risk factors that are most predictive of suicide attempts among American high schoolers, compare and select the most predictive machine learning model, and identify prevention strategies to reduce the overall rate of suicide attempts. The purpose of the study is not to build a model accurate for future datasets. Rather, it is to understand if one machine learning model could be better than another.

Literature review

Suicidal attempts among high schoolers can be influenced by numerous risk factors, including substance abuse, social isolation, traumatic life events, psychosocial stressors, cultural factors, etc. (3). Sleep patterns, physical activity and perceptions of school safety, including experiences of bullying and the presence of weapons and drugs on campus, are evidenced as risk factors by a national study involving adolescents (4). Early sexual activity has a positive correlation with adverse mental health outcomes, such as increased suicidal thoughts. This type of activity can lead to feelings of sexual guilt, concerns about early pregnancy, and the risk of experiencing violent victimization (5,6). For male adolescents, their sexual orientation directly correlates with suicide attempts, regardless of other variables. However, in female adolescents, the relationship between sexual orientation and suicidal inclinations may be affected by drug consumption, as well as encounters with violence and victimization (7).

Previous studies of suicide attempts for high schoolers have reported substance abuse (e.g., cigarette smoking, using vapor products, and illegal drug abuse) as one of the most common risk factors. Furthermore, in 2019, 48.7% of high school students who reported alcohol use additionally reported themselves previously considering committing suicide (8). Early or frequent alcohol consumption is linked to youth suicidal behaviors (9). Social connections (e.g., family closeness and connection to friends) also affect vulnerability to suicide attempts, due to feeling dislocated from their safe spaces (10). Recent research indicates that homeless youth are particularly vulnerable to severe suicidal thoughts (11). A recent analysis of literature found that mental health declined in tandem with increased social distancing measures during the COVID-19 pandemic quarantine (12).

A parallel investigation into YRBS using 2019 data found that individuals classified as obese had 1.65 times higher odds of attempting suicide compared to those who were classified as non-obese (13). Multiple research findings indicated that exercising exhibited a positive effect on alleviating depressive symptoms

among individuals, significantly reducing the frequency of suicide attempts (14).

Past literature provided insights into adolescent suicide risk factors, but ranking the most predictive factors requires a quantitative analysis. By applying multiple machine learning methodologies, we aimed to select the most predictive machine learning model, rank predictive factors, and propose effective early prevention strategies for high schooler suicide ideators.

Methods

Study data

The Study data for machine learning was extracted from the 2021 Youth Risk Behavior Survey (YRBS) (15). YRBS was collected from a nationally representative sample of 17,232 American students in grades 9 through 12, from 209 systematically selected, public, Catholic, and other private schools in the 50 States and the District of Columbia. The overall response rate (school response rate × student response rate) was 57.5% (8). Most of the 2021 survey questions have demonstrated a high degree of test-retest reliability (16).

Data processing and data reduction

Our literature review indicated that twelve themes of high school suicide risk factors were relevant for this study as outlined in Figure 1. The 2021 YRBS contained 99 survey questions, many of which were highly correlated and not suitable for regression modeling.

Theme #	Theme Topic	Survey Questions
Theme 1	Unsafe school environment, Physical fighting, Sexual assault, Bullying	14-24
Theme 2	Smoking	30-39
Theme 3	Alcohol	40-44
Theme 4	Drug Abuse	5-49, 51-56, 88-89
Theme 5	Sexual Behavior	57-63
Theme 6	Physical Activity	77-80, 92
Theme 7	COVID-related Impacts	93-94
Theme 8	Closeness with Friends and Family	96-97
Theme 9	Sexual Identity	64-65
Theme 10	Sleep	86
Theme 11	Homelessness	87
Theme 12	BMI	1,2,6,7

Figure 1. Twelve themes for high schoolers' suicide factors

These 99 survey questions were reduced to 12 questions, each corresponding to one of twelve risk factor themes. A pairwise Pearson correlation test was used to identify highly correlated questions among the 12 themes. Based on the high correlation we eliminated 2 variables, leaving 10 variables representing 10 themes. The details of this process are explained in Step 3 in the Machine Learning Procedure Section.

These 99 survey questions were reduced to 12 questions, one question per theme. For each theme, we used Linear Regression to find which of the theme questions exhibited the highest slope with respect to suicide attempts.

This identified the question in the theme that best predicted suicide attempts. For example, Theme 5, Sexual Behavior Theme, had 7 questions (Q57-Q63) from the survey. Figure 2 presents the correlation matrix from the 7 questions. Several variables, such as Q57 and Q61, for example, were highly correlated with a Pearson correlation coefficient of 0.90. We conducted linear regressions between each of these 7 variables and the target variable. The results of 7 linear regressions are shown in Figure 3.

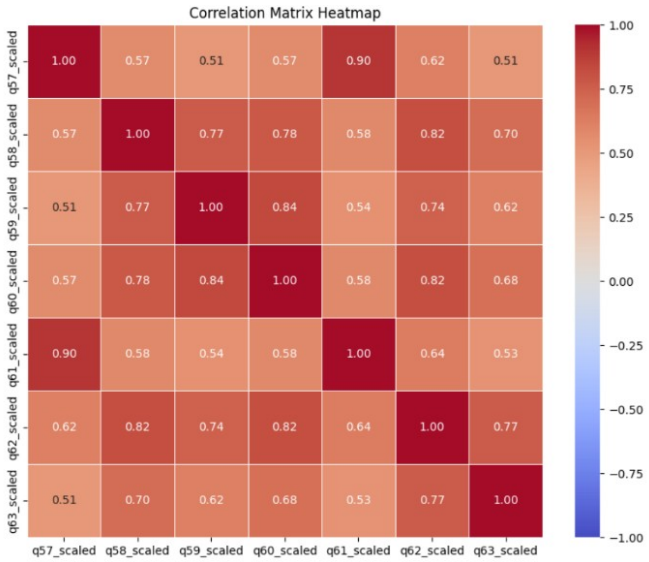

Figure 2. Theme 5 pairwise Pearson correlation results

We chose survey question q60_scaled to represent Theme 5 because the absolute value of its regression coefficient was the largest as shown in Figure 3. The same process was repeated for the 9 themes which constituted multiple questions in each theme. We hence obtained 9 questions to represent the 9 of the 12 themes. Themes 10 and 11 were represented with single-question themes which did not require correlation or data reduction. Theme 12 combined height and weight to compute BMI with the following formula shown in Figure 4.

Survey Questions	Coefficient	P-value	Conf_Lower	Conf_Upper	Intercept	Intercept_P-value
q60_scaled	0.2571	0.0000	0.2304	0.2837	0.0554	0.0000
q58_scaled	0.2437	0.0000	0.2273	0.2600	0.0454	0.0000
q59_scaled	0.2052	0.0000	0.1868	0.2235	0.0569	0.0000
q63_scaled	0.1617	0.0000	0.1421	0.1814	0.0509	0.0000
q62_scaled	0.1367	0.0000	0.1231	0.1504	0.0422	0.0000
q61_scaled	0.1302	0.0000	0.1163	0.1441	0.0601	0.0000
q57_scaled	0.0802	0.0000	0.0697	0.0907	0.0643	0.0000

Figure 3. Theme 5 linear regression results

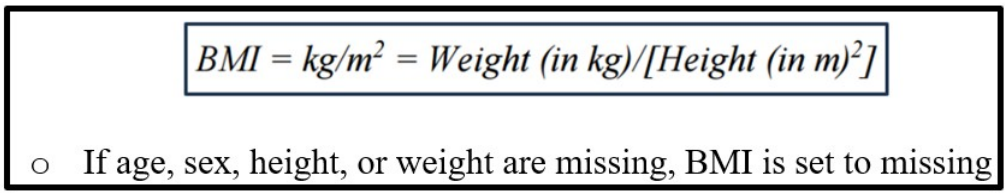

Figure 4. BMI calculation formula

Model variables

The model target variable was created using Question 28; "During the past 12 months, how many times did you actually attempt suicide?" The answer choices were, A. 0 times, B. 1 time, C. 2 or 3 times, D. 4 or 5 times, E. 6 or more times. The response "A. 0 times" to survey question 28 was classified as 0. The respondents who responded with choices "B" through "E" were classified as 1. The independent variables used in the machine learning models were the 10 questions selected from the variable reduction steps. The 10 questions were determined in Step 3 in the Machine Learning Procedure Section. Our modeling sample contained 17,232 observations, of which 1,753 were suicide attempts, thus our model target rate was 10.2%.

Model methodology

There are many machine learning models available to analyze the data. We aimed to balance interpretability, robustness, and predictive accuracy, ensuring that we captured both linear and non-linear patterns in our imbalanced dataset. For example, we needed to rank the feature importance in this study. Neural networks are difficult to explain and interpret, and do not explicitly output feature importance. Thus we rejected neural networks for this study.

The following three machine learning models were chosen for further study: Logistic Regression, Random Forest, and Light Gradient Boosting Machine (LightGBM). Each methodology has its own strengths and weaknesses, making them complementary in many scenarios.

Logistic Regression is a model usually chosen for its simplicity and interpretability (17). It works well when the relationship between the features and the target variable is approximately linear. Logistic Regression is also computationally efficient and provides a clear interpretation of the output, which can be useful for understanding the influence of individual features. Random Forest is an ensemble learning method which is powerful for capturing non-linear relationships and feature interactions (18). Random Forest is robust to overfitting due to its use of multiple decision trees and averaging of their results. It handles a large number of features well and can provide insights into feature importance, making it a good choice for various data types and complexities. LightGBM is a gradient boosting framework which is highly efficient and scalable, making it suitable for large datasets and high-dimensional data. LightGBM excels in handling non-linear relationships and complex feature interactions (18). It also offers advantages in terms of speed and memory usage. The gradient boosting machine family has GBM, LightGBM and XGBoosting. They are similar in methodology (19) and we

arbitrarily chose LightGBM due to its faster runtime.

The performance of each machine learning methodology depends on factors such as the type of data, data quality, feature engineering, model validation strategy, evaluation metrics, and proper hyperparameter tuning. In our study, we employed stratified k-fold cross-validation for model validation, alongside AUPRC, AUC, and K-S statistic as evaluation metrics.

Our full model dataset was imbalanced, with a target rate of 10.2%, compared to the ideal 50% in balanced data. Given this imbalance, stratified k-fold cross-validation was chosen because it ensured that each fold maintained the same class distribution as the original dataset. This approach is recommended for imbalanced datasets like ours, where the target rate is ~10%. It provides a more reliable evaluation of the model's performance by preserving class balance across folds, leading to more consistent and accurate assessment compared to a one-fold train-test split, which sometimes may not adequately address the class imbalance (20). Our 3-fold cross validation was performed with stratification which kept both the training and test sets at a similar target rate of 10.2%.

	Training					Test					
Fold	Training%	Training_FoldSize	Target=1	Target=0	Target Rate	Test%	Test_FoldSize	Target=1	Target=0	Target Rate	Training+Test
1	66.67%	11,488	1,168	10,320	10.17%	33.33%	5,744	585	5,159	10.18%	17,232
2	66.67%	11,488	1,169	10,319	10.18%	33.33%	5,744	584	5,160	10.17%	17,232
3	66.67%	11,488	1,169	10,319	10.18%	33.33%	5,744	584	5,160	10.17%	17,232

Figure 5. Cross validation training + test fold size

Python Code:

```
cv = StratifiedKFold(n_splits=3, shuffle=True, random_state=42)
for train_index, test_index in cv.split(X, y):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
    print(f"Training data shape: {X_train.shape}, Test data shape: {X_test.shape}")
```

Python Log:

```
Training data shape: (11488, 10), Test data shape: (5744, 10)
[LightGBM] [Info] Number of positive: 1168, number of negative: 10320 in training
[LightGBM] [Info] Number of positive:  585, number of negative:  5159 in test

Training data shape: (11488, 10), Test data shape: (5744, 10)
[LightGBM] [Info] Number of positive: 1169, number of negative: 10319
[LightGBM] [Info] Number of positive:  584, number of negative:  5160 in test

Training data shape: (11488, 10), Test data shape: (5744, 10)
[LightGBM] [Info] Number of positive: 1169, number of negative: 10319
[LightGBM] [Info] Number of positive:  584, number of negative:  5160 in test
```

Figure 6. Python code and log to create 3-fold cross validation

Figure 5 shows the 3-fold cross validation datasets. In each fold, a model is trained on 66.7% of data and evaluated on the withheld 33.3%. The final evaluation is the average

performance of all folds. Figure 6 presents the Python code and log for our LightGBM model 3-fold cross validation. The same process was used for all the three machine learning methodologies.

Machine learning procedure

Figure 7 presents the overall project flow chart, which are introduced below as 8 steps.

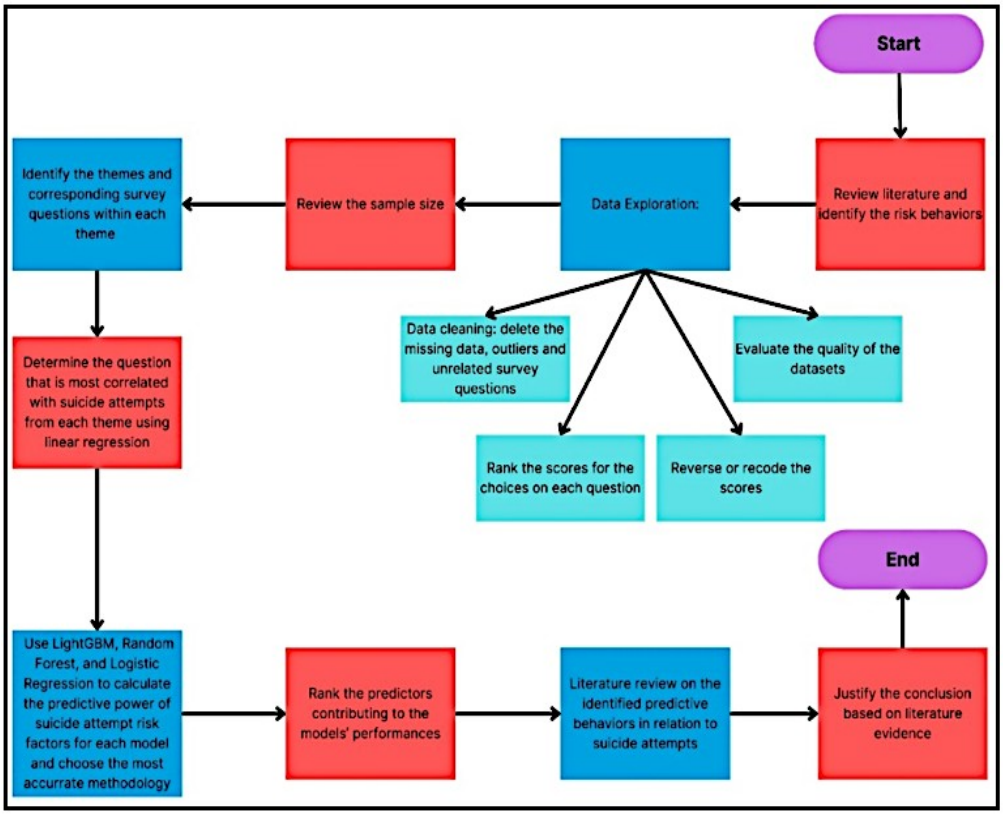

Figure 7. Project flowchart

In STEP 1, Exploratory Data Analysis (EDA) was performed on the risk behaviors of interest in the survey data. This encompassed a] Cleaning up the data (including processing the missing data, reversing or recoding the scores), b] Categorizing the risk behaviors into twelve themes, c] Calculating the score sum/mean/percentage for each risk behavior variable based on the risk factor themes, d] Standardizing the data variables.

Within STEP 2, Figure 8 shows the results of the variable reduction step to select 12 variables that best predicted suicide attempts from the 99 survey questions. The BMI value was calculated from 4 survey questions for theme 12.

Theme	Description	The Best Survey Question to Represent the Theme
1	Unsafe school environment, Physical fighting, Sexual assault, Bullying (14-24)	Q15: During the past 12 months, how many times has someone threatened or injured you with a weapon such as a gun, knife, or club on school property?
2	Smoking (30-39)	Q32: During the past 30 days, on how many days did you smoke cigarettes?
3	Alcohol (40-44)	Q41: During the past 30 days, on how many days did you have at least one drink of alcohol?
4	Drug Abuse (45-49, 51-56, 88-89)	Q54: During your life, how many times have you used ecstasy (also called MDMA or Molly)?
5	Sexual Behavior (57-63)	Q60: During the past 3 months, with how many people did you have sexual intercourse?
6	Physical Activity (77-80, 92)	Q77: During the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day? (Add up all the time you spent in any kind of physical activity that increased your heart rate and made you breathe hard some of the time.)
7	COVID-related Impacts (93-94)	Q93: During the COVID-19 pandemic, how often was your mental health not good? (Poor mental health includes stress, anxiety, and depression.)
8	Closeness with Friends and Family (96-97)	Q97: How often do your parents or other adults in your family know where you are going or with whom you will be?
9	Sexual Identity (64-65)	Q65: Which of the following best describes you? Straight, Gay or Lesbian, Bisexual, I describe my sexual identity some other way .....
10	Sleep (86)	Q86: On an average school night, how many hours of sleep do you get?
11	Homelessness (87)	Q87: During the past 30 days, where did you usually sleep?
12	Body Mass Index (1, 2, 6, 7)	Calculated using BMI formula in Figure 3

Figure 8. Risk factor theme questions and associated descriptions

In STEP 3, we further reduced the 12 theme variables to 10 to eliminate multicollinearity. Using 0.55 correlation as the cut-off, we found two pairs with high correlation, as shown in Figure 9, and decided to remove one variable from each pair. From between Q54 *vs* Q87 with correlation of 0.61, we selected Q54. Between Q93 *vs* Q97 with correlation of 0.57, we selected Q93. The variable we selected showed a higher Pearson correlation with the target variable.

In STEP 4, the following machine learning models were run on the full study population: Logistic regression; Random Forest (with hyperparameter tuning); LightGBM (with hyperparameter tuning), from which our model feature importance originated. This method of modeling the full data and evaluating with k-folds has been used successfully by other statisticians (21).

In STEP 5, the machine learning models were run with stratified 3-fold cross validation strategies to explore the varied effects of distinct risk behaviors on suicidal attempts through three methodologies. Stratified Cross-fold validation is vital as it evaluates the model's generalizability in imbalanced data, mitigates overfitting risks, and generates performance estimates across diverse data subsets. Each fold built a model on a 67% training split and evaluated on a withheld 33% test split.

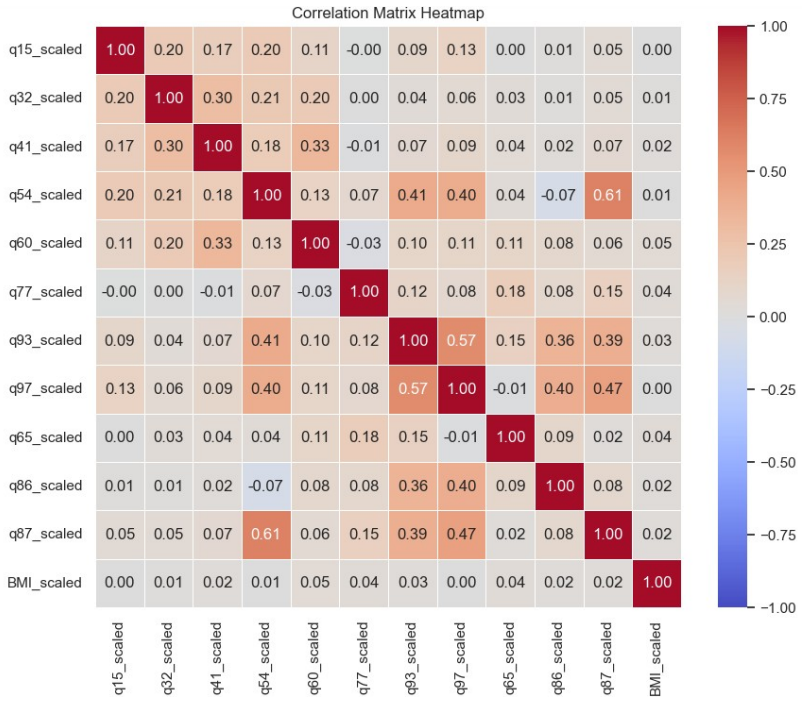

Figure 9. Pearson correlation among 12 variables from Figure 8

In STEP 6, The best model was chosen based on each model's AUC value, AUPRC value, K-S statistic, and Accuracy calculated from the Confusion Matrix (22). The AUC, or the "Area under the ROC Curve" measures the area underneath the receiving operating characteristic curve (ROC). The ROC Curve plots two parameters: True Positive rate and False Positive rate. The larger the AUC value the better the target prediction. For that reason, we compared the AUC value for each model with the best model possessing the highest AUC value. The AUPRC stands for the "Area under the Precision – Recall Curve". It is a performance metric used to evaluate binary classification models, particularly in cases where the class distribution is imbalanced. The larger the AUPRC value the better the target prediction. For that reason, we compared the AUPRC value for each model with the best model possessing the highest AUPRC value. The K-S statistic; the Kolmogorov-Smirnov statistic; is the maximum difference between the cumulative distribution of the predicted probabilities for the positive class and the cumulative distribution of the predicted probabilities for the negative class (23). Mathematically, it is defined as:

$$\text{K-S Statistic} = \max_x |\text{FPR}(x) - \text{TPR}(x)|$$

Where x is a value on the horizontal axis which represents the fraction of total data population, FPR(x), also known as 1-Specificity, is the cumulative proportion of the negative class (in our study non-suicide-attempts), TPR(x), also known as Sensitivity, is the cumulative proportion of the positive class (in our study suicide attempts). The K-S statistic indicates the maximum separation between the two distributions. The larger the K-S statistic the better is the target prediction (Figure 10). For that reason, we compared the K-S statistic for each model with the best model possessing the highest K-S statistic (24).

Kolmogorov-Smirnov Statistic	Model Predictiveness
KS statistic < 0.2	Poor
$0.2 \leq$ KS statistic < 0.4	Fair
$0.4 \leq$ KS statistic < 0.6	Good
KS statistic ≥ 0.6	Excellent

Figure 10. K-S statistic *vs.* model predictiveness

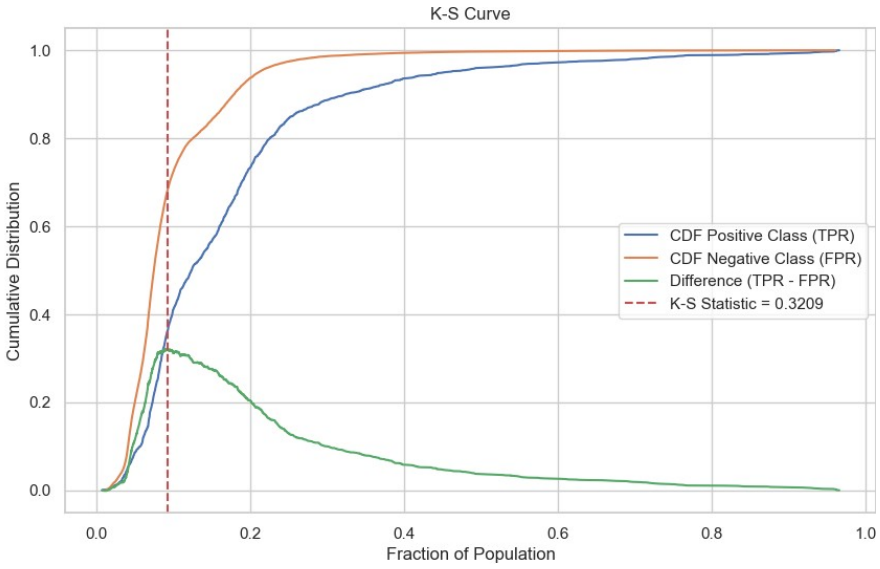

Figure 11. K-S curve for Logistic Regression model

The K-S Curve is visual representation of the K-S statsistic. Our logistic regression model K-S curve is shown in Figure 11. The Confusion Matrix is a performance measurement for machine learning and is useful for measuring accuracy, recall, and specificity. Figure 12 shows a Confusion Matrix for Binary Classification. A False Positive is also called a Type I Error, where the model predicts a positive outcome, when in reality, the outcome is negative. A False Negative is called a Type II Error, where the model predicts a negative outcome, when in reality, the outcome is positive (22).

From the Confusion Matrix we computed the accuracy as shown as Figure 13. The larger the accuracy the better the target prediction. For that reason, we compared accuracy for each model with the best model possessing the highest accuracy value.

		Predicted Class	
Actual Class		True Positive (TP)	False Negative (FN)
		False Positive (FP)	True Negative (TN)

Figure 12. Confusion matrix definition

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

Figure 13. Statistical model accuracy formula

In STEP 7, The model findings were compared with the existing literature, which also included the BMI post-analysis. In STEP 8, a discussion of the study's strength and weakness was presented.

Results

Models could not be produced using all correlated 99 survey questions

The following model logs in Figure 14 show the results of sending all 99 survey questions directly into machine learning methodologies, indicating that a valid model could not be produced. To mitigate this problem, the 99 survey questions were reduced into 10 risk factor themes prior to running the regression.

Logistic Regression *on the full population using 99 survey questions*

Current function value: inf

LinAlgError: Singular matrix

Random Forest model *on full population using 99 survey questions*

Mean Squared Error: 0.0

Model AUC: 1.0 (in reality it is impossible to have a perfect model)

LightGBM Model *on full population using 99 survey questions*

Mean Squared Error: 0.0

Model AUC: 1.0 (in reality it is impossible to have a perfect model)

Figure 14. Model log shows models cannot be solved with all 99 raw survey questions

LightGBM is the most predictive model for the study population

Three machine learning methodologies (i.e. Logistic Regression, LightGBM, and Random Forest) were performed separately, each with 3-fold cross-validation and basic-level hyper-parameter tuning. Based on the AUC values, AUPRC values, Accuracy, and K-S statistic, LightGBM was selected as the best model methodology for this study. Figure 15 presents the details of the model results.

Model Types	Run Time	Accuracy	K-S Score	Area Under the Curve (AUC)	Area Under the Precision-Recall Curve (AUPRC)
Logistic Regression	0.81s	0.8995	0.3209	0.7091	0.2670
Random Forest	90s	0.8822	0.3881	0.7581	0.2723
LightGBM	677s	0.9138	0.4437	0.7984	0.3366

Figure 15. Comparing three model methodologies using 3-fold cross-validation average

Logistic Regression model results

Logistic Regression, developed in the early 20th century, is a statistical model used for binary classification that predicts the probability of a binary outcome based on one or more predictor variables. When all 10 variables were input into the Logistic Regression model, two variables exhibited insignificant P-values using 0.05 as the cut-off (Figure 16) (25). Thus, we removed Q15 (Unsafe school environment, etc.) and Q60 (Sexual Behavior), and ran the logistic regression on the remaining 8 significant variables. Figures 17 and 18 show the Logistic Regression model results on these 8 variables.

Logistic Regression Model Predictor Candidates		P-values
Alcohol	q41_scaled	0.0017
Body Mass Index	BMI_scaled	0.0000
COVID-related Impacts	q93_scaled	0.0000
Drug Abuse	q54_scaled	0.0000
Physical Activity	q77_scaled	0.0000
~~Sexual Behavior~~	~~q60_scaled~~	~~0.2477~~
Sexual Identity	q65_scaled	0.0000
Sleep	q86_scaled	0.0000
Smoking	q32_scaled	0.0140
~~Unsafe school environment, Physical fighting, Sexual assault, Bullying~~	~~q15_scaled~~	~~0.1514~~

Figure 16. Initial Logistic Regression run model coefficients P-Values

Logistic Regression Model		Coefficient	P-Value	Feature Importance
Risk Factors	Intercept	-4.4614		
Smoking	q32_scaled	1.8577	0.0359	1
Alcohol	q41_scaled	1.8364	0.0001	2
Drug Abuse	q54_scaled	1.7901	0.0000	3
Sexual Identity	q65_scaled	1.7771	0.0000	4
Body Mass Index	BMI_scaled	0.5047	0.0000	5
Sleep	q86_scaled	-0.4571	0.0000	6
COVID-related Impacts	q93_scaled	0.4403	0.0000	7
Physical Activity	q77_scaled	0.3384	0.0000	8

Figure 17. Logistic Regression model feature importance

Model Predictor Variable P-Values:

Smoking	0.0359
Alcohol	0.0001
Drug Abuse	<0.0001
Physical Activity	<0.0001
COVID-related Impacts	<0.0001
Sexual Identity	<0.0001
Sleep	<0.0001
BMI	<0.0001

A p-value that is less than 0.05 indicates that there is a relationship between the predictor variable and the target variable

Cross Validation Mean AUC=0.7091
Cross Validation Mean AUPRC = 0.2670
Execution Time: 0.81 seconds

Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	15,430	49
Actual Positive	1,682	71

There were 15,430 true negatives (TN), 71 true positives (TP), 49 false positives (FP), and 1,682 false negatives (FN). The accuracy calculated by (TP + TN)/ (TP+TN+FP+FN) * 100% = 89.95%

Logistic Regression K-S Statistic: 0.3209
The Logistic Regression model fit is fair since the K-S statistic is between 0.2 and 0.4

Figure 18. Logistic Regression model evaluation statistics

LightGBM model results

LightGBM, created by Microsoft in 2016, is a gradient boosting framework that uses tree-based learning algorithms, designed for efficiency and speed, especially with large datasets (26). Figures 19 & 20 show our LightGBM model results including the model feature importance.

Cross-Validation AUC (Area Under the Curve) Scores:

AVERAGE AUC:
0.7984

Cross-Validation AUPRC (Area Under the Precision-Recall Curve) Scores:

AVERAGE AUPRC:
0.3366

LightGBM Full Population Model Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	15,395	84
Actual Positive	1,402	351

There are 15,395 true negatives (TN), 351 true positives (TP), 84 false positives (FP), and 1,402 false negatives (FN). The accuracy calculated by (TP + TN)/(TP+TN+FP+FN) * 100% = 91.38%

K-S Statistic: 0.4437
Execution Time: 90 seconds

The LightGBM Model was fitted Good since the K-S statistic is between 0.2 and 0.4

Figure 19. LightGBM model evaluation statistics

The top five predictors of high school suicide attempts as returned by the LightGBM model were high body mass index, sleep disturbances, COVID Impact, alcohol consumption at an early age and a low level of physical activity.

LightGBM models require hyperparameter tuning to optimize model performance (27). Hyperparameters are parameters that control the machine learning process, but are not model independent variables (28). A pre-set grid of hyperparameter values was input to evaluate the LightGBM model's performance using cross-validation for all combinations of hyperparameter values. The hyperparameter

tuning that was used for the current stage LightGBM model involved, a] Max_depth: the maximum depth of each decision tree in the gradient-boosting machine. The Best max_depth returned by the grid search process was 5. b] Min_child_sample: minimum number of data points required to be in a leaf node. Best Min_child_sample returned by the grid search process was 7. c] N_estimators: the number of decision trees to be used in the gradient-boosting machine. Best n_estimators returned by the grid search process was 100.

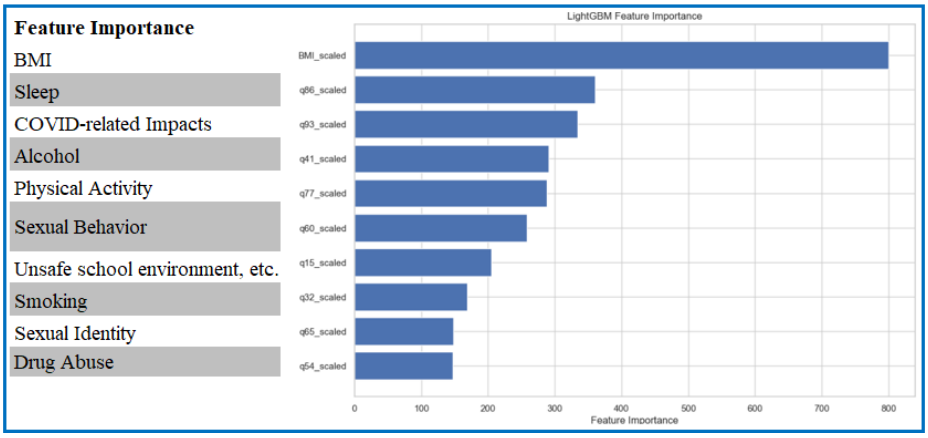

Figure 20. LightGBM model feature importance

Random Forest model results

Random Forest, introduced by Leo Breiman in 2001, is a machine learning method that constructs multiple decision trees during training and outputs the mode of their predictions for classification tasks (29).

Cross-Validation AUC (Area Under the Curve) Scores:

AVERAGE AUC:
0.7581

Cross-Validation AUPRC (Area Under the Precision-Recall Curve) Scores:

AVERAGE AUPRC:
0.2723

Random Forest Confusion Matrix:

	Predicted Negative	Predicted Positive
Actual Negative	15,060	419
Actual Positive	1,611	142

There are 15,060 true negatives (TN), 142 true positives (TP), 419 false positives (FP), and 1,611 false negatives (FN). The accuracy calculated by (TP + TN)/(TP+TN+FP+FN) * 100% = 88.22%

K-S Statistic: 0.3881
Execution Time: 677 seconds

The Random Forest Model was fitted Fair since the K-S statistic is between 0.2 and 0.4

Figure 21. Random Forest model evaluation statistics

Body Mass Index (BMI)	BMI_scaled	0.3759
Physical Activity	q77_scaled	0.1293
Sleep	q86_scaled	0.0934
Alcohol	q41_scaled	0.0758
Sexual Behavior	q60_scaled	0.0753
COVID-related Impacts	q93_scaled	0.0716
Unsafe school environment, etc.	q15_scaled	0.0531
Sexual Identity	q65_scaled	0.0457
Smoking	q32_scaled	0.0424
Drug Abuse	q54_scaled	0.0376

Figure 22. Random Forest feature importance

Figures 21 and 22 show our Random Forest model results including the model feature importance. The top five predictors of high school suicide attempts for the Random Forest model were high body mass index, less level of physical activity, sleep disturbances, alcohol consumption at an early age and early age sexual engagement.

Random Forest models also require hyperparameter tuning to optimize model performance. All pre-set hyperparameter values were searched through to evaluate the model's performance (30) for each combination of hyperparameters. The hyperparameters tuned during the current stage study were, a] Max_depth: the maximum depth of each decision tree in the Random Forest. Best max_depth returned by the grid search process was 10. b] Min_samples_split: the minimum number of samples required to split a tree node. Best min_samples_split returned by the grid search process was 8. c] N_estimators: the number of decision trees to be used in the Random Forest. Best n_estimators returned by the grid search process was 300.

Selection of the best model: LightGBM

The LightGBM model returned the highest AUC value, AUPRC value, accuracy and K-S statistic, indicating that it consistently performed better when compared to the Random Forest and Logistic Regression models.

BMI was one of the highest ranked risk factors for predicting suicide attempts. A high BMI is usually indicative of high body fat. BMI was the top predictor in the both the LightGBM and Random Forest models. Following this ranking, the following 5-step post-model BMI statistics analysis was performed (31).

In Step 1, a histogram of BMI values from the YRBS survey data was prepared and the BMI value ranges classification was retrieved (Figure 23). In Step 2, the suicide rates for the two BMI classification groups were calculated (Figure 24).

BMI	Weight
Below 18.5	Underweight
18.5 – 24.9	Normal Weight
25.0 – 29.9	Overweight
30.0 and Above	Obesity

Classification of BMI Ranges into Weight Status Categories

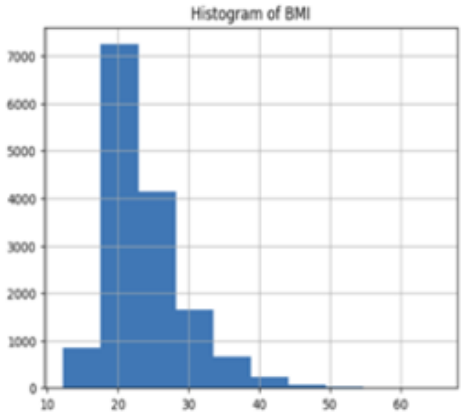

Figure 23. BMI histogram and classification

BMI Classification	# of Survey Respondents	# of Respondents with Suicide Attempts	Suicide Rate
Normal	10,391	935	9.00%
overWeight+Obese	4,505	543	12.05%

Figure 24. Suicide attempt rates for BMI classification

In Step 3, a two-sample t-test for proportions was performed (31). The Null Hypothesis (H0), was formulated as; 'There is no significant difference in the proportions of respondents with suicide attempts between the "Normal" and "Overweight + Obese" groups'. The Alternative Hypothesis (H1), was formulated as; 'There is a significant difference in the proportions of respondents with suicide attempts between the "Normal" and "Overweight + Obese" groups.'

The Sample Proportions (P) were calculated using the variables X1 = # of respondents with suicide attempts in the "Normal" group: 935; X2 = # of respondents with suicide attempts in the "Overweight + Obese" group: 543; N1 = Total number of respondents in the "Normal" group:10,391; N2 = Total number of respondents in the "Overweight + Obese" group: 4,505; P1=X1/N1=935/10391=0.090; P2=X2/N2=543/4505=0.1205.

$$SE = \sqrt{\frac{\hat{p}_1 \times (1-\hat{p}_1)}{N1} + \frac{\hat{p}_2 \times (1-\hat{p}_2)}{N2}}$$

$$SE = \sqrt{\frac{0.090 \times 0.91}{10391} + \frac{0.1205 \times 0.8795}{4505}}$$

$$SE \approx \sqrt{0.00000765 + 0.00003233} \approx \sqrt{0.00004098} \approx 0.0064$$

Figure 25. Standard error calculation in t-test

The standard error, SE was calculated (Figure 25) as was the t-statistic, as t = (p1-p2)/SE = (0.090-0.1205) /0.0064 = -0.0305/0.0064 = -4.77. In Step 4, The Degrees of Freedom (df), were calculated as, df = N1+N2-2 = 10391+4505-2 = 14894. With a t-statistic of -4.77, and 14,894 degrees of freedom, the two-tailed p-value was close to 0. Since the p-value < 0.05, the null hypothesis was rejected. In Step 5, we concluded that the above two-sample t-test found a significant difference in suicide attempt rates between high school students categorized as "Normal" (9%) and "Overweight + Obese" (12%), indicating that there were significantly more attempts by students in the overweight group to commit suicide than those in the normal weight group.

Discussion

Risk factor ranking for suicidal attempts

The top 5 predictor risk factor themes as output by the LightGBM model in order of model performance were, a] Theme 12: BMI, body mass index (obesity), b] Theme 10: Sleep (excess sleep hours), c] Theme 7: COVID – related Impacts (stress, anxiety, depression), d] Theme 3: Alcohol (early age usage), e] Theme 6: Physical Activity (number of hours)

The risk factors identified in the study are consistent with the findings reported in the literature studies investigating suicide attempts in youth. The model indicated that high schoolers who are obese are more likely to attempt suicide. This finding is consistent with the odds of suicide attempts being 1.65 times higher for high schoolers classified in the obese category (13). The model found that high schoolers who have disordered sleep patterns, especially excess sleep hours are more likely to attempt suicide. This finding is also consistent with a national study involving adolescents (4).

The model further found that poor mental health (stress, anxiety and depression) during the COVID-19 pandemic was associated with an increased risk of suicide attempts among high schoolers. Literature reviews found that a person's mental health deteriorated after they engaged in social-distancing during the COVID-19 quarantine (12). The model also found that high schoolers who used alcohol at an early age were more likely to attempt suicide. The published literature states alcohol use (9) as a common risk factor for individuals considering suicide. The model found that high schoolers who were less physically active were more likely to attempt suicide. This finding is consistent with a national study involving adolescents (4).

Further study using machine learning models

The models were run using the full population and 3-fold stratified cross-validation model datasets. Our model AUPRC was less than 50%. A low AUPRC indicates that the model is having difficulty in identifying true positives without also including a high number of false positives. Future modeling attempts can be improved by considering the following (32); a] Implementing over-sampling and under-sampling techniques; b] Tuning the decision threshold to maximize the AUPRC, and monitoring recall, precision and F1 score model metrics (Figure 26); c] Creating new features and applying additional transformations to existing features; d] Utilizing transfer learning, where a model trained on one task is adapted to improve

performance on a related task using a pre-trained model.

$$Recall = \frac{TP}{TP + FN} \qquad Precision = \frac{TP}{TP + FP} \qquad F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figure 26. Calculating recall, precision and F1 value

Survey data limitations and assumptions

The responses were most likely influenced by the following factors, that may have resulted in inaccuracies. a] Cultural/Social Bias: Respondents may have answered based on what they believe are considered their social/cultural norms, instead of expressing their true beliefs (33). b] Nonresponse Bias: Respondents may have feared repercussions for their answers and doubted the survey's anonymity (34), leading them to withhold their true beliefs. c] Memory Bias (35): Respondents may have poor memory recall for a variety of factors, leading to inaccurate or inconsistent responses. d] Teenage Immaturity (36): The respondents are all teenagers which may lead to inaccurate responses due to their immaturity.

It was assumed that the survey sample represented a random selection of all American high school students and the survey conclusions could hence be extrapolated for all American high school students. Also, it was assumed that respondents were given ample privacy to answer independently and did not feel pressured to answer based on others' choices.

Conclusion

This study compared three machine learning models for their accuracy in predicting suicide attempts, and identified the top risk factors, which, in turn, suggested mitigation strategies to reduce youth suicide attempts. The LightGBM model outperformed the Random Forest and Logistic Regression models in terms of prediction accuracy. An important finding was that the models could be rendered functional by first reducing the 99 survey questions into 10 non-correlated risk factor themes. Utilizing stratified cross-validation along with AUPRC and AUC metrics provided an effective approach to evaluate the imbalanced data with a 10.2% target rate. The top model predictors returned by the LightGBM model were BMI, sleep, COVID related impacts, alcohol early age use and physical activity.

Corresponding mitigation strategies suggested to reduce youth suicide attempts were to increase physical activity to improve BMI; to addressing alcohol abuse through counseling and educational programs on the risks of alcohol; to maintain routine sleeping habits and to foster social connections through more virtual and in-person school activities. The latter could be addressed by school administrators and mental health professionals identifying at-risk students earlier and providing support and referrals to treatment services, as well as adding more education classes for promotion of the mental wellness of adolescents.

Acknowledgments

We would like to express our sincerest thanks to our mentors Nick Tschaika, Yan Zhao, and Dr. Manju Daniel for investing their time to help us improve our project.

References

1. National Institute of Mental Health. (2021). *Major Depression*. U.S. Department of Health and Human Services, National Institude of Health. https://www.nimh.nih.gov/health/statistics/major-depression

2. National Alliance on Mental Illness. (2023*). The Reality of "High Functioning" Depression*. https://nami.org/Blogs/NAMI-Blog/October-2023/The-Reality-of-High-Functioning-Depression

3. Hink, A. B., Killings, X., Bhatt, A., Ridings, L. E., Andrews, A. L. (2022). *Adolescent Suicide-Understanding Unique Risks and Opportunities for Trauma Centers to Recognize, Intervene, and Prevent a Leading Cause of Death*. Current trauma reports, 8(2), 41–53. https://doi.org/10.1007/s40719-022-00223-7

4. Pfledderer, C. D. (2019). *School environment, physical activity, and sleep as predictors of suicidal ideation in adolescents: Evidence from a national survey*. Journal of Adolescence, 74, 83-90. https://www.sciencedirect.com/science/article/abs/pii/S0140197119300910

5. Maurya, C., Muhammad, T., Thakkar, S. (2023). *Examining the relationship between risky sexual behavior and suicidal thoughts among unmarried adolescents in India. Scientific reports*, 13(1), 7733. https://doi.org/10.1038/s41598-023-34975-2

6. Baiden, P., Panisch, L. S., Kim, Y. J., LaBrenz, C. A., Kim, Y., Onyeaka, H. K. (2021). *Association between First Sexual Intercourse and Sexual Violence Victimization, Symptoms of Depression, and Suicidal Behaviors among Adolescents in the United States: Findings from 2017 and 2019 National Youth Risk Behavior Survey*. International Journal of Environmental Research and Public Health, 18(15), 7922. https://doi.org/10.3390/ijerph18157922

7. Garofalo R., Cameron Wolf R., Wissow, L. (1999). *Sexual Orientation and Risk of Suicide Attempts Among a Representative Sample of Youth*. Arch. Pediatr. Adolesc. Med., 153(5), 487-493. https://jamanetwork.com/journals/jamapediatrics/fullarticle/346930

8. Centers for Disease Control and Prevention. (2023). *CDC releases the Youth Risk Behavior Survey Data Summary & Trends Report: 2011-2021*. https://www.cdc.gov/nchhstp/dear_colleague/2023/dsrt-dcl.html

9. Lee, J. W., Kim, B. J., Lee, C. S., Cha, B., Lee, S. J., Lee, D., et al. (2021). *Association Between Suicide and Drinking Habits in Adolescents*. Soa Chongsonyon Ohongsin Uihak, 32(4), 161-169. https://doi.org/10.5765/jkacap.210024

10. Weissinger, G., Rivers, A. S., Atte, T., Diamond, G. (2023). *Suicide risk screening in the school environment: Family factors and profiles.* Children and Youth Services Review, 145, 1-6. https://www.sciencedirect.com/science/article/abs/pii/S0190740922004029

11. Slesnick,N., Zhang, J., & Walsh, L., (2020). *Youth Experiencing Homelessness with Suicidal Ideation: Understanding Risk Associated with Peer and Family Social Networks*. Community Ment. Health J., 57(1), 128-135. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7572482/

12. Pathirathna, M.L., Nandasena, H.M.R.K.G., Atapattu, A.M.M.P., Weerasekara, I. *Impact of the COVID-19 pandemic on suicidal attempts and death rates: a systematic review*. BMC Psychiatry 22, 506 (2022). https://doi.org/10.1186/s12888-022-04158-w

13. Iwatate, E., Folefac, D., Atem, E.C., Jones, J.L., Hughes, T.Y., E. M.S. (2023). *Association of Obesity, Suicide Behaviors, and Psychosocial Wellness Among Adolescents in the United States*. Journal of Adolescent Health, 72(4), 526-534. https://doi.org/10.1016/j.jadohealth.2022.11.240

14. Ning, K., Yan, C., Zhang, Y., Chen, S. (2022). *Regular Exercise with Suicide Ideation, Suicide Plan and Suicide Attempt in University Students: Data from the Health Minds Survey 2018-2019*. International journal of environmental research and public health, 19(14), 8856. https://doi.org/10.3390/ijerph19148856

15. Centers for Disease Control and Prevention. (2021). *2021 National Youth Risk Behavior Survey: National High School Questionnaire.* *https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2021/2021-YRBS-National-HS-Questionnaire.pdf*

16. Mpofu, J.J., Underwood, J.M., Thornton, J.E., Brener, N.D., Rico, A., Kilmer, G., et al. (2023). *Overview and Methods for the Youth Risk Behavior Surveillance System — United States, 2021*. MMWR Suppl, 72(Suppl-1), 1–12. http://dx.doi.org/10.15585/mmwr.su7201a1

17. Hosmer, D. W., Lemeshow, S., Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley, Hoboken, NJ, USA. https://www.wiley.com/en-us/Applied+Logistic+Regression%2C+3rd+Edition-p-9780470582473

18. Hastie, Trevor, et al. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer, NY, NY, USA. https://link.springer.com/book/10.1007/978-0-387-84858-7

19. Khandelwal (2023). *Which algorithm takes the crown: Light GBM vs. XGBOOST?* https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/

20. Machine Learning Mastery. (2020). *Cross-validation for imbalanced classification*. https://machinelearningmastery.com/cross-validation-for-imbalanced-classification/

21. Jones, C., Selvanathan, S., Brown, J. R., Moosa, R., Dennison, A. R. (2023). *Artificial intelligence in colorectal cancer: Current applications and future directions*. Frontiers in Medicine*, 10*, 125378. https://doi.org/10.3389/fmed.2023.125378

22. Narkhede, S. (2018). *Understanding Confusion Matrix*. https://www.scirp.org/reference/referencespapers?referenceid=2796936

23. Kirkman, T.W. (1996). *Kolmogorov-Smirnov Test*. http://www.physics.csbsju.edu/stats/KS-test.html

24. Siddiqi, N. (2015). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Wiley, Hoboken, NJ, USA. https://books.google.com/books?id=SEbCeN3-kEUC&printsec=frontcover#v=onepage&q&f=false

25. Nahm, F. S. (2017). *What the P values really tell us*. The Korean Journal of Pain, 30(4), 241–242. https://doi.org/10.3344/kjp.2017.30.4.241

26. Wikipedia contributors. (2024). *LightGBM. In Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/wiki/LightGBM

27. Amazon. (2023). *Tune a LightGBM model*. https://docs.aws.amazon.com/sagemaker/latest/dg/lightgbm-tuning.html

28. Mandot, P. (2017). *What is LightGBM, how to implement it? How to fine tune the parameters?* https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

29. Breiman L., (2001). *Random Forests*. Machine Learning, 45, 5-32. https://link.springer.com/content/pdf/10.1023/A:1010933404324.pdf

30. Saxena, S. (2023). *A beginner's guide to random forest hyperparameter tuning*. https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/

31. Anupuma S. (2023). *T-test: Definition, formula, types, applications*. https://microbenotes.com/t-test/

32. He, H., Ma, Y. (Eds.). (2013). *Imbalanced Learning: Foundations, Algorithms, and Applications*. Wiley, Hoboken, NJ, USA. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118646106

33. Groves, R. M., Fowler, F. J., Couper, M.P., Lepkowski, J.M., Singer, E., Tourangeau, R. (2009). *Survey Methodology (2nd ed.)*. Wiley, Hoboken, NJ, USA. https://www.wiley.com/en-us/Survey+Methodology%2C+2nd+Edition-p-9780470465462

34. Groves, R. M., Peytcheva, E. (2008). *The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis*. Public Opinion Quarterly, 72(2), 167-189. http://dx.doi.org/10.1093/poq/nfn011

35. Bradburn, N. M., Rips, L. J., Shevell, S. K. (1987). *Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys*. Science, 236(4798), 157-161. https://doi.org/10.1126/science.3563494

36. Hargittai, E., Hinnant, A. (2008). *Digital Inequality: Differences in Young Adults' Use of the Internet*. Communication Research, 35(5), 602-621. https://doi.org/10.1177/0093650208321782