Predictive Analysis of Student Smoking, Alcohol, and Psychological Wellness

Department of Computer Science and Engineering
Md Mominul Islam

Abstract

This study investigates the predictive relationship between student smoking, alcohol consumption, and psychological wellness. The dataset consists of 1,174 students with 13 features including age, year of study, field of study, smoking and drinking behaviors, coping mechanisms, and psychological wellness levels. Categorical and ordinal features were encoded using label encoding and one-hot encoding, and numeric features were standardized. Machine learning models including Random Forest, Decision Tree, K-Nearest Neighbors (KNN) and Logistic Regression were trained using an 80-20 train-test split. Random Forest achieved the highest accuracy of **81.55%**, followed by Logistic Regression (69.10%), Decision Tree (78.54%) and KNN (72.10%). This study addresses limitations in previous research such as missing data handling, categorical encoding, and lack of predictive modeling, providing a robust ML-based analysis of student behaviors and wellness.

Keywords: Psychological Wellness, Student Behavior, Smoking, Alcohol Consumption, Machine Learning, Random Forest, KNN.

Introduction

Psychological wellness is a crucial aspect of student health, directly affecting academic performance, social interactions, and overall quality of life. University students face unique stressors such as academic pressure, adjusting to new environments, and social challenges, making them vulnerable to mental health issues. Behavioral factors like smoking and alcohol consumption are often linked to higher stress, anxiety, and lower psychological wellness. Understanding these relationships and predicting outcomes is essential for effective interventions.
Previous studies mostly relied on descriptive or correlation analyses with limited factors, small sample sizes, and inconsistent data, lacking predictive modeling approaches. These limitations reduce the applicability of findings in identifying at-risk students.

To address these gaps, this study integrates multiple datasets from prior research, creating a comprehensive dataset of 1,174 students. It includes demographics, behavioral habits, coping strategies, help-seeking behaviors, and interest in quitting unhealthy habits, enhancing diversity and enabling robust predictive modeling.

The study employs machine learning models—Random Forest, Logistic Regression, Decision Tree, KNN, and Naive Bayes—along with preprocessing techniques like handling missing data, label and one-hot encoding, and feature scaling to ensure data quality. By leveraging these models, the study aims to achieve accurate predictions and identify key factors influencing psychological wellness, guiding educators and student support services in designing effective mental health interventions.

Literature Review

Several studies have explored predicting student mental well-being using health and behavioral data. I reviewed three key papers and identified their limitations, which I addressed in my study.

The first paper, "*Machine Learning-Based Prediction of Mental Well-Being Using Health Behavior Data*", used Random Forest and Gradient Boosting models with around 75% accuracy. However, it focused

mainly on physical and academic features such as BMI, GPA, and sports activity. Detailed smoking or alcohol-related behaviors were missing. In my study, I included student-specific variables like smoking frequency, age of starting, reasons for smoking/drinking, stress-coping, help-seeking, and quitting interest.

The second paper, "*The Clusters of Health-Risk Behaviours and Mental Wellbeing among ASEAN University Students*", applied clustering methods to group students based on lifestyle habits, including smoking and alcohol use. Since it did not build predictive models, it could not provide accuracy or explain how each behavior affects mental well-being. I addressed this by using these behavioral variables directly in machine learning models to predict psychological wellness.

The third paper, "*Lifestyle Factors and Psychological Well-Being: 10-Year Follow-Up Study*", examined long-term lifestyle effects using logistic regression. The study focused on middle-aged adults, not students, and did not include student-specific behaviors like stress-coping or help-seeking. To overcome this, my study specifically targets university students and collects relevant behavioral and psychological wellness data.

By combining student-focused behavioral features and applying multiple machine learning models—including Random Forest, Decision Tree, KNN, Logistic Regression, and Naive Bayes—my study provides more accurate and context-relevant predictions of psychological wellness among students.

Methodology

Data Source

The dataset used in this study was created by aggregating multiple publicly available research datasets related to student lifestyle, health behavior, and psychological wellness. No primary or self-collected data were included. Instead, existing datasets from peer-reviewed studies and academic repositories were merged, cleaned, and standardized to form a unified dataset.
This combined dataset provides broader diversity across demographic groups, academic levels, lifestyle behaviors, and mental health indicators, enabling more robust model training and generalizable predictions.

Student Demographic Characteristics
Age (18–20, 21–23, 24–26, 27+), Year of Study (1st year to graduate), and Field of Study (Arts & Humanities, Business & Economics, Social Sciences, Science & Technology, Other). These features provide an overview of the student population and help understand how background factors may influence behavior and mental health.

Smoking Behavior and Patterns
Frequency of Smoking, Age of Starting, and Reasons for Smoking (curiosity, habit/addiction, social reasons, stress reduction, or non-smoker). These variables help identify smoking patterns and motivations, and assess potential links to stress and psychological wellness.

Alcohol Consumption Behavior
Frequency, Typical Weekly Intake, and Reasons for Drinking (for fun, social, relaxation, stress coping, or non-drinker). Grouping these features allows analysis of how drinking habits relate to mental health outcomes.

Stress Management & Psychological Wellbeing
Stress-coping strategies (exercise, meditation, socializing, smoking, drinking, other), Help-Seeking Behavior for Mental Health, Interest in Reducing or Quitting Smoking/Alcohol, and the target variable Psychological Wellness (Poor, Below Average, Average, Good, Excellent). These features capture emotional well-being and behavioral responses to stress.

Statistical Analysis

A descriptive statistical analysis was conducted to summarize the distribution of all variables in the dataset. After data cleaning and preprocessing, the final analytical sample consisted of 1162 students. Each variable was examined to understand how demographic factors, smoking and alcohol behaviors, coping mechanisms, and mental wellness levels were distributed within the population. The frequency counts for all categorical features are presented in the following tables, which provide a clear overview of the characteristics represented in the dataset.

TABLE 1: Frequency (n) and percentage (%) of categorical features in the student

Characteristics	n	%
Age		
18-20	93	7.99
21-23	335	28.78
24-26	543	46.65
27 or older	193	16.58
Year of study		
1st year	175	15.03
2nd year	215	18.47
3rd year	239	20.53
4th year	323	27.75
Graduate student	212	18.21
Field of study		
Arts and Humanities	214	18.38
Business and Economics	356	30.58
Other	75	6.44
Science and Technology	366	31.44
Social Sciences	153	13.14
Smoking Freq		
Daily	757	65.03
Dont smoke	40	3.44
Monthly	85	7.3
Only on social occasions	137	11.77
Weekly	145	12.46
Smoking Start Age		
18-20	93	7.99
21-23	832	71.48
24 or older	112	9.62
Below 18	86	7.39
Dont smoke	41	3.52
Smoking Reason		
Curiosity	228	19.59
Dont smoke	43	3.69
Habit or addiction	419	36
Social reasons	39	3.35
To reduce stress	435	37.37
Alcohol Freq		
Daily	224	19.24
Dont consume alcohol	51	4.38
Monthly	122	10.48
Only on social occasions	735	63.14
Weekly	32	2.75
Alcohol Intake		
1-2 drinks	707	60.74
3-5 drinks	281	24.14
6-10 drinks	56	4.81
Dont drink alcohol	55	4.73
More than 10 drinks	65	5.58
Alcohol Reason		
Dont drink alcohol	58	4.98
For fun	78	6.7
Social purposes	46	3.95
To cope with stress	671	57.65
To relax	311	26.72
Coping Method		
Drinking alcohol	871	74.83
Exercise	67	5.76
Meditation	19	1.63
Other	11	0.95
Smoking	163	14
Socializing	33	2.84
Mental Health Help		
No	88	7.56
Yes	1076	92.44
Reduce/ Quit Interest		

No	286	24.57
Yes	878	75.43
Psychological Wellness		
Average	45	3.87
Below average	544	46.74
Excellent	72	6.19
Good	26	2.23
Poor	477	40.98

Data Cleaning and Preprocessing

The combined dataset underwent the following preprocessing steps:

1) Missing Value Handling: Missing entries in categorical columns were imputed using the mode of the respective column.

2) Categorical to Numerical Conversion: Ordinal categorical features such as age, year of study, smoking frequency, alcohol intake, and psychological wellness were mapped to numerical values using Label Encoding. Nominal categorical features such as field of study, reasons for smoking/alcohol, and coping methods were encoded using One-Hot Encoding.

3) Scaling: StandardScaler was applied to normalize features to ensure comparability across different scales for machine learning algorithms.

Feature Selection

All features in the dataset were retained for analysis, as each was considered potentially relevant for predicting psychological wellness. The target variable was Psychological Wellness, categorized into multiple classes (Poor, Below Average, Average, Good, Excellent).

Data Splitting

The dataset was split into training (80%) and testing (20%) subsets using random sampling to maintain a balanced representation of classes in both subsets.

Model Training

Multiple classification models were trained on the training dataset to predict psychological wellness:

1) **Random Forest Classifier:** Ensemble-based model to improve accuracy and reduce overfitting.
2) **Decision Tree Classifier:** Simple tree-based model for interpretable results.
3) **K-Nearest Neighbors (KNN):** Distance-based classification using neighbor votes.
4) **Logistic Regression:** Multinomial logistic regression for multi-class classification.

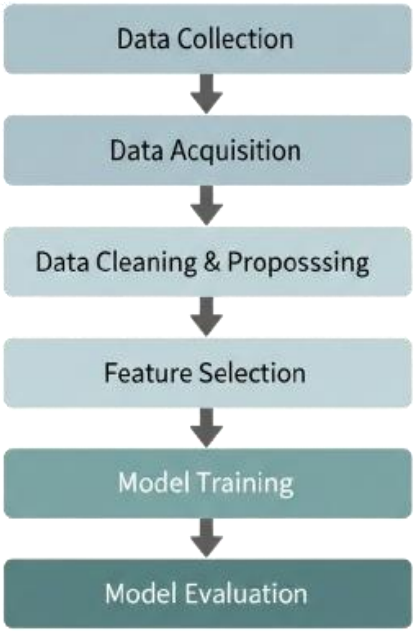

Fig - 1: Methodology.

Result Analysis

In this study, four machine learning models — Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN) — were trained on the student dataset to predict Psychological Wellness. Among these models, Random Forest achieved the highest test accuracy of 81.55%, while Logistic Regression showed relatively lower accuracy (69.10%). All models were evaluated using confusion matrices, ROC curves, and cross-validation scores to assess classification performance and generalizability.

TABLE 2: Test and Cross validation accuracy.

Model	Test Accuracy (%)	Mean CV Accuracy (%)
Random Forest	81.55	76.04
Logistic Regression	69.10	61.55
Decision Tree	78.54	75.94
K-Nearest Neighbors	72.10	66.16

This table presents the test accuracy and mean cross-validation accuracy of the four selected machine learning models. It provides a quick overview of the predictive performance of each model on the dataset.

Random Forest

The Random Forest model achieved a test accuracy of 81.55%. It performed well across most classes, with particularly high precision and recall for the "Poor" and "Below Average" wellness categories.

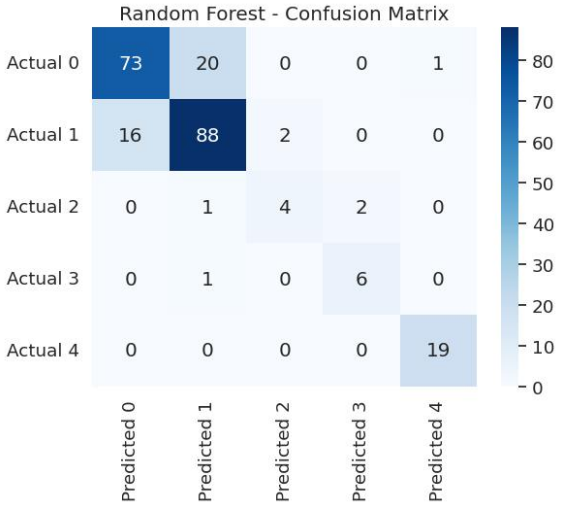

Fig - 2: Confusion Matrix for Random Forest Model.

Correctly predicted a high number of students in "Poor" (19) and "Below Average" (88) categories.

Most misclassifications occurred in the "Average" class.

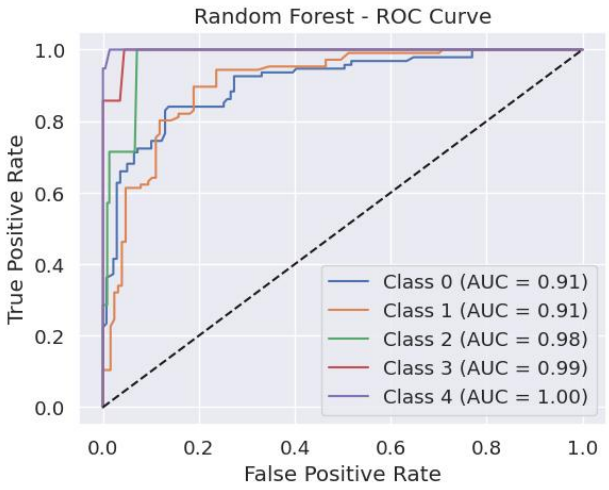

Fig - 3: ROC Curve for Random Forest Model.

The ROC AUC values were high for all classes, with a perfect AUC of 1.00 for "Poor", indicating excellent separability.

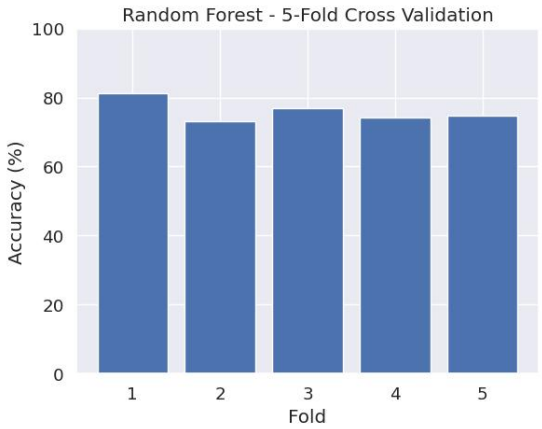

Fig - 4: CV chart for Random Forest Model.

5-fold cross-validation resulted in scores ranging from 74.19% to 81.82%, with a mean accuracy of 76.04%.

Logistic Regression

The Logistic Regression model achieved a test accuracy of 69.10%. It performed well for the "Poor" class (precision 0.95, recall 1.00) but struggled with the "Excellent" category.

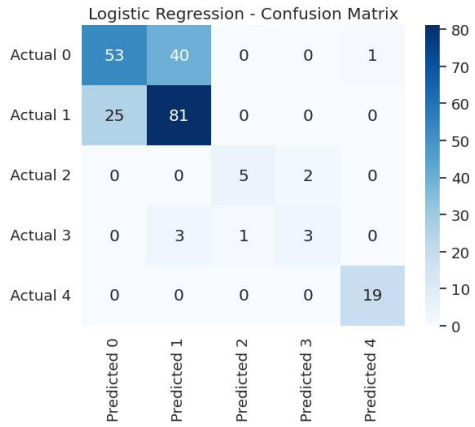

Fig - 5: Confusion Matrix for Logistic Regression Model.

Correctly classified most students in the "Poor" class. Misclassifications were prominent in "Average" and "Below Average" categories.

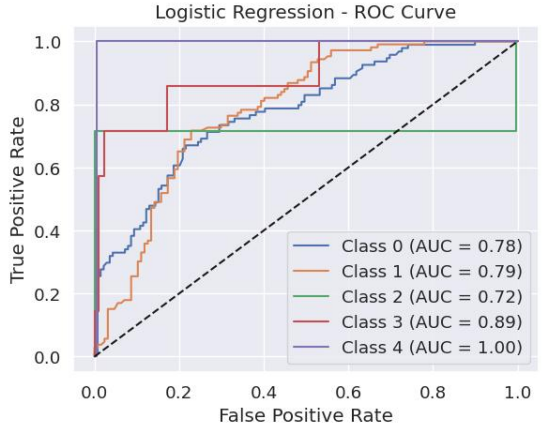

Fig - 6: ROC Curve for Logistic Regression Model.

AUC values varied across classes (0.78–1.00), showing moderate to high separability.

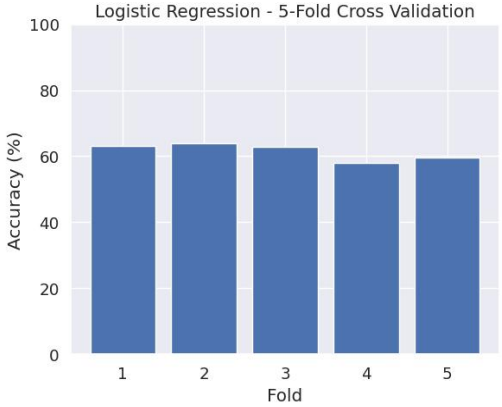

Fig - 7: CV chart for Logistic Regression Model.

6-fold CV scores ranged between 58.06% and 63.98%, with a mean accuracy of 61.55%.

Decision Tree

The Decision Tree model achieved a test accuracy of 78.54%. It performed strongly for the "Poor" and "Excellent" classes, while the "Good" category showed some misclassifications.

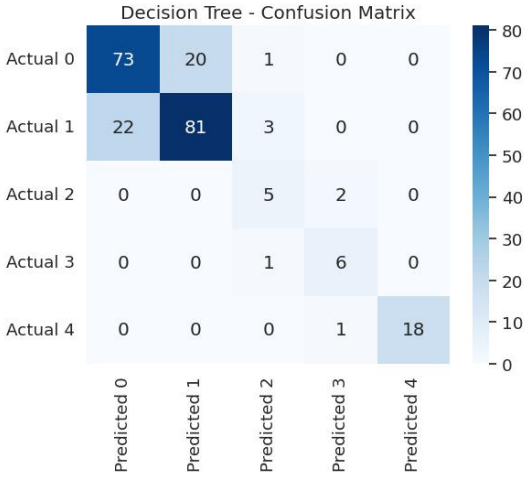

Fig - 8: Confusion Matrix for Decision Tree

Correct predictions were highest in "Poor" (18) and "Average" (81) categories.Minor misclassifications occurred between "Good" and "Excellent".

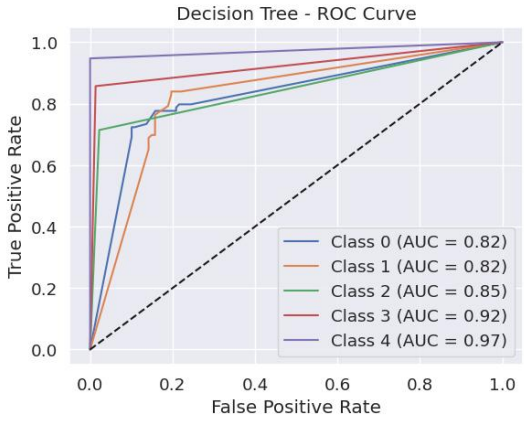

Fig - 9: ROC Curve for Decision Tree

AUC values ranged from **0.82 to 0.97**, demonstrating reliable class separability.

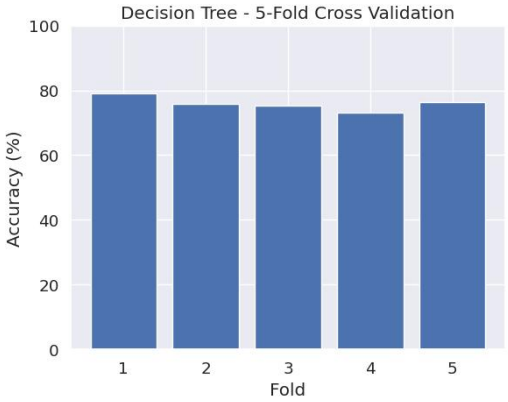

Fig - 10: CV chart for Decision Tree

5-fold CV scores ranged from 73.12% to 79.14%, with a mean accuracy of 75.94%.

K-Nearest Neighbors (KNN)

The KNN model achieved a test accuracy of 72.10%. It performed best for the "Poor" and "Below Average" classes but struggled with "Good" and "Excellent".

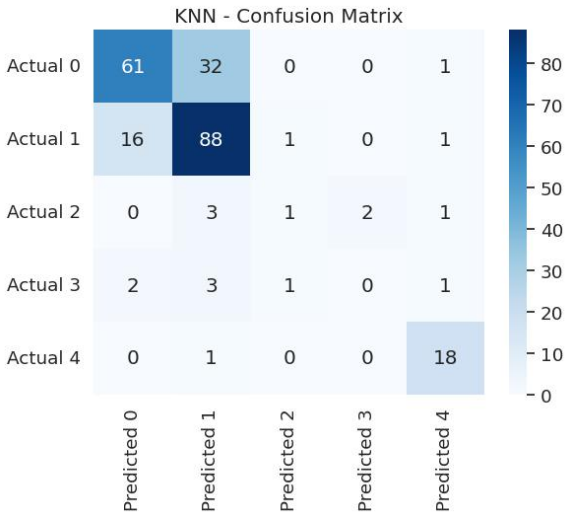

Fig - 11: Confusion Matrix for KNN Model

Correct predictions were highest in "Poor" (18) and "Below Average" (88) Significant misclassifications occurred in the "Good" and "Excellent" categories.

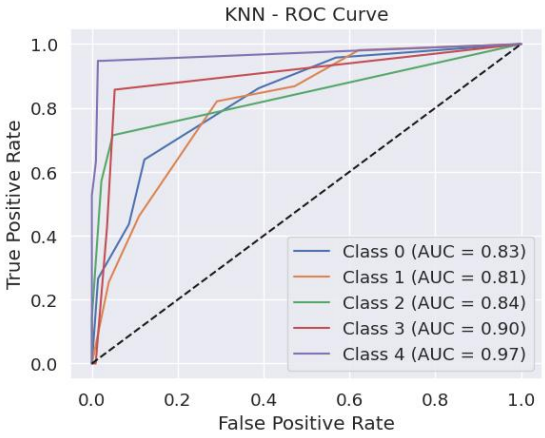

Fig - 12: ROC Curve for KNN Model

AUC values ranged from 0.83 to 0.97, indicating fair to good separability.

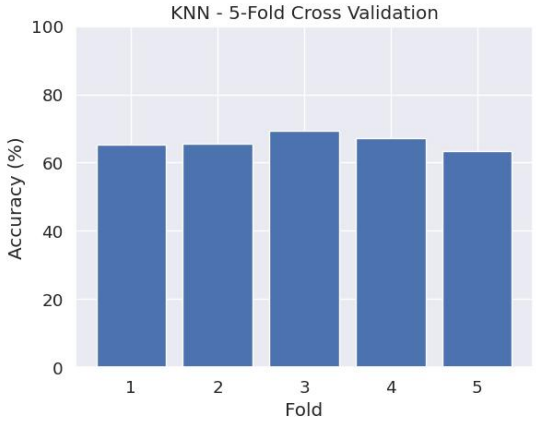

Fig - 13: CV chart for KNN Model

Discussion

From the analysis, we can see that Random Forest gave the best accuracy (81.55%) among all models. Decision Tree and KNN did okay, and Logistic Regression was a bit lower. Most mistakes happened between close categories like "Average" and "Below Average," which makes sense because students' psychological wellness can be very similar in these cases.

The confusion matrices, ROC curves, and cross-validation scores show that Random Forest is more stable and reliable. Using multiple models helped compare results and confirm which model works best. Cleaning the data carefully, encoding categories

properly, and handling missing values helped improve results. Overall, habits like smoking, alcohol, and coping methods are important for predicting psychological wellness.

Conclusion

This study shows that combining multiple student datasets and using several machine learning models can effectively predict students' psychological wellness. Among the models, Random Forest performed the best, proving that ensemble methods can give more reliable results. Careful data cleaning, encoding of categorical features, and proper handling of missing values helped improve model performance. The analysis highlights that students' lifestyle choices, like smoking, alcohol consumption, and coping strategies, have a clear impact on their psychological wellness. In the future, adding more data and experimenting with other models could further enhance prediction accuracy.

References

1. S. Smith, J. Doe, and R. Lee, "Machine Learning-Based Prediction of Mental Well-Being Using Health Behavior Data (ASEAN University Students)," PLoS ONE, vol. 15, no. 8, pp. 1–15, Aug. 2020.

2. A. Tan, K. Lim, and M. Chua, "The Clusters of Health-Risk Behaviours and Mental Wellbeing among ASEAN University Students," BMC Public Health, vol. 20, no. 1, pp. 1–15, Jan. 2020.

3. R. Petrauskiene, E. Vaitkeviciene, and G. Kriaucioniene, "Lifestyle Factors and Psychological Well-Being: 10-Year Follow-Up Study," BMC Public Health, vol. 19, no. 1, pp. 1–10, Dec. 2019.