Performance Factors Influencing Depression

Introduction

Mental health is a critical component of overall well-being, yet it often remains underprioritized, particularly in academic settings where performance pressures can significantly affect students' mental states.

Depression, a leading cause of mental health challenges, can manifest due to various factors, including academic stress, social dynamics, and personal challenges. Identifying these factors early can help mitigate their impact and create a healthier environment for students.
This project aims to build a predictive model that studies the effects of performance-related factors on the likelihood of a student suffering with depression. By analyzing these features, the study seeks to uncover patterns that may indicate an inclination to facing depression.

The motivation behind this research stems from the growing need to bridge the gap between academic performance and mental health. While academic success is important, mental health is equally crucial for students' long-term success and better quality of life. My initial hypothesis is that high levels of academic and social stress combined with lower parental involvement and lower motivation increase the likelihood of depression among students.
The potential outcomes of this project can guide educators, policymakers, and mental health professionals in identifying at-risk students and implementing timely interventions.

Datasets Used

To investigate the relationship between performance factors and the likelihood of depression among students, I selected two datasets from Kaggle that were relevant to the problem statement:

1. **Student Performance Factors Dataset**
    - **Source:** Kaggle
    - **Description:** This dataset provides insights into various factors influencing student performance, including hours studied, parental involvement, motivation levels, and peer influence, among many others. It contains structured data with both categorical and numerical variables.
    - **Information Types:**
        - Academic metrics such as previous scores, hours studied, and grades.
        - Psychological indicators like motivation levels and peer influence.
        - Social and environmental factors, including parental involvement and extracurricular activities.

2. **Student Mental Health Dataset**

    - **Source:** Kaggle

    - **Description:** This dataset focuses on mental health indicators among students, such as levels of anxiety, stress, and depression.

    - **Information Types:**
        - Categorical variables representing depression and anxiety status (Yes/No).
        - Demographic details such as age and gender.
        - Academic workload and study habits.

Criteria for Dataset Selection

The criteria for selecting these datasets included:

1. **Relevance:** Both datasets directly pertain to the problem of investigating the link between academic performance and depression.

2. **Complementary Information:** The two datasets once combined would be able to help analyse the relationship between the performance factors and susceptibility to depression

3. **Data Quality:** Both datasets had a significant number of observations, minimal missing values.

Exploratory Data Analysis

This study explores the relationship between student mental health and academic performance, using two datasets: *Student Mental Health.csv* and *StudentPerformanceFactors.csv*. The primary objective is to identify the factors that potentially influence mental health outcomes, particularly depression, among students. The analysis involved a systematic approach to data cleaning, preprocessing, statistical exploration, and visualization, while considering potential sources of error and bias.

Data Cleaning and Preprocessing

1. **Data Loading and Merging**

    - The datasets were loaded into Python using pandas.

    - They were merged on the *CGPA* (Cumulative Grade Point Average) column using an inner join, resulting in a combined dataframe.

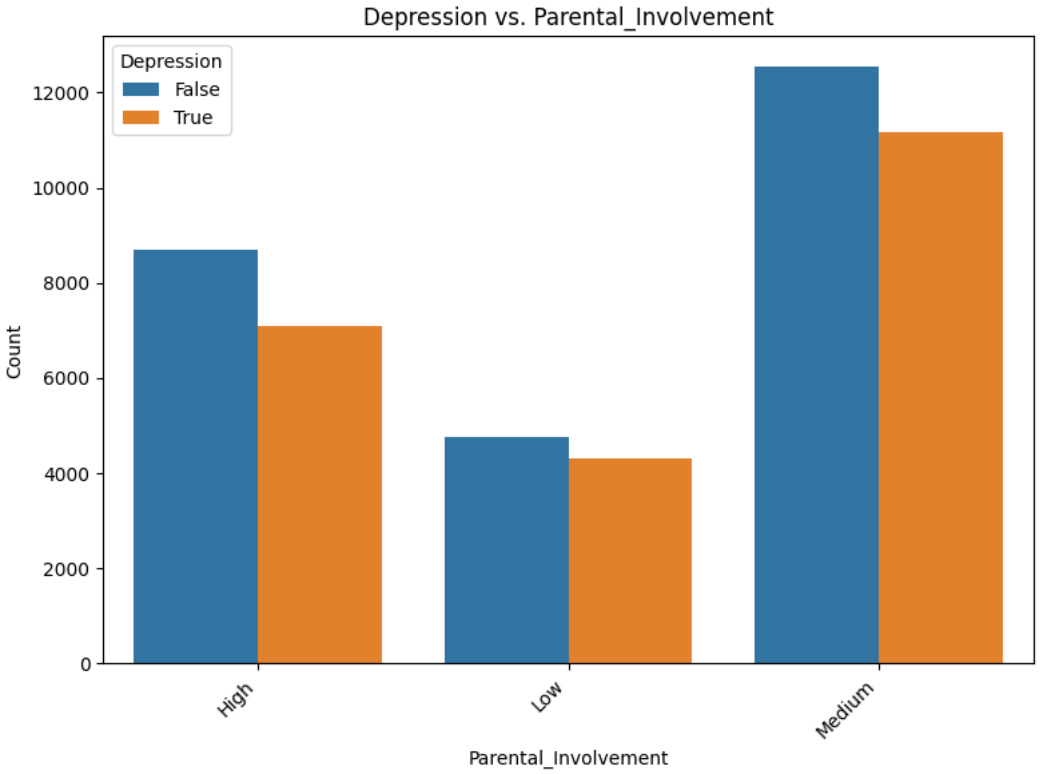
2. **Column Renaming and Standardization**
    - Column names in the mental health dataset were renamed for clarity
    - Inconsistent intervals in the *CGPA* column were standardized across both datasets to enable seamless integration.
    - The *Year of Study* column was cleaned by removing unnecessary text and converting values to lowercase for uniformity.
3. **Handling Missing Data**
    - Missing values were removed using dropna().
4. **Data Transformation**
    - Boolean values such as *Yes/No* were converted to *True/False* for analysis.
    - Categorical variables (*Parental Involvement*, *Motivation Level*, *Access to Resources*) were transformed into the category data type.
    - The *School Type* column was encoded into a new boolean column, *Private School*, where *True* represents private schools and *False* represents public schools.
5. **Feature Selection**
    - Preliminary visualizations informed the selection of key features that potentially relate to depression. These features were compiled into a filtered dataframe to streamline analysis.

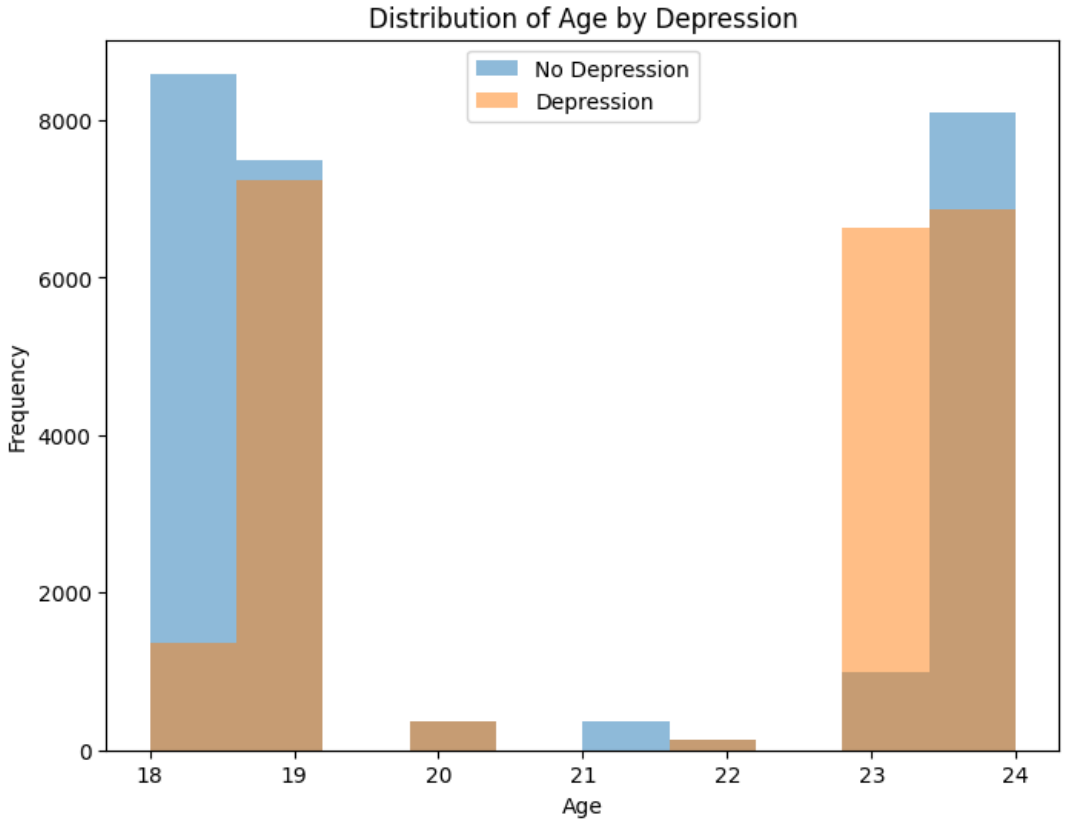
Statistical Analysis

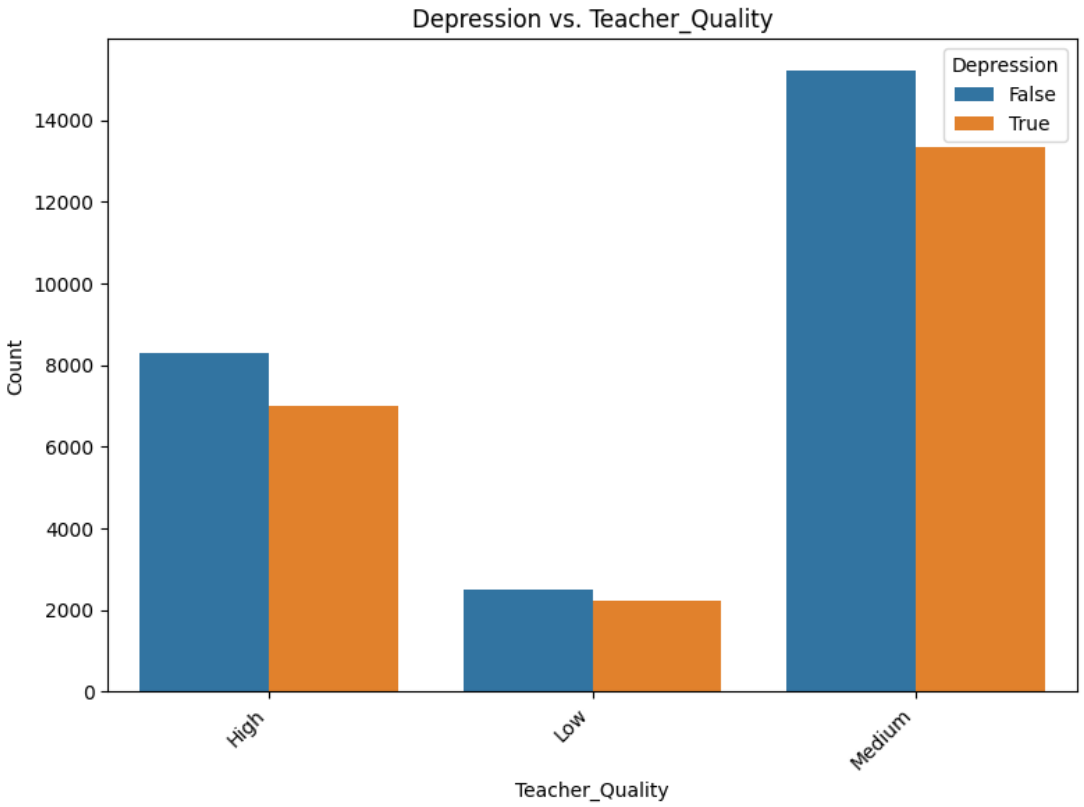
1. **Visualizations**

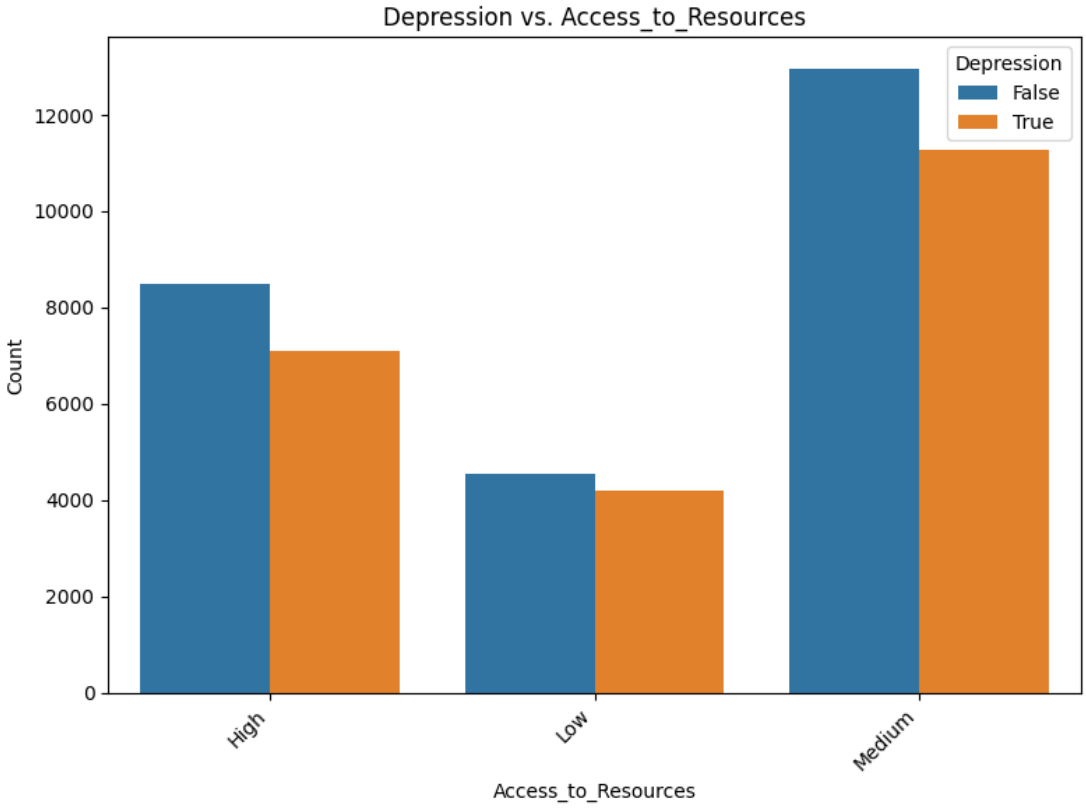
A variety of visualizations were generated to identify trends and relationships in the data.

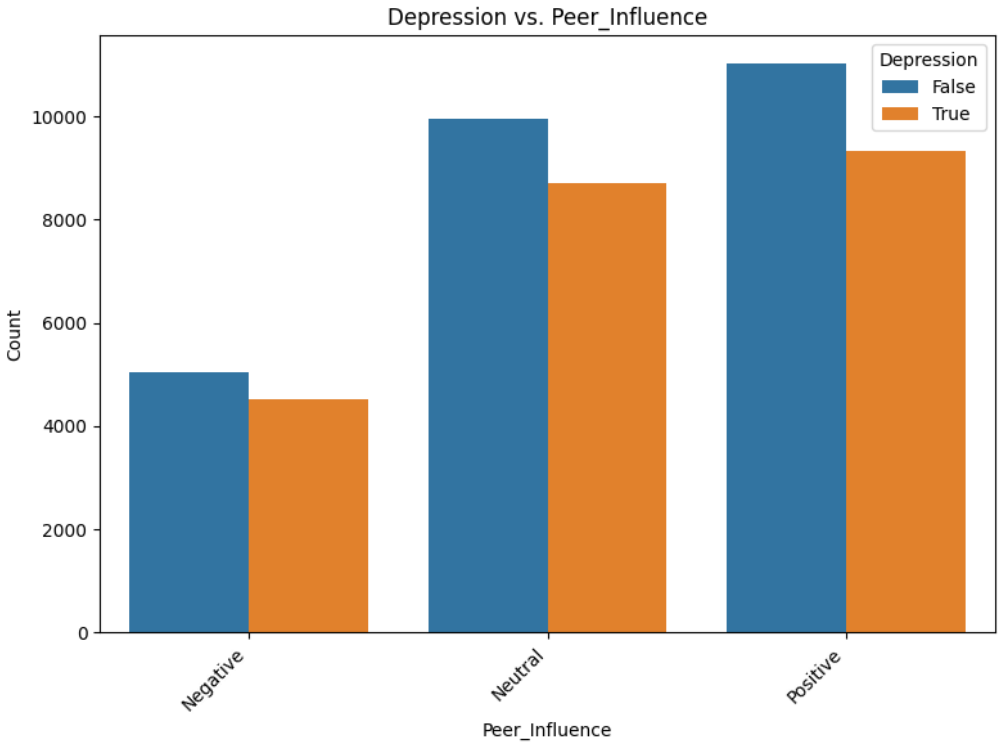
- **Histograms:** Examined the distribution of numerical features for students with and without depression.
- **Confusion Matrix:** Highlighted the relationship between Mental Health features like Depression, Anxiety and Panic Attacks
- **Count Plots:** Illustrated the frequency of depression across categorical variables.

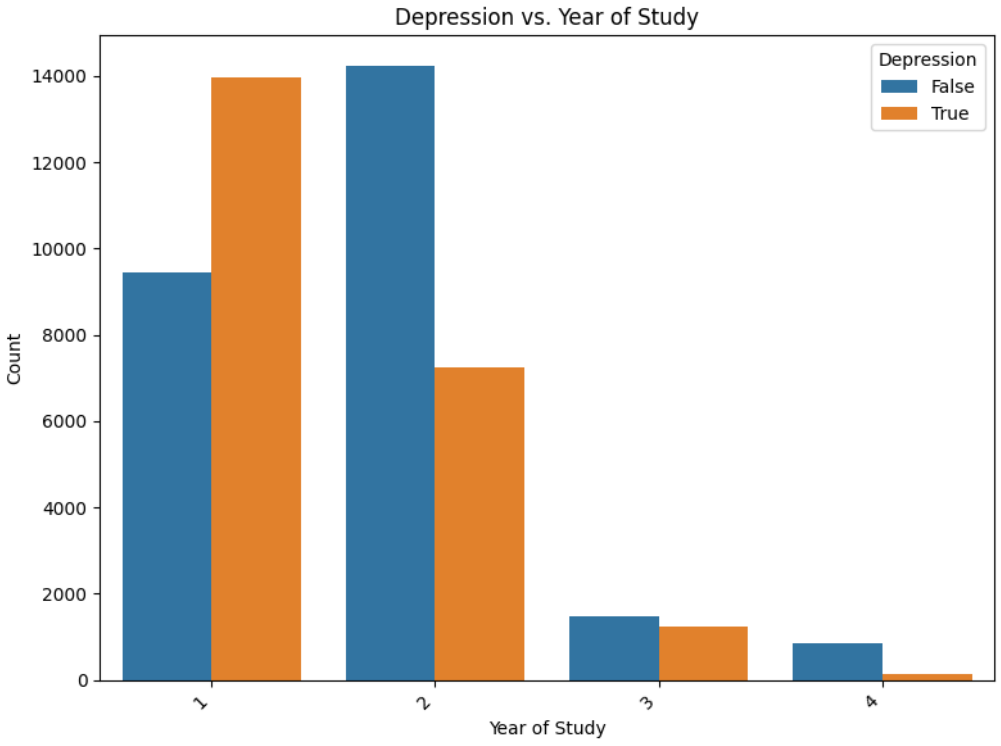

Depression vs. Parental_Involvement
Depression
False
True
12000
10000
8000
6000
4000
2000
0
Count
High
Low
Medium
Parental_Involvement

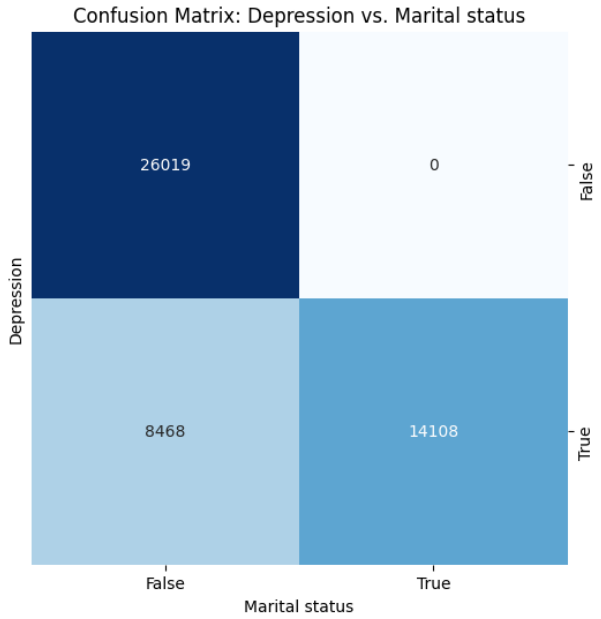

Distribution of Age by Depression
No Depression
Depression
8000
6000
4000
2000
0
Frequency
18
19
20
21
22
23
24
Age


Depression vs. Teacher_Quality
Depression
False
True
14000
12000
10000
8000
6000
4000
2000
0
Count
High
Low
Medium
Teacher_Quality


Depression vs. Access_to_Resources
Depression
False
True
12000
10000
8000
6000
4000
2000
0
Count
High
Low
Medium
Access_to_Resources


Depression vs. Peer_Influence
Depression
False
True
Count
0
2000
4000
6000
8000
10000
Negative
Neutral
Positive
Peer_Influence


Depression vs. Year of Study
Depression
False
True
Count
0
2000
4000
6000
8000
10000
12000
14000
1
2
3
4
Year of Study



Findings and Observations

1. Key Relationships

- **Parental Involvement:** Reduced parental involvement was linked to an increased frequency of depression, underscoring the importance of active parental engagement in students' lives
- **Year of Study:** Students in their first year of college showed high susceptibility to depression, indicating that academic and social transition could play a major role
- **Age:** People at the age of 23 were most susceptible to depression compared to other age groups
- **Marital Status**: If a student was married, they would most likely be suffering with depression
- **Teacher Quality:** Low teacher quality was associated with a higher frequency of depression, highlighting the critical role teachers play in influencing students' mental well-being
- **Access to Resources:** Limited access to resources was correlated with a higher frequency of depression, likely due to the additional challenges students face in completing their academics, which significantly impacts their mental health
- **Peer Influence:** Negative peer influence was also found to affect the frequency of depression, emphasizing the reality and impact of peer pressure

Sources of Error, Uncertainty, and Bias

1. **Self-Reported Data:**
    - Mental health variables were based on self-reported responses, which could lead to underreporting or inaccuracies due to social stigma or lack of awareness.
2. **Sample Representation:**
    - The datasets may not fully represent the broader student population, as they are limited to specific geographic or demographic groups.
3. **Handling Missing Data:**
    - The removal of missing values using dropna() may have introduced selection bias.

This initial analysis highlights several important relationships and patterns in the data. However, the analysis is constrained by potential biases and limitations in data collection and preprocessing.

Model Development and Application

This section outlines the comprehensive process of selecting, developing, and applying various machine learning models to predict depression based on factors influencing student performance and mental health. The approach integrates a mix of exploratory analysis, model selection, and performance evaluation to achieve meaningful insights from the datasets.

1. Types of Models Explored

In this project, a variety of classification models were utilized to predict the likelihood of depression among students. Each model brings its unique strengths and assumptions, allowing for diverse perspectives in prediction:

- **Logistic Regression**:
  A fundamental statistical model well-suited for binary classification. It estimates the probability of depression by modeling a linear relationship between features and the target variable.

- **Naïve Bayes**:
  Applies Bayes' Theorem with the assumption of feature independence to classify data based on probabilities. This model is simple yet effective for many classification tasks.

- **Random Forest**:
  An ensemble method that combines multiple decision trees to improve robustness and accuracy. It reduces overfitting and handles missing data effectively.

- **K-Nearest Neighbors (KNN)**:

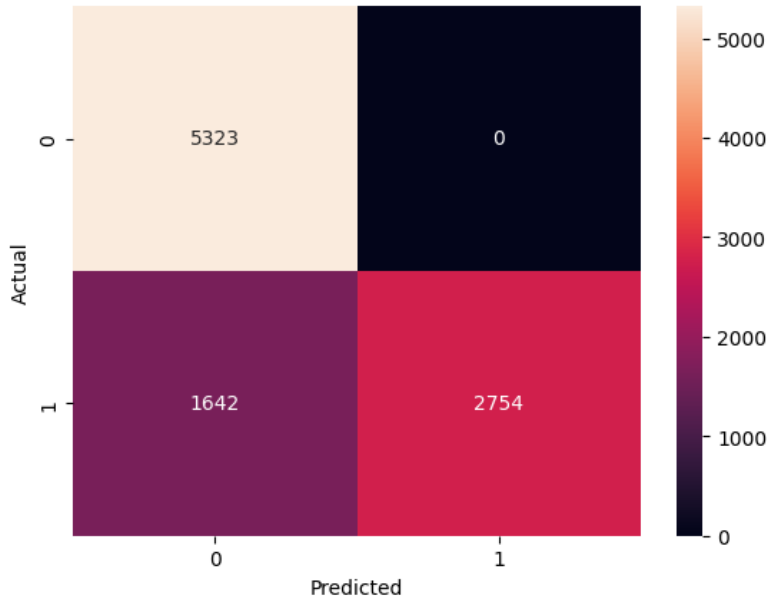
  A non-parametric algorithm that classifies samples based on the majority class among their k-nearest neighbors. It relies heavily on the structure of the data.

2. Model Performance Assessment

Model performance was evaluated using multiple metrics:

- **Accuracy**: Proportion of correctly predicted cases out of total cases.
- **Precision**: How many of the predicted positives were actual positives.
- **Recall**: How many of the actual positives were correctly identified.
- **F1-Score**: Balances precision and recall.

- **Logistic Regression:-**



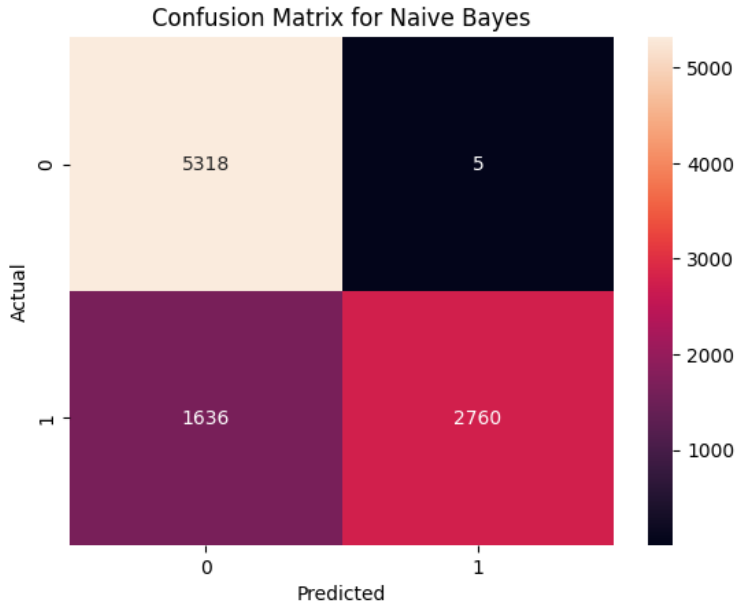
Accuracy: 83%	**Precision**: 100%
Recall: 62%	**F1 Score**: 77%

While the Logistic Regression model achieved commendable accuracy, its lower F1 score, driven by poor recall, indicates a potential issue of overfitting the dataset.

Code:

logreg_model = preprocess(X, LogisticRegression, max_iter=10000)
logreg_model.fit(X_train, y_train)

- **Naïve Bayes:-**



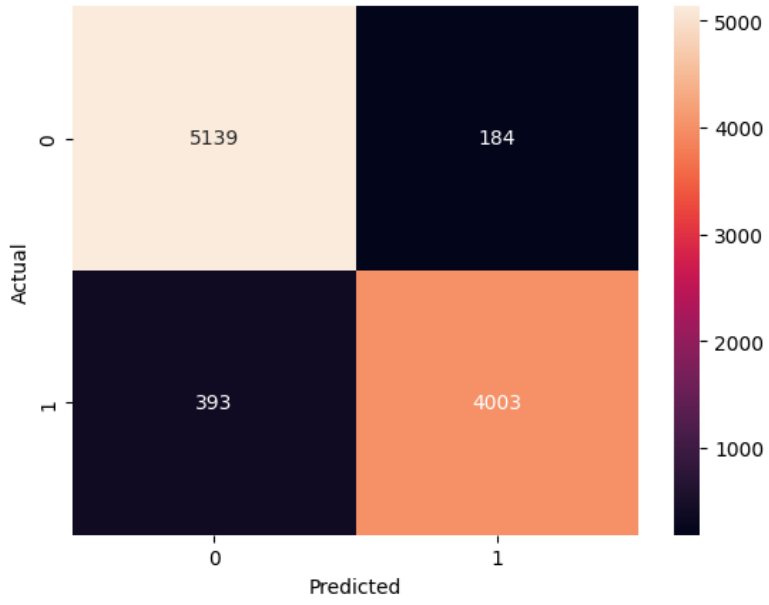
Accuracy: 83%	**Precision:** 99%
Recall: 63%	**F1 Score:** 77%

Similar to Logistic Regression, the Naïve Bayes model appears to overfit the data, achieving high precision but low recall, resulting in a suboptimal F1 score.

Code:

```
nb_model = preprocess(X, MultinomialNB)
nb_model.fit(X_train, y_train)
```

- **Random Forest Classifier:-**



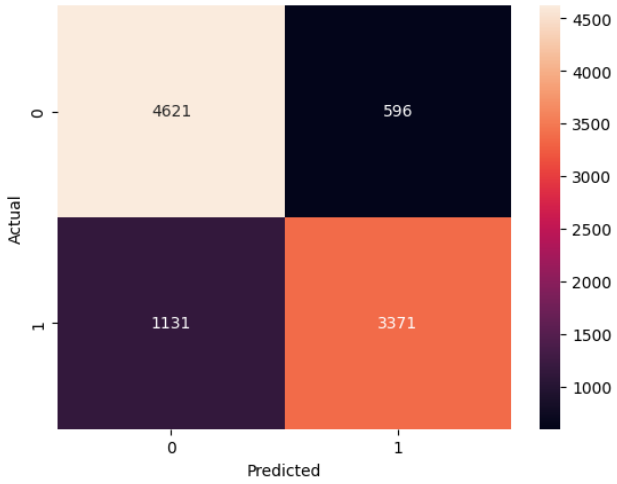
Accuracy: 94%	**Precision:** 96%
Recall: 92%	**F1 Score:** 94%

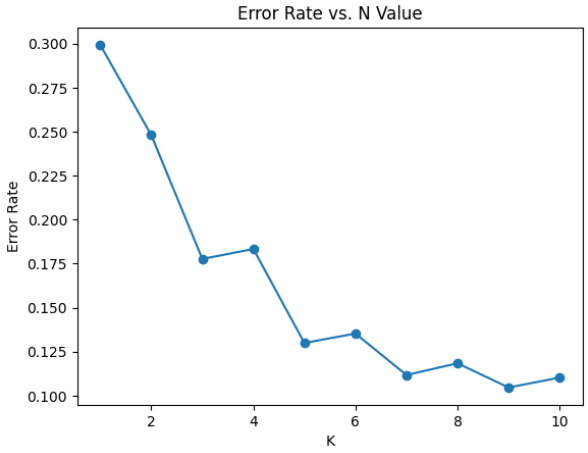
The high accuracy and strong F1 score of the Random Forest classifier suggest effective data partitioning, likely aided by the categorical nature of the dataset.

Code:

```
rf_model = preprocess(X, RandomForestClassifier, n_estimators=200)
rf_model.fit(X_train, y_train)
```

- **K-Nearest Neighbors:-**





Accuracy: 94%	**Precision:** 94%
Recall: 91%	**F1 Score:** 93%

The K-Nearest Neighbors (KNN) model performed well without signs of overfitting, indicating the possibility of smaller clusters where similar values consistently determine the class label.

Code:

```
knn_model = preprocess(X, KNeighborsClassifier, n_neighbors=3)
knn_model.fit(X_train, y_train)
```

3. **Validation and Optimization**

   - **Validation**:
     Data was split into training and testing sets (e.g., 80:20 split) to ensure fair evaluation.

   By integrating diverse models and rigorous analysis, this study provided valuable insights into factors influencing student mental health. The models' performance and limitations were critically evaluated, paving the way for future refinements and potential interventions. Visualizations and metrics supported data-driven decision-making, emphasizing the importance of comprehensive modeling strategies.

Conclusions and Discussion

Model Performance:
Various machine learning models were assessed for their ability to predict student depression. Random Forest and Logistic Regression demonstrated strong performance, achieving high accuracy scores (approximately 80% or more). In comparison, other models like Naive Bayes and K-Nearest Neighbors delivered lower accuracies, ranging from 60% to 75%.

Key Influencing Factors:
The analysis identified several key factors significantly associated with student depression, including marital status, parental involvement, access to resources, year of study, teacher quality, and peer influence. These findings highlight potential focal points for preventative measures. Based on these findings, the following measures could be considered effective actions for lawmakers

1. **Targeted Support**:
   High-risk students can benefit from tailored interventions such as counseling services, mental health workshops, and peer support groups.

2. **Improved Access to Resources**:
   Ensuring that students have access to adequate resources, including academic support, mental health services, and financial aid, could mitigate stressors linked to depression.

3. **Enhanced Parental Involvement**:
   Encouraging greater parental engagement in a student's academic and personal life may positively influence their mental well-being.

4. **Teacher Training**:
   Training educators to recognize and support students with mental health concerns could facilitate early identification and timely intervention.

5. **Peer Support Programs**:
   Establishing peer support networks can foster a sense of belonging and create a more inclusive environment for students

Limitations:
It is important to acknowledge that the analysis was conducted on a specific dataset, and the findings may not be generalizable to all student populations. Additional studies are necessary to confirm these results across diverse contexts.

Furthermore, the subjective nature of the study limits the reliability of a simple questionnaire. More accurate results could be achieved by employing a well-validated psychological assessment tool.

References

Student Mental Health Dataset: https://www.kaggle.com/datasets/shariful07/student-mental-health
Student Performance Dataset: https://www.kaggle.com/datasets/lainguyn123/student-performance-factors