Table of Contents:

Introduction	3
Dataset Description	3
Dataset Preprocessing	5
Feature Scaling	8
Dataset Splitting	8
Model Training and Test	9
Model Select and Comparison Analysis	13
Conclusion	15

Introduction

Depression among students has become an increasingly critical concern in educational institutions worldwide, significantly impacting academic performance, social relationships, and overall well-being. Early detection of depression symptoms can facilitate timely intervention and support for affected students. This project leverages machine learning techniques to develop a predictive model for identifying students at risk of depression, utilizing four distinct classification algorithms: K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Decision Tree. The rising prevalence of mental health issues in academic environments, particularly during and after the global pandemic, underscores the importance of developing automated screening tools. Machine learning approaches offer a promising solution by analyzing patterns in student behavioral data, academic performance indicators, and self-reported symptoms to identify potential cases of depression before they become severe. By comparing these different approaches, we aim to identify the most effective model for early depression detection in students. The findings of this study could contribute to the development of automated screening tools for educational institutions, helping identify at-risk students and enabling proactive mental health support.

Dataset Description

Dataset Link: Dataset

The data set is on Student Depression. The dataset consists of 18 columns and 27901 datapoints. There are three data types present in the dataset which are Object, Float, Int64.
The dataset consists of 18 columns. However after preprocessing the column number decreased to 14. So, the dataset has 13 features.
The dataset consists of categorical and quantitative data both.
Initially the Categorical features are: ‘Gender', 'City', 'Profession', 'Sleep Duration', 'Dietary Habits', 'Degree', 'Have you ever had suicidal thoughts ?', 'Family History of Mental Illness'
Quantitative features are: 'id', 'Age', 'Academic Pressure', 'Work Pressure', 'CGPA', 'Study Satisfaction', 'Job Satisfaction', 'Work/Study Hours', 'Financial Stress', 'Depression'
The target data is Depression which is formed with only 0 and 1 as value. So, it is a classification problem.

A heatmap using the Seaborn library is generated to visualize the correlation. There are correlation between Depression, academic pressure, and Have you ever had suicidal thoughts.

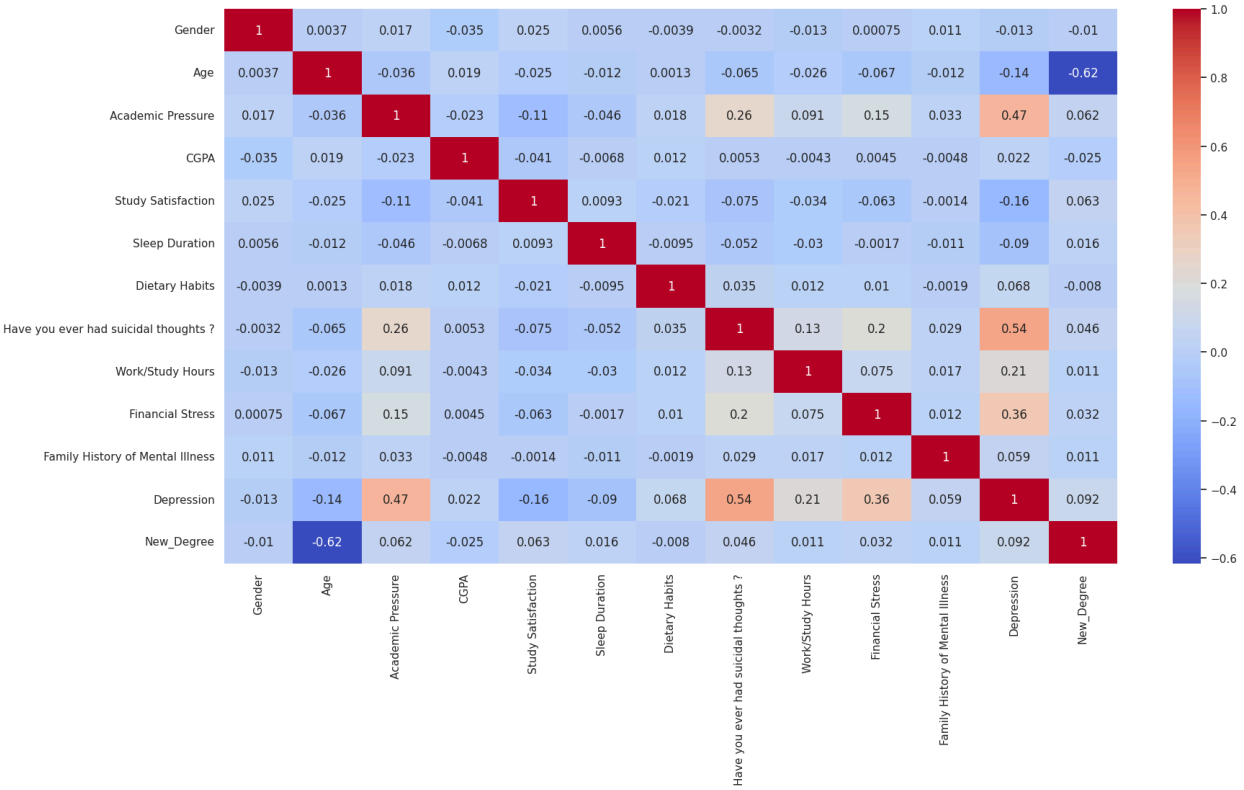
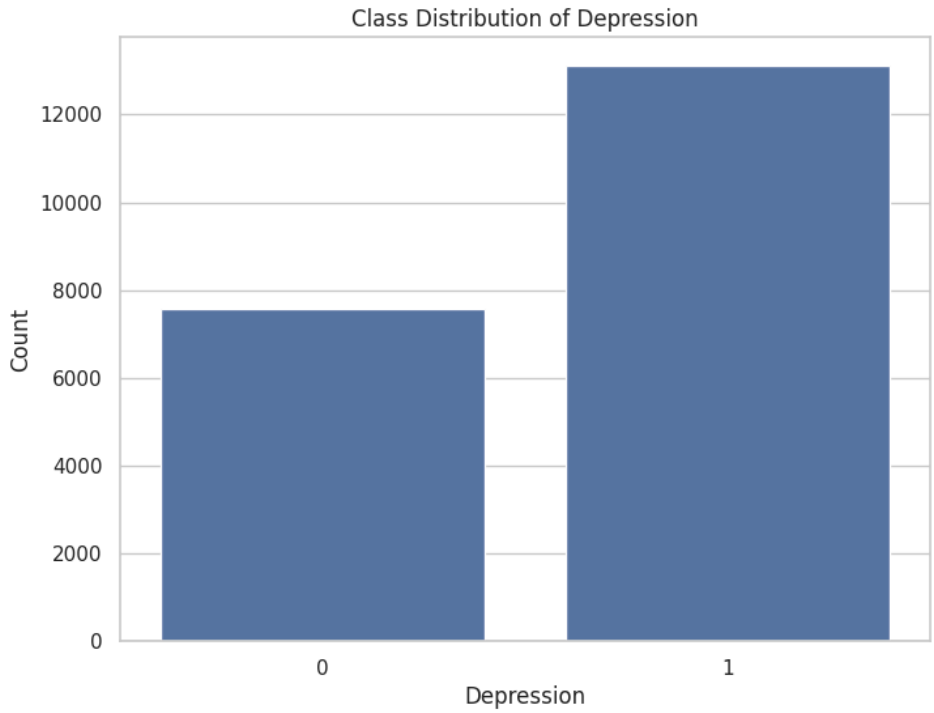


Fig: Correlation Heatmap

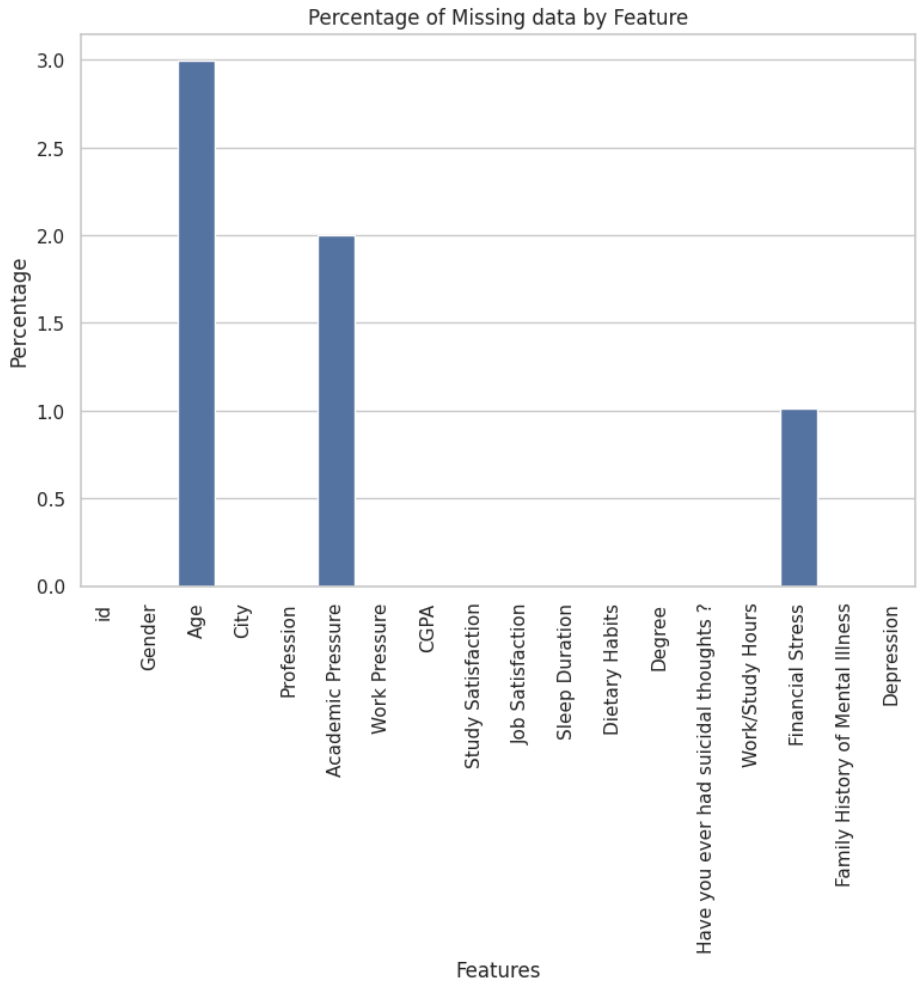
There is an imbalance in the dataset. The cases of depression are larger than the cases of not depression.



Dataset Imbalance

Dataset Preprocessing

There are many null values present in the dataset. In the preprocessing several steps were taken to deal with the NULL values.



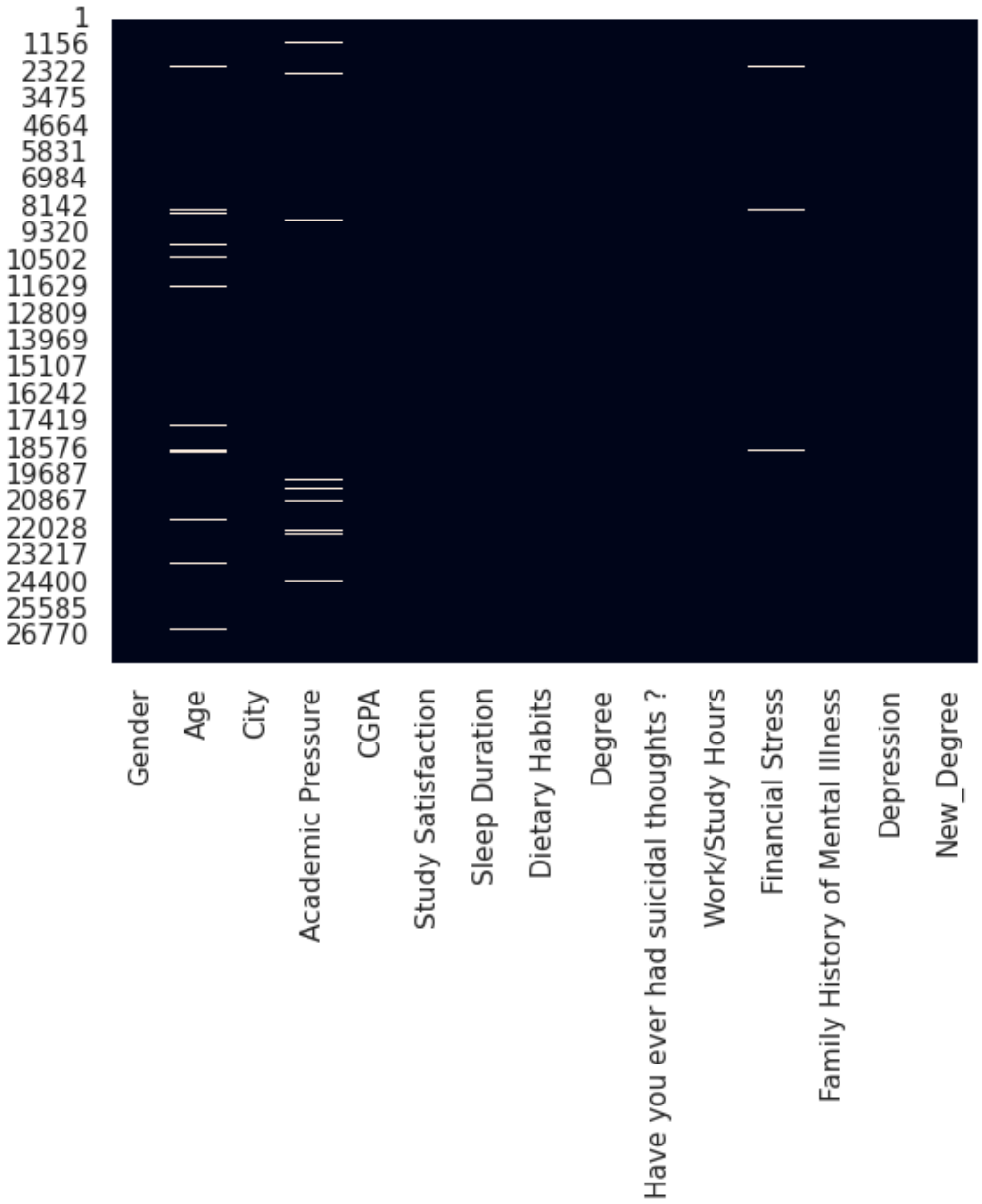
Missing Values in the dataset

As there were categorical and quantitative data, One hot encoding, binary encoding was used.

In the data preprocessing, it was observed that id, Work Pressure, Job satisfaction had no impact on the decision making, so all of these columns were dropped. Again a very large majority in the dataset have one profession which was student. This feature was one value based. So, the column was dropped to avoid unwanted data.

There are many categorical values in the dataset, Those values were converted to neumerical values for the model to understand the relation such as Sleep Duration, Dietary Habit, Have you ever had suicidal thoughts, Family History of Mental Illness. All of these column values are converted to numerical values.

City and Degree features of the dataset are categorical with many unique values. Firstly, cities with less than 400 hundred students were dropped to ensure no extreme value integers with our decision making. Secondly, One hot encoding is used to deal with many unique values of the City data column.. Binary encoding is used for columns with binary values like yes or no, 0 or 1. Lastly for filling up the missing value the mean was used in the dataset.



Feature Scaling

The data preprocessing part mentioned clearing extreme values, adjusting data points for categorical values with many unique values. The dataset underwent feature scaling as a crucial preprocessing step, normalizing the varying ranges of input variables to a common scale. This standardization process was essential to optimize the machine learning models' performance and ensure robust predictive capabilities.

Dataset Splitting

The dataset was split as follows:

Training Set: 70% of the data

Testing Set: 30% of the data

Model Training and Test

The following models were trained and evaluated:

Logistic Regression:

Logistic Regression Accuracy: 83.81%

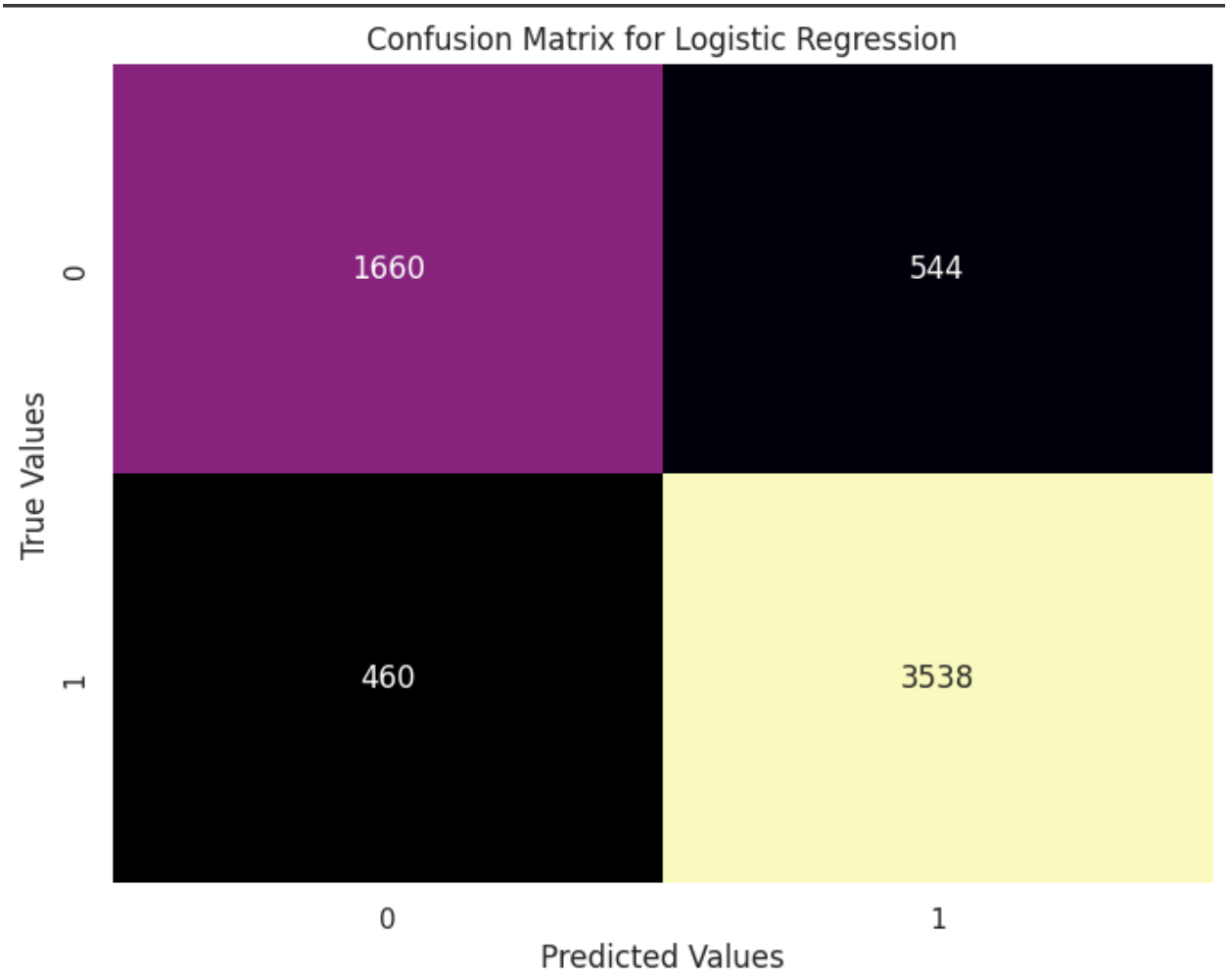


Fig. Confusion matrix for Logistic Regression

Decision Tree:

Decision Tree Accuracy: 77.44%

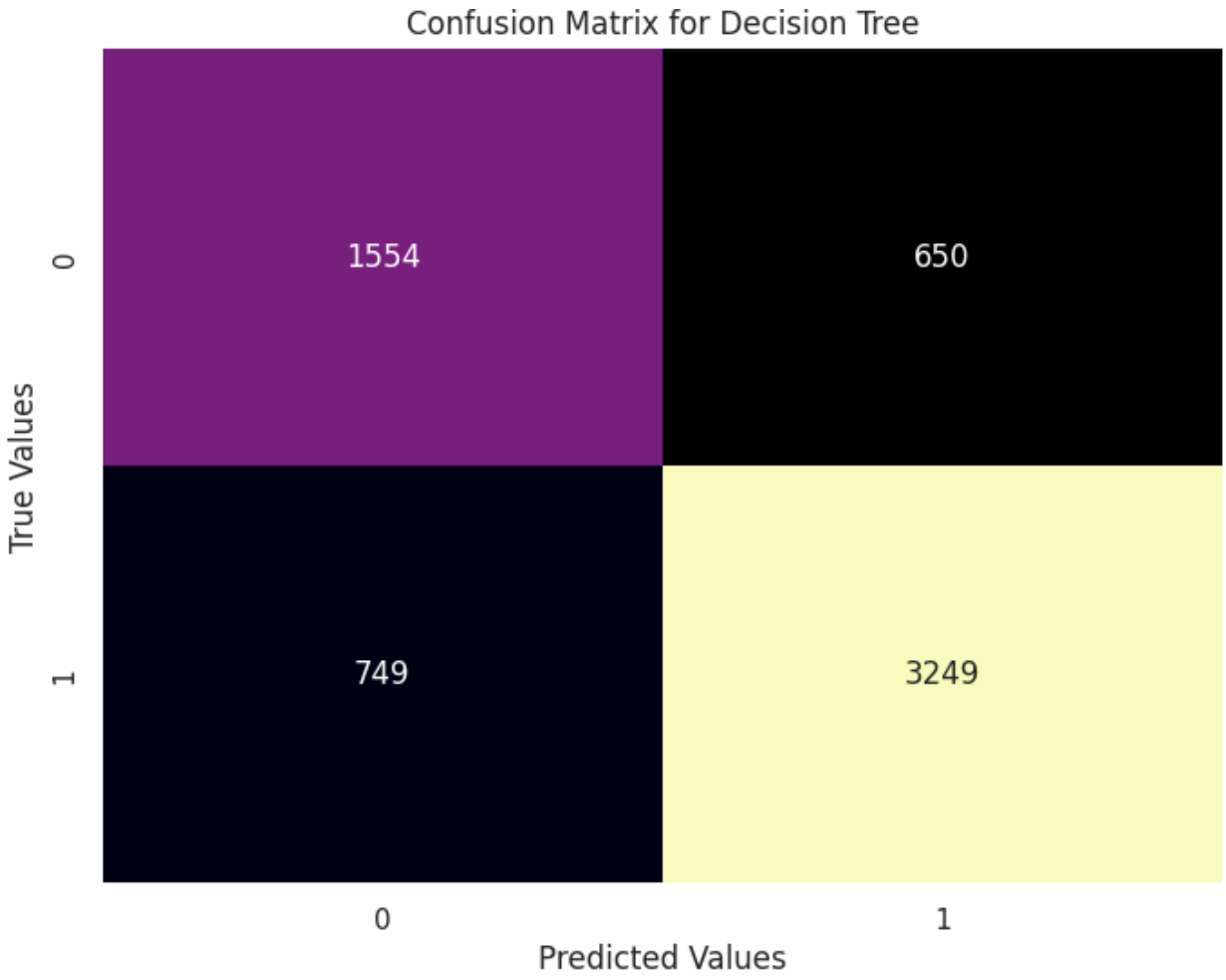


Fig. Confusion matrix for Decision Tree

K-Nearest Neighbors (KNN):

K-Nearest Neighbors Accuracy: 80.62%

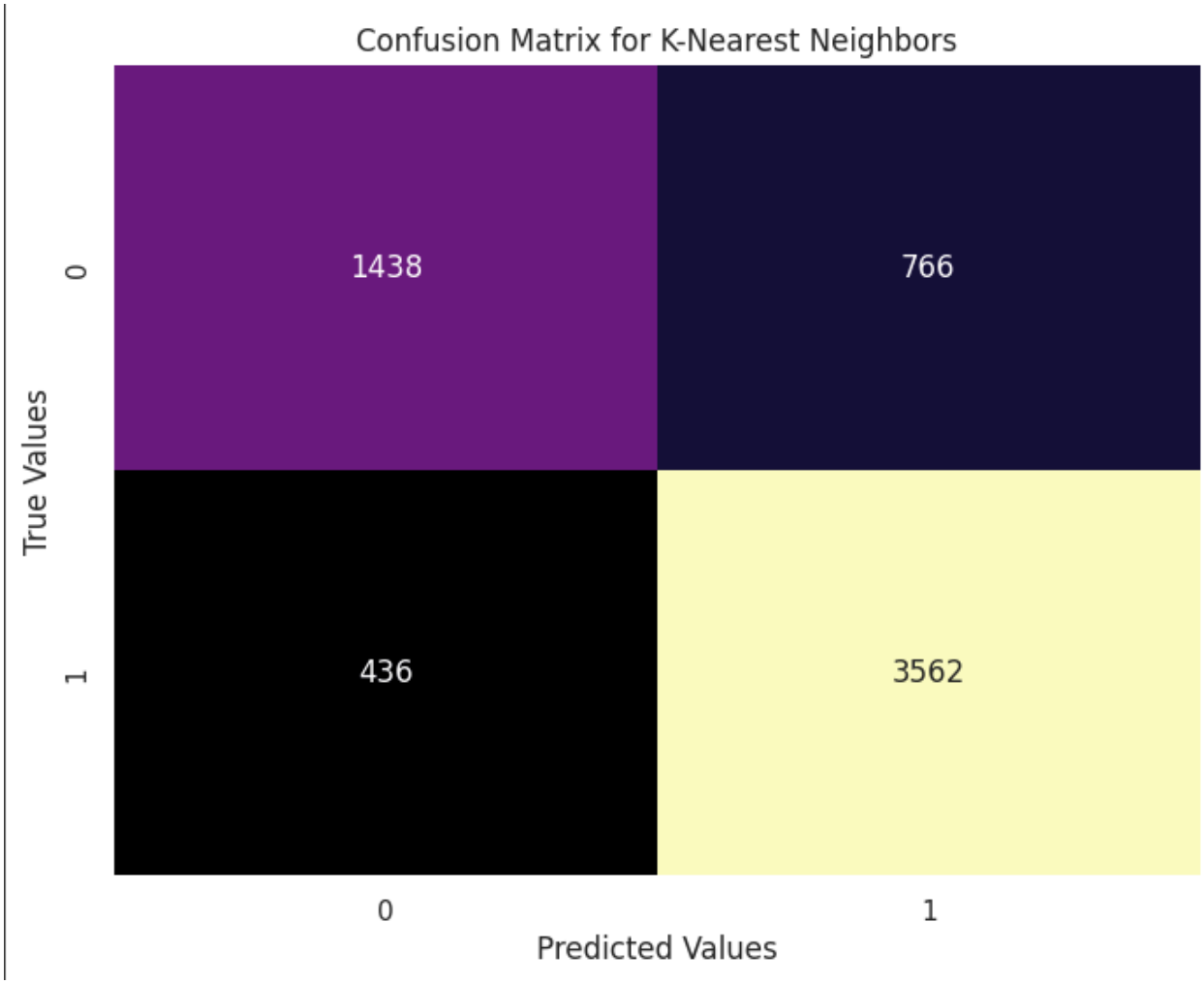


Fig. Confusion matrix for K-Nearest Neighbors

Naive Bayes:

Naive Bayes Accuracy: 75.78%

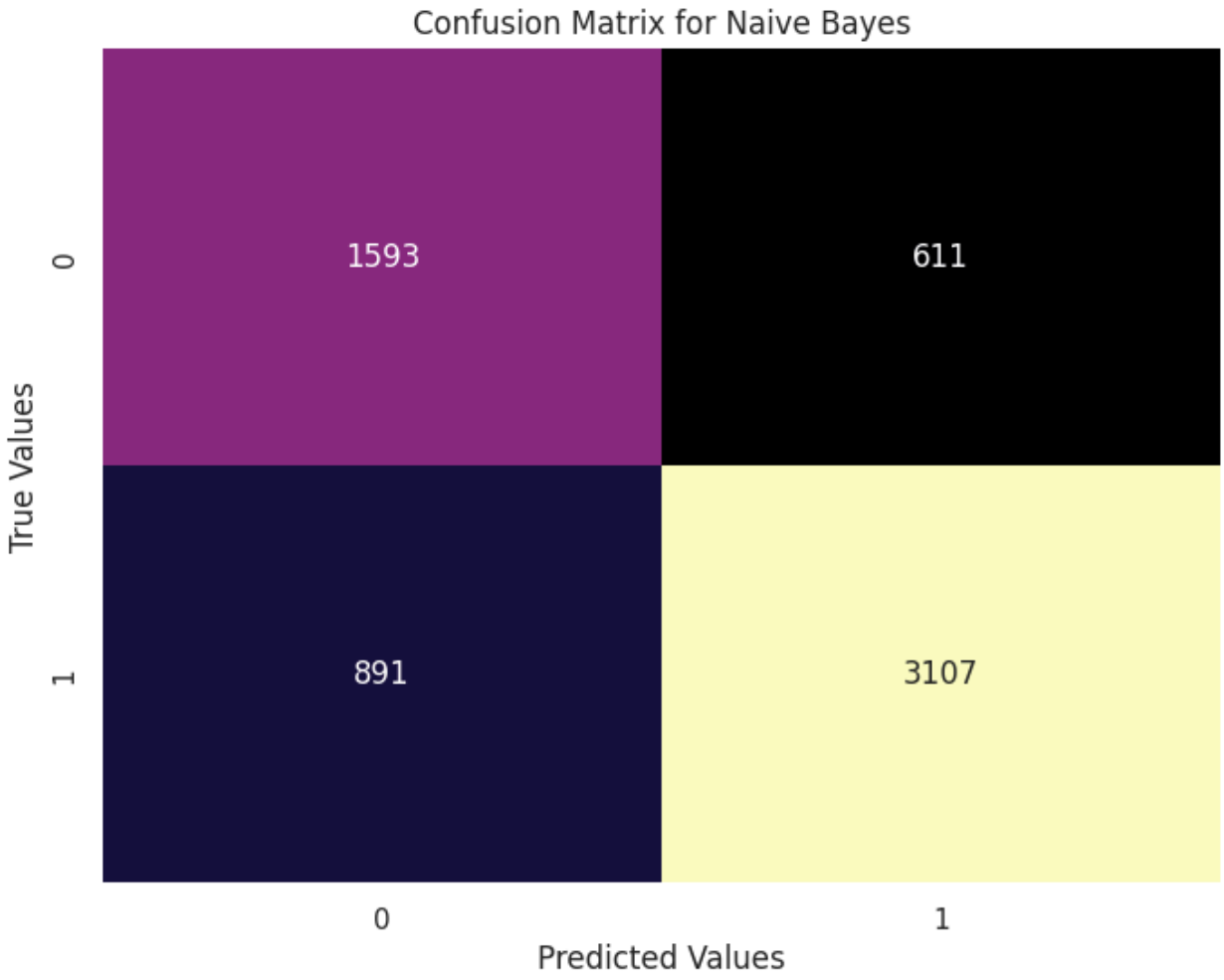


Fig. Confusion matrix for Naive Bayes

Model Select and Comparison Analysis

Accuracy Comparison:

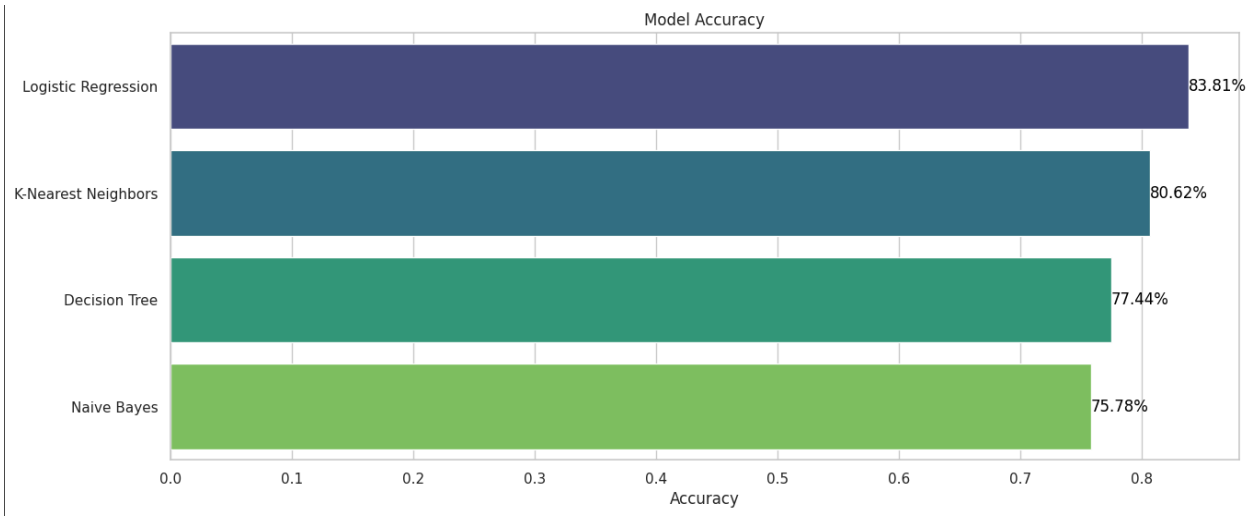
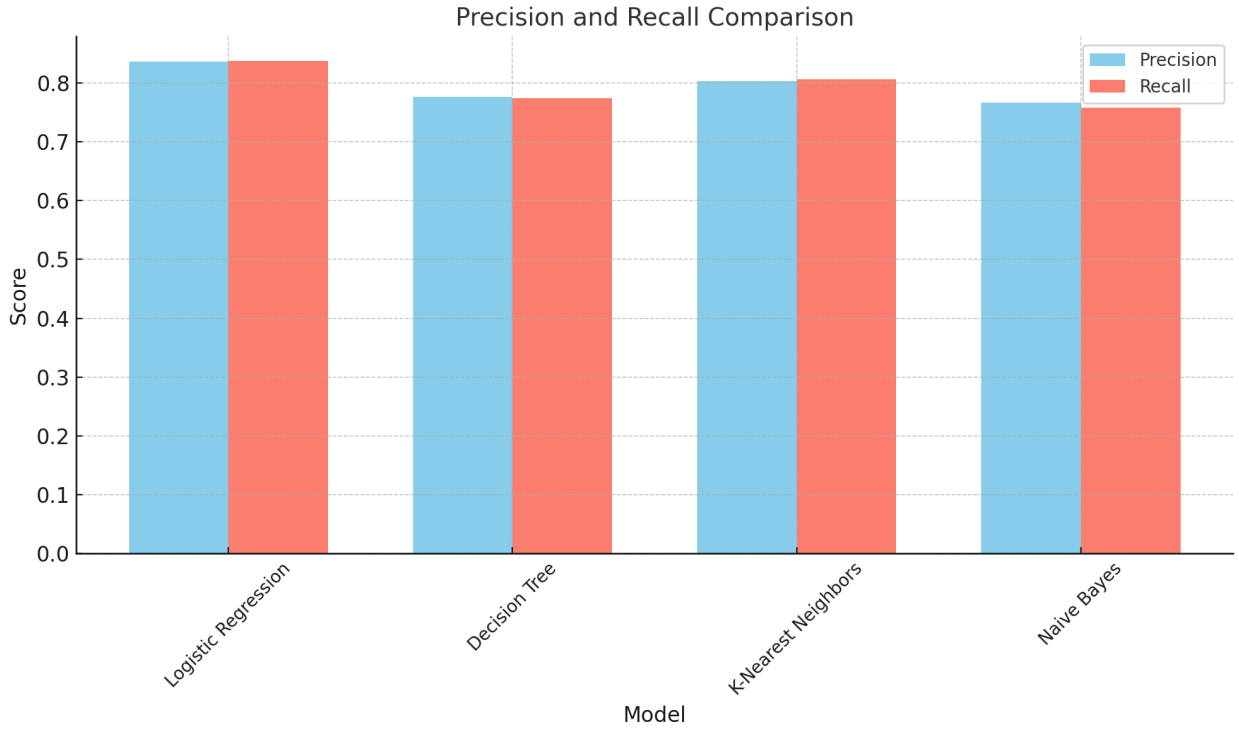


Fig: Accuracy Comparison between models

From the bar chart, it can be clearly visualized that, among the 4 models used, Logistic Regression has the highest prediction accuracy of 83.81%. KNN is following just behind with 80.62%. Decision tree has 77.44% accuracy and naive Bayes shows the worst accuracy performance with 75.78%.

Precision and Recall:



Conclusion

The project successfully developed and evaluated multiple machine learning models, including Logistic Regression, Decision Tree, K-Nearest Neighbors, and Naive Bayes. Logistic Regression achieved the highest precision (0.836983) and recall (0.838117), making it the most effective model for this dataset. K-Nearest Neighbors followed closely, with balanced performance metrics. Decision Tree and Naive Bayes had comparatively lower metrics, highlighting areas for improvement.

Future work could focus on hyperparameter tuning, exploring ensemble methods, and incorporating additional features or advanced techniques like Neural Networks to enhance performance further.