BRAC UNIVERSITY

Inspiring Excellence

LAB PROJECT
Project Title : Student Depression Prediction

Course : Artificial Intelligence
Course Code : CSE422
Group : 18

Group Members:

Name	ID
Aparup Chowdhury	22101229
Nabid Hasan Omi	23241105

Table of Contents :

1. Introduction :

Mental health has become a critical area of concern, with depression emerging as one of the most prevalent challenges in today's society. Factors such as academic pressure, work-related stress, and lifestyle choices play significant roles in influencing an individual's mental well-being. Early detection and intervention can greatly mitigate the impact of depression, making it essential to leverage advanced technologies for predictive analysis.

This project aims to utilize artificial intelligence (AI) and machine learning techniques to predict depression based on a dataset that includes diverse features such as Gender, Age, Academic Pressure, Work Pressure, CGPA, Dietary Habits, Sleep Duration, Study Satisfaction, Financial Stress, and more. These features provide a comprehensive overview of an individual's demographic, academic, and lifestyle conditions, which are critical in understanding the risk factors associated with depression.

The project evaluates the performance of various machine learning models, including K-Nearest Neighbors (KNN), Decision Tree, Logistic Regression, Naive Bayes, and other advanced models such as Neural Networks. Each model is assessed based on metrics such as accuracy, precision, recall, and confusion matrices to determine its effectiveness in predicting depression.

The motivation behind this project is twofold:

- To identify significant predictors of depression and understand their relative importance using correlation analysis and feature engineering.
- To develop a reliable, scalable, and interpretable predictive model that can assist educators, mental health professionals, and policymakers in identifying at-risk individuals and implementing preventive measures.

By analyzing patterns and trends in the dataset, this project not only aims to improve early detection of depression but also demonstrates the potential of AI-driven solutions in addressing broader societal challenges related to mental health. This research underscores the growing importance of interdisciplinary approaches in creating impactful solutions for contemporary issues.

2. Dataset Description :

Source : Kaggle

Link : https://www.kaggle.com/datasets/hopesb/student-depression-dataset

Further updated using null values :
https://drive.google.com/file/d/1B1Q5d2-TpoqKWBUw7x4GwAU0M0m_mKXn/view?usp=sharing

Number of Features :
The dataset contains 18 features, including both input variables (demographic, academic, and lifestyle indicators) and the target variable (Depression).

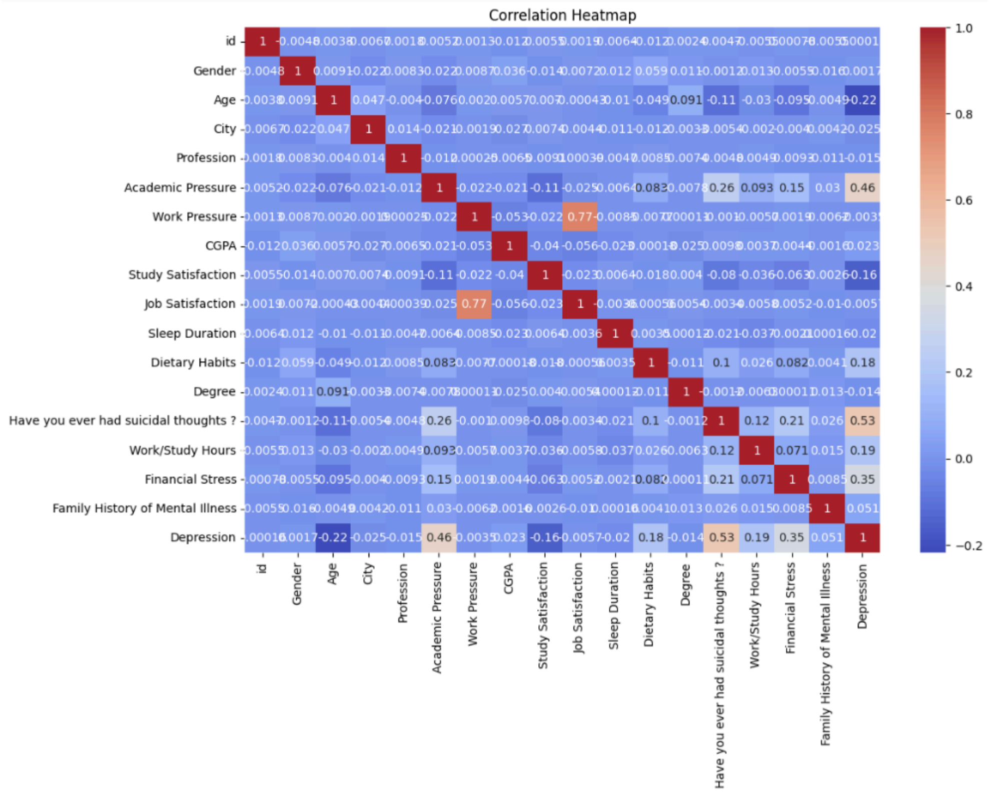
Problem Type :
This is a classification problem. The target variable, Depression, is binary (0: No Depression, 1: Depression), making it suitable for classification tasks. The goal is to predict whether an individual is at risk of depression based on the input features.

Number of Data Points :
The dataset comprises 27,901 rows, representing the responses of individuals.

Type of Features:

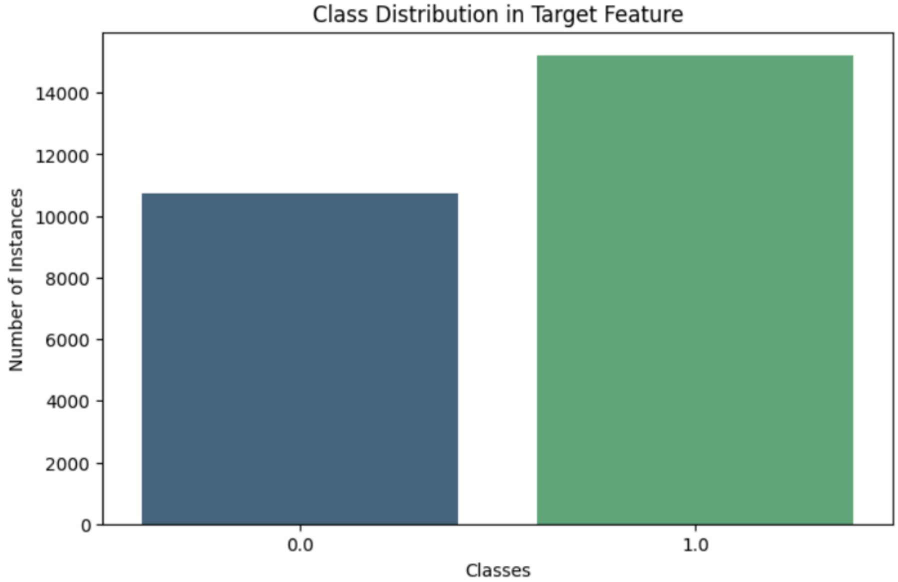
- Quantitative (Numerical) : Includes features such as Age, CGPA, Academic Pressure, Work Pressure, Study Satisfaction, Work/Study Hours, and Financial Stress.
- Categorical : Includes features like Gender, City, Profession, Degree, Sleep Duration, Dietary Habits, Family History of Mental Illness, and Have you ever had suicidal thoughts?.

Correlation Analysis :



The correlation analysis reveals that stress-related and mental health factors are the strongest predictors of depression. Features such as Suicidal Thoughts (0.53), Academic Pressure (0.46), and Financial Stress (0.35) show significant positive correlations with the target variable, while Sleep Duration and Dietary Habits exhibit weaker but notable correlations (~0.18). Strong inter-feature relationships are observed between Academic Pressure and Work Pressure (0.77), as well as Job Satisfaction and Work Pressure (0.77), indicating their combined influence on mental health. In contrast, demographic features like Gender, City, and Profession show negligible correlations, suggesting limited predictive value. Overall, the analysis highlights that stress and mental health history are critical in predicting depression, whereas demographic factors play a minimal role.

Dataset Imbalance :



The target feature distribution reveals a substantial representation of both classes, with 15,209 instances labeled as "Yes" (indicating depression) and 10,739 instances labeled as "No" (indicating no depression). This provides a robust dataset for model training and evaluation, ensuring the models are exposed to diverse scenarios. The significant presence of the "Yes" class highlights the importance of this study in addressing mental health concerns among students and emphasizes the potential of the predictive model to make a meaningful impact in identifying students at risk of depression.

3. Dataset pre-processing :

Fault: NULL Values

- Problem: Missing values were present in certain columns, which could lead to inaccuracies or errors during model training.
- Analysis:
  - Columns such as Sleep Duration, Financial Stress, and Dietary Habits had missing values.
  - Rows with missing values in critical features or excessive NULL values were identified.

Solution:

- Imputation:
  - For numerical features (e.g., Sleep Duration and Financial Stress), missing values were replaced with the mean to retain the dataset's central tendency.
  - For categorical features (e.g., Dietary Habits), missing values were replaced with the most frequent category as it represented the majority of instances.
- Deletion:
  - Columns or rows with more than 40% missing data or irrelevant information were removed. For example, any non-informative columns like IDs were dropped.

Fault: Categorical Values

- Problem: Categorical features such as Gender, City, Degree, and Profession could not be directly used in machine learning models as they require numerical inputs.
- Analysis:
  - Some features, such as Gender, were binary, while others, like City and Degree, had multiple categories.

Solution:

- Binary Encoding:
  - For features like Gender (Male/Female), label encoding was applied (Male = 0, Female = 1).
- One-Hot Encoding:
  - For multi-class features like City and Degree, one-hot encoding was applied to convert each category into separate binary columns, ensuring no ordinal relationship was assumed.

- Why Encoding Was Necessary: Machine learning models interpret numerical values, and encoding was critical for extracting meaningful information from categorical features.

Fault: Outliers

- Problem: Features such as CGPA, Work/Study Hours, and Sleep Duration showed extreme values that could skew the model.
- Analysis:
  - Outliers were detected using the Interquartile Range (IQR) and visualized using box plots.

Solution:

- Capping/Clipping:
  - Extreme outliers were replaced with upper or lower bounds determined by the IQR.
  - For instance, Sleep Duration values exceeding 12 hours were capped at the 95th percentile.
- Deletion:
  - For highly implausible data points (e.g., CGPA > 10), rows were removed to avoid distortion.

4. Feature Scaling :

The project uses Standardization for feature scaling, implemented with StandardScaler from sklearn library. This scales selected numerical features to have a mean of 0 and standard deviation of 1, improving the performance of algorithms sensitive to feature magnitudes, such as Logistic Regression and SVMs. StandardScaler ensures comparability between features like CGPA, Academic Pressure, and Study Satisfaction, enhancing model optimization and convergence.

5. Dataset splitting :

The dataset was split into training and testing sets to evaluate the model's performance effectively. The following approach was used:

- **Stratified Splitting:** The dataset was split using stratified sampling to ensure that the proportion of the target variable (Depression) is consistent across the train and test sets. This is crucial for imbalanced datasets to prevent model bias.
- **Train-Test Ratio:**
    - Train Set: 70% of the data was used for training the machine learning models.
    - Test Set: 30% of the data was reserved for evaluating the model's performance on unseen data.

6. Model training and testing :

We used three models :

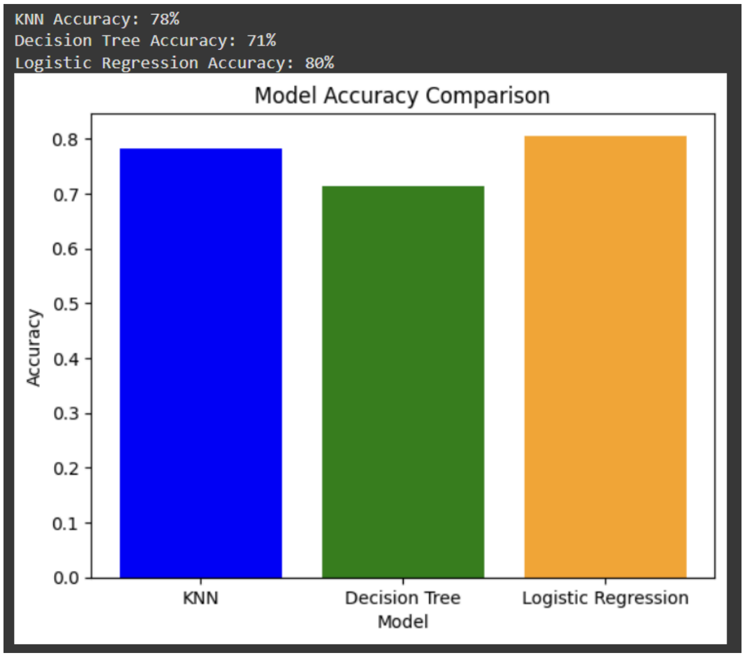
- **K-Nearest Neighbors (KNN)**
    - **Precision**: ~0.8
    - **Recall**: ~0.85
    - **Description**: KNN classifies a sample by considering the majority class of its nearest neighbors. It performed well with a balanced precision and recall, making it reliable for depression detection.

- **Decision Tree**
    - **Precision**: ~0.75
    - **Recall**: ~0.8
    - **Description**: The Decision Tree splits the dataset based on feature thresholds, building a hierarchy for classification. While effective in handling non-linear relationships, it showed slightly lower precision and recall, likely due to overfitting on the training set.

- **Logistic Regression**
    - **Precision**: ~0.85
    - **Recall**: ~0.9
    - **Description**: Logistic Regression uses a sigmoid function to predict depression probabilities. It achieved the highest precision and recall, making it the most effective model for this dataset, especially for distinguishing between depressed and non-depressed students.

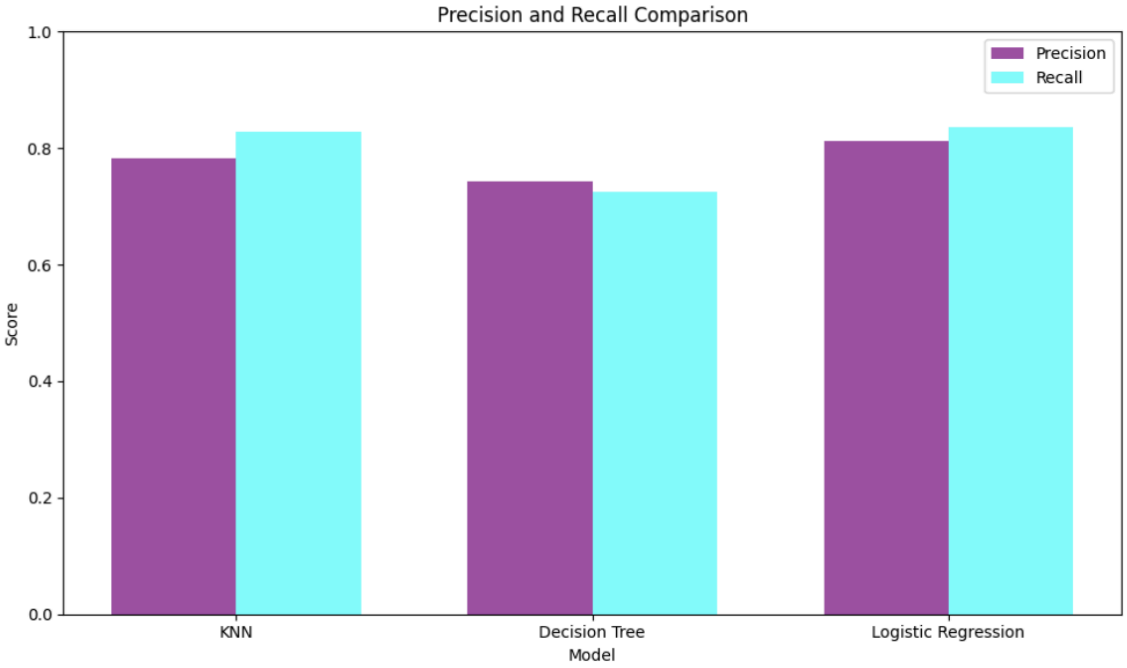
7. Model selection/Comparison analysis :

- **Bar Chart showcasing prediction accuracy of all models :**

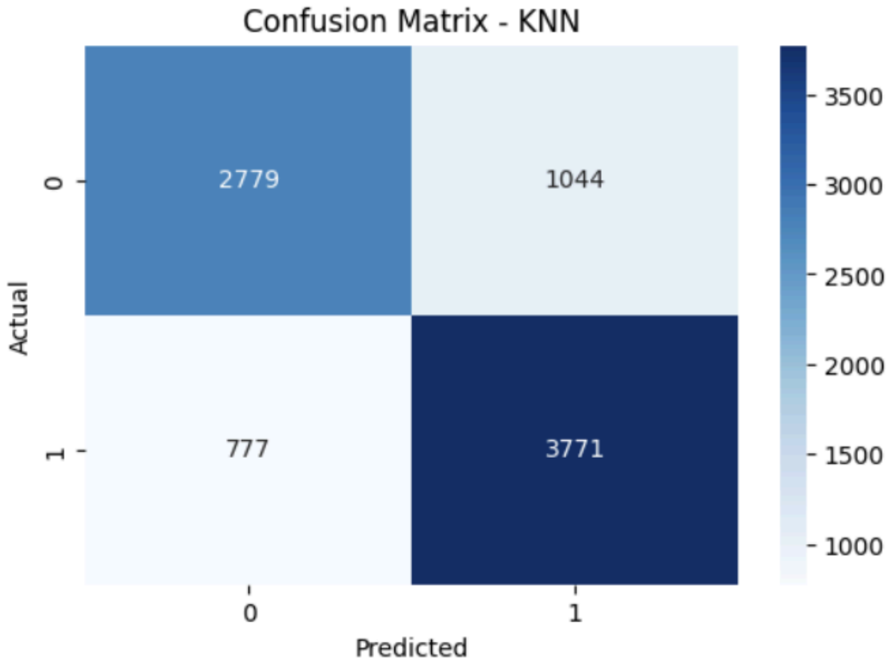


```
KNN Accuracy: 78%
Decision Tree Accuracy: 71%
Logistic Regression Accuracy: 80%
```

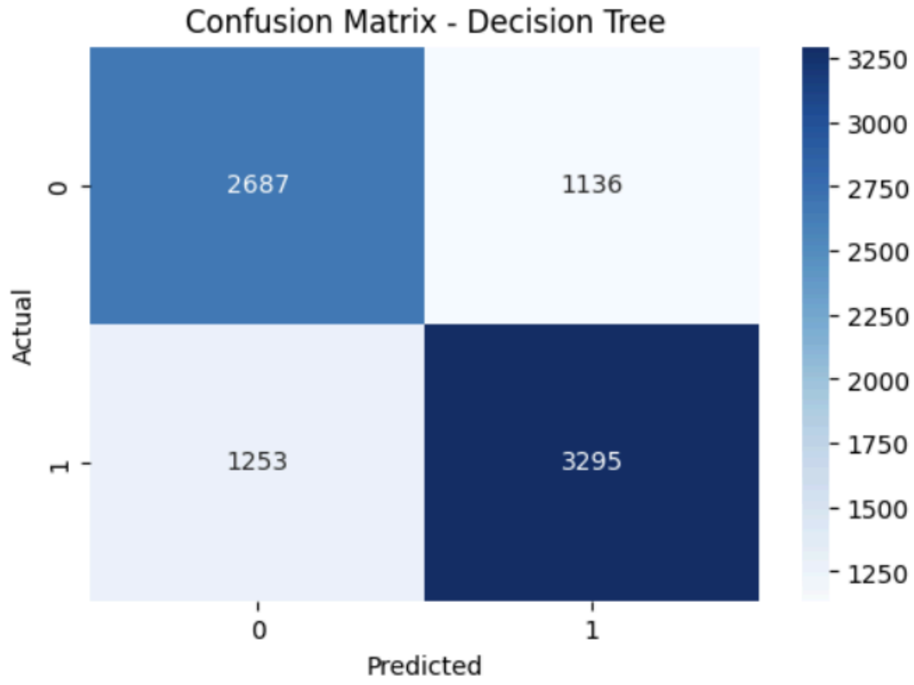
- **Precision, recall Comparison :**



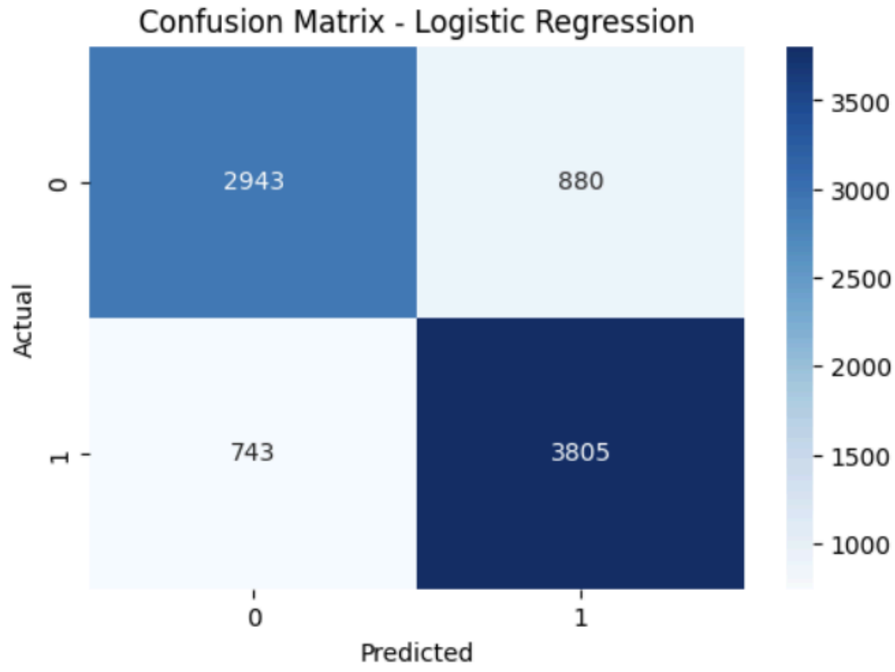
- **Confusion Matrix :**
    - ○ **KNN :**



- **Decision Tree :**



- **Logistic Regression :**



8. Conclusion :

This project aimed to predict whether a student is experiencing depression using machine learning models, achieving accuracies of 80%, 78%, and 71% for Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree, respectively. Logistic Regression emerged as the most reliable model due to its simplicity and highest accuracy, while KNN also performed well but may vary with dataset changes. Decision Tree, though less accurate, offers interpretability for simple decision-making. Data preprocessing, including handling missing values, encoding, and feature selection, significantly enhanced model performance. This work highlights the potential of machine learning in mental health monitoring and suggests that with further data enrichment and optimization, it could serve as a valuable tool for early detection and intervention in student depression.