**Project Report**

Depression Prediction among Students

**1. Problem Definition**

Problem Statement

The goal of this project is to predict whether a student is likely to experience depression based on various factors such as academic pressure, financial stress, study hours, and personal habits. By analyzing the features provided in the dataset, this model can assist in identifying at-risk students, providing insights that can be used by mental health professionals and educational institutions to intervene early.

**Objectives**

* Analyze the data to uncover trends and relationships between various factors and depression.
* Build and evaluate machine learning models to predict depression in students.
* Provide actionable insights for mental health interventions based on data findings.
* Significance
* Student Well-being: Depression in students can severely impact their academic performance and overall health. Predicting depression can enable timely interventions.
* Educational Support: Schools and colleges can use the results to develop better support systems for students facing academic and financial stress.
* Mental Health Research: This study can contribute to the broader research on factors affecting student mental health.

**2. Asking the Right Questions**

To guide the analysis, the following key questions were formulated:

* What are the key factors influencing depression among students?
* How can we predict depression levels using features like academic pressure, study hours, and financial stress?
* What is the relationship between sleep duration, diet, and depression?
* Can machine learning models accurately predict depression based on available data?

**3. Data Collection**

**Dataset Source**

**Source:** The dataset is sourced from Kaggle, specifically from the Depression Student Dataset.

**Description:** This dataset includes survey responses from students about their academic pressure, study habits, sleep duration, family history of mental illness, and other lifestyle factors.

**Features:** Age, Academic Pressure, Study Satisfaction, Study Hours, Financial Stress, Gender, Suicidal Thoughts, Family History of Mental Illness, Dietary Habits, Depression (target).

**4. Data Wrangling (Preprocessing)**

**Steps Performed**

**Missing Value Handling:** Missing values were checked from the dataset to ensure a clean dataset for modeling.

**Feature Encoding:** Categorical variables such as Gender, Suicidal Thoughts, and Dietary Habits were encoded using LabelEncoder to convert them into numerical format for machine learning models.

**Feature Selection:** The relevant numeric features were selected, including Age, Academic Pressure, Study Hours, and Financial Stress, which were most correlated with the target variable Depression.

**Outlier Detection:** Outliers in the numeric features were detected using the IQR method, and box plots were used to visualize these outliers.

**5. Exploratory Data Analysis (EDA)**

Key Insights from EDA

**Distribution of Numeric Features:** Plots showed that variables like Age and Academic Pressure had varying distributions, which might indicate their relationship with depression.

**Categorical Feature Distribution:** Count plots revealed the distribution of categorical features like Gender, Suicidal Thoughts, and Dietary Habits. The majority of respondents had varying levels of study satisfaction, and gender distribution was fairly balanced.

**Correlation Analysis:** A heatmap of the correlation matrix showed strong positive correlations between Academic Pressure, Study Hours, and Depression. This suggests that higher academic pressure and more study hours may correlate with an increased likelihood of depression.

Features like Financial Stress and Age also showed correlations with depression.

**6. Predictive Modeling**

Models Used:

**Decision Tree Classifier:**

A Decision Tree model was trained on the dataset to predict depression. The model achieved an accuracy of approximately 0.77, and the classification report showed good precision and recall for the depression target.

Decision Tree Visualization: The decision tree visualized the most important features such as Study Hours and Financial Stress, which were used to split the data.

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

The KNN model, with 5 neighbors, also predicted depression with an accuracy of 0.72. The confusion matrix helped assess the true positives and false positives for both classes.

KNN Confusion Matrix: The confusion matrix was plotted to evaluate the model’s performance on the test set, showing how well it classified students into "Depressed" vs. "Not Depressed" categories.

**Random Forest Classifier:**

The Random Forest classifier performed the best, with an accuracy of 0.80, suggesting that ensemble methods improve the predictive capability over single models like Decision Tree and KNN.

**7. Conclusion and Future Work**

**Summary**

The analysis identified several key factors influencing depression in students, including Academic Pressure, Study Hours, and Financial Stress. The Random Forest model was the most accurate, showing the importance of using ensemble techniques for classification tasks.

**Key Insights:**

**Model Performance:** Random Forest outperformed Decision Tree and KNN in terms of accuracy, making it the most reliable model for this dataset.

**Factors Affecting Depression:** Features such as Study Hours and Academic Pressure were significant predictors of depression.

**Future Work:** Improved Model Tuning: Hyperparameter tuning could be applied to further improve the performance of the models.

**External Validation:** Using an external dataset to validate the findings and generalize the results.

**Feature Expansion:** Additional features such as social media usage or sleep patterns could be incorporated to improve predictions.