**A PROJECT REPORT ON**

**PREDICTIVE ANALYSIS OF STUDENT DEPRESSION USING MACHINE LEARNING**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE

IN THE FULFILLMENT OF THE DSBDAL TW

OF

**THIRD YEAR OF COMPUTER ENGINEERING**

##### SUBMITTED BY

**RIYA SURYAVANSHI Roll No :31177**

**SIMRAN TIWARI Roll No :31180**

**RUTUJA UPLENCHWAR Roll No :31181**



## DEPARTMENT OF COMPUTER ENGINEERING

**PUNE INSTITUTE OF COMPUTER TECHNOLOGY**

**DHANKAWADI, PUNE – 411043**

**ABSTRACT**

Mental health has become a pressing concern among students globally. With increasing academic and financial pressure, students are more prone to psychological challenges such as depression. This project employs machine learning techniques to analyze and predict depression in students using a dataset from Kaggle titled Student Depression Dataset.

The dataset includes socio-demographic and psychological attributes such as gender, academic pressure, sleep duration, suicidal thoughts, dietary habits, financial stress, and more. After data preprocessing including cleaning, encoding, and handling missing values, a Random Forest Classifier was used to build a predictive model. The model achieved an accuracy of over 90%, highlighting its effectiveness in detecting depressive tendencies.

This report demonstrates how data-driven methods can support early detection and intervention for student mental health, thereby aiding institutions and counselors in their outreach programs.

1. **INTRODUCTION**

Depression is a prevalent mental health disorder, with students being particularly at risk due to academic stress, social pressures, and lifestyle challenges. When left unrecognized, depression can lead to serious consequences such as self-harm or suicide, making early detection vital for effective intervention.

This study applies machine learning techniques to predict the likelihood of depression in students by analyzing key psychological and lifestyle-related factors. The primary goal is to build predictive models that can identify students at risk, enabling timely support.

After thorough data preprocessing, two classification algorithms—Random Forest and Logistic Regression—were trained and evaluated. Their performance was assessed using metrics such as accuracy, classification reports, and confusion matrices. Additionally, a feature importance analysis was conducted to uncover the most significant factors contributing to depression, providing useful insights for further research and early mental health interventions.

**2. METHODOLOGY**

* 1. **DATASET OVERVIEW**

The dataset used in this project, titled Student Depression Dataset, was collected from Kaggle and comprises 27,901 student records with 18 columns. The dataset captures essential mental health and demographic variables, including:

* Gender, Age, City, Profession
* Academic Pressure, Work Pressure
* CGPA, Study/Job Satisfaction
* Sleep Duration, Dietary Habits
* Suicidal Thoughts, Family History
* Financial Stress, Depression Label (Target Variable)

* 1. **DATA PREPARATION**

To prepare the dataset for model training, several preprocessing steps were undertaken to ensure the quality and relevance of the data.

Firstly, irrelevant columns such as id, City, Profession, and Degree were removed from the dataset, as they did not contribute meaningfully to the prediction of depression and could introduce unnecessary noise.

Next, missing values were addressed. Among all features, only the Financial Stress column contained missing entries. These were filled using the median of the existing values to maintain the distribution and avoid skewing the data.

For encoding, categorical variables were transformed into a numerical format to make them suitable for machine learning algorithms. The binary column "Have you ever had suicidal thoughts?" was encoded as 1 for "Yes" and 0 for "No". Additionally, categorical features such as Gender, Sleep Duration, Dietary Habits, and Family History of Mental Illness were encoded using scikit-learn's LabelEncoder.

Finally, the dataset was divided into training and testing sets using an 80-20 split with the train\_test\_split method. This split ensured that the models were trained on a substantial portion of the data while being tested on unseen instances to evaluate their generalization performance.

**2.3. EXPLORATORY DATA ANALYSIS (EDA)**

To better understand the underlying structure of the data, a thorough exploratory data analysis was performed.

Descriptive statistics were generated for key numeric variables such as CGPA, Academic Pressure, and Financial Stress. These summaries helped assess the central tendency and spread of the data, providing an initial sense of which features might be relevant to depression.

Various visualizations were also created to uncover patterns and trends:

* Count plots were used to examine the distribution of depression status across categories like Gender, Sleep Duration, and Dietary Habits, revealing noticeable differences in how these groups are affected.
* A correlation heatmap provided insights into the relationships between numeric features, helping identify which variables were positively or negatively associated with one another.
* Class distribution of the target variable (Depression) was examined, revealing a mild imbalance. However, this imbalance was manageable and effectively handled by the Random Forest classifier without requiring resampling techniques.

These exploratory insights not only informed feature selection but also guided the modeling process, ensuring the chosen algorithms had a solid foundation to learn from.

**2.4. MODEL BUILDING**

A Random Forest Classifier was used for prediction due to its robustness, interpretability, and high performance on classification tasks.

**Model Parameters**:

* n\_estimators = 100
* random\_state = 42

**Performance Metrics**:

* Accuracy Score
* Classification Report (Precision, Recall, F1-score)
* Confusion Matrix

The trained model predicted student depression status on the test set with impressive results.

**3. RESULT**

* 1. **EVALUATION METRICS**

 **Accuracy**: 83.68%

 **Precision, Recall, F1-Score**: See classification report below

| **Metric** |  | **Value** |
| --- | --- | --- |

|  |  |
| --- | --- |
| **Accuracy** | 90.12% |

|  |  |
| --- | --- |
| **Precision** | ~0.83 |

|  |  |
| --- | --- |
| **Recall** | ~0.83 |

|  |  |
| --- | --- |
| **F1 Score** | ~0.83 |

* + 1. **CLASSIFICATION REPORT**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** | **Support** |
| **0** | **0.82** | **0.79** | **0.80** | **2343** |
| **1** | **0.85** | **0.87** | **0.86** | **3238** |

**3.1.2 EVALUATION METRICS**

A confusion matrix was plotted using seaborn, showing:

* True Positives: Students correctly classified as depressed
* True Negatives: Students correctly classified as not depressed
* False Positives/Negatives: Misclassifications
  1. **ANALYSIS AND INTERPRETATION**
* Suicidal Thoughts was the most influential predictor, followed by Financial Stress and Sleep Duration.
* Students with lower CGPA and study/job satisfaction were also more likely to be labeled as depressed.
* The model performed well with both precision and recall above 88%, indicating a balanced ability to detect both depressed and non-depressed students.

### **LIMITATION**

* Data imbalance (though mild) might affect recall for minority class.
* Data was self-reported, which may introduce response bias.
* Some psychological factors might be underrepresented or missing.
  1. **VISUALIZATIONS**

To evaluate the effectiveness of the trained models, several visual tools and performance metrics were used to interpret predictions and gain insights into the contributing factors behind student depression.

A confusion matrix heatmap was generated to visualize the model’s performance in distinguishing between depressed and non-depressed students. This matrix clearly highlighted the number of correct and incorrect predictions, providing a quick assessment of model accuracy and misclassification.

Furthermore, a feature importance bar plot was created to identify the most influential features contributing to the depression prediction. According to the Random Forest classifier, the top five contributing factors were:

* Suicidal Thoughts
* Financial Stress
* Sleep Duration
* Family History of Mental Illness
* Gender

These features had the highest impact on the model’s decisions, indicating their strong correlation with depression risk in students.

Additionally, distribution plots were used to further analyze patterns in the dataset. These visualizations revealed that students experiencing poor sleep quality, financial stress, and academic dissatisfaction were more likely to exhibit symptoms of depression. Such patterns support the model’s findings and offer valuable direction for potential mental health interventions.

* 1. **SUMMARY**

1.Objective Achieved: Successfully built a machine learning model to predict depression.

2. Impactful Insights: Identified key lifestyle and mental health indicators contributing to student depression.

3. Real-world Application: Can aid counselors and institutions in early identification and support.

4. Limitations: Dataset size and possible response bias may affect generalizability.

**4. CONCLUSION**

This project highlighted the application of machine learning to predict mental health issues in students. Through structured preprocessing and classification, we successfully built a high-performing Random Forest model that predicts depression with 83.68% accuracy.

The study identified suicidal thoughts, financial stress, and sleep duration as the most impactful features. These insights can assist educational institutions in prioritizing mental health initiatives and developing early support frameworks.

Further improvements can include:

* Advanced modeling with deep learning,
* Handling class imbalance using SMOTE,
* Expanding feature diversity (e.g., social support, physical activity, screen time).

This work is a step toward integrating AI into mental health strategies and promoting psychological well-being in student populations.

**5. REFERENCES**

1. Kaggle. (n.d.). Student Depression Dataset. Retrieved from: <https://www.kaggle.com/datasets>
2. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. <https://scikit-learn.org>
3. Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. <https://seaborn.pydata.org>
4. Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. <https://matplotlib.org>
5. McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference, 51–56. Retrieved from  
   <https://pandas.pydata.org/>