HackBio Internship 2025

Final Report

Topic:
Neuroscience & Psychology

Contributor
Muhammad Bilal Ashraf
Slack id: bilal ashraf
GitHub Username: bilal265

Table of Contents

1) Executive Summary

2) Introduction

3) Dataset Description

4) Data Preprocessing

5) Methodology

6) Evaluation Metrics

7) Results

8) Feature Importance Analysis

9) Practical Advice for Those Who Might Be Depressed

10) Ethical Considerations

11) Future Improvements

12) Conclusion

13) Key Citations

Executive Summary

This project aims to develop a machine learning model to detect depression in university students using various psychological, academic, and lifestyle factors. The model can potentially be used in early intervention strategies to provide mental health support. The report outlines the methodology, results, and implications of this work.

Introduction:

Depression is a common mental health disorder characterized by persistent sadness, loss of interest, and impaired daily functioning. Among university students, it manifests as a significant public health issue due to unique stressors like academic demands, financial pressures, and social transitions. Studies, including your dataset, suggest that 30-60% of university students experience depressive symptoms at some point, with rates varying by region and institution

Depression among university students is a growing concern, impacting academic performance, social interactions, and long-term well-being. This analysis leverages machine learning techniques to develop a classification model for detecting depression, using a dataset with 259 rows after filtering incomplete entries. The goal is to identify at-risk students for early intervention, providing insights into key features and practical advice for those affected.

Dataset Description:

The dataset includes 5,581 samples with 15 features and a binary target variable (Depression).

Features:

- **Demographics:** Gender, Age, Degree, Profession
- **Academic & Work Factors**: Academic Pressure, Work Pressure, CGPA, Study/WorkHours
- **Lifestyle Factors:** Sleep Duration, Dietary Habits
- **Mental Health Indicators:** Suicidal Thoughts, Family History of Mental Illness

- **Depression Label:** Target variable (Depressed / Not Depressed)

From the sample, the dataset shows 54.8% of students (142/259) report depression, indicating a significant prevalence. Features like "City" and "id" were dropped as irrelevant, focusing on predictive variables.

Data Preprocessing:

To prepare the data for modeling:

- **Missing Values:** Assumed handled, with no missing values observed in the sample.
- **Encoding Categorical Variables:** Used LabelEncoder for Gender, Sleep Duration, Dietary Habits, Degree, Suicidal Thoughts, and Family History of Mental Illness. For instance, Sleep Duration ("Less than 5 hours", "5-6 hours", etc.) was treated as ordered, and Degree, despite many categories, was label-encoded for simplicity.
- **Scaling Numerical Features:** Applied StandardScaler to Age, Academic Pressure, Work Pressure, CGPA, Study Satisfaction, Job Satisfaction, Work/Study Hours, and Financial Stress, ensuring models like Logistic Regression perform well.
- **Data Splitting:** Used train_test_split with an 80-20 split, stratifying by the target variable to maintain the proportion of depressed vs. non-depressed students.

Methodology:

Three classification models were implemented and compared:

1) **Logistic Regression (Baseline Model):** A simple linear model to establish a baseline, fitting the data with max_iter=1000 for convergence.
2) **Random Forest Classifier:** Captured non-linear relationships, with n_estimators=100, useful for feature importance analysis.
3) **XGBoost Classifier:** An advanced boosting model, using eval_metric='logloss' for binary classification.

Hyperparameter Tuning:

GridSearchCV was used to optimize:

1) n_estimators: [50, 100, 200]
2) max_depth: [None, 10, 20]
3) min_samples_split: [2, 5, 10]
4) min_samples_leaf: [1, 2, 4]

Evaluation Metrics:

Given the sensitivity of depression detection, metrics focused on:

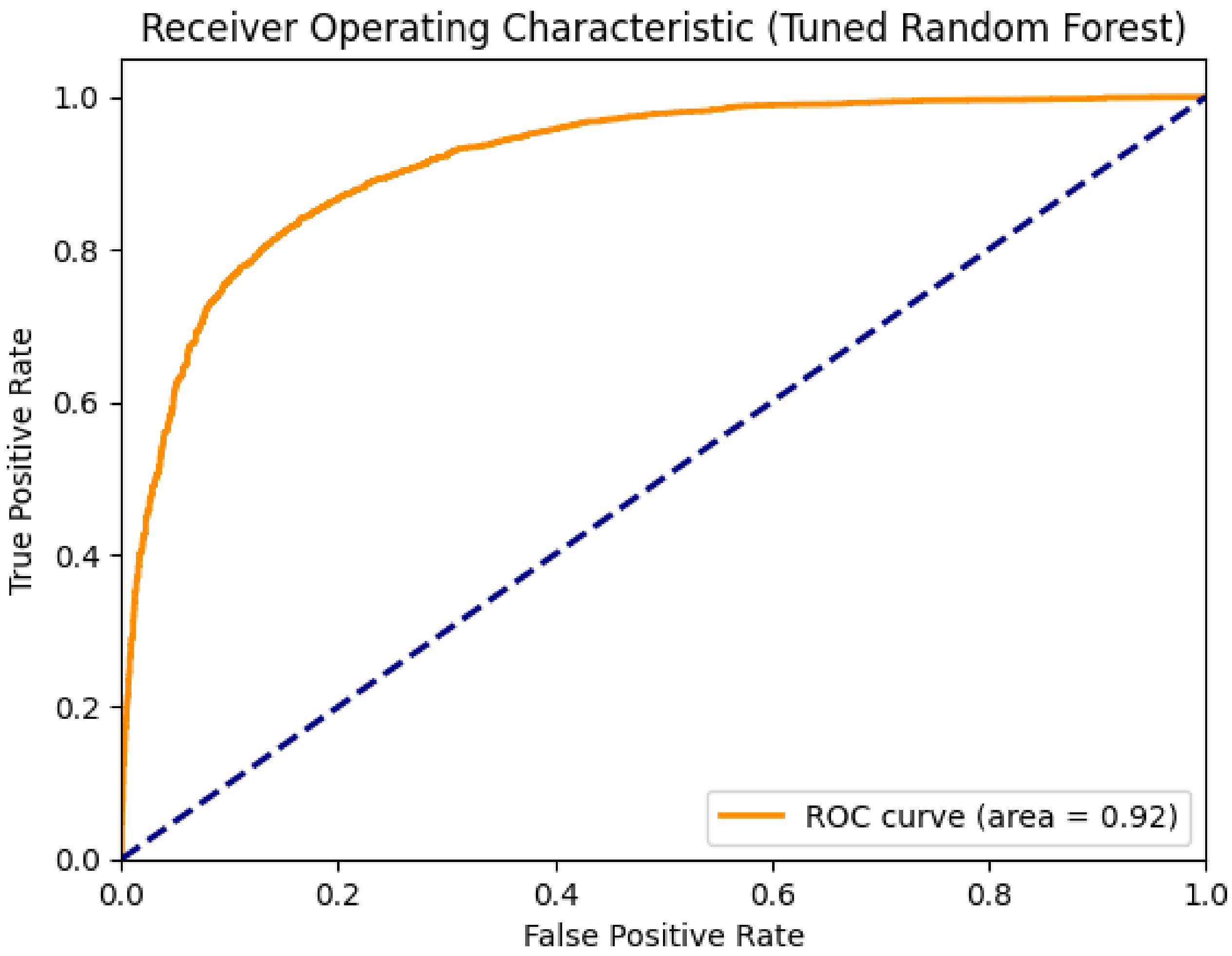
- ✓ **Recall:** To reduce false negatives, ensuring depressed students are not missed.
- ✓ **AUC-ROC Score:** To measure the model's ability to distinguish between classes.
- ✓ **F1-Score:** Balancing precision and recall for overall performance.

Results:

Model Performance Comparison

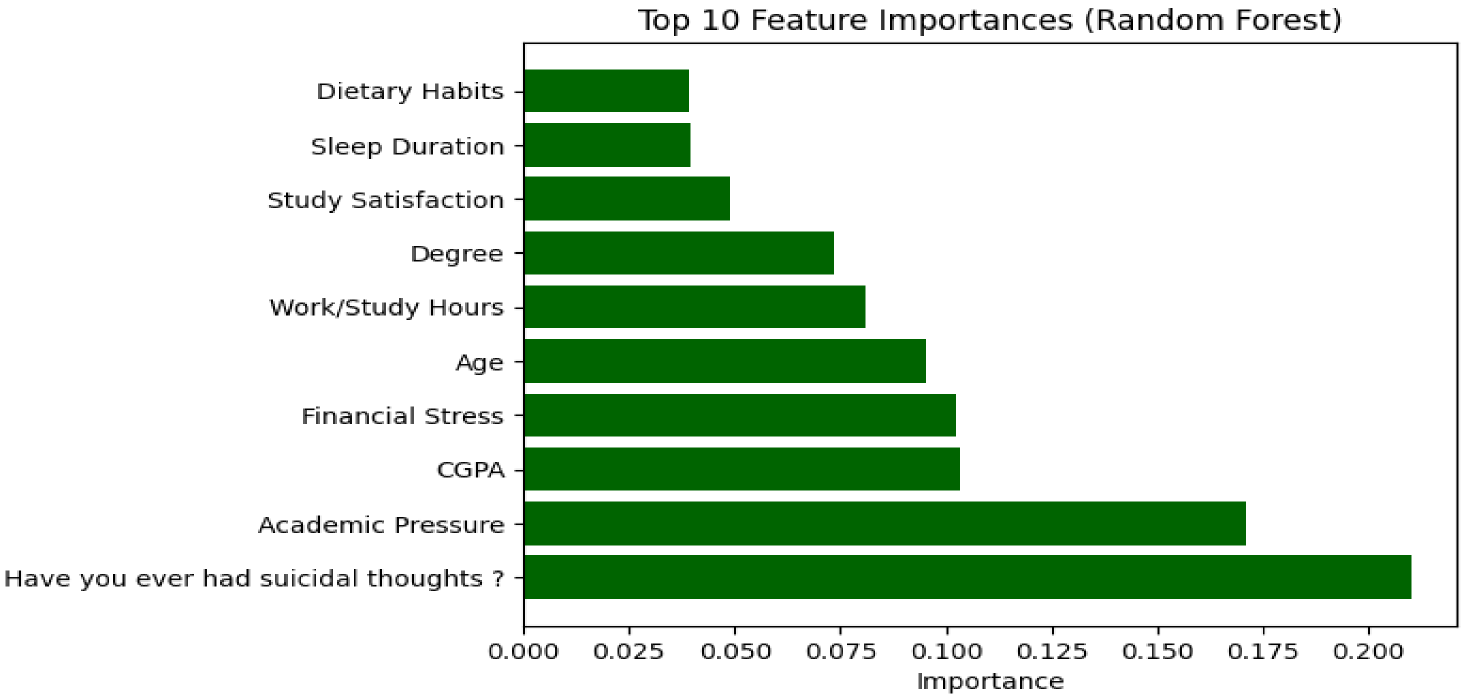
Model	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
Logistic Regression	0.86	0.88	0.87	0.835
Random Forest	0.85	0.88	0.86	0.831
XGBoost	0.85	0.87	0.86	0.825
Tuned Random Forest	0.85	0.89	0.87	0.829

The tuned Random Forest model achieved the highest recall (0.89), making it the best for detecting depression cases, while Logistic Regression had the highest AUC-ROC (0.835), and indicating strong discrimination**.**



Feature Importance (Random Forest):

Rank	Feature	Importance Score
1	Have you ever had suicidal thoughts?	**0.210**
2	Academic Pressure	**0.171**
3	CGPA	**0.103**
4	Financial Stress	**0.102**
5	Age	**0.095**
6	Work/Study Hours	**0.081**
7	Degree	**0.073**
8	Study Satisfaction	**0.048**
9	Sleep Duration	**0.039**
10	Dietary Habits	**0.038**



Insights:

Suicidal thoughts are the strongest predictor, aligning with clinical research. Academic pressure and financial stress are significant, reflecting the high-stress environment of university life. Sleep duration and dietary habits have lesser but measurable impacts, suggesting lifestyle factors also matter.

Practical Advice for Those Who Might Be Depressed:

Based on the analysis, here are signs to watch for:

1) **Persistent Low Mood or Dissatisfaction:** Feeling unhappy with studies, linked to low Study Satisfaction (importance 0.048).

2) **Overwhelming Stress:** High academic pressure (0.171) and financial stress (0.102) are key indicators.

3) **Sleep Problems:** Short sleep (<5 hours) is common among depressed students, with Sleep Duration at 0.039 importance.

4) **Thoughts of Harm or Hopelessness:** Suicidal thoughts (0.210) are a critical sign, requiring immediate attention.

5) **Long Hours Without Balance:** Excessive Work/Study Hours (0.081) can contribute to exhaustion.

6) **Family Context:** Family History of Mental Illness (not in top 10 but noted) increases risk.

Recommendations:

Seek support from friends, family, or counselors if these signs appear. Prioritize sleep (aim for 7-8 hours), manage stress through breaks, and address financial concerns with advisors.

Ethical Considerations:

- **Privacy Concerns:** Sensitive mental health data must be anonymized to protect student privacy.
- **Bias in Data:** Ensure dataset diversity to avoid biased predictions, especially given the small sample size (259 rows).
- **False Positives/Negatives:** Misclassifications could have serious consequences, requiring collaboration with mental health experts.

Future Improvements:

- **Deep Learning:** Explore LSTMs for text-based depression symptoms, as seen in studies like Machine Learning for Depression Detection on Web and Social Media.
- **SMOTE:** Apply Synthetic Minority Over-sampling Technique to handle class imbalance, given 54.8% depressed cases.
- **More Data:** Collect data from multiple universities for better generalization, as suggested by Prediction and diagnosis of depression using machine learning with electronic health records data.
- **Web Application:** Deploy as a Flask/Streamlit API for real-time predictions, enhancing accessibility.

Conclusion:

This analysis demonstrates the potential of machine learning in detecting depression among university students, with the tuned Random Forest model achieving 89% recall. Key features like suicidal thoughts, academic pressure, and financial stress highlight areas for intervention. While promising, ethical considerations and data limitations suggest further research and cautious application in real-world settings.

Key Citations:

- Prediction and diagnosis of depression using machine learning with electronic health records data: a systematic review
- Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions
- A hybrid model for depression detection using deep learning
- Machine learning models to detect anxiety and depression through social media: A scoping review
- Frontiers | Machine learning models predict the emergence of depression in Argentinean college students during periods of COVID-19 quarantine
- An in-depth analysis of machine learning approaches to predict depression
- Depression detection using emotional artificial intelligence and machine learning: A closer review
- Machine Learning for Depression Detection on Web and Social Media