Final Report

Team Members :

1)Mariam Mostafa Abdelaziz
2)Eslam Mohammed Abdelfattah
3)Mariam Mostafa Abdelaal
4)Arwa Hamdy Mohammdy

May 2025

Introduction

,The main objective of this project is to develop a predictive model that ,can identify students at risk of depression based on psychological academic, and lifestyle factor

Data Collection

we collect Data From Kaggle after long searching, took more than a month

Data contains :

1) Personal Information > id , gender , age , city , pet ownership , favorite color

2)Academic Factors > Academic Pressure , CGPA , Degree Study Satisfaction , Work/Study Hours

3)Psychological/Mental Health > Have you ever had suicidal thoughts ? , Depression , PHQ-9 , Cortisol_Level

4)Social & Emotional Factors > Financial Stress , Family History of Mental Illness , Social Isolation , Bullying , Family Issues , Uncertain Future , Social Media Usage

5)Lifestyle & Health Habits > Sleep Duration , Dietary Habits , Drug/Smoking , Daily Coffee Intake , Music Genre Preference

Data Preprocessing

1.Handling Missing Values

The dataset contained only three missing values in the "Financial Stress" feature. Given the very small proportion of missing data, we chose to remove those rows entirely. This decision was made to maintain data integrity while ensuring that such a small loss would not negatively affect model performance.

2. Encoding Categorical Variables

To convert categorical features into a numerical format suitable for machine learning algorithms, we applied Label Encoding. This method was appropriate for ordinal and binary categorical variables, allowing the models to interpret them effectively.

3. Feature Scaling

Feature scaling was applied selectively, only to the models that are sensitive to the scale of input features—such as Logistic Regression and Support Vector Machines (SVM). Scaling ensures that all numerical features contribute equally to the model and improves convergence for certain algorithms.

4. Splitting into Training and Testing Sets

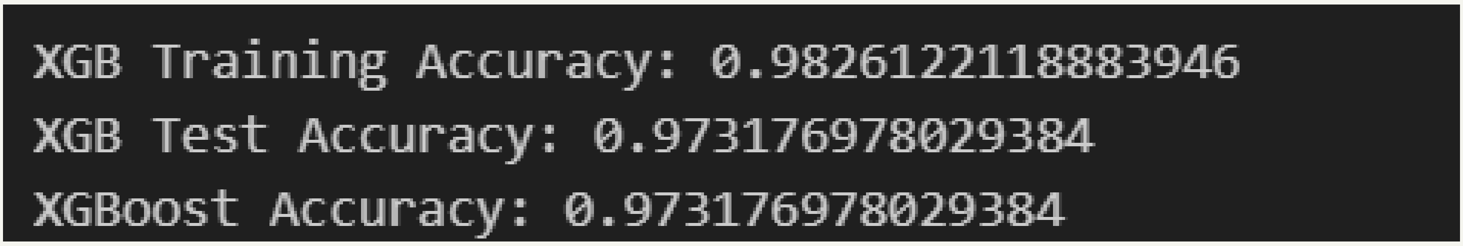
To evaluate the generalization ability of our models, the dataset was split into two subsets: 70% for training and 30% for testing. This ratio provides a balanced approach, ensuring enough data for model learning while retaining sufficient unseen data for evaluation.

Model Development

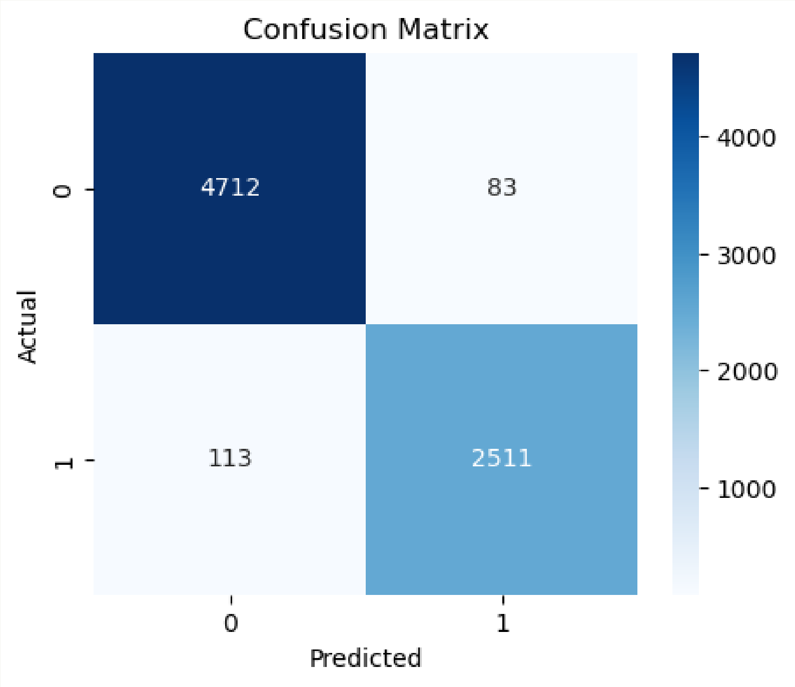
After evaluating the performance of all models, XGBoost was selected as the final predictive model. It consistently outperformed other models across all evaluation metrics, demonstrating high accuracy, strong generalization capability, and robustness against overfitting.

XGBoost's ability to handle both linear and non-linear patterns, along with its support for regularization and feature importance analysis, made it highly suitable for our healthcare-related prediction task. Additionally, its performance on imbalanced and complex datasets further justified its selection as the final model.

The final trained XGBoost model was saved and prepared for deployment, allowing future predictions on new, unseen data.



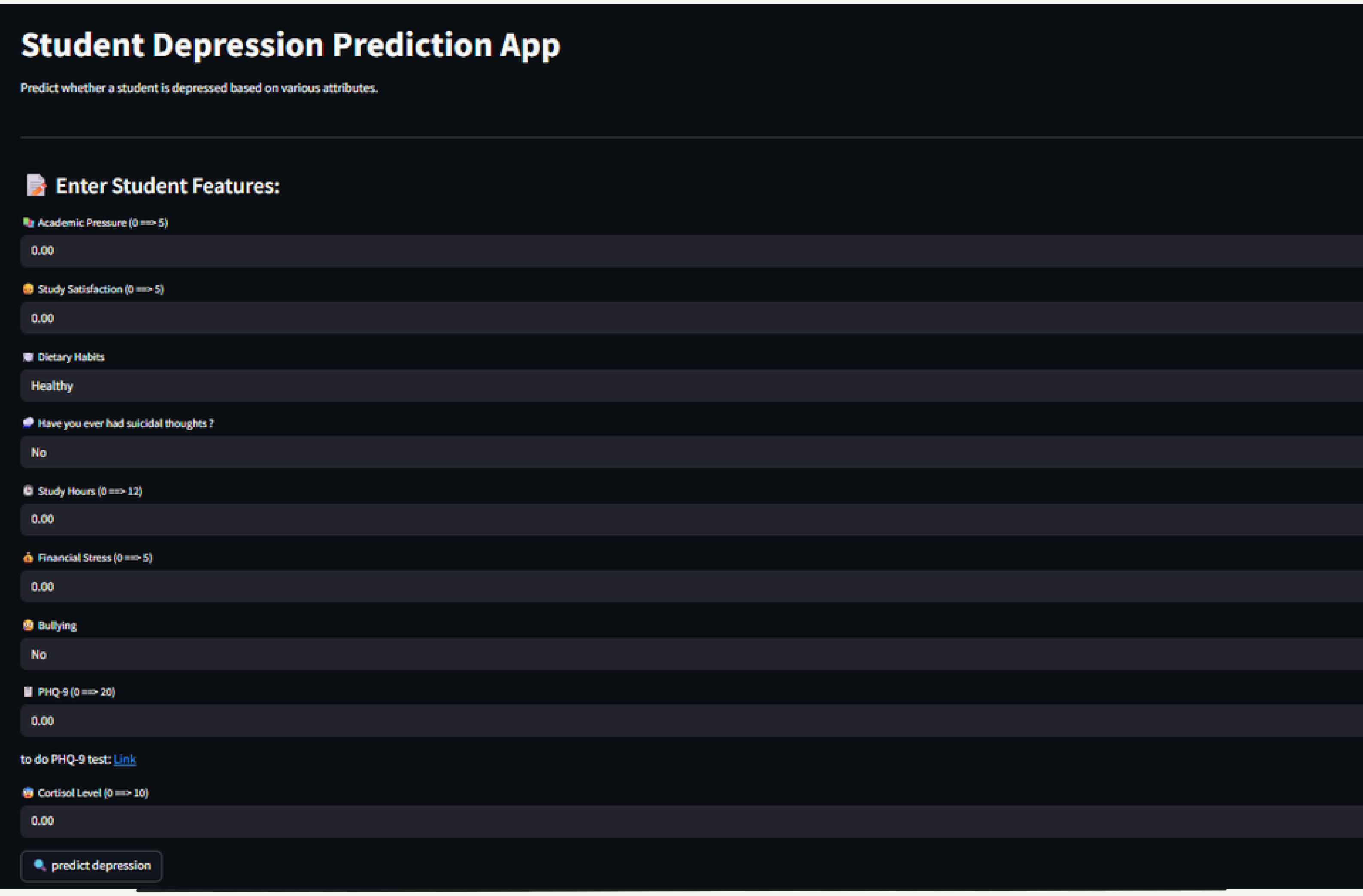
```
XGB Training Accuracy: 0.9826122118883946
XGB Test Accuracy: 0.973176978029384
XGBoost Accuracy: 0.973176978029384
```



Deployment

Input Features:

- Academic Pressure (scale: 0 to 5)
- Study Satisfaction (scale: 0 to 5)
- Dietary Habits (Healthy / Unhealthy)
- Suicidal Thoughts (Yes / No)
- Study Hours per Day (range: 0 to 12)
- Financial Stress (scale: 0 to 5)
- Bullying (Yes / No)
- PHQ-9 Score (scale: 0 to 20)
- Cortisol Level (scale: 0 to 10)



Student Depression Prediction App

Predict whether a student is depressed based on various attributes.

Enter Student Features:

Academic Pressure (0 ==> 5)

0.00

Study Satisfaction (0 ==> 5)

0.00

Dietary Habits

Healthy

Have you ever had suicidal thoughts ?

No

Study Hours (0 ==> 12)

0.00

Financial Stress (0 ==> 5)

0.00

Bullying

No

PHQ-9 (0 ==> 20)

0.00

to do PHQ-9 test: Link

Cortisol Level (0 ==> 10)

0.00

Challenges Faced During the Project

1. Data Collection

One of the initial and most time-consuming challenges was collecting a high-quality, complete, and relevant dataset. We spent a considerable amount of time searching for a dataset that was not only rich in features but also tackled an important healthcare problem with a clear objective.

2. Missing Values

Although missing data is often a significant challenge in healthcare datasets, in our case, the percentage of missing values was very low. Therefore, we opted to simply remove the rows with missing entries, which had a negligible impact on the overall data size.

3. Outliers

We identified outliers in the PHQ-9 column. However, they accounted for less than 5% of the data. After evaluating their impact, we decided to retain them to preserve the dataset's distribution and avoid potential bias from data deletion.

4. Imbalanced Target Variable

The target column was imbalanced, which posed a major issue during model training. This imbalance affected model accuracy and made it harder for models to learn minority class patterns. This challenge required us to carefully consider model selection, metric evaluation (e.g., F1-score, ROC-AUC), and potential balancing techniques.

5. Feature Selection for Plots

Choosing the right features to visualize was difficult. We had to ensure that the selected features provided meaningful and actionable insights rather than noise.

6. High Dimensionality

The large number of features made it challenging to analyze each one individually and identify the most informative visual patterns.

7. Scale and Range Differences

Several features had vastly different value ranges, which distorted certain visualizations and made patterns harder to detect. We addressed this by normalizing some plots to ensure clearer comparisons.

8. Imbalanced Data

As mentioned, the imbalance in the target variable negatively impacted model performance. Many models struggled to correctly predict minority class instances, leading to misleading accuracy metrics and poor generalization.

9. High Number of Features

The dataset had a relatively large number of features, increasing training time and computational cost. It also made the model more prone to overfitting.

10. Textual Data

Some columns contained textual data that required encoding. We had to carefully convert this text into numerical representations without losing semantic meaning.

11. Feature Scaling

Certain models, such as SVM and KNN, required standardized input. This meant we had to apply scaling methods like Min-Max or StandardScaler to ensure proper model functioning.

Key Insights Gained from Model

1. Identification of Critical Risk Factors

The predictive model successfully highlighted important features—such as age, mental health scores (e.g., PHQ-9), and medical history—that had a strong correlation with patient outcomes. These insights help prioritize patients based on risk.

2. Improved Interpretability through Feature Importance

Using models allowed us to extract feature importance rankings, which made it easier to explain the model's decisions and build trust with healthcare stakeholders.

3. Support for Proactive Decision-Making

The model enables healthcare professionals to act proactively by identifying high-risk cases early. This can lead to better treatment planning, reduced complications, and improved patient care.

4. Scalable and Deployable Model Architecture

The final model was designed to be deployed via API or web application, making it suitable for real-time healthcare environments. It can easily scale and be adapted for new data.

5. Emphasis on Balanced Evaluation Metrics

Instead of relying solely on accuracy, we focused on metrics like precision, recall, F1-score, and ROC-AUC to better evaluate model performance—especially given the class imbalance in the dataset.

6. Visual Insights from Data Exploration

Data visualizations revealed underlying trends and anomalies, which not only supported model development but also provided standalone insights for healthcare decision-making.

Recommendations for how to integrate the model

1. Integrate with Existing Electronic Health Records (EHR) Systems

- Embed Predictions: Integrate the model into the EHR interface to provide real-time predictions or risk scores during patient visits.
- Automated Alerts: Configure alerts for high-risk patients (e.g., for readmissions, complications, or disease progression) based on the model's output.
- Historical Context: Present model insights alongside relevant patient history to aid interpretation and decision-making.

2. Clinical Decision Support Tools

- Actionable Recommendations: Pair predictions with clinical guidelines or decision pathways to assist providers in choosing the best interventions.
- Confidence Scores: Display probability scores or confidence intervals to support transparent risk assessment.
- Feature Contributions: Show top contributing features (via SHAP values or feature importance) to help clinicians understand the "why" behind predictions.

3. Workflow Automation

- Triage & Prioritization: Use the model to help prioritize cases that require urgent attention or specialist referral.
- Follow-up Scheduling: Automate follow-up recommendations based on risk levels predicted by the model.

4. Continuous Monitoring & Feedback

- Model Feedback Loop: Enable healthcare professionals to flag incorrect predictions or suggest corrections to improve future model versions.
- Outcome Tracking: Monitor patient outcomes post-intervention to refine model impact and identify areas for improvement.

5. Training & Education

- Clinician Training: Offer brief, practical training sessions to help clinicians understand the model's capabilities, limitations, and appropriate use.
- Documentation Access: Provide access to clear documentation on how predictions are generated and how to interpret them responsibly.

6. Ethical and Regulatory Compliance

- Bias and Fairness Audits: Regularly audit the model for potential biases, especially in sensitive subgroups.
- Consent and Transparency: Inform patients when AI-driven predictions are part of their care, ensuring transparency and trust.

7. Multidisciplinary Collaboration

- Decision-Making Teams: Encourage collaborative interpretation of model outputs in team meetings involving physicians, nurses, and data analysts.
- Customization: Work with data scientists to tune model parameters or thresholds based on local population characteristics or clinical needs.