# Predicting Student Depression Using Machine Learning :

## A Comparative Analysis

Presented by Mohamed Yassine El Wardani & Abderrahmen Youssef

24 April 2025

# Introduction

## Why this subject?

- Student depression is a growing problem worldwide.
- It affects academic performance, social life, and overall well-being.
- Early detection is crucial for timely support and intervention.

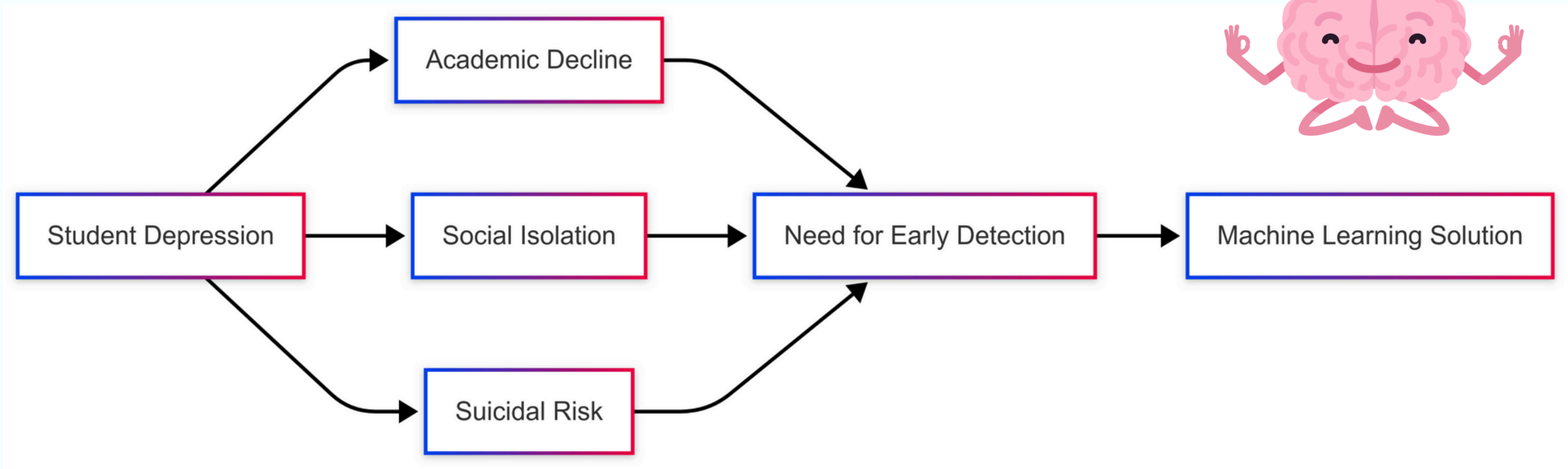
## Scientific Motivation

- Traditional detection methods are often subjective and slow.
- Data science and machine learning can help identify at-risk students more objectively and efficiently

# Introduction

## Project Objective

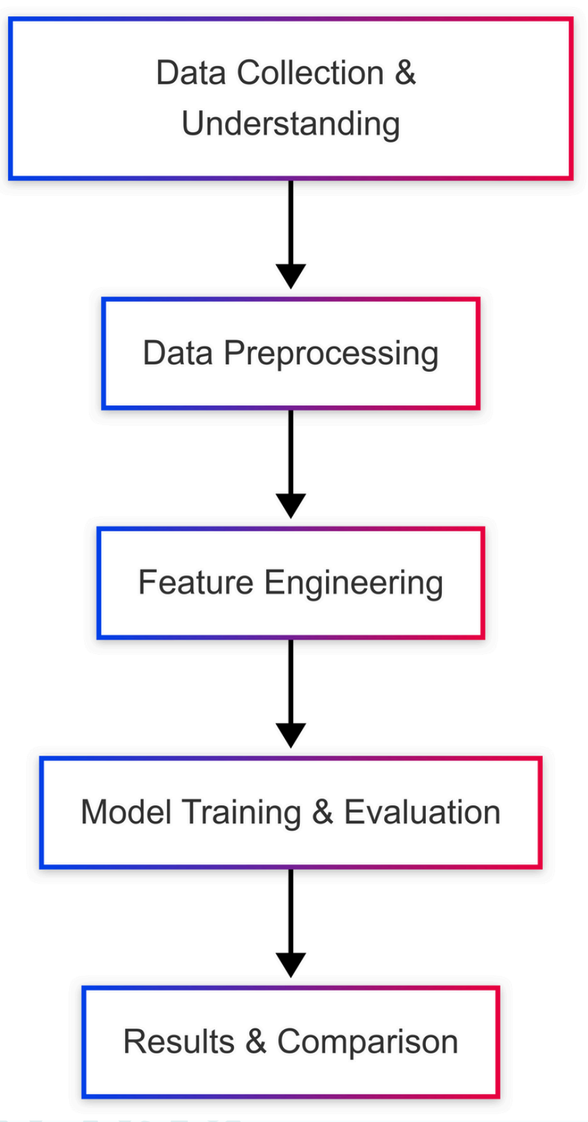
- **Build and evaluate predictive models to detect depression risk among students using real survey data.**



# Research Workflow Overview

- **Our scientific workflow followed these main steps:**

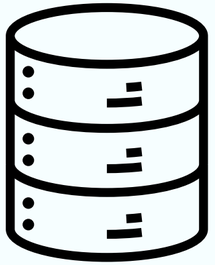
1. **Data Collection & Understanding**
2. **Data Preprocessing**
3. **Feature Engineering**
4. **Model Training & Evaluation**
5. **Results & Comparison**



# Data Collection & Understanding

## Dataset Overview

- **Source: Student depression survey dataset (27,901 records) from kaggle**
- **Each row = 1 student**
- **Features include:**
    1. **Age, Gender, City, Profession**
    2. **Academic Pressure, Work Pressure, CGPA (Academic performance)**
    3. **Study Satisfaction, Job Satisfaction**
    4. **Sleep Duration, Dietary Habits, Degree**
    5. **Suicidal Thoughts, Work/Study Hours, Financial Stress**
    6. **Family History of Mental Illness**
    7. **Target: Depression**



# Data Collection & Understanding

| | id | Gender | Age | City | Profession | Academic Pressure | Work Pressure | CGPA | Study Satisfaction | Job Satisfaction | Sleep Duration | Dietary Habits | Degree | Have you ever had suicidal thoughts ? | Work/ Study Hours | Financial Stress | Family History of Mental Illness | Depression |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2 | Male | 33.0 | Visakhapatnam | Student | 5.0 | 0.0 | 8.97 | 2.0 | 0.0 | '5-6 hours' | Healthy | B.Pharm | Yes | 3.0 | 1.0 | No | 1 |
| **1** | 8 | Female | 24.0 | Bangalore | Student | 2.0 | 0.0 | 5.90 | 5.0 | 0.0 | '5-6 hours' | Moderate | BSc | No | 3.0 | 2.0 | Yes | 0 |
| **2** | 26 | Male | 31.0 | Srinagar | Student | 3.0 | 0.0 | 7.03 | 5.0 | 0.0 | 'Less than 5 hours' | Healthy | BA | No | 9.0 | 1.0 | Yes | 0 |
| **3** | 30 | Female | 28.0 | Varanasi | Student | 3.0 | 0.0 | 5.59 | 2.0 | 0.0 | '7-8 hours' | Moderate | BCA | Yes | 4.0 | 5.0 | Yes | 1 |
| **4** | 32 | Female | 25.0 | Jaipur | Student | 4.0 | 0.0 | 8.13 | 3.0 | 0.0 | '5-6 hours' | Moderate | M.Tech | Yes | 1.0 | 1.0 | No | 0 |

# Data Collection & Understanding

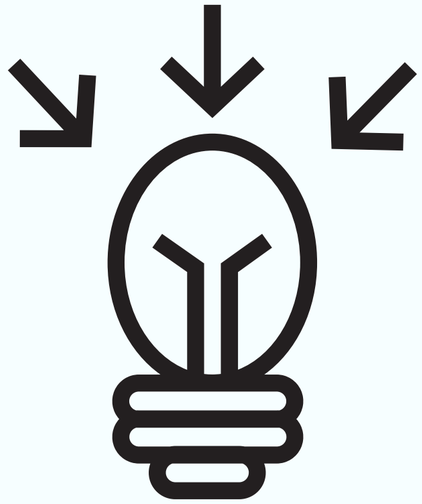
## Difficulties

- **Missing values and inconsistent formats (profession)**
- **Categorical variables (text, yes/no)**
- **Some irrelevant columns (ID, city)**

# Data Preprocessing

## Key Steps

- **Column Renaming: Standardized column names (lowercase, underscores, removed special characters).**
- **Cleaning: Removed extra quotes and whitespace from text fields.**
- **Categorical Mapping:**
  - **Converted sleep duration and yes/no answers to numbers.**
  - **Label encoded degrees, one-hot encoded dietary habits.**
- **Missing Values:**
  - **Replaced “?” with NaN.**
  - **Imputed missing numerical values with the median.**
- **Dropped Irrelevant Columns:**
  - **Removed ID, city, profession, and gender.**



# Data Preprocessing

```
RangeIndex: 27901 entries, 0 to 27900
Data columns (total 19 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   age                    27901 non-null  float64
 1   academic_pressure      27901 non-null  float64
 2   work_pressure          27901 non-null  float64
 3   cgpa                   27901 non-null  float64
 4   study_satisfaction     27901 non-null  float64
 5   job_satisfaction       27901 non-null  float64
 6   sleep_duration         27901 non-null  int64  
 7   degree                 27901 non-null  int64  
 8   suicidal_thoughts      27901 non-null  int64  
 9   work/study_hours       27901 non-null  float64
 10  financial_stress       27901 non-null  float64
 11  illness_history        27901 non-null  int64  
 12  depression             27901 non-null  int64  
 13  diet_Moderate          27901 non-null  bool   
 14  diet_Others            27901 non-null  bool   
 15  diet_Unhealthy         27901 non-null  bool   
 16  academic_stress_combo  27901 non-null  float64
 17  burnout_index          27901 non-null  float64
 18  wellness_score         27901 non-null  float64
dtypes: bool(3), float64(11), int64(5)
memory usage: 3.5 MB
```

# Feature Engineering

## What we did

- Created new features to better capture student stress and wellness:
  - Academic Stress Combo: Academic Pressure × Financial Stress
  - Burnout Index: Academic Pressure × Work/Study Hours
  - Wellness Score: Study Satisfaction + Sleep Duration

## Scientific Rationale

- Combining related variables can reveal deeper patterns and improve model performance.
- These engineered features help quantify complex aspects of student life.

| | academic_stress_combo | burnout_index | wellness_score |
|---|---|---|---|
| 0 | 7.0 | 21.0 | 13 |
| 1 | 14.0 | 32.0 | 13 |
| 2 | 7.0 | 14.0 | 13 |
| 3 | 5.4 | 21.0 | 13 |
| 4 | 14.4 | 40.5 | 12 |

# Data Exploration & Correlation Analysis

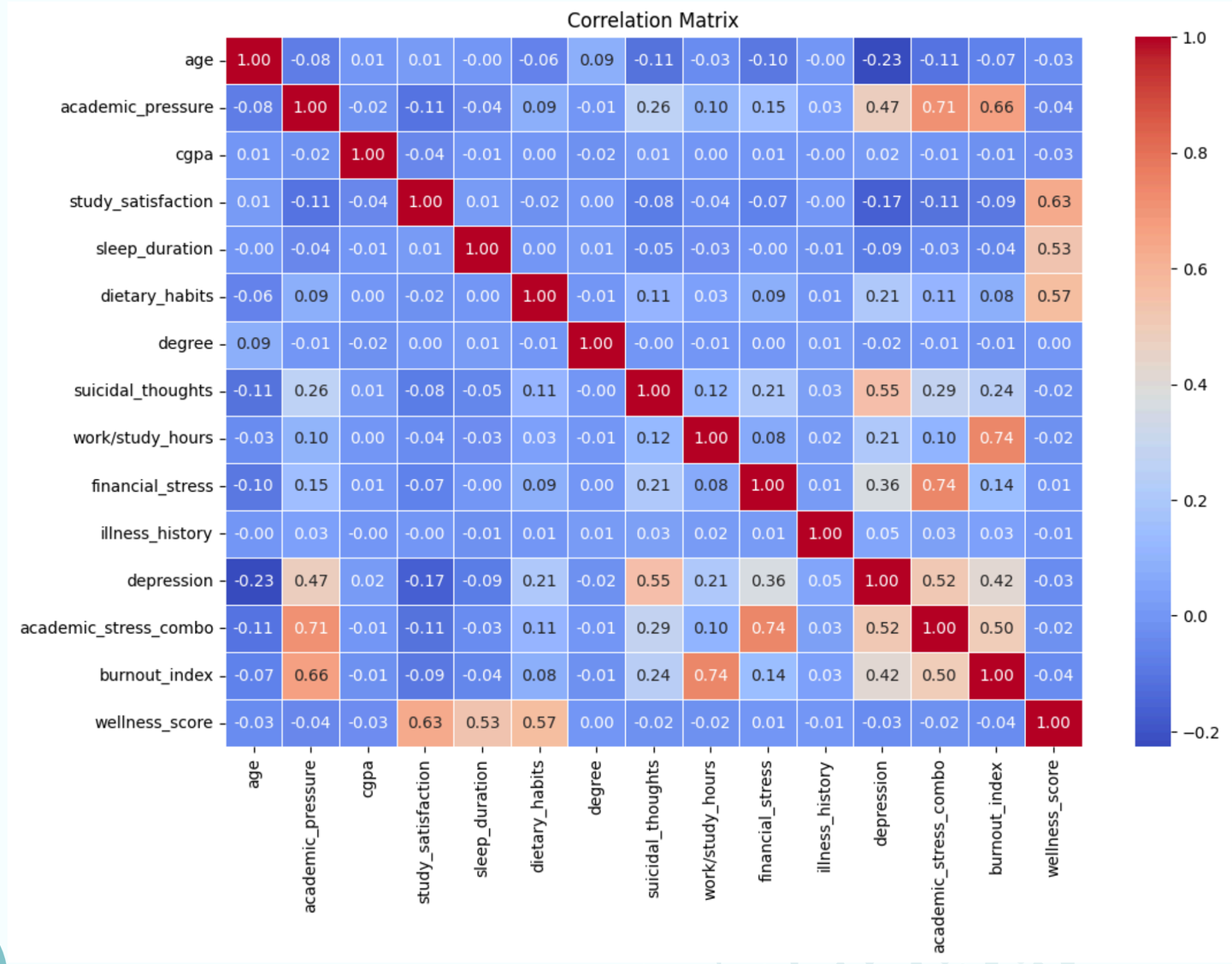
## What we did

- Explored relationships between variables using a correlation matrix.
- Identified which features are most strongly linked to depression.
- Used heatmaps for visual understanding of data structure.

## Scientific Rationale

- Correlation analysis helps in selecting the most relevant predictors for the model.
- For example, academic pressure, sleep duration, and financial stress showed strong correlations with depression.

# Data Exploration & Correlation Analysis



# Model Training

## Models Used

- Random Forest
- XGBoost
- Logistic Regression

## Scientific Rationale

- Data split into training and test sets (80/20, stratified for balance).
- Numerical features scaled for better model performance.
- Hyperparameter tuning with GridSearchCV and cross-validation for each model.
- Each model trained to predict depression risk.

# Model Evaluation

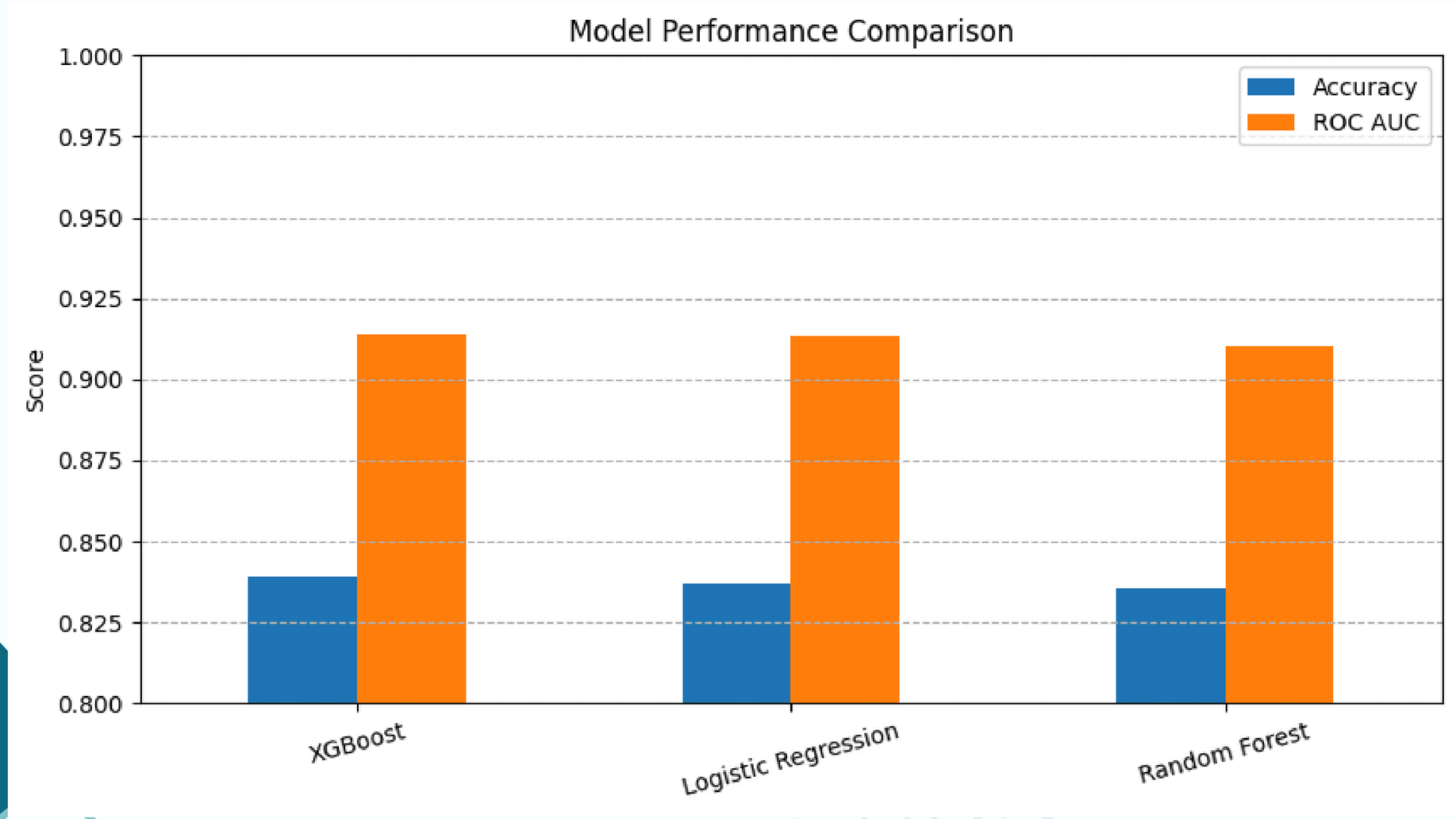
## How we evaluated the models

- **Metrics Used:**
  - **Accuracy**
  - **ROC AUC (Area Under the Curve)**
  - **Precision, Recall, F1-Score (from classification report)**
  - **Cross-validation scores for robustness**

## Scientific Rationale

- **Evaluated each model on the test set and with cross-validation to ensure results are reliable and not due to chance.**

# Model Evaluation



# Results Comparison

## Comparison of Model Performance

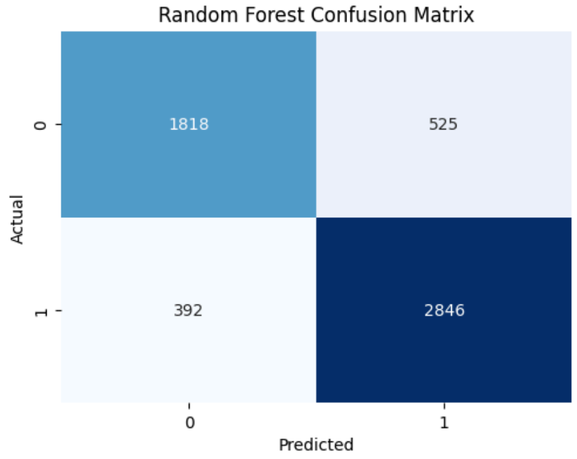
- Compared Random Forest, XGBoost, and Logistic Regression.
- Evaluated using:
  - Test Accuracy
  - Test AUC (Area Under Curve)
  - Mean Cross-Validation Accuracy
  - Mean Cross-Validation AUC

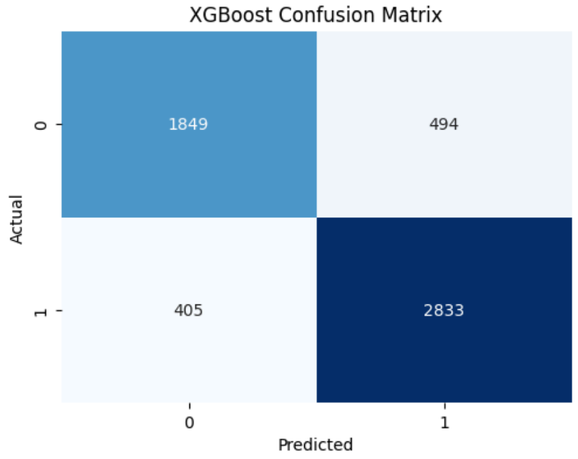
## Scientific Rationale

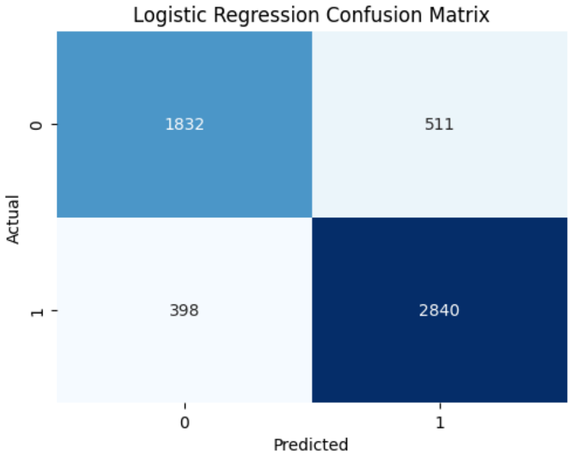
- XGBoost achieved the highest accuracy and AUC.
- Each model has strengths: Logistic Regression had higher precision, Random Forest better recall.

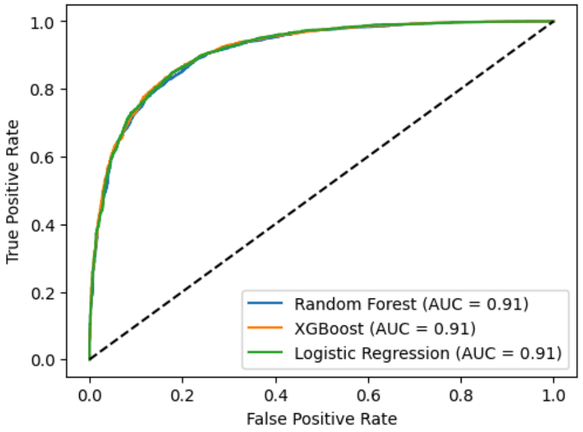
# Results Comparison

## Model Performance Benchmarks




Metrics Comparison

| Model | Accuracy | Precision | Recall | F1-Score | ROC AUC |
|---|---|---|---|---|---|
| Random Forest | 0.8357 | 0.8443 | 0.8789 | 0.8612 | 0.9103 |
| XGBoost | 0.8389 | 0.8515 | 0.8749 | 0.8631 | 0.914 |
| Logistic Regression | 0.8371 | 0.8475 | 0.8771 | 0.862 | 0.9136 |

# Challenges & Scientific Difficulties

## Data Quality Issues

- Missing values and inconsistent formats in the dataset
- Categorical variables with different representations (e.g., 'Yes', 'No', extra quotes)

## Modeling Challenges

- Class imbalance: fewer students with depression than without
- Avoiding overfitting while tuning models
- Choosing the right evaluation metrics for a fair comparison

# Scientific Impact & Future Work

## Scientific Contributions & Improvements

- **Improved Model Performance:**
  - **Achieved higher accuracy and AUC compared to basic models by:**
    - **Careful data cleaning and preprocessing**
    - **Feature engineering (stress combo, burnout index, wellness score)**
    - **Hyperparameter tuning and cross-validation**
- **Robust Evaluation:**
  - **Used mean cross-validation accuracy and AUC to ensure results are reliable and generalizable.**
- **Best Results:**
  - **XGBoost achieved the highest test accuracy and AUC.**
  - **All models showed improved mean CV accuracy and AUC after tuning.**

# Scientific Impact & Future Work

## Impact

- Demonstrates the value of machine learning for early depression detection in students.
- Provides a scientific workflow that can be reused or extended for similar mental health studies.

## Future Work

- Collect more diverse and larger datasets for better generalization.
- Explore advanced models (e.g., deep learning).
- Collaborate with mental health professionals for real-world deployment and validation.

# Conclusion

- **Machine learning can effectively predict depression risk among students using survey data.**
- **Careful data cleaning, feature engineering, and model tuning led to significant improvements in accuracy and reliability.**
- **XGBoost achieved the best results, but all models benefited from scientific workflow and robust evaluation.**

# References


1. World Health Organization, “Depression,” WHO Fact Sheets, 2021.
2. A. Sau and I. Bhakta, “Predicting anxiety and depression in elderly patients using machine learning technology,” Healthcare Technology Letters, vol. 6, no. 1, pp. 29–33, 2019.
3. T. Richter et al., “Using machine learning-based analysis for behavioral differentiation between anxiety and depression,” Scientific Reports, vol. 10, no. 1, pp. 1-12, 2020.
4. R. P. Auerbach et al., “WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders,” Journal of Abnormal Psychology, vol. 127, no. 7, pp. 623–638, 2018.
5. A. Friedrich and A. A. Schlarb, “Let’s talk about sleep: a systematic review of psychological interventions to improve sleep in college students,” Journal of Sleep Research, vol. 27, no. 1, pp. 4–22, 2018.
6. Y. Li et al., “Associations of dietary patterns with the risk of depression: A systematic review and meta-analysis,” Front. Nutr., vol. 7, p. 581226, 2020.
7. L. Wang et al., “Ensemble learning for the prediction of student depression based on electronic health records,” Journal of Medical Imaging and Health Informatics, vol. 10, no. 10, pp. 2391–2399, 2020.
8. I. Fatima et al., “Analysis of user-generated content from online social communities to predict and detect depression,” International Journal of Human–Computer Interaction, vol. 35, no. 1, pp. 57–69, 2019.
9. J. Chen et al., “Early detection of depression: Using machine learning methods to predict depression based on health check data,” Frontiers in Psychiatry, vol. 12, p. 661610, 2021.

# Thank You