Machine Learning Report: Predicting University Students' Depression Levels

May 10, 2025

Abstract

This report presents a detailed application of machine learning methods for predicting and classifying depression levels among university students. We follow the CRISP-DM methodology, clearly outlining the processes of data understanding, data preparation, modeling, and evaluation. We also discuss practical implications and recommendations based on our analysis.

1 Business Understanding

1.1 Problem Definition

The identified problem focuses on the **mental health of university students**, specifically on the early detection and classification of depression levels among students. Mental health significantly impacts students' academic performance and overall well-being, making accurate understanding and proactive management essential.

1.2 Objectives and Desired Outcomes

The primary objectives of this project are:

- **Predicting students' depression levels** using relevant features, including:
  - Suicidal thoughts,
  - Measured depression values,
  - Other psychological variables (anxiety, stress), and academic factors.
- **Segmenting student profiles** into homogeneous groups to better understand their specific needs and propose tailored interventions.
- **Recommending career paths** aligned with each student's psychological and academic profiles to enhance their motivation and future well-being.

1.3 Business Goals

Specific, measurable business goals achievable through the Machine Learning application are:

- **Quantifiable improvement in student well-being:**
  - Achieving at least a 20% reduction in severe depression levels through targeted interventions.
  - Increasing the utilization of university mental health support resources by 25% via improved identification of at-risk students.
- **Personalized academic and career pathways:**
  - Increasing student satisfaction by 30% regarding career or academic paths recommended based on their psychological and academic profiles.
  - Reducing academic dropout rates through early identification and targeted support by 15%.

1.4 Requirements and Constraints

Key project constraints and limitations include:

- **Data privacy:** Ensuring strict confidentiality and ethical handling of sensitive mental health and academic data.
- **Technical constraints:** Limited computational resources that may influence model complexity and algorithm choices.
- **Resource availability:** Restrictions on the frequency and extent of interventions and recommendations due to institutional resource limitations.

2 Data Understanding

2.1 Data Collection

The dataset used in this project was sourced from a publicly available repository on Kaggle. This dataset provides comprehensive mental health data collected via surveys from Bangladeshi university students.

- **Data Volume and Structure:**
  - **Rows (Records):** 1,978 student responses.
  - **Columns (Features):** 39 distinct features capturing demographic, academic, and mental health-related information.
- **Data Types:**
  - **Integer:** 29 columns (e.g., anxiety scores, depression scores, stress values).
  - **Categorical (String):** 9 columns (e.g., Age range, Gender, Department, University, Depression Label).
  - **Boolean:** 1 column (Scholarship/Waiver Status).

2.2 Exploratory Data Analysis (EDA)

A comprehensive EDA was conducted to better understand the characteristics, distributions, and relationships within the data. The following analyses were performed:

- **Visualizations:**
  - **Histograms** to analyze the distribution of numeric variables such as Anxiety, Depression, and Stress scores.
  - **Boxplots** to detect outliers and visualize distribution patterns for scores and GPA values.
  - **Correlation Heatmaps** to reveal significant correlations between psychological factors (anxiety, stress, depression) and academic performance (GPA).
- **Statistical Summaries:**
  - Calculated descriptive statistics, including **mean, median, standard deviation**, and **variance** for numeric variables.
  - Frequency distributions for categorical features (e.g., gender, academic year, university).
  - Analysis of missing values conducted to evaluate data completeness.

2.3 Data Quality Assessment

An in-depth data quality assessment was performed, resulting in the following findings:

- **Missing Values:**
  - Minimal missing values were found, primarily within categorical demographic features, requiring minor interventions (imputation or removal).
- **Inconsistencies:**
  - Some inconsistencies in data entries (e.g., varied categorical labeling) required standardization.
- **Outliers:**
  - Boxplot visualizations identified a few outliers, particularly in stress and depression scores. Legitimate outliers were retained to preserve authentic data variability.
- **Potential Data Biases:**
  - The dataset exhibited demographic bias, predominantly representing students from specific universities and departments (mostly computer engineering students). This may limit generalizability to broader student populations.

3 Data Preparation

3.1 Data Cleaning

Data cleaning was a critical initial step to ensure high-quality inputs for modeling. The following processes were performed:

- **Handling Missing Values:**
  - Missing values, primarily within categorical features, were addressed through appropriate imputation techniques:
    - **Categorical Features:** Missing entries were filled using the mode (most frequent value).
    - **Numerical Features:** Numerical variables with minimal missing values were handled using median imputation to minimize distortion caused by outliers.
- **Removing Duplicates and Correcting Inconsistencies:**
  - Duplicate rows were identified and removed to maintain data integrity.
  - Standardization was applied to categorical labels (e.g., Gender, Academic Year, Depression Labels) to correct inconsistencies.

3.2 Feature Engineering

Feature engineering involved creating meaningful transformations and enhancements to optimize model performance:

- **Creating New Features:**
  - Aggregated scores for anxiety, stress, and depression were computed by summing responses to respective survey questions, resulting in composite scores:
    - **Anxiety Value** (0–21),
    - **Stress Value** (0–40),
    - **Depression Value** (0–27).
  - Derived categorical labels (Minimal, Mild, Moderate, Severe) were created based on numeric thresholds of these composite scores.
- **Encoding and Normalization:**

  - **Categorical Encoding:** Categorical variables (e.g., Gender, Department, Academic Year, Depression Labels) were encoded using Label Encoding and One-Hot Encoding techniques, based on the nature of data and algorithmic requirements.
  - **Normalization and Scaling:** Numerical features (Anxiety Value, Stress Value, Depression Value, GPA scores) were scaled using standard normalization (*StandardScaler*), transforming data to have a mean of 0 and standard deviation of 1, improving model convergence.

3.3 Data Splitting

The prepared dataset was split into clearly defined subsets to effectively train, validate, and test machine learning models:

- **Training Set (80%):** Used to train and optimize model parameters.
- **Validation Set (10%):** Used during model development to select hyperparameters and prevent overfitting.
- **Test Set (10%):** Reserved exclusively for evaluating final model performance on unseen data, ensuring an unbiased assessment of generalization capability.

Splitting proportions were justified based on dataset size and standard machine learning practices to ensure robust and generalizable results.

4 Modeling

4.1 Algorithm Selection

For this project, supervised learning algorithms tailored specifically for classification tasks were chosen, considering the categorical nature of the target variable (depression level). The selected algorithms were:

- **Random Forest Classifier:** Selected for robustness, interpretability, and resistance to overfitting. It effectively handles mixed data types and feature interactions.
- **XGBoost (Extreme Gradient Boosting):** Chosen for its superior performance, efficiency, and capability to handle nonlinear relationships and complex feature interactions.

4.2 Brief Theoretical Background

- **Random Forest:** Random Forests are ensemble methods consisting of multiple decision trees trained on bootstrapped data samples. Predictions are aggregated (majority voting for classification tasks), significantly reducing variance and enhancing prediction accuracy.
- **XGBoost:** XGBoost is a gradient boosting framework that iteratively builds new decision trees to correct residual errors from previous models. It optimizes an objective function consisting of a loss term (accuracy) and regularization term (prevention of overfitting), resulting in highly effective yet generalizable models.

4.3 Model Building

- **Model Parameters and Hyperparameters:**
  - **Random Forest:**
    - Number of trees (`n_estimators`): tested values (50, 100, 150).
    - Maximum depth (`max_depth`): tested values (5, 10, 15, None).
    - Criterion (`criterion`): Gini impurity, Entropy.
  - **XGBoost:**
    - Number of trees (`n_estimators`): (50, 100, 150).
    - Learning rate (`learning_rate`): (0.01, 0.05, 0.1, 0.2).
    - Maximum depth (`max_depth`): (3, 5, 7, 9).

        - Subsample ratio: fraction of samples used per tree (0.6, 0.8, 1.0).

- **Training Process:**
    - Models were trained using the training dataset (80% of data) with stratified 5-fold cross-validation, ensuring balanced class representation and reliable performance estimation.

4.4 Hyperparameter Tuning

- **Techniques Applied:**
    - **Grid Search Cross-Validation:** systematically applied for Random Forest due to its moderate hyperparameter search space.
    - **Randomized Search Cross-Validation:** utilized for XGBoost to efficiently explore its larger parameter space while minimizing computational effort.
- **Parameter Configurations Tested and Selected:**
    - **Random Forest:** Optimal parameters chosen:
        - `n_estimators = 100`
        - `max_depth = 10`
        - `criterion = "gini"`
    - **XGBoost:** Optimal parameters chosen:
        - `n_estimators = 100`
        - `learning_rate = 0.1`
        - `max_depth = 5`
        - `subsample = 0.8`

These optimized parameters provided an effective balance between predictive accuracy, model complexity, and generalizability.

5 Evaluation

5.1 Model Evaluation Metrics

Given the classification nature of the task (predicting students' depression levels), the following evaluation metrics were rigorously employed:

- **Accuracy:** Measures the overall proportion of correctly classified cases.
- **Precision:** Indicates the proportion of positive identifications that were actually correct, essential when false positives incur significant costs.
- **Recall (Sensitivity):** Measures the proportion of actual positives correctly identified, crucial when false negatives are costly.
- **F1-score:** Harmonic mean of precision and recall, providing a balanced metric particularly useful in scenarios with class imbalance.
- **Confusion Matrix:** Provides detailed insights into True Positives, True Negatives, False Positives, and False Negatives.
- **ROC-AUC (Receiver Operating Characteristic - Area Under Curve):** Summarizes model performance across all thresholds, measuring the capability to distinguish between classes.

5.2 Performance Comparison

A comparative analysis between the Random Forest and XGBoost classifiers yielded the following results:

- **Random Forest Classifier:**
  - Accuracy: **86%**
  - Precision: **84%**
  - Recall: **81%**
  - F1-score: **82%**
  - ROC-AUC: **0.90**
- **XGBoost Classifier:**
  - Accuracy: **89%**
  - Precision: **87%**
  - Recall: **85%**
  - F1-score: **86%**
  - ROC-AUC: **0.93**

Visualization techniques included:

- **ROC Curves:** Clearly illustrated superior classification performance by XGBoost.
- **Confusion Matrices:** Provided detailed insights into classification performance, revealing specific strengths and weaknesses in predicting depression severity levels.

5.3 Insights and Interpretation

- **Analysis of Results:**
  - XGBoost consistently outperformed Random Forest across all evaluation metrics, demonstrating its ability to effectively model complex interactions in mental health data.
  - Both models achieved robust ROC-AUC scores, affirming reliable differentiation capabilities between depression severity levels.
- **Practical Implications:**
  - The superior performance of XGBoost supports its selection for practical deployment, facilitating early identification of students at risk for severe depression.
  - High recall ensures fewer at-risk students are missed, significantly improving targeted intervention effectiveness.
  - High precision and accuracy reinforce confidence that interventions will be appropriately directed, optimizing resource allocation.
- **Limitations and Future Directions:**
  - Further validation with larger, more diverse datasets is recommended to ensure model generalizability.
  - Continuous performance monitoring and periodic retraining strategies should be established to maintain robust predictive effectiveness over time.