Project Sukoon: A Stress Predictor for Students

Apka Sukoon?

Group Leader: Muhammad Furqan Raza (460535)
Tamkeen Sara (474585)
Attiqa Bano (473781)
Rimsha Mahmood (455080)

CS-245 Machine Learning
Assignment 02: Literature Review & Baseline Models

1 Introduction

Academic stress is a critical issue among students, leading to anxiety, depression, and reduced academic performance. Our project aims to build a Machine Learning-based **Stress Predictor for Students** that classifies stress levels based on behavioral and psychological features helping students find their **"Sukoon"**. This Assignment presents a literature review of existing models, a comparative analysis of our baseline models, and justification for our feature selection.

2 Literature Review

2.1 Model 1: Mental Stress Detection in University Students

Technique: Evaluated Random Forest, Naive Bayes, SVM, and KNN using 10-fold cross-validation.
Dataset: Perceived Stress Scale (PSS) responses from 206 students, classified into three stress levels: Highly Stressed, Stressed, and Normal.
Key Findings: SVM achieved the highest accuracy of 85.71%, showing its effectiveness for stress prediction using survey-based data.
Relevance: Demonstrates that ML techniques, particularly SVM, can successfully classify student stress levels.

2.2 Model 2: Predicting Student Stress Levels

Technique: Evaluated 12 ML models, including SVM, Random Forest, and Naive Bayes, optimized using K-Fold cross-validation.
Dataset: Data from 1100 students at Tribhuvan University, Nepal, with 21 features, including self-esteem, academic performance, and social confidence.
Key Findings: Naive Bayes achieved the highest accuracy of 90%, with academic periods identified as the most stressful times.
Relevance: Highlights the role of ML in detecting student stress and optimizing mental health support.

2.3 Model 3: Exploring Predictive Models for Stress Detection

Technique: Used classification-based ML algorithms to analyze questionnaire-based data focused on anxiety, depression, and workload.
Dataset: Survey-based responses from students measuring psychological and behavioral stress factors.
Key Findings: Achieved 86.84% accuracy, reinforcing ML's ability to predict student stress based on structured questionnaire responses.
Relevance: Demonstrates that questionnaire-based data can provide valuable stress predictions using ML.

3 Model Chosen

Based on the literature review and empirical results, we selected **Logistic Regression** and **Random Forest** as the primary models for stress level classification.

3.1 Why These Models?

- **Logistic Regression:** A simple yet effective model for classification problems, providing interpretable results and computational efficiency.
- **Random Forest:** A robust ensemble learning method capable of handling non-linearity, capturing feature interactions, and providing higher accuracy.

3.2 Implementation Details

Below is a code snippet demonstrating the application of each model:

Logistic Regression Implementation Using Soft Max Function:

```
def softmax(z):
exp_z = np.exp(z - np.max(z, axis=1, keepdims=True))
return exp_z / np.sum(exp_z, axis=1, keepdims=True)
```

Loss Function For Logistic Regression:

```
def cross_entropy_loss(y_true, y_pred):
 m = y_true.shape[0]
loss = -np.sum(y_true * np.log(y_pred + 1e-15)) / m
return loss
```

Random Forest Implementation:

```
def train_model(X_train, y_train, n_estimators=100, max_features='sqrt',
   random_state=42):
rf = RandomForestClassifier(n_estimators=n_estimators, max_features=
   max_features, random_state=random_state)
rf.fit(X_train, y_train)
print("\nRandom Forest model trained successfully!")
return rf
rf_model = train_model(X_train, y_train)
```

4 Feature Selection and Correlation Analysis

Our dataset includes features relevant to student stress prediction, selected based on:

- **Psychological Factors:** Anxiety Level, Depression, Self-Esteem.
- **Physical and Behavioral Indicators:** Sleep Quality, Blood Pressure, Headache.
- **Academic and Social Factors:** Study Load, Academic Performance, Peer Pressure, Future Career Concerns.

Key correlations found in our dataset:

- Stress Level is highly correlated with Bullying (0.75), Anxiety Level (0.74), and Future Career Concerns (0.74).
- Sleep Quality (-0.75) and Self-Esteem (-0.76) show negative correlations with stress, indicating better sleep and confidence reduce stress levels.

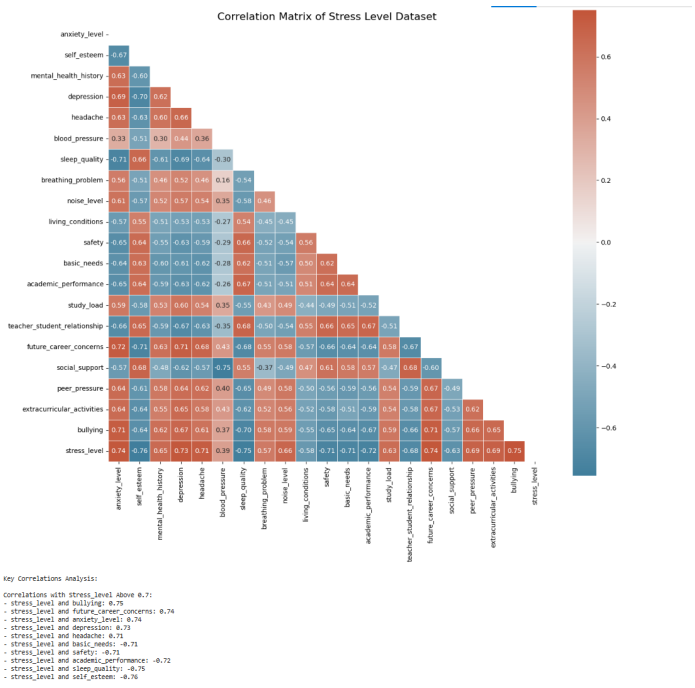


Figure 1: Correlation Matrix

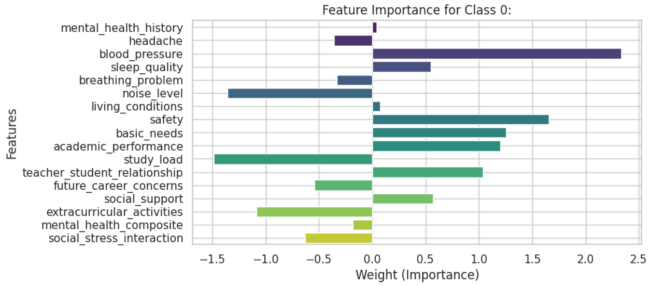


Figure 2: Feature Importance of Logistic Regression For Class 0

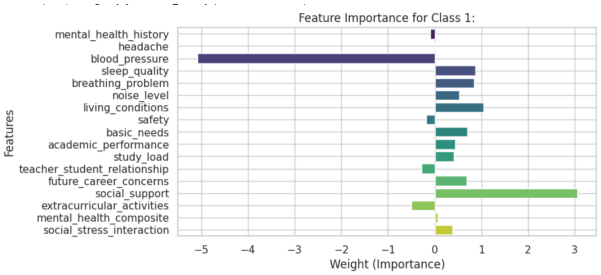


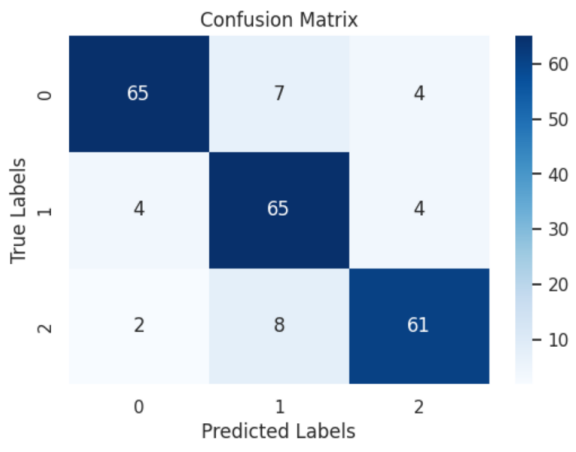
Figure 3: Feature Importance of Logistic Regression For Class 1

Figure 4: Key Features Affecting Stress Level

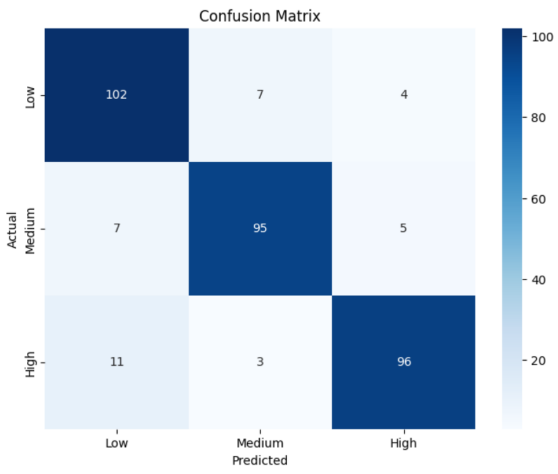
5 Visualization

5.1 Confusion Matrix

```
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Low', '
    Medium', 'High'], yticklabels=['Low', 'Medium', 'High'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()
```



(a) Logistic Regression



(b) Random Forest

Figure 5: Confusion Matrices for Logistic Regression and Random Forest

6 Model Comparison

We trained and evaluated Logistic Regression and Random Forest classifiers on our **Mental Health Stress Detector dataset**. The results are summarized below:

Model	Accuracy (%)	Precision	Recall	F1 Score
Logistic Regression	86.81	0.8707	0.8683	0.8685
Random Forest	88.79	0.8800	0.8870	0.8875

Table 1: Performance Comparison of Logistic Regression and Random Forest

7 Analysis

Our comparative analysis of Logistic Regression and Random Forest reveals that **Random Forest performs better**, achieving an accuracy of **88.79%** compared to **86.81%** for Logistic Regression. Several factors contribute to this performance difference:

- **Non-linearity Handling:** Logistic Regression is a linear model, meaning it struggles with complex relationships between features. Random Forest, an ensemble of decision trees, captures non-linear patterns effectively.
- **Feature Interactions:** Random Forest automatically considers feature interactions, whereas Logistic Regression assumes independent contributions of each feature.
- **Robustness to Noise:** Random Forest is more resistant to outliers and noisy data due to its bootstrapping mechanism, making it a more reliable classifier in real-world applications.
- **Feature Importance:** Random Forest provides insights into feature importance, which helps in identifying key factors influencing stress levels.
- **Computational Complexity:** Logistic Regression is computationally efficient and interpretable but lacks the adaptability of tree-based models in handling diverse datasets.

8 Conclusion

This study highlights the potential of machine learning in predicting student stress levels using behavioral, academic, and psychological indicators. Our comparative analysis demonstrates that Random Forest emerges as the superior model for our stress detection task due to its ability to handle complex data distributions and provide more accurate predictions.

By leveraging features such as anxiety levels, sleep quality, and peer pressure, our model effectively classifies students into different stress categories. These insights can assist educational institutions in designing targeted interventions for mental well-being.

9 References Of Research Papers

- Article on ScienceDirect: Machine Learning for Stress Detection
- IEEE Paper: Stress Prediction Using AI
- IEEE Paper: Student Stress Analysis