Course Code: CSE422

Course Title: Artificial Intelligence

Semester: Fall 2024

Project Title: Student Stress Level Prediction

Group: 7

Section: 12

Group Information:

- **Anika Bushra ID: 21201068**
- **Md. Jihadul Amin ID: 24241200**

Table of Contents:

Introduction

Modern lifestyles expose individuals to numerous stress-inducing factors that significantly influence their physical and mental well-being. By understanding the interplay between lifestyle choices and stress, we can uncover valuable insights that lead to healthier living. This dataset explores the relationship between various lifestyle factors and stress levels, providing a comprehensive view of how daily habits and environmental conditions contribute to overall well-being.

The primary objective of this analysis is to examine key lifestyle attributes, such as anxiety level, self-esteem, mental health history, sleep quality, academic performance, and social support, and their impact on stress. Stress is a critical factor affecting productivity, health, and happiness. Analyzing the dataset aims to uncover actionable patterns and correlations, empowering individuals and institutions to make informed decisions about stress management.

Dataset Description

The dataset provides detailed information on 21 variables collected from a diverse group of individuals, encompassing mental health indicators, lifestyle habits, and environmental conditions. It contains 1,100 rows, each representing an individual's profile. These attributes are designed to offer insights into lifestyle patterns and their potential impact on stress levels.

Dataset Link: [Student Stress Level Dataset]

Key features include:

- **Anxiety Level, Self-esteem, and Depression:** Quantitative indicators of mental health status.
- **Physical Symptoms:** Metrics such as blood pressure, headache frequency, and breathing problems.
- **Sleep and Noise Levels:** Environmental and behavioral factors impacting well-being.
- **Academic and Social Attributes:** Study load, academic performance, teacher-student relationships, and social support.

- **Environmental and Safety Factors:** Living conditions, safety perceptions, and basic needs fulfillment.
- **Stress Level:** The target variable, representing the individual's perceived stress level.

The dataset includes both numeric and categorical variables, offering a rich foundation for exploratory analysis and predictive modeling. By leveraging this dataset, we aim to develop predictive models and insights that inform strategies to enhance lifestyle choices and reduce stress.

Libraries Used:

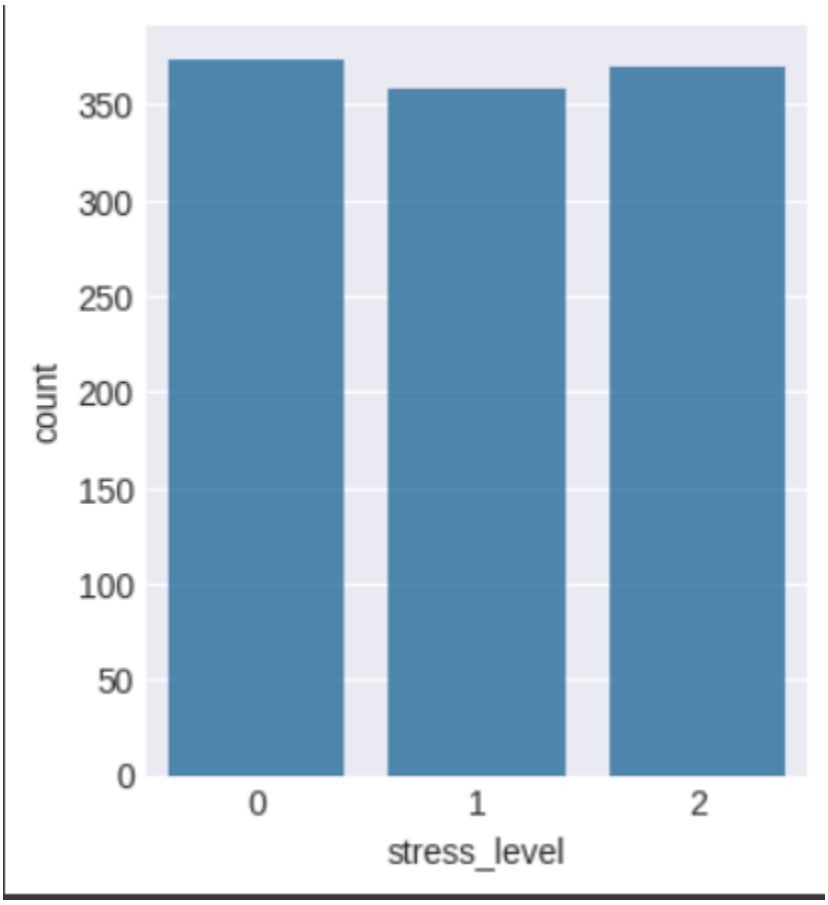
- The project uses the following libraries:
- Pandas - for data manipulation and analysis
- Numpy - for numerical operations
- Sklearn - for machine learning models and preprocessing
- Matplotlib - for data visualization
- Seaborn - for enhanced visualization

Dataset Characteristics

- Number of Features: 20 (including anxiety_level, self_esteem, mental_health_history, depression and 16 measurements)
- Type of class/label: multi-class classification
- Number of data points: 1100
- Types of features: Quantitative (continuous numerical values)

Dataset Distribution (Balanced)

The class distribution is balanced:



Feature Correlation:

Data Splitting and Feature Scaling

- Training split: 70% (770 samples)
- Testing split: 30% (330 samples)

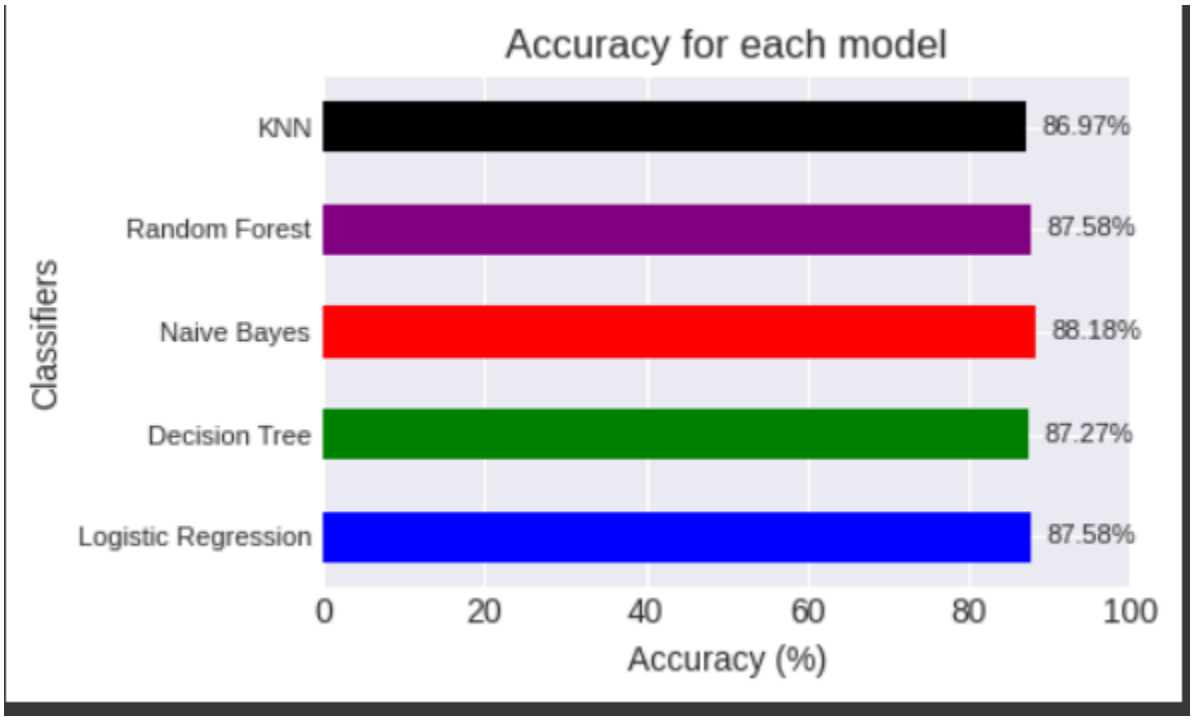
- Used stratified splitting to maintain class distribution. Applied StandardScaler for feature normalization

Machine Learning Model Used for Prediction

- Logistic Regression
- Decision Tree
- Naive Bayes
- Random Forest
- KNN

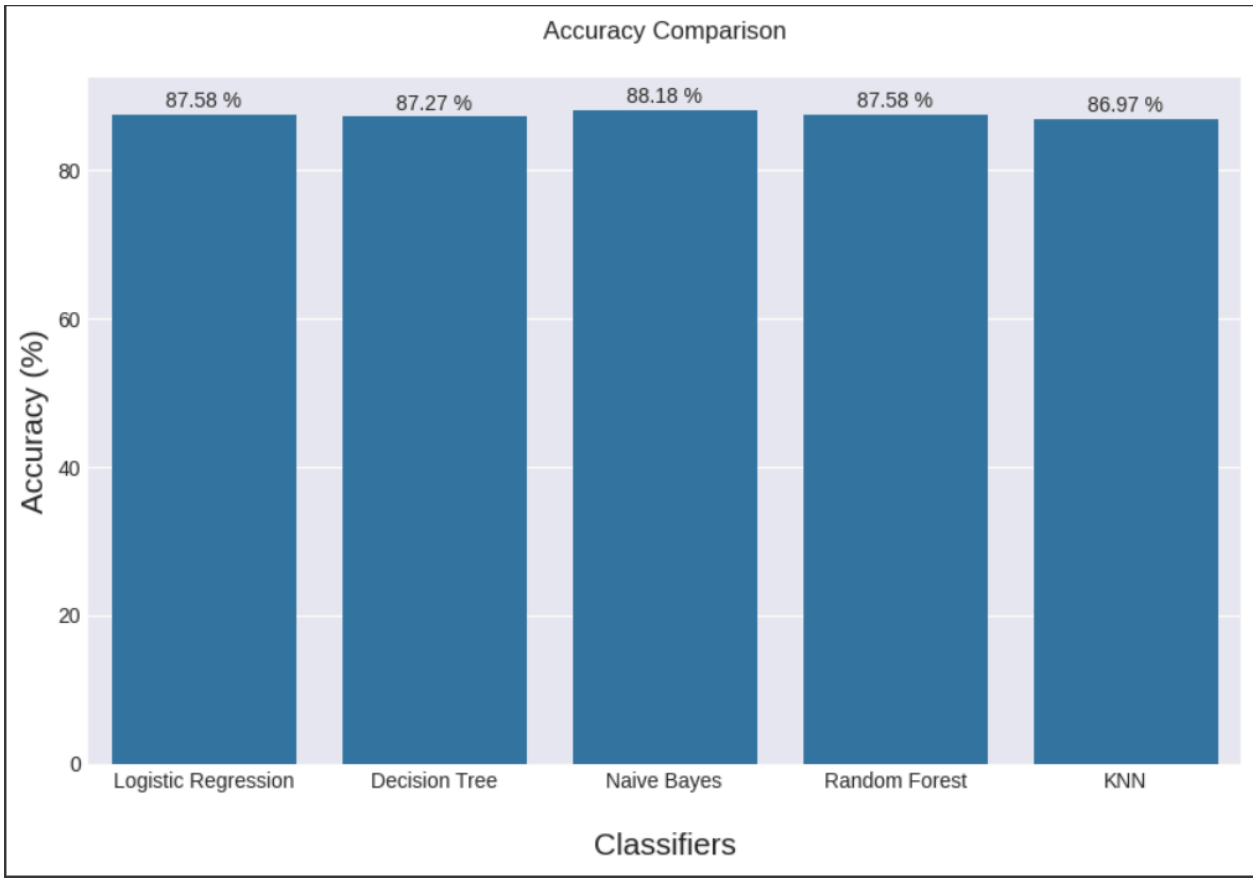
Result Model Selection/Comparison Analysis

1. Bar Chart Showcasing Prediction Accuracy

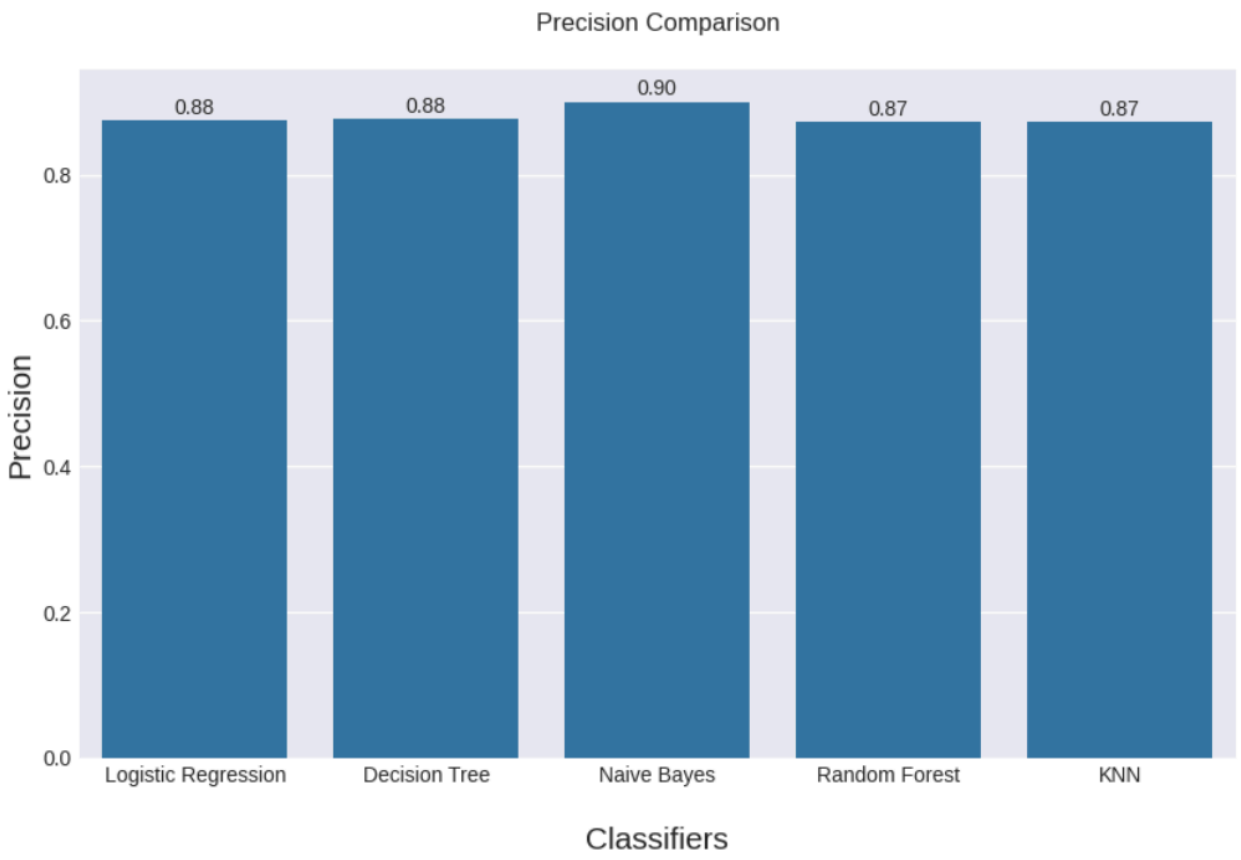


Model Comparison Analysis:

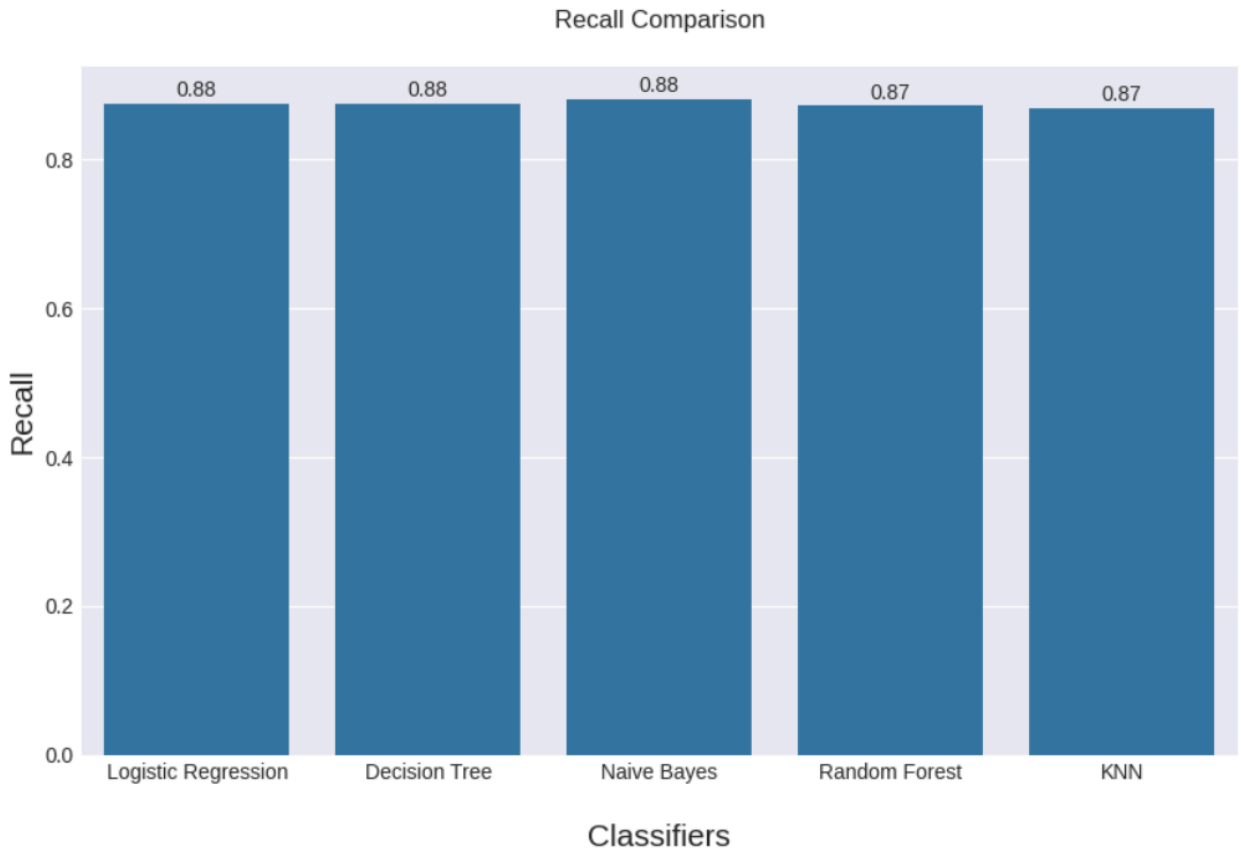
All models showed strong performance:



Precision Comparison:



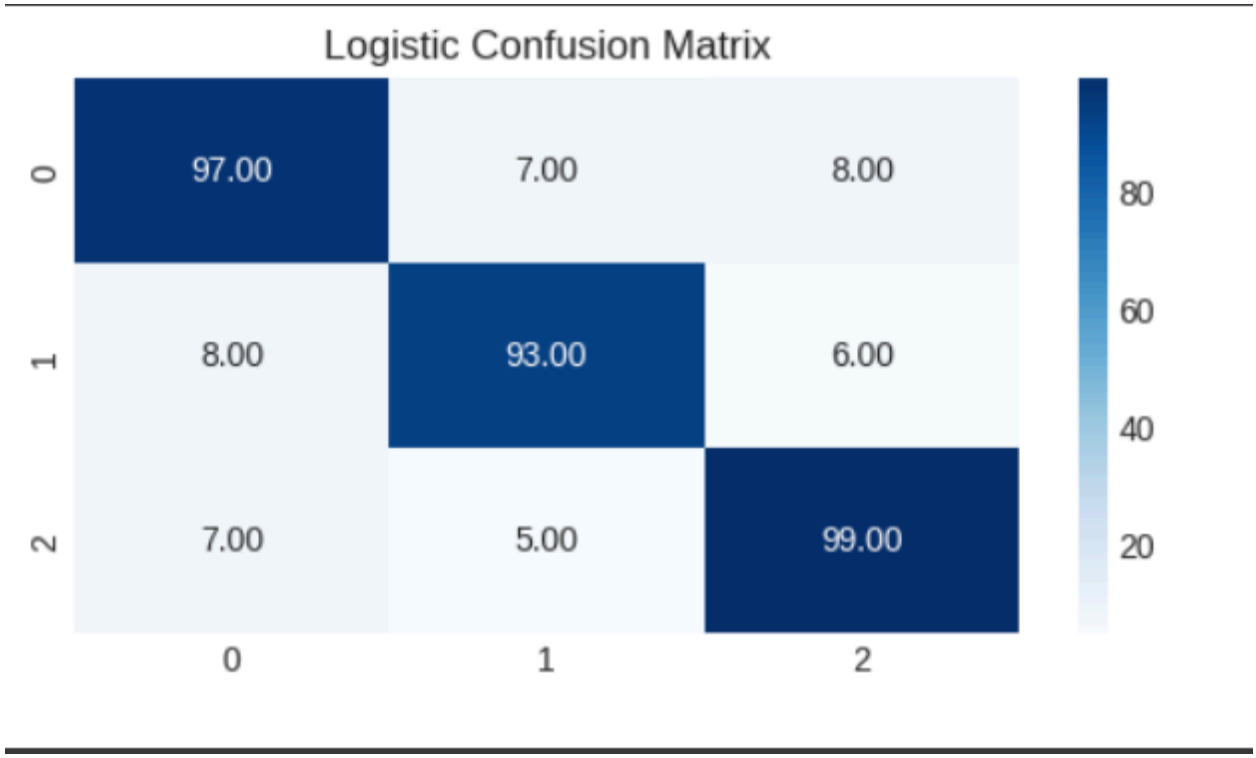
Recall Comparison

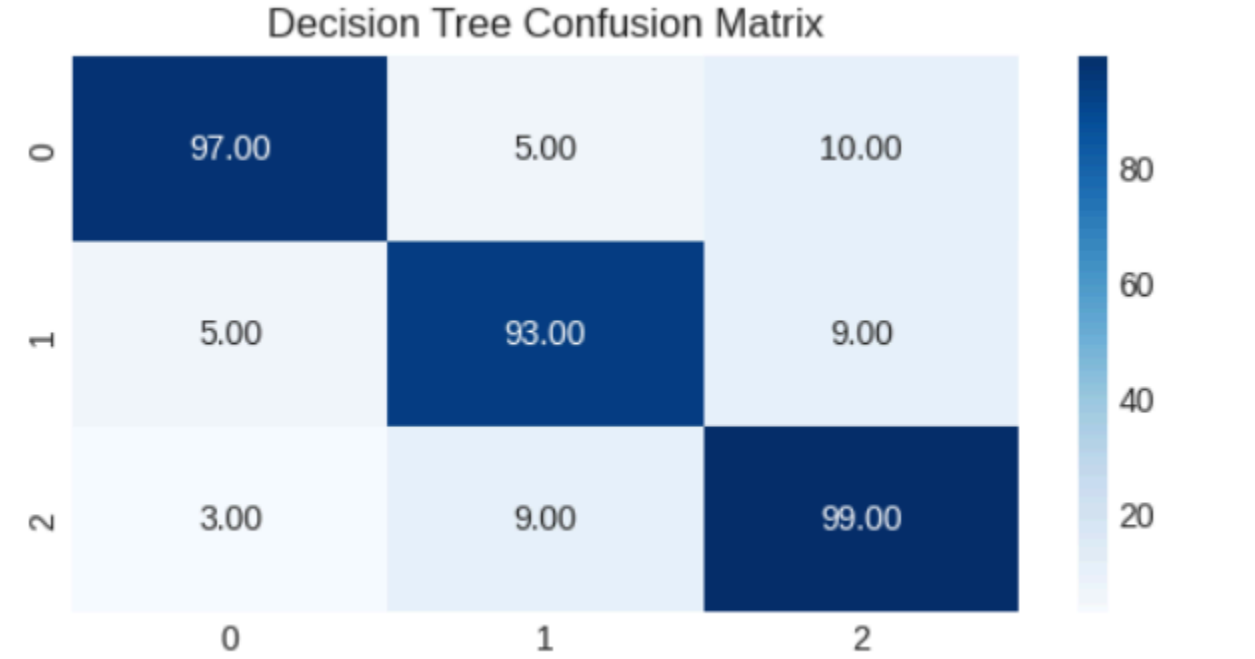


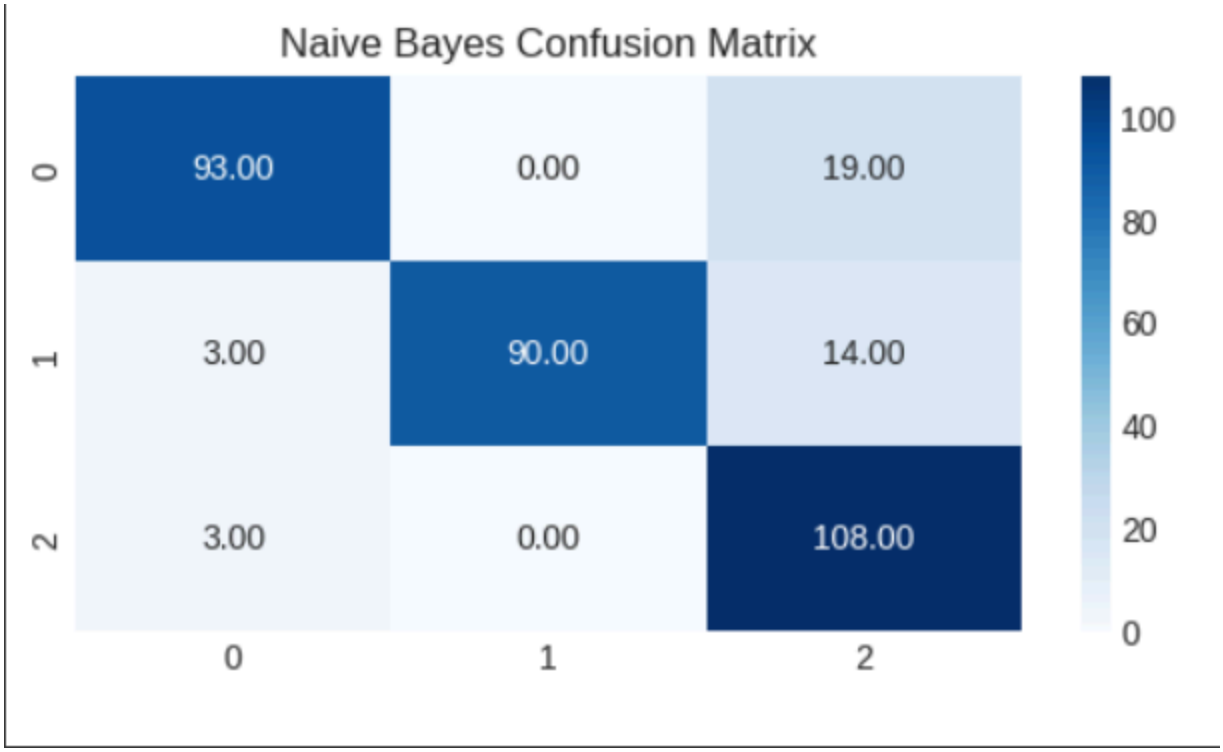
- **Confusion Matrix Results:**

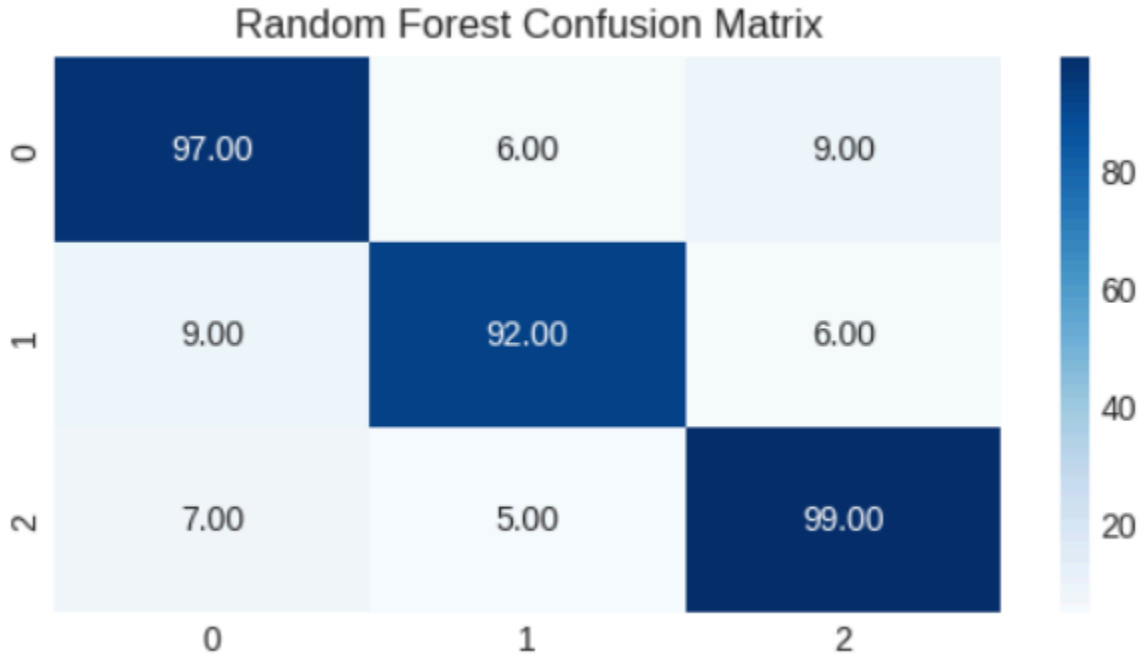
  The confusion matrices for all models reflected strong performance, with very few misclassifications across all datasets, further confirming the accuracy of the predictions.

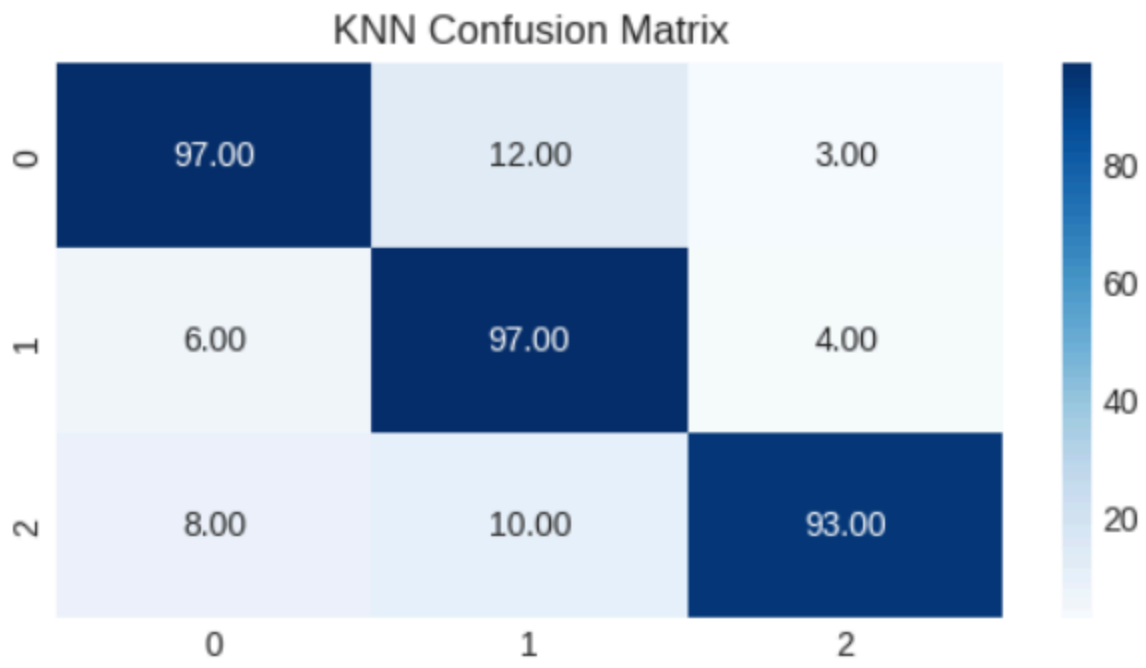
  These results underscore the exceptional performance of all models, with Logistic Regression, Decision Tree, Naive Bayes, Random Forest, KNN closely following.


Logistic Confusion Matrix
97.00
7.00
8.00
8.00
93.00
6.00
7.00
5.00
99.00
0
1
2
0
1
2
80
60
40
20


Decision Tree Confusion Matrix
97.00
5.00
10.00
5.00
93.00
9.00
3.00
9.00
99.00
0
1
2
0
1
2
80
60
40
20


Naive Bayes Confusion Matrix
93.00
0.00
19.00
3.00
90.00
14.00
3.00
0.00
108.00
0
1
2
0
1
2
100
80
60
40
20
0


Random Forest Confusion Matrix
97.00
6.00
9.00
9.00
92.00
6.00
7.00
5.00
99.00
0
1
2
0
1
2
80
60
40
20



Conclusion

The remarkable consistency across multiple machine learning models, all achieving accuracy rates above 86%, reveals something fascinating about the nature of stress prediction. While Naive Bayes marginally led the pack at 88.18%, the negligible performance gap between different approaches from the straightforward Logistic Regression 87.58% to the more complex Random Forest 87.58% suggests that stress patterns may be more structured and detectable than previously thought. This convergence of results, regardless of model complexity, hints at clear underlying patterns in how stress manifests in measurable indicators. Such findings not only validate the reliability of machine learning in stress detection but also open intriguing possibilities for real-world applications, from preventive healthcare to workplace wellness programs. The high precision and recall scores further strengthen the case for implementing these models in sensitive real-world scenarios where false positives could be as concerning as missed cases

References

1. Scikit-learn documentation: https://scikit-learn.org/

2. Dataset: https://www.kaggle.com/datasets/rxnach/student-stress-factors-a-comprehensive-analysis

3. Seaborn visualization library: https://seaborn.pydata.org/