Faculty of Engineering and Technology

Electrical and Computer Engineering Department

MACHINE LEARNING AND DATA SCIENCE – ENCS5341

Assignment#3 (Project) – Report

Prepared By:-

Alaa Saleem-1200001

JanaAbuNasser-1201110

Section#: 1

Instructor: Dr. Yazan Abu Farha

Date of Submission: 26.Jan.2024

BIRZEIT

Jan – 2024

Introduction

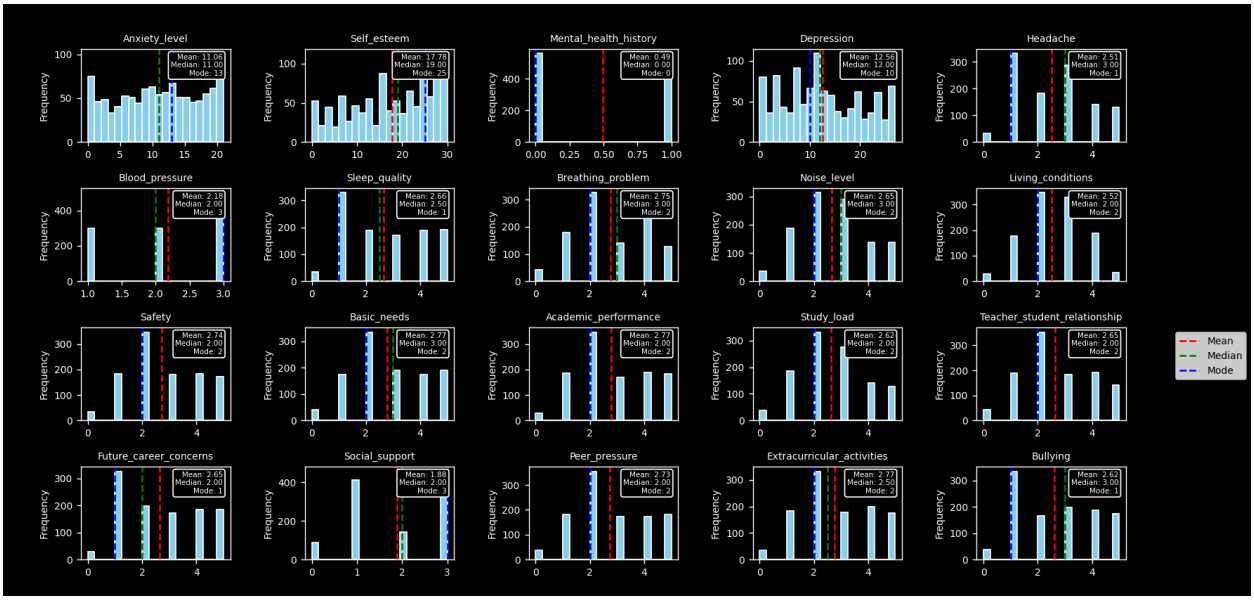
In this project, we tackle 3-class classification task that involves the analysis of various models to determine the most effective approach for our project. Our primary objective is to identify the best-performing model among Logistic Regression, Decision Trees, Support Vector Machines (SVM), Ensemble Methods (such as Adobos and Bagging), and Random Forest. We aim to apply these methods to our dataset and evaluate their respective performance to make an informed choice between Logistic Regression and Random Forest after we did the Baseline model, as indicated by the classification reports. Throughout our evaluation, we employ a set of well-defined evaluation metrics, including Accuracy, Precision, Recall, F1 Score, and the Confusion Matrix, to gain a comprehensive understanding of each model's capabilities and ultimately select the most suitable one for our specific classification task.

Data Set

The dataset contains information that looks at the things that make students stressed. It covers things like feeling anxious or having low self-esteem, as well as the pressure of schoolwork and how they interact with others. This collection gives a complete picture of what challenges students face today. The main file of this dataset is named " StressLevelDataset.csv" containing 1100 examples and 20 feature for each example (with no missing values for all features) with one discrete target variable “stress level”.The goal is to determine the stress level using 3 values (0,1 and 2).

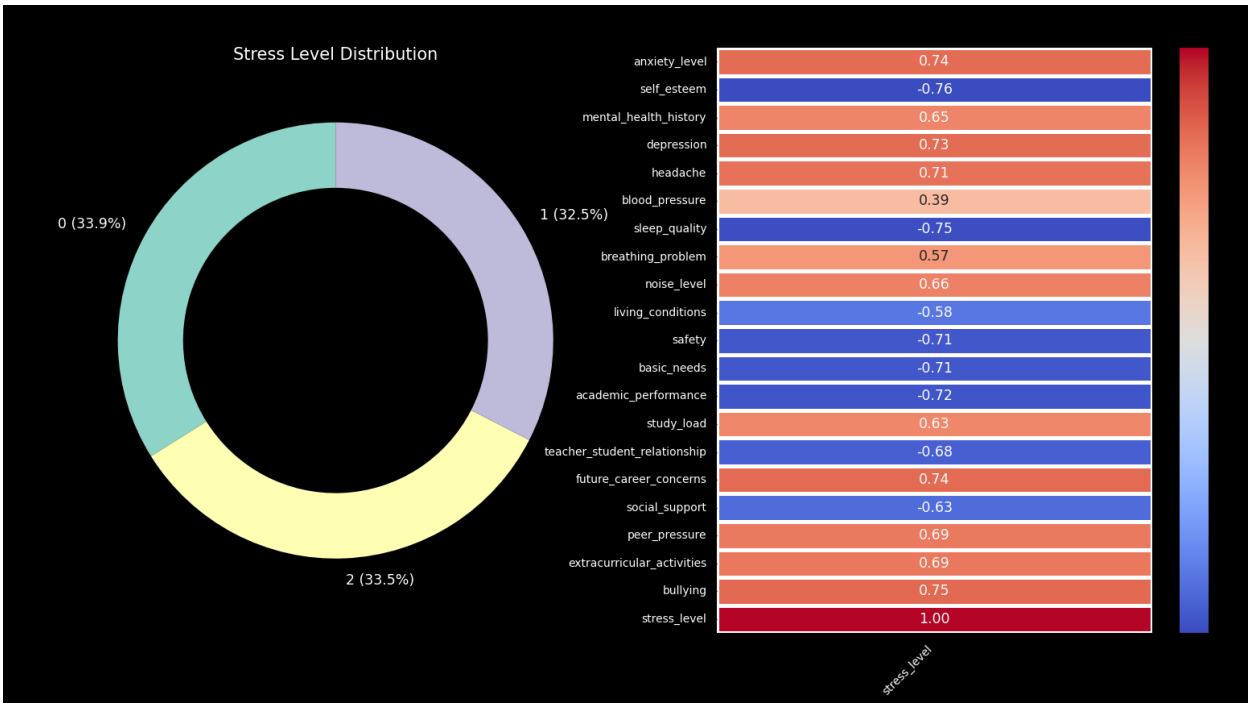
Link of the dataset: [Student Stress Factors: A Comprehensive Analysis (kaggle.com)]

Descriptive Statistics



This appears to be a comprehensive collection of histograms, each representing the frequency distribution of responses or measurements across a range of psychological and environmental features. Each histogram displays mean, median, and mode values for its respective variable. For instance, in the anxiety level histogram, the mode is 13, and the mean is slightly lower. Similarly, self-esteem has a mode of 25, indicating the most common response, with a slightly lower mean. Mental health history is binary, with a mode at 0, indicating most respondents lack such history. Depression's most common value is 10, but the mean is slightly higher. Headaches range from 0 to 5, with 1 as the most frequent value. Blood pressure has three levels, with level 3 being the most common. Other parameters use a 5-point scale, each with different

mean, median, and mode values. These histograms feature mean (red line), median (green line), and mode (blue line) markers, offering insights into data distribution. Close alignment of mean, median, and mode suggests symmetry, while wider spreads indicate data skewness.



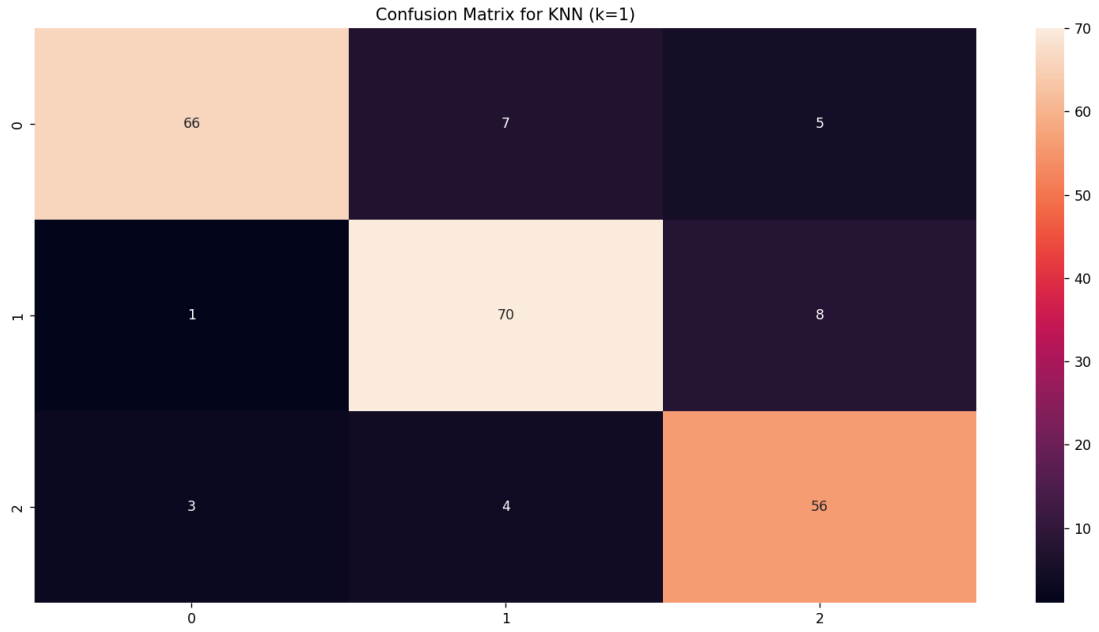
The visual data representation consists of a donut chart and a bar graph, both conveying information about stress levels. The donut chart is divided into three segments, representing stress levels 0, 1, and 2, with roughly equal distribution among respondents, indicating no predominant stress level. On the right, the bar graph shows the importance of various features in relation to stress levels. Features like "anxiety_level," "depression," and "bullying" have strong positive importance, indicating a direct relationship with higher stress levels. Conversely, "self-esteem," "sleep_quality," and "basic_needs" have strong negative importance, suggesting an association with lower stress levels. The color scale on the right represents correlation strength, with red indicating strong positive correlation and blue a strong negative correlation. Together, these visualizations provide a comprehensive overview of stress distribution and its influencing factors.

Models used

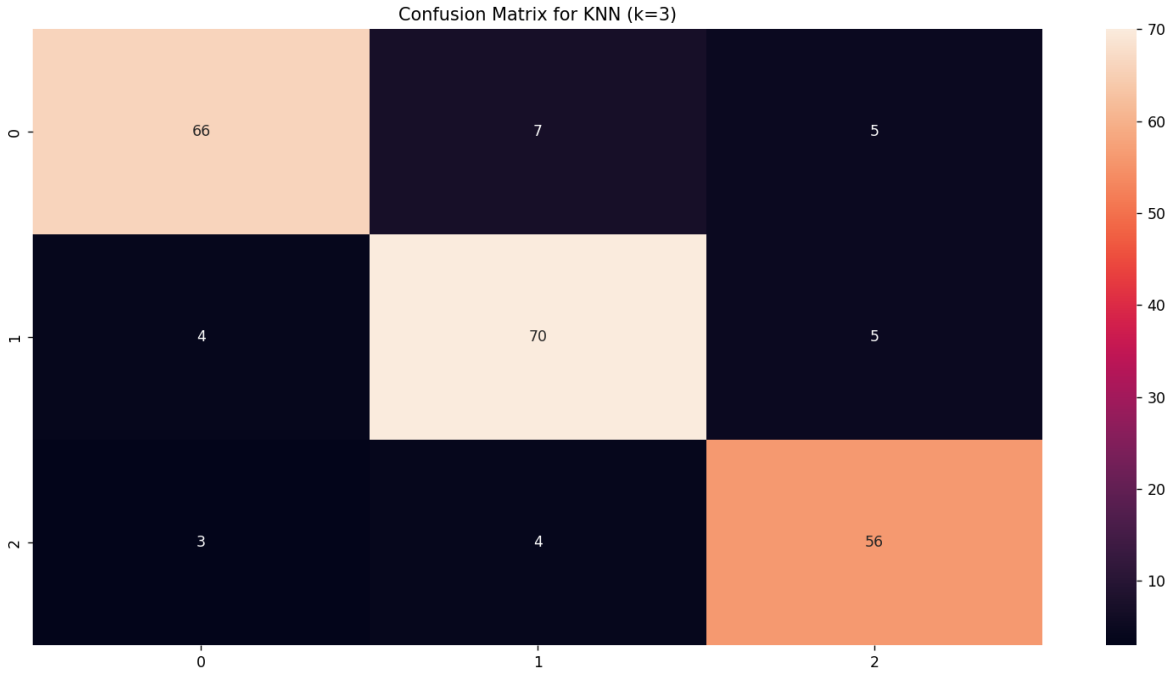
Before applying models, we have split our data using 60% training, 20% validation and 20% testing sets, and since we want to compare and select the best model.

1. KNN (K=1 and K=3)

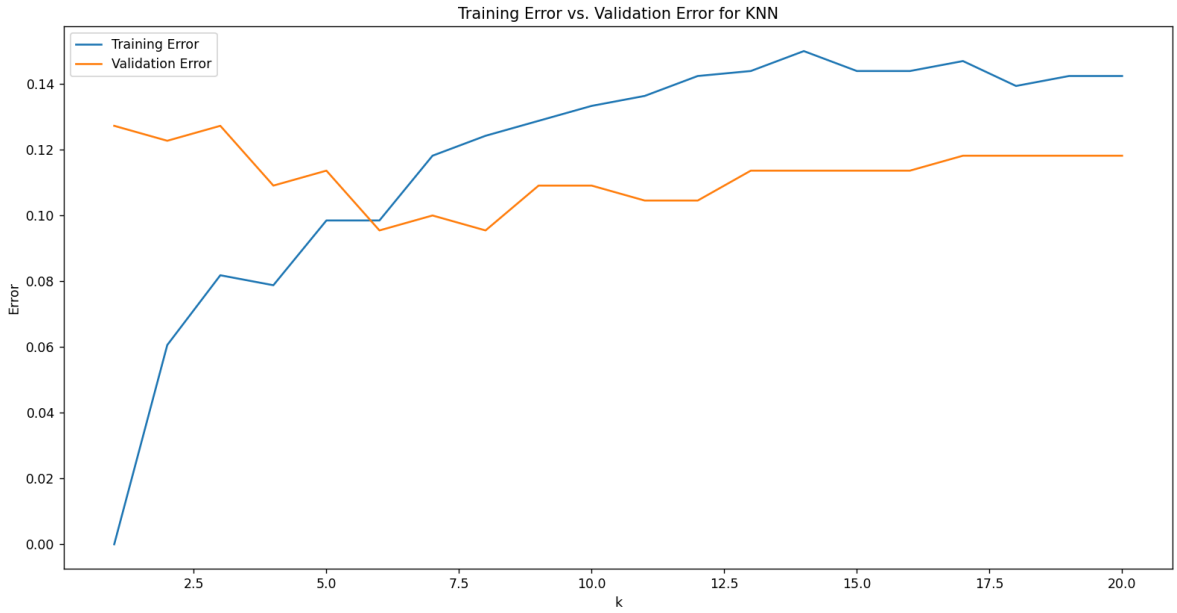
For the distance choice we decide to use the default distance which is the Minkowski distance with p=2 (Euclidean) since the features has numeric data, we did a standard scale.



The confusion matrix for a k-Nearest Neighbors (kNN) classifier with k=1 provides a comprehensive overview of the model's classification performance. It consists of diagonal elements (66, 70, 56) representing true positives for each class ('0,' '1,' '2') where the predicted label matches the true label. Off-diagonal elements indicate misclassifications, e.g., '7' in the first row and second column signifies 7 instances falsely predicted as class '1' instead of '0.' The color intensity represents instance counts, aiding in identifying confusion between classes. Overall accuracy can be computed from the diagonal elements, while the off-diagonal elements offer insights into specific model errors, crucial for model improvement.

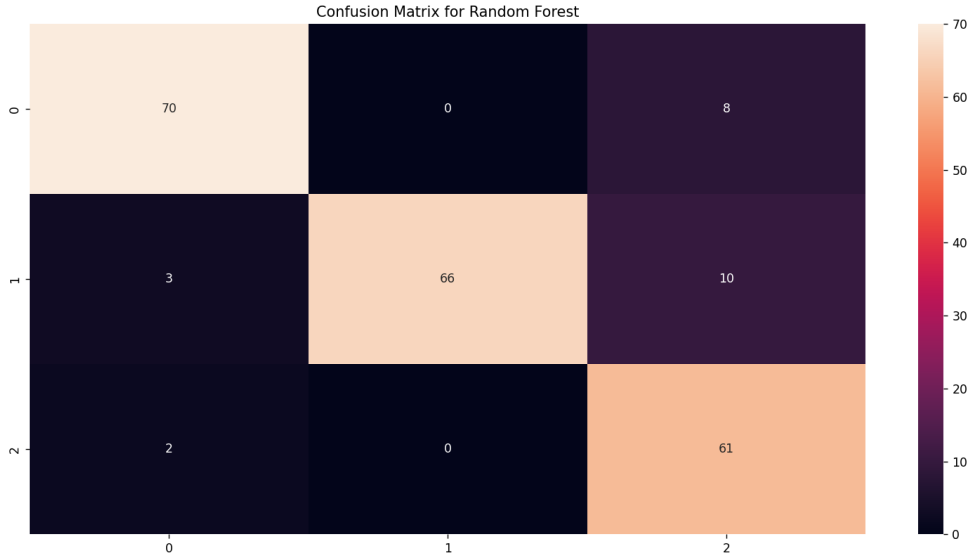


The confusion matrix for a k-Nearest Neighbors (kNN) classifier with k=3 evaluates the model's classification performance. It shows true positives (66, 70, 56) for classes '0,' '1,' and '2,' representing correct predictions. The non-diagonal cells reveal misclassifications, like '7' for class '0' incorrectly predicted as '1.' Another '4' in the second row, first column indicates four '1' instances wrongly classified as '0.' Additionally, '5' in the first row, third column, and the second row, third column illustrate instances of class '0' and '2' being misclassified as each other. The color gradient aids in identifying higher instance counts. Comparing this matrix to k=1, the diagonal remains consistent, suggesting unaltered accuracy for correct predictions. However, the distribution of misclassifications may vary, offering insights into the kNN classifier's performance with different k values or datasets.



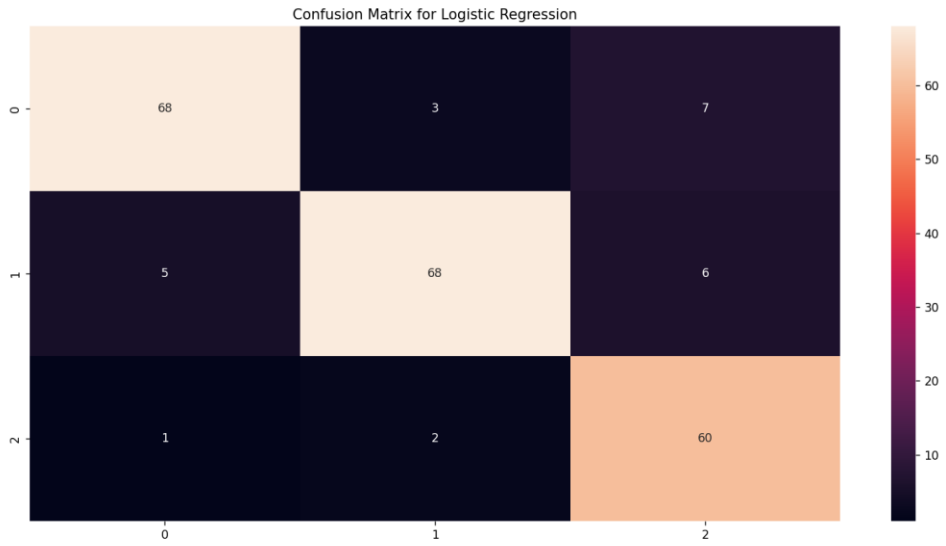
The graph represents a K-Nearest Neighbors (KNN) model's performance with different 'k' values, representing the number of neighbors considered. "Training and Validation Errors vs. k," contrasts training and validation error rates (y-axis) for different 'k' values (x-axis). At k=1, the training error is 0%, signaling overfitting, while the validation error is higher. Conversely, at k=3, the model balances complexity and generalization better.

2. Random Forest



The confusion matrix for the Random Forest classifier visualizes the model's performance. The diagonal elements (70, 66, 61) represent accurate predictions for classes '0,' '1,' and '2.' For example, class '0' was correctly predicted 70 times. Off-diagonal elements indicate misclassifications, such as '8' instances of class '0' wrongly predicted as '2.' Class '1' had '3' misclassifications as '0' and '10' as '2.' Class '2' had '2' misclassifications as '0.' Interestingly, no misclassifications occurred between classes '0' and '1.' This result suggests that these two classes may be more separable in the dataset or that the Random Forest algorithm excels in distinguishing them.

3. Logistic Regression



The confusion matrix for the Logistic Regression classifier provides a visual representation of its classification performance. It comprises diagonal numbers (68, 68, 60) denoting correct predictions for each

class (true positives): '0' was predicted accurately 68 times, '1' 68 times, and '2' 60 times. Off-diagonal numbers indicate misclassifications, such as 3 '0' instances misclassified as '1,' 7 '0' instances misclassified as '2,' 5 '1' instances misclassified as '0,' 6 '1' instances misclassified as '2,' 1 '2' instance misclassified as '0,' and 2 '2' instances misclassified as '1.' The color gradient visually represents instance counts, with darker shades indicating higher numbers, aided by the gradient scale on the right side.

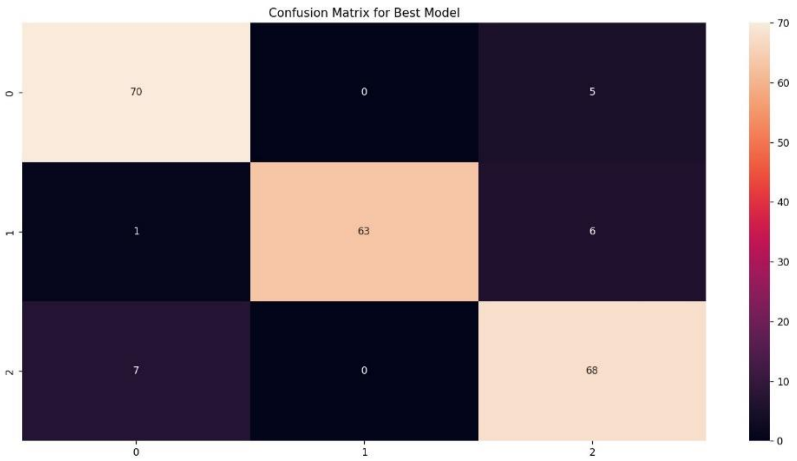
We change the hyperparmeter for each model and the best model was chosen, and the accuracy was improved.

4. Performance analysis

In this part, the performance of the best model from the previous part is analyzed. As mentioned before, we used 3 models one of them is the baseline model (KNN) one with k=1 and the other is with k=3, and the others are Random Forest and Logistic Regression. As we progressed in the project, we followed the metrics Accuracy, Precision, recall, and F1, the Random Forest and logistic regression have the highest accuracy and recall but the Random Forest has the highest precision and F1 so it is the best model.

```
------------------------- KNN -------------------------
   k  Accuracy  Precision    Recall        F1
0  1  0.872727   0.877022  0.872727  0.873396
1  3  0.872727   0.873849  0.872727  0.872764
------------------------- Random Forest -------------------------
Best parameters: {'max_depth': 1, 'n_estimators': 100}
Best accuracy score: 0.8651515151515152
   Accuracy  Precision    Recall        F1
0  0.895455   0.911116  0.895455  0.897348
------------------------- Logistic Regression -------------------------
Best parameters: {'C': 1, 'max_iter': 100}
Best accuracy score: 0.8681818181818182
   Accuracy  Precision    Recall        F1
0  0.890909   0.895662  0.890909  0.891191
------------------------- Best Model -------------------------
Accuracy in Best Model  0.9136363636363637
Precision in Best Model  0.9175666106045854
Recall in Best Model  0.9136363636363637
F1 Score in Best Model  0.9144409394901937
Classification Error for Best Model: 8.64%
```

The performance metrics for three different algorithms, namely k-Nearest Neighbors (KNN), Random Forest, and Logistic Regression, The best model is Random Forset and The classification error for the best model is 8.64%, making it the top-performing model among the three algorithms.



5. Conclusion

In conclusion, this project successfully explored various machine learning models, including Logistic Regression, and Random Forest, for a 3-class classification task focused on determining stress levels in students. The dataset, comprising 1100 examples with 20 features each, provided a comprehensive picture of factors contributing to student stress. The models were evaluated using metrics like Accuracy, Precision, Recall, F1 Score, and the Confusion Matrix. Ultimately, Random Forest emerged as the best-performing model based on its high precision and F1 score, outperforming others in accurately predicting stress levels. This project not only highlights the effectiveness of Random Forest in handling complex classification tasks but also underscores the importance of machine learning in understanding and addressing student stress. we faced several limitations. Firstly, the dataset size was relatively small, which may not fully represent the diversity of student stress factors. Secondly, the dataset's features were limited, potentially overlooking other relevant variables affecting stress levels. Additionally, the binary nature of some features oversimplified complex issues. The models' performance could also be influenced by the imbalance in class distribution, leading to biases.