Stuyding the Impact on Student Stress level from different factors

MD. Farhan Sadik
ID: 21-44403-1
Dept Name:CSE
Institute Name:American International University-Bangladesh
Dhaka, Bangladesh
farhansadik701@gmail.com
01992501067

TAIFUR MOHAMMAD TANIM
ID: 20-43338-1
Dept Name: CSE
Institute Name: American International University-Bangladesh
Dhaka, Bangladesh
taifurmohammadt@gmail.com
01970891284

MAHNAZ HOSSAIN
ID: 20-43835-2
Dept Name:CSE
Institute Name: American International University-Bangladesh
Dhaka, Bangladesh
mahnazhossain258@gmail.com
01875202263

Rahman khandoker Alvee
ID: 19-41492-3
Dept Name:CSE
Institute Name: American International University-Bangladesh
Dhaka, Bangladesh
alveerahman47@gmail.com
01300773448

Abstract
This project aimed to study in-depth of what causes stress in students through multiple factors and also aid in understanding the core factors among many others leading to an analytical result. Methodologies that were utilized were feature reduction techniques, ensemble learner, specifically stacking which utilized multiple base models, resulting in better accuracy.

I. Introduction

A machine learning project was conducted to understand the factors contributing to students' elevated stress levels. The project used ensemble learning, a powerful paradigm that combines the predictions of multiple base models to enhance performance and robustness. The ensemble consisted of four diverse base models: Logistic Regression, Random Forest, Naive Bayes, and K-Nearest Neighbors. Each model provided unique insights into linear relationships between features and stress levels. Logistic Regression, a linear model, provided insights into linear relationships between features and stress levels. Random Forest, an ensemble of decision trees, excelled at capturing non-linear relationships and complex interactions. Naive Bayes, a probabilistic model, considered the likelihood of different factors contributing to stress. K-Nearest Neighbors, a distance-based method, considered the proximity of instances in the feature space. The project aimed to predict stress levels accurately and identify the most influential features contributing to heightened stress, offering actionable insights for educational institutions, policymakers, and support services.

II. Motivation of the Project

Through the use of machine learning techniques, particularly ensemble learning, this project seeks to address mental health issues that students encounter in the context of education. With early warning systems for focused interventions, the project offers a prediction tool for determining stress levels. Through the identification of significant elements, it also helps to create mental health awareness and assistance inside educational institutions. In order to establish a more encouraging learning environment, the analysis supports individual students as well as institutional behaviors and policy. The initiative supports a comprehensive approach to mental health in educational institutions and is consistent with a dedication to using technology for the good of society.

III. Objective of the Project

- **Predictive Accuracy:** Develop an ensemble learning model that accurately predicts stress levels in students by combining the strengths of Logistic Regression, Random Forest, Naive Bayes, and K-Nearest Neighbors. The primary goal is to provide a reliable tool for early detection of elevated stress.

- **Feature Identification:** Gain insight into the intricate interactions between academic, social, and personal aspects by discovering the most significant elements influencing stress levels. This will make it possible to comprehend the underlying reasons why students experience stress on a deeper level.

- **Interpretability:** Use Logistic Regression as the meta-model to optimize stress prediction interpretability. The goal is to provide educators, administrators, and support services with simply understandable and applicable actionable insights and recommendations.

- **Supportive Learning Environment:** Contribute to the creation of a supportive learning environment by providing educators and institutions with the tools to proactively address stress-related challenges. The project seeks to enhance mental health awareness within educational settings.

- **Contribution to Research:** Advance the field of machine learning in the context of mental health by exploring ensemble techniques for stress prediction. The project aims to contribute valuable findings to the broader research community working on mental health and educational analytics.

- **Ethical Considerations:** Prioritize ethical considerations in data usage and model deployment, ensuring that the project aligns with privacy norms and guidelines. The goal is to develop a responsible and ethical framework for stress prediction in the educational domain.

- **Societal Impact:** Increase the influence to improve educational procedures and policies, not just individual pupils. The project's objectives are to promote proactive steps to enhance students' well-being and to spark more widespread conversations in society on mental health in education.

IV. METHODOLOGY

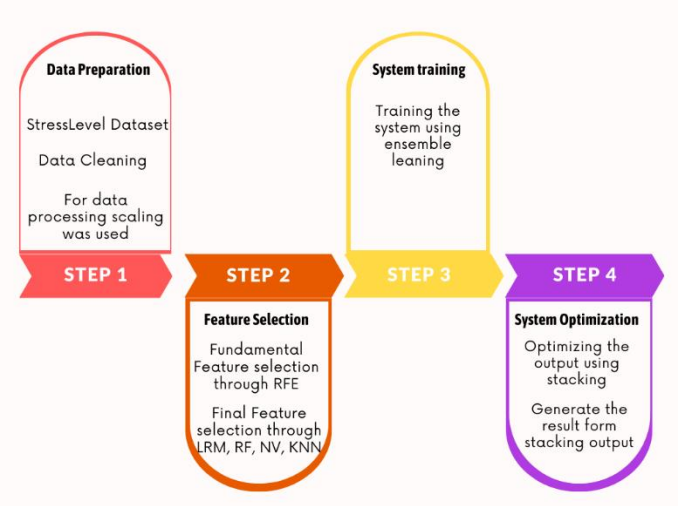

Fig 1: Methodology flow diagram

A. Data Collection

The data was collected from Kaggle's large collection of datasets, more specifically from [1]. This dataset is all about understanding the inherent causes and their impact on students today through multiple real-life factors such as sleep quality, study load, and even bullying.

B. Data processing

With 'stress_level' as the target variable, the dataset presents a valuable opportunity for beginners and seasoned data enthusiasts to explore the dynamics affecting student well-being. During data exploration, the dataset is split into training and testing sets using train_test_split with a test size of 20% and training size of 80% of the data with a random state of 42 for reproducibility.

The usage of 'StandardScaler' normalizes the features especially those that have a wider range of values when compared to other values. Additionally, feature reduction methods are utilized to reduce the dimensionality of the dataset which in turn reduces the complexity and makes it easier to perform visualization of the dataset through different plots.

C. Dataset description

This dataset contains around 20 features that create the most impact on the Stress of a Student. The features are selected scientifically considering 5 major factors, they are Psychological, Physiological, Social, Environmental, and Academic Factors. Some of them are:
Psychological Factors => 'anxiety_level', 'self_esteem', 'mental_health_history', 'depression',

Physiological Factors => 'headache', 'blood_pressure', 'sleep_quality', 'breathing_problem

Environmental Factors => 'noise_level', 'living_conditions', 'safety', 'basic_needs',

Academic Factors => 'academic_performance', 'study_load', 'teacher_student_relationship', 'future_career_concerns',

Social Factor => 'social_support', 'peer_pressure', 'extracurricular_activities', 'bullying'.

Below is one such plot that shows how much the selected 5 factors affect the stress level of a student

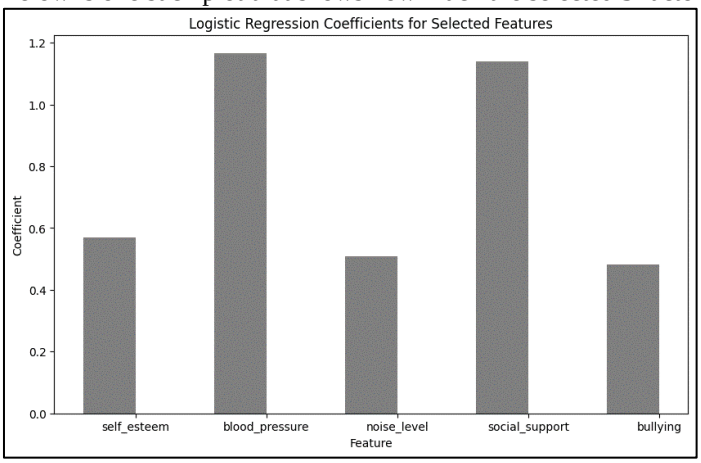

Fig 2: Top 5 factors on stress level, filtered through RFE and trained using a logistic regression model

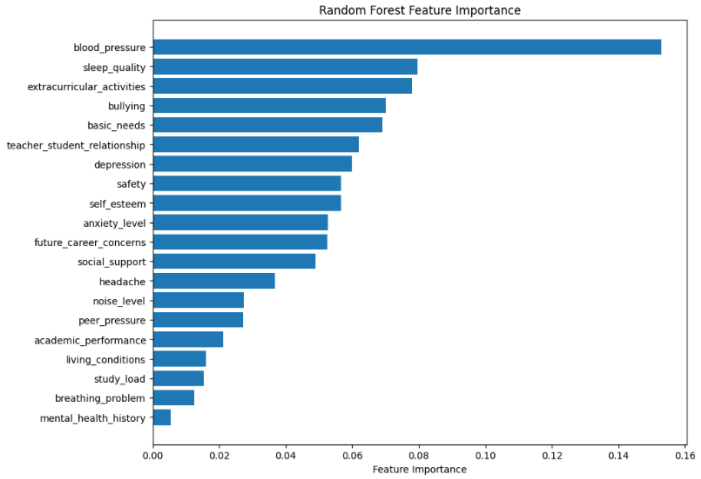

Fig 3: All factors importance at a glance (ascending) from the random forest model

Additionally, the distribution of stress levels across the total no. of students were also observed

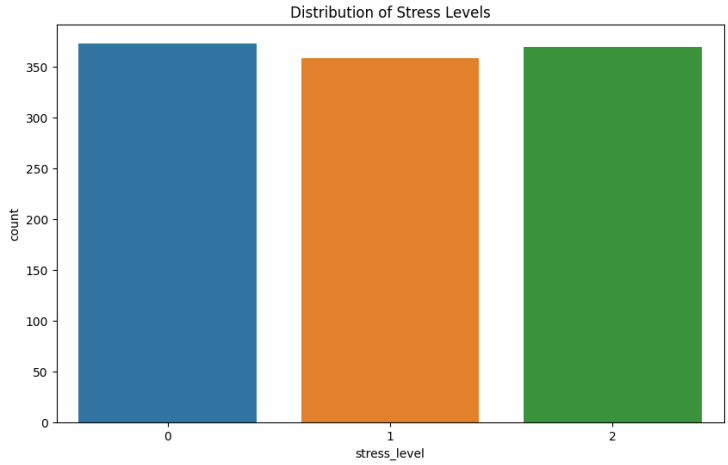

Fig 4: Stress level distribution of the total no. of students (total no. of instances)

D. ***Machine Learning model development and evaluation***

Considering the dataset is composed of numerical data, the chosen base models for the stacking ensemble learning approach were logistic regression, naïve bayes, random forest classifier & k-nearest neighbor to observe the impact on the target feature "stress_level". Some notable common observation across all the models were trained on same the percentage of training, test data and also the random state. The reason for this was to maintain consistency in the result. Some common libraries used for creating dataframes, plots, training and testing sets, classifiers, performance metrics, feature selection/reductions, absolute errors, confusion matrix, scalers, heatmaps were pandas, numpy, train_test_split, sklearn.linear_model, sklearn.ensemble, sklearn.naive_bayes, sklearn.neighbors, sklearn.metrics(accuracy_score, matthews_corrcoef, f1_score, confusion_matrix modules) and sklearn.preprocessing, seaborn repectively. Additionally, for f1 scores, the average had to be 'weighted' since the dataset has multi-class classification.

Logistic Regression Model:

The logistic regression model was created after performing the aforementioned data processing using a 'StandardScaler' object post data input into a pandas dataframe to have all the feature values within a common range. Afterwards, the dataset is divided to exclude the target feature and store its value in a separate dataframe. The data is split using train_test_split with 20% being test data and 80% training data. The iteration count for logistic regression was increased to accommodate the large dataset of 1000+. Fitting RFE to extract top 5 important features then taking absolute coefficients of the features creating a list to store them. Once predictions have been generated, the accuracy, mcc and f1 score values are calculated using the modules accuracy_score, matthews_corrcoef, f1_score and the confusion matrix is generated using confusion_matrix module. Finally, Fig 1 is generated showing the coefficients (feature importances) of the selected features. A heatmap is also generated to determine the effectiveness of the model in predicting the values (plot will be shown in the results section).

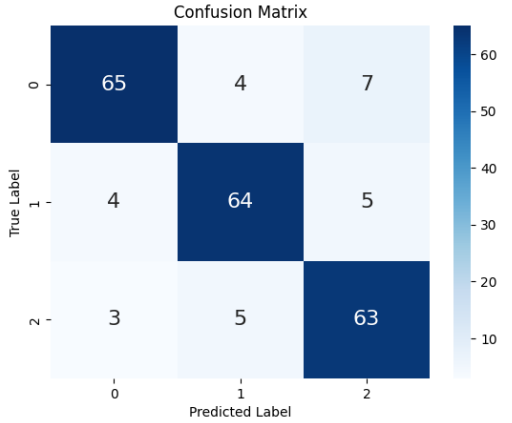

Fig 5: Heatmap for Logistic Regression Model

Random Forest Classifier:

The random forest classifier utilized the same strategy for splitting the dataset into training and test set but it set its estimators (decision trees) is set to 100 (n_estimators=100) and a random state of 42 for reproducibility. The classifier is trained on the training set using random_forest_classifier.fit(X_train, y_train). The trained model is used to make predictions on the test set (predictions = random_forest_classifier.predict(X_test)). The accuracy, mcc and f1 of the model were computed using rf_accuracy, rf_mcc and rf_f1 respectively. The feature importance of each variable is visualized using a horizontal bar chart. This provides insights into which features contribute more to the model's predictions. Once again, the confusion matrix and its corresponding heatmaps were generated using the same strategy as logistic regression.

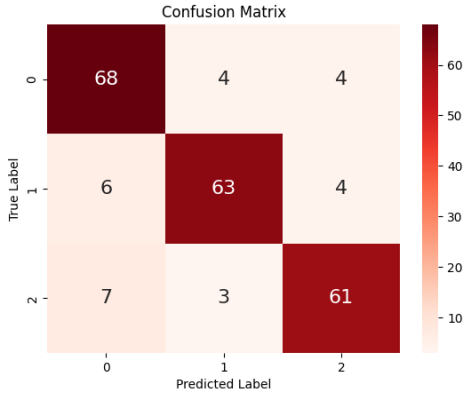

Fig 6: Heatmap for random forest classifier

Naïve Bayes:

This was the only model which had to utilized correlation matrix to eliminate the highly correlated features excluding the target feature since strong correlation can lead to worse results. The procedure followed the same strategy for getting the training and testing data. Following this, the remaining features (after filtering the highly correlated ones with a set threshold value) are saved into a separate csv file for continued usage. Gaussian Naive Bayes from the module 'sklearn.naive_bayes' to the Student Stress Factors dataset provides a comprehensive exploration of various factors influencing student stress, especially considering that the values are continuous which the gaussian naïve bayes can easily deal with. The resulting accuracies, mcc and f1 score values are found in the same way as before and then the confusion is generated and for better visualization the seaborn module is used to create the heatmap for it.

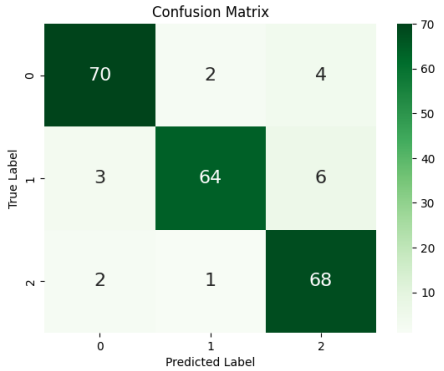

Fig 7: Heatmap of Gaussian Naïve Bayes

KNN:

Our K-NN analysis empowers us to comprehend the multidimensional nature of student stress, shedding light on the interplay of psychological, physiological, environmental, academic, and social factors through proximity-based classification.). The KNN model uses the same ratio of training and testing dataset and also the 'StandardScaler' is again used to extract the scaled features. The hyperparameter 'k' was tuned by first having all the possible k values in a dictionary and then using cross validation of 5-fold, the best value for k was determined and then was used in the training phase and later the classifier was passed onto do the predictions from which again, the accuracy, mcc and f1 scores were recorded to measure its performance later on. Confusion matrix and heatmap to show it was implemented using the same libraries as before just with some color changes for variety purposes.

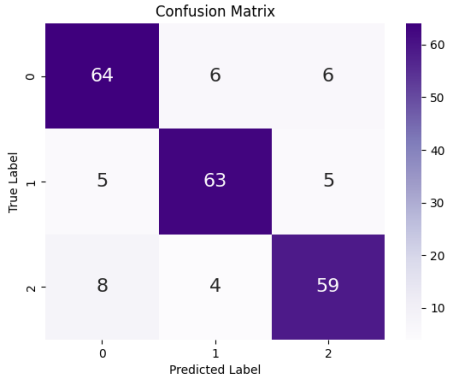

Fig 8: Heatmap of KNN

Stacking Model:

This is the meta model which utilized the previously mentioned base models and learns how to best combine their output and fit the predictions on the same dataset that we started off with. Unlike the other models, this model takes a list of predictions (generated from base models) as inputs and then uses a separate model as its final estimator model, which in this case was logistic regression since we are dealing with continuous numeric values and stacking method does in fact work well with probability values. Post training, the model's accuracy, mcc and f1 score was taken for both its training and test dataset. Just like before, a heatmap representing the confusion matrix was created using the libraries mentioned in the base models. An important point is that here, "StackingClassifier" module is used to create the stack model which is part of "sklearn.ensemble" library.

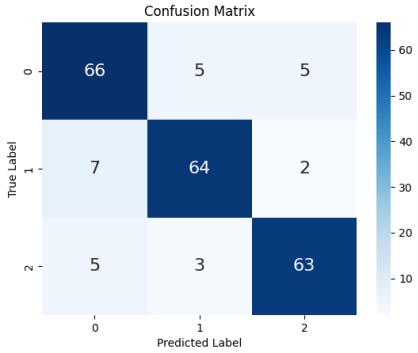

Fig 9: Heatmap for stack model

V. RESULTS

From the heatmaps for the models above, we are able to easily visualize the confusion matrix to understand the effectiveness of the models built. In the case of logistic regression model's heatmap the model was able to predict a total of 195 correctly and 25 unsuccessfully and with the values of MCC= 0.83 (high level of agreement between predicted and actual values) and F1 score of 0.89 (precise, has good recall and thus effective). Random forest classifier's heatmap, the model was able to predict a total of 186 correctly and 34 unsuccessfully and with the values of MCC= 0.83 (high level of agreement between predicted and actual values) and F1 score of 0.89 (precise, has good recall and thus effective). Gaussian naïve bayes heatmap, the model was able to predict a total of 202 correctly and 18 unsuccessfully and with the values of MCC= 0.88 (relatively higher level of agreement between predicted and actual values) and F1 score of 0.91 (high precision with has good recall and relatively more effective than the models). KNN's heatmap, the model was able to predict a total of 186 correctly and 34 unsuccessfully and with the values of MCC= 0.79 (relatively lower level of agreement between predicted and actual values but still adequate) and F1 score of 0.85 (relatively less precise than the rest of the models so recall is satisfactory and effectiveness maybe not be as good as the rest)

Finally, the stack model's heatmap shows that it was able to predict a total of 193 correctly and 27 unsuccessfully and with the values of MCC= 1.0 (highest level of agreement between predicted and actual values) and F1 score of 1.0 (highest precision possible with great recall and most effective).

	Accuracy	MCC	F1	Precision	Recall
lrm	0.89	0.83	0.89	0.89	0.89
rf	0.87	0.81	0.87	0.87	0.87
nb	0.92	0.88	0.92	0.92	0.92
knn	0.85	0.77	0.85	0.92	0.92
stack	1.00	0.82	0.88	0.88	0.88

Fig 10: Tabular view of accuracy, mcc, f1 scores, precision and recall values of all models

Below is a plot of the performance metrics (prediction accuracy, MCC and F1 scores) considered of all the models, including the stack model.

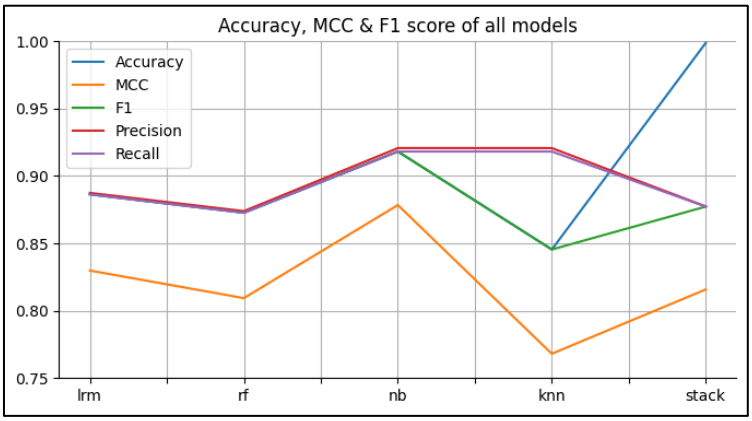

Fig 11: A line graph showing the performance metrics of different models on the dataset

To better understand threshold values are important when it comes to correctly classify the different stress levels of students, a ROC (Receiver Operating Characteristic) curve was plotted along with that the AUC values as part of the legend. Additionally, it also helps in understanding the trade-off between sensitivity (True Positive Rate) and specificity (False Positive Rate).

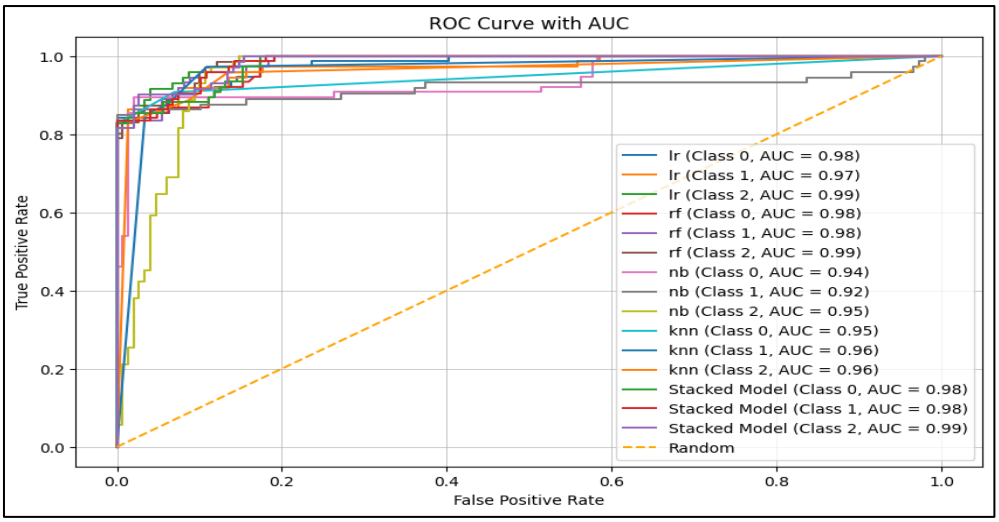

Fig 12: ROC curve of all stress level classes along with their AUC values

VI. REFERENCES

[1] HTTPS://WWW.KAGGLE.COM/DATASETS/RXNACH/STUDENT-STRESS-FACTORS-A-COMPREHENSIVE-ANALYSIS

[2] https://www.w3schools.com/python/matplotlib_grid.asp

[3] https://colab.research.google.com/drive/1npyKNNQ_fjvzEjVUQOT7_ayXc-0RDOUJ#scrollTo=yFjCiiK5mhZX

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

[5] https://scikit-learn.org/stable/modules/preprocessing.html

[6] https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression