ECS 171 Group 10 Project

Leader: Ryan Yu

Amira Basyouni, Calvin Chen, Alexis Lydon, Tianming Tan

Github Repository: `https://github.com/ryyu444/ECS-171-Group-Project`

November 28, 2023

Introduction and Background

As a student, we constantly grapple with numerous stressors while striving for academic success. Deadlines, exams, and the intrinsic self-asserted pressure to excel, they all contribute to the challenges we face. These factors can have a negative impact on our quality of learning and thus hinder our progress toward acquiring a degree to ultimately enter the workforce. In contrast, sleep is an activity that we all partake in and seems relaxing. So we posed a question: does sleep have any connections to stress? And if so, how are they related? Our goal is to construct predictions regarding the effects that sleep quality and sleep duration can have on our stress levels. By analyzing our data, we will be able to determine the correlation between stress and those two sleep-related attributes. It's worth noting that our data includes individuals of ages 27 to 59. Although the range does not fall within the typical college student age, our results can still provide valuable insights and promote awareness that can benefit us for years to come.

In this machine learning project centered on the intersection of sleep health and stress, our primary objective is to examine the correlation between key attributes such as sleep duration and sleep quality, and their influence on stress levels. This analysis will help us better understand and predict the dynamics between stress and other factors. Thus, it will help identify the attributes that have the most significant impact on stress levels. As an integral part of the project, we will develop accurate models for stress prediction based on the highly correlated attributes in our dataset.

Literature Review

Machine Learning has great capabilities within the realm of pattern recognition and categorization; two fields that we hope to utilize as we explore the connection between sleep and stress. It is important to address some of the pre-existing studies which were conducted with a similar objective. Presented below are examples of such research and methodologies as well as their overall findings.

Jayawickrama and Rupasingha, researchers from Sabaragamuwa University of Sri Lanka, conducted a study to examine the impact of sleep habits on stress levels.(2) They applied various machine learning models which included Naïve Bayes, Random Forest, Decision Trees, Multi-layer Perceptrons (MLPs), Support Vector Machines (SVMs), and Logistic Regression mixed with cross-validation. It was revealed that five of the six models achieved accuracy rates exceeding 80%, suggesting that there is a strong correlation between sleep and stress. Notably, Naïve Bayes was the best predictor with an accuracy of 91.27%.

Minhazur Rahman and a different group of researchers predicted stress levels using data collected from sleeping participants.(1) Some of the models they utilized include Gradient Boosting, Decision Trees, Random Forest, Gaussian Naive Bayes, and Linear Support Vector Machine. Their models achieved an accuracy rate of over 95%, with Naive Bayes and SVM as the top performers.

Overall, these studies highlight the efficacy of machine learning models in predicting stress levels based on sleep data, with Naive Bayes and SVM demonstrating high performance. We hope to utilize these research metrics as the basis of judgment for the models that we will be building and testing.

Dataset Description and Exploratory Data Analysis

The dataset being used in this project is titled "Sleep Health and Lifestyle Dataset" and was made by Laksika Tharmalingam, accessible on Kaggle. The dataset comes in the form of a CSV file and contains 400 rows and 13 columns, consisting of various sleep and lifestyle variables (e.g. gender, age, occupation, sleep duration, sleep quality, and stress levels). These variables encompass the very aspects of sleep and health that can be useful in predicting stress. Despite the abundant source of data, a limitation of this dataset is its synthetic nature, generated artificially rather than from real-world observations. Although it may lack certain real-world nuances, high-quality synthetic data proves effective in emulating real datasets. It serves as a cost-efficient and time-saving approach in comparison to gathering real-world observations to use for training and testing machine learning models.

Initially, we conducted essential data preprocessing to convert all categorical data into numerical values, enhancing their compatibility with prediction models. This involved applying label encoding to categorical attributes from the dataset like BMI and Gender. We opted for label encoding over one-hot encoding to both maintain consistency in the number of columns in our dataset and to enhance clarity in visualizing a pair plot and correlation matrix.

Following data preprocessing, we generated a pair plot matrix (reference to figure) to investigate any potential linearly separable relationships between the attributes and Stress Level. The pair plot revealed notable linear associations, particularly between Sleep Quality and both Sleep Duration and Age in predicting Stress Level. Additionally, Sleep Duration and Gender also have somewhat of a linear relationship in predicting stress.

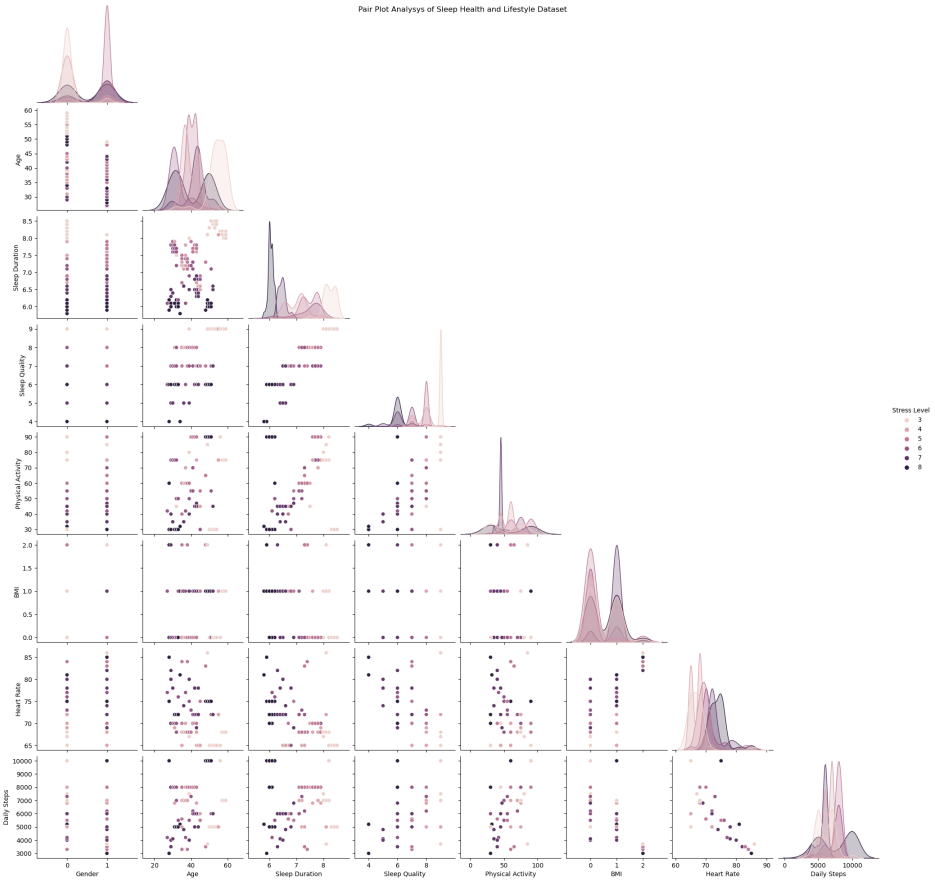

Figure 1: Pair Plot Matrix

Subsequently, a correlation matrix (reference to figure) was constructed to examine high correlations among attributes, with a special focus on their connection to Stress Level as the dependent variable in our models. The correlation matrix reinforced the findings from the pair plot, highlighting strong negative correlations of -0.9 between Sleep Quality and Stress Level, and -0.81 between Stress Level and Sleep Duration.

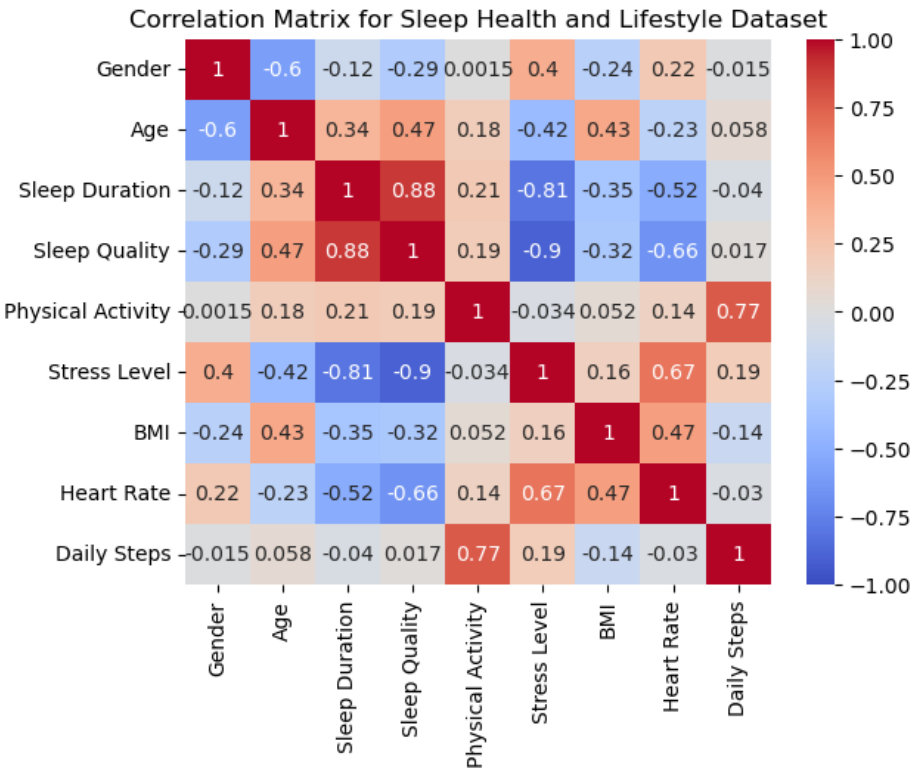

Figure 2: Correlation Matrix

From these insights, it became evident that Sleep Quality and Sleep Duration were pivotal predictors of stress in our models. The potential synergistic predictive power of using these two attributes together became apparent and, for that reason, we opted to use Sleep Quality and Sleep Duration as our predictors.

We also performed some outlier detection on these important variables using box plot models. We found no significant outliers neither in the predictor variables nor in the dependent variable Stress Level. Although Stress Level could be a value from 1 to 10, the only outcome values available from this dataset were between 3 and 8.

Proposed Methodology

Given our objective to predict stress levels across six categories (levels 3-8), we are dealing with a multi-class classification problem. Upon reviewing the pair plots (reference to figure), certain graphs, such as Sleep Quality vs Sleep Duration, Sleep Duration vs Gender, and Sleep Duration vs Age indicated optimal separability in the data.

Managing the numerous levels of our outcome variable, Stress Level, made it difficult to discern a clear distinction between our labels. To simplify the problem, we contemplated grouping levels (e.g. combining 3 and 4, 5 and 6, 7 and 8 to create three stress levels instead of six). However, we ultimately decided against grouping to avoid potential oversimplification and instead retained the

original range of 3-8.

After having finalized our predictors and prediction values, we plan to construct four different models using Linear Regression (LR), Multinomial Logistic Regression, a Naive Bayes Classifier (NB), and a Linear Support Vector Machine (SVM). We will then proceed to split the preprocessed data with an 80-20 train-test ratio and train them accordingly using cross-validation for LR and MLR. Cross-validation can pose computational challenges, particularly with an increasing number of folds because the process may become time-consuming, especially if the model is intricate or computational resources are constrained.

Afterward, to evaluate the models, we will be using MSE and R^2 for LR and a classification report consisting of Precision, Recall, F1-Score, and Accuracy for the other three models. These metrics serve as indicators to evaluate the quality and performance of our models. With these values in mind, we will ultimately decide which model is the best predictive model for our dataset and compare our findings with those from previous research. Lastly, we will build a frontend to display our model predictions based on the provided user inputs.

Experimental Results

Linear Regression:

```
Results for Linear Regression Model:

Mean Squared Error Train: [0.36320511 0.25092829 0.64591507 0.68595341 0.731697   0.52184439
 0.48877131 0.88277543 0.91570423 0.85754598 0.45701947 0.71880336
 0.47552581 0.61161942 0.35861328]

R^2 Train: [0.88081867 0.93643362 0.77936291 0.78067037 0.6744396  0.76781117
 0.85995091 0.62071947 0.74895018 0.74703658 0.83010429 0.80293259
 0.78286493 0.80568088 0.84514426]

Average Mean Squared Error Train: 0.5977281049731364
Average R^2 Train: 0.7908613628260651

Mean Squared Error Test: [1.29807995 0.44473663 0.97142352 0.42314097 0.30259433 0.40275331
 0.51707457 1.16017129 1.8097533  0.70787939 0.24351678 0.05034095
 0.29989956 0.7349889  1.19452387]

R^2 Test: [ 0.29452177  0.8610198   0.80725724  0.68886694  0.91203653  0.92485946
  0.78455226  0.60805024 -6.54063874  0.73186387  0.88726075  0.97752636
  0.92023948  0.7812533  -3.97718278]

Average Mean Squared Error Test: 0.7040584873984371
Average R^2 Test: -0.0225675681268673
```

Figure 3: Classification Report of Linear Regression

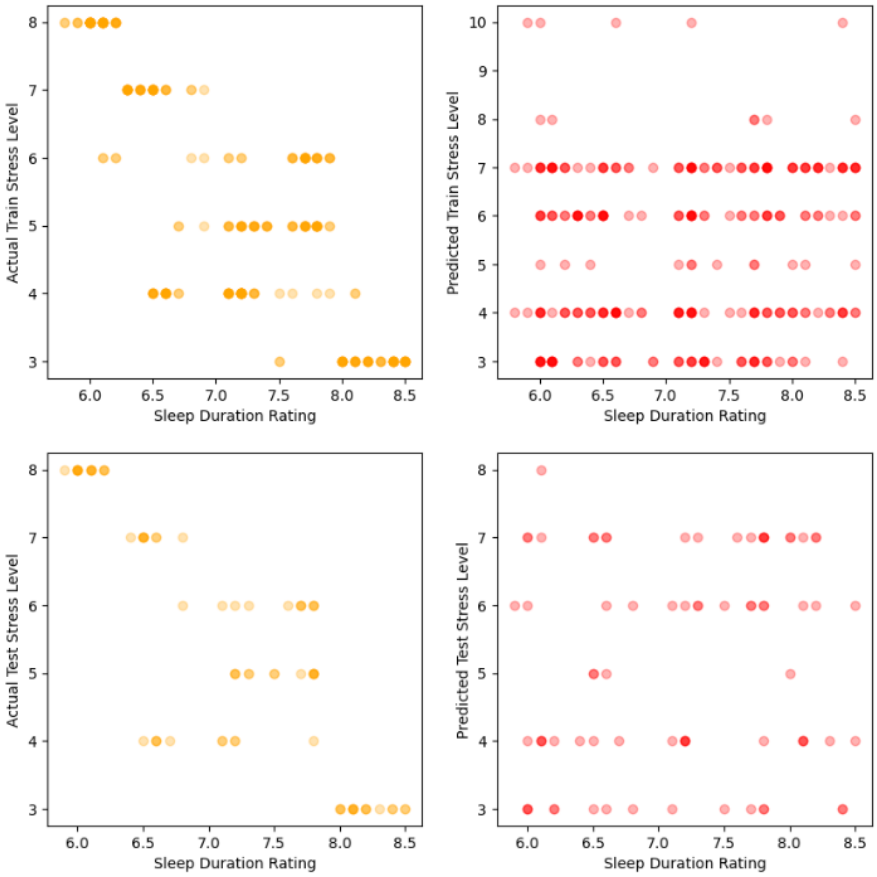

Figure 4: Scatterplot of Linear Regression

The model showed decent training performance but struggled to generalize to the test data, indicating potential overfitting. Further analysis and alternative modeling approaches may be needed for improvement.

In the training data, MSE values varied between 0.2509 and 0.9157, averaging at 0.5977. This implies an average deviation of approximately 0.598 units between the model's predictions and the actual values. The R^2 values for the training data ranged from 0.6207 to 0.9364, with an average of 0.7909. This indicated that the model accounts for about 79.1% of the variance in the target variable. Shifting to the test data, MSE values ranged from 0.0503 to 1.8097, averaging at 0.7041. This slightly exceeds the training data average, suggesting potential overfitting. The R^2 values for the test data spanned from -6.5406 to 0.9775, with an average of -0.0226, highlighting the limited explanatory power of the model for the test data.

In summary, the model showed decent training performance but struggled to generalize to the test data, indicating potential overfitting. Further analysis and alternative modeling approaches may be needed for improvement.

Multinomial Logistic Regression:

MLR Training Classification Report:

	precision	recall	f1-score	support
3	0.96	0.96	0.96	52
4	0.62	0.88	0.73	56
5	0.80	0.52	0.63	54
6	0.85	0.78	0.81	36
7	1.00	0.90	0.95	40
8	0.95	1.00	0.98	61
accuracy			0.84	299
macro avg	0.86	0.84	0.84	299
weighted avg	0.86	0.84	0.84	299

Accuracy: 0.842809364548495

MLR Testing Classification Report:

	precision	recall	f1-score	support
3	1.00	1.00	1.00	19
4	0.67	1.00	0.80	14
5	1.00	0.77	0.87	13
6	1.00	0.80	0.89	10
7	1.00	0.70	0.82	10
8	0.90	1.00	0.95	9
...				
macro avg	0.93	0.88	0.89	75
weighted avg	0.93	0.89	0.90	75

Accuracy: 0.8933333333333333

Figure 5: Classification Report of Multinomial Logistic Regression

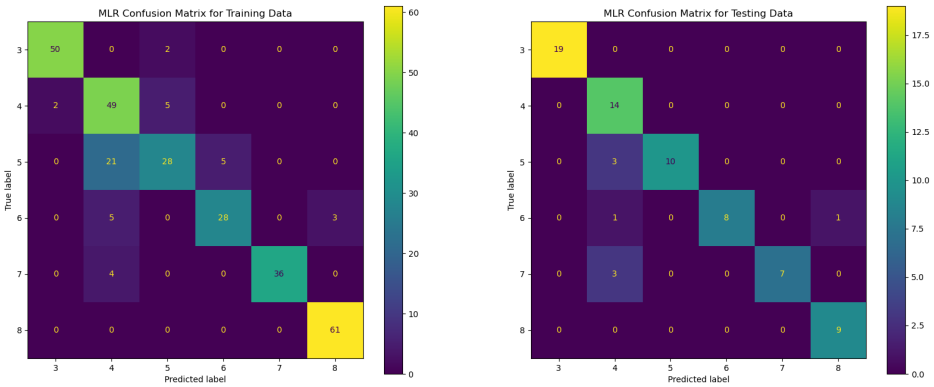

Figure 6: Confusion Matrix of Multinomial Logistic Regression

The Logistic Regression model, built using sklearn's LogisticRegressionCV function, demonstrated strong performance across most classes in training and test sets. From the classification report, classes 3, 7, and 8 have high precision, recall, and f1-score in both sets indicating the precise classification of these classes. However, class 4 showed a lower precision than the other classes in both sets, but a higher recall indicates a higher likelihood of false positives.

The overall accuracy of the training set is approximately 84.28%, while the test set accuracy is 89.33%. However, the lower precision of class 4 impacts the overall accuracy of both sets, so there is room for improvement. These high accuracy scores demonstrate the model's ability to correctly classify class labels for a significant portion of the instances in both seen and unseen data.

Naive Bayes:

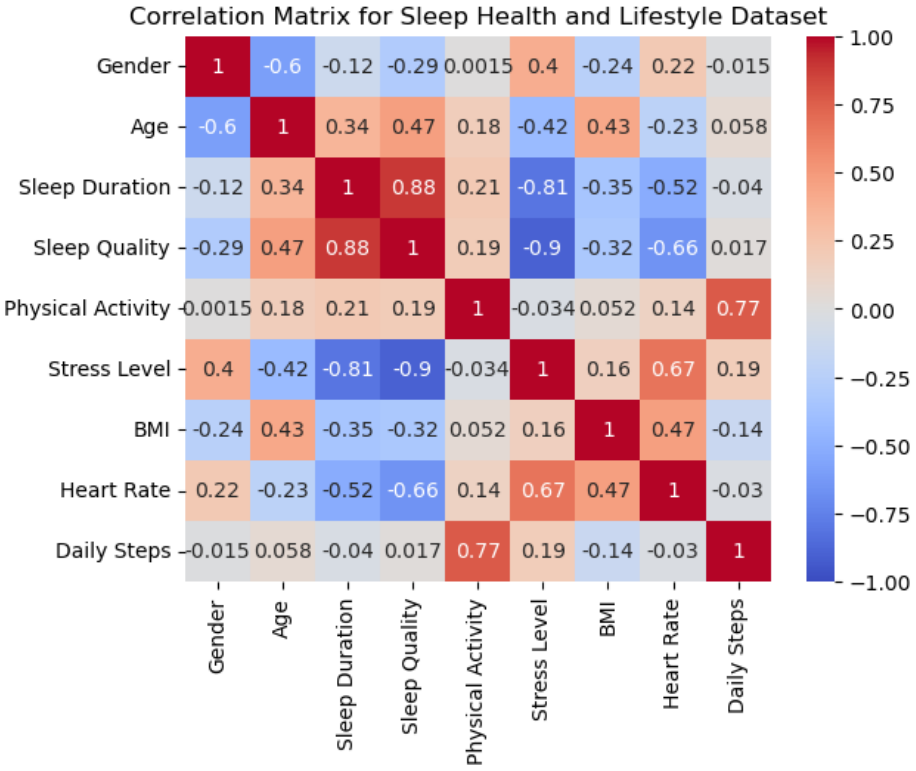

Figure 7: Correlation Matrix

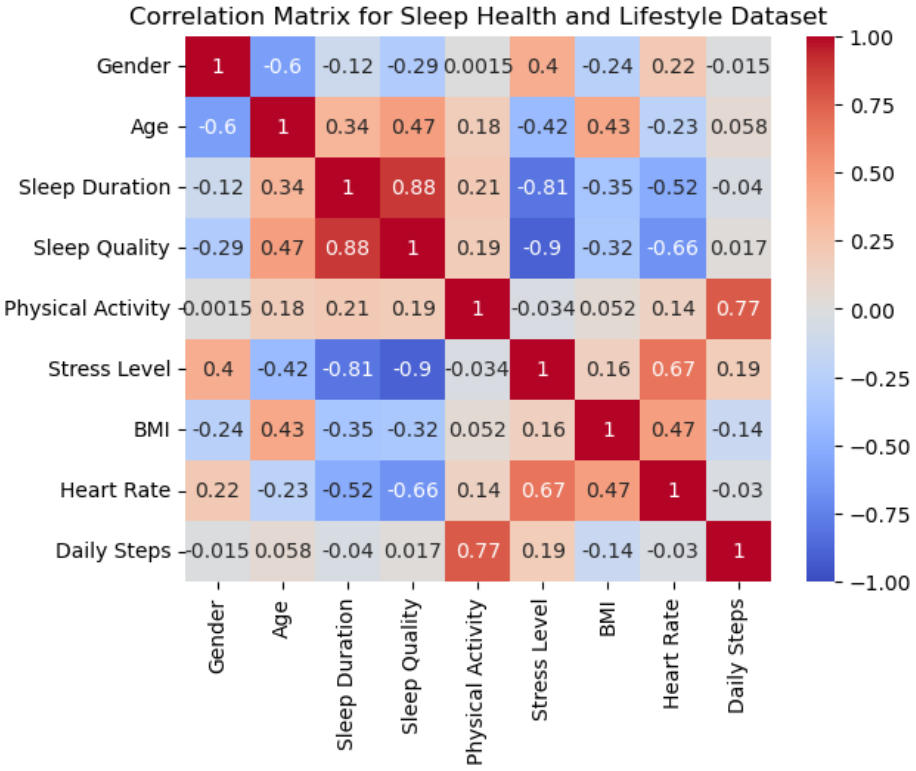

Figure 8: Correlation Matrix

The Naive Bayes model, built using sklearn's GaussianNB function demonstrated mixed performance across classes in training and test datasets. From the classification report, classes 3 and 8 have high precision, recall, and f1-score that is all above 0.9 in both sets. This indicates strong and precise classification of these classes. However, classes 4 and 5 showed a lower precision, with class 4 also having low recall, indicating inaccurate classification for these classes.

The overall accuracy of the training set is approximately 80.60%, while the test set accuracy is 82.67%. This suggests a relatively better performance when classifying unseen data. There is room for enhancing the model's performance, particularly in accurately classifying instances for classes 4 and 5.

Linear Support Vector Machine:

SVM Classification Train Report:

	precision	recall	f1-score	support
3	0.96	0.96	0.96	56
4	0.61	0.89	0.72	57
5	0.84	0.56	0.67	55
6	0.91	0.74	0.82	39
7	1.00	0.84	0.91	38
8	0.93	1.00	0.96	54
accuracy			0.84	299
macro avg	0.87	0.83	0.84	299
weighted avg	0.86	0.84	0.84	299

Accuracy Train: 0.8394648829431438

SVM Classification Test Report:

	precision	recall	f1-score	support
3	1.00	1.00	1.00	15
4	0.67	0.92	0.77	13
5	0.88	0.58	0.70	12
6	0.86	0.86	0.86	7
7	1.00	0.92	0.96	12
8	1.00	1.00	1.00	16
...				
macro avg	0.90	0.88	0.88	75
weighted avg	0.91	0.89	0.89	75

Accuracy Test: 0.8933333333333333

Figure 9: Classification Report of Linear Support Vector Machine

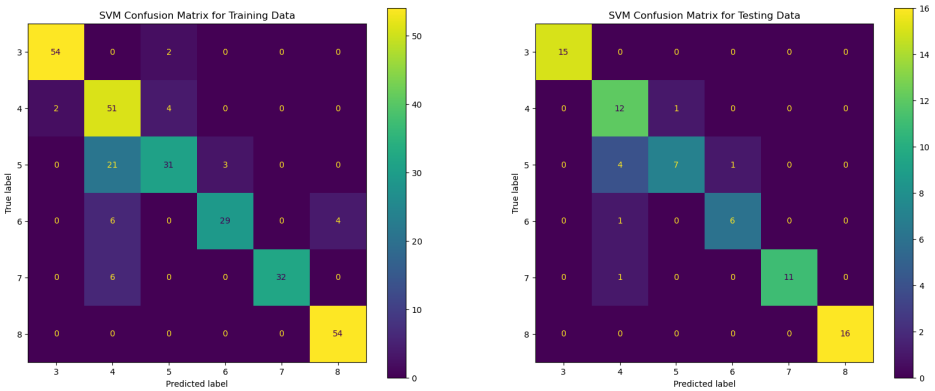

Figure 10: Confusion Matrix of Linear Support Vector Machine

We utilized sklearn's SVC function to implement the Support Vector Machine (SVM) for our analysis. The SVM model employs a Radial Basis Function (RBF) kernel.

In the training set, the model demonstrated a high precision of 0.97 for class 3, with commendable recall and F1-score, also at 0.97. Conversely, class 4 exhibited a lower precision of 0.64, coupled with a high recall of 0.91. The overall training set accuracy was 83.95%. In the test set, the model maintained strong precision, achieving a perfect 1.00 for classes 3 and 8. However, class 4 had lower precision at 0.67, with a high recall of 0.92, echoing the training set trend. The overall test set accuracy was 89.33%.

In summary, the SVM model displayed commendable performance in classifying the dataset, with elevated precision, recall, and F1-scores across most classes. There is still room for improvement in the model's performance for class 4. Despite that, the model's consistent accuracy on both the training and test sets suggests its effectiveness and ability to generalize well to unseen data.

Software Implementation:

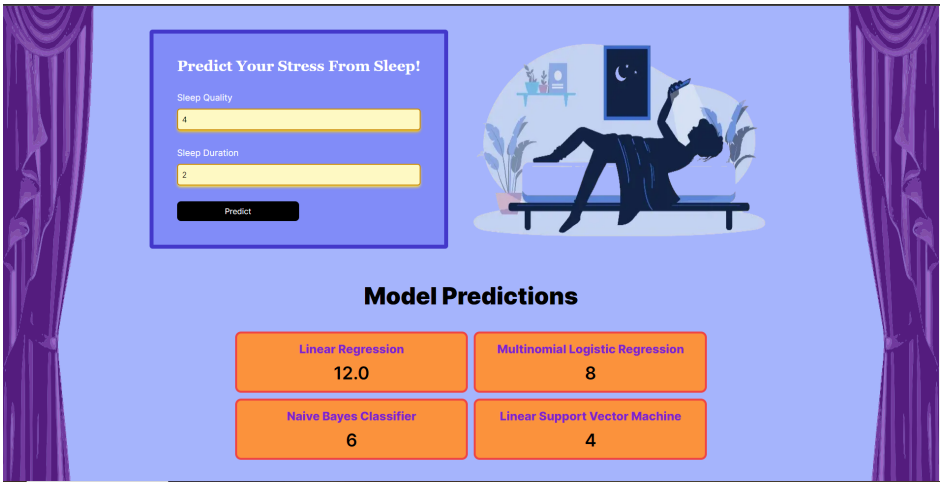

Figure 11: Demo of Frontend

For software implementation, we utilized Next.js and Flask to construct our web page for testing and running our model. The component receives two inputs, namely sleep quality and sleep duration. It subsequently initiates a POST request to the '/api/predict' API endpoint with these inputs. The API then employs four distinct machine learning models (i.e. Linear Regression, Multinomial Logistic Regression, Naive Bayes, and Linear Support Vector Machine) to make predictions. These models, having been pre-trained, are loaded from a pickle file. Following the prediction process, the results are returned to the front end as a JSON response and displayed.

Conclusion and Discussion

Through the use of "Sleep Health and Lifestyle Dataset" that contains various sleep and lifestyle variables, and the use of four different models (i.e. Linear Regression, Logistic Regression, Naive Bayes, and Support Vector Machine), we were able to form multiple models that predict Stress Level. Despite our initial hopes for a strong predictive Linear Regression model, the high MSE and R2 values brought it out of the race. On the other hand, our Multinomial Logistic Regression,

Naive Bayes, and SVM models were able to achieve over 80% accuracy. In terms of stress levels, these three were strongest at predicting levels 3, 6, 7, and 8, but performed much poorer with the remaining values. Overall, our best predictive models consist of Multinomial Logistic Regression and SVM, which corroborates with what prior research found in our literature review. In stark contrast, however, we found that Naive Bayes was the worst out of the three contenders in comparison to external studies that concluded with Naive Bayes having the best predictive capabilities amongst all of their models. This can be attributed to many factors (e.g. our way of categorization, training and testing, etc).

For the future of stress prediction with machine learning, it is recommended that more robust models, along with a larger training set (possibly from real data), are used for practicality and to get less ambiguous trends. Furthermore, creating and testing other types of models, such as Decision Trees or Multilayer Perceptrons (MLPs), may yield slightly more accurate predictions than the models developed in this project. Regardless of what the future might bring, we take pride in what we've accomplished over the span of 10 weeks, transitioning from a collection of ideas to creating predictive models with accuracies exceeding 80%, and ultimately developing a fully functioning frontend to display their capabilities.

References

References

[1] M. M. Rahman, A. Mohaimenul Islam, J. Miah, S. Ahmad, and M. Mamun, "Sleepwell:stress level prediction through sleep data. are you stressed?," 2023 IEEE World AI IoT Congress (AIIoT), 2023. doi:10.1109/aiiot58121.2023.10174306.

[2] S. Jayawickrama and S. Rupasingha, "Predicting stress levels using sleep data," 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 2019. doi:10.1109/bibe.2019.000-9.

[3] L. Tharmalingam, "Sleep Health and Lifestyle Dataset," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/lahiripremarathne/sleep-health-and-lifestyle-dataset. [Accessed: 10-Jun-2021].