REPORT/WHITE PAPER

STRESS LEVEL PREDICTIVE MODEL

Joseph Choi | Jenny Overby | Daniel Meier | Serge Nane

DSC450 Applied Data Science
Spring 2024

TABLE OF CONTENTS

INTRODUCTION	3
BUSINESS PROBLEM	4
METHODS/ANALYSIS	4
RESULTS	8
RECOMMENDATIONS	9
CONCLUSION	9
REFERENCES	10
APPENDIX	11

INTRODUCTION

In today's educational landscape, mental health is a major concern, especially for students. As mental health issues continue to rise, it has become evident that managing stress levels is essential as it can significantly impact one's life and lead to negative outcomes if not handled properly. With students facing constant pressure throughout their day-to-day, it is imperative that we fully understand the factors influencing their stress levels and to what degree. Therefore, our team has decided to analyze the psychological, physiological, social, environmental, and academic stressors faced by students to ultimately develop a predictive model that can forecast stress levels for our third and final project.

To promote our initiatives, we used a comprehensive dataset from Kaggle containing a wide range of factors influencing student stress levels to conduct our analysis and build out our model. The Stress Level dataset consists of 20 influential features categorized into five groups. It is worth noting that all features in this dataset are quantified using standardized scales or categorical ratings.

Here are the descriptions of each feature, including details about the scale and ratings used:

- **PSYCHOLOGICAL FACTORS**:
  - **anxiety_level**: Scale from 0 to 21 based on the GAD-7 scale
  - **self_esteem**: Scale from 0 to 30 using the Rosenberg Self-Esteem Scale
  - **mental_health_history**: Presence (1) or absence (0) of mental health history
  - **depression**: Level from 0 to 27, based on the Patient Health Questionnaire (PHQ-9)
- **PHYSIOLOGICAL FACTORS**:
  - **headache**: Rated from 0 (none) to 5 (high)
  - **blood_pressure**: Levels categorized from 0 (low) to 5 (high)
  - **sleep_quality**: Rated from 0 (poor) to 5 (excellent)
  - **breathing_problem**: Rated from 0 (none) to 5 (severe)
- **ENVIRONMENTAL FACTORS**:
  - **noise_level**: Rated from 0 (quiet) to 5 (loud)
  - **living_conditions**: Rated from 0 (poor) to 5 (excellent)
  - **safety**: Rated from 0 (unsafe) to 5 (very safe)
  - **basic_needs**: Rated from 0 (unmet) to 5 (fully met)
- **ACADEMIC FACTORS**:
  - **academic_performance**: Rated from 0 (poor) to 5 (excellent)
  - **study_load**: Rated from 0 (light) to 5 (heavy)
  - **teacher_student_relationship**: Rated from 0 (poor) to 5 (excellent)
  - **future_career_concerns**: Rated from 0 (none) to 5 (extreme)
- **SOCIAL FACTORS**:
  - **social_support**: Rated from 0 (none) to 5 (strong)
  - **peer_pressure**: Rated from 0 (none) to 5 (high)
  - **extracurricular_activities**: Rated from 0 (none) to 5 (high)
  - **bullying**: Rated from 0 (none) to 5 (frequent)
- **stress_level**: Rated from 0 (none) to 2 (extreme)

Link: Student Stress Factors: A Comprehensive Analysis (kaggle.com)

BUSINESS PROBLEM

The main focus of our project revolves around two fundamental research questions:

1. What are the key factors influencing stress levels among students?
2. Can we build a predictive model that can accurately predict stress levels based on these factors?

Throughout the project, we made sure to keep our focus on our business problems, ensuring that our data wrangling, science, and visualization tasks stayed aligned with these objectives to avoid going off track.

The insights gathered from this project will be key to maintaining students' well-being and performance in school. That is why our team would like to research this topic and determine how best to support students and implement effective stress management strategies.

METHODS/ANALYSIS

For this project, our team split the project into three parts. Here are the descriptions for each segment, outlining the methods used and the insights gathered from our analysis:

1. DATA WRANGLING

For this phase of the workflow, we started by loading the dataset into a DataFrame and performing an initial EDA to assess the quality of our data. From the assessment, we found that our dataset contains 1,100 entries and 21 features, three of which are categorical and the rest numerical. This output was alarming as all features within our dataset should be numerical. Therefore, we converted object-type features into float by removing symbols. This process was applied to three features: anxiety level, self-esteem, and depression.

We also checked for missing values in the dataset and confirmed that there were none. In addition, we also found that most features showed a balanced distribution around the midpoint of their respective ranges. This was indicated by comparing the mean of each feature with its range to understand central tendency and variability.

2. DATA SCIENCE

After preparing the Stress Level dataset for further analysis and model training, we moved on to the data science phase of the project, which consisted of three parts: EDA, Model Selection, and Model Training.

EXPLORATORY DATA ANALYSIS (EDA)

During the EDA phase, our goal was to fully understand the trends and patterns of each input feature and the target variable. We focused on extracting insights to inform and guide our approach to feature selection, feature engineering, and model selection.

The first task we completed during the EDA phase was to build histograms to understand the distribution of the data. Once the histograms were plotted, we labeled each feature as either **Evenly Distributed** or **Skewed** (Fig. 1). We did this to categorize our features as evenly distributed features are beneficial for predictive modeling as they provide diverse input data, while skewed features will require normalization to ensure the model treats these features without bias.

Features	Category	Remark
Anxiety Level	Evenly Distributed	Level is even throughout students
Self-Esteem	Skewed	Most students rated high
Mental Health History	Evenly Distributed	Level is even throughout students
Depression	Evenly Distributed	Level is even throughout students
Headache	Skewed	Most students rated either low or high
Blood Pressure	Skewed	Most students rated high
Sleep Quality	Skewed	Most students rated either high or low
Breathing Problem	Skewed	Most students rated either high or medium
Noise Level	Evenly Distributed	Level is even throughout students
Living Conditions	Evenly Distributed	Level is even throughout students
Safety	Skewed	Most students rated either high or medium
Basic Needs	Skewed	Most students rated either high or medium
Academic Performance	Skewed	Most students rated either high or medium
Study Load	Evenly Distributed	Level is even throughout students
Teacher-Student Relationship	Skewed	Most students rated either high or medium
Future Career Concerns	Skewed	Most students rated either high or medium
Social Support	Skewed	Most students rated either high or low
Peer Pressure	Skewed	Most students rated either high or medium
Extracurricular Activities	Skewed	Most students rated either high or medium
Bullying	Skewed	Most students rated either high or low
Stress Level	Evenly Distributed	Level is even throughout students

Fig. 1: Histogram Categories by Features

After plotting and reviewing our histograms, we built a correlation matrix to better comprehend the relationships between the input features and stress levels among students (Fig. 2). It revealed which features have the strongest positive and negative correlation with stress levels. From the visual, we concluded that bullying, future career concerns, and anxiety levels have high correlations with stress levels, indicating that as these factors increase, so do the stress levels. On the other hand, features like self-esteem, sleep quality, and academic performance show strong negative correlations, indicating that higher values in these areas bring out lower stress levels. Reviewing the matrix, we observed that all features have some level of correlation with the stress levels of students. This suggests that each feature contributes valuable information to the learning process. Therefore, we are leaning towards retaining all features in our model for a more accurate prediction.

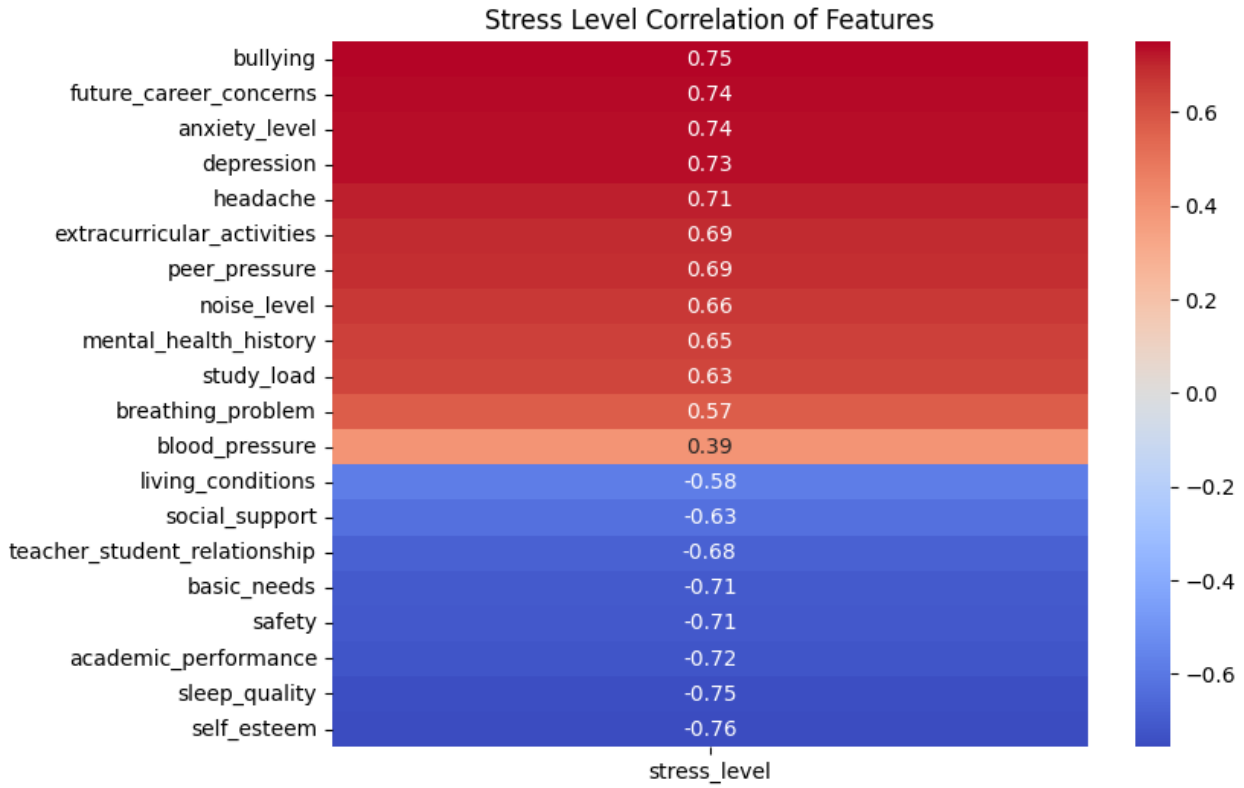

Fig. 2: Correlation Matrix

MODEL SELECTION

For our model selection process, we took the time to carefully analyze our dataset to select the most suitable machine learning algorithms for our model training process. As mentioned in the Introduction section, our dataset consists of features with standardized scales and categorical ratings (all numerical data types). Keeping this in mind, we decided to use the Random Forest Classifier and K-Nearest Neighbors (KNN).

We chose the Random Forest Classifier because it excels at capturing non-linear relationships between input and target variables. Since our dataset contains 20 influential features grouped into five categories, we figured this model would help capture this complex pattern throughout the learning process. In addition, we chose to use a Classifier rather than a Regressor as our target variable is categorical, even though the categories are numerically encoded. We selected KNN because it performs well with numerical data. Since our features are already quantified using scales and ratings, we figured KNN would efficiently calculate distances between data points.

MODEL TRAINING

During the model training phase, we experimented with two models: Random Forest Classifier and KNN. For each model, our predefined plan for each iteration was as follows:

1. We trained the model, including all 20 features (unscaled). This helped us understand the models' baseline performance in their raw unadjusted scaling form.
2. We used Grid Search to adjust the hyperparameters based on the performance from the first iteration. This helped us optimize the models with unscaled data.

3. We normalized all 20 features and retrained the models. This was done as algorithms like KNN and even Random Forest, to some extent, perform better when features are on a similar scale.
4. We used Grid Search to adjust the hyperparameters based on the performance from the third iteration. This helped us optimize the models with normalized data.
5. We adjusted the training and testing set split ratios (70/30, 80/20, 90/10) to see if we can further improve our model's performance.

We documented and evaluated the performance for each iteration and selected the best-performing model as our final choice. Please refer to the appendix for more detailed documentation of the model training for each iteration.

3. DATA VISUALIZATION

During the data science phase, our team also built additional visuals to extract deeper insights into our Stress Level dataset.

One of the visuals we created was a set of correlation matrices (Fig. 3) that show how different groups of factors (psychological, physiological, environmental, academic, and social) correlate with student stress levels. Unlike the original correlation matrix (Fig. 2) that included all 20 features together, these focused matrices allowed us to drill down on the relationships within each group more clearly. As expected, the order of correlation within each group matched the patterns we saw in the original matrix. This detailed view helped us understand how each factor group influences stress levels in students. Based on the color scaling of the matrices, we concluded that psychological factors have the strongest influence on stress levels, while environmental factors are the least influential.

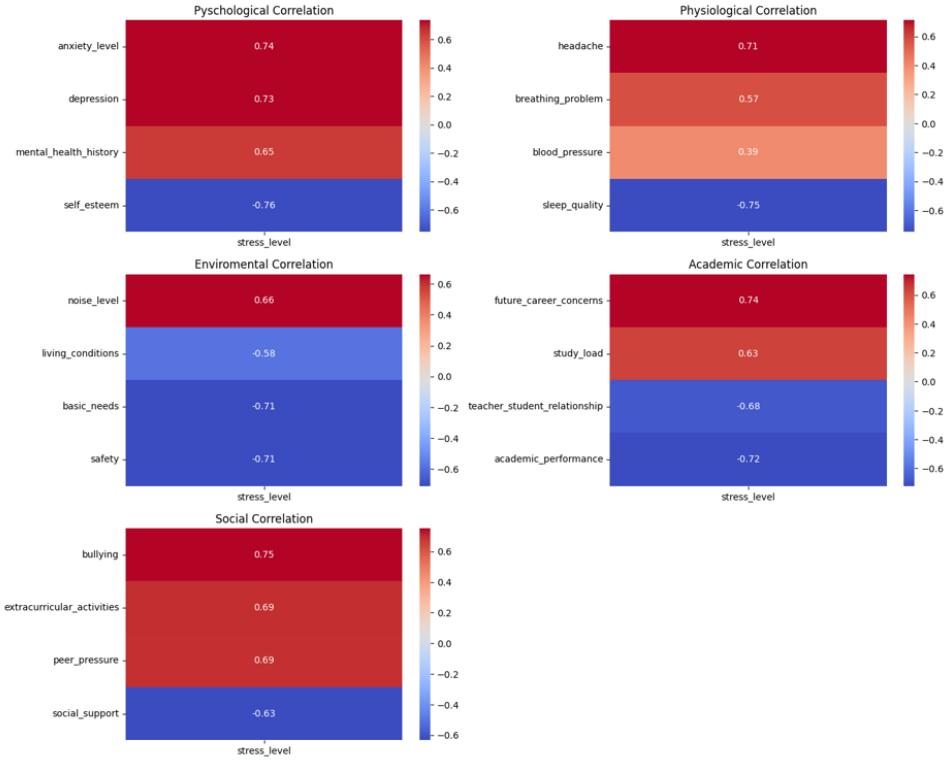

Fig. 3: Correlation Matrices by Factor

In this report, we also want to display the return of feature importance as determined by our best-performing model. Our previous visuals mainly focused on initial correlations identified during the EDA phase before we trained the model. We thought it would be insightful to see which features the model identified as most important and compare these with our initial analysis. Based on the plot, we found that academic performance and depression are the top predictors of stress levels according to the model. While depression aligns with its strong positive correlation observed initially in Fig. 2, academic performance shows significant model importance despite its negative correlation with stress. This suggests the complex interactions among the features and their impact on predicting student levels.

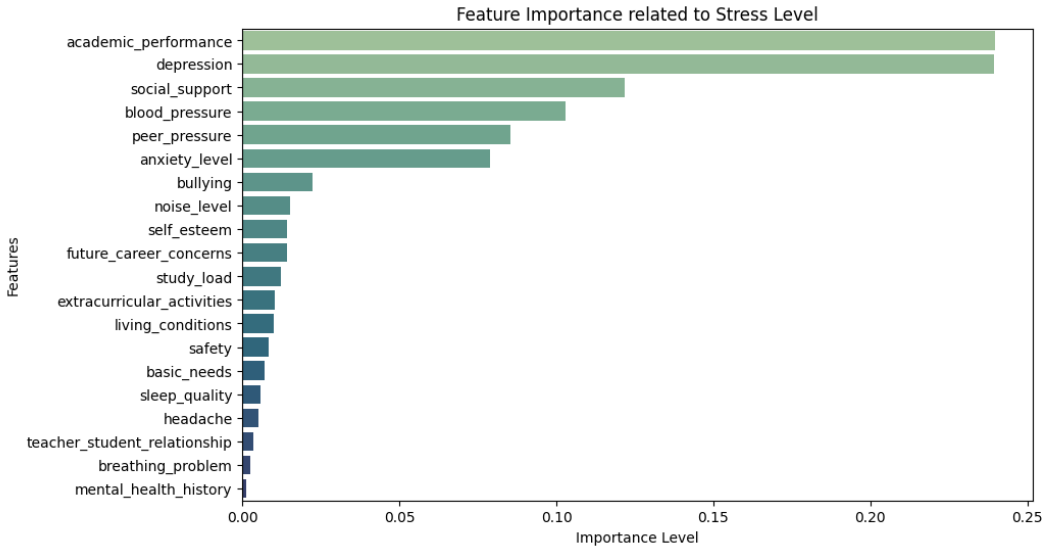

Fig. 4: Feature Importance Chart

RESULTS

Referencing our analysis and the visuals we've created, we concluded that **bullying**, **future career concerns**, and **anxiety levels** are the top key factors influencing stress levels. On the other hand, factors like **self-esteem**, **sleep quality**, and **academic performance** have the strongest negative impact on stress levels. The negative impact means that higher values in these negative factors are related to lower stress levels, while positive correlation coefficients are associated with higher stress levels.

After the data science phase, we selected the Random Forest Classifier as our final machine learning algorithm for this project because it performed better than the KNN models. In terms of iterations, we initially leaned towards the first iteration as it produced the highest accuracy score. However, after further discussion and research, we ultimately decided to go with the second model iteration. Here are the outputs for both Random Forest iterations:

- **Iteration #1: Non-scaled dataset with no hyperparameter tuning**
  - Training: 100%
  - Testing: 88.7%
- **Iteration #2: Non-scaled dataset with hyperparameter tuning**
  - Training: 89.2%
  - Testing: 86.4%

We ended up selecting the second iteration because the mode's performance on the test set is more reliable and less likely to be overfitted. Although the first iteration had a higher testing accuracy, its perfect training accuracy indicates that it may be overfitting the training data. This typically suggests that the model will not be able to generalize well to new data. The second iteration, with slightly lower accuracy but better generalization, provides a more balanced and reliable model, which is crucial for the practical application of the model in the long run.

RECOMMENDATIONS

Based on our results from the correlation matrix (Fig. 2), we identified the top factors that increase stress levels and those that help reduce stress levels. For the factors that increase stress, we need to find ways to reduce and mitigate them. For the factors that reduce stress, we need to focus on improvements to further help lower stress levels. Taking this into account, here are some simple strategies to manage and reduce student stress:

1. **Bullying**: Schools should create a safe environment with clear policies and programs to prevent and address bullying.
2. **Future Career Concerns**: Schools should offer more career guidance and counseling services to help students navigate their future career paths. In addition, schools can also organize workshops on resume writing, job searching, and interview skills to boost students' confidence after graduation.
3. **Anxiety Levels**: Schools should teach mindfulness practices to help them manage anxiety and properly articulate the importance of managing anxiety levels.
4. **Self-Esteem**: Schools should create an environment where teachers use positive reinforcement and provide constructive feedback to help build student's confidence.
5. **Sleep Quality**: Schools should teach their students about the importance of good sleep practices and encourage maintaining a regular sleep schedule.
6. **Academic Performance**: Schools should offer more tutoring services and academic support to help students who are struggling with their classes.

Once our model is tested further and finalized, we would also like to recommend that students, parents, and teachers use our Stress Level Predictive Model as having some sort of metric to rely on can help visualize a student's mental state. We believe that continuous tracking of this metric will allow students to take action to reduce their stress and seek help if their levels reach a high predetermined threshold.

CONCLUSION

Tying everything together, our project successfully identified the key factors influencing student stress levels and developed a predictive model that can accurately predict stress levels based on these factors. We deeply analyzed all the related stressors like psychological, physiological, social, environmental, and academic and found that bullying, future career concerns, and anxiety levels are the most significant contributors to student stress. On the other hand, higher self-esteem, sleep quality, and academic performance help reduce stress. Through various trial and error experiments, we chose the Random Forest Classifier as our final model. In conclusion, we highly encourage the use of our Stress Level Predictive Model for stress management as we believe implementing this can create a supportive environment that promotes student mental health.

REFERENCES

Al-Atawi, A. A. (2023). Stress monitoring using machine learning, IoT and wearable sensors. Sensors, 23(21), 8875. https://doi.org/10.3390/s23218875

Bay Atlantic University. (2019). Effects of stress. Retrieved from https://bau.edu/blog/effects-of-stress/

Córdova Olivera, P. (2023). Academic stress as a predictor of mental health in university students. Cogent Education, 10(1). https://doi.org/10.1080/2331186X.2023.2232686

Li, H. (2022). Stress prediction using micro-EMA and machine learning during COVID-19 social isolation. Smart Health, 23, Article 100242. https://doi.org/10.1016/j.smhl.2021.100242

Maxhuni, A. (2016). Stress modelling and prediction in presence of scarce data. Journal of Biomedical Informatics, 63, 344-356. https://doi.org/10.1016/j.jbi.2016.08.020

Muñoz, S. (2022). Prediction of stress levels in the workplace using surrounding stress. Information Processing & Management, 59(6), Article 103064. https://doi.org/10.1016/j.ipm.2022.103064

Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2019). The impact of stress on students in secondary school and higher education. International Journal of Adolescence and Youth, 24(1), 104-112. https://doi.org/10.1080/02673843.2019.1596823

Rana, A. (2019). Stress among students: An emerging issue. University of Delhi. Retrieved from https://www.researchgate.net/publication/334835276_Stress_among_students_An_emerging_issue

Ribeiro, I. J. S., Pereira, R., Freire, I. V., de Oliveira, B. G., Casotti, C. A., & Boery, E. N. (2017). Stress and quality of life among university students: A systematic literature review. Health Professions Education, 3(2), 77-87. https://doi.org/10.1016/j.hpe.2017.03.002

Slimmen, S. (2022). How stress-related factors affect mental wellbeing of university students: A cross-sectional study to explore the associations between stressors, perceived stress, and mental wellbeing. PLOS ONE, 17(11), Article e0275925. https://doi.org/10.1371/journal.pone.0275925

APPENDIX

Here are the reported outcomes from each of our model training iterations:

RANDOM FOREST CLASSIFER:

Iteration 1: Non-scaled w/ no hyperparameter tuning

90/10 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	85.4%	87.2%	85.5%
Training Accuracy Score	100%	100%	100%
Difference	14.6	12.8	14.5

80/20 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	88.1%	86.4%	85%
Training Accuracy Score	100%	100%	100%
Difference	11.9	13.6	15

70/30 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	88.7%	88.5%	87.6%
Training Accuracy Score	100%	100%	100%
Difference	12.3	11.5	12.4

Iteration 2: Non-scaled w/ hyperparameter tuning

90/10 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	85.5%	87.3%	86.4%
Training Accuracy Score	93.2%	93.8%	89.2 %
Difference	7.7	6.5	2.8

80/20 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	95%	97.8%	87.7%
Training Accuracy Score	86.6%	85.5%	99.1%
Difference	8.4	12.3	11.4

70/30 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	87.6%	87.6%	87.3%
Training Accuracy Score	99.2%	95.3%	99.7%
Difference	11.6	7.7	12.4

Iteration 3: Scaled w/ no hyperparameter tuning

90/10 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	89.1%	85.5%	90%
Training Accuracy Score	100%	100%	100%
Difference	10.9	14.5	10

80/20 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	85.6%	86.4%	85.5%
Training Accuracy Score	100%	100%	100%
Difference	14.4	13.6	14.5

70/30 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	88.5%	89.1%	87.6%
Training Accuracy Score	100	100%	100%
Difference	12.5	10.9	12.4

Iteration 4: Scaled w/ hyperparameter tuning

90/10 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	89.1%	85.5	90%
Training Accuracy Score	100%	100%	100%
Difference	10.9	14.5	10

80/20 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	85.6%	86.4%	85.5%
Training Accuracy Score	100%	100%	100%
Difference	14.4	13.6	14.5

70/30 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	88.5%	89.1%	87.6%
Training Accuracy Score	100%	100%	100%
Difference	12.5	10.9	12.4

KNN:

Iteration 1: Non-scaled w/ no hyperparameter tuning

90/10 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	88.2%	N/A	N/A
Training Accuracy Score	92%	N/A	N/A
Difference	3.8	N/A	N/A

80/20 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	87.3%	N/A	N/A
Training Accuracy Score	91.9%	N/A	N/A
Difference	4.6	N/A	N/A

70/30 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	88.2%	N/A	N/A
Training Accuracy Score	90.9%	N/A	N/A
Difference	2.7	N/A	N/A

Iteration 2: Scaled w/ no hyperparameter tuning

90/10 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	85.5%	N/A	N/A
Training Accuracy Score	88.3%	N/A	N/A
Difference	2.8	N/A	N/A

80/20 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	86.4%	N/A	N/A
Training Accuracy Score	88.9%	N/A	N/A
Difference	2.5	N/A	N/A

70/30 Ratio:

	First Attempt	Second Attempt	Third Attempt
Testing Accuracy Score	87.6%	N/A	N/A
Training Accuracy Score	88.4%	N/A	N/A
Difference	0.8	N/A	N/A