An Exploration of Stress Among High-School Students

Behavioral Data Science Final Project Report

Reno Malanga & Christian Arndt

Cognitive Science Department
Rennsalaer Polytechnic Institute

April 26, 2025

Contents

1 Introduction

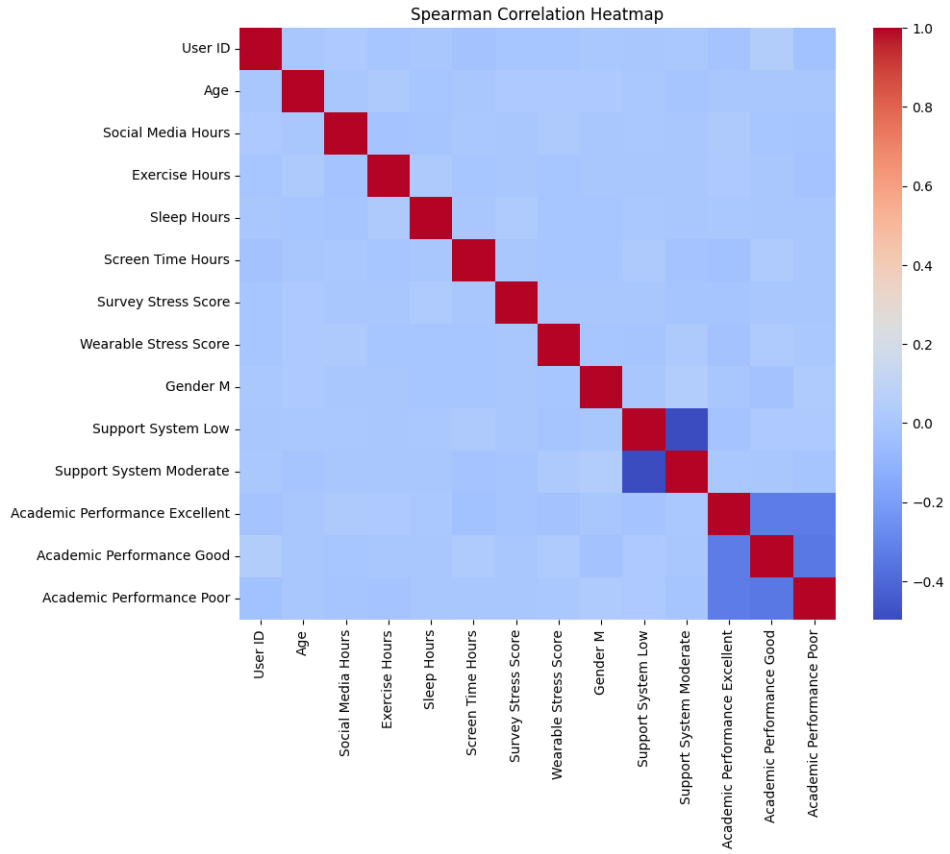
Being in a class called Behavioral Data Science, we did not want to choose any topic simply concerning itself with behavior, so we decided to work with a topic that was relevant to our research. For a topic of interest, Reno's research is more aligned with behavior, as he is currently researching stress, so we opted to look for datasets concerning that. After some searching, we found a couple of interesting datasets based on their specific topic and the data itself. The two datasets in which we were interested included the mental health of adolescents and stress factors from students called "Mental_Health_Analysis_Among_Teenagers" and "Student stress factors" respectively. With these two datasets, the research question we were interested in focused on asking if there is a relationship between one's stress and their lifestyle, and if there is, what does it look like?

2 Data

Choosing a dataset wasn't technically difficult, but we did run into some problems. Our first dataset turned out to be unusable, for reasons discussed below. After some secondary exploration, we went with the "Mental_Health_Analysis_Among_Teenagers" first as it showed interesting trends.

2.1 The Dud

This first dataset had a sample size of 5000, but each dimension was uniform. This immediately made us suspicious, but we were curious about the data, so we ran a Spearman correlation matrix to observe if there were any links between dimensions. The resulting heatmap is below:



As you can see, there were minimal correlations between dimensions. Based on these two metrics, we decided to cut our losses and find another dataset with more promising results.

Looking into other stress data, we found one with a lower usability score (9.41), but it showed more Gaussian distributions within the dimensions, making this a more viable dataset to work with.

2.2 A New Hope

Before getting into analyzing the data, we wanted to get an idea of what the study was especially concerned with and what the participant population is. The data we will be using for this analysis is called ”Student stress factors” which had participants complete an online six-question survey. The data itself can be found

here:

`https://www.kaggle.com/datasets/samyakb/student-stress-factors`.

According to the website where this code was provided, the majority of participants are engineering students. Since there are no demographic questions relating to the participants themselves, we can only assume that they are college students. The first five questions asked participants about their lifestyle (sleep quality, headache frequency, academic performance, study load, and extracurricular practice frequency), and the last question asked about their perceived level of stress. Each question in the survey is listed on a Likert scale from 1 to 5.

Since the survey is only six questions long, and has no demographic information, we are rather limited in what we can do for our methods, along with the implications of what our models can say. When it comes to stress research, collecting demographic information is very important, as everyone experiences stress differently. Some factors that can impact one's level of stress include age, gender, sex, and socioeconomic status.

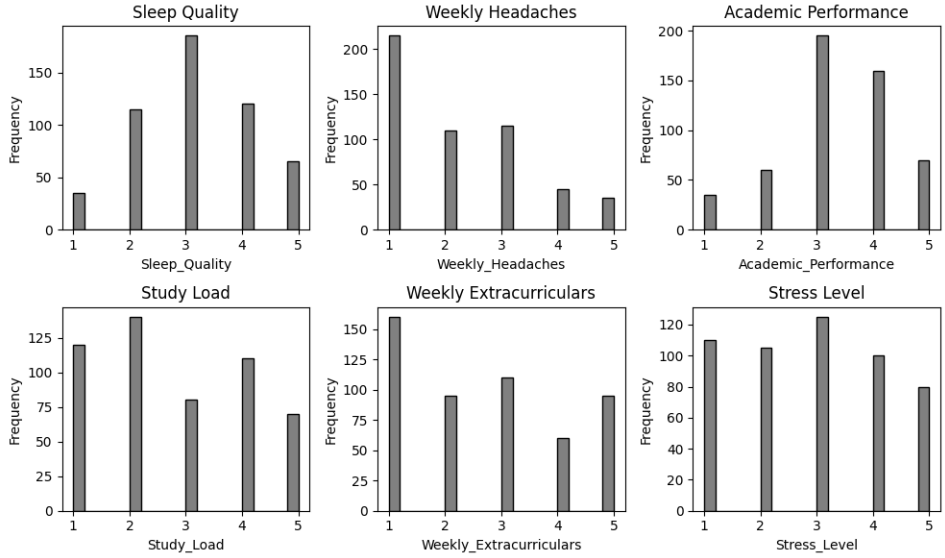
A deeper dive into the questionnaire allows us to see what exactly the participants are answering. The first question relates to sleep quality, as mentioned above these are on a scale of 1 to 5, with no labels on either side which has the potential to confuse participants. Another question has a misalignment of what is being asked and what the potential answer is. For example, question 2 asks, "How many times a week do you suffer from headaches?" With the answer on a scale from less to more, the question is not being explicitly answered. The last question asks about one's perceived level of stress, however the ordering of the Likert scale is from "Very High" to "Very Low", 1 to 5, respectively. When running this study, researchers may need to reverse the test scores for this question so they can accurately reflect an appropriate result.

Below is a table containing a few summary statistics for each of the columns in our dataset.

Feature	Mean	SD	MAD
Sleep Quality	3.125	1.099	1
Weekly Headaches	2.183	1.247	1
Academic Performance	3.327	1.061	1
Study Load	2.750	1.372	1.5
Weekly Extracurriculars	2.683	1.471	1
Stress Level	2.875	1.358	1

Looking at the summary of the data, a couple of things stuck out to us. First, let us note that the participants' 'Weekly Headaches' does not have a normal distribution. This makes sense given the nature of the question; we would expect fewer people to have more frequent headaches and vice versa. It's also the only feature for which the median absolute deviation (MAD) was larger than the standard deviation (SD). Based on the MAD being larger than the SD for study load, we may see the data in this dimension not being normally distributed; this can be been in the data histograms (plotted below).

The other thing we noted was that, for all features but 'Weekly Headaches' the mean very close 3, which is the expected value of likert-scale data with a normal distribution. This is hopeful; let us now consider the histograms.



We can see that, as we expected, the 'Weekly Headaches' feature is not distributed normally; it appears to exhibit some kind of decay. This lines up with our expectations.

We find the histogram for 'Study Load' and 'Weekly Extracurriculars' to be more suprising; these features also appear not to be normally distributed. They don't have exactly the same shape, but they both exhibit a sort of dip in the middle, with concentrated portions on both sides. It's possible that this shape is indicative of two groups of students: those who push themselves to study and do a lot, and those who don't care that much. Such an underlying structure could easily explain the shapes here, but it's just speculation.

This gives us hope that the clustering we plan to run might produce interesting results. It also provides a great demonstration of why summary statistics on their own can be misleading.

3 Methods

There are two stages in our experiment.

First, we will attempt to distill the information we can out of the lifestyle features. Our plan here is to use exploratory factor analysis to attempt a dimensionality reduction. If that fails, we'll instead try to run a k-means clustering algorithm. We would prefer the latent variable model to the clustering model, as it will give us more informative inputs to our regression models, but we have some concerns about the data's reducibility; there are only 5 lifestyle features to start with, so there simply isn't much room for reduction.

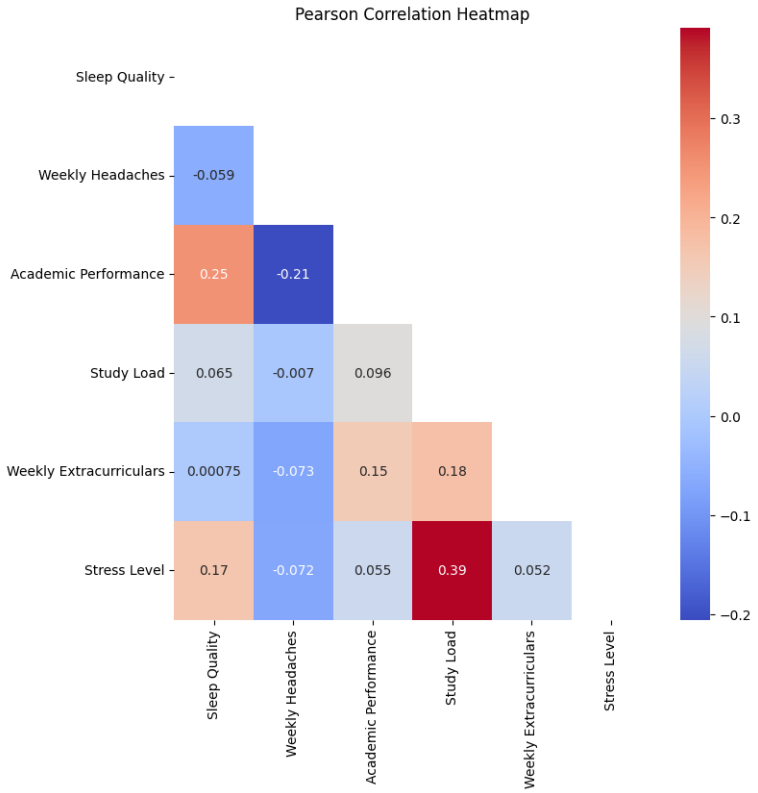
Once we've done what we can with the lifestyle features, we will fit a number of regressions from the lifestyle features onto stress. This exploration of the relationship between lifestyle and stress is the primary goal of our project; all of our various messings-around with the lifestyle features are simply to ensure that we're getting the best regression model we can.

Also, as a general note: we standardized the data before all analysis. The only graphics in this report that don't use standardized data are the data summary table and histogram in section 2.2.

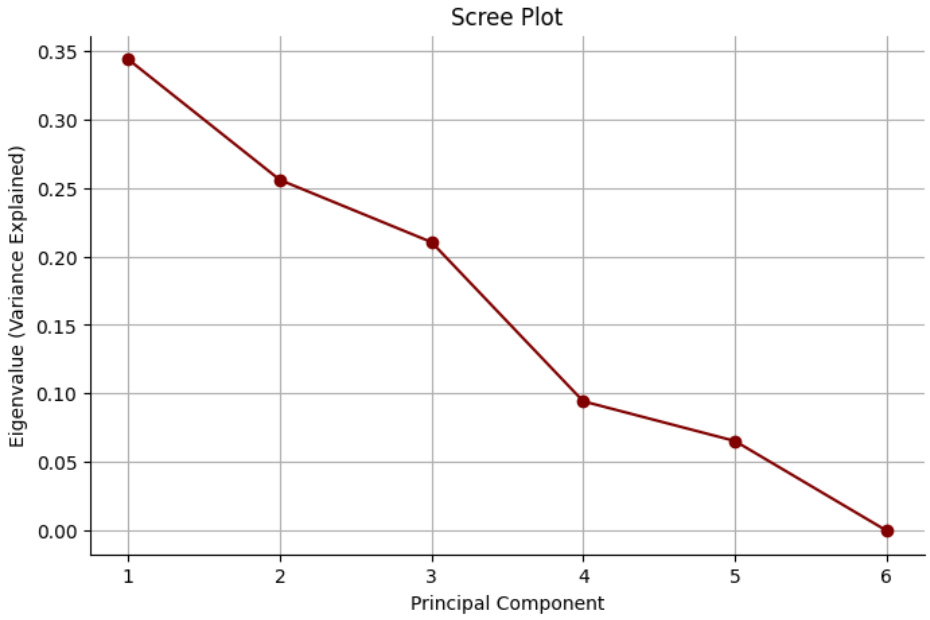
3.1 Lifestyle Data Distillations

3.1.1 Latent Variable Models

As mentioned previously, there are five questions relating to the lifestyle of each participant. As an initial exploration into the effectiveness a latent variable model might have on these features, we first ran a Pearson correlation matrix to determine if there were any correlated factors within the study. Based on the figure below, there are a couple of notable correlations that are present in the heatmap.



We see that there is a positive correlation between stress level and study load, and a negative correlation between academic performance and weekly headaches. Following the correlation matrix, we generated a scree plot to determine if the factors were reducible:

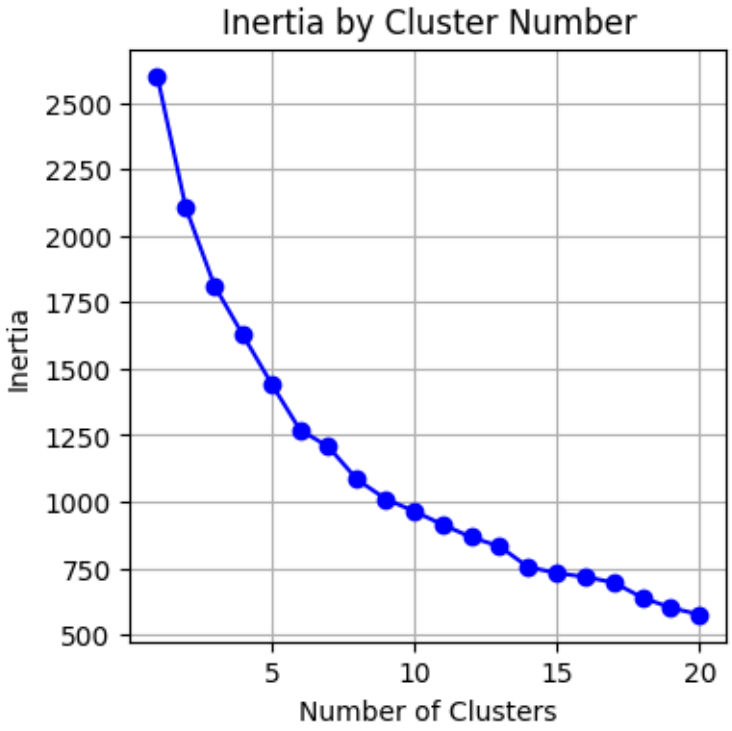


The above plot indicates that our data is basically irreducible. There's no obvious elbow, and more importantly, the eigenvalues are all well below 1. Instead of pursuing a latent variable model, we will move on to using k-means clustering to analyze the lifestyle portions of the data.

3.1.2 K-Means Clustering

Because clustering is unsupervised, we elected not to do a train/test split. Never fear: we will use a split for our later regression models. This may be a bit of a spoiler, but as a post-hoc justification for this decision: even using all the data to fit our clustering model, our best regression models were those which used only the basic data, excluding clustering labels.

We used sklearn to fit a number of clustering models to the data. Below is a plot of those fits' inertia over the number of clusters.



Like the scree plot with PCA, this plot helps determine the number of clusters we should have in our model. Inertia is a measure of how well the clusters capture the data; the lower the inertia, the better the clustering fit. So, the goal is to look for an elbow. This is the point at which adding more clusters stops effectively capturing the shape of the data. There's no obvious elbow, but after some discussion and re-runs of the clustering algorithm with different seeds, we decided that both $k = 3$ and $k = 8$ exhibited elbow-like behavior more than any other k. We added the labels from these clustering models to our data and prepared for the regression.

The cluster centers for the $k = 3$ clustering model are as follows (split into two tables for formatting readability):

Cluster Centers for $k = 3$

Feature	C1	C2	C3
Sleep Quality	0.536	-0.364	0.076
Weekly Headaches	1.342	-0.587	-0.109
Academic Performance	-0.622	-0.285	0.569
Study Load	-0.338	-0.200	0.352
Weekly Extracurriculars	-0.594	-0.651	0.895

There are a few things to say here. The third cluster of students clearly represents the high achievers; these students participate in a lot of extracurriculars and have high academic performance and study load, and low sleep quality.

The other two clusters are harder to parse for meaning. The first cluster has a very high headache level, and a high sleep score, with low study load, academic performance, and extracurricular participation. The second cluster, on the other hand, is just below average on all counts. There is no obvious archetypal student lifestyle that we think fits into these categorizations.

The cluster centers for the $k = 8$ clustering model are as follows (split into two tables for formatting reasons):

Cluster Centers for $k = 8$ (clusters 1–4)

Feature	C1	C2	C3	C4
Sleep Quality	1.251	-0.167	1.039	0.455
Weekly Headaches	-0.547	-0.665	-0.734	1.607
Academic Performance	0.352	0.579	0.320	-0.720
Study Load	-0.692	0.782	0.279	-0.319
Weekly Extracurriculars	-0.804	-0.624	1.212	-0.549

Cluster Centers for $k = 8$ (clusters 5–8)

Feature	C5	C6	C7	C8
Sleep Quality	-0.341	-0.324	-0.941	-1.388
Weekly Headaches	0.789	-0.393	0.072	-0.307
Academic Performance	-0.072	0.054	0.634	-1.345
Study Load	1.336	-0.939	-0.546	-0.328
Weekly Extracurriculars	0.726	-0.621	1.205	-0.464

Having 8 clusters makes it more difficult to identify clusters with particular ways of life, so we'll simply report the centers and move on.

One final note on our clustering: we plotted the inertia of our clustering fits over the number of clusters as a means of finding the best clustering models for our data, but this is a purely relative metric. Inertia is calculated by squaring the distance between each point in a cluster and that cluster's center, and then summing that value across all the data points. This means that inertia scales with basically everything: number of data points, number of dimensions, and feature scale all contribute magnitude to the inertia of a clustering model.

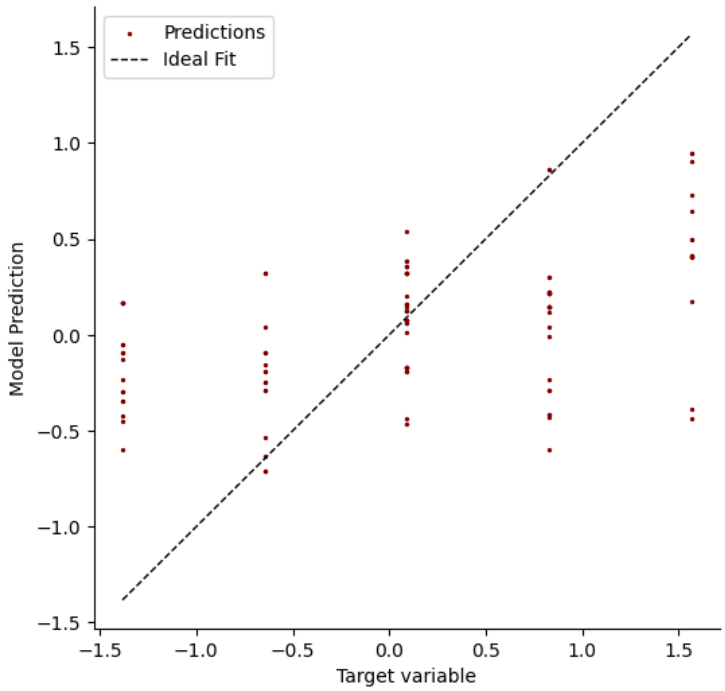
So, while the above graph gives us a method for choosing the best clustering models, we still don't know whether or not those models are actually *good*. There are other methods we could use for a normalized evaluation of clustering models, but we decided they were unnecessary for our project. Our main goal is simply to increase the number of inputs to our regression model; if the cluster labels prove very effective in increasing our regression models' performance, we can intuit that they're good clustering models, and if not, they're not very relevant to our topic of interest.

3.2 Regression Models

With our attempts at squeezing as much as we can out of our dataset's lifestyle features out of the way, let us move on to the regression model fitting and evaluation. As a global note: all regression model evaluations (plots and RMSE measurements) are generative.

3.2.1 The Baseline Model

As a baseline, let us first consider the simplest model: a linear regression using only the basic dataset. This model has a RMSE of roughly $.9853$. Below is a plot of the model's predictions over the true value:



This is not a particularly impressive model. It is somewhat obvious from the plot, the model basically failed to capture a meaningful relationship between lifestyle and stress. It is possible that such a relationship is just very weak, but let us consider some other regression models before making that conclusion.

3.2.2 The Model Bonanza

Including the baseline model, we fit 12 total models, by varying two model features:

- We fit models of the following model types:
  - Linear regressions
  - Random forest regressions
  - Deep neural nets
- For each model type, we fit a model with each of the following subsets of our data:

- the basic dataset, with no clustering labels
- the basic dataset with the labels from the $k = 3$ model
- the basic dataset with the labels from the $k = 8$ model
- the basic dataset with labels from both clustering models

Below is a table containing each model's RMSE. Columns represent the different model types, while rows represent the training data used.

	Linear Models	Random Forests	Deep NNs
Basic data	0.9853	0.5154	0.2077
Incl. k = 3	0.9459	0.5408	0.2062
Incl. k = 8	0.9143	0.5496	0.2004
Incl. k = 3, 8	0.8945	0.5474	0.2100

This table allows us to make a few interesting statements.

First, let us state that there is a clear hierarchy of model quality here: the linear models perform the worst, the deep neural nets perform the best, and the random forest models' performance falls almost exactly in the middle. Clearly there is a relationship here; describing that relationship just requires more functional expressiveness than is possible with linear regressions.

It is fairly obvious that the best model here is the basic neural network. The performance of all the neural networks is almost the same, so we feel most confident in choosing the simplest one.

Another thing we can see after examining the above table is that the clustering models add basically nothing to the regressions. The linear and neural-network based regressions both benefited some very small amount from the labels, but there are a few reasons not to consider this gain as meaningful:

- First, adding the cluster labels makes the models less intuitively explainable. This could be justified with a large increase in performance, but the performance gains seen here just aren't that big. We don't think they warrant the increased model opacity.
- Second, it seems odd that, for the neural nets, adding the $k = 8$ cluster labels improved performance, while adding both sets of labels *decreased*

performance. If our clustering inertia plot made it obvious that the data is most faithfully represented with 8 clusters, this behavior might make sense, but that is not the case: the plot made it clear that the $k = 8$ model is only marginally better at capturing the data than the other clustering models. Given this unintuitive result, we feel unable to make any confident statements about the usefulness of clustering labels.

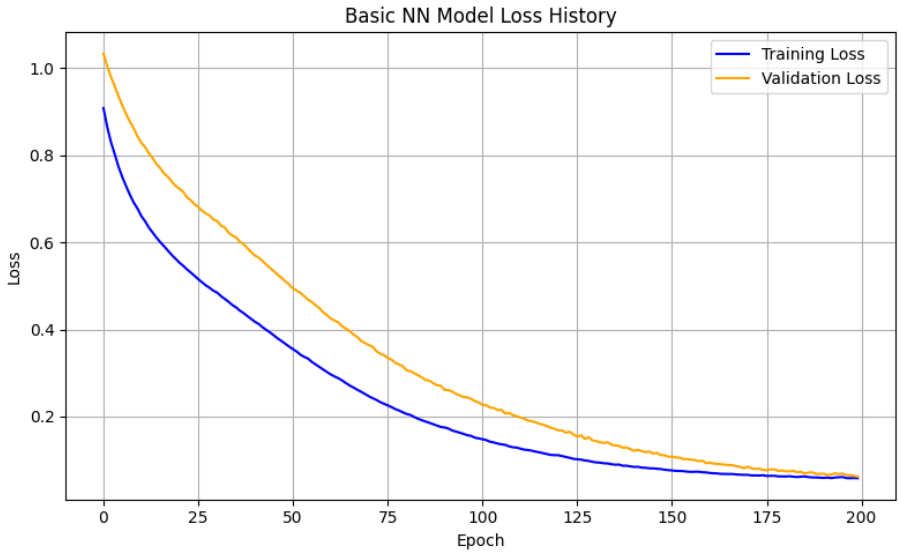
It's also worth noting that including clustering labels had a universally negative effect on the performance of the random forest models. All of this taken together means that our clustering models either:

- Completely fail to capture the data, or
- Capture parts of the data that are already mostly accessible through the underlying features and are thus redundant for the regression fitting.

3.2.3 The Best Model

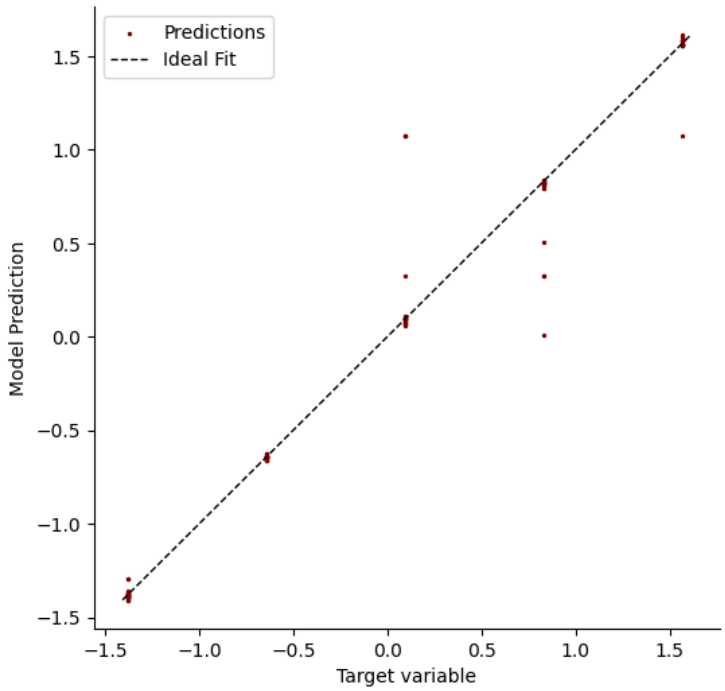
As mentioned above, the best regression model is the basic neural net. Let us now examine this model in more detail.

As a brief sanity check, let us examine the loss history for our basic neural network's fitting:



This is generally what we would expect: the loss decreases over time until it 'bottoms out', indicating that the model's fit has stabilized and the model has learned all it can.

We've already established that this model's RMSE on the test set is around .21. Below is a plot of the model's predictions over the target variable's true value:



Comparing this plot to the analogous plot for the basic linear model makes it incredibly obvious just how much better this model is at predicting stress levels.

3.2.4 An Experiment in Rounding

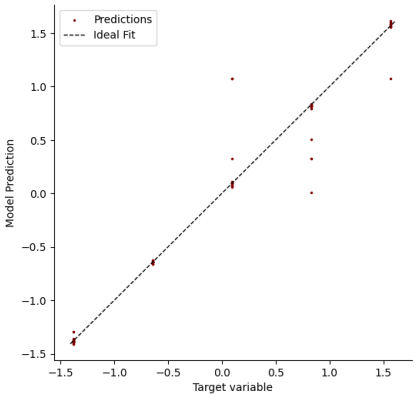
One thing that struck us while looking at the prediction-vs-truth plots is that the neural network seems much better at predicting integer values. We thought this might be the neural network's capacity to learn non-linear functions, like rounding. Since the target variable is on a likert scale (and thus, only contains integer values), it occurred to us that a more fair RMSE comparison between the models might first round the predictions. So we took the predictions for each model, inverse scaled them, rounded them, and then re-scaled them before recalculating the RMSE. Below is a table containing the results from this idea:

	Original RMSE	Rounded-prediction RMSE
Linear, Basic	0.9853	1.0440
Linear, $k = 3$	0.9459	0.9849
Linear, $k = 8$	0.9143	0.9499
Linear, $k = 3, 8$	0.8945	0.9049
RF, Basic	0.5154	0.5356
RF, $k = 3$	0.5408	0.5547
RF, $k = 8$	0.5496	0.5911
RF, $k = 3, 8$	0.5474	0.5452
NN, Basic	0.2077	0.1911
NN, $k = 3$	0.2062	0.1911
NN, $k = 8$	0.2004	0.1911
NN, $k = 3, 8$	0.2100	0.1911

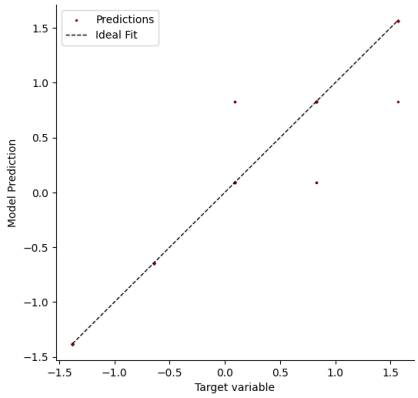
We were wrong about this; rounding the predictions almost universally increased the RMSE. The only models for which this was *not* the case are the neural networks, where rounding increased performance by some small margin.

While it goes agaiinst our initial assumption, this makes some intuitive sense: rounding should remove any error contribution from data points which the model predicts within 0.5 of the true value (on the non-standard scale), an increase error contributions for most of the data points outside that range. So if a model is already predicting values close to the line of best fit, rounding will improve performance; and if it's not, rounding will likely decrease performance. This is easily visualized by comparing the prediction-v-truth plots for the basic neural net both with and without rounding:

Non-Rounded Predictions



Rounded Predictions



We can see that the rounded predictions are quantized to the grid, and thus, more of the predictions are equal to the target variable than before.

We are not sure whether rounding the outputs of a regression is generally good practice, but we think it makes sense for this particular application discrete dataset. One important note, though: rounding does little to answer our research question. Rounding might make it better to compare the performance of our models on this particular data set, but it likely doesn't actually improve their capacity to understand the relationship between lifestyle and stress. This is because, in humans, stress is not discrete; the datas' non-continuous nature is the result of the collection method, and not a representation of reality.

4 Results

We have chosen a model as our ideal; let us now see what we can say with it. The first thing we should state is that we believe our basic neural network makes it clear that there *is* a relationship between lifestyle and stress. Our model does a pretty good job of predicting stress based on the lifestyle features. With that out of the way, let us consider our model in more detail.

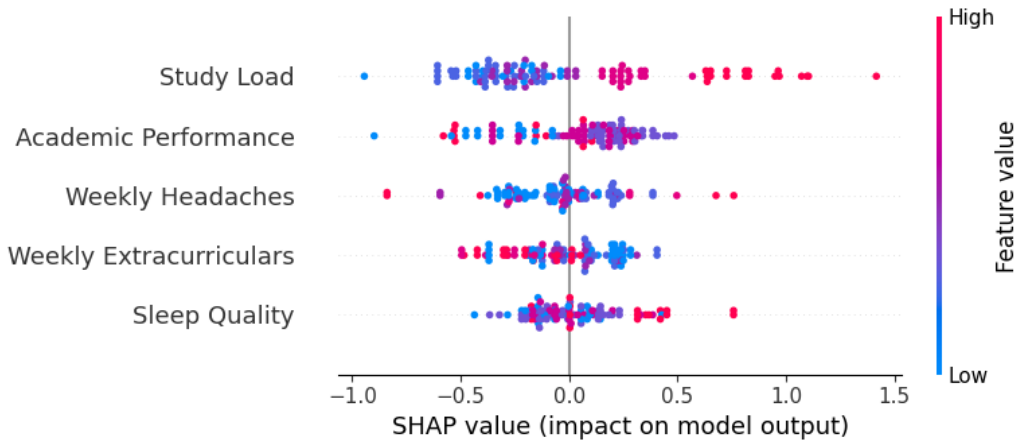
4.1 Shapley Values and their Interpretation

We used the `shap` library to analyze feature importance with the basic neural network.

These `shap.Explainer` objects calculate the impact of each feature for each data point; to get shapley value estimates, we need to take the mean across the data. But, because the variables have a scale that includes negative numbers, taking the pure mean could result in the negative and positive contributions canceling each other out. To get around this, we can take the mean of the absolute values. This scrubs the values of information about the kind of correlation, but should accurately reflect the features' relative importance. The value estimates we calculated this way are as follows:

Feature	Shapley Estimate
Study Load	0.3466
Academic Performance	0.2770
Weekly Headaches	0.2099
Weekly Extracurriculars	0.1928
Sleep Quality	0.1828

We can see that study load is the most important feature, which is a fairly intuitive conclusion. Let us now consider the following summary plot for some more detailed analysis:



This plot is information-dense. Each dot represents a single dimension of one data point in our test set. The color of the point encodes the value of that dimension for that data point; pinks are high values (4s or 5s on the likert scale), while blues are low values (1s or 2s). The y-axis contains each of the input features in descending order of importance from top to bottom, and the x-axis encodes the impact that feature had on the output for that data point.

All of this taken together indicates a few things:

- First, features with a large number of points far from 0 affect model output more than those clustered at 0.
- Second, the features with an obvious gradient have an obvious correlation with the model output; if pink is on the right and blue is on the left, the correlation is positive, and in the inverse case it's negative.

Understanding these two things makes it obvious that study load is by far the most important feature for predicting stress. Study load features both a large spread and an obvious gradient, indicating that study load is directly correlated with stress.

The other features are more difficult to interpret. It's clear from the absolute shapley value estimates that all of the features contribute to the regression, but the plot is not easy to parse. Particularly confusing is the 'Sleep Quality' feature; according to the shapley summary plot, it looks like high sleep quality is positively correlated with stress. This makes no sense to us. We considered that the question may have been asked in the negative, with 5 on the likert scale being the worst sleep quality, but after re-examining the form used to collect data that doesn't seem to be the case.

We can also see a confusing relationship between 'Weekly Headaches' and stress. It seems like the a high value for this feature indicates either a strong positive impact on the stress prediction, or a strong negative impact on the stress prediction. Low values for this feature are clustered near 0. This also seems unintuitive; we would expect more headaches to correlate strongly with more stress.

The other features are less egregious. While it's difficult to interpret the plot, there are no obviously unintuitive conclusions to draw.

The layout of the 'Academic Performance' feature, for example, could be easily explained by considering the different mindsets that students might have about scholastics. Most of the students with middling academic performance are clustered just to the right of 0, indicating that their academic life contributes somewhat to their stress. There are also a number of students with high academic performance in the same zone; these are likely the students who place a high value on academics, and so stress about their performance even when they're doing well. Then, to the left of 0, we see two extremes: students who do very well, and students who do very poorly. Those who do well are likely confident about their performance, while those who do poorly simply don't care about school.

The last feature, 'Weekly Extracurriculars', is nearly as straightforward as 'Study Load'. There's an obvious gradient, this time with the high values on the left; this means that students who participate in extracurriculars are less stressed.

5 Conclusions

Let us be clear in our conclusions: there is a relationship between lifestyle and stress.

This is an eminently obvious conclusion. Any human being could probably have told you as much without the hours of coding, debugging, and confusion this project wrought upon us. Still, we found this project validating both to our experience as students of data science, and also to our experience as periodically stressed-out students of data science. With that in mind, we're more than happy to state our obvious conclusion confidently and with great rigor.

Having experienced both much learning and much stress in the past week, we feel a particular affinity for those stressed-out students with a high study load. We're still confused by students with low sleep quality, though; this relationship likely requires more exploration to fully untease. We plan to personally explore the sleep problem space after submitting this report.