College Student Stress

IEOR 135/290

Group 25 Data-X Final Report

Veda Gadhiya, Jared Gutierrez, Mark Hashimoto, Swetha Prabakaran, Gayatri Dutt, Nafisa Nikhath

1. Problem

Stress has been shown to have both positive and negative effects. On the positive side, stress is necessary because "our brains are wired such that it's difficult to take action until we feel at least some level of this emotional state." Furthermore, intermittent stress "entices the brain into growing new cells responsible for improved memory." On the opposite end of that, "prolonged stress causes degeneration in the area of the brain responsible for self-control" and "can wreak (havoc) on one's physical and mental health" (Bradberry, 2019). There exists an optimal level of stress such that you achieve optimal performance.

College students are increasingly experiencing prolonged stress from a variety of stressors and increased demands such as social, academic, growing pains, figuring out their future, and finding balance. All of this stress can impact students in various ways including academic performance and burnout. The American College Health Association - National College Health Assessment (ACHA – NCHA) found that in 2018 at 31.9% stress was the biggest factor to affect student academic performance (Fall 2018 Reference Group). On a more severe note, stress can affect student mental health. A World Health Organization and Columbia University global study of 14,000 first-year college students from eight countries (including the US) "found that 35 percent struggled with a mental illness" with depression and anxiety being the top illnesses (Hess, 2018). These are conditions that need proper care and counseling at the right time to avoid overall physical health consequences, further mental health deterioration, and ultimately, potentially life-threatening or harmful situations such as suicide attempts, self-harm, or violence.

2. Approach

The overall objective of this project is to help students manage stress by identifying potentially dangerous stress levels and provide recommendations on how to balance/optimize stress and performance.

2.1. Goals

The primary goal is to address the negative effects of stress, as those effects can be very damaging and possibly life altering. According to Professor Randy P. Auerbach (Columbia University Psychology) and Dr. Sherry Benton (VP of the Society of Counseling Psychology of the APA): "Internet-based clinical tools may be helpful in providing treatment to students who are less inclined to pursue services on campus or are waiting to be seen" and "It's essential that students know when it's time to reach out for help" (Hess, 2018).

As such, the desired end product for this project is a web application where students enter current habits/mental state, and receive recommendations as well as predictions. The main goal is to use machine learning techniques with a student's habits/ state as inputs to predict their stress level and then classify them as having an acceptable or critical level of stress. This prediction could then be used to advise students to seek professional help and especially encourage timid students to access those services, all of which can prevent crisis episodes. The predictions could also be used to educate students on lifestyle changes that can help them. Overall, by predicting and classifying stress levels, attempts to prevent mental or physical health deterioration can be made.

The ultimate goal is to help students figure out their level of stress for optimal performance in order to help them prioritize and balance their demands such that the stress is "worth it" and they do not reach the exceedingly negative effects of stress in the first place.

2.2. Dataset

To keep scope manageable but achieve high relevance, an initial focus on high performing colleges like UC Berkeley or similar colleges in the United States was done. In the search for datasets, it was found that due to the sensitive nature of health and psychological data, finding and gaining access to such datasets is difficult. In addition, tradeoffs between relevance, reliability, and feature/response/observation richness were discovered.

This project ultimately chose to use the StudentLife dataset from Dartmouth College, with contributions from researchers at The University of Texas at Austin and Northeastern University. This dataset was found via a reference in a 2014 ACM publication while continuing research into the problem statement. The paper is titled "Student Life: assessing mental health, academic performance and behavioral trends of college students using smartphones". The links to the dataset and paper are in the appendix. The data for this set was collected from about 50 computer science students during the spring semester at Dartmouth College via an app on their phones and surveys administered before and after the semester. Due to privacy issues, all of the data related to student identities was removed and replaced by user id numbers (UID).

This dataset was chosen because it was publicly available from a reliable source and could serve as a good starting point as it was collected from a high caliber university with students from a high stress major. The dataset contains 12 GB data with observations as students and features as student life aspects. The table below highlights some of the contents from this dataset.

Table 1. Highlights (examples) of the Contents of Dataset

Student Life Aspects	Survey Data	App/Sensor Data
Personality (ex. mood, outlook)	Big five personality assessment, Flourishing scale (optimism/outlook), Loneliness scale, PANAS (mood)	Mood, Behavior, Social
Activities/Lifestyle (ex. appetite, sleep)	PSQI (sleep quality), VR-12 (overall health), PHQ-9 (functioning)	Activity, Exercise, Sleep, Conversation
Academics (ex. grades)	GPA, Classes, Piazza usage, Deadlines	
Stress	Perceived stress	Stress checks

2.3. Data Cleaning

After selecting a dataset, several challenges in regard to data cleaning and wrangling were observed. In general, the balance between number of features, number of observations, and quality of data was difficult to manage. The first major challenge was the file/data structure of this set. The data was decentralized and separated in many directories that were stored in various types of files (CSV, JSON), which also had temporal differences (seconds, hours, daily, semester, etc.). There were also inconsistencies in row sizes for different users. For example, the flourishing scale for a user, that measures things like optimism, was administered before/after the semester and stored on a CSV with all users; while, the sensor data on the accelerometer, measuring activity, was taken in intervals of a few seconds and app questionnaire about stress was on a daily basis, both of which were stored separate JSON files per feature per user. All of this made merging and organizing the files problematic. In order to address these challenges, this project created functions to merge by UID, converted everything to a uniform time, and made aggregator functions.

Another challenge was the fact that there were many mixed data types. There were strings in the pre and post survey responses (ex. agree/disagree, frequently/not often). The app and sensor data contained a mix of strings, floats, and integers (ex. yes/no, hrs of sleep, mins of activity, # of people). Coupled with this challenge was the fact that there are positive and negative aspects of stress collected in the data. For example, the PANAS (Positive and Negative Affect Schedule) that measures mood asks about both ends of the spectrum like if you are upset/guilty (negative) or proud/enthusiastic (positive). Moreover, there were several non-categorical responses. This project took several steps to translate the data so that it could be used. The survey data was translated into different numeric scales; for example, the strongly agree to strongly disagree responses were scaled 1-5. The app and sensor data was handled via creating keys/bins to translate. All of the questions asked in the data were looked at to see if they suggested a positive or negative outlook and then the numeric values were inverted if they were negative.

Lastly, the problem of missing data arose. There were different missing UIDs (observations) in various files. In addition, there were certain feature values missing in files and the missing values were of different types. This led to the predicament of the tradeoff between more features, more observations, and quality. One technique used to tackle this was to first drop UIDs missing from many files as well as drop columns with too many NaNs (via a threshold), as it would be difficult and inaccurate to fill in so many missing values. Next, the remaining NaNs were filled using various methods based on what was deemed fit on a case by case evaluation. For example, in the Big 5 Personality questionnaire, the 1-5 strongly disagree to strongly agree scale, missing values were filled in with 3 as it was the neutral neither agree nor disagree position that would affect the data the least. Another example is for the PSQI sleep data, the NaNs were filled in with the median as that would have the least impact on quality. Note that data cleaning is done in the EDA notebooks as well as the EMA and processing notebooks located on the GitHub repo with a link provided in Appendix 5.4.

2.4. Exploratory Data Analysis (EDA)

Some of the interesting takeaways from initial data analysis are as follows. In general, students have a similar outlook when comparing their pre and post self-perceived success (flourishing scale survey), can see that their mean is approximately the same around 42.8 with the total range being 15-56. The sensor and app data can help make up for this in modeling.

The pre and post survey stress levels were also similar. Both levels were highly concentrated in the middle with mean and median both around 18 with a range of 3-34. The recorded stress levels from the app can help make up for this. The app recorded stress levels ranged 1-5 with a mean of 2.26. It was interesting to note that the mean stress was slightly higher during the middle of the day, later in the week, and ramped up late in the semester. Along those lines, it was interesting to note that physical activity level was very low in morning hours until around 11 am and then picked up.

The most correlated questions to stress levels were those that were related to mood (ex. PANAS) and daily functioning (ex. PHQ and VR-12). Within those questions, top negative correlations were noted from conscientiousness, interested, active, alert, flourishing score (optimism), and VR-12 mental and physical scores. Top positive correlations were noted from upset, neuroticism, PHQ score, nervous, and hostile.

In terms of academics, upper end of deadlines showed a stronger correlation to stress levels as well. The range of total deadlines for the semester per student was 7-84, which meant weekly it was 1-12 with Monday and Sunday having the most deadlines. The overall GPAs for the students ranged 2.4-3.947 with mean around 3.4. Also, more stress was reported on campus locations like around engineering and libraries (the corresponding map is in the appendix). There are several EDA graphs/charts in Appendix 5.1. Also, more EDA graphs/charts and comments are in the same notebooks on GitHub as mentioned above with data cleaning.

3. Solution

This project was ultimately able to use the Dartmouth dataset to develop models that predict the levels of stress students are under with about a 70% chance of accurately predicting a student's stress level (out of 5). The stress ranking is based on a number of surveys and continuous data recording points that gave a deep level understanding of what a student is going through. This project's solution particularly targets the average college/university student as much of the data points revolve around a student's life. This includes data points such as physical activity, self-perception, campus involvement, sociability with other students, number of deadlines etc. This project tested and tailored the models to predict accurately and correctly so that students would be able to gain a better understanding of how their stress levels compare to their fellow students. This would then encourage the students on the higher end to seek help and get it before their health deteriorates.

3.1. Models

In order to address this solution, we developed several models that allowed us to use the dataset that we had in order to predict student stress levels and the habits/features that play the largest affect in the outcome. We began by creating a linear regression model that took in all ~120 features and trained it across a large subset of the student population ultimately getting an MSE of 19.035. We then performed a logistic regression that would classify students under 5 bins of stress levels that showed how stressed they were compared to the average student at Dartmouth. With a C variable = 1, we were able to accurately predict the stress categorization of students up to 72% accuracy. Wanting to make our models more accurate we used lasso regression and feature correlation in order to eliminate uninfluential features and group questions into categories that more broadly expressed the students habits. This resulted in a slight increase in the accuracy of both our linear and logistic regressions. We found that our MSE was able to be reduced by 17% and our classification accuracy increased ~2%.

- A table summary of model performance is in Appendix 5.2.
- A highlight of the correlation matrix formed for feature reduction found in Appendix 5.2.

3.2. System Overview

A figure representing the system overview is in Appendix 5.3. The system uses 80-120 parameters and numerical variables as inputs into the regression models. These inputs are related to personality, activities/lifestyle, and academics and were chosen via feature selection and filtering to have the most impact. The inputs were run thru several several types of models, such as linear modeling and classification on a subset of the students in order to train and fit the model. The resulting output of the system is a student's predicted/projected level of stress and how it compares to their fellow students. Additionally, similar models (including stress as a parameter) were run in order to predict both a student's grade (semester and overall). With this information students would be able to see whether or not the stress they are enduring is worth the GPA that they are receiving.

In the future, the model can be linked with this project's working web application such that the system can add in feedback from students that can then continually reinforce the models and ultimately create a more accurate predictor. Moreover, additional data visualizations to the user interface, so students can visualize where they lie on the average student stress curve, can be added. Lastly, more recommendations on how students can improve either their grades via better study habits and making adjustments in their daily lives to reduce their stress level can also be added to the system. The hope is that this product can be used to overall better the lives of students.

3.3. User Interface

A link to the prototype user interface along with images from it are in Appendix 5.3. The general idea behind the user interface is to make the main interface a short survey that students can take quickly so they are more likely to complete it. This survey is a reduced subset of the questions asked within the surveys that were found (through EDA) to be the most effective questions in predicting a student's stress level (about 10-14 questions). Eventually (after linking the model to the user interface), similar to what was mentioned in the system overview, the survey answers would feed into the model and output not only stress levels but also visualizations and recommendations. This quick, easy, and accessible interface would ultimately be able to help a lot of students get the help they need and improve/optimize their stress.

4. Conclusion

4.1. Learning Path

This project has been a learning experience, the highlights of which are as follows. Finding the right dataset is very difficult and yet crucial part of the project that can dictate the direction it goes in. In regards to finding data, some challenges faced were gaining access, finding a reliable source, relevance, and feature/response richness. Cleaning the data is also very difficult, depending on the dataset, and involves careful maneuvering to balance features, observations, quality, scope, and manageability. Also, in cleaning the data, a lot of research had to be done into each of the survey types and project theme of stress in order to make sense of the data so that strategies to clean it could be correctly implemented. Some of the cleaning challenges were mentioned in section 2.3 above. Next, EDA was found to be an important step that could help understand the data and useable features. Building the models in this case were somewhat challenging as stress is an abstract/subjective concept that is difficult to assign a numeric value to and categorize. There were was no clear direction on how to build the models so there was a lot of trial and error on this. In addition, the metrics and tuning parameters for model improvement also involved a lot of trial and error. Finally, creating the user interface prototype took more time and brainstorming than anticipated to figure out how to make it the most user friendly as possible so that people could actually use the developed tool in the future.

Overall over the course of the semester, this project allowed for skill development not only in technical areas but also in teaming and project management; all of which are valuable in many industries and across disciplines. The biggest lesson learned was that taking a general idea, turning that idea into a vision, and developing something representing that vision is a very difficult thing to do with a lot of uncertainty and room for improvement. Ultimately, being able to produce a product that not only worked, but that gave insights into the effects that students' demands and habits have on stress and overall health was rewarding.

4.2. Future Exploration

Moving forward this project could add in more data to not only reinforce the data models with more data points, but would also provide a more general view of students. Continuing similar studies to the Dartmouth one are being done at the University of Texas - Austin and Northeastern University, both of which could be potential sources of additional data. Additionally, the group that created the original dataset used is conducting a more recent and more in depth study that could be used to see if anything has changed over the past few years given differences in how post secondary education has changed the mindset of young people.

Given the importance of this topic in recent years, this project also found that companies are also conducting similar studies within the workforce. As such, it would be valuable to be able to expand the tools created in this project to fit the general population and even make comparisons to see how workers perform and their stress levels, across various companies, industries, and careers. The inclusion of a workforce dataset would also allow for more exploration to see whether age, location, etc. has an effect on an individual's stress and wellbeing.

In order to improve on the current implementation, additional data points and variables to the model can be added in order to increase accuracy and consider any outlying factors that the current dataset fails to cover. This could be done by searching for more datasets across universities and companies and even conducting the surveys ourselves using the online platform and feedback. Also building upon the academic performance prediction, optimization of stress/performance can be achieved via more data, research, and time. Lastly, improving the website's UI and marketing the product to the general student population via ads will indirectly help increase the accuracy of the models.

5. Appendix

5.1. Exploratory Data Analysis (EDA) Graphs and Charts

Figure 5.1.1. Feature Importance

LASSO

	feature	weight
0	vr_12_ment	-0.817026
1	flour	-0.073805
2	vr_12_phys	0.039093
3	Conscientiousness	-0.735498
4	Active	-0.275755
5	Alert	-0.956694
6	Neuroticism	0.413749
7	Upset	-0.000000
8	Nervous	0.000000
9	phq_total	0.144005
10	Hostile	0.798189

Correlation Matrix

type	level_0	type	post
0	vr_12_ment	post	-0.765781
1	Interested	post	-0.520546
2	Active	post	-0.507939
3	Alert	post	-0.434042
4	flour	pre	-0.411800
5	Conscientiousness	post	-0.387363
6	Attentive	post	-0.386443
7	vr_12_ment	pre	-0.380732
8	vr_12_phys	post	-0.351661
9	flour	post	-0.349798

type	level_0	type	post
0	Hostile	post	0.431651
1	Neuroticism	pre	0.461548
2	Distressed	post	0.476272
3	Nervous	post	0.478903
4	Upset	post	0.541151
5	Neuroticism	post	0.581124
6	phq_total	post	0.632501
7	stress	pre	0.634861
8	phq_total	pre	0.699325

Figure 5.1.2. Perceived Stress Pre/Post

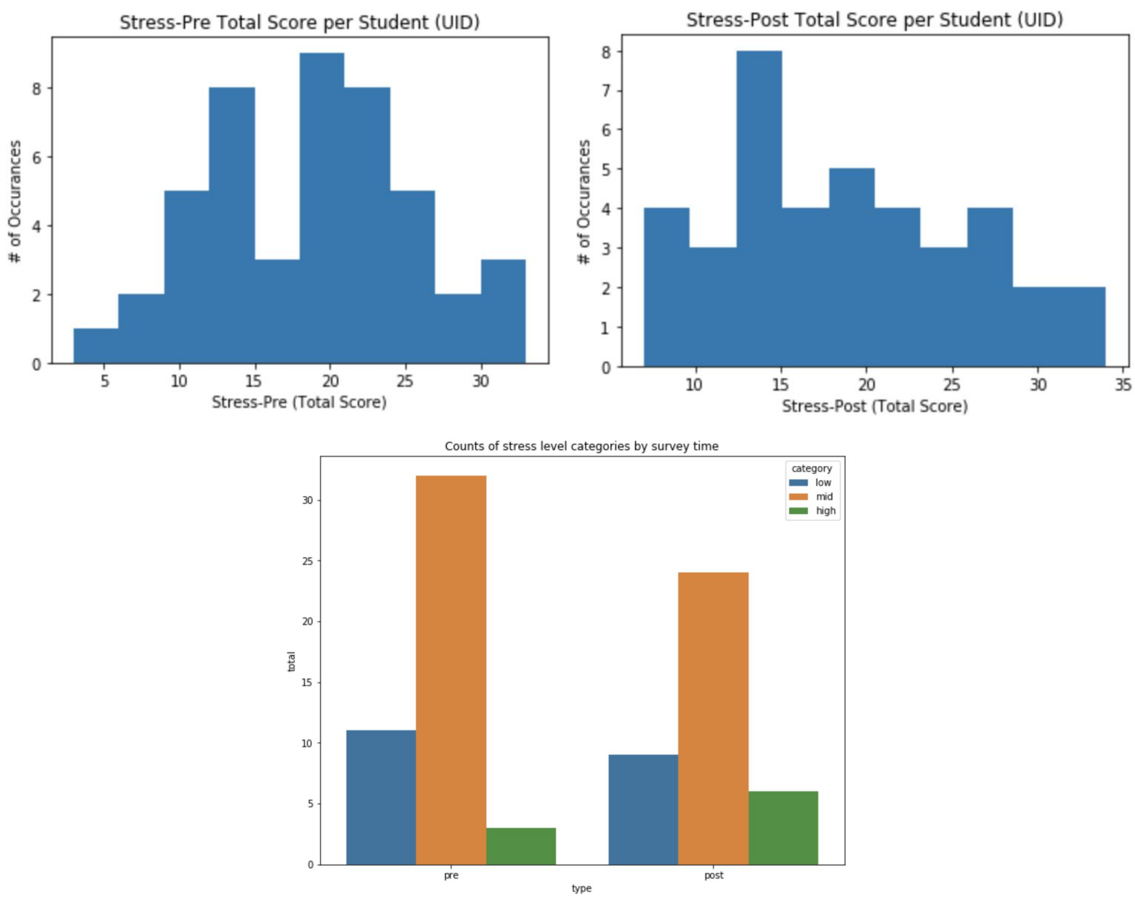


Figure 5.1.3. Stress Levels via App/Sensor

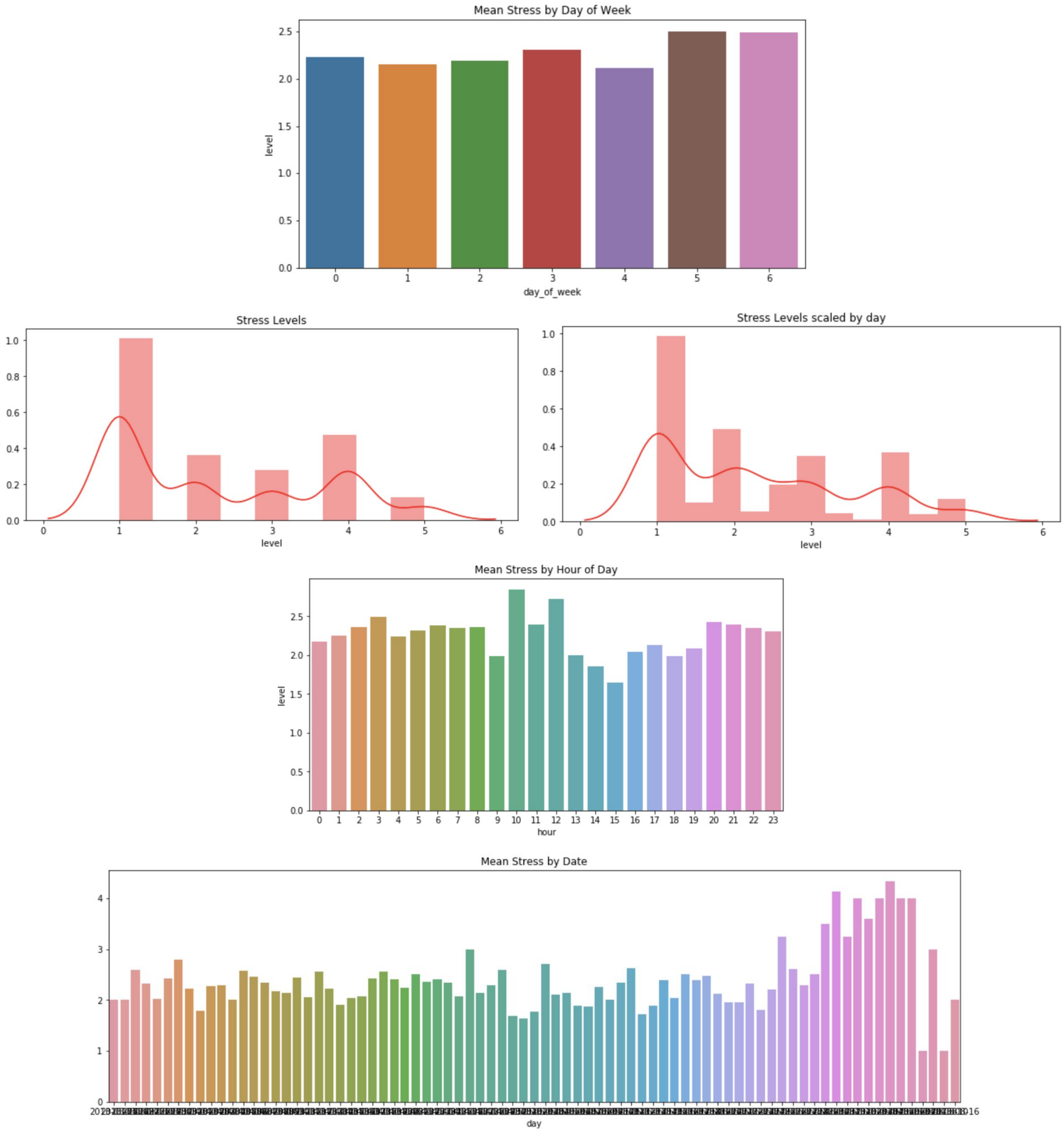


Figure 5.1.4. Activity Levels

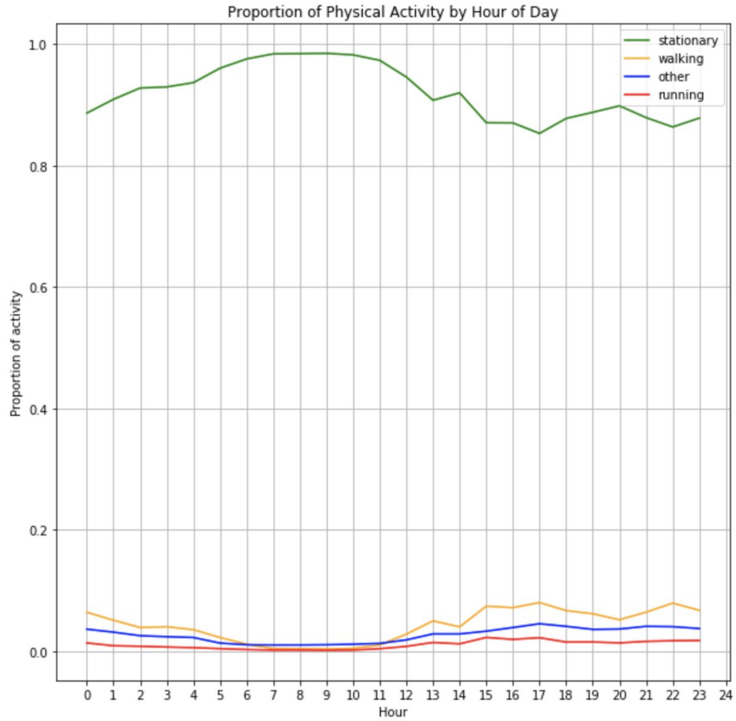


Figure 5.1.5. Stress Mapping

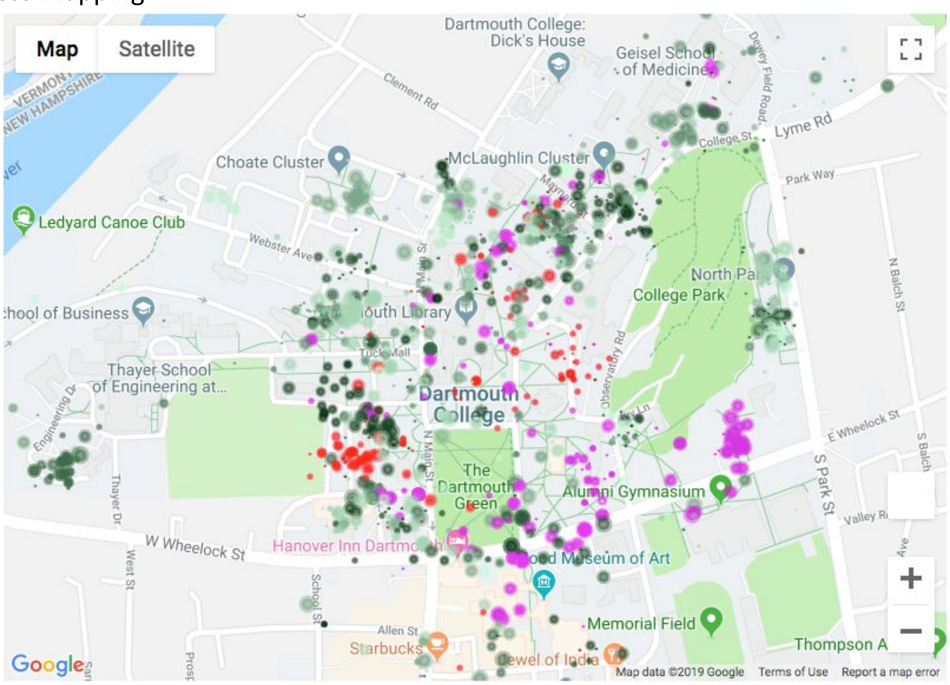


Figure 5.1.6. Flourishing Scale Pre/Post

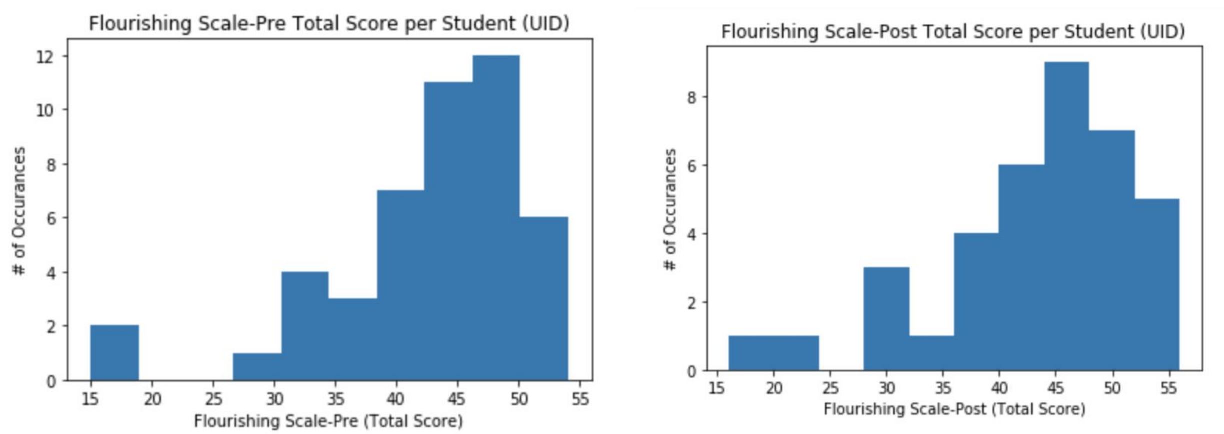


Figure 5.1.7. Deadlines

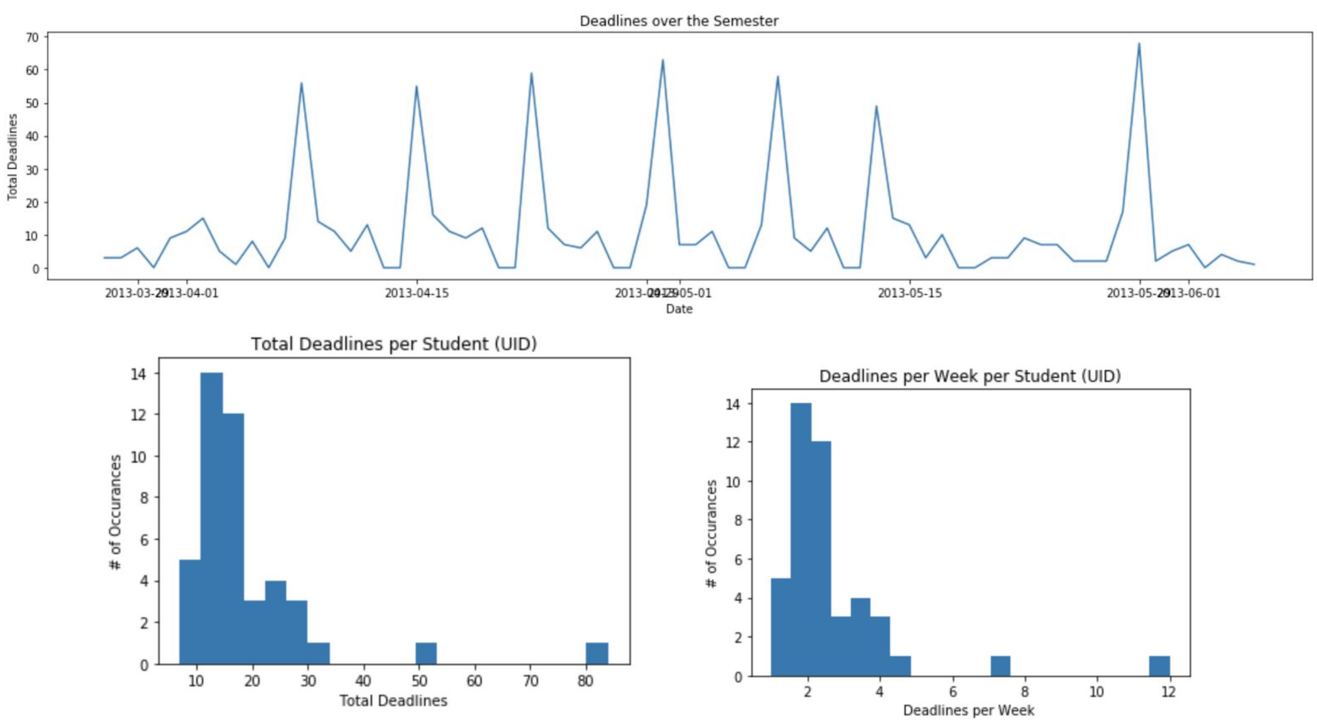
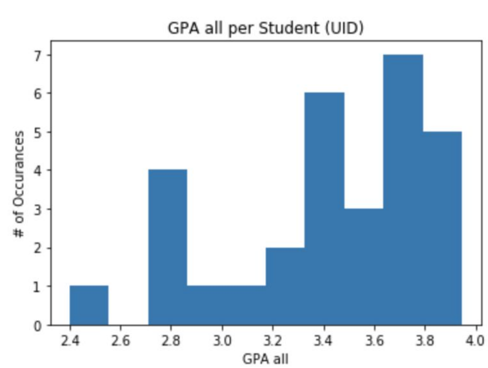


Figure 5.1.8. GPA



5.2. Model Performance Tables

Table 5.2.1. Linear Regression (Stress Prediction) Performance

Model	Type	Hyperparameter	MSE	MAE
Linear Regression	All Survey		19.035	3.602
		Lasso Alpha = 0.004	15.792	3.713
	Feature Reduction		15.877	2.985
		Lasso Alpha = 0.17	14.420	2.779

Table 5.2.2. Logistic Regression (Classification of Stress) Performance

Model	Type	Hyperparameter	Accuracy	Notes
Logistic Regression	All Survey	C = 1	72%	No predictions were over 1 off of the true value. Test: 0.2 Train: 0.8
	Feature Reduction	C = 10	73%	Predictions were skewed toward the middle. Test: 0.2 Train: 0.8

Table 5.2.3. Correlation Matrix created for feature reduction (partial table)

		flour		lonely		Interested		Distressed		Upset		Strong		Guilty	
	type	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre	post	pre
	type														
flour	post	1	0.549	-0.467	-0.423	0.124	0.365	-0.347	-0.0189	-0.31	0.0645	0.165	0.00643	0.128	-0.0512
	pre	0.549	1	-0.527	-0.424	0.4	0.269	-0.282	-0.0697	-0.369	0.109	0.248	0.0998	-0.133	0.153
lonely	post	-0.467	-0.527	1	0.802	-0.298	-0.298	0.501	0.0667	0.132	-0.088	-0.274	-0.147	0.0213	-0.0867
	pre	-0.423	-0.424	0.802	1	-0.273	-0.0694	0.469	0.163	0.157	0.00804	-0.231	-0.124	-0.0109	-0.105
Interested	post	0.124	0.4	-0.298	-0.273	1	0.239	-0.377	-0.031	-0.403	0.058	0.42	-0.152	-0.00408	0.336
	pre	0.365	0.269	-0.298	-0.0694	0.239	1	-0.193	0.0803	-0.104	-0.000731	0.0217	0.432	0.0466	-0.0342
Distressed	post	-0.347	-0.282	0.501	0.469	-0.377	-0.193	1	0.287	0.68	0.0894	-0.276	-0.0433	0.0908	-0.326
	pre	-0.0189	-0.0697	0.0667	0.163	-0.031	0.0803	0.287	1	0.27	0.741	-0.061	-0.0326	0.132	0.194
Upset	post	-0.31	-0.369	0.132	0.157	-0.403	-0.104	0.68	0.27	1	-2.54e-17	-0.283	0.0645	0.0663	-0.177
	pre	0.0645	0.109	-0.088	0.00804	0.058	-0.000731	0.0894	0.741	-2.54e-17	1	0.0694	-0.103	0.102	0.169
Strong	post	0.165	0.248	-0.274	-0.231	0.42	0.0217	-0.276	-0.061	-0.283	0.0694	1	0.133	0.00826	0.15
	pre	0.00643	0.0998	-0.147	-0.124	-0.152	0.432	-0.0433	-0.0326	0.0645	-0.103	0.133	1	0.24	-0.0891
Guilty	post	0.128	-0.133	0.0213	-0.0109	-0.00408	0.0466	0.0908	0.132	0.0663	0.102	0.00826	0.24	1	0.195
	pre	-0.0512	0.153	-0.0867	-0.105	0.336	-0.0342	-0.326	0.194	-0.177	0.169	0.15	-0.0891	0.195	1
Scared	post	-0.067	-0.417	0.0975	0.176	-0.305	-0.0224	0.198	0.294	0.357	0.143	-0.101	0.0973	0.388	0.0544
	pre	-0.129	-0.0111	-0.156	-0.0867	0.252	0.168	-0.103	0.356	0.0377	0.463	0.112	0.0888	0.0163	0.405
Hostile	post	-0.311	-0.554	0.422	0.361	-0.239	-0.222	0.393	0.323	0.442	0.19	0.0708	0.258	0.0526	-0.251
	pre	0.118	0.134	-0.00203	-0.0695	0.186	0.256	-0.0156	0.435	3.11e-17	0.47	0.141	0.362	0.231	0.453
Enthusiastic	post	0.0445	0.0266	-0.146	0.0663	0.459	0.305	-0.166	0.0379	0.0154	-0.0312	0.251	0.216	-0.151	0.0668
	pre	0.229	0.229	-0.366	-0.178	-0.0016	0.469	-0.136	-0.098	0.106	-0.0954	0.277	0.523	-0.138	-0.068
	post	0.244	0.206	-0.173	0.0871	0.533	0.25	-0.256	0.141	-0.199	0.0417	0.448	0.118	-0.0963	0.192

LASSO

	feature	weight
0	vr_12_ment	-0.817026
1	flour	-0.073805
2	vr_12_phys	0.039093
3	Conscientiousness	-0.735498
4	Active	-0.275755
5	Alert	-0.956694
6	Neuroticism	0.413749
7	Upset	-0.000000
8	Nervous	0.000000
9	phq_total	0.144005
10	Hostile	0.798189

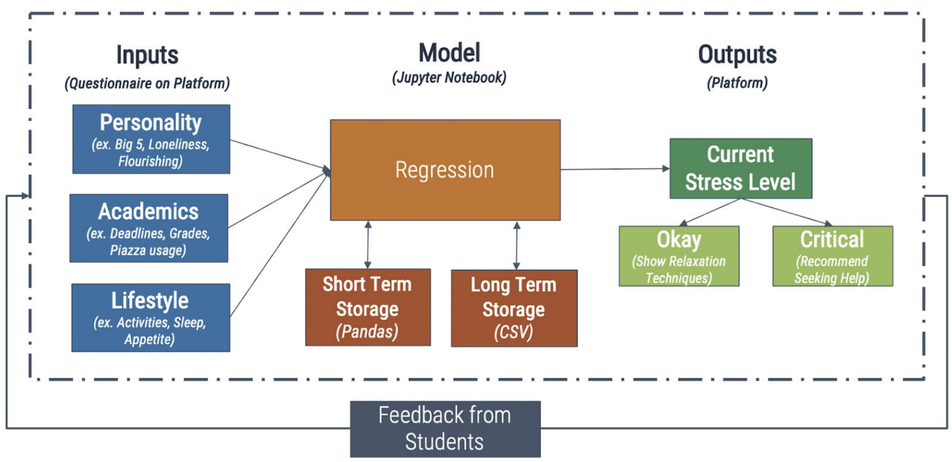
Correlation Matrix

type	level_0	type	post
0	vr_12_ment	post	-0.765781
1	Interested	post	-0.520546
2	Active	post	-0.507939
3	Alert	post	-0.434042
4	flour	pre	-0.411800
5	Conscientiousness	post	-0.387363
6	Attentive	post	-0.386443
7	vr_12_ment	pre	-0.380732
8	vr_12_phys	post	-0.351661
9	flour	post	-0.349798

type	level_0	type	post
0	Hostile	post	0.431651
1	Neuroticism	pre	0.461548
2	Distressed	post	0.476272
3	Nervous	post	0.478903
4	Upset	post	0.541151
5	Neuroticism	post	0.581124
6	phq_total	post	0.632501
7	stress	pre	0.634861
8	phq_total	pre	0.699325

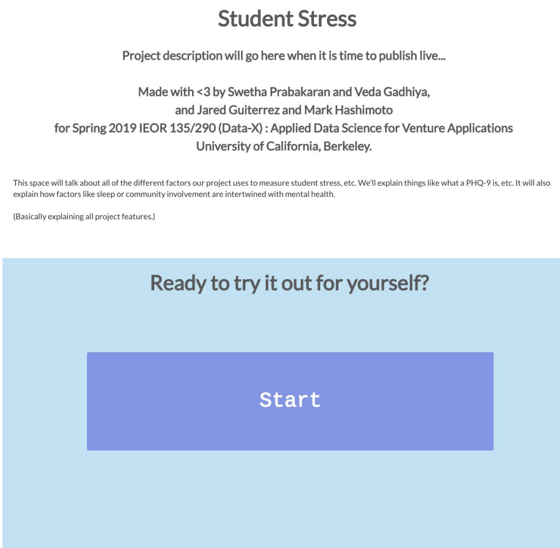
5.3. System

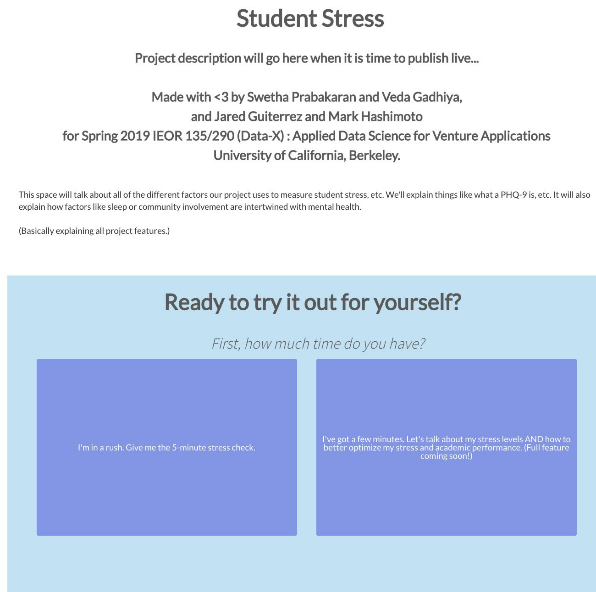
5.3.1. Overview



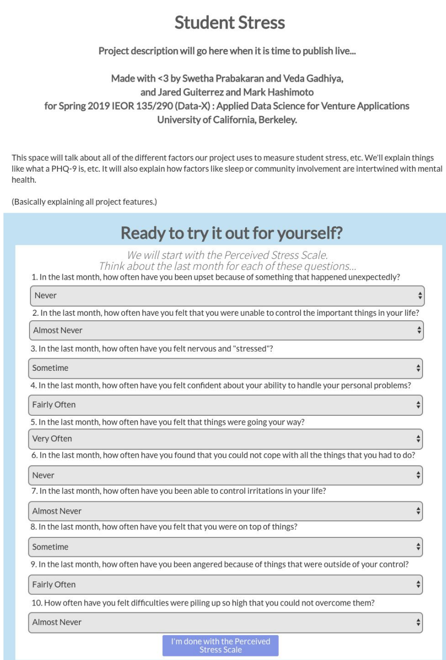
5.3.2. Prototype

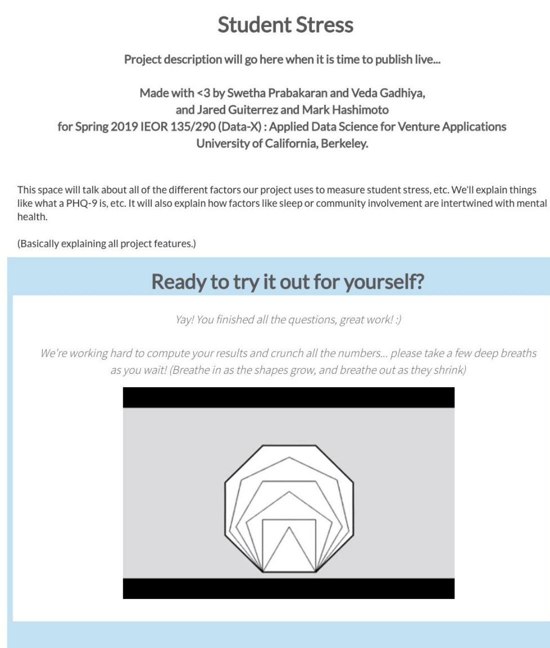
Steps 1-2) Enter the platform and Pick what you want to do





Steps 3-4) Answer Questionnaire and Get recommendation





5.4. Links

5.4.1. GitHub

- Current: https://github.com/mkhash/Modelling-Student-Stress
- Initial: https://github.com/vgadhiya/College-Student-Stress

5.4.2. Prototype

- https://dataxfinalproject.bubbleapps.io

5.4.3. Dataset

- Data: https://studentlife.cs.dartmouth.edu/dataset.html
- Paper Publication: https://dl.acm.org/citation.cfm?id=2632054

5.5. References

Bradberry, T. (2019, February 6). How Successful People Stay Calm. Forbes. Retrieved April 8, 2019, from https://www.forbes.com/sites/travisbradberry/2014/02/06/how-successful-people-stay-calm/amp/

Hess, A. (2018, October 04). Massive survey finds 1 in 3 college freshmen struggle with mental health. Retrieved April 8, 2019, from https://www.cnbc.com/2018/10/04/4-ways-to-be-proactive-about-your-mental-health-in-college.html

Fall 2018 Reference Group Executive Summary. American College Health Association National College Health Assessment. Retrieved April 8, 2019, from https://www.acha.org/documents/ncha/NCHA-II_Fall_2018_Reference_Group_Executive_Summary.pdf