Predicting Stress and Depression of College Students Using Machine Learning

Authors - Jeenal Shah, Tanishqua Dave Mentors - Sagar Kavaiya, Dr. Dhaval Patel, *Member, IEEE*

***Abstract*—This paper conducted a case study that how M.L. Model (Machine Learning) can help us predict Stress and Depression among college students. The acquired data is a result of 10 week-long longitudinal studies[1]. The longitudinal time period (whole semester) of data acquisition helped us to avoid the hypes (which could create biased results) faced by the student during their semester and able to predict the average stress and depression level during the semester. We did a correlation between factorial data sets (which was collected on a daily basis)and survey scale data sets (Conducted pre semester and post semester) and also applied Linear regression as the M.L. model. Hence, this helped us to understand college students Mental Health during the semester.**

***Index Terms*—Machine Learning Model, linear regression, correlation,Mental health, college students, stress, depression, factorial data sets, survey data sets.**

I. Introduction

A. Background

A survey conducted by Gregg Henriques (A psychology professor at James Madison University), shows that there is an increase in mental health issues (like Depression and Anxiety) in American College Students. This survey compares the data for the years 2007/08 and 2017, which is almost a decade and we found an increase in mental health issues.[2] Now, it is considered more than a crisis because it is increasing for a decade.

You might be thinking, the students must seek help from the universities mental health services by seeking therapies and medications from the experts. Yes, we can definitely take help from them but we are lacking at the initial stage and that is to recognize or identify that the students are facing particular mental health problems and another important step where they are generally misinterpreted is to consider the stressful situation a new normal because of peer pressure maybe. Once the problem is detected we can support them with available therapies and medications.

B. Motivation

For the last one and a half years, we are going through the Covid-19 Pandemic and this pandemic has made our work online which created a work from home situation. Working from home makes us feel isolated and lonely not only this shifting work online also creates a lot of challenges. Every field was struggling with there own problems similarly the education field was also suffering from the online atmosphere. Being a student I had an opportunity and exposure to understand the challenges faced by the educational field in detail. We found college students of Maharastra facing depressive symptoms and changes in their routine which created irregularity in their work and sleeping patterns which affect our stress levels.This motivated us to work on stress and depression levels faced by college students.[3]

C. Methodology

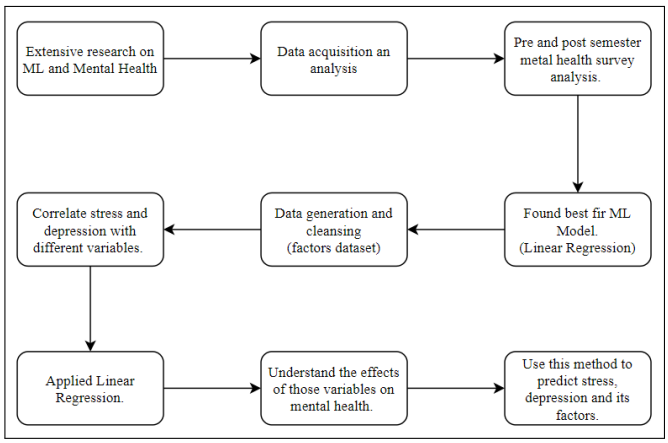

Fig. 1: Block Diagram of our work process

II. Data Acquisition, Generation and Cleansing

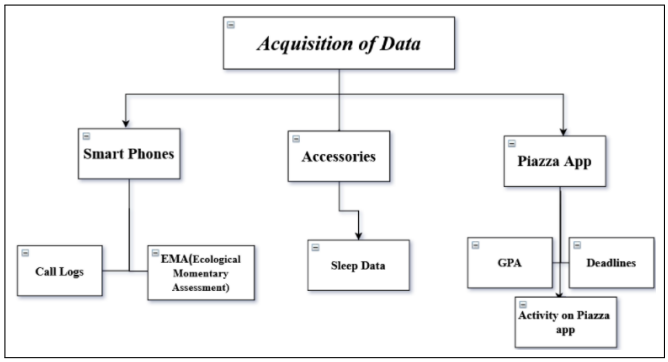

Fig. 2: Acquired Data

We acquired the data from the Dartmouth University website. The data was a result of a longitudinal study (of 10 weeks) of the spring semester. The dataset consists of two types of data, one is the survey data and another is the factorial data. The survey data is the result of PHQ-9 (depression survey), Perceived Stress Scale (PSS: Stress survey) and UCLA (Loneliness survey). The surveys were taken before and after the semester (pre and post semester). To find out students' mental state before and after is the same or not. Factorial data is the data collected while the semester was on and the

factors we selected and implemented while reproducing the results are EMA Stress data (Ecological Momentary Data), Activity on Piazza application and Sleep data. There were a total of 60 participants when they started the study but only 38 participants completed the study. So, we conducted analysis on the 38 participants who completed the study. We got average and total of 10 weeks data of every student for each factorial. After selecting the data and the factorials we did data cleansing, by removing the faulty responses.

III. Numerical Results

A. Work on Survey Data (PHQ-9, PSS and UCLA Scale)(Reproduced Results)

1) PHQ-9 data-set

Consists of 9 questions with 0-3 answer scale each.[5]

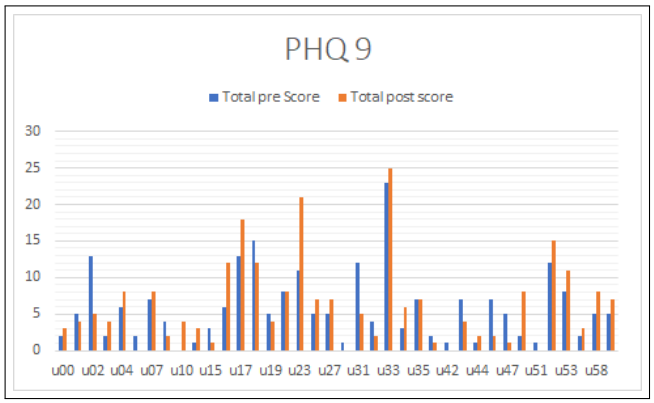

Fig. 3: Comparison of depression values before and after the semester

Figure 3 shows the graph of depression values in pre-semester and post-semester surveys.

Depression severity	minimal	minor	moderate	Moderately Severe	severe
Score	1-4	5-9	10-14	15-19	20-27

Depression	Pre Semester	Post Semester
Mean	5.521739	6.263158
Standard Deviation	4.612732	5.838753

Fig. 4: Scale and Comparison

Figure 4: Table 1 shows the scale of depression.
Table 2 shows the pre and post semester depression value comparison using mean and standard deviation. Here we could notice that depression is little higher in the post semester survey.

2) PSS data-set

Consists of 10 questions with 0-4 answer scale each.[4]
Figure 5 shows the graph of stress values in pre-semester and post-semester surveys.

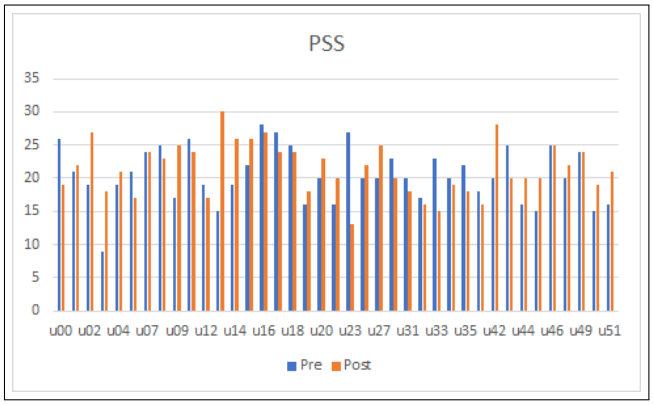

Fig. 5: Comparison of stress values before and after the semester

Stress Scale	low	moderate	high
Score	0-13	14-26	27-40

PSS	Pre Semester	Post Semester
Mean	18.42	18.9
Standard Deviation	6.8	7.1

Fig. 6: Scale and Comparison

Figure 6: Table 1 shows the scale of stress level.
Table 2 shows the pre and post semester stress value comparison using mean and standard deviation. Here we could notice that stress is little higher in the post semester survey.

3) UCLA Scale data-set

Consists of 20 questions with 1-4 answer scale each.[6]
Figure 7 shows the graph of loneliness degree values in pre-semester and post-semester surveys.

Figure 8: The table shows the pre and post semester loneliness degree comparison using mean and standard deviation. Higher the score, higher is the loneliness degree. Here we could notice that loneliness is little lower in the post semester survey.

4) Work and Inferences

The PSS, PHQ-9 and UCLA Scale survey data was available in the categorical form then we converted the data into numerical data and divided it as per the scale: None, Mild, Moderate, Moderately severe and Severe. After converting the primary data into secondary data, we compared the pre-semester and post semester data of PSS and PHQ-9 surveys.

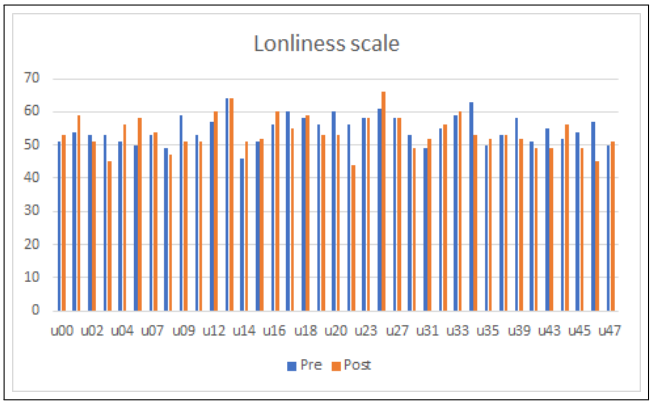

Fig. 7: Comparison of Loneliness Degree before and after the semester

Loneliness Scale	Pre Semester	Post Semester
Mean	54.23941018	53.38824773
Standard Deviation	4.122988457	5.095484664

Fig. 8: Comparison

PSS Scale consists of 10 questions with 0-4 answer scale each, PHQ-9 Scale consists of 9 questions with 0-3 answer scale each and UCLA scale consists of 20 questions with 1-4 answer scale each.

We got the total score of all participants in the survey data for pre and post semester surveys. Compared pre and post semester scores using mean and standard deviation. Depression and stress were seen a little higher and loneliness was seen a little lower in the post semester survey as compared to the pre semester survey.

B. Analysis Using ML Model

The factorial data which were collected on a daily basis are sleep data, call logs, active days on piazza application and Ecological Momentary Assessment Stress data. As these data were collected throughout the 10 week semester, there were many faulty data found because of lack of battery in participant's gadgets affecting the data, also EMA data were collected around several times a day (between 8 to 13 times). Due to the high frequency of EMA pop-ups, students used to miss the EMA's due to external factors(like battery low, forgot the phone in the room or might be busy with some other things). Hence, to correlate we need secondary data of these factorial data and to get secondary data we had to cleanse the data by recognizing and removing faulty data manually. After cleaning the data, we counted the average (of per participant) for EMA stress data and Sleep data of the students for data analysis. We did a total for the call logs and active days on the Piazza app to obtain secondary data. We also summed up the data of deadlines to correlate. The GPA data was already available in the secondary form.

Linear Regression is considered one of the best predictors for correlation. Hence, we used Linear Regression as our Machine Learning Model. These are the few independent variables which affect stress and depression the most.

1) Stress (EMA stress level), No. of calls
2) Stress, sleep hours
3) Stress, No. of deadlines
4) Stress, GPA
5) Stress, No. of active days on piazza
6) Depression (PHQ-9), No. of calls
7) Depression, sleep hours
8) Depression, No. of deadlines
9) Depression, GPA
10) Depression, No. of active days on piazza

1) Stress Level and Sleep Hours

$R^2 = 0.004421174890784418$
Intercept = 1.9565409234292304
Regression Coefficient = [0.04329695]

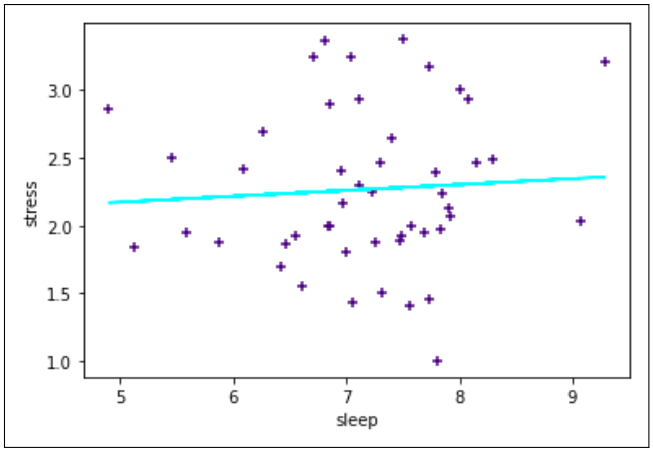

Fig. 9: Linear Regression between Stress and Sleeping Hours

Figure 9 shows Linear Regression between stress and sleep. Stress seems to increase with an increase in sleeping hours. We could say that with sleeping more time, more work is being compromised which increases the stress.

2) Stress Level and Call Logs

$R^2 = 0.007847019804197553$
Intercept = 2.067857624375249
Regression Coefficient = [0.00013672]

Figure 10 shows Linear Regression between stress and No. of Calls in three months. Stress is more with increasing number of calls. To cope with high stress social interaction strategy is used. The result we got shows the possibility that more stressed participants use this strategy to reduce their stress.

3) Stress Level and Deadlines

$R^2 value = 0.084677$
Intercept = 1.990443
Regression Coefficient = [0.014243]

Figure 11 shows Linear Regression between stress and Deadlines. Here we could notice that participants having

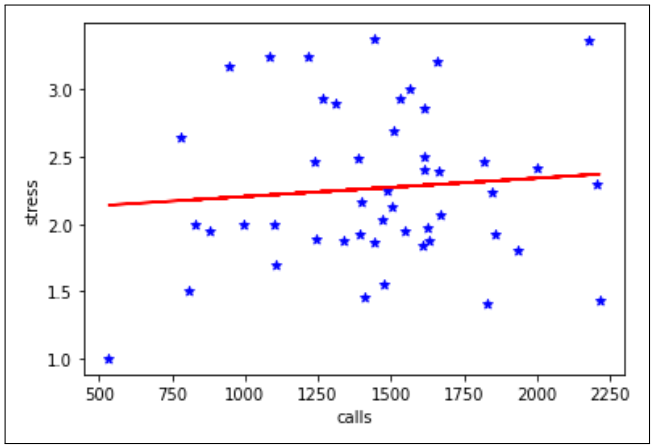

Fig. 10: Linear Regression between Stress and No. of Calls

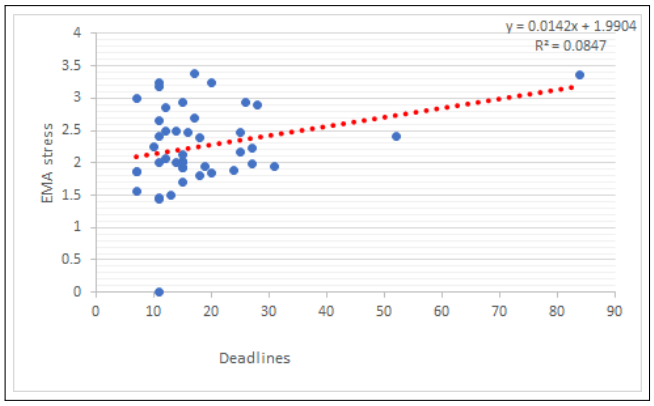

Fig. 11: Linear Regression between Stress and No. of Deadlines

more deadlines are more stressed. Stress hormones are released when we try to meet the deadlines. Hence, more deadlines cause more stress.

4) Stress Level and GPA

$R^2value = 0.023442$

Intercept = 3.137086

Regression Coefficient = [0.120776]

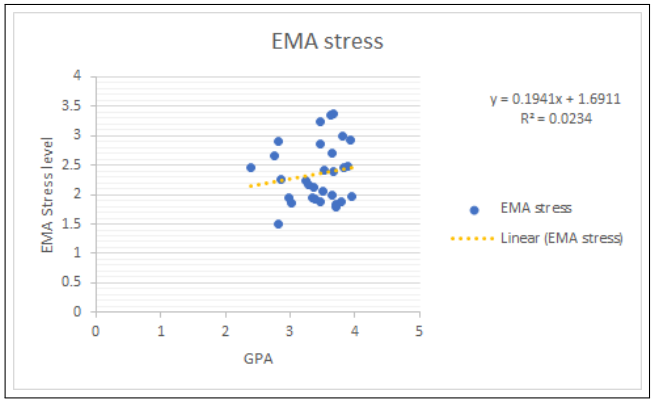

Fig. 12: Linear Regression between Stress and GPA

Figure 12 shows Linear Regression between stress and GPA. With an increase in GPA, the stress levels seem to rise. This could be due to studying for a longer time which causes more stress.

5) Stress Level and No. of active days on Piazza

$R^2value = 0.0071$

Intercept = 2.3375

Regression Coefficient = [-0.0028]

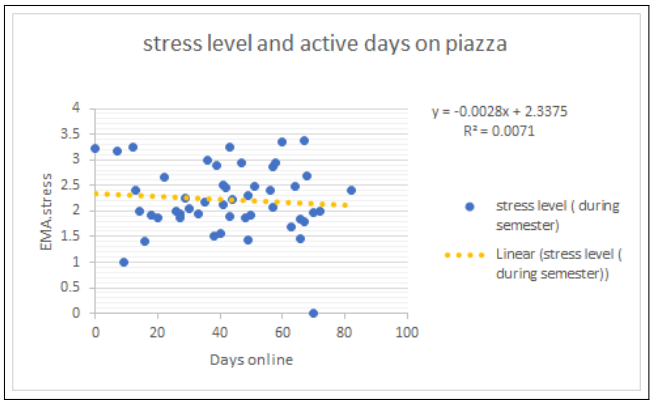

Fig. 13: Linear Regression between Stress and No. of active days on Piazza

Figure 13 shows Linear Regression between stress and No. of active days on Piazza. The stress level seems to decrease with more active days on the piazza. This could be because less stressed students show higher degrees of activeness.

6) Depression and Sleep Hours

$R^2value = 0.093065$

Intercept = 20.11626

Regression Coefficient = [-1.93661]

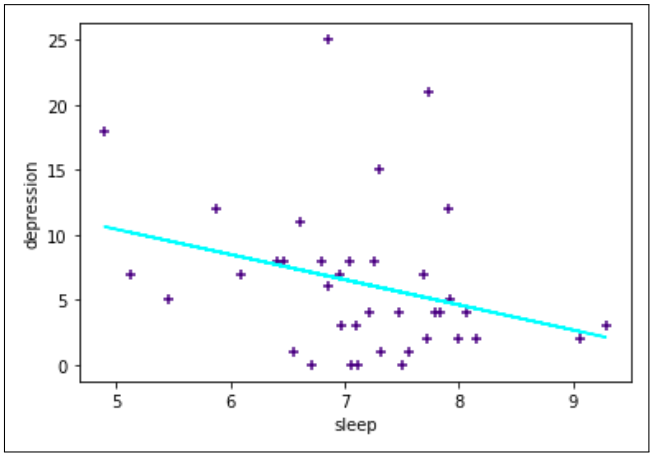

Fig. 14: Linear Regression between Depression and Sleeping Hours

Figure 14 shows Linear Regression between depression and sleep. Depression is seen less in participants who sleep for more hours. This could be because depressed people could have trouble falling asleep. This could also mean that sleep deprivation may lead to depression.

7) Depression and Call Logs

$R^2value = 1.5E-05$

Intercept = 6.165855

Regression Coefficient = [0.005907]

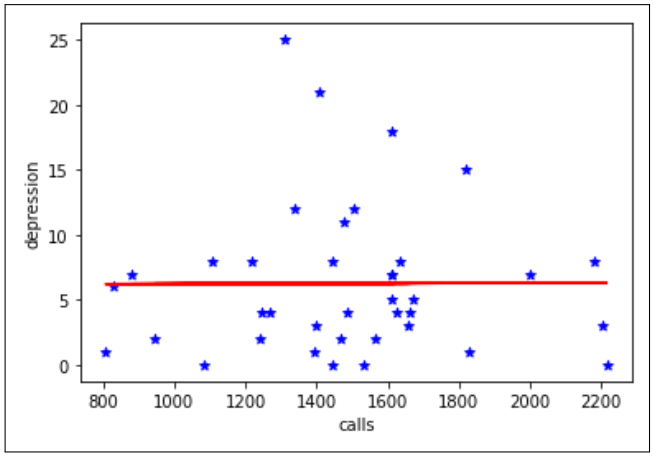

Fig. 15: Linear Regression between Depression and No. of Calls

Figure 15 shows Linear Regression between Depression and No. of Calls in three months. The more depressed participants show a higher number of calls. This could be because more depressed students seek more social affection.

8) Depression and Deadlines

$R^2value = 0.004599$

Intercept = 6.153554

Regression Coefficient = [0.028266]

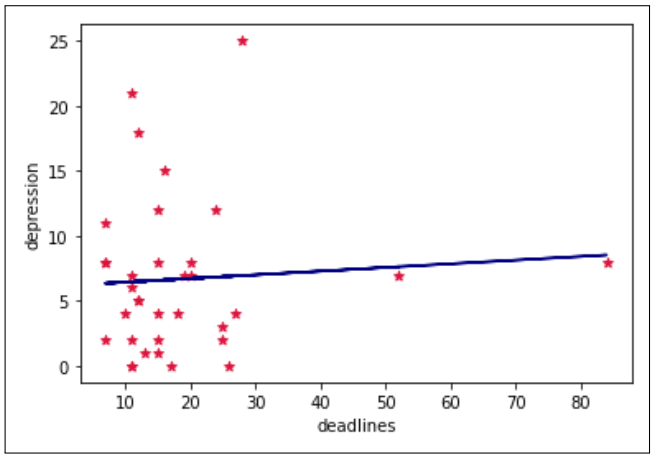

Fig. 16: Linear Regression between Depression and No. of Deadlines

Figure 16 shows Linear Regression between depression and Deadlines. Here we could notice that participants having more deadlines are more depresses. Stress hormones are released when we try to meet the deadlines. Hence, more deadlines cause more stress and depression.

9) Depression and GPA

$R^2value = 0.22092$

Intercept = 30.64107

Regression Coefficient = [-6.90694]

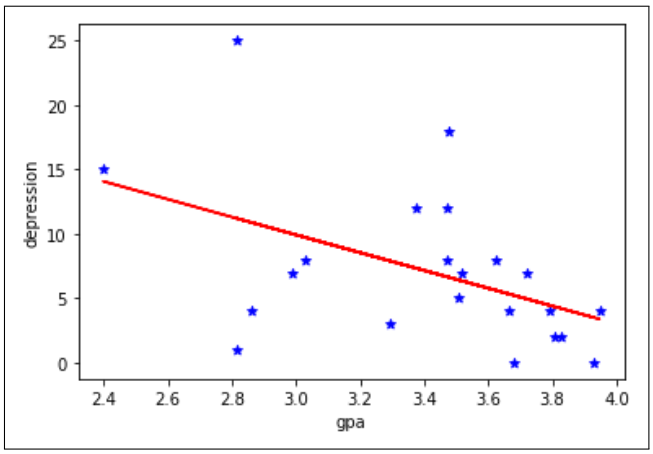

Fig. 17: Linear Regression between Depression and GPA

Figure 17 shows Linear Regression between stress and GPA. With an increase in GPA, the stress levels seem to rise. The participants who got a higher GPA showed lower levels of depression. This could be because lower GPA could cause more depression.

10) Depression and No. of active days on Piazza

$R^2value = 0.0497$

Intercept = 4.0359

Regression Coefficient = [0.0696]

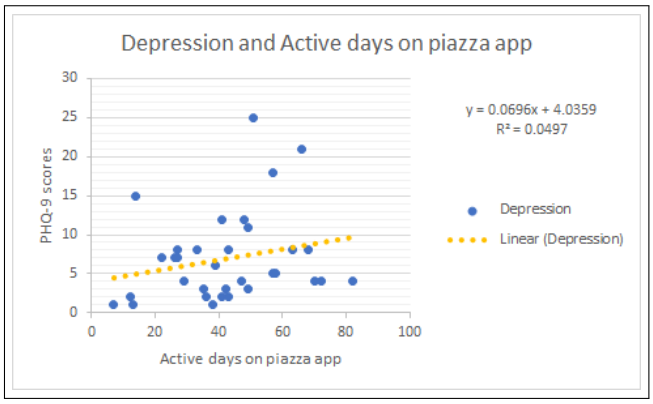

Fig. 18: Linear Regression between Depression and No. of active days on Piazza

Figure 18 shows Linear Regression between Depression and No. of active days on Piazza. The depression seems to increase with more active days on the piazza. This could be because more work load could cause more depression.

IV. Conclusion

In our paper, we have tried to understand stress and depression levels of the students using Machine Learning. Initially

we tried to understand the difference between the pre-semester survey and post-semester survey values. We generated some data-sets using the acquired data sets to perform Linear regression on them and understand which factors affect stress and depression. The independent factors used to correlate with stress are average Sleep Hours in three months, No. of calls in three months, no. of deadlines in three months, GPA and No. of active days on piazza in 3 months of each participant. The results and inferences we achieved after applying Linear Regression could be highly helpful for understanding the stress and depression reasons for students and how they try to handle it.

REFERENCES

1) Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., . . . Campbell, A. T. (2017). StudentLife: Using Smartphones to Assess Mental Health and Academic Performance of College Students. Mobile Health, 7-33. doi:10.1007/978-3-319-51394-2-2 https://www.andrew.cmu.edu/user/fanglinc/pubs/ubicomp14-studentlife.pdf

2) Henriques, G., 2021. The College Student Mental Health Crisis (Update). [online] Psychology Today. Available at: https://www.psychologytoday.com/us/blog/theory-knowledge/201811/the-college-student-mental-health-crisis-update

3) Moghe, K., Kotecha, D., amp; Patil, M. (2020). COVID-19 and Mental Health: A Study of its Impact on Students in Maharashtra, India. https://doi.org/10.1101/2020.08.05.20160499

4) State of New Hampshire Employee Assistance Program. (n.d.). Perceived Stress Scale. https://doi.org/https://das.nh.gov/wellness/docs/percieved

5) INSTRUCTION MANUAL. (n.d.). Instructions for Patient Health Questionnaire (PHQ) and GAD-7 Measures . https://doi.org/https://www.pcpcc.org/sites/default/files/resources/instructions.pdf

6) Russel , D. W. (1996). UCLA Loneliness Scale. https://doi.org/https://depts.washington.edu/uwcssc/sites/default/files/hw00/d40/uwcssc/sites/default/files/UCLA

7) CHIKERSAL, PRERNA (2021). Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection. https://doi.org/https://dl.acm.org/doi/pdf/10.1145/3422821

8) Theme, A. (2020). Machine Learning in Mental Health: A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems . https://doi.org/https://dl.acm.org/doi/pdf/10.1145/3398069