Student Behavior Analysis Using Data Mining Techniques

Harish Gonnabattula
University of California Riverside
hgonn001@ucr.edu

Vishal Surya Madhavan
University of California Riverside
vmadh001@cs.ucr.edu

Achyuth Madhav Diwakar
University of California Riverside
adiwa001@cs.ucr.edu

ABSTRACT

Stress levels in students seemed to be on the rise through the years. Succumbing to the pressure and sacrificing time for recreation or sleep has skyrocketed the stress levels when they are occupied with deadlines of homework and projects. The addiction towards email and continuous cell-phone usage is also one of the factors. According to a study conducted by Dartmouth College, they were able to understand how mental health, academic performance and behavioral trends changed as the term progressed with the data collected from the sensors of the mobile phones to understand the relations among sleep , mood , sociability, activity, stress levels .

KEYWORDS

Behavior analysis, Student health, StudentLife, mobile sensing, Perceived Stress Scale(PSS), College Students, Mental health

1. INTRODUCTION

Smartphones offer the hope that depression can be detected using passively collected data from the phone sensors. The aim of this study to identify We show how passive sensing data from phones can infer mental well being of students for the term. We collected 5 days of data for 18 participants: a wrist sensor (accelerometer and skin conductance), mobile phone usage (call, short message service, location and screen on/off) and surveys (stress, mood, sleep, tiredness, general health, alcohol or caffeinated beverage intake and electronics usage). We applied correlation analysis to find statistically significant features associated with stress and used machine learning to classify whether the participants were stressed or not. In comparison to a baseline 87.5% accuracy using the surveys, our results showed over 75% accuracy in a binary classification using screen on, mobility, call or activity level information (some showed higher accuracy than the baseline). The correlation analysis showed that the higher- reported stress level was related to activity level, SMS and screen on/off patterns.

2. OBJECTIVE

To design a model that uses passive sensing data and self-reports from students' smartphones to understand various factors that affect the students' stress level over the period of 10 weeks and to find correlations between stress scores and smartphone usage as well as various sensor data, pointing towards innovative ways for automatic stress measurements. Given a student's activities our model outputs the mental state of the student or rather show how a given activity can affect his mental health.

3. LITERATURE SURVEY

[1]Mental health in students is a very sensitive and important topic. Not only have rates of stress skyrocketed in United States in the last year, but a greater number of them are students. Subjecting students to great stress has lead them to depression which is a major cause for suicide. Long-term conditions with high stress can be chronic and people may be less likely to notice whether they are under high stress or may be generally less sensitive to stressors. Stress detection technology could help people better understand and relieve stress by increasing their awareness of heightened levels of stress that would otherwise go undetected.

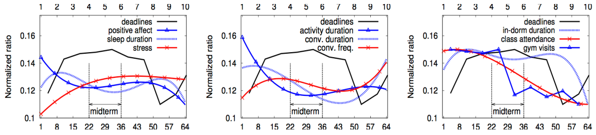


Figure 1. Previous work

There are various factors that could contribute to a student's mental health and stress levels in a single academic term. Particularly, the progress of the academic term and the physical activity of the student could have a marked effect on his/her mental health and behavior[2]. The article takes two factors into consideration: physical activity behaviors and Sociability behaviors. The authors take issue with the fact that the most of the data gathered by the existing studies on physical activity is self-reported and not sensor based. More importantly the guidelines of the American College Health Association are used as a yardstick to evaluate the physical activity of the students. The key idea is that unlike previous research, the usage of mobile sensor data would help to actually get data with a low bias and in real-life situations rather than a controlled experimental setup. Further research was then done by creating latent curve growth models, both conditional and unconditional for the physical activity and sociability. It is then concluded that a set of students tend to have decreased physical and social activity in the initial stages of the term with an increase in the later stages. More importantly this paper stresses on the advantages obtained by using mobile sensor data.

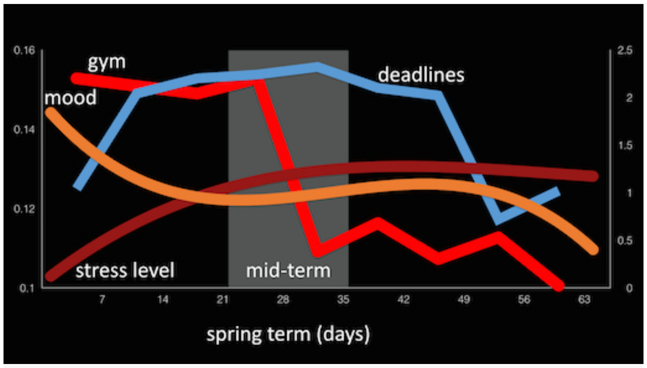


Figure 2. Indicating the relationship between mood of the student and deadlines.

[7]Aims to detect the signs of depression based on the data obtained from the mobile sensors of the Dartmouth College students. The PHQ-9 questionnaire (Patient Health Questionnaire 9-item) is used as a self report measure of the severity of the subjects' depression. The research also takes into account the location features, specifically the time and place where the students have been over a ten week period. The differences in weekends and weekdays is also taken into account. [7]Linear correlations between location variance, Circadian movement, Speed mean, speed variance, total distance, etc and the PHQ-9 scores are computed in order to visualize the relationships between those attributes. The article stresses on the separation of workday and non-workday data separation in the student location data. The researchers observed significant differences due this separation as it resulted in depressions indications on weekend data but not weekdays. The limitations of the dataset and the models chosen are then discussed in detail. However, even in this study the authors advocate the use of mobile sensor data due its low bias advantages and also point out that with the availability of much larger datasets of a similar structure, the models may produces better results.

[8]Discusses about the mixed effects linear modeling of the sensor derived geospatial data and the sleep levels and the variability in geospatial data with the stress levels.

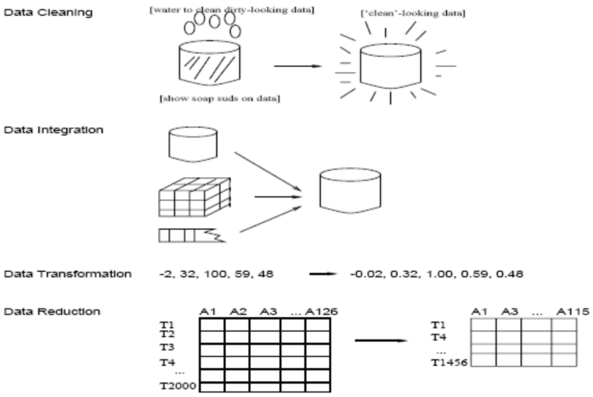


Figure 3. Indicating the data cleaning process.

4. LIMITATIONS

A limitation of this work is that it does not consider individual differences; for example, extracurricular activities could make some students happy or be stressful for other students. Another limitation of the models used to predict the mental state is they aren't considering actions that have long term effect on the behavior. Our model seeks to implements few of these feature evaluators and classifiers on the data set and classify data. The student sample that has been considered is a predominantly male and the student cohort is about 48 students over a ten week term. The sample space selected is very small and is particular to a particular Ivy league university and a particular ten week term. So the stressors experienced might be different compared to students at larger university with less rigor and academic demand. The sensors might incorrectly infer certain micro behaviors and especially when the phone is kept on the table or inside the bag and may come to wrong conclusions that a person is involved in a conversation while he might be engaged in another activity.

5. IMPLEMENTATION

The data used in this study was obtained from the StudentLife Dataset from Dartmouth College. This consists data of 48 students over a period of 10 weeks. The key process in this project was to the various parameters in this dataset namely mobile phone usage, sleep patterns, academic deadlines, physical activity and a few others to relate to the self-declared and sensor obtained stress levels of the students. The stress was also quantified from the responses of the students to obtain a numerical value that the model could be trained upon. A randomized sample of data is taken each time to obtain a fair distribution of data points for the training algorithm. This is performed to check the unbalanced nature of the dataset.

The primary challenge faced during the initial phases of this work was that feature selection. The dataset is a very comprehensive set of various parameters in the daily lives of the students at Dartmouth College. The approach followed to select the features and predict the stress of the students is discussed in the further sections.

5.1 Data Pre-Processing

Data in the real world is dirty and incomplete: with missing attribute values, lack of certain attributes of interest, or containing only aggregate data.

5.1.1 Data Cleaning

The StudentLife dataset, though was cleaned up Dartmouth college, still had many missing values for various attributes of multiple students. To handle the missing values in the data we used the attribute mean for all the samples belonging to the same class as the given tuple.

Ex: In our data we were missing some important samples of students who joined the survey but had to leave in between. As a result the data which was calculated over a period of 10 weeks was not consistent and so we took the mean values of that particular attribute of all students and used it to fill the missing ones.

We found that most of our data were not in the desired form for data mining. Values were not of same types and range. To get the data suitable for applying data mining methods we used data transformation.

Ex: PHQ-9 [11] survey results were a list of Q&A where each value was assigned a specific weight. We mapped the answers to the values and converted the survey into a range of 0-27 in the order of less stressed to most stressed.

5.1.2 Feature Selection

Correlations were computed for each of the latent variables which helped to determine the features that have a profound effect on the stress levels of the students. The sensor and phone data were analyzed and the following features were extracted from them:

- Phone usage/screen on-off time
- SMS count
- Call durations
- Mobility
- Geo spatial data
- Duration of sleep
- Class Deadlines

Different feature evaluation methods like Linear correlation coefficients (r), Greedy measure etc were calculated between these features and Mean Stress values. Significant features were identified and various classification techniques are applied on them. The resultant classification is compared against the Mean Stress and the accuracy is measured.

5.1.3 Training Data

To train our model we considered a split of 80-20%. The students were randomly selected and their data was used to train the model. We tried changing the ratio of the train and test set and we found that we had a best accuracy at a split of 80-20.

The training data of was an integration of all the features affecting the mean stress levels. The features were identified in the previous section. Data was read from each individual file and an outer join was performed to integrate the attributes. This operation clubbed all the students data into one file which is our Training and Test file.

5.1.4 Stress Level Prediction

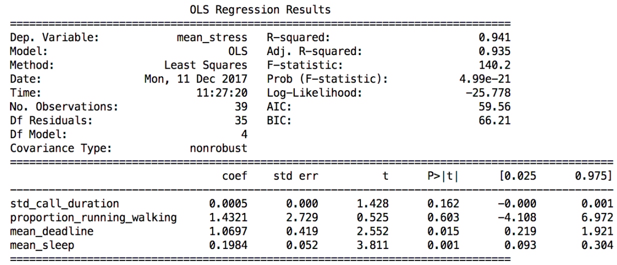
The stress of the students was indicated everyday for 60 days as a part of the data collection for the StudentLife. The stress levels are numbered as follows:

[1]A little stressed
[2]Definitely stressed
[3]Stressed out
[4]Feeling good
[5]Feeling great

The mean stress level of each student is required so that one can get a picture of the general stress pattern of the complete set of 48 students over the 10 week period. With the data splitting approach detailed in the previous section the linear regression model was trained with 80% of the mean stress values against the selected features that have a good correlation with the stress levels. The model then predicts the stress level of the remaining 20% of the students as a numerical value from 1 to 5. This value indicates the mean stress level of the student being tested and can be compared with the EMA questions to gauge the stress of the student. The predicted values and their comparison to the actual values in the dataset is discussed in further sections.

6. RESULTS AND DISCUSSION

Analyzing the results and outcomes from our respective data mining models:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            mean_stress   R-squared:                       0.941
Model:                            OLS   Adj. R-squared:                  0.935
Method:                 Least Squares   F-statistic:                     140.2
Date:                Mon, 11 Dec 2017   Prob (F-statistic):           4.99e-21
Time:                        11:27:20   Log-Likelihood:                -25.778
No. Observations:                  39   AIC:                             59.56
Df Residuals:                      35   BIC:                             66.21
Df Model:                           4
Covariance Type:            nonrobust
=============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------
std_call_duration           0.0005      0.000      1.428      0.162      -0.000       0.001
proportion_running_walking  1.4321      2.729      0.525      0.603      -4.108       6.972
mean_deadline               1.0697      0.419      2.552      0.015       0.219       1.921
mean_sleep                  0.1984      0.052      3.811      0.001       0.093       0.304
==============================================================================
```

• Multiple Linear Regression

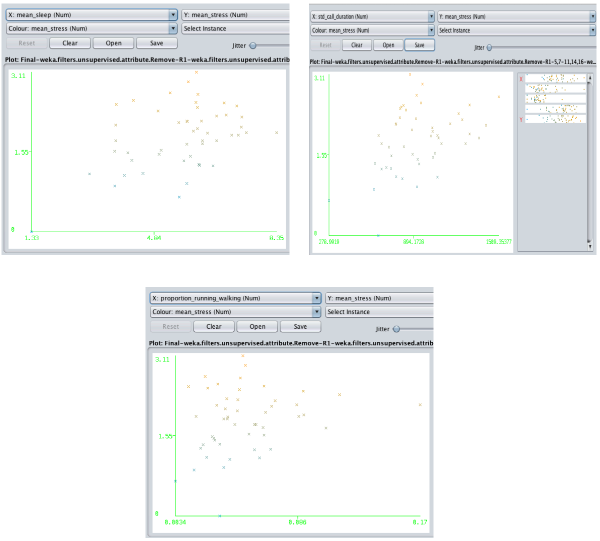
Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y. The population regression line for p explanatory variables x1, x2, ... , xp is defined to be y = 0 + 1x1 + 2x2 + ... + pxp. This line describes how the mean response y changes with the explanatory variables. The observed values for y vary about their means y and are assumed to have the same standard deviation . The fitted values b0, b1, ..., bp estimate the parameters 0, 1, ..., p of the population regression line.

Based on the definition of multiple regression above we formed a regression line for dependent variable mean_stress and explanatory variables std_call_duration, proportion_running_walking, mean_deadline, mean_sleep. We used linear_model from scikit library to design a model for our data. Training it using our test data we found the model Fig 3, to be 94% accurate.

Mean_Stress(y) = 0.0005* Std_call_duration + 1.4321*proportion_running_walking + 1.0697*mean_deadline + 0.1984*mean_sleep

Out of all the explanatory attributes we have found that mean_sleep is the most significant attribute of all($p < 0.005$). This has been verified in the previous works also.

The relation between the attributes was further verified with the help of the correlation graphs between them Fig 4.



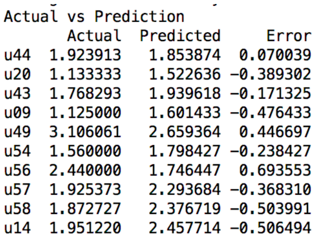
X: mean_sleep (Num)
Y: mean_stress (Num)
Colour: mean_stress (Num)
Select Instance
Plot: Final-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attrib

X: std_call_duration (Num)
Y: mean_stress (Num)
Colour: mean_stress (Num)
Select Instance
Plot: Final-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R1-5,7-11,14,16-we...

X: proportion_running_walking (Num)
Y: mean_stress (Num)
Colour: mean_stress (Num)
Select Instance
Plot: Final-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attrib

Figure 4: Correlation graphs

Now that we have established strong relation between the attributes and the dependent variable we can safely say that our model can be used to predict the output. We have done so in Fig 5 and you can find that the error range is very low.



```
Actual vs Prediction
        Actual  Predicted     Error
u44  1.923913   1.853874  0.070039
u20  1.133333   1.522636 -0.389302
u43  1.768293   1.939618 -0.171325
u09  1.125000   1.601433 -0.476433
u49  3.106061   2.659364  0.446697
u54  1.560000   1.798427 -0.238427
u56  2.440000   1.746447  0.693553
u57  1.925373   2.293684 -0.368310
u58  1.872727   2.376719 -0.503991
u14  1.951220   2.457714 -0.506494
```

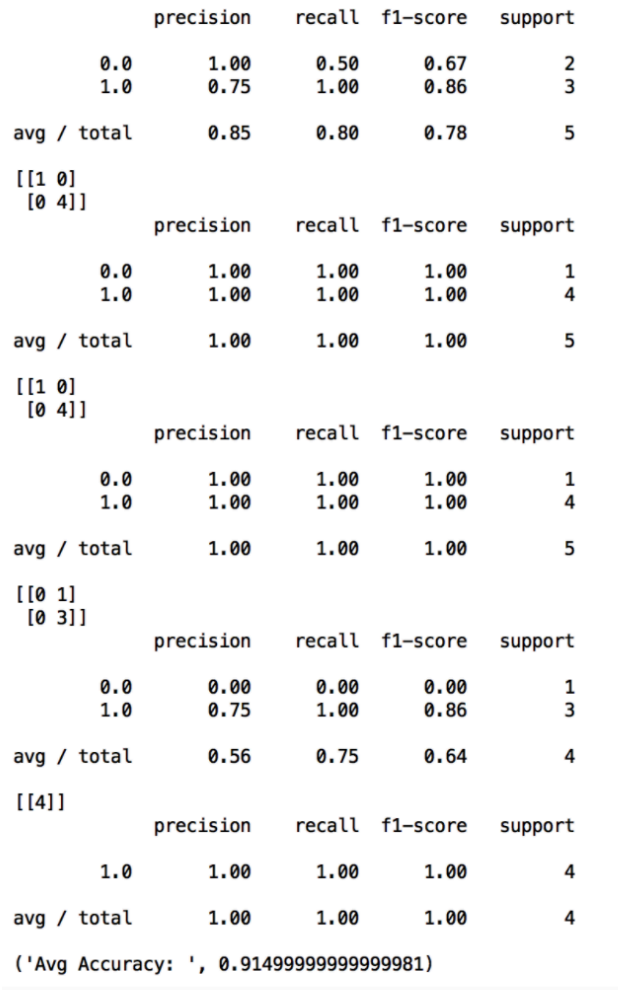
Figure 5: Actual vs Predicted

7. FUTURE WORK

The good accuracy in stress prediction obtained from the linear regression model is encouraging, however, the study calls for other classification models to be applied on the StudentLife dataset. The cornerstone of the linear regression model is that it assumes a linearly distributed data and makes the predictions based on those assumptions. Therefore it can be seen from the results obtained using said model that the incorrectly predicted values show drastic variation from their actual ones. This further emphasizes the need for testing other classification models with this dataset.

7.1 Random Forest

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way. The following technique was described in Breiman's original paper[12] and is implemented in the R package randomForest.[13] The first step in measuring the variable importance in a data set is to fit a random forest to the data. During the fitting process the out-of-bag error for each data point is recorded and averaged over the forest. To measure the importance of the -th feature after



```
             precision    recall  f1-score   support

        0.0       1.00      0.50      0.67         2
        1.0       0.75      1.00      0.86         3

avg / total       0.85      0.80      0.78         5

[[1 0]
 [0 4]]
             precision    recall  f1-score   support

        0.0       1.00      1.00      1.00         1
        1.0       1.00      1.00      1.00         4

avg / total       1.00      1.00      1.00         5

[[1 0]
 [0 4]]
             precision    recall  f1-score   support

        0.0       1.00      1.00      1.00         1
        1.0       1.00      1.00      1.00         4

avg / total       1.00      1.00      1.00         5

[[0 1]
 [0 3]]
             precision    recall  f1-score   support

        0.0       0.00      0.00      0.00         1
        1.0       0.75      1.00      0.86         3

avg / total       0.56      0.75      0.64         4

[[4]]
             precision    recall  f1-score   support

        1.0       1.00      1.00      1.00         4

avg / total       1.00      1.00      1.00         4

('Avg Accuracy: ', 0.91499999999999981)
```

Figure 6: Partial output

training, the values of the -th feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the -th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences.

Features which produce large values for this score are ranked as more important than features which produce small values.

The results from the Random Forest classification are shown above Fig 6.

7.2 Multi-class SVM

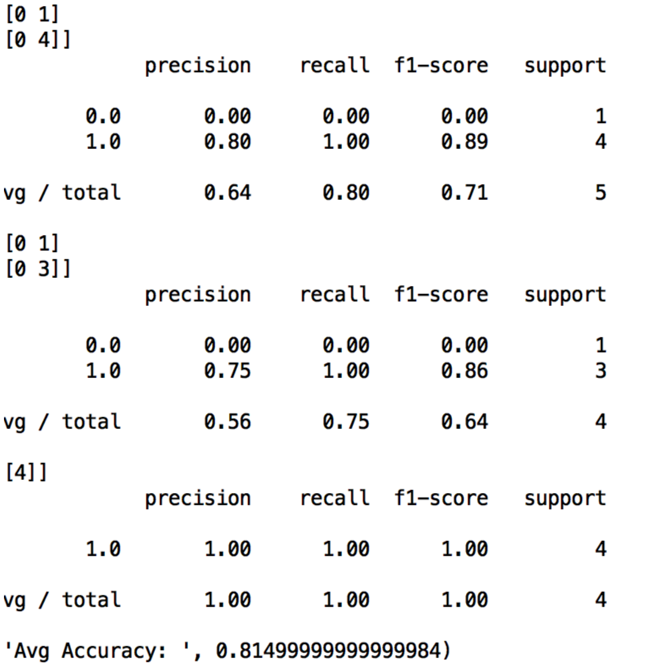
Multi-class SVM aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements.

The approach followed for doing so is to reduce the single multi-class problem into multiple binary classification problems. •

Building binary classifiers which distinguish (i) between one of the labels and the rest (one-versus-all) or (ii) between every pair of classes (one-versus-one). Classification of new instances for the one-versus-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class. For the one-versus-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with the most votes determines the instance classification.

The results obtained from the SVM classification are shown below.

Kernel Trick used : the (Gaussian) radial basis function kernel



```
[0 1]
[0 4]]
             precision    recall  f1-score   support

        0.0       0.00      0.00      0.00         1
        1.0       0.80      1.00      0.89         4

vg / total       0.64      0.80      0.71         5

[0 1]
[0 3]]
             precision    recall  f1-score   support

        0.0       0.00      0.00      0.00         1
        1.0       0.75      1.00      0.86         3

vg / total       0.56      0.75      0.64         4

[4]]
             precision    recall  f1-score   support

        1.0       1.00      1.00      1.00         4

vg / total       1.00      1.00      1.00         4

'Avg Accuracy: ', 0.81499999999999984)
```

8. CONCLUSIONS

In summary, we have performed an experimental analysis on the StudentLife dataset to study possible relation between various aspects of the college students life and their respective mental health levels. Several attributes have been identified and their correlation with the mean stress levels has been measured by various data mining techniques. Using these attributes as factors we predicted the variability of students mental health. The experimental results have been successfully interpreted by the root mean square error value which shows the accuracy of our model. The predicted values were compared against the ground truth values, which is our Mean Stress data.

ACKNOWLEDGMENTS

The StudentLife Dataset used in this work was provided by Dartmouth College. This consists of passive and automatic sensing data from the phones of a class of 48 Dartmouth students over a 10 week term to assess their mental health (e.g., depression, loneliness, stress), academic performance (grades across all their classes, term GPA and cumulative GPA) and behavioral trends (e.g., how stress, sleep, visits to the gym, etc. change in response to college workload -- i.e., assignments, midterms, finals -- as the term progresses).

REFERENCES

[1] Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. (2016) The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 4:e2537

[2] Harari, G. M., Gosling, S. D., Wang, R., Chen, F., Chen, Z., Campbell, A. T. (2017). Patterns of Behavior Change in Students Over an Academic Term: A Preliminary Study of Activity and Sociability Behaviors Using Smartphone Sensing Methods. Computers in Human Behavior, 67, 129-138.

[3] Andrew Campbell " My brother Ed: Mental illness was not his choice. ", Presented at the ACM UbiComp workshop on mental health , September 2016

[4] Wang, Rui, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. "StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones." In Proceedings of the ACM Conference on Ubiquitous Computing. 2014

[5] http://studentlife.cs.dartmouth.edu

[6] Guyon I, Elisseeff A. An introduction to variable and feature selection. J of Mach Learning Research. 2003;3:1157–1182

[7] A. Sano R. W. Picard "Stress recognition using wearable sensors and mobile phones" Affective Computing and Intelligent Interaction (ACII) 2013 Humaine Association Conference, pp. 671-676 2013.

[8] The relationship between mobile phone location sensor data and depressive symptom severity Sohrab Saeb1,2, Emily G. Lattie1, Stephen M. Schueller1, Konrad P. Kording2 and David C. Mohr1

[9] Department of Preventive Medicine, Northwestern University, Chicago, IL, United States Rehabilitation Institute of Chicago, Department of Physical Medicine and Rehabilitation, Northwestern University, Chicago, IL, United States

[10] Harari, Gabriella M., et al. “Patterns of Behavior Change in Students over an Academic Term: A Preliminary Study of Activity and Sociability Behaviors Using Smartphone Sensing Methods.” Computers in Human Behavior, vol. 67, 2017, pp. 129–138., doi:10.1016/j.chb.2016.10.027.

[11] http://www.phqscreeners.com/sites/g/files/g10016261/f/201412/PHQ-9_English.pdf

[12] Breiman, Leo (2001). "Random Forests". Machine Learning. **45** (1): 5–32. doi:10.1023/A:1010933404324.

[13] Liaw, Andy (16 October 2012). "Documentation for R package randomForest”.