RESEARCH

Introduction of Profile Areas of Data Science :Project 8

Suresh Kumar Choudhary, Frenny Macwan and Aman Jain

Full list of author information is available at the end of the article

Abstract

Goal of the Project: To find the co-relation between the activity and the stress for the students studying in a university, along with the other stress correlated features and training a classifier model to classify the mental state of a student.

Main Result of the Project: We were able to build the adaboost classifer model with weighted F1-score 0.54 to predict the student's stress. We found the negative co-relation between the stress and activity and tabulated other co-related features with stress.

Personal Key Leanings: Prophet Library, Using SQLLite, SQLAlchemy, Pandas_gbq

Estimate working hours: 48 Hours in Total

Project Evaluation: 2

Number of Words: 1792

Keywords: Time Series data trend; Seasonality; Corelation; Classification; Studentlife

Scientific Background

Student-Life continuous sensing app assess day to day activity of the students of a batch of 48 students who were using android devices and were voluntarily recruited from a CS batch . The app tracks the sleep, mood, social life, academic performance almost everything, things which are already informed to the students and before signing the consent they are given a walk through of the system, they get some incentives in return like Google Nexus Phones, T-Shirts etc. The general trend what is being observed is term starts with good sleep patterns, good confidence, but when we move towards the end of the term, the sleep pattern, changes, stress increases. The App integrates MobileEMA STATES stress mood across the term. There was a strong co-relation between automatic sensing data and a broad data of Mental Well being, depression etc. Using call records, Bluetooth proximity detect the social life and daily activity. and If there is no response to EMA, a mail was being sent to the student to connect over the wifi. The data were SSL encrypted so there is no privacy breach and also the data would be removed once the people leave the study. Physical activity detetor detects the movements, and classifies run, jog, walk cycling etc. Conversation Detection, which detect a conversation. Sleep Detection sees the light features, phone usage features including the phone lock state during the night. Mobile EMA is like a short survey system, which is being popped on the screen regularly on the student's phone and if the student misses filling the feedback questionare an email is being sent to him/her for reminder. We discuss a number of insights into behavioral trends, and importantly, correlations between objective sensor data from smartphones and mental well-being and academic performance for a set of students at Dartmouth College.

Goal

To reanalyze parts of the Student Life Study by :

1 Analyzing if Stress is co-related to Activity.
2 What else is stress correlated to?
3 Predicting a Student's stress level.

Data

The StudentLife dataset contains the EMA (Survey Responses), the Automatic Sensing data which collects a lot of information.The data set contain four types of data : Sensor Data, EMA Data, Pre and Post Survey Responses and Educational Data. Sensor Data is further divided into 10 different sensor data, physical activity, audio inferences, conversation inferences, Bluetooth scan, light sensor, GPS, phone charge, phone lock, WiFi, WiFi location which are all stored in CSV Files. Inside each sensor data, the information for each student like activity_u01 will be provided in the seperate csv files. EMA Responses are stored in EMA folder where the responses of stress are been given. All the Pre and Post Survey responses are being stored in the survey folder. Educational data, which include classes taken during 2013 Spring term, deadlines for each participants, grades and Piazza usage for CS65, is stored under education folder.

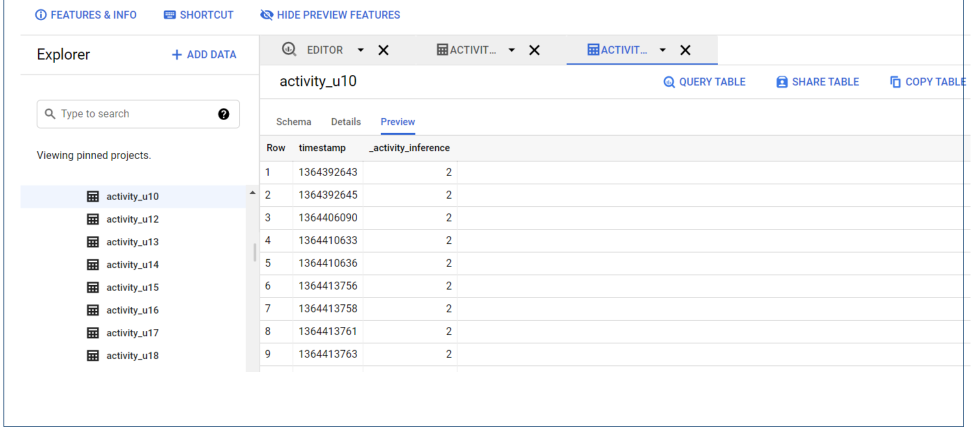
Result

Task 1 : Is stress co-related to activity level?

1 Google cloud platform - Big Query was being used for this task. A dataset 'student' was created in which we created tables for activity and EMA Stress individually for all the students using automated python script. Firstly we loaded all files from the data directory , after that loaded each file in pandas dataframe and finally created the tables in bigquery with same data structure like loaded dataframe and saved file's data in that table with the help of pandas_gbq package functionality. To import the tables in the dataframe we used the statement gbq.read_gbq('select * from Students.activity_u10',project_id='ipads2020assignment8'), with this a dataframe in python notebook was created and all the data was imported. (Can be seen in the Figure 1

2 As per the analysis, most of the students felt stressful sometimes and there were very less students who almost never felt stressful (Figure 2). We studied the data for two students u10 and u16 and the study shows that the student u10 feels stressful fairly often whereas u16 feels stressful very often. Student u16 is more active compared to student u10 where 6% of the u16 students time goes in activity, whereas just short of it activity for u10 is 4.8% , which states student u16 is more active than student 10, where the average stationary % among these set of students are 94.58 and being active is 5.42%, hence student u16 is more active than the average students of the study batch.

Figure 1 Activity Tables in Google BigQuery



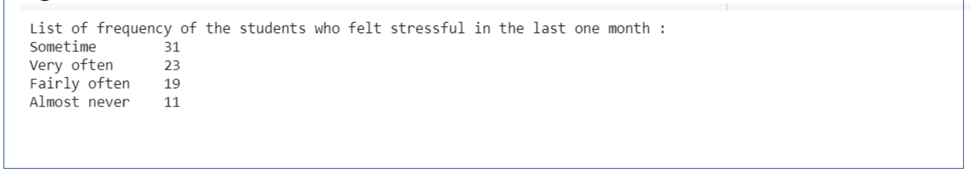
Viewing pinned projects.

activity_u10
activity_u12
activity_u13
activity_u14
activity_u15
activity_u16
activity_u17
activity_u18

activity_u10

Row	timestamp	_activity_inference
1	1364392643	2
2	1364392645	2
3	1364406090	2
4	1364410633	2
5	1364410636	2
6	1364413756	2
7	1364413758	2
8	1364413761	2
9	1364413763	2

Figure 2 Stress Distribution of all the Students



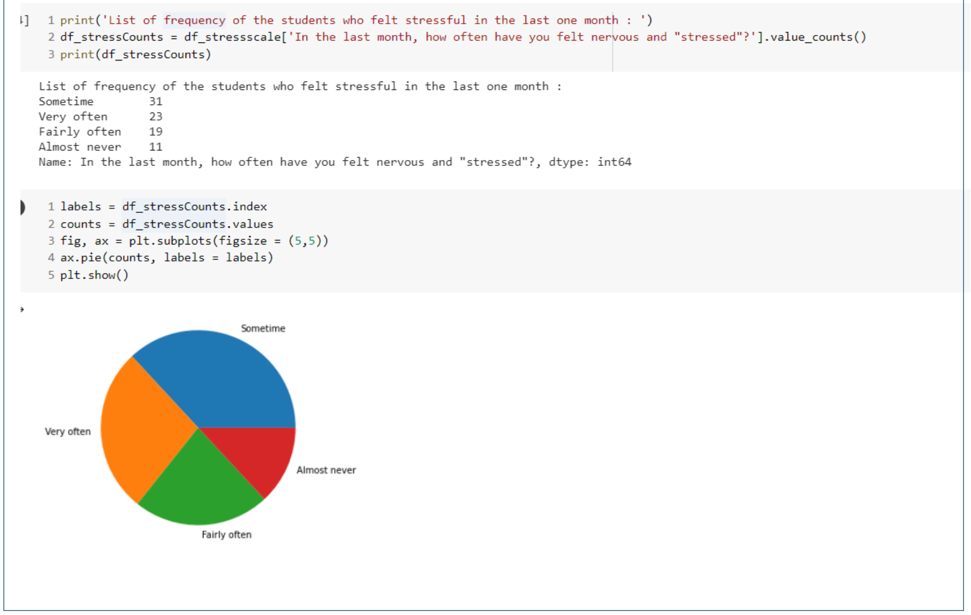
```
List of frequency of the students who felt stressful in the last one month :
Sometime        31
Very often      23
Fairly often    19
Almost never    11
```

3 The activity data frame of student u10 and u16 has been merged into one data frame to perform the time series analysis, and the columns timestamp and the activity_inference are renamed to y and ds to fit the dataframe signature inlign to FBProphet Library. We were unable to use the SQL to impute the dataset and hence we used Pandas to proceed with the imputation of the data. The stressScale table was amputated with only two columns, the *student id* and the column *In the last month, how often have you felt nervous and "stressed"?.*

- Distribution of the stress level of the students is showed in Figure : 3, where we can get the exact count of the students and their stress level for the past one month by the responses they have sent via EMA Api.
- Activity Distribution of the student_u10 is showed in Figure : 4
- Activity Distribution of the student_u16 is showed in Figure : 5
- Through the data set we also concluded that Student U_16 was more active than StudentU_10, which can be seen in Figure : 6

4. Time series for the students u10 and u16 were analysed with the help of Prophet library by extracting components such as trend, weekly and daily Fig 7 and Fig 8. For both students, the trend of stress development is growing with time. Looking at weekly component, it looks similar for both students, that the stress is high in the beginning and goes down until Friday. By the student u10 the stress grows on weekend again, by student u16 the stress grows on Saturday and goes down on Sunday. Looking at the daily component, it can be noticed that for the student u10 the stress is in the beginning of the day high and goes gradually down. For the student u16 the stress increases for period for 10 till 17 and is low for the rest of time. For both students, the trend of activity development is growing with time.

Figure 3 Stress Distribution of all the Students

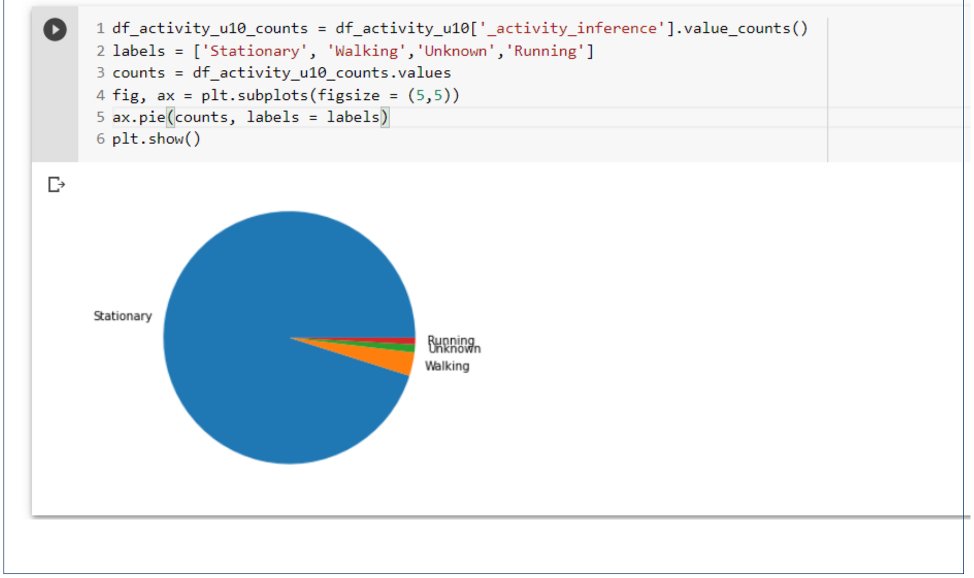


```
print('List of frequency of the students who felt stressful in the last one month : ')
df_stressCounts = df_stressscale['In the last month, how often have you felt nervous and "stressed"?'].value_counts()
print(df_stressCounts)
```

```
List of frequency of the students who felt stressful in the last one month :
Sometime        31
Very often      23
Fairly often    19
Almost never    11
Name: In the last month, how often have you felt nervous and "stressed"?, dtype: int64
```

```
labels = df_stressCounts.index
counts = df_stressCounts.values
fig, ax = plt.subplots(figsize = (5,5))
ax.pie(counts, labels = labels)
plt.show()
```

Figure 4 Activity Distribution of Student U10

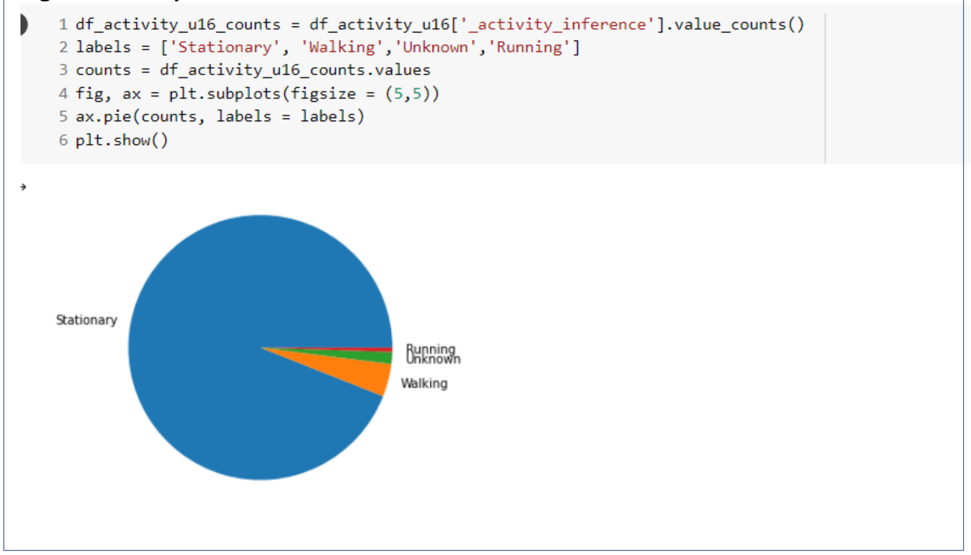


```
df_activity_u10_counts = df_activity_u10['_activity_inference'].value_counts()
labels = ['Stationary', 'Walking','Unknown','Running']
counts = df_activity_u10_counts.values
fig, ax = plt.subplots(figsize = (5,5))
ax.pie(counts, labels = labels)
plt.show()
```

5. We have used Pearson Co-relation coefficient to find how the factors stress and activity are related .The relation ranges from 1 to -1, and 0 signifies there is no co-relation between the factors. We found negative corelation between stress and activity, that means, a student who was more involved in activities, feels less stressed and who didn't involve in activities, feels more stressed or nervous. The code screentshot for the same can be seen in figure 9.

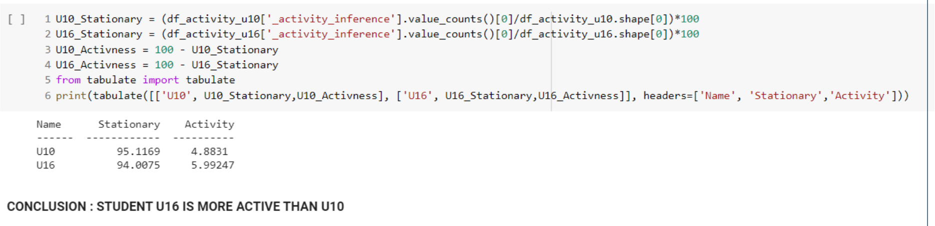
Task 2 : What else is stress correlated to? Table 1 shows the correlations between sensor data, ecological momentary assessment(EMA) data and perceived stress scale (PSS). To calculate the correlation with the stress

Figure 5 Activity Distribution of Student U16



```
df_activity_u16_counts = df_activity_u16['_activity_inference'].value_counts()
labels = ['Stationary', 'Walking','Unknown','Running']
counts = df_activity_u16_counts.values
fig, ax = plt.subplots(figsize = (5,5))
ax.pie(counts, labels = labels)
plt.show()
```

Figure 6 Activity Comparison of U10 and U16

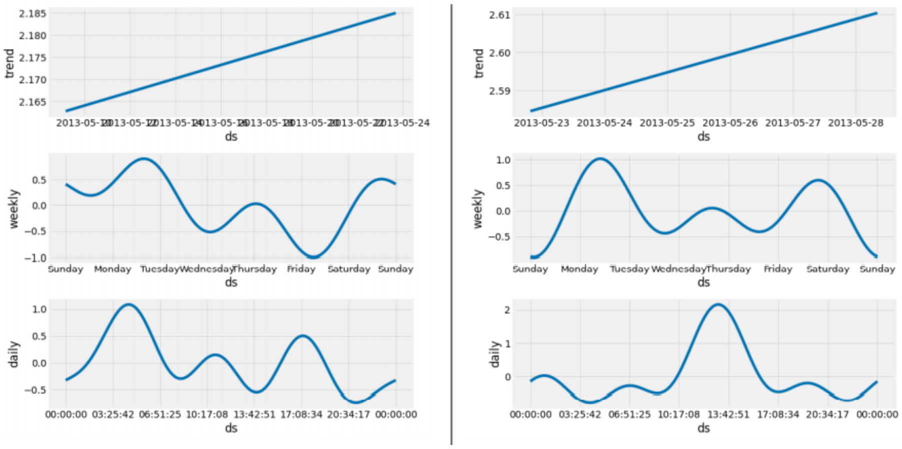


```
U10_Stationary = (df_activity_u10['_activity_inference'].value_counts()[0]/df_activity_u10.shape[0])*100
U16_Stationary = (df_activity_u16['_activity_inference'].value_counts()[0]/df_activity_u16.shape[0])*100
U10_Activness = 100 - U10_Stationary
U16_Activness = 100 - U16_Stationary
from tabulate import tabulate
print(tabulate([['U10', U10_Stationary,U10_Activness], ['U16', U16_Stationary,U16_Activness]], headers=['Name', 'Stationary','Activity']))
```

```
Name      Stationary    Activity
------  ------------  ----------
U10          95.1169     4.8831
U16          94.0075     5.99247
```

CONCLUSION : STUDENT U16 IS MORE ACTIVE THAN U10

Figure 7 students u10 and u16 - Time series analysis for stress data



, we have used the features of automatic sensor data and EMA data like conversation, phonecharge, phonelock, sleep duration, social activity, class

Figure 8 students u10 and u16 - Time series analysis for activity data

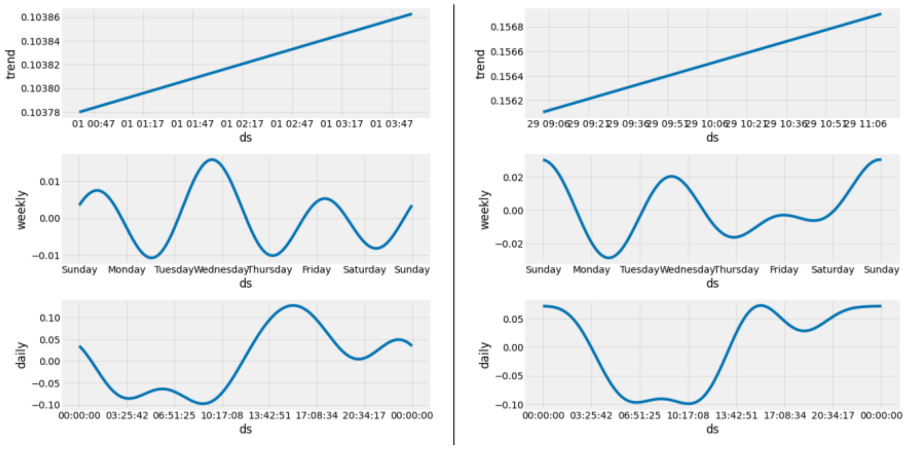
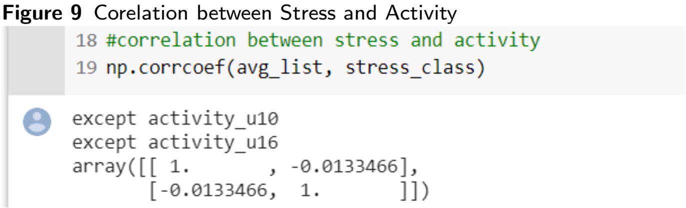


Figure 9 Corelation between Stress and Activity



```
#correlation between stress and activity
np.corrcoef(avg_list, stress_class)
```

```
except activity_u10
except activity_u16
array([[ 1.        , -0.0133466],
       [-0.0133466,  1.        ]])
```

enjoyment, time spent on coursework outside the class, assignment due, time spent in lab etc. The stress scale was encoded in numbers {'Almost never':0,'Sometime':1,'Fairly often':2,'Very often':3}, the information for 44 students were imported in the GoogleBigQuery and tables/schemas were created for each of the student. Conversation (in day), sleep duration, Assignment due(yes:1,no:2) and activity shows negative correlation with pre and post perceived stress. This would in general mean, a student indulged in more conversations has less stress and student who don't converse a lot, feels more depressed or alone. A student who sleeps more, has less stress and student who sleep less , feels more nervous and stressed. One who has no due assignment feels less stressed and who has due assignment feels stressed. Class enjoyed by students (neutral:1, strongly agree:2, agree:3, disagree:4, strongly disagree:5), Lab enjoyed by students (Strongly agree:1, agree:2, neutral:3, disagree:4, strongly disagree:5) and time spent in lab show the positive correlation with pre and post perceived stress. That means, who enjoyed the class and lab have less stress, while who didn't enjoy the class and lab feel more stressed. However, one who spent more time in lab feel more stressed.

Task 3 : Predicting Student's Stress As a part of input data to classifier, we used automatic sensor data (conversation duration(day time), phonecharge duration, phonelock duration and activity) and EMA data (sleep duration, sleep rating, class and lab enjoyment, social acitvity, time spent in

Table 1 Correlations between automatic sensor data, EMA data and perceived stress scale (PSS)

automatic sensing and EMA data	Correlation (r)
Assignment Due (pre)	-0.35
Assignment Due (post)	-0.165
Lab Enjoyed (pre)	0.09
Lab Enjoyed (post)	0.25
Time Spent in Lab (pre)	0.10
TIme Spent in Lab (post)	0.22
conversation duration (pre)	-0.035
conversation duration (post)	-0.08
Phonecharge (pre)	0.132
Phonecharge (post)	-0.13
Phonelock (pre)	0.191
Phonelock (post)	-0.07
Activity (pre)	-0.06
Activity (post)	-0.013
sleep duration (pre)	-0.325
sleep duration (post)	-0.24
overall sleep rating (pre)	0.176
overall sleep rating (post	0.20
social activity (pre)	0.198
social activity (post)	-0.05
enjoyed the class (pre)	0.08
enjoyed the class (post)	0.06
Time spent on coursework outside class (pre)	-0.13
Time spent on coursework outside class (post)	0.28

lab, time spent in coursework outside class) to predict the pre and post perceived stress. To get the idea of input features to model, we took the help of research papers of Mikelsons et al. [1] and Shaw et al. [2]. We calculated the average conversation duration, phonecharge duration, phonelock duration, sleep duration, time spent for lab and coursework for each student and for features like activity, assignment due, class and lab enjoyed , we calculated the most frequent category for each student. We randomly splitted the dataset into the training and test set with ratio 8:2 for both pre and post perceived stress prediction. We used 38 student's data for post preceived stress prediction model and 44 student's data for pre preceived stress prediction model. We build the decision tree and adaboost classifier to predict how often have you (student) felt nervous and stressed in the last month (Almost never:0,Sometime:1,Fairly often:2,Very often:3). We used the scikit-learn library for decison tree classifier with the parameters criterion = ”entropy”, max_depth=10, min_samples_leaf=1 and for adaboost classfier with parameter n_estimators=30. As a part of quality measure, we used weighted F1-score as there were 4 classes in perceived stress variable (response variable). Among these two models, Adaboost model was performing better for both pre and post preceived stress prediction with weighted F1-score as 0.70 and 0.54 respectively. The result can be seen in table 2 for the both models for pre and post cases. The screenshot of code for model result can be found in figure 10.

Table 2 Classification Model to Predict Student Stress

Classifier	F1-Score
Decision Tree (pre)	0.48
Adaboost (pre)	0.70
Decision Tree (post)	0.26
Adaboost (post)	0.54

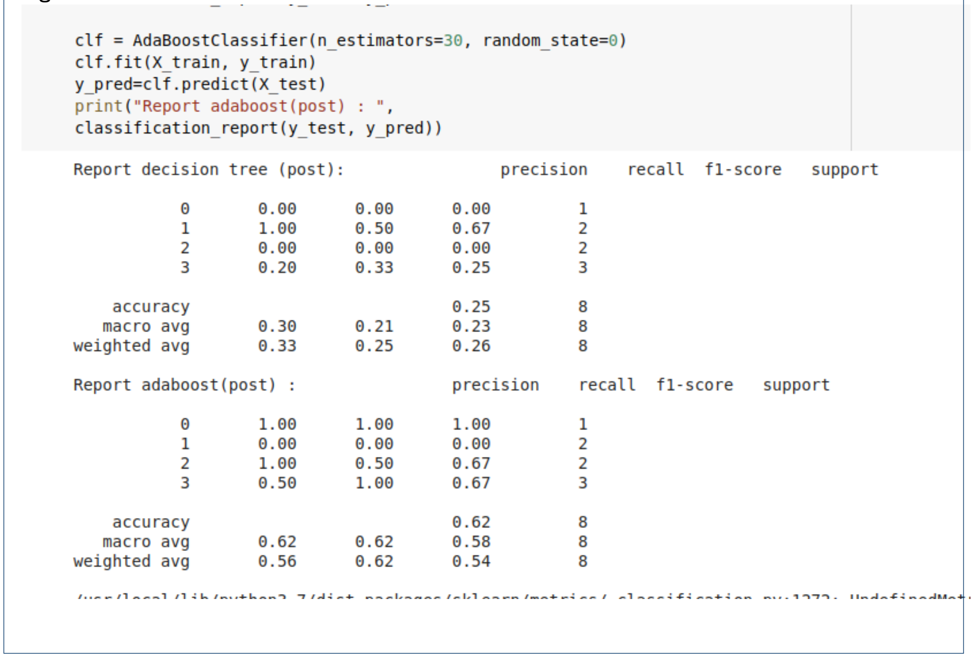


Figure 10 Adaboost classfier code screenshot

```
clf = AdaBoostClassifier(n_estimators=30, random_state=0)
clf.fit(X_train, y_train)
y_pred=clf.predict(X_test)
print("Report adaboost(post) : ",
classification_report(y_test, y_pred))
```

```
Report decision tree (post):                precision    recall  f1-score   support

           0       0.00      0.00      0.00         1
           1       1.00      0.50      0.67         2
           2       0.00      0.00      0.00         2
           3       0.20      0.33      0.25         3

    accuracy                           0.25         8
   macro avg       0.30      0.21      0.23         8
weighted avg       0.33      0.25      0.26         8

Report adaboost(post) :                precision    recall  f1-score   support

           0       1.00      1.00      1.00         1
           1       0.00      0.00      0.00         2
           2       1.00      0.50      0.67         2
           3       0.50      1.00      0.67         3

    accuracy                           0.62         8
   macro avg       0.62      0.62      0.58         8
weighted avg       0.56      0.62      0.54         8

/usr/local/lib/python3.7/dist-packages/sklearn/metrics/_classification.py:1272: UndefinedMet
```

Discussion

Mental well being of a student is of the most priority. University can surely take this study and find out how the mental state of the student vary during the course of the semester. Through the data the university can get ample amount of data or about the lifestyle of the students. As we move towards the end of semester the habits break, the sleep patterns become very irregular, the diets becomes bad, and due to which the students become more stressful. Studies might show, a good amount of activity, talking to personals over phone calls, having a good social life, less social media can reduce the stress of the student. University this way, can surely circulate some guidelines to the students during the course of semester.

Appendix

Code -

- Aman - Task 1.1, 1.2, 1.3
- Suresh - Task2, Task3
- Frenny - Task 1.4, 1.5

Report -

- Aman - Scientific background, Data, Task1
- Suresh - Discussion, Task2, Task 3
- Frenny - Abstract, Goal and Task1.4,1.5.

Author details

References

1. Mikelsons, G., Smith, M., Mehrotra, A., Musolesi, M.: Towards deep learning models for psychological state prediction using smartphone data: Challenges and opportunities. arXiv preprint arXiv:1711.06350 (2017)
2. Shaw, A., Simsiri, N., Deznaby, I., Fiterau, M., Rahaman, T.: Personalized student stress prediction with deep multitask network. arXiv preprint arXiv:1906.11356 (2019)