



Depression Forecast for Supporting Students at University



Executive Summary

Due to the lockdown for pandemic, the percentage of depression in Germany increased dramatically especially for students in university. In this report, based on the provided dataset, we are aiming to use machine learning algorithms to predict level of depression from the features about lifestyle. After the analysis and visualisation of data, we preprocess and transform the data into suitable form, then select the features and balance the original dataset. We use Random Forest, K Nearest Neighbor (KNN), and Support Vector Machine (SVM) as our models and achieve relatively high accuracy after hyper-parameter tuning. The results indicate that our model could possibly be used to help psychologists with depression prediction in university.

Contents

1. Introduction	4
1.1. Problem statement.....	4
1.2. Research under discussion and gaps	4
1.3. Outlines of our report.....	4
2. Dataset	5
2.1. Origin of the data	5
2.2. Data record description	5
2.3. Data preparation and visualization.....	5
2.3.1. Ratio scale	5
2.3.2. Interval scale	5
2.3.3. Nominal scale.....	6
2.3.4. Ordinal scale.....	6
2.3.5. Visualization.....	6
2.4. Feature selection and Data normalization	6
2.5. Balanced dataset.....	7
3. Procedure and Analysis.....	7
3.1. Statistics overview (OLS)	7
3.2. Classification.....	7
3.2.1. Random Forest	7
3.2.2. KNN	8
3.2.3. SVM	8
4. Results and Discussion.....	8
4.1. Comparison of the results.....	8
4.2. Discussion	9

1. Introduction

1.1. Problem statement

During the pandemic, almost everyone's life changed dramatically, and faced multiple uncertainties. University students, in particular, the young adults under a lot of social, financial, and daily life pressure must study at home, and be away from peers and normal social life during the lockdown period, which makes them more vulnerable to depression. For instance, a study in Germany has shown that 23 % of students had depressive symptoms during the lockdown, compared to 13 % before the pandemic. There are also many types of research focusing on potential determinants of Anxiety and Depression Among University Students in different cultures and backgrounds.

1.2. Research under discussion and gaps

A study from Pakistan (Hamid Saeed, 2017) [1] has found significant differences in frequency distribution regarding age, marital status, living status, and reasons affecting mental health between annual and semester system students. Multivariate analysis demonstrated a significant association between depression in male students, age ≤ 22 years, and living status. Another similar Study (Baye Dagnew, 2020) [2] from Ethiopia focusing on Medical and Health Science students reports that the odds for depression are higher among students who came from rural families, who experienced tooth grinding, who had night sleep disturbances, who reported daytime sleepiness, who had reported stress, and those studying Health Science.

Several studies in East Asia have come to similar conclusions. According to Md.Ashraful Islam (2016) [3] and their study from Malaysia. The risk of depression was higher in second-year students compared to first-year students, and higher in students staying outside campus compared to students staying inside the campus. Lower economic status, sleeping issues, and PTSD also seem to be factors influencing depression among university students in Malaysia. When considering the impact of the COVID-19 pandemic on depression among university students, a recent study from China (Yanling Yu, 2022) [4] may provide some insight. Their population-based research indicates that males, upper grades, low parental education, low physical activity levels, and irrational eating habits are risky factors for complaining of depression.

1.3. Outlines of our report

In this report, we would use machine learning methods to predict the depression level of university students based on their lifestyle. In section 2, we would pre-process our data. In this part, the data set will be cleaned, modified, encoded, and balanced. In section 3, we analyze different algorithms. In section 3.1, we use OLS Regression to overview the whole dataset. In section 3.2, Random Forest, k-Nearest Neighbor, Support Vector Machine will be used to conduct classification. In section 4, we compare different methods and analyze the results, also a discussion is followed.

2. Dataset

2.1. Origin of the data

The dataset used has 754 samples, and each sample has 34 features. Based on the dataset we can assess the psychological functioning of the participants [5] and fulfill the depression forecast further.

2.2. Data record description

In general, within the 34 features of the dataset, it is included different aspects of information for the participants. First of all, some basic background is collected, like gender, age, number of siblings, hobbies, subject, academic year, etc. We also care about the mood and personality of these underground students. Besides, there are life routine features, such as the time to get up, when to have breakfast, etc. And other lifestyle variables may be correlated with depression, like sports, alcoholic or not, drug-addicted condition, etc., consisting of the rest of the dataset.

From statistical perspective, these features can be divided into different scales as following table:

Category	features	
Ratio scale	age, semester, siblings, study_hours, movies_per_week, music hours, friends, <i>depression</i>	numeric values, no natural zero, no division and multiplication
Interval scale	wakeup, breakfast, launch, dinner, sleep	numeric values, natural zero
Nominal scale	mood, gender, subject, living, employment, personality, relationship, hobbies, smoker, alcoholic, drug-addicted, medication	Categorical variables and cannot be ordered, no average
Ordinal scale	hangout, social_events, pray, sports, exercise, video_games, meditation, phone-hours, <i>category</i>	Meaningful order exists

After overviewing the features, we need to preprocess these features before the data mining and analysis. The two features which are italic in the table above are dependent features of our dataset.

2.3. Data preparation and visualization

2.3.1. Ratio scale

Since these features are numeric, we can easily check the null value, and outliers (we set limits for them), transfer non-numeric numbers to null and drop these samples.

2.3.2. Interval scale

For the features that fall in this category, we use “wakeup” and “sleep” producing a new feature “sleep_duration”, which is a ratio scale and more related to depression from an empirical view.

2.3.3. Nominal scale

We use one hot encoding method to transfer the 12 features to different matrixes. Besides, we combine wrongly written subjects and similar ones to make “subject” uniform and choose the 8 most common hobbies to make life easier.

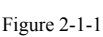
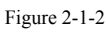
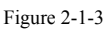
2.3.4. Ordinal scale

Since these features can be ordered, we change the frequency of them to numbers, and the bigger number means the higher frequency.

2.3.5. Visualization

We use histograms for ratio scale features (**Appendix 2-1**), and we can get some clues about the distribution of these features. To be more quantitative, we visualize the new feature *sleep_duration* using a scatter plot and histogram. From the figures (**Appendix 2-2**), we cannot see obvious relation between *sleep_duration* and *depression* but it seems like following Gaussian distribution.

For nominal and ordinal scale features, before encoding, we show histograms for the feature themselves. At the same time, we visualize the relations between them and depression using a boxplot and swarm plot (**Appendix 2-3**). Although most of the figures don’t show any correlation intuitively, three features do show some properties, helping the further analysis.

From  figure 2-1-1 we can  see that doing sports  sometimes may decrease your depression [6], since the higher frequency of doing sports, the lower the average depression score shows; and figure 2-1-2 shows that drug addiction can increase the depression level a lot; in the right figure, we can see that the *happy* mood has an overall lower score of depression.

2.4. Feature selection and Data normalization

After we transfer as many as possible features to ratio scale data, we still have to select features and normalize them to get our data ready. We follow these steps to remove features:

- 1) Remove the improper interval scale features, since they make no sense anymore;
- 2) Drop the two target columns, which are *category* and *depression*;
- 3) Remove correlated features;

Besides, we check the feature importance for random forest, see more detail in section 3.2.1.

2.5. Balanced dataset

After all these works, we still have one concern about our dataset. We can see that most of the data in the dataset fall into the category “*Moderate*”. However, from the practical view, we need a balanced dataset to make the machine learning algorithms work better, so we produce a balanced dataset using SMOTE algorithm. There may be bias and noise exist because of the relatively small dataset. And our OLS (ordinary least squares) regression result verifies this, see more detail in section 3.1.

3. Procedure and Analysis

3.1. Statistics overview (OLS)

After all procedures in section 2. We have 65 features as our independent variables and “*depression*” as the dependent variable for regression and “*category*” as another dependent variable for classification.

To understand better the correlation with the independent variables and feature “*depression*”, we set up the OLS model to get an overview. The following table shows a simplified version for comparing the output of the original dataset and the balanced dataset.

	Original Dataset	Balanced Dataset
R-squared	0.384	0.599
Adj. R-squared	0.308	0.570

The balanced dataset obviously improves the overall correlation. In addition, in the original dataset, there are only 10 out of 65 features have significance with the 95% confidence interval. But this number increase to 23 out of 65 when we use a balanced dataset. Based on these reasons, we use the balanced dataset (1143 samples) for the following analysis.

3.2. Classification

To infer better the relationship between independent variables and the feature “*category*” and accomplish the prediction in the end, we introduce 3 machine learning classifiers. Since there are three categories in the dependent feature, we use multiply classifications in our machine learning models. We split the dataset into training and test parts first. For all models, we turn the hyper-parameters based on precision and recall criteria [7] and record the accuracy score.

3.2.1. Random Forest

A random forest is an estimator that fits decision tree classifiers on various sub-samples of the full dataset and uses averaging to improve the predictive accuracy and control over-fitting. And bootstrap is the default method here to obtain the sub-samples. Besides, we also set the hyper-parameter, which is the depths (nodes) of every decision tree, in a series of numbers and search for the best one.

In addition, we check the feature importance for random forest (**Appendix 2-4**). It shows that there are 4 features non-significant within the top 10 important features. They are sleep_duration(4th), movies_per_week(6th), study_hours(7th), meditation(8th). Therefore, there are more than correlations when describing the relationship between dependent and independent features.

3.2.2. KNN

KNN classification is an estimator that the query data point is assigned the class which has the most representatives within the nearest neighbors of the point. In the algorithm, we need to set the numbers of neighbors manually. Besides, we use the Euclidean distance as the criterion to choose neighbors in the high-dimension space.

3.2.3. SVM

SVM finds the hyperplanes that represent the largest separation, or margin, between classes. We choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. In general, the kernel is used to transfer samples into a higher dimension and classify data easily. In this project, we choose the RBF kernel.

4. Results and Discussion

4.1. Comparison of the results

The accuracy scores are shown below:

	Random Forest	KNN	SVM
Testing accuracy	0.79	0.60	0.84
Training accuracy	1.00	0.99	0.88

According to the table, SVM and Random Forest perform much better than KNN. See the full scores table for every algorithm in the output of code. In our view, the reasons behind this are following:

1) For KNN method, we know that it would work well if the data is relatively clustered and that the same class samples are closer in the vector space. However, if the data mess up together, it cannot work out very well. In this case, SVM can be effective, since it can use the kernel to increase dimensions.

2) It is reasonable that Random Forest also performs well since it is the property for ensemble learning.

Besides, we use the original data run the same algorithms and the outputs are worse. Therefore, they are not shown in the code and here are two possible reasons:

1) From the sample's perspective, the original dataset is relatively small and bias exists. Though we balance the dataset, since the algorithms would perform better if there are more real data from depression classes "None" and "Severe".

2) From the feature's perspective, maybe there is too much noise in the dataset. In section 3.1, even for the balanced dataset, there are still more than half of the features that are non-significant and the value of R-square and Adjusted R-square is not that ideal. Under this circumstance, collecting other features which are more significant maybe a good choice.

4.2. Discussion

Based on the results and analysis, we can see that our model has the ability to predict depression category in a relatively high accuracy. Thus, our model could possibly be used for depression prediction in university or at least be an assistance to the Psychological Counseling Department.

In addition, since the most important features which cause depression are identified, we could implement a system and let the students do the testing by themselves then decide if to make an appointment to a psychologist.

In the future, we would try to collect more data, use other algorithms and feature selection strategies to get better performance, in order to generate a more applicable model.

Bibliography

- [1] Hamid Saeed, 2017. Determinants of Anxiety and Depression Among University Students of Lahore
- [2] Baye Dagne, 2020. Depression and Its Determinant Factors Among University of Gondar Medical and Health Science Students, Northwest Ethiopia: Institution-Based Cross-Sectional Study
- [3] Md. Ashraful Islam, 2016. Factors Associated with Depression among University Students in Malaysia: A Cross-sectional Study
- [4] Yanling Yu, 2022. Prevalence and Related Factors of Depression, Anxiety and Stress in University Students: an Extensive Populationbased Survey in China.
- [5] Ahnaf Atef Choudhury, 2019. Predicting Depression in Bangladeshi Undergraduates using Machine Learning
- [6] Rachel Jewett, 2014. School Sport Participation During Adolescence and Mental Health in Early Adulthood
- [7] Wikipedia https://en.wikipedia.org/wiki/Precision_and_recall

Appendix

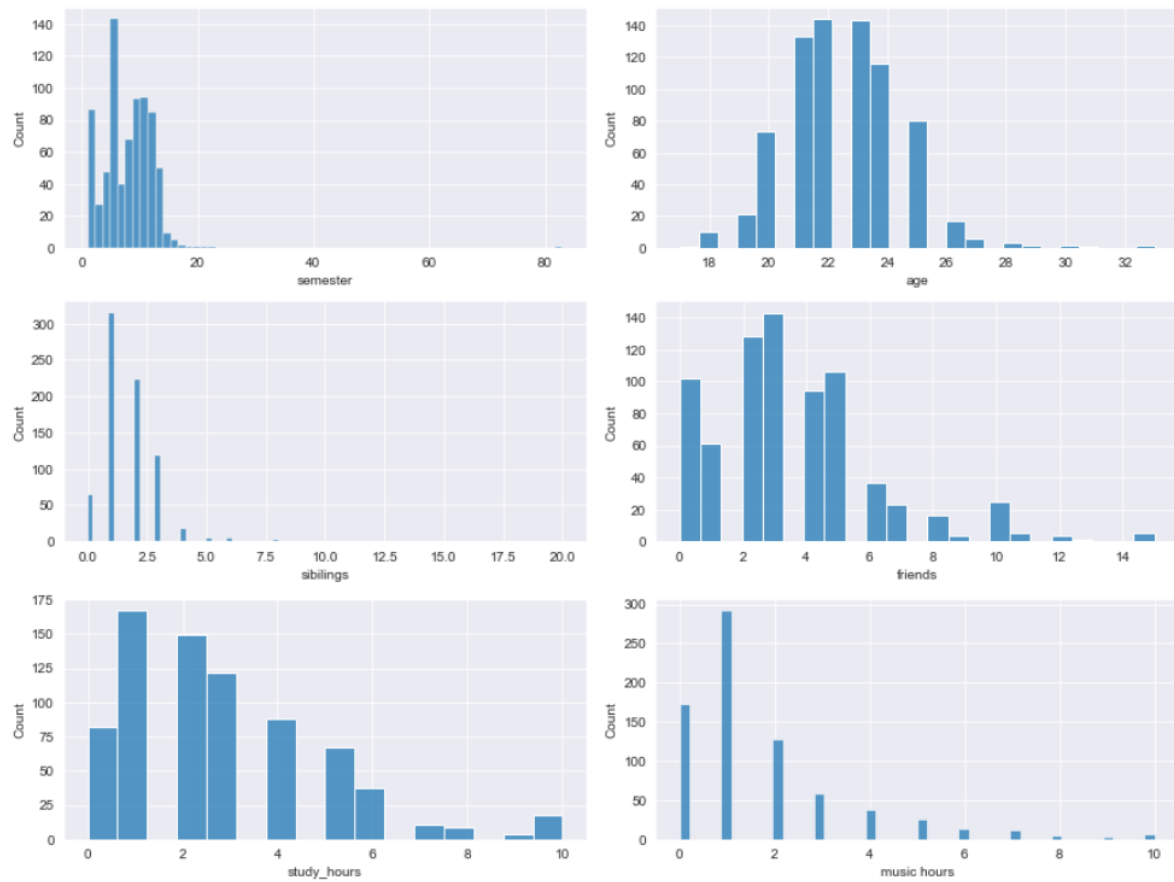
1 Description of the Code

Since we add enough comments in the code (Jupyter Notebook) so we believe it can basically explain itself. The description of the data organization and architecture can be seen at the beginning the code. However, there are some order differences comparing with the report and the code, thus we describe the code structure below:

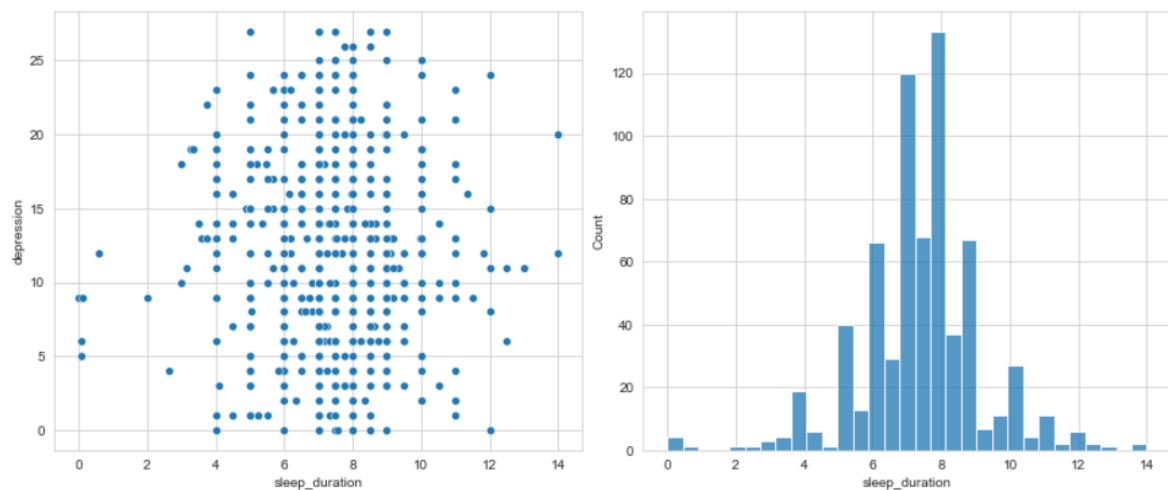
- ◇ Problem Statement, Task, Dataset
- ◇ Data Checking: overview of the dataset about number of students and features, feature statistics
- ◇ Data Visualization: the visualization of distribution for some important features
- ◇ Data Pre-processing: Data cleaning for NaN and unreasonable values; Feature Transformation for hobbies, subject and sleep duration which we created to replace *wakeup* and *sleep* feature; Some descriptive analysis before the Feature Transformation for categorical features to numerical and one hot encoding; Outlier Detection specifically on four features (age, semester, friends and siblings).
- ◇ Feature Selection: Remove Improper Features including features which make no sense, are replaced and are targets; Remove Correlated Features based on correlation matrix
- ◇ Data Normalization to scale the value of features to 0-1; Split dataset to training set and testing set; Data Balancing to get more balanced data for training
- ◇ Model Creation and Evaluation: we use Random Forest, KNN and SVM to classify the category of depression. The Feature Importance is used to Random Forest and the number of features has then been reduced. We also do the hyper-parameters for every algorithm.
- ◇ Further Analysis and Exploration: Meaning of Balancing Data, we use OLS Regression to get R-squared scores for three different ways to balance data and compare with the original data; PCA is used to SVM in order to get better results but we don't get satisfied output for now (**Appendix 3**).

Appendix 2 Figures

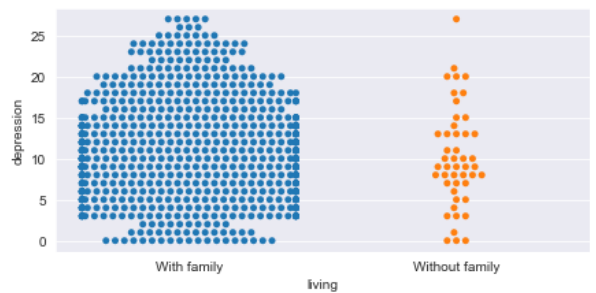
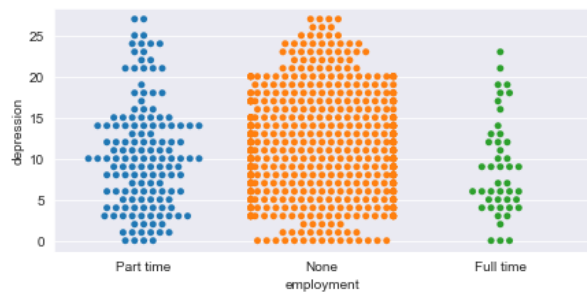
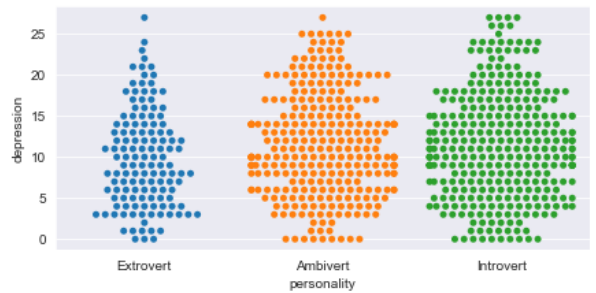
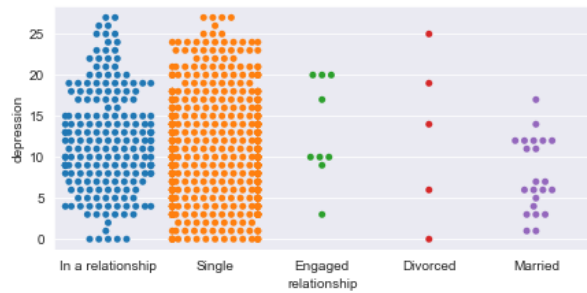
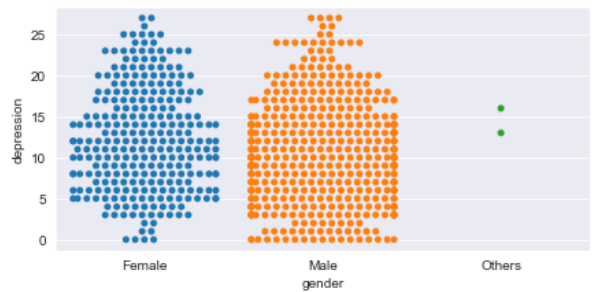
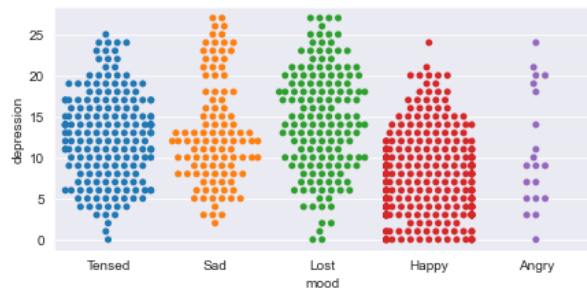
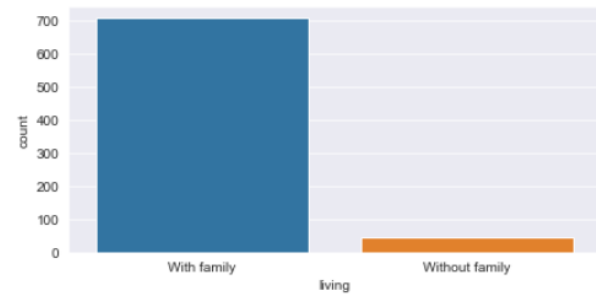
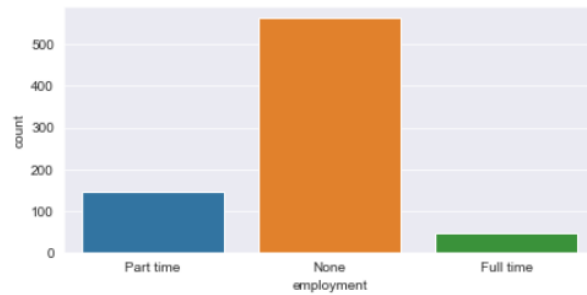
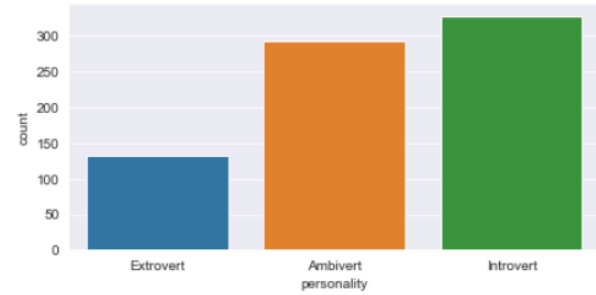
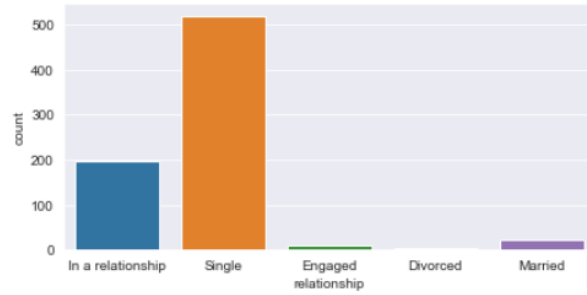
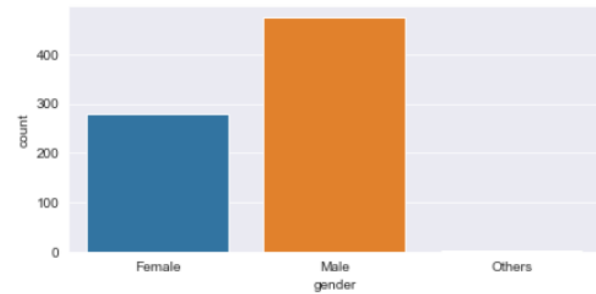
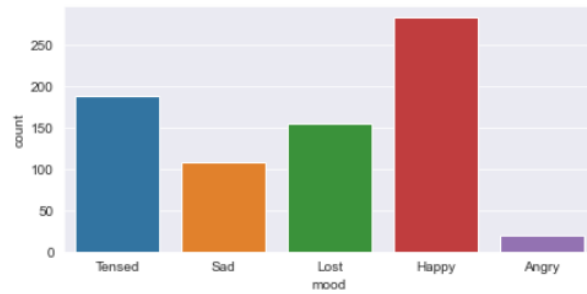
Appendix 2-1

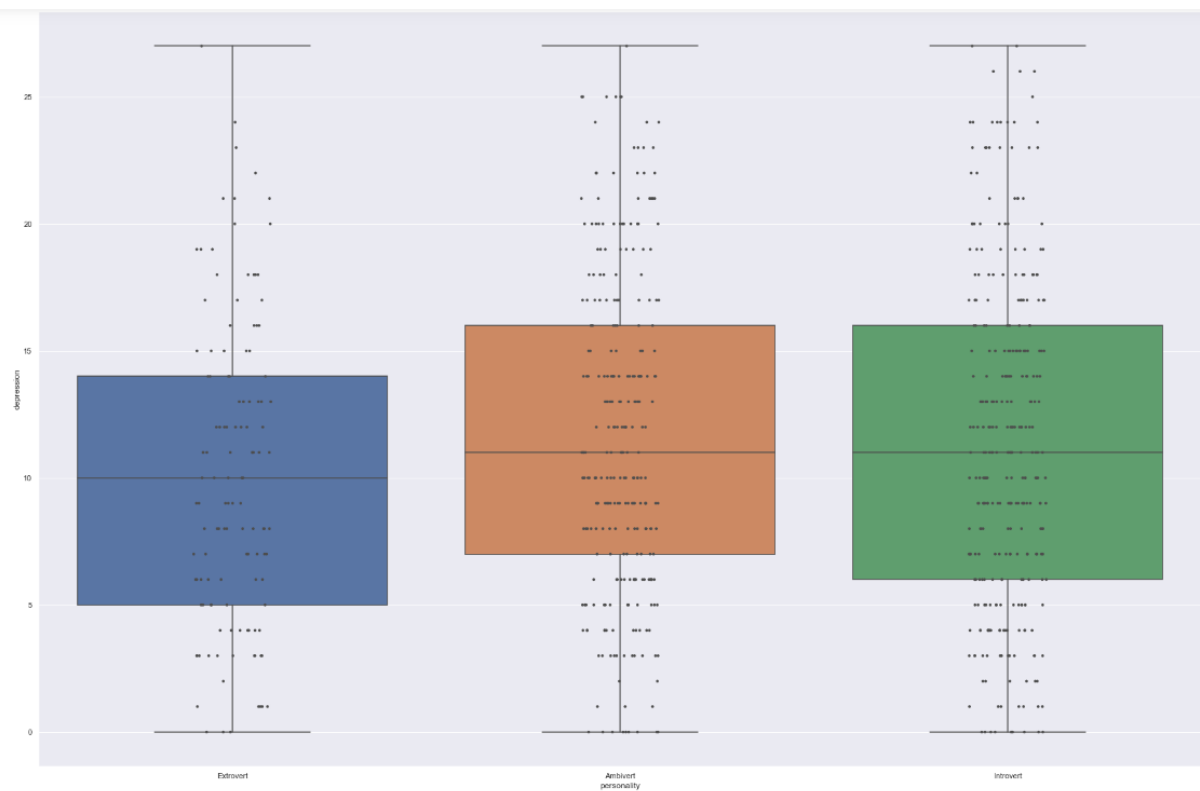
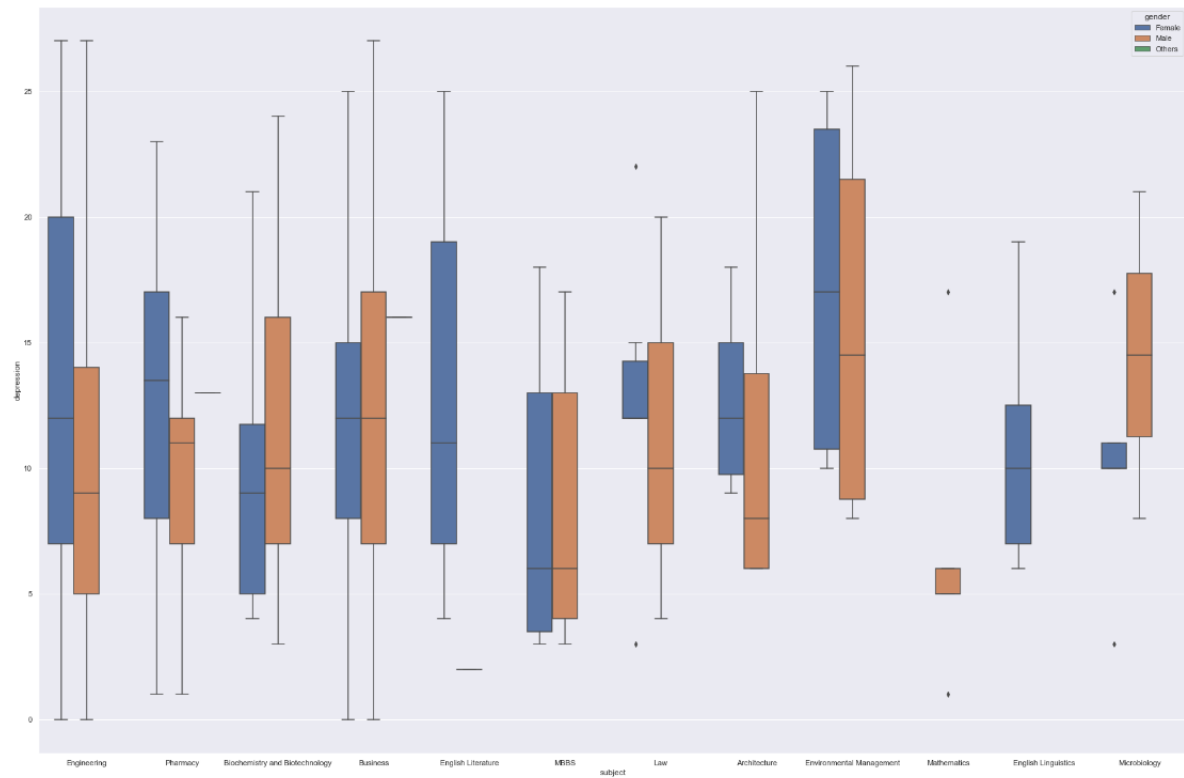


Appendix 2-2

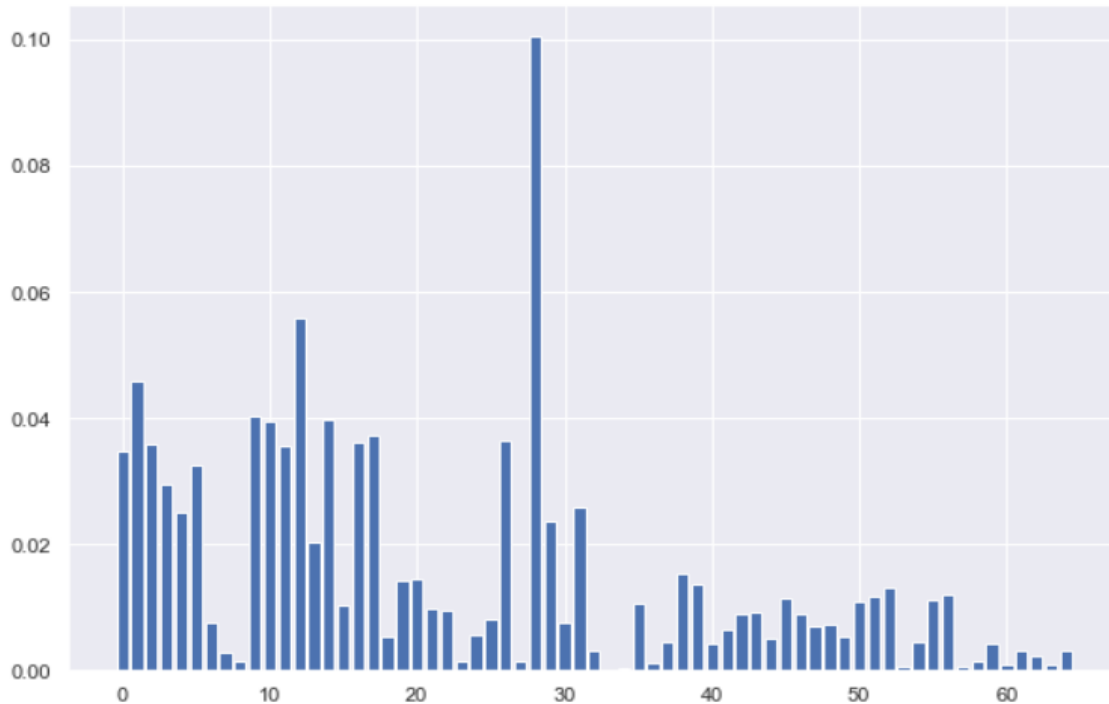


Appendix 2-3





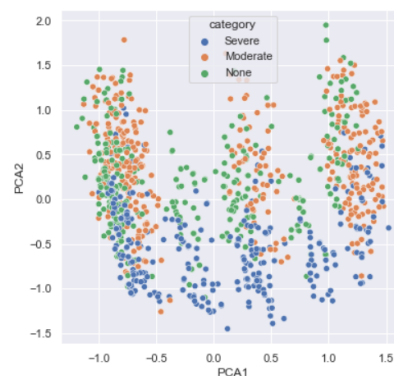
Appendix 2-4



```
(1143, 18)
Final Features:
Index(['age', 'gender', 'semester', 'sibilings', 'employment', 'personality',
      'friends', 'pray', 'sports', 'exercise', 'meditation', 'study_hours',
      'movies_per_week', 'music hours', 'sleep_duration', 'mood_Happy',
      'mood_Lost', 'mood_Tensed'],
      dtype='object')
```

Appendix 3 PCA

PCA is a widely used method to reduce the dimension of data when we assume there is a linear relation between features. And we try to find a new orthogonal space to represent as much variance as possible from the original data. Here the hyperparameter is the number of components which is set to 2. The output is visualized in below. Then we use the SVM as the classifier. Since the result is not ideal, we did not go further.





Declaration of Authorship

I affirm that I have produced the work independently, that I have not used any aids other than those specified and that I have clearly marked all literal or analogous reproductions as such.

Location, Date
