Depression, Anxiety and Stress Scale

Dăscălescu Dana

Group 407 - Artificial Intelligence

January 18, 2022

Introduction, Motivation and Objectives

Introduction

In this project, we are exploring ways to analyse the dataset, pre-process the data and extract relevant features for the clustering algorithms. The main objective is to segment patients according to their mental health problems and the major dimensions(severity) of mental illness based on the symptoms they experience.

Composition of the dataset and task description

Everyone experiences symptoms of stress, anxiety or depression at some time. For example, some people state that stress only moderately affects their lives, but it usually disappears when the stressors disappear. However, chronic mental illnesses due to stress can affect our well-being and cause symptoms such as headaches, uncontrollable shaking, nervousness, palpitations and an unsteady body temperature. For this reason, it is important to identify the symptoms and assign them to the corresponding illness, and then provide appropriate treatment. The DASS report and its variations are commonly used in the clinical setting by researchers and clinicians to establish a change in a person's emotional state and to make a diagnosis. The ability to automatically distinguish between the three related states of depression, anxiety and stress should be useful to identify them in a relatively short time, which is also the main motivation for choosing this topic for the project. However, decisions based on this particular score profile should only be made by experienced clinicians who have conducted a further appropriate clinical examination.

The DASS-42 report (see **??**) is designed to determine a person's emotional disturbance over a short period of time, to measure the negative emotional states of depression, anxiety and stress, and to distinguish between the three related states. Because of its properties, the test has been widely studied and used in research. However, concerns have been raised that it does not measure a single trait but a variety of loosely related traits, so the result is not very meaningful in classifying people according to their primary emotional state.

The **Depression, Anxiety and Stress Scales** dataset consists of approximately 39,600 DASS-42 online surveys[1] completed by people who agreed to also conduct the full research survey and answered " *Yes*" to the question " *Did you give accurate answers and may they be used for research?*". Initially, the dataset was used to predict the overall DASS score. Later, it was used for a classification task on a three-point scale (predicting the primary mental illness - anxiety, depression or stress) or a five-point scale (predicting the intensity of the core problem - normal, mild, moderate, severe or extremely severe). In this project, we will use the dataset in an unsupervised environment to compare different clustering algorithms.

The data includes the 42-item questionnaire, the position of the question in the survey, the time taken to answer in milliseconds and other durations measured by the server, the responses to the Ten Item Personality Inventory test[5], a detailed demographic survey and technical information describing the conditions under which the user completed the research survey. In total, the dataset contains 170 columns.

Overview and organization

The rest of the documentation is structured as follows: Section 3 describes in detail the dataset and the investigations conducted with the data and provides a summary of the main features and patterns discovered, Section 4 presents in detail a comparison of different clustering models and different types of features, the steps taken to set the model parameters and the evaluation protocol, and the last section is a summary of the information presented and the conclusions drawn.

Exploratory Data Analysis

Before building the models, it was necessary to perform an analysis of the dataset, steps for cleaning and pre-processing the data.

Data Cleaning

Firstly, we generate descriptive statistics about our dataset without performing any pre-processing step. We wanted to summarize the central tendency, dispersion, shape of the dataset's distribution, and check for duplicates and missing values.

Since the responses recorded in the DASS survey are the most important information, we decided to remove all rows with missing values. We also checked the percentage of missing values in the other columns and found that 0.1% of the values were missing in the 'country' column and 28.68% in the 'major' column. Given the high percentage of missing data, we decided to remove the 'major' (field of study) column from the dataset and replace the missing values with the tag 'Unknown' for the 'country' column.

The following list of words was presented as a checklist and the subjects were instructed to tick all the words whose definitions they were sure they knew:

VCL1 - boat	VCL2 - incoherent	VCL3 - pallid	VCL4 - robot
VCL5 - audible	VCL6 - cuivocal	VCL7 - paucity	VCL8 - epistemology
VCL9 - florted	VCL10 - decide	VCL11 - pastiche	VCL12 - verdid
VCL13 - abysmal	VCL14 - lucid	VCL15 - betray	VCL16 - funny

Table 1: List of words used for validity check

A value of 1 means check and 0 uncheck. The words used for **VCL6**, **VCL9** and **VCL12** are not real words, so we use these columns for a validity check of the data provided by the survey participants. We will eliminate all the samples which have a 1 on either one of the VCL6, VCL9, or VCL12 columns. Since this word list does not contain any information about the emotional state of the participant, we will also remove all 16 columns.

We choose to also eliminate columns with the duration measures related to time spent on the landing page (*introelapse*), time spent on the DASS report (*testelapse*) and time spent responding to the other parts of the survey (*surveyelapse*), the screen size of the device from which the user took the survey (*screensize*), the user's network (*uniquenetworklocation*), how the user found the report (*source*), as well as the columns regarding the position of the survey question, as this information is redundant.

Number of samples remaining after completion of the data cleaning process: $34,583$.
Number of columns remaining after removing the redundant data: 106.

	age	gender	education	religion	orientation	race	married	familysize
count	34583.000000	34583.000000	34583.000000	34583.000000	34583.000000	34583.000000	34583.000000	34583.000000
mean	23.527542	1.795188	2.517162	7.728017	1.633635	30.029205	1.151751	3.557817
std	22.818171	0.438131	0.879199	3.469664	1.361671	25.689312	0.434728	2.117563
min	13.000000	0.000000	0.000000	0.000000	0.000000	10.000000	0.000000	0.000000
25%	18.000000	2.000000	2.000000	4.000000	1.000000	10.000000	1.000000	2.000000
50%	21.000000	2.000000	3.000000	10.000000	1.000000	10.000000	1.000000	3.000000
75%	25.000000	2.000000	3.000000	10.000000	2.000000	60.000000	1.000000	5.000000
95%	41.000000	2.000000	4.000000	10.000000	5.000000	70.000000	2.000000	7.000000
99.9%	73.000000	3.000000	4.000000	12.000000	5.000000	70.000000	3.000000	12.418000
max	1998.000000	3.000000	4.000000	12.000000	5.000000	70.000000	3.000000	133.000000

Figure 1: Statitical summary of the demographic data

Data Preprocessing

Remove Outliers

We are interested in identifying outliers or particular patterns that might indicate the validity of the data. From the statistical summary of the demographic data (figure 1), we can see that there are some outliers in the variable 'familysize' that might indicate that the people in the survey did not provide reliable information. Often the outliers contain valuable information about the process under study or the process of data collection, but in our case these outliers can be used as a measure of validity. For this reason, we will keep only the rows with the data from 'familysize' column above the 99.9% percentile.

DASS scores and categories

The 42-item questionnaire contains 14 items for each of the three scales, which are divided into subscales, with each subscale containing between 2 and 5 questions with similar content. Subjects were able to answer each question using a four-point frequency scale to rate the extent to which they had experienced each affirmation in recent weeks. Each answer is associated with a score and the scores for the depression, anxiety and stress scales are calculated by summing the scores for the corresponding items. The overall DASS score is the average of the scores of the three scales. For the calculation of the DASS score and the categories, we followed the methodology described in [2], [3].

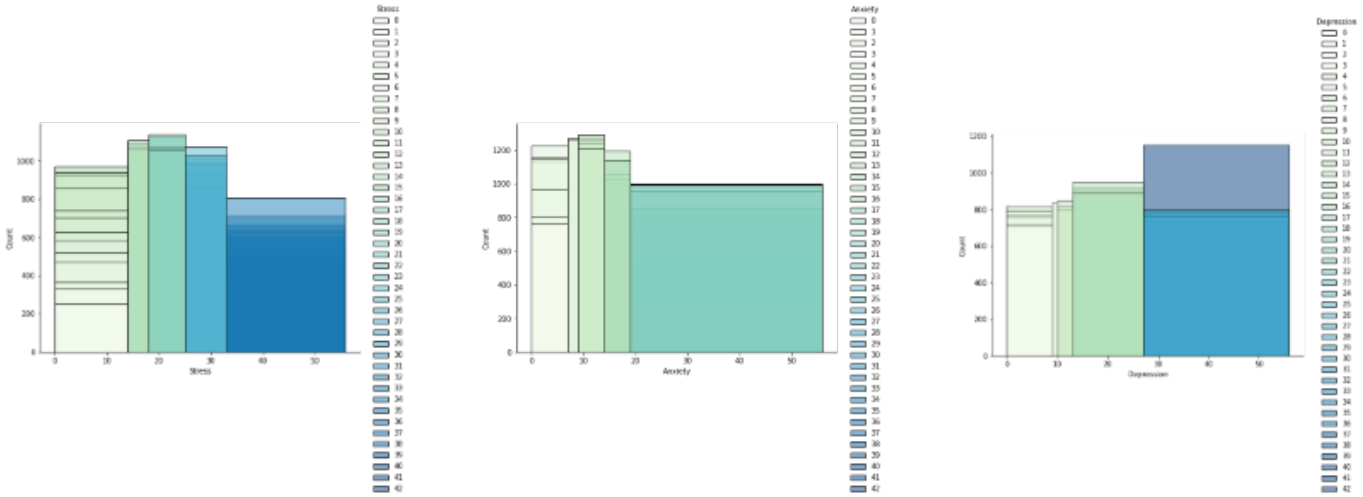
Personalities Types

Personality types were assigned as described in [5]. For each affirmation, the subject was asked to rate how much they saw themselves on a seven-point scale.

Descriptive Statistics

Distribution of Depression, Anxiety and Stress Severity

As you can see from the distribution of depression severity (figure 2c), a high number of participants fall into the extremely severe category. In contrast to the distribution of the severity of depression, the distributions of the severity of anxiety and stress (figures 2b and 2a) shows that most people fall into the categories of mild, moderate, and severe.



(a) Distribution of Stress Severity (b) Distribution of Anxiety Severity (c) Dsitribution of Depression Severity

Figure 2: Dsitribution of DASS Scales

DASS categories and Personality Traits

As can be seen in all three figures 3, 4, 5, the average score for all five personalities decreases as the severity of the core symptoms increases.

The mean score of the Emotional Stability trait has the steepest downward slope of all the other traits, suggesting that emotional instability is linearly correlated with the severity of mental illness. Likewise, the figures 4, 3 shows that openness to new experiences can be a good indicator of the severity of anxiety or depression, since the symptoms of the two are strongly correlated.

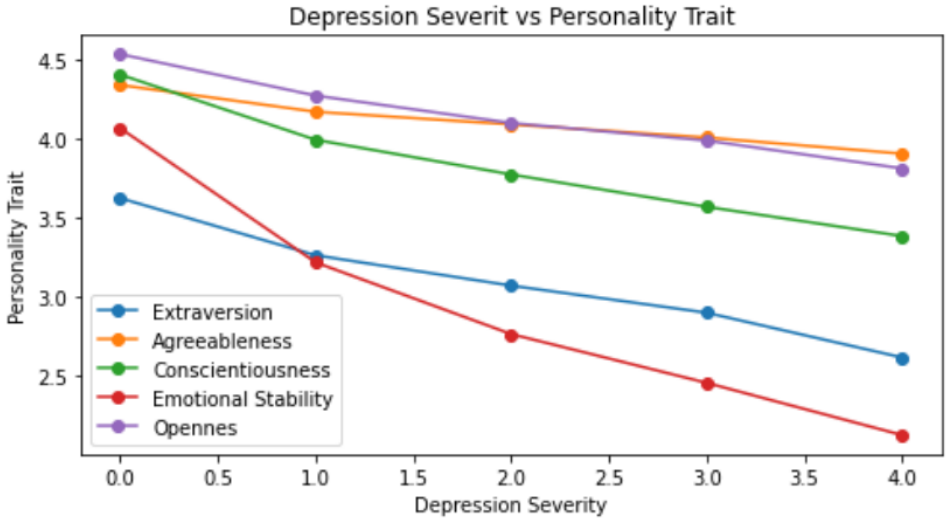


Figure 3: Depression Severity and Personalitiy Trait

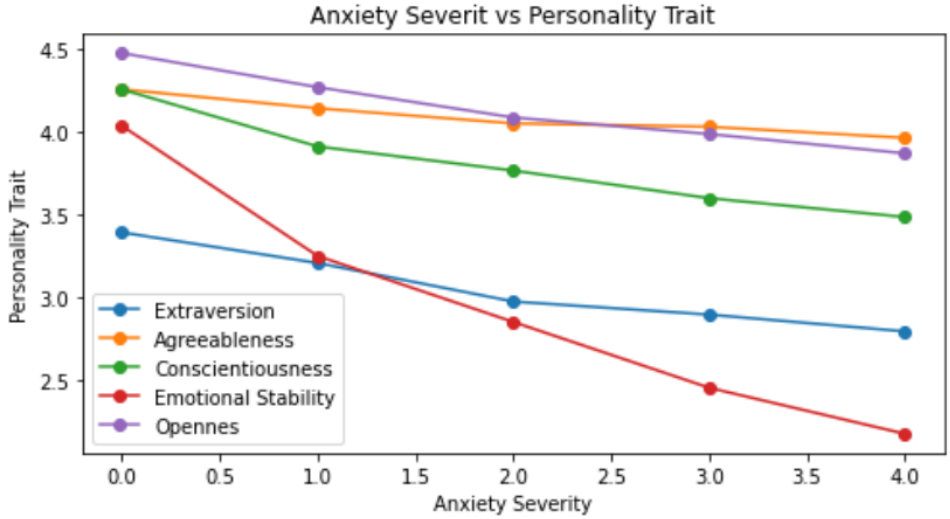


Figure 4: Anxiety Severity and Personalitiy Trait

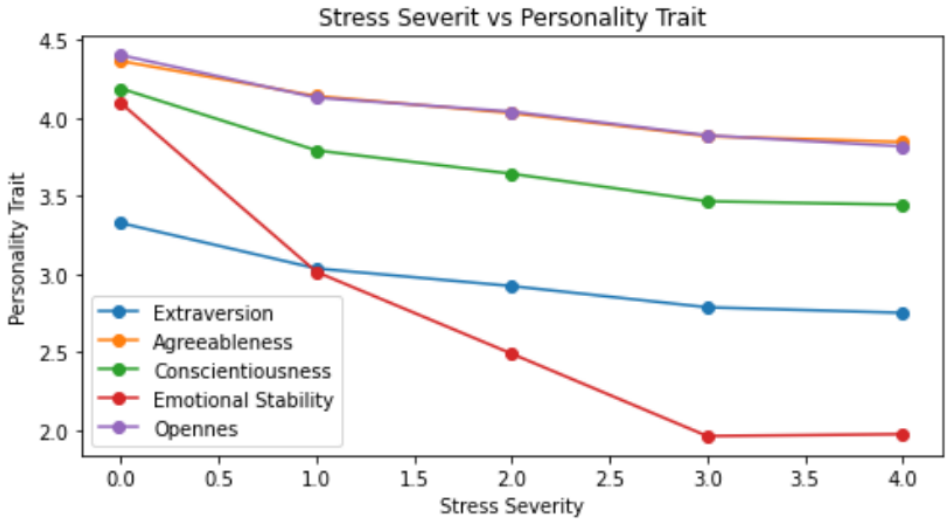


Figure 5: Stress Severity and Personalitiy Trait

Correlation matrix

We plot the correlation matrix (figure 6) for a subset of columns to see which variables can be selected as good features.

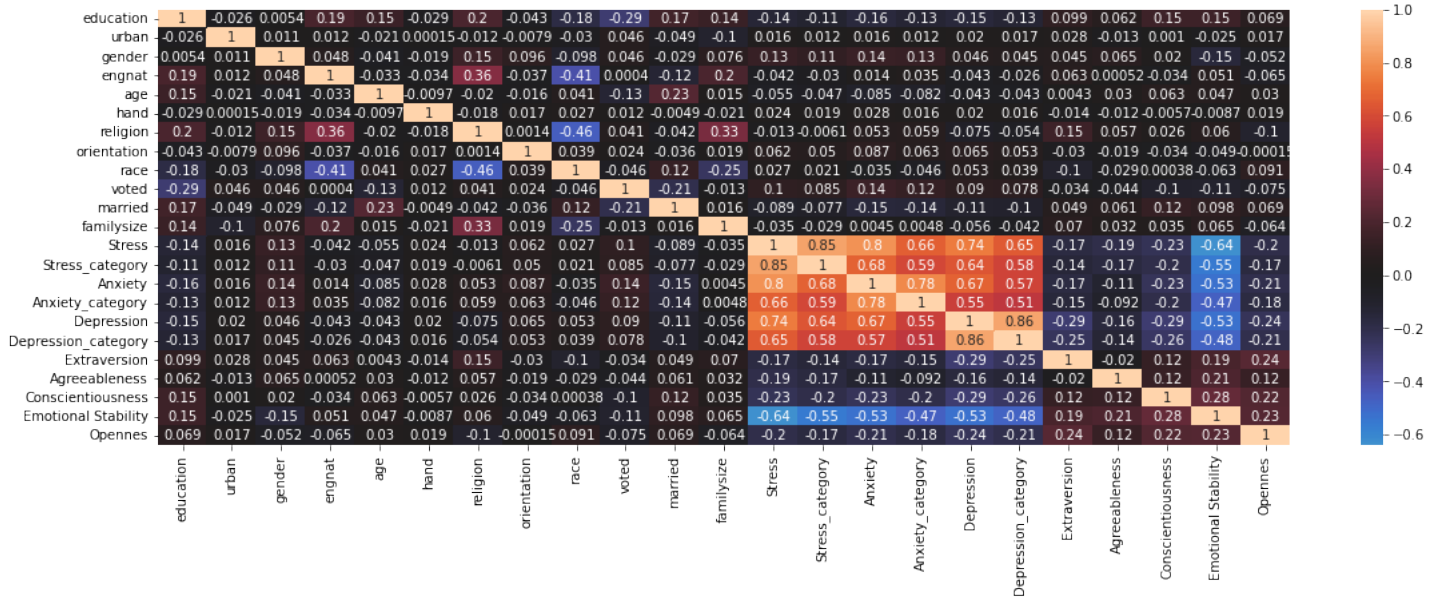


Figure 6: Correlation matrix for numerical features

The following steps are applied to pre-process the data:

1. Label the categorical features (the answers to questions Q1A to Q42A are 'Never', 'Sometimes', 'Almost Always', 'Always', and the country column has values of string type)
2. Scaling the features using the Standard Scaler
3. Creating a subset of the data and applying the dimensionality reductions algorithms.

Dimensionality Reduction

In dimensionality reduction process, the number of random variables considered is reduced by identifying a set of principal variables. This step is also important for visualisation purpose with the PCA and t-SNE algorithms, but we will also use the output of the PCA algorithm as a feature for our models.

Experiments

To ensure that all is alright with our data and the pre-processing phase, and to have a baseline for comparing our model, we will compare the results of the clustering algorithms with the random chance and the results of a Support Vector Machine model.

Cluster patients with depression based on the severity of their core symptoms

KMeans

First, we will use the Elbow method to determine the number of clusters to be formed.

The Elbow method (figure 7) shows that five is the optimal number of clusters for the K-Means algorithm, which is also the known number for the severity scales of the core symptoms of depression.

The first clustering algorithm we will examine is the K-means++ algorithm. We set the number of cluster to five and used the default hyperparameters. We used all the columns as features and record the results for the unormalized data, normalized data, unormalized data reduced to 3 components using PCA algorithm, and normalized data reduced to 3 components using PCA algorithm.

For the evaluation part, we match the unordered clusters labels with the corresponding class by calculating the confusion matrix confusion_matrix. confusion_matrix[i][j] is the number of samples of class i that were assigned to

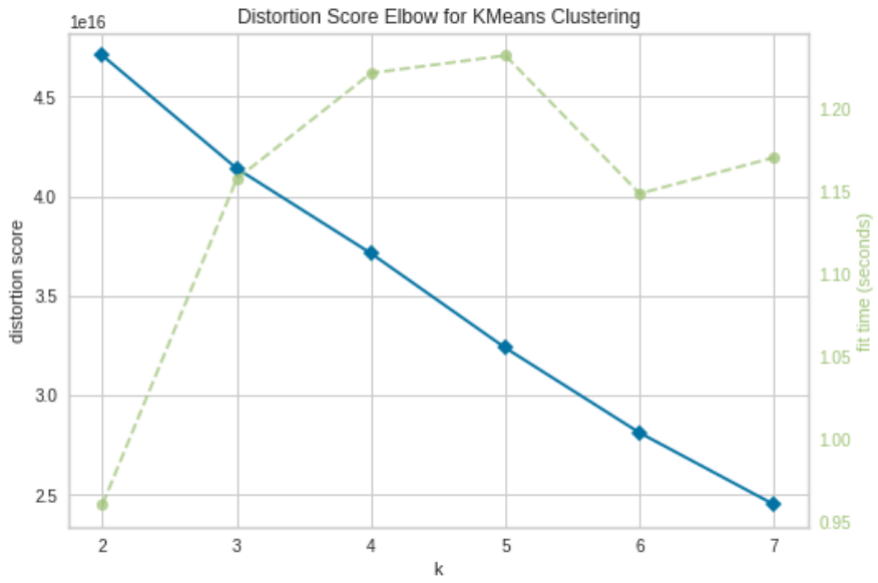


Figure 7: Elbow method for determining the number of clusters

Method	Accuracy
Random chance	0.19995
Supervised method (Support Vector Machines)	0.43186
KMeans (default parameters and unnormalized data)	0.43186
KMeans (default parameters and normalized data)	0.35336
Kmeans (default parameters and dimension reduction on unnormalized data)	0.42989
Kmeans (default parameters and dimension reduction on normalized data)	0.29223

Table 2: Results for the KMeans method

cluster i in the training set, and we perform the matching using the linear_sum_assigment method[4]. This method is commonly used as a solution to the classical assignment problem where the cost matrix is rectangular. This method requires a cost function to calculate the best matching. To do this, we feed the negation of the confusion matrix $(-1. * confussion\_matrix)$.

The results for the test data (randomly selected data from 25% of the entire data set) are shown in the table 3.

DBSCAN

The second clustering algorithm we will examine is the DBSCAN algorithm.
Steps taken for the evaluation of the DBSCAN model:

1. Evaluate the distribution of the samples in the training data and use them as a reference point when deciding on the hyperparameters model. The chosen metric for this task is the *Euclidian distance*.
2. Perform grid-search to find the distance for the optimal number of clusters.
3. Evaluate models allowing multiple clusters to have assigned the same label.

Method	Accuracy
Random chance	0.19995
Supervised method (Support Vector Machines)	0.43186
DBSCAN (unnormalized data)	0.426768
DBSCAN (dimension reduction on normalized data)	0.42979

Table 3: Results for the DBSCAN algorithm

Interpretation of the results

To evaluate the two models, we drew random samples from the clusters after clustering and calculated the score for these clusters using only the depression questions. Most of the samples had scores from the same bin, which means that the algorithm succeeded in clustering the patients correctly.

Conclusions

In the case of the DBSCAN algorithm, instead of using the high-dimensionality dense data, reducing the dimensionality of the features to three components improved the performance of the algorithm.

From the experiments performed, both algorithms are sensitive to the number of clusters (setting the cluster number parameter in the case of KMeans and setting the eps hyperparameter according to the desired number of clusters in the case of DBSCAN). The best performance is obtained by setting the number of clusters equal to that found by the Elbow method.

References

[1] Depression, anxiety and stress scale, url: http://www2.psy.unsw.edu.au/dass/.

[2] Depression, anxiety, and stress scales, url: https://www.psytoolkit.org/survey-library/depression-anxiety-stress-dass.html.

[3] Overview of the dass and its uses, url: http://www2.psy.unsw.edu.au/dass/over.htm.

[4] Reference guide -scipy.optimize.linear_sum_assignment - scipy v0.18.1.

[5] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. A very brief measure of the big-five personality domains. *Journal of Research in personality*, 37(6):504–528, 2003.