Depression, Anxiety and Stress: Tentative Prediction of Mental Health Condition

Noemi Carolina Guerra Montiel
CS577 Principles and Techniques of Data Science
SDSU San Diego State University
San Diego, USA
nguerramontiel0193@sdsu.edu

Nahla Mohamed Elshafey
SDSU San Diego State University
San Diego, USA
nelshafey2214@sdsu.edu

***Abstract*—In the current society, all over the world, mental health problems are rising among people. Depression, Anxiety and Stress are some of the top psychological illnesses ailing humanity. Thankfully, access to mental health support is increasingly available through a multitude of resources, individually or in groups, virtually or in clinics. This project aims to use data science principles and techniques to provide a tentative prediction of the condition of individuals regarding their susceptibility to symptoms of depression, anxiety and stress. This prediction is implemented through the Depression, Anxiety and Stress Score (DASS), facilitated through the response of the identified individual to a questionnaire and collecting personal and demographic information about this individual. Achieving a 92% accuracy for this prediction using a Random Forest Classifier, this implies a successful approach that can be taken further by using it in applications and by clinical professionals.**

Keywords—DASS-42; depression; anxiety; stress; data science; Notebook; predicting; modeling

I. Introduction

According to the World Health Organization and the World Bank, mental illness is a leading cause of disability worldwide [1]. In America, major depression is the leading cause of disability and it is estimated that 1 in 5 Americans experience a mental illness in a given year [2]. In order to take action on the significant impact on society and public health associated with mental health, it becomes relevant to analyze tools to detect early symptoms that can help diagnose and treat people in an effective manner. One of these tools is a survey called DASS-42, which is a 42 item questionnaire that measures the emotional states of depression, anxiety and stress. It was designed to provide an introductory clinical assessment on the severity of a patient's symptoms and it has shown excellent reliability and validity on clinical and non-clinical samples.

Throughout this project, a dataset containing different responses of the DASS-42 was used to create some models that help predict the severity of a given illness through the calculation of its total score on the questionnaire. This analysis and prediction was developed in a Jupyter Notebook where each step of the process will be detailed, starting by preparing and understanding the data through data cleaning and transform techniques like feature engineering and construction. Then, some visuals were created to compare and evaluate the similarities between variables, as well as finding their correlation, which led to the discovery of patterns and features of interest. Finally, some predicting models were implemented and evaluated to find the one that presents the highest levels of efficiency, precision and accuracy. It was discovered that between Gaussian Naive Bayes, Random Forest and AdaBoost Classification, the model that offered the best predictions was Random Forest.

II. DASS-42 Questionnaire

The DASS-42 questionnaire is widely used as a first assessment tool to measure the probabilities of having a mental health condition as depression, anxiety, or stress. Its main objective is to assess the severity of the reported given condition and be a means by which a patient can get its appropriate treatment. If the patient reports a high score on one or more areas, it can be interpreted as a sign of high intensity levels of the patient's anguish or pain, and similarly if the scores are low, there may not be much distress, thus a strong treatment may not be encouraged [3].

A. Questions

As the name suggests, there are 42 questions with a 4-point scale that go from 0 ("Did not apply to me at all") to 3 ("Applied to me very much, or most of the time"). Each question refers to a different illness and so the following key is used to calculate the total score of each condition [3]:

Depression: 3, 5, 10, 13, 16, 17, 21, 24, 26, 31, 34, 37, 38, 42

Anxiety: 2, 4, 7, 9, 15, 19, 20, 23, 25, 28, 30, 36, 40, 41

Stress: 1, 6, 8, 11, 12, 14, 18, 22, 27, 29, 32, 33, 35, 39

When the corresponding scores are added, the results indicate the intensity of the condition according to the following metrics represented in Table 1:

TABLE I. Severity of each condition

Severity	**Condition**		
	Depression (D)	***Anxiety (A)***	***Stress (S)***
Normal	0-9	0-7	0-14
Mild	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Severe	21-27	15-19	26-33
Extremely Severe	28+	20+	34+

It is relevant to mention that the dataset that was used included the results for the 42 questions plus two additional features including the time in milliseconds that the person took to answer the question and the position on the survey. The first one can be used to filter out the people whose answers took too long or too little, but the second one is mainly useless for the purposes of this work.

B. *Demographic and personal information*

In addition to the 42 questions used to get a score regarding each condition, the survey includes a section of questions that consider some demographic and personal characteristics of the survey taker. These include the subjects major, country of residence, age, education, religion, marital status, race, among others. These features will be of great importance later to analyze their correlation between the resulting scores. They could represent a relevant factor to use in the models in charge of making the predictions with the highest accuracy possible. Additionally, there are 10 personality options named TIPI that were also included in the modeling to see their relationships with the score results.

Among the less important features of this category that were not included in the analysis of the data were:

1) Validity check words (VCL list): a list of real and made-up words that were added to improve the validity of the survey.

2) Recorded durations: some features that included the time spent on each page.

3) Others: vote, family size, screen size, unique work location, and hand used to complete the survey.

III. Methodology

The project was divided into three main sections.

C. *Data importing and Preprocessing*

The dataset that was used is the "Depression Anxiety Stress Scales Responses" dataset by *Open psychometrics* [4]. Its data was collected from 2017 to 2019 to measure the three negative emotional states/disorders of depression, anxiety, and stress. In addition, it has a total of 172 features and 39775 records, and it mainly consists of numerical features. However, it possesses 2 string values, which are *major* and *country*.

For the data preprocessing, the data was cleaned and canonicalized to prepare it for further analysis. For the cleaning stage: the people that answered the survey questions too slowly or too quickly were eliminated, the extreme ages were replaced, the columns with additional information about the 42 questions (like their time and position) were deleted, the VCL and other irrelevant features were dropped, and the resulting categories for the demographic and personal questions were standardized with values from 0 to 3.

Furthermore, for the canonicalization stage: missing, nan, and incomplete values in the "major" feature were replaced for the string "No degree". In addition, all values were transformed to lowercase strings and the 5 most frequent majors had as much string standardization as possible.

D. *Data Analysis and visualization*

The first thing in this section is the official calculation of the DAS score and the creation of a classification feature that stores the severity of each condition. To perform the DAS score calculation, the dataset was divided into three, with each part representing a mental health condition. Figure 1 represents the results of the number of answers associated with each intensity. Most answers present severely extreme intensity levels of depression and anxiety, followed shortly after by normal levels of each illness. This indicates that the given scores are balanced enough and that there is enough distribution of data to perform good predictions.

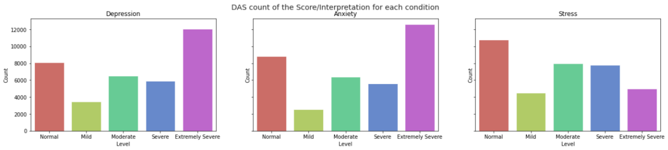

Figure 1. Count of each intensity classification for every condition

Next, a correlation matrix for each data frame was constructed between all the features to see which of them presented the most similarities with the condition level classification since it is the only independent variable aside from the score. Besides the obvious strong correlations with the 42 questions, it was found that some of the personality questions or TIPI were negatively correlated to the scores. For example, the one that represents calmness and emotional stability had the highest negative correlation with the score and condition intensity, which is not surprising since the less emotionally stable a person is, the higher their chances of getting a higher score. Other personality traits that also presented higher negative correlations were: TIPI1 (Extraverted-enthusiastic), TIPI3

(Dependable-self-disciplined) and TIPI5 (Open to new experiences). As for personal information, the most negatively correlated features were age and education, probably because people get jobs depending on their age and education level, which has a direct impact on their lifestyle and overall happiness.

Afterwards, a comparison and visualization of some analyzed features is done with a more intense approach on the demographic and personal questions. The created graphs were separated into the following categories:

a) *Education level:* Figure 2.1 shows that in the education levels prior to high school, there is not much mental illness and students tend to be more or less relaxed. However, the numbers increase greatly when they enter high school and undergraduate college, and as an example the graph shows that the three analyzed illnesses tend to present the highest number of cases on the 'Extremely Severe' intensity level during this level of education. After high school, the DAS intensity levels start to slowly decrease and during graduate studies they seem to resemble the relaxed intensity levels of the first school years.

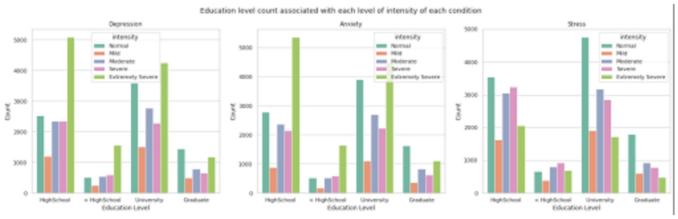

Figure 2.1. Count of each intensity classification related to education level

b) *Gender:* The people that took the survey were mostly female and they present a high rate of "Extremely severe" depression and anxiety. Even though there are not many males represented, the graph shows that they tend to have lower scores than women in all three conditions.

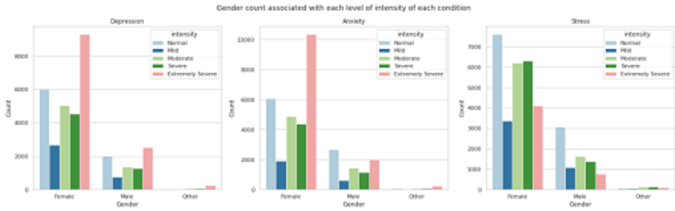

Figure 2.2. Count of each intensity classification related to gender

c) *Age group:* The age group feature is formed with 8 different categories that go from people with less than 20 to more than 60 years old. Figure 2.3 shows that the majority of survey respondents are young adult people from around 20 and 24 years old. It also shows that younger people tend to have higher scores (especially with depression and anxiety) and that the scores decrease as people get older.

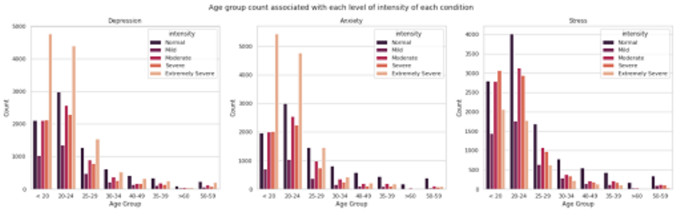

Figure 2.3. Count of each intensity classification related to age group

d) *Religion:* Most people who took the questionnaire were Muslim (class 10) and figure 2.4 shows the same behavior as the previous features where anxiety and depression have higher intensity levels than stress.

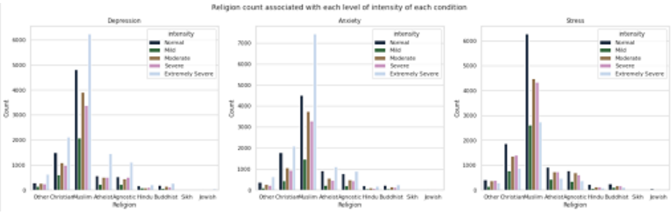

Figure 2.4. Count of each intensity classification related to religion

e) *Sexual orientation:* Most respondents are heterosexual. In Figure 2.5, it can be seen that heterosexual people tend to have more "Normal", "Mild" and "Moderate" condition levels scores than the ones that belong to the LGBT community.

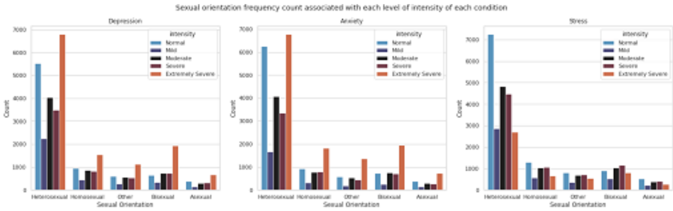

Figure 2.5. Count of each intensity classification related to sexual orientation

f) *Race:* Most respondents are Asian, followed by white and others. Figure 2.6 shows that the predominant categories have similar distributions for the different condition intensities.

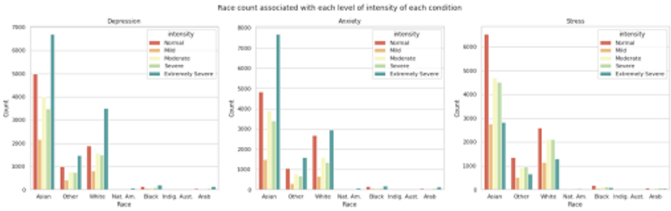

Figure 2.6. Count of each intensity classification related to race

g) Marital status: Figure 2.7 shows that most respondents are single and that they tend to have higher levels of depression and anxiety than married people.

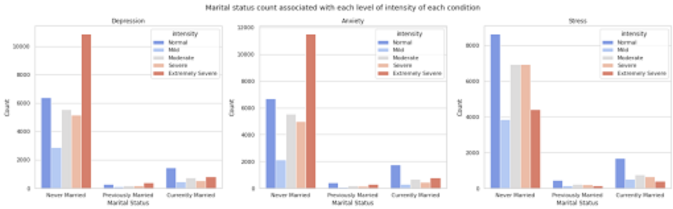

Figure 2.7. Count of each intensity classification related to marital status

Finally, the three previously analyzed data frames are merged to form one unique dataset with a total DAS score to try to make more general predictions of the severity of a person's mental health. In order to use classification models, the total score gets divided into the same 5 categories that have been mentioned throughout the project.

E. Data Analytics

In this section, different modeling techniques are used to find the best possible predictions of the severity of each condition, as well as an approximate prediction of the intensity level of the complete situation when the sum of the DAS scores is taken into consideration. There were three supervised categorical evaluation methods taken into consideration: Gaussian Naive Bayes, Random Forest Classification, and AdaBoost Classification. The procedure to implement each technique was the following:

a) Split the data into training and testing set

The first step to start modeling is to separate the data into a training and testing set in order to avoid overfitting and calculate important evaluation metrics like accuracy and f1 score. The 'y' or ground truth label is the classification of the severity of the condition, while the 'X' set contains the rest of the features that were previously reviewed. After defining 'X' and 'y', their contents get split with 80% going to the training set and the remaining 20% to the testing.

b) Apply a modeling technique (Gaussian Naive Bayes, Random Forest, and AdaBoost Classification)

The sklearn library was used to apply the different models into the training and testing set, which makes the training and learning process easy. The model is chosen, it gets fitted with the 'X' and 'y' train data and the predictions are made with the 'X' testing set. The same tool can be used to get the evaluation metrics and assess the model's performance.

This procedure was applied four times in total, one for each condition and an additional one for the complete cumulative DAS score.

IV. Evaluation

Evaluation metrics have a huge role in achieving the optimal model [5]. The main evaluation metrics used to measure the certainty and efficiency in the classification predictions were accuracy and f1 score.

F. Accuracy

The accuracy evaluation metric measures the ratio of correct predictions over the total number of instances evaluated. Its formula can be described as:

$$\frac{tp+tn}{tp+fp+tn+fn} \tag{1}$$

Where $tp = true\ positive$, $tn = true\ negative$, $fp = false\ positive$, and $fn = false\ negative$

Table 2 shows the accuracy of each model as it is applied to its corresponding mental health illness data frame.

TABLE II. Accuracy Evaluation Metric

Data frame	**Model**		
	Gaussian NB	***Random Forest***	***AdaBoost***
Complete DAS	39	42	42
Depression	86	92	75
Anxiety	81	86	70
Stress	84	90	70

Figure 3.1 represents a visualization of the different accuracies of each model.

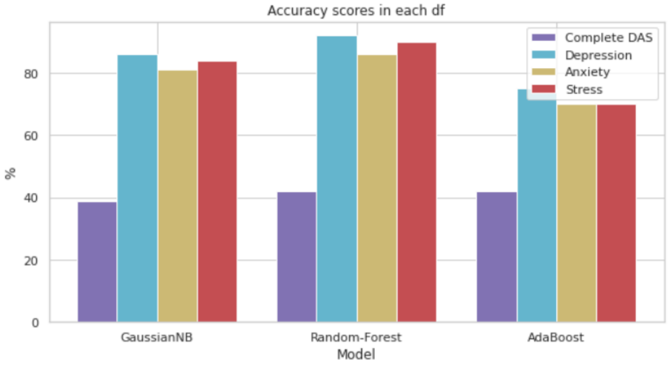

Figure 3.1. Accuracy scores for each DAS data frame

It can be seen that the most accurate model is Random Forest, followed by Gaussian Naive Bayes and lastly the AdaBoost Classifier.

G. F1 score

The F1 score evaluation metric represents the harmonic mean between the recall and precision values. Recall measures the fraction of positive patterns that are correctly classified, while precision measures the positive patterns that are correctly predicted from the total predicted patterns in a positive class.

The recall and precision formulas can be described as:

$$r = \frac{tp}{tp+tn}, \qquad p = \frac{tp}{tp+fp} \tag{2}$$

Thus, the F1 score formula is:

$$F1 = \frac{2*p*r}{p+r} \tag{3}$$

Table 3 shows the f1 scores of each model as it is applied to its corresponding mental health illness data frame.

TABLE III. F1 Score Evaluation Metric

Data frame	Model		
	Gaussian NB	***Random Forest***	***AdaBoost***
Complete DAS	38	42	42
Depression	87	92	76
Anxiety	82	84	72
Stress	85	89	71

Figure 3.2 represents a visualization of the different f1 scores of each model.

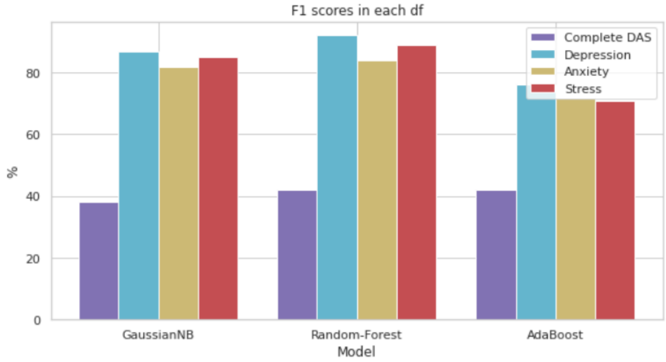

Figure 3.2. F1 scores for each DAS data frame

It is observed that the model with the highest F1 score is also Random Forest with AdaBoost being the lowest.

V. Related Work

H. DAS Prediction

We compared our implementation with the results found in python notebook [6], titled DAS Prediction, which used the same dataset to predict the condition of each survey answer regarding their DAS scores. Table 4 demonstrates that we were able to achieve better or similar accuracy and/or F1-score for the two models: Random Forest and Gaussian NB.

TABLE IV. COMPARISON WITH SIMILAR WORK

Model Name	**Random Forest**						**Gaussian NB**					
Evaluation Metric	Accuracy			F1-Score			Accuracy			F1-Score		
	D	A	S	D	A	S	D	A	S	D	A	S
Other	92	84	89	92	82	88	87	81	85	87	81	85
Ours	92	87	89	87	82	85	86	81	84	92	84	89

Conclusions

This work presents a 92% accuracy in predicting Depression, Anxiety, and Stress conditions in individuals of various ages, according to their replies to questions about their personality, situation and hypothetical scenarios. This could be a stable foundation for professional clinical psychologists to work on/with to assess the individual more thoroughly.

Future work can focus on the questions regarding the top 10 personality traits and how these responses can affect the prediction and if personal characteristics can generally dictate mental health issues. Improved methods or other threshold basis can be further investigated as well to predict the outcome of the total DAS score with higher accuracy.

References

[1] World Health Organization, “Depression”, https://www.who.int/news-room/fact-sheets/detail/depression (accessed Dec. 10, 2022).

[2] G. Norquist and S. Hyman, “Advances in understanding and treating mental illness: Implications for policy”, *Health Affair*s, vol. 18, no. 5, September 1999, https://www.healthaffairs.org/doi/abs/10.1377/hlthaff.18.5.32.

[3] S.H. Lovibond and P.F. Lovibond, "Manual for the Depression Anxiety Stress Scales", *Psychology Foundation*, (2nd ed.), 1995.

[4] L. Greenwell “Depression Anxiety Stress Scales Responses.” https://www.kaggle.com/datasets/lucasgreenwell/depression-anxiety-stress-scales-responses, 2020, (accessed Nov. 18, 2022).

[5] M. Hossin and M.N. Sulaiman, “A review on evaluation metrics for data classification evaluations”, *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, vol.5, no.2, March 2015.

[6] Teju, “DAS Prediction”, Kaggle Notebook, https://www.kaggle.com/code/teju4405/das-prediction#Model-Creation-for-predictions, (accessed Nov. 18, 2022)