## GitHub URL

<https://github.com/doneill83/UCDPA_DerekONeill.git>

## Abstract

Mental health is a growing issue in today’s society. This has been magnified by the emergence of a global pandemic and all the societal changes that has brought.

This project aims to use a data bank of an online questionnaire to see if depression and level of depression can be predicted by the answers given or the demographics of respondents.

Exploratory data analysis will be used along with correlation matrices. Three supervised learning models will be used to ascertain if it is a viable prediction model.

## Introduction

Depression, stress and anxiety can have a high impact on how a person’s quality of life. They can have both physical and psychological such as impacting self-esteem, fatigue, and affecting the immune system. The effects of depression, stress or anxiety can be wide ranging and quite damaging.

Depression is a common illness worldwide. Approximately 3.8% of the population of the world are affected by depression with 5% among adults. Depression is one of the leading causes of disability and the second leading cause of death among 18-29 year olds.1

A recent World Health Organisation report has concluded that while there is an increase in mental health awareness, there is a worldwide failure to provide mental health support.2

The Covid-19 pandemic has brought significant changes to the lives of billions. There is however a real lack of empirical studies to understand the impact this pandemic may have had on depression, anxiety or stress and may continue to have as the pandemic subsides and we move towards a ‘new normal’. One such way to quickly and cheaply try to understand the changing needs of the population and aim to effectively and efficiently bridge the gaps in mental health support may be to use self-reporting questionnaires as a guide.

## Dataset

The dataset I have chosen for my project is called Predicting Depression, Anxiety and Stress and is available on the Kaggle website.3

The dataset utilises responses to an online version of the Depression, Anxiety Stress Scale (DASS) which is based on the Taylor Manifest Anxiety Scale. The online survey was available to anyone. At the end of the survey, participants were given the option to complete a short research survey. The dataset comes from those who completed the research survey and gave consent for their data to be used for research purposes. The data was collected between 2017 and 2019.

The data contains 42 questions with answers on a 4 point scale. The time taken to answer each question is also recorded. The total time of the test along with time on the landing page is also recorded. Responses to the Ten Item Personality Inventory and a variety of other questions were also included. There are 39,775 rows in the dataset with 172 columns.

I work in a public health research environment, so this dataset really piqued my interest. Mental health has been a growing concern as awareness develops and funding remains underwhelming. Anecdotal evidence within my work suggests that Covid-19 and the societal changes attached has had a major impact on mental health across all ages. Exploring this dataset form a predictive and machine learning standpoint could perhaps lead to more sophisticated self-reporting in the future and a means to quickly identify those most in need of support. The dataset had a high usability score and plenty of notebooks and comments available to try to learn from.

## Implementation process

In order to do any analysis of the dataset it first needs to be downloaded and imported. The file was available in csv format and imported into a pandas dataframe. Other datasets relating to mental health and other public health areas were explored before ultimately deciding to use this dataset. This stage of the process was very much about trying to understand the vast amount of packages and libraries available in python. Some of the datasets I was investigating required importing the json library and using the json.load command. I also experimented with using the beautifulsoup4 package. This allowed for web scraping by passing a html/xml document through the relevant parser, turning the content into a complex tree of python objects. This will be useful for me in future when scraping sports statistics from websites for analysis. The data contained some bad rows so I skipped them in the upload using on\_bad\_lines =’skip’.

For this project, I utilised a lot of packages based on code I was sampling from other notebooks. It was a great learning experience to do this and an eye opener to the sheer variety and power of the python language.

As the dataset was quite large and impossible to decipher without the codebook, it was important to get a basic understanding of the shape of the data and the columns and format. I used standard commands for this but changed the options to allow me to read the whole file. A check for nulls and a basic description of the file was performed before moving on to some cleaning.

Rows were removed where participants took too long or too short a time in answering questions. Outlier ages were also changed to the mean using the loc function.

A new variable was then created to group ages into category. This was an ordered list computed through the ‘age’ column. List is appropriate here as it is an ordered value. A lambda function was also used here. The lambda function is useful in data manipulation.

The shape of the data is then output in a formatted style showing total participants and the number of questions.

Target DASS columns were then created using a dictionary. A dictionary was appropriate here as a dictionary holds key:value pair. Each key is separated by a comma with the value separated by a colon. This allowed me to map questions to Depression, Anxiety and Stress and create bins based on the scoring to mimic the DASS model. This required a for loop iterator to subtract one from the scores so they matched the original scoring breakdown. Regex, map function and filter were used to achieve this. Personality types were then added to the data. This involved calculation of the big five personality types according to Gosling et al.4 Using the tipi columns and the DASS values calculations were made to create character variables.

Table

Description automatically generated

Figure 1: Interpretation of Dass scores.

Some basic data visualisation was then done. This involved using the variables to make pie charts showing the distribution of participants by various demographics. This may show the possibility of bias in the data.

A correlation function was then done using the character and DASS variables previously created to create a correlation heatmap. A second correlation heatmap was done using the DASS score and variables. This was to try to view significant variables.

Moving on from this I then created several box plots to try determine significant variables.

Several functions were created to parse the codebook and map questions using dictionaries. These are used in the machine learning testing.

For the machine learning I did supervised learning using the questions in the codebook. The targets were Depression, Anxiety and Stress and their associated created categories. The test size was 0.3. I used three supervised models.

The first model I used was Ridge regression. In the case where linear regression models suffer from overfitting, it is possible that the features (variables) included in the model are highly correlated with other variables in the model. A model that suffers from multicollinearity produces biased estimates and it does not generalise well to a test data set. A ridge regression is a technique for analysing multiple regression data that suffer from multicollinearity, which leads to overfitting. I also used K-nearest neighbours modelling and support vector regression model.

Lastly I used xgboost to boost the modelling and attempted to visualise the feature importance in predicting depression.

I also tried to do feature importance analysis and while I did get outputs I struggled to interpret them.

## Results

Graphical user interface

Description automatically generated with low confidence

Looking at the distribution by certain demographic variables there is some cause to be wary of bias in the data. Age is predominantly in the 20-24 category. Gender is heavily skewed towards female. Race is almost 60% Asian and, perhaps linked to this, over one in two of participants define their religion as Muslim.

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Correlation tests were run. Our computed variables from the DASS model and the 5 major personality types were used to determine correlations. Depression, Anxiety and Stress are all highly correlated.

Graphical user interface, application

Description automatically generated

A heatmap was also done to determine how the correlation between DASS scores and variables.

Chart

Description automatically generated

Boxplots were then run which did hint at some correlations around ethnicity and gender however these are inconclusive especially with the heavy weighting towards female and Asian in the data.

Trying to use machine learning to predict depression led to robust modelling.

Ridge regression modelling had an ROC-AUC score of 0.997 which is excellent, while SVR was 0.614 and KNN was 0.532. Using xgboost the model was boosted to 0.9981 with an accuracy score of 0.98.

I had also tried to perform feature importance analysis but had difficulty understanding how to determine redundant variables.

## Insight

* Higher education appears to mean lower depression scores.
* Age seems to also be a factor in depression scoring with older ages scoring lower.
* Depression, anxiety and stress are all highly correlated
* The machine learning models and the boosted models suggest there is a robust possibility to predict depression levels based on this line of questionnaire. While SVR and KNN are not much above random, Ridge regression produced an excellent score with the boosting aiding it.
* The models may suffer overfitting due to high correlations between depression, anxiety and stress but ridge regression accounts for this.

## Personal insights

* The sheer number of options with regards to plotting and models to use can be overwhelming.
* The language seems to be constantly moving. So much of code I tried to use, pycharm gave back errors saying it was obsolete or about to be obsolete.
* The python community is amazing. There is so much code and articles available to try help.
* This is just a starting point. It requires a lot of practice on a regular basis.

## References

1. <https://www.who.int/news-room/fact-sheets/detail/depression>
2. <https://www.who.int/news/item/08-10-2021-who-report-highlights-global-shortfall-in-investment-in-mental-health>
3. <https://www.kaggle.com/yamqwe/depression-anxiety-stress-scales>
4. <http://gosling.psy.utexas.edu/wp-content/uploads/2014/09/JRP-03-tipi.pdf>.