Predicting Mental Health Conditions Using Student Demographic Information

Ashwith Yarram

4/18/2024
Inspirit AI research paper

Abstract

Type your abstract here, no more than 150 words

This paper describes models that predict three mental health conditions in students. These models can be used to detect early warning signs for panic attacks, anxiety, and depression and therefore help students get the help they need. Classification data was obtained on student demographic information and whether each student had depression, anxiety, and/or panic attacks. The data was from kaggle and it showed the students and their year of study, what major, and their relationship status, age, and cgpa. A classification model was built. Results showed a 0.71 for panic attack accuracy in predicting whether the person had stress or panic attacks.

1. Introduction

Motivation for why detecting mental health conditions is important since it can help detect if someone has depression, anxiety, or panic attacks. It is not easy for the human mind to figure out if a person has these symptoms. It is possible, but it takes a lot of work and time to figure out. The models would make the job easier and in the future, we could implement solutions into the programs on how to treat each type of mental illness.

Unfortunately, a lot of people have a tough time finding mental health care in the medical industry. Many people can't find spots in therapy sessions, and many don't have enough money to afford these sessions. Implementing this code will help them figure out what they are going through.

We are working with supervised data and this is classification data. With this, we will train the data to extrapolate what symptoms the user has based on their demographics and academic/social life. This can help understand what symptoms a person has of certain characteristics. Age, gender, course year, course study, cgpa, and marital status are the inputs of the data. Whether the person is more prone to depression, anxiety, or panic attacks is the output of the model.

2. Background

Anxiety is long term stress and worrying about what will happen. Depression is feeling sad about what has happened. Panic attacks are when you get scared and you start panicking when a problem occurs.

The data is centered around college students aged 18-22 and their information, such as age, gender, course year, course study, cgpa, and marital status. These would indicate what factors could influence these mental health symptoms. Most of the data is in categorical variables. cgpa is in numerical variables.

3. Dataset

To find a data set, we had to search on kaggle.com, a website that was a hub for data. We searched for a long time and eventually found a good, accurate dataset which was relevant to our question and time period.

The dataset is a survey published by a researcher Shariful Islam that shows the person and their ‘input’ information. We would use this library to allow us to upload the data to our code:

```
from google.colab import files
data = files.upload()
```

Then, we would choose our file to upload once we ran this code. Here, we would select our .csv file. It is important to use a csv file, since this file type is compatible with the code.

Then, we would have to access our data, so we would use this function:

```
mentalhealth_data = pd.read_csv("Student Mental health (1).csv")
```

```
    (The pd.read_csv would
```

We had to ‘clean’ the data, which meant that we had to make sure that the unnecessary data was deleted, such as the order of the people. This was basically that number 1 was assigned to the first person of the data set, number 2 for the second person of the data set, and etc. We also had to delete the date of when the data for each person was surveyed because it wasn’t necessary in the regression analysis. Then, we had to change the data sets to fit into our model. This meant that the courses needed to be categorized based on the broader field of study, then changed to numerical data, e.g. assigning 1 to STEM field, 2 to english field, etc.

Once we have done this to every data that doesn’t follow numerical data rules, we can safely use the dataset for the regression analysis.

Here is the link to our dataset: Student Mental Health by Shariful Islam

4. Methodology / Models

The model uses sklearn library, which provides necessary models for splitting datasets into training.

Logistic regression is the process of modeling outcomes based on probable events, based on an input. For instance, if I have an input name, the model could predict what gender the person is based on the data that the suffix of the name resonates with many names that have the same suffix from a specific gender.

The output of a logistic regression is a probability that is from 0 to 1. As the probability is higher, the more accurate the model is at predicting the data.

Mathematically, this would be the logistic regression model.

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$$

P(Y = 1|X) is the probability of the dependent variable Y being equal to 1 given the X value, independent.

e is the natural base log, approximately 2.51.

B0, B1, B2, …, Bn are the coefficients associated with the independent variable X1, X2, …, Xn respectively.

The model changes the coefficients such that it maximizes the probability of the observed data.

Next, the model uses cross validation. Cross validation is when data split into mini samples, then the one sample is not regarded and the rest is studied. Then, the model runs a test using the sample that was disregarded and returns an accuracy score. The accuracies of each of the mini-samples will be averaged to get a more accurate estimate of the data.

This is the mathematical formula:

$$\text{Avg\_Performance} = \frac{1}{k}\sum_{i=1}^{k} E^{(i)}$$

k is the number of folds (groups)

E^(i) is the accuracy of the ith fold (group)

5. Results and Discussion

Show your classification reports for all three models. Compare them.

```
Anxiety acc = 0.7619

#Depression acc = 0.7619

#Panic Attack acc = 0.7142
```

The Anxiety and Depression acc are more accurate than the panic attack acc.

These models show more than a 70% accuracy in their conclusions, which means that their conclusions for over 70% of the data were correct. This is a great step into showing how the regression models can accurately predict whether the person has certain mental health conditions for most of the data sets. There is obviously room for improvement, since sometimes a female who studies medicine may not have experienced panic attacks, even though the data was trained to identify a female that practices medicine to have experienced panic attacks. The categorical data was grouped for course of study, so we don't know the specifics in each specific course of study, but we get a general sense based on the subject area of the course (e.g. medicine, STEM, arts and humanities, business). The model does a very good job weighing multiple factors that can influence a person's chance of experiencing certain mental health conditions.

6. Conclusions

Our model is still rusty, but fairly well at predicting the outcome of a person's mental health based on their personal data. With this, there will be more information to add input to the model to create more accuracy in finding anxiety and depression, and panic attacks.

With more training of data, more exposure to data, and adding more inputs in the future, we could be able to accurately predict the chance of mental health conditions based on demographics. This could help many people understand what type of mental health care they need based on their demographic/personal data. We can help connect people to mental health care through these models. This project serves to show that we can push the boundaries and use these regression models to predict conditions for future patients with mental health conditions.

Acknowledgments

I would like to thank my Inspirit AI+X Mentor John Basbagill, for helping me with my project. I would like to thank the Inspirit AI+X team for giving me this opportunity to do research. I would like to thank my parents for encouraging me to research more in the STEM field.

References

[1]

Islam, Shariful."Student Mental Health." *www.kaggle.com*, August 7, 2020. www.kaggle.com/datasets/shariful07/student-mental-health/data. Accessed March 24, 2024.

[2]

Kim J, Kim H. "Demographic and Environmental Factors Associated with Mental Health: A Cross-Sectional Study." *Int J Environ Res Public Health.* April 17, 2017. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5409632/. Accessed May 5, 2024.