A Comparative Study of Machine Learning Techniques for Health Prediction.

Rosemary Varghese, Anila S, Shyama R

Assistant Professor, CSE
Adi Shankara Institute of Engineering and Technology
Kalady, Kerala

Abstract—Personal well-being refers to both physical as well as mental fitness. In the current scenario of expeditious commercial growth and pandemics, the human race is also challenged by immense psychological pressures. This paper presents the prediction of the most pertinent psychological issues identified by the World Health Organization – Anxiety, Stress, and Depression. Machine Learning algorithms are used for the prediction of the same. The data was previously collected from people in various economic, cultural, and social situations through the Depression, Anxiety, and Stress Scale Questionnaire (DASS21). Three supervised learning algorithms were applied and corresponding confusion matrices were calculated. The accuracies of each model were compared and found that the model with the best accuracy is K-Nearest-Neighbor. In addition, analysis of the results divulged that the models were sensitive to negative results.

Keywords: Decision Tree (DT); Depression, Anxiety, Stress (DASS-21); K Naïve Bayes(NB); Machine Learning.

INTRODUCTION

Humans, in today’s fast-paced modern world, have become ambitious, so as to grow and excel professionally through every possible opportunity. Anxiety, depression, and stress are factors that have become common their daily professional life. The World Health Organization (WHO) results states depression to be the most prevalent mental disorder, and the increase in severity of the condition has led to many studies being focused in this area. Differentiating anxiety, depression, and stress from one another is a difficult task even for machines.

The initial diagnosis is the Patient Health Questionnaire (PHQ); while the Depression, Anxiety and Stress Scale (DASS21), which has 21 questions, is used for screening the patients with symptoms relative to this mental illness.

The main symptoms of Depression are loss of memory; lack of concentration; inability to make decisions, loss of interest in recreational activities, low appetite and weight; feeling of guilt, worthlessness, helplessness and irritation, and also suicidal thoughts as well. The symptoms of Anxiety include irritability, nervousness, and gastro intestinal problems, sense of impending danger, rapid breathing, difficulty concentrating and increased heart rate. The symptoms of Stress are feeling upset or agitated, inability to relax, low energy levels, chronic headaches and cold infections.

Thus, anxiety, depression, and stress have many common symptoms, all of which makes classifying these symptoms, a challenging task for machines.

METHODOLOGY

The paper presents the detection of anxiety, stress, and depression using the DASS 21 questionnaire. The data for training was previously collected via a survey from around 35000 participants and was available in online data repositories. Classifications on the test set have been done using three machine learning algorithms—specifically Decision Tree, Naïve-Bayes, and K-Nearest Neighbors.

I. Participants

There were a total of 39775 responses previously collected via survey from people with heterogeneous situations. The data was put on OpenPsychometrics.org for making it available for psychological research purposes. In addition, data of real-time participants were collected to make the prediction.

II. Questionnaire

The required data was formerly collected through DASS-21 questionnaire and was available online. The questionnaire consists of 21 questions of which 7 are allotted to predict each illness uniquely. In addition to this, a GUI was designed for taking responses from the real-time user. For each question, four possible answers are possible, which are saved as numeric responses, depicted as below:

1. Did not apply to me at all.
2. Applied to me to some degree, or some of the time.
3. Applied to me to a considerable degree, or a good part of the time.
4. Applied to me very much.

The questions are described below:

Anxiety

Dryness of mouth
Difficulty in breathing.
Experience trembling
Felt scared without any good reason.
Close to panic
Aware of the action of the heart in absence of physical exertion.
Worried about panic and make a fool of themselves.

Depression
Felt that there is nothing to look forward to.
Felt wasn’t worth much a person.
Difficult to work up the initiative to do things
Feeling that life was meaningless.
Unable to become enthusiastic.
Felt down-hearted and blue.
Could not experience a positive feeling.

Stress
Touchy
Overreact to situations.
Difficult to relax.
Found hard to wind down.
Intolerant to getting interrupted from what I was doing.
A lot of nervous energy
Getting agitated.

Afterward, responses were numerically encoded with values scaling from 1 to 4. From the saved responses, scores were calculated for each illness class by summation of the values associated with each categorical question.

Once the score calculation is done, the class for which the maximum score was obtained by each individual is identified, and is assigned to the resulting class.

For real-time prediction, inputs are taken from a single person at a time, whose responses are numerically encoded and stored in a file. These values are then passed as a test case to each model.

III. Classification

The machine learning algorithms were implemented in Python using Anaconda Version 3. The models predict whether the person shows symptoms of Anxiety, Stress, or depression. The dataset was preprocessed to filter the responses of 21 questions. 70% of the processed data was taken as the training set and the rest as the test set. The working of the employed algorithms is explained in the successive subsections.

IIIa. Decision Tree

It is a supervised learning technique. This algorithm can be used to solve classification as well as regression problems. A decision tree is a flow-chart-like tree structure. Each internal node or non-leaf node denotes a test on an attribute. The outcome of the test is represented by branches of the tree. The leaf node or terminal node holds a class label. Any number of choices greater than two is possible at each decision point. The following figure depicts a decision tree for the classification of a life form into fish, bird, or mammal.

IIIb. Naïve Bayes

It is a statistical classifier that performs probabilistic prediction, i.e., predicts class membership probabilities. The foundation of the algorithm is based on Bayes Theorem. It can be used in a variety of classification tasks. It assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. That is why the model is called naive. The formula for Naïve Bayes is given as follows:

\[ p(H|D) = \frac{p(H) \cdot p(D|H)}{p(D)} \]  

Where,

- \( p(H|D) \) is the posterior probability
- \( p(D|H) \) is the likelihood of seeing that evidence
- \( p(H) \) is the prior probability of a proposition.
- \( P(D) \) is the prior probability of evidence.

IIIc. K-Nearest Neighbor (K-NN)

A k-nearest neighbor classifier searches the pattern space for the k training instances that are closest to the unknown instance. The training instances are described by n attributes. Each instance represents a point in an n-dimensional space. The values of each attribute are normalized in advance to prevent attributes with initially large ranges from outweighing attributes with initially small ranges. The nearest neighbor can be defined in terms of distance equations of choice. An example is depicted below.
F1 Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (8)

Where,
TP (True positive)-Diagonals of matrix
FN (False negative)-Sum of consistent rows for class (excluding TP of that class)
FP (False Positive)-Sum of corresponding columns for class (excluding TP of that class)
TN (True negative)-Sum of all row and column (excluding row and column of that class)

RESULTS
The three Machine Learning techniques – i.e. Decision Tree (DT), Naïve Bayes (NB) and K-Nearest Neighbour (KNN) were used for the predicting classes of Anxiety, Depression and Stress. The training accuracy was verified using the confusion matrix formulated in Tables 1, 2 and 3.

I. Naïve Bayes
Accuracy = 0.760
Confusion Matrix = 
\begin{bmatrix}
2043 & 134 & 520 \\
564 & 297 & 2737 \\
\end{bmatrix}

TABLE 1
CLASSIFICATION REPORT

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety</td>
<td>0.75</td>
<td>0.76</td>
<td>0.75</td>
<td>2697</td>
</tr>
<tr>
<td>Depression</td>
<td>0.75</td>
<td>0.76</td>
<td>0.76</td>
<td>1660</td>
</tr>
<tr>
<td>Stress</td>
<td>0.75</td>
<td>0.76</td>
<td>0.77</td>
<td>3598</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td>0.76</td>
<td>7955</td>
</tr>
<tr>
<td>Macro avg</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>7955</td>
</tr>
<tr>
<td>Weighted avg</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>7955</td>
</tr>
</tbody>
</table>

II. K-Nearest Neighbors (KNN)
Accuracy = 0.788
Confusion Matrix = 
\begin{bmatrix}
2149 & 64 & 484 \\
503 & 167 & 2928 \\
\end{bmatrix}

\[
P(A/B) = \frac{p(B/A) \times p(A)}{p(B)} \quad (2)
\]

Where,
P(A/B) = Probability of A occurring given evidence B has occurred. This is called posterior probability of A given B.
P(B/A) = Probability of B occurring given evidence A has occurred. This is called posterior probability of B given A.
P(A) = Probability of occurrence of A. This is called prior probability of A.
P(B) = Probability of occurrence of B. This is called prior probability of B.

The equation for calculating the entropy is as follows:
\[
E(S) = \sum_{i=0}^{n} p_i \log_2(p_i) \quad (3)
\]

Where,
P_i = probability of occurrence of predictor I and n stands for total number of attributes.

Accuracy = \frac{\text{Sum of diagonal (TP)}}{\text{Total number of instances}} \quad (4)

Error rate = 1 – Accuracy \quad (5)

Precision = \frac{TP}{TP+FP} \quad (6)
Recall = \frac{TP}{TP+FN} \quad (7)
The comparison revealed that the confusion matrices formed are highly unbalanced. F1 scores are to be considered as an indicator of performance.

Table 1 shows the precision, recall, F1-Score, Support, and Accuracy of the Naïve Bayesian model. The F1-scores obtained for anxiety, stress, and depression in this model are 0.75, 0.76, and 0.77. The average accuracy of the model was found to be 0.70.

Table 2 depicts the performance parameters of the K-NN algorithm. The F1-scores obtained were 0.78, 0.77, and 0.80. Thus the average accuracy calculated was 0.79.

For real-user prediction, the expected accuracy ranges from 0.7 – 0.8.

It was clear that the K-NN model has the highest accuracy and will be the best for the classification in this scenario.

**CONCLUSION**

In the proposed system, three major algorithms were used to predict the different symptoms of anxiety, depression and stress. A particular data has been collected based on a questionnaire which analyzes the common symptoms of anxiety, depression and stress (DASS - 21).

Mainly three different classification techniques were implemented - Decision Tree (DT), K-nearest neighbors (K-NN) and Naïve Bayes. On the basis of comparisons performed, performance of K-NN classifier was found to be the best.

As we are dealing with an imbalanced set of data, the F1 score is considered for evaluating performance of the system.
REFERENCES

3. https://adaa.org/understandinganxiety/depression/symptoms


6. https://doi.org/10.1371/journal.pone.0230389/Ashley E Tate, Ryan C McCabe, Henrik Larsson, Sebastian Lundstrom, Paul Lichtenstein, Ralf Kuja-Halkola (2020) "Predicting mental health problems in adolescence using machine learning techniques".
7. Nan Shi, 1 Dongyu Zhang, 1 Lulu Li, 1 and Shengjun Xu 2 1School of Software, Dalian University of Technology, Dalian 116620, China. 2021” Predicting Mental Health Problems with Automatic Identification of Metaphors”.