2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) | 979-8-3503-9763-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEEICT56924.2023.10157693

Anxiety, Depression and Stress prediction among College Students using Machine Learning Algorithms

Shahid Shabeer Malik
Department of computer science and Engineering
School of Engineering Science and Technology
Jamia Hamdard New Delhi-110062
malikshahid6769@gmail.com

Aneeque Khan
Department of computer science and Engineering
School of Engineering Science and Technology
Jamia Hamdard New Delhi-110062
malikshahid6769@gmail.com

***Abstract*—** Anxiety, Stress and Depression have become common psychological health issues in today's life. While these health issues have severely affected every age group of people, large number of students are suffering from these issues. The most surprising thing about these issues is that people suffering from them cannot figure out if they have one of these health problems. In this paper, we predicted anxiety, stress and depression using machine learning algorithms. In this paper, we predicted the severity levels of anxiety, stress and depression in college students using machine learning algorithms. DASS21 was used to collect data from 400 students. It is a standard questionnaires used to measure the common signs of anxiety, stress and depression. The severity levels were mild, normal, moderate, severe and extremely severe. The classification algorithms that were applied are Support Vector Machine, KNN, logistic regression, decision tree and naive Bayes. Different calculation matrices like accuracy, precision, specificity were used to compare the models. F1 score measure was included because it was found that the classes were imbalanced in the confusion matrix. Thus it helped find the best model for prediction of these psychological problems. After implementing all the algorithms, we found that K-Nearest Neighbour performed best followed by logistic regression.

Keywords—component, Anxiety, Depression, Stress, KNN, Naïve Bayes, Logistic Regression, Decision Tree, Support Vector Machine

I. INTRODUCTION

The ambition of humans to grow professionally, relationship issues, academic failures are some of the common factors that contribute to anxiety, depression and stress in humans. According to World Health Organization, depression is the most common mental illness and has affected more than 300 million people around the globe [1]. Different models have been proposed about how to predict these mental disorders, however it is very difficult for machines to differentiate anxiety, stress and depression from each
Others. Therefore, there is a need of an accurate model for accurate diagnosis.

We have used the Depression, Anxiety and Stress scale (DASS21). It contains 21 questions and is used for filtering the signs associated to these mental issues [4-5].

The signs of depression [2] are memory loss; not able to make decisions; lack of concentration; loss of interest in task and hobbies including sex; increase in body weight and overeating; feeling guilty, worthlessness, low appetite and weight loss; feeling helpless, feeling irritated and restless; suicidal thoughts. Irritability, feeling nervous, weariness, sleeplessness, gastrointestinal problems, fear, high heart rate, sweating, breathing problems and lack of concentration are the main symptoms of Generalized Anxiety Disorder(GAD) [2].

The symptoms of stress [3] are, an inability to relax, feeling upset or agitated, tired, frequent overreaction, chronic Headaches and constant colds or infections.

In this paper we applied machine learning algorithms like naïve Bayes, KNN, logistic regression, decision tree, Support Vector Machine for predicting these mental illnesses and classifying them in five classes in terms of severity.

Some research works most related to ours are discussed in the following section. Section 3 incudes materials and methodology followed by the results in Section 4 that were obtained after using the machine learning algorithms. At last the conclusion is discussed in section 5.

II. LITERATURE REVIEW

Many researchers have proposed different machine learning models for predicting anxiety, stress and depression.

Reece et al. [6] used Hidden Markov Model (HMM) to predict the increase in the probability of depression and Post Traumatic Stress Disorder (PTSD) among Twitter users. About 31.4% of users were affected by depression and and 24% of users were affected by PTSD respectively.

Braithwaite et al. [7] used decision tree to evaluate suicide risk on the tweets of 135 members appointed from Amazon Mechanical Trunk. It was found that the accuracy was 92%.

In a work carried out by Sau et al. (2017), data was gathered manually from the Medical College and Hospital of Kolkata, India. It included 630 old aged persons. Among them 520 were put in special care. The classification algorithms that were applied are Bayesian Network, multiple layer perceptron, logistic, random forest, naïve Bayes, random tree, sequential random optimization, J48, K star and random sub-space. It was concluded that random forest produced the best accuracy rate of 91% and 89% among the two datasets of 10 and 520 people, respectively. They used WEKA tool for feature selection and classification [1].

Anu priya et al. (2020) predicted anxiety, stress and depression levels among people using DASS21 questionnaire. They used

classification algorithms like naïve Bayes, KNN, random forest, decision tree and support Vector machine (SVM). It was concluded that the accuracy rate of KNN was highest, but random forest was considered to be the best model because classes were imbalanced. So the best model was chosen by considering F1 score.

Hou et al. [8] predicted depression of people using a big data approach based on their reading habits. A book classifier was developed based on the features of the Chinese text. The five classification algorithms were applied and naïve Bayes was found to perform best.

III. MATERIALS AND METHODOLOGY

In our work, we used 5 machine learning algorithms to predict the severity levels of anxiety, stress and depression using DASS21 questionnaire. The data was collected from 400 college students. The classification algorithms that were applied are naïve Bayes, KNN, logistic regression, decision tree, random forest.

A. Questionnaires

DASS-21, the Depression, Anxiety and Stress Scale questionnaire, was used to gather the data for the research work.

It includes 21 questions, equally shared between Anxiety, Depression and Stress.

The possible answers for each are as follows:

0 means it did not applied to me

1 means it applied to me to certain degree, or sometimes.

2 means it applied to me to a reasonable degree or a handsome number of times.

3 means it applied to me very much or most frequently.

After the data is collected, numeric values of 0 to 3 were used for encoding participant's response, and calculation of scores was done by summation of the values for each question and the mathematical expression:

$$\text{Score} = 2^* \text{ Addition of grading points of each class} \quad (1)$$

After the calculation of final scores is done, these were classified as per severity of the disease – i.e. Normal, Mild, Moderate, Severe and extremely severe (see Table 2).

TABLE 1. Questions on anxiety, depression and stress.

	Depression	Anxiety	Stress
Normal	0-7	0-9	0-14
Mild	8-9	10-13	15-18
Moderate	10-14	14-20	19-25
Severe	15-19	21-27	26-33
Extremely severe	20+	28+	33+

Table 2. Levels of Severity.

	Depression	Anxiety	Stress
Normal	0-7	0-9	0-14
Mild	8-9	10-13	15-18
Moderate	10-14	14-20	19-25
Severe	15-19	21-27	26-33
Extremely severe	20+	28+	33+

B. Participants

This work was carried out on 400 students aged between 16-29 years, both men and women from diverse backgrounds and cultures.

C. Classification

We used python programming to implement the classification algorithms in Visual studio version --. The algorithms predicted the levels of anxiety, stress and depression. We divided the dataset into the ratio 70:30 for training and testing respectively. Following subsections discusses the working of classification algorithm.

a) K- Nearest Neighbour (K-NN):

K-NN is one of the simple but a very powerful classification algorithms use in machine learning. KNN takes information or data and classifies based on closest measures, latest information points. The data is then assigned to the class with the primary closest neighbour. K-NN classifies the new data points based on the similarity measure of the earlier stored data points.

Fig 1. K- nearest neighbour representation.

b) Decision Tree: It is most powerful and popular classification algorithms used for classification and prediction. It is tree like structure. In Decision Tree, a class label is assigned for each leaf node. The non-terminal nodes, which include the root and other internal nodes, contain attribute test conditions to separate records that have different characteristics for eg. YES or NO, age>20 or age<45. In this example, we have divided the question into yes or no (2 options) into two branches (yes and no).

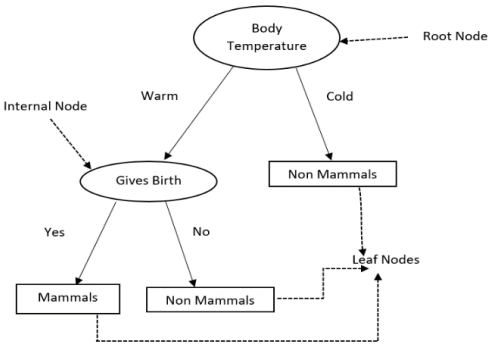

Fig 2. Decision Tree example.

c) Naïve Bayes: A very easy algorithm to implement. The important thing to mention here is that naïve Bayes gives result in probabilities. These probabilities depict how confident the algorithm is. Bayesian classifiers use Bayes theorem. Naïve Bayes is robust isolated noise points and also to irrelevant attributes. Its formula is given as follows:

$$p(H|D)\frac{p(H)p(D|H)}{p(D)} \tag{2}$$

Where,

P (H|D) is the posterior probability of class (target) given predictor (attribute).

P (H) represents the prior probability of class.

P (D) is the prior probability of predictor.

P (D|H) is the likelihood which is the probability of predictor given class.

P (D) represents the prior probability of predictor.

d) Support Vector Machine (SVM):

A Support Vector Machine is a machine learning algorithm that used for both classification and regression work but is primarily used in classification. A simple linear SVM classifier makes a straight line between two classes and hence classifies the data. So the data points on the two sides of the line represent two different categories. Therefore, an infinite number of lines can be choose. It is widely used because of its presentation quality and the ability of classification. In this technique the data is linearly divided into two separate classes (also known as hyperplanes), With the two classes at maximum distant from each other.

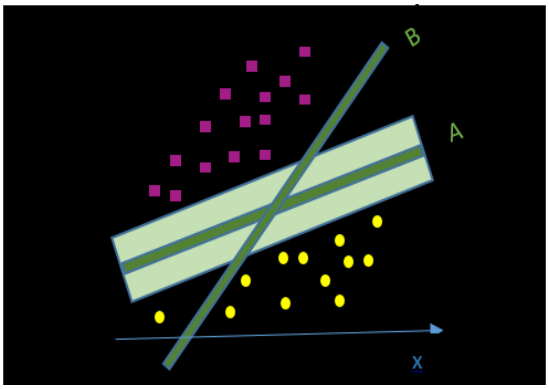

Fig 3. Support vector machine representation.

e) *Logistic Regression:* It is a supervised learning classification algorithm that is used to predict the probability of a target variable. It is based on the concept of probability. It is used when the dependent variable or target value is categorical. It is commonly used in when dependent variable is dichotomous, means there would be only two possible classes. Logistics regression uses the sigmoid function to return the probability of a label.

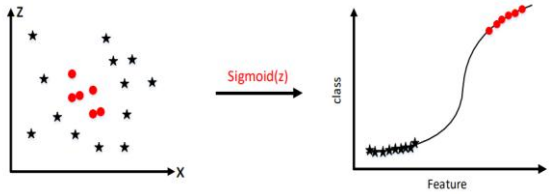

Fig 4. Logistic regression Representation.

IV. Results

The application of all five classification algorithms to all three classes of Anxiety, Depression and Stress, produced confusion matrices shown in Table 3. Actual classes are depicted by the rows of the confusion matrices and the predicted classes are shown by the columns. Normal, mild, moderate, severe and extremely severe cases, respectively are represented by the numbers 1, 2, 3, 4 and 5 in the rows and columns. The calculation of error rates, accuracy, recall, precision and specificity in each confusion matrix was done by using following equation.

$$Accuracy\ Rate = \frac{Sum\ of\ diagonals(TP)}{Total\ numbers\ of\ instances} \tag{3}$$

$$Error\ Rate = 1 - Accuracy\ Rate \tag{4}$$

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{6}$$

$$F1\ Score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{7}$$

$$Recall = Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

Whereas,

TP (True positive), FN (False Negative), FP (False Positive), TN (True Negative)

TABLE 3. Confusion Matrix.

	Anxiety	Depression	Stress
Support Vector Machine	34 0 0 0 0 0 20 0 0 0 0 9 29 0 0 0 0 0 23 0 0 0 0 0 25	20 0 0 0 0 0 35 1 0 0 1 1 25 0 7 0 10 0 14 0 0 0 0 0 26	14 0 0 0 0 0 15 9 0 0 0 0 22 0 7 0 13 0 48 0 0 0 0 0 12
Naive Bayes	34 0 0 0 0 0 20 0 0 0 0 0 38 0 0 0 0 0 23 0 11 0 0 0 14	20 0 0 0 0 0 19 17 0 0 1 1 25 0 7 0 10 0 14 0 0 0 0 0 26	14 0 0 0 0 0 24 0 0 0 0 0 22 0 7 0 0 0 61 0 0 0 0 0 12
Logistic Regression	34 0 0 0 0 0 20 0 0 0 0 9 29 0 0 0 0 0 23 0 11 0 0 0 14	20 0 0 0 0 0 19 17 0 0 1 1 32 0 0 0 10 0 14 0 0 0 0 0 26	14 0 0 0 0 0 24 0 0 0 1 0 28 0 0 0 0 0 61 0 1 0 0 0 11
Decision Tree	34 0 0 0 0 0 20 0 0 0 0 0 38 0 0 0 0 0 23 0 11 0 0 0 14	20 0 0 0 0 0 35 1 0 0 1 1 32 0 0 0 10 0 14 0 0 0 0 0 26	14 0 0 0 0 0 24 0 0 0 0 0 22 0 7 0 0 0 61 0 0 0 0 0 12
K-NN	34 0 0 0 0 0 20 0 0 0 0 0 38 0 0 0 0 0 23 0 0 0 0 0 25	20 0 0 0 0 0 36 0 0 0 1 1 32 0 0 0 0 0 24 0 0 0 0 0 26	14 0 0 0 0 0 24 0 0 0 0 0 28 0 1 0 0 0 61 0 0 0 0 0 12

TABLE 4. Values of different parameters for different Algorithms.

Classifier	Mental illness	Error Rate	Precision	Recall	Specificity	F1 Score	Accuracy
Naïve Bayes	Anxiety	0.079	0.951	0.912	0.915	0.974	0.921
	Depression	0.258	0.793	0.769	0.921	0.765	0.742
	Stress	0.05	0.926	0.951	0.983	0.927	0.95
Support Vector Machine	Anxiety	0.065	0.937	0.952	0.979	0.936	0.935
	Depression	0.143	0.892	0.858	0.954	0.856	0.857
	Stress	0.208	0.775	0.830	0.935	0.793	0.792
K Nearest Neighbour	Anxiety	0.020	0.972	0.967	0.991	0.982	0.935
	Depression	0.015	0.985	0.988	0.995	0.986	0.985
	Stress	0.008	0.984	0.993	0.997	0.988	0.992
Decision Tree	Anxiety	0.089	0.951	0.912	0.974	0.915	0.921
	Depression	0.093	0.936	0.899	0.969	0.904	0.907
	Stress	0.050	0.926	0.951	0.983	0.927	0.950
Logistic Regression	Anxiety	0.143	0.889	0.864	0.954	0.852	0.857
	Depression	0.208	0.847	0.810	0.935	0.811	0.792
	Stress	0.015	0.975	0.976	0.995	0.974	0.985

Table 4. Shows the accuracy, recall, error rate, specificity precision, and F1 score of each class calculated by the different methods. As we can see in Table 4, the KNN achieved the maximum accuracy for all three scales of anxiety, depression and stress. Also, confusion matrices in Table 3 shows that classes were not balanced, because there were 23, 14 and 61 instances of normal but 20, 30 and 24 occurrence of mild, respectively in the confusion matrices of anxiety, depression and stress classes . Also, 38, 42 and 22 occurrence of moderate were there; 14, 33 and 19 occurrence of severe; and 45, 21 and 14 occurrence of extremely severe for the scales of Anxiety, Depression and Stress respectively. So, it was not feasible to measure accuracy alone rather we used F1 score to determine best model. Because in situations where classes are imbalanced the model with higher F1 score is considered to be best even if it has lower accuracy. The F1 score of KNN was the maximum for Stress and anxiety and depression.

V. CONCLUSION

In this paper, we predicted the severity levels of anxiety, stress and depression in college students using machine learning algorithms. DASS21 was used to collect data from 400 students. It is a standard questionnaires used to measure the common signs of anxiety, stress and depression. The severity levels were mild, normal, moderate, severe and extremely severe. The classification algorithms that were applied are Support Vector Machine, KNN, logistic regression, decision tree and naive Bayes. The accuracy of KNN was found to be the highest followed by logistic regression. We also added F1 score as the problem produced imbalanced classes. The KNN was found to be the best model in terms of F1 score. It produced F1 score of 0.988.

The above results show that KNN has performed best out of all the algorithms. So in future it can be implemented online in the form of a website where users would need to answer these 21 questions and their mental health would get predicted.

REFERENCES

[1] https://www.webmd.com/balance/stress-management/stress-symptoms-effects_of-stress-on-the-body#1.

[2] Oei, T. P., Sawang, S., Goh, Y. W., Mukhtar, F. (2013) "Using the depression anxiety stress scale 21 (DASS-21) across cultures." International Journal of Psychology 48 (6): 1018-1029.

[3] Kroenke, K., Spitzer, R. L., Williams, J. B. (2001) "The PHQ - 9: validity of a brief depression severity measure. "Journal of general internalmedicine 16 (9): 606-613.

[4] Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., Langer, E. J.(2016) "Forecasting the Onset and Course of Mental Illness with Twitter Data."Scientific reports 7 (1): 13006.

[5] Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., Hanson, C. L. (2016) "Validating machine learning algorithms for Twitter data against established measures of suicidality."JMIR mental health 3 (2): e21.

[6] Hou, Y., Xu, J., Huang, Y., Ma, X. (2016) "A big data application to predict depression in the university based on the reading habits."3rd IEEE International Conference on Systems and Informatics (ICSAI): 1085-1089.

[7] Laijawala, V., Aachaliya, A., Jatta, H., & Pinjarkar, V. (2020). Mental Health Prediction using Data Mining: A Systematic Review. In Proceedings of the 3rd International Conference on Advances in Science

& Technology (ICAST) 2020. KJ Somaiya Institute of Engineering and Information Technology.

[8] A. J. Xu, M. A. Flannery, Y. Gao, and Y. Wu, “Machine learning for mental health detection,” 2019, https://digitalcommons.wpi.edu/mqp-all/6732/.

[9]] A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, “Predicting mental health problems in adolescence using machine learning techniques,” PLoS One, vol. 15, no. 4, Article ID e0230389, 2020.

[10] P. Morillo, H. Ortega, D. Chauca, J. Proaño, D. Vallejo-Huanga, and M. Cazares, “Psycho web: a machine learning platform for the diagnosis and classification of mental disorders,” in Advances in Neuroergonomics and Cognitive Engineering, pp. 399–410, Springer International Publishing, Berlin, Germany, 2019.

[11] A. M. Chekroud, R. J. Zotti, Z. Shehzad et al., “Cross-trial prediction of treatment outcome in depression: a machine learning approach,” The Lancet Psychiatry, vol. 3, no. 3, pp. 243–250, 2016.

[12] A. Ahmed, R. Sultana, M. T. R. Ullas, M. Begom, M. M. I. Rahi, and M. A. Alam, “A machine learning approach to detect depression and anxiety using supervised learning,” in Proceedings of the 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2019.

[13] A. Sau and I. Bhakta, “Predicting anxiety and depression in elderly patients using machine learning technology,” Healthcare Technology Letters, vol. 4, no. 6, pp. 238–243, 2017.

[14] A. E. Tate, R. C. McCabe, H. Larsson, S. Lundström, P. Lichtenstein, and R. Kuja-Halkola, “Predicting mental health problems in adolescence using machine learning techniques,” PLoS One, vol. 15, no. 4, Article ID e0230389, 2020.

[15] Sau, A., Bhakta, I. (2017)"Predicting anxiety and depression in elderly patients using machine learning technology."Healthcare Technology Letters 4 (6): 238-43.

[16] https://adaa.org/understanding-anxiety/depression/symptoms.