Healthcare Recommendation System For Depression Using Machine Learning Algorithms

[1]Jerusha S, [2]Deepika R, [3]Hemalatha G, [4]Keerthiga D

[1]Assistant Professor, [2,3,4]
Department of Information Technology
Sri Venkateswara College of Engineering
Chennai, India.

Abstract- **Over the years, stress and anxiety causes major effects in people's minds worldwide. New technological advancements are changing the future of the healthcare system. Lifestyle is something which defines an individual the best. Lifestyle including factors like income, age group, marital status, child, alcohol consumption and many more affect the quality of life of an individual. Identification of factors that are responsible for causing depression may lead to new experiments and treatments. Because depression as a disease is becoming a leading community health concern worldwide. NHANES is a program of studies designed to assess the health and nutritional status of adults and children. NHANES conducted a survey for analyzing depression and more than 1200 responses were recorded. These responses are used for training the model and for predicting the depression level. Using machine learning techniques this project presents a complete methodological framework to process and explore the heterogeneous data and to better understand the association between factors related to quality of life and depression. With the identified features, different models were chosen and trained accordingly. Random Forest, Logistic Regression and KNN have been chosen. By evaluating the results of various ML algorithms, Random Forest Classifier outperformed all other algorithms in predicting the levels of depression. The RF based prediction model is more accurate and informative in predicting. The final outcome received was 81.22%. Then, according to the level of depression, a recommendation will be given for future therapy. The recommendation will also consider the parameters responsible for the depression. This will help in assistance to other researchers and clinicians with the recognition of risk related to depression and other psychological disorders.**

Keywords- **depression, random forest, KNN, logistic regression, quality of life.**

I. INTRODUCTION

Healthcare is one of the major problems faced by the entire world regardless of the situation whether the country is developing or developed. As a leading interest worldwide, smart, efficient, and secure healthcare systems are developed to Improve the quality of life. Identifying the mental health issues of a patient is an enduring challenge to doctors and healthcare organizations and especially among younger people, is not a new phenomenon. Recent advances in the field of machine learning and deep learning have shown its power to identify the psychological disorders of individuals as well as recognize the impact of such disorders on their lifestyle. All over the world, the most important growth-related change among people is a change in mental health. That is why depression and anxiety are considered the two most important disorders related to the age factor. Both badly affect the QoL in patients and weaken the decision-making system which in result causes a high level of distress, so finally the patient attempt suicide. "Depression" is considered as one of the most complex and serious psychological problems, having a negative impact and foremost cause of disease burden amongst all diseases. That is why many researchers and medical staff had led their research work towards the investigation of depression. After successful treatment, the impact of depression is still there and continues its struggle in the reduction of performance and generally affecting the QoL of an individual. The term "Quality of Life" highlights different features of an individual's life such as emotional, physical, and psychological well-being. These features explain the experience of living of an individual and are under consideration by different researchers and health experts.

II. RELATED WORKS

Nazmun Nessa, Moon, Asma Mariam and Shayla Sharmin (2021), proposed a prediction model [2] for depression in job sectors. Satisfaction level was one of the main features to detect depression in jobs. For their known accuracy rate Random Forest Classifier, Random Forest Regressor, Naïve Bayes, and K Neighbors Classifier Algorithms are used to determine which sources of stress predict stress-related symptoms in people exploring job satisfaction on the training data set to build a model. This study focuses on predicting depression and showing which sex is most unhappy and happy with their work. To get reliable results, the authors gathered data from both males and women. The data used by

the auto-report questionnaires were used to determine the sources of stress that predict stress symptoms among people who are exploring job satisfaction as expected and work depressed by the following: age, monthly income, gender, occupation, children, city, job before, marital status, satisfaction level of the current job, face harassment to the job, satisfaction level, opinion of a colleague and relative. Random Forest Regressor, Random Forestry Classifying Algorithms were used.

Jelti Chung and Jason Teo have proposed a Mental Health Prediction [1] model that involves psychological testing, and machine learning algorithms such as support vector machine, linear discriminant analysis, and K-nearest neighbor have been used to classify the intensity level of the anxiety and depression, which consists of two data sets. This model categorizes the collected research articles based on the mental health problems such as schizophrenia, bipolar disorder, anxiety and depression, post traumatic stress disorder, and mental health problems among children. flexible algorithms will become the main challenge toward mental health because of heterogeneity in the input data. This paper presents a recent systematic review of machine learning approaches in predicting mental health problems. Furthermore, we will discuss the challenges, limitations, and future directions for the application of machine learning in the mental health field. We collect research articles and studies that are related to the machine learning approaches in predicting mental health problems by searching reliable databases. Moreover, we adhere to the PRISMA methodology in conducting this systematic review. We include a total of 30 research articles in this review after the screening and identification processes. e categorize the collected research articles based on the mental health problems such as schizophrenia, bipolar disorder, anxiety and depression, post traumatic stress disorder, and mental health problems among children. Discussing the findings, we reflect on the challenges and limitations faced by the researchers on machine learning in mental health problems. Additionally, we provide concrete recommendations on the potential future research and development of applying machine learning in the mental health field.

W. J. Zhang, C. Yan, D. Shum, and C. P. Deng, have proposed a paper titled [4] “Responses to academic stress mediate the association between sleep difficulties and depressive/anxiety symptoms in Chinese adolescents”. This research aimed to explore whether stress responses mediate the concurrent and future connection between sleep difficulties and symptoms of depression/anxiety in Chinese adolescents. A total of 17,946 adolescents aged 14 to 18 completed assessments, including the Pittsburgh Sleep Quality Index, the Center for Epidemiologic Studies Depression Scale, Revised Children's Manifest Anxiety Scale, and the Responses to Stress Questionnaire. Additionally, 710 participants completed one-year follow-up assessments. Structural equation models were employed to examine the concurrent and prospective mediation effects of stress responses, considering the moderating influence of gender and age. Involuntary engagement and disengagement responses, along with engagement coping, were found to significantly mediate the cross-sectional association between sleep difficulties and symptoms of depression/anxiety. Furthermore, baseline sleep difficulties predicted increased involuntary engagement responses but decreased use of engagement coping strategies one year later, resulting in heightened levels of depressive/anxiety symptoms. Lastly, females and younger adolescents experiencing greater sleep difficulties were more likely to exhibit maladaptive stress responses. These findings underscore the crucial role of stress responses in the link between sleep difficulties and symptoms of depression/anxiety.

III. PROPOSED SCHEME

In Medical Science, identifying the mental health issues of a patient is an enduring challenge to doctors and healthcare organizations, especially among younger people.That is why depression and anxiety are considered the two most important disorders related to the age factor.Both badly affect the quality of life (QoL) in patients and weaken the decision- making system. The main goal of this project is to Develop an appropriate and smart model frame work for

the identification of QoL factors that are responsible for causing depression and provide appropriate remedies based on the results. The architecture of depression prediction is shown in Figure 1.

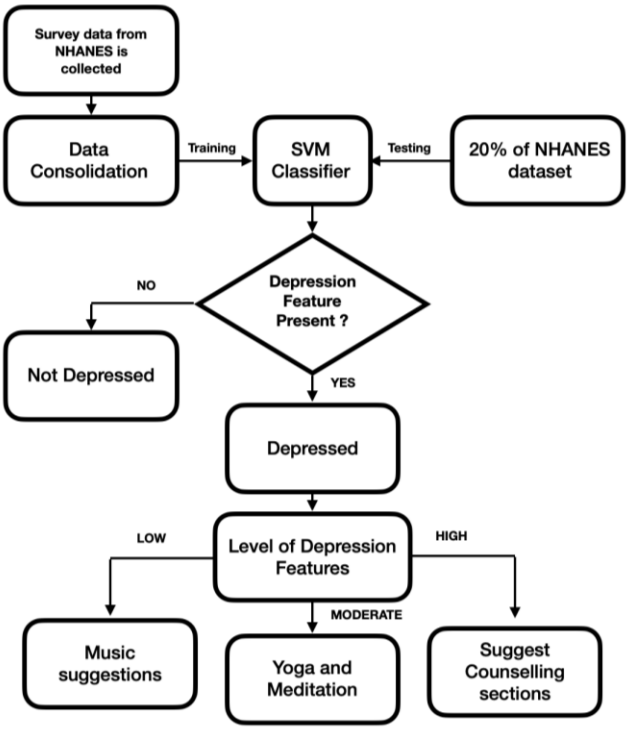

Figure 1 Depression prediction architecture

A. Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous. Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio level independent variables. The logistic regression model uses the logistic function to squeeze the output of a linear equation between 0 and 1. It is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. It is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. It can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

B. *KNN Algorithm*

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using KNN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. The KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data. it does no training at all when you supply the training data. At training time, all it is doing is storing the complete data set but it does not do any calculations at this point. The size of the neighbourhood needs to be set by the analyst or can be chosen using cross-validation (we will see this later) to select the size that minimises the mean-squared error.While the method is quite appealing, it quickly becomes impractical when the dimension increases, i.e., when there are many independent variables.

C. *Random Forest Algorithm*

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. The forest it builds is an ensemble decision of trees usually trained with the bagging method. It can be used for both Classification

and Regression problems. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. The random forest technique has a capability to focus both on observations and variables of a training data.

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It works with data having discrete labels known as a class. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Also, the random forest algorithm makes it easier to measure the relative importance of each feature on the prediction.

Assumptions for random forest algorithm are There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result and the predictions from each tree must have very low correlations.

IV. Experiments and Results

A.Dataset and Implementation

The Dataset is acquired from NHANES(National Health and Nutrition Examination Survey) dataset which contains more than 10,000 records of data about depression. These images are made available to public in-order to improve depression diagnosis. The Missing data from the dataset are analysed using the isnull() function. The columns with highest missing data are then checked if necessary since it reduces the accuracy. They are dropped using the drop function. Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important preprocessing step for the structured dataset in supervised learning.The features such as age and other numerical values are replaced with the mean values. The gender values are trained with standard words.

B.Training the model

The data correlation is studied using the correlation matrix. Then it is presented with the heatmap representation. The dataset is visualized with the help of density graph. The Age values are then transformed to fit using the minmaxscaler().

The dataset is splitted into X_train,X_test, y_train, y_test with the help of train_test_split imported from the sklearn model selection function with the test_size = 0.30 and the random_state value 0. The relationship between age vs family history mental depression is represented using graph in the below figure 2.

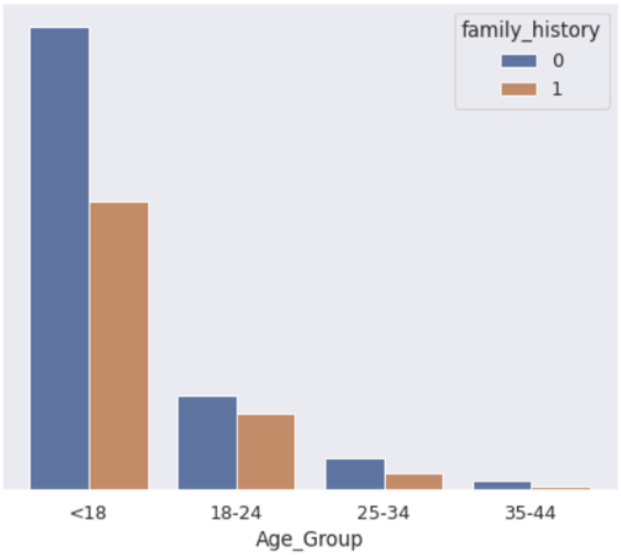

Figure 2 Relationship between Age vs Family

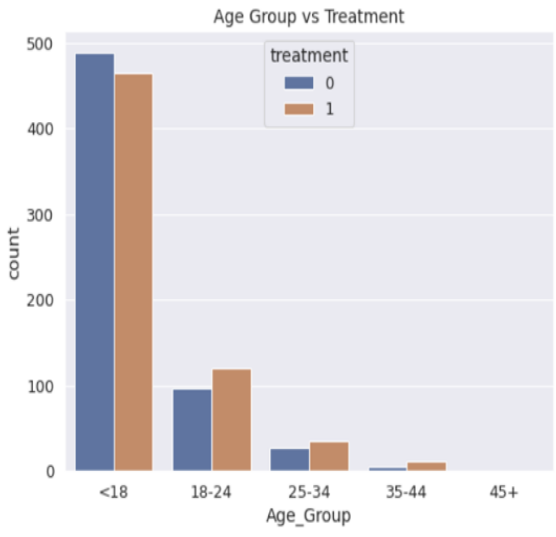

Figure 3 Age-wise Depression range

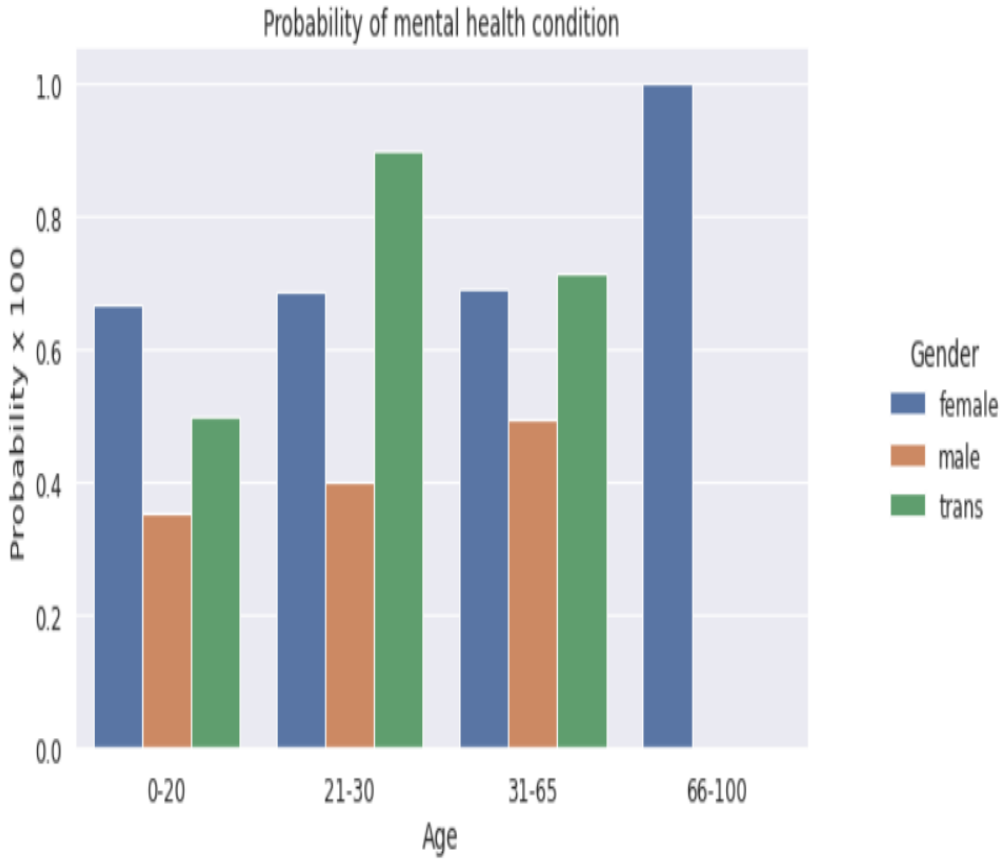

Figure 4 Probability of Mental health condition

The distribution of depression range in the categorization of age group is represented in the below figure 3. The probability of depression vs gender classification is considered in the figure 4. The ExtraTreesClassifier with estimators 250 and random_state = 0 are used to find the feature importances. The features are Age, Gender, Family_history, Benefits, Care_options, Anonymity, Leave and Work_Interference.

C.Feature selection

For the machine learning methods to be able to learn, instead of using all the features in training the model, the important features have to be extracted from dataset. Only those features will determine the result. In our model, features like Age, Gender, family_history, benefits, care_options, anonymity, leave, work_interfere were found to be the most important features in predicting depression by RandomizedSearchCV. These features are selected based on the relationship between all other features by building a covariance matrix. Among numerous columns in the dataset, these

features affect the end result considerably. The covariance matrix based on the importance of features are represented in figure 5. Also, the important features are selected and are represented in the below figure 6.

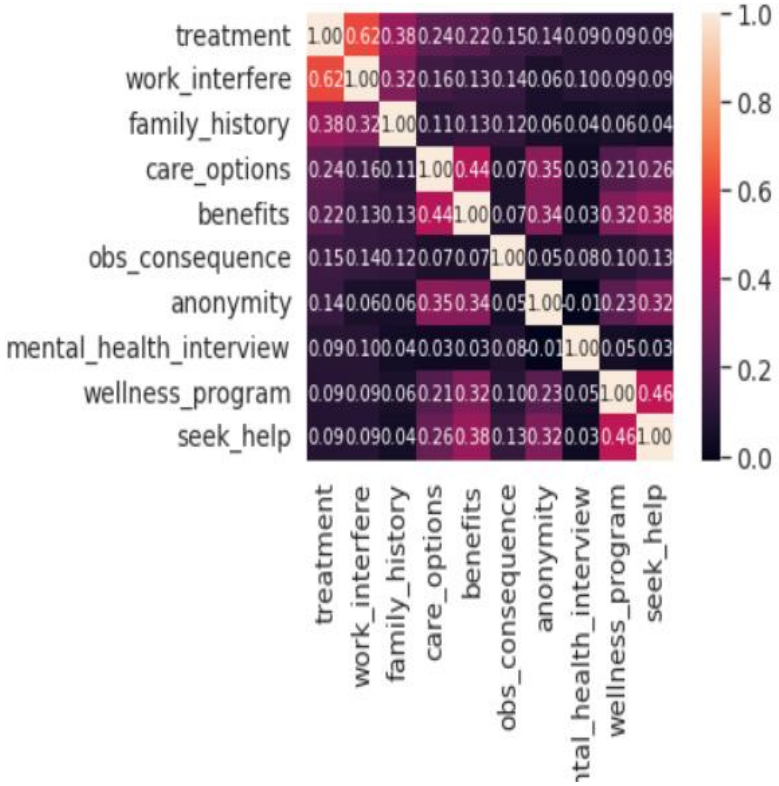

Figure 5 Covariance matrix

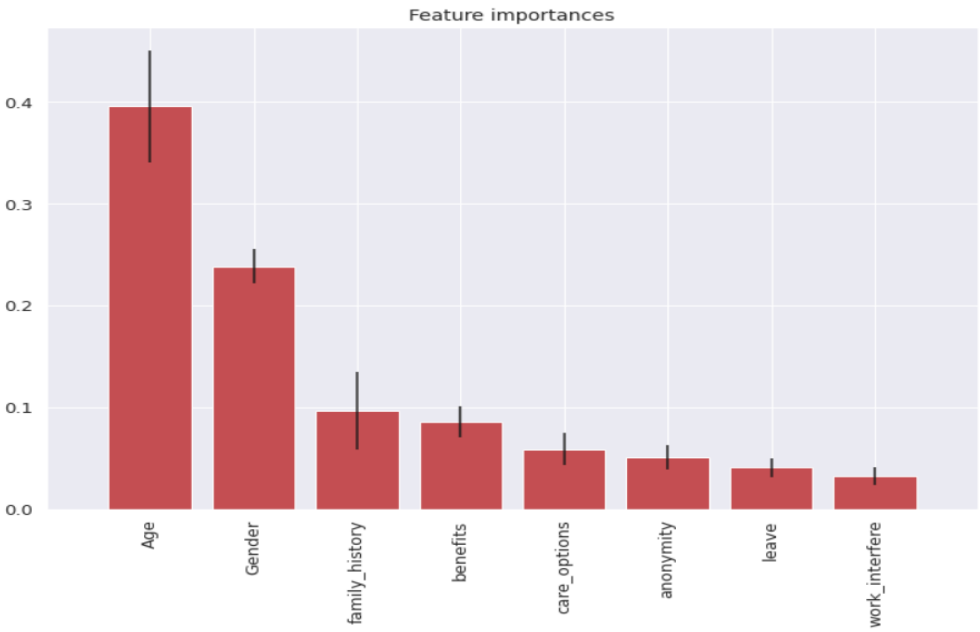

Figure 6 Feature importances

The logistic regression model is setup using LogisticRegression() from sklearn.linear_model and its accuracy is predicted. The KNN model is defined with k_range 1 to 31, the weight option is assigned to uniform. The Random Forest model is built with criterion gini and entropy, min_samples_split and leaf to be a random number between 1 to 9 and maximum depth 3.

D. Tuning the model

While evaluating a Classification model, the Null accuracy, percentage of ones, percentage of zeros are considered. The crossover validation score tuning is used to find the accuracy using the KNN model. The GridSearchCV and RandomizedSearchCV are used to find the accuracy of Random Forest Algorithm.

E. Choosing the Best Model

The Logistic Regression model is trained and got the accuracy of 79.62%. Then, The KNN model is trained with accuracy of 75.99%. Finally, RandomForest model achieved an accuracy of 81.88% which is greater than the above two models. The confusion matrix for the random forest model is shown in the below figure 7.

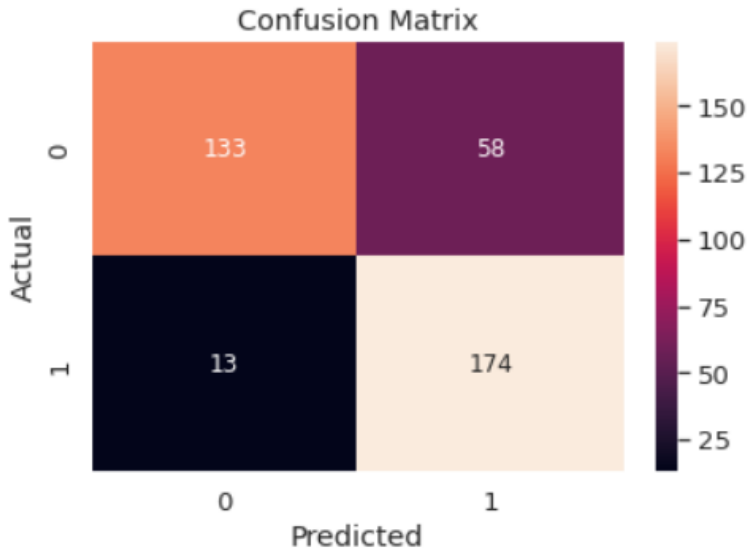

Figure 7 Confusion matrix

The random forest model have a classification error of 18.78%, AUC Score of 81.34 and Cross-validated AUC of 89.34%. The histogram of predicted probabilities vs frequency is shown in the below figure 8.

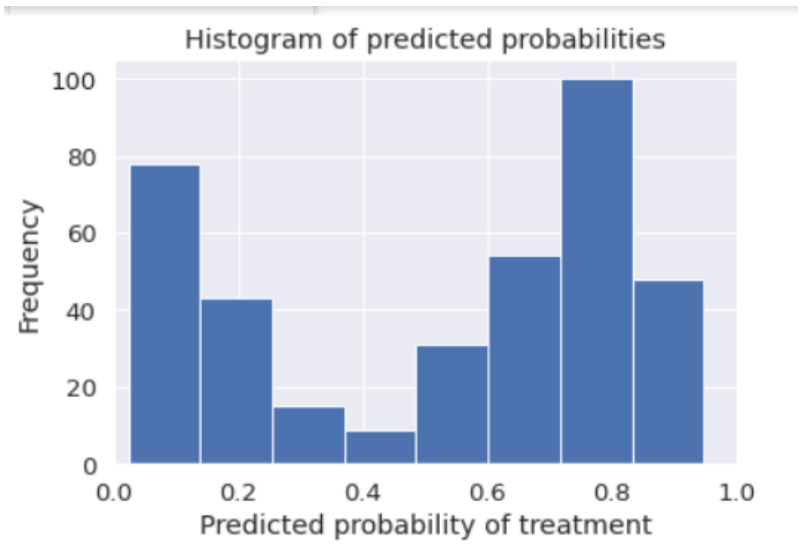

Figure 8 Predicted Probabilities

Based on the probability of the output the stage of the depression is defined as No depression, Initial, Intermediate and Extreme level depression. Also, the remedies for the particular stage is also prescribed. The remedies include movie recommendation system for initial level depression, yoga practices with asana poses for intermediate level depression and psychologist details with the mental health page for extreme level depression is given.

V. CONCLUSION

Mental Health, being one of the major problems for the upcoming generation, this depression prediction system will diagnose the depression and predict the severity of the depression which will be helpful for further treatments.

REFERENCES :

1. Jetli Chung, Jason Teo, "Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges", *Applied Computational Intelligence and Soft Computing*, vol. 2022, Article ID 9970363, 19 pages, 2022. https://doi.org/10.1155/2022/9970363

2. Moon NN, Mariam A, Sharmin S, Islam MM, Nur FN, Debnath N. Machine learning approach to predict the depression in job sectors in Bangladesh. Curr Res Behav Sci 2021;2:1–10. https://doi.org/10.1016/j.crbeha.2021.100058
3. Milicetal.,"Tobacco smoking and health-related quality of life among university students: Mediating effect of depression," PLoS One, vol. 15, no. 1, pp. 1–18, 2020, doi: 10.1371/journal.pone.0227042.
4. W. J. Zhang, C. Yan, D. Shum, and C. P. Deng, "Responses to academic stress mediate the association between sleep difficulties and depressive/anxiety symptoms in Chinese adolescents," J. Affect. Disorders, vol. 263, pp. 89–98, Nov. 2019, doi: 10.1016/j.jad.2019.11.157.
5. Chung, L. Salvador-Carulla, J. A. Salinas-Pérez, J. J. Uriarte-Uriarte, A. Iruin-Sanz, and C. R. García-Alonso, "Use of the self-organizing map network (SOMNet) as a decision support system for regional mental health planning," Health Res. Policy Syst., vol. 16, no. 1, pp. 1–17, 2018, doi: 10.1186/s12961-018-0308-y.
6. K. P. Linthicum, K. M. Schafer, and J. D. Ribeiro, "Machine learning in suicide science: Applications and ethics," Behavioral Sci. Law, vol. 37, no. 3, pp. 214–222, Sep. 2018, doi: 10.1002/bsl.2392.
7. Dibeklioglu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," IEEE J. Biomed. Health Informat., vol. 22, no. 2, pp. 525–536, Mar. 2018, doi: 10.1109/JBHI.2017.2676878.
8. Ledesma, M. A. Ibarra-Manzano, E. Cabal-Yepez, D. L. AlmanzaOjeda,andJ.G.Avina-Cervantes,"Analysis Of Datasets With Learning Conflicts for machine learning," IEEE Access, vol. 6, no. 1, pp. 45062–45070, 2018, doi: 10.1109/ACCESS.2018.2865135.
9. M. Masseroli, A. Canakoglu, and S. Ceri, "Integration and querying of genomic and proteomic semantic annotations for biomedical knowledge extraction," IEEE/ACM Trans. Comput. Biol. Bioinf., vol. 13, no. 2, pp. 209–219, Mar./Apr. 2016, doi: 10.1109/TCBB.2015.2453944.
10. H.-G.Kreßel, "Pairwise classification and support vector machines," in Advances in Kernel Methods: Support Vector Learning, Cambridge, MA, USA: MIT Press, 1999, pp. 255–268.
11. Y.S.Erin, L.Allwein, and R.E.Schapire, "Reducing multiclass to binary: A Unifying approach for margin classifiers," J. Mach. Learn. Res.,vol.1, pp. 113–141, 2000, doi: 10.1002/9783527677320.ch16.
12. H.Qian,Y.Mao,W.Xiang,and Z.Wang,"Recognition of human activities using SVM multi-class classifier," Pattern Recognit. Lett., vol. 31, no. 2, pp. 100–111, 2010, doi: 10.1016/j.patrec.2009.09.019.
13. Amin, R.; Al Ghamdi, M.A.; Almotiri, S.H.; Alruily, M. Healthcare Techniques Through Deep Learning: Issues, Challenges and Opportunities. IEEE Access 2021, 9, 98523–98541.
14. He, L.; Niu, M.; Tiwari, P.; Marttinen, P.; Su, R.; Jiang, J.; Guo, C.; Wang, H.; Ding, S.; Wang, Z.; et al. Deep learning for depression recognition with audiovisual cues: A review. Inf. Fusion 2021, 80, 56–86.
15. Patel, H.H.; Prajapati, P. Study and Analysis of Decision Tree Based Classification Algorithms. Int. J. Comput. Sci. Eng. 2018, 6, 74–78.
16. Li, X.; Zhang, X.; Zhu, J.; Mao, W.; Sun, S.; Wang, Z.; Xia, C.; Hu, B. Depression recognition using machine learning methods with different feature generation strategies. Artif. Intell. Med. 2019, 99, 101696.
17. Tao, X.; Chi, O.; Delaney, P.J.; Li, L.; Huang, J. Detecting depression using an ensemble classifier based on Quality of Life scales. Brain Inform. 2021, 8, 2.