A Framework for Performing an Analysis on Behavioral traits using Machine Learning

Christy Jacqueline[1], K. Ranjith Singh [2]

{christy.jacqueline@gmail.com[1]}

Research Scholar, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore[1], Assistant Professor & Research Guide, Department of Computer Science, Karpagam Academy of Higher Education, Coimbatore [2]

Abstract. Personality is a combination of an individual's behavior, emotion, motivation and characteristics of their thought pattern. The main aim of this work is to find a better solution for identifying behavioral characteristics systematically using methods such as KMeans and Agglomerative Hierarchical clustering. In the first phase, clustering is used to identify the different personality traits. In the second phase different machine learning algorithms such as Naive Bayes, Logistic Regression, K Nearest Neighbor, Decision Tree and Random Forest are used. A real-time dataset is used for building the prediction models. To evaluate the effectiveness of the proposed framework, a step-by step model evaluation is done. The metrics such as accuracy, precision, recall, ROC AUC Score are used to evaluate the performance. The accuracy rate for the Random Forest and Decision tree was higher. Random Forest has slightly better .95 accuracy rate when compared to Decision Tree with 93.

Keywords: Mental Health, Behavioral traits, K-Means, Agglomerative Hierarchical clustering, Classifiers.

1 Introduction

Behavioral problems are caused by poor mental health. Serious issues can be prevented with early discovery and treatment. An individual's motivation and capacity to deal with situational demands may be out of balance in the most extreme situations, which can lead to psychological imbalance. Mental or emotional distress results from difficult or challenging circumstances. Later, this can result in stress, anxiety, or depression. As a result, a detailed investigation of behavioral disorders, their causes, effects, and treatments both for therapy and prevention is required.

We attempt to identify people who exhibit similar behavioral characteristics together and categorize them into various groups. It is also crucial to create groupings with meaning within the data. It recognizes and organizes the information that is largely homogeneous within itself and largely heterogeneous between itself. Therefore, the purpose of this study is to use

IACIDS 2023, November 23-25, Lavasa, India
Copyright © 2024 EAI
DOI 10.4108/eai.23-11-2023.2343255

unsupervised and supervised ML techniques to predict behavioral issues on a targeted population [1].

To the best of our knowledge, the clustering analysis of research on behavioral features in the field of mental health is not widely recognized. The main objective is to provide a thorough breakdown and categorization of the behavioral characteristics within the context of mental health.

The paper has the following main contributions:

1. An in-house dataset is created for analyzing the behavioral features.

2. Performing cluster analysis to group the data into five behavioral traits

3. To find the performance of each classifier in the in-house dataset.

4. Generates inferences from the results obtained.

The rest of the paper is organized in the following manner: Section II covers the background study; Section III analyzes prominent Unsupervised ML algorithms. Section IV presents the observation and results. Section V includes the discussions and the paper concludes with Section VI.

2 Background Study

2.1 Related Work

The main purpose of this research work is to identify the individuals who are having behavioral issues in the targeted population. These individuals need special attention in order to rectify behavioral issues at the right time. Hence, we have used a benchmarked questionnaire and rate them based on the responses. There are five different behavioral characteristics such as extraversion, neuroticism, agreeableness, conscientiousness, and openness.

A technique that incorporated network estimate and cluster identification was presented by Kashihara et al., [14]. Four transdiagnostic clusters were found, and these clusters were used to construct clinical hypotheses.

Elgendi et al. [13] claim that studying bio signals can be used to predict driving stress. Three unsupervised ML techniques are applied in the longitudinal analysis: interaction principal component analysis, connectivity-based clustering, and K-Means clustering.

Zhengai Yang et al. [15] present a method employing machine learning to screen for depression at large scale for targeted populations using certain norms. Using K-Means Clustering, four levels of depressive symptoms are determined.

The author Riya Paul et al. suggested a clustering strategy to identify major depressive disorder using a mixed model [11]. A model-based clustering approach is used to identify the major depressive disorder treatment response class.

By developing a framework for comprehending mental health, Mohanavalli Subramaniam et al. [12] work makes it possible for numerous target groups to intuitively comprehend the mental health of individuals. K-Means, an Agglomerative approach to hierarchical clustering, and K-Medoids were the clustering techniques used.

2.2 Clustering

Clustering is the process of grouping like things into sensible groups so that they are more similar to one another than they are to those in other groups. Some of the prominent clustering techniques include partitional, hierarchical, density-based, grid-based, and model-based ones.

K Means Clustering

Non-hierarchical clustering is the kind used in this situation. K-means clustering is one of the most prevalent examples of this kind of clustering. In this kind of clustering, a starting set of cluster means is established, and each case is then given the closest cluster mean [1]. Iterative algorithms like the K-Means algorithm are frequently employed.

Hierarchical Agglomerative Clustering

This study emphasizes agglomerative hierarchical clustering as well. This bottom-up clustering technique is effective in locating small groups [3]. There are two different types of hierarchical cluster analysis and they are agglomerative or divisive[7].

Classifiers

For supervised machine learning problems, classifiers are machine learning algorithms [2]. Based on a set of attributes, this approach classifies data as belonging to a specific class or group. Some of the prominent classifiers used are naïve bayes, logistic regression, decision tree, KNN, Random Forest [5].

Classifier Performance Measures

The foundation for computing the performance measures is the confusion matrix[6]. A confusion matrix can be used to express the evaluation of the most effective response during classification training. The anticipated class is presented in the table's row, while the actual class is shown in the column. For determining accuracy, precision, recall, and F1-score can be used [4].

3 The Proposed Model

3.1 Proposed Framework

The proposed framework is summarized in Fig. 1. The framework illustrates the steps of the presented approach.

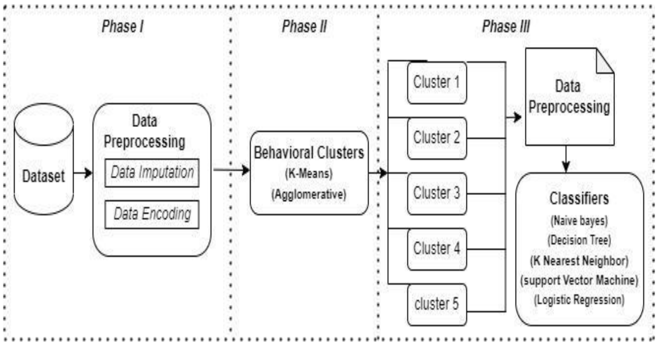


Fig 1. Overview of the Approach

The entire work is carried on in three phases. In Phase I, a new dataset is used for performing the analysis. In the second phase, the preprocessed data is grouped based on the similarity measure. The pre-processing steps are shown in figure 2. Five clusters are identified in order to determine the class labels to build prediction models in the third phase[7].

The main objective is to identify the individuals with different behavioral characteristics. A benchmarked 20-item questionnaire is used to identify the behavioral characteristics. The consent from the participants were also collected. The questions were prepared based on the guidelines given by the domain expert. Each item on the form has five options. The weights given to the responses range from 1 to 5. There are scores for each option. The behavioral features are computed and predicted based on the participant scores. A total of 725 samples were gathered from various age groups.

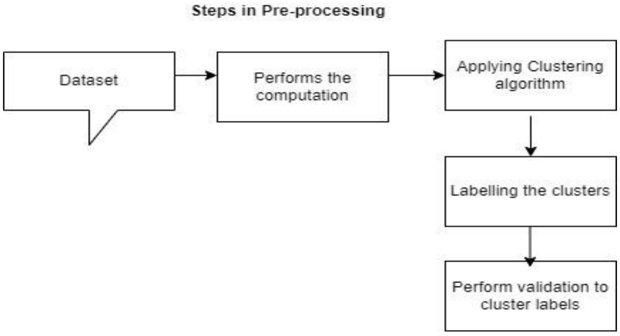


Fig 2. Steps in Preprocessing Stage

3.2 Phase II & Phase III

Phase II includes segmentation of data. Behavioral trait features and demographic features are included in the dataset. The dataset is clustered using k-means and agglomerative hierarchical clustering methods. Phase III, includes the classification process. First, based on behavioral and demographic characteristics, behavioral disorders are predicted. 725 samples make up our dataset, which also has 20 attributes and 5 class labels. The ratio of training to testing was set at 70:30, respectively. After splitting, there are 218 samples in the testing set and 507 samples in the training set. The classifier model was built employing classifiers as the subsequent phase.

Naive Bayes, Logistic Regression, KNN, Decision Tree, and Random Forest are the classifiers that are employed[10].

The models were built after the training set had been fed into several classifiers. The test set data is used to assess each classifier's performance. Split-validation is employed to validate the prediction model, and evaluation metrics including accuracy, precision, recall, and F1-score are computed for each model. These measures are used to assess the effectiveness of prediction models[8][9].

4 Results and Discussion

Dataset Description

A questionnaire was developed to ascertain a person's behavioral characteristics. A 20-item survey was utilized. Data related to personality traits, demographics, education, and employment were stored in the database. An initial data preparation process has been completed. After processing, the categorical variables are encoded using Sci-Kit Learn's OneHotEncoder. After scaling the input features to lie between 0 and 1, normalizing them to lie between 0 and 1, and leaving any missing features as NaN, the features were finally transformed. The dataset's missing values have been calculated. and eliminated the rows with empty fields.

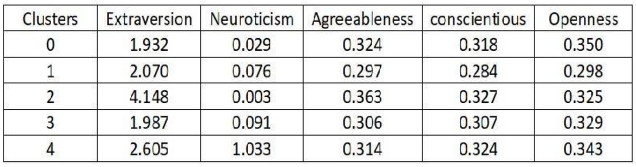
In this study, clustering is primarily used to find potential behavioral traits in the population that is being studied. The dataset is subjected to clustering techniques like K-Means and Agglomerative Hierarchical Clustering. The grouping phase concentrated on features for which information might be gathered on identifying various behavioral traits.

Analyzing the effectiveness of cluster analysis and validation in separating dissimilar samples from comparable ones is the main objective. The dataset is subjected to two different kinds of clustering analysis. Agglomerative Hierarchical Clustering with K-Means.

KMeans Clustering

The dataset is subjected to KMeans clustering in order to identify the person who shares comparable behavioral traits. The mean scores for each behavioral trait, including extraversion, neuroticism, agreeableness, conscientiousness, and openness, are shown in Table 1 for each cluster.

Table I. Means Scores For Individual Cluster



Clusters	Extraversion	Neuroticism	Agreeableness	conscientious	Openness
0	1.932	0.029	0.324	0.318	0.350
1	2.070	0.076	0.297	0.284	0.298
2	4.148	0.003	0.363	0.327	0.325
3	1.987	0.091	0.306	0.307	0.329
4	2.605	1.033	0.314	0.324	0.343

It presents the combined scores for several behavioral qualities before doing a cluster analysis to group the data points into clusters. As shown in Table 1, we have also determined the average scores for each behavioral feature within each cluster.

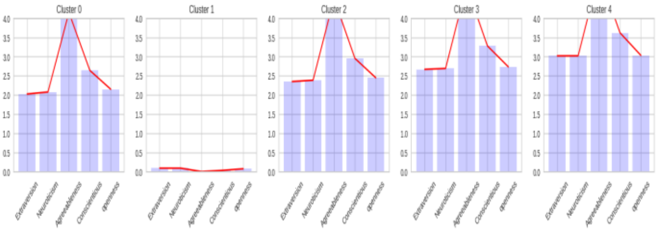


Fig 3. Visualizing the Mean score of each cluster

In the figure 3, the mean scores are represented by bars in each of the five subplots, and a red line connects the mean values. The relevant cluster number is written on each subplot's label. We utilized Principal Component Analysis (PCA) to visualize the data in a 2D graph. It helps to project the data points onto a 2D plane by reducing the dimensions to two. Fig. 4 illustrates how PCA was used to convert the original data into a 2D representation.

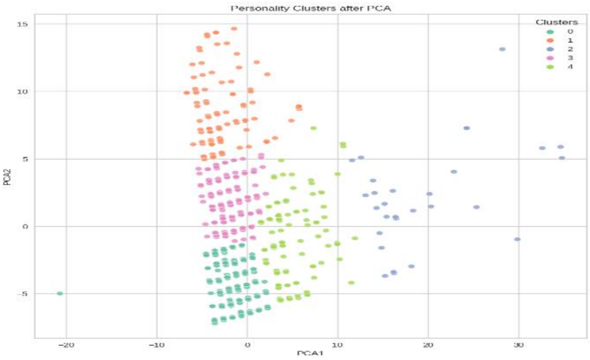


Fig 4. Personality clusters after PCA

In figure 4, each data point is represented by its coordinates on a PCA plane in a 2D scatter plot created from the original data. Based on their clustering, the data points are dispersed and separated in the condensed 2D space in this graph. Based on where each data point belongs in the cluster, it is coloured and given a label.

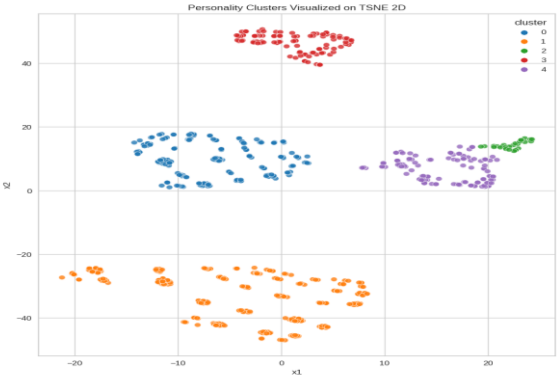


Fig 5. Clusters visualized on t-SNE 2D

The scatter plot aids in the visualization of the behavioral clusters found using the t-SNE (t-distributed stochastic neighbor embedding) method in a 2D space. In Fig. 5, the scatter plot's x and y axes are set to the two dimensions ('x1' and 'x2') obtained from the t-SNE. The data points are coloured according to their cluster assignment and are represented by their coordinates in the t-SNE space.

Agglomerative Hierarchical clustering

We have also used the agglomerative clustering strategy, which successively creates behavioral groupings by methodically combining clusters that are similar. The three most widely used linkage metrics are single linkage, complete linkage, and average linkage. The hierarchical link between the data points is shown in a dendrogram that is displayed using the hierarchical clustering method. It demonstrates how, during the hierarchical clustering process, the data points are combined into clusters.

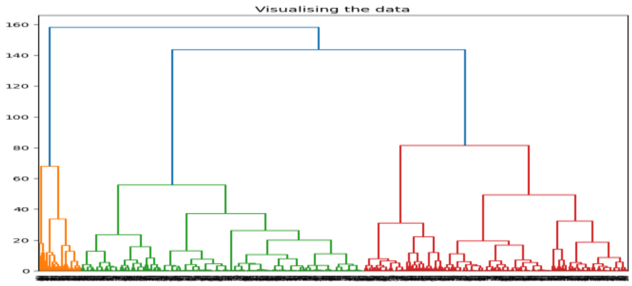


Fig 6. Visualizing the data using the dendrogram

The horizontal lines symbolize the clustering of data points, and the vertical lines the distance at which the data points are merged, as shown in figure 6. The ideal number of clusters was established by examining the heights of the horizontal lines.

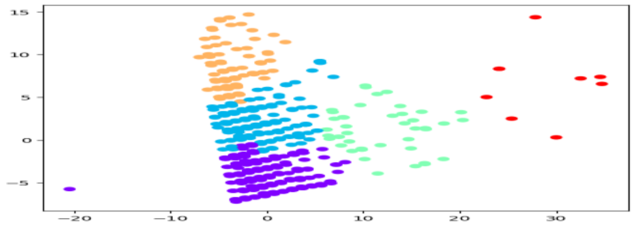


Fig 7. Agglomerative hierarchical clustering with PCA

The vertical lines represent the distance at which the data points are combined, while the horizontal lines represent the clustering of the data points. The heights of the horizontal lines were examined to determine the optimum number of clusters.

There is more than one value for each variable in the clusters. This implies that when one or both of the clusters has more than one case, we must determine the optimal method for calculating an exact distance measure between the two clusters for each variable. A linkage measure was used for this objective. Based on the minimum or biggest distance that may be discovered between pairs of instances, linkage measures determine the distance between two clusters. Single, total, and average are the three types of linkage measurements.

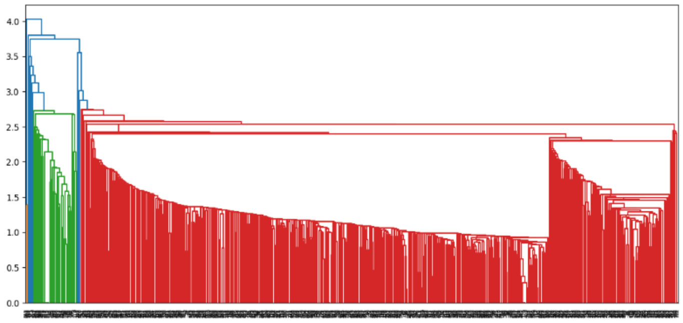


Fig 8. Dendrogram of hierarchical cluster -Single linkage

The distance between two clusters is defined by single linkage as the smallest distance discovered between one case from each cluster. This approach has the drawback of occasionally causing chaining between the clusters. The cluster solution may suffer from this chaining effect. The analysis utilizing a single linkage is displayed in Fig. 8. The dendrogram makes it very evident how connection can lead to chaining.

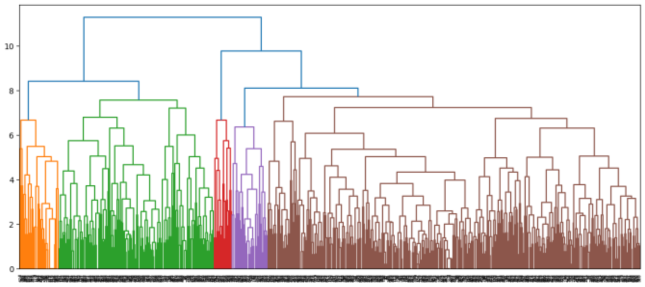


Fig 9. Dendrogram of hierarchical cluster Complete linkage

The dendrogram shown in figure 9 is the analysis using complete linkage. Five clusters were derived from this analysis.

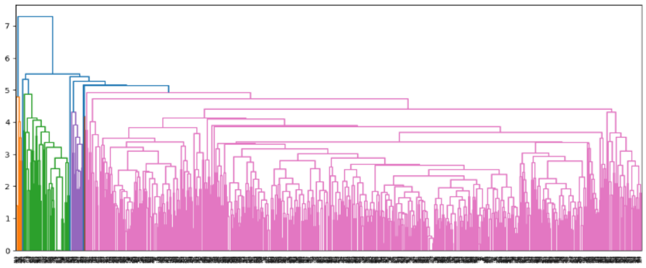


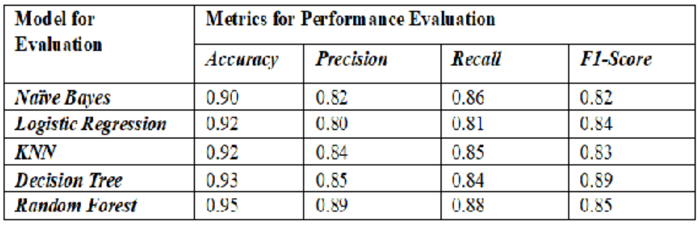
Fig 10. Dendrogram of hierarchical cluster Average linkage

Figure 10 depicts the dendrogram of hierarchical cluster average linkage. For the dataset in use, average linkage was the best choice. It is crucial to note that each dataset will require different actions to be taken and results to be obtained.

Classifiers

The supervised machine learning algorithms and the performance evaluation metrics for classifiers is shown in Table II.

Table II. Performance Evaluation And Detailed Results Analysis



Model for Evaluation	Metrics for Performance Evaluation			
	Accuracy	*Precision*	*Recall*	*F1-Score*
Naïve Bayes	0.90	0.82	0.86	0.82
Logistic Regression	0.92	0.80	0.81	0.84
KNN	0.92	0.84	0.85	0.83
Decision Tree	0.93	0.85	0.84	0.89
Random Forest	0.95	0.89	0.88	0.85

Table II. displays the recall, accuracy, and precision. Naïve Bayes correctly classifies most of the instances in the dataset and also indicates that both positive instances and positive predictions are accurate. But when Naïve Bayes is compared to other models the F1-score is comparatively less. Logistic regression appears to perform slightly better in terms of accuracy and F1- score. KNN has the highest accuracy, best precision. It is evident that when compared to other models, the accuracy rate for the Random Forest and Decision tree was higher. Random Forest has shown slightly better accuracy and precision and lower F1-score when compared to Decision Tree.

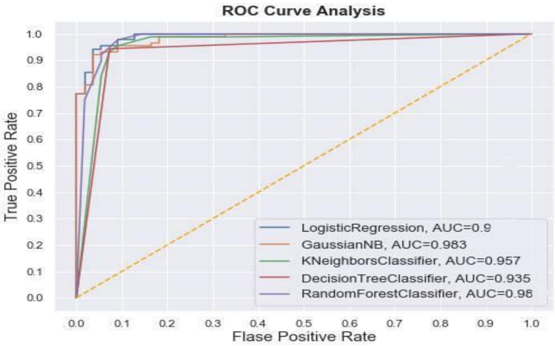


Fig 11. ROC-AUC curve of the classifiers

Figure 11 shows the overall ROC-AUC curve study that we conducted. The performance of various classifiers used on the novel dataset may be visually compared. By separating positive and negative occurrences on the roc-auc curve, we were able to locate the model. As can be seen from the analysis, Decision Trees and Random Forest give the best performance as individual classifiers and the other models provided fair performance.

5 Conclusion

Machine Learning has excellent potential to determine behavioral traits that can be used for identifying mental health issues at the right time. This framework was used to build prediction models. The clustering algorithms are used to identify the optimal number of clusters. The experiments have demonstrated that Naive Bayes, Logistic Regression, KNN, have provided a fair performance. Also, the classifiers such as Decision Tree and Random Forest were found to provide significant improvement in performance. This framework can be used as a secondary tool to assess the behavioral issues of an individual. It can also be used by a large community which will result in more data samples. The accuracy obtained using classifiers can be improved using ensemble methods. In this we have used machine learning techniques to know the robustness of the model. In future, we will adopting an enhanced super learner that will outperform the existing approach of classification.

References

[1] Tawhid, M.A., Ibrahim, A.M. An efficient hybrid swarm intelligence optimization algorithm for solving nonlinear systems and clustering problems. Soft Computing 27, 8867–8895 (2023). https://doi.org/10.1007/s00500-022-07780-8

[2] Li, M., Zhu, Y., Shen, Y. et al. Clustering-enhanced stock price prediction using deep learning. World Wide Web 26, 207–232 (2023). https://doi.org/10.1007/s11280-021-01003-0

[3] Suren A. Tatulian, Challenges and hopes for Alzheimer's disease, Drug Discovery Today,Volume 27, Issue 4,2022, Pages 1027-1043, ISSN 1359-6446.

[4] Tuichievna, M. O., Elmurodova, L. K., & Rasulovna, K. B. (2023). The Main Age-Related Diseases and Conditions Common among Elderly Men and Women. Scholastic: Journal of Natural and Medical Education, 2(3), 37–43. Retrieved http://univerpubl.com/index.php/scholastic/article/view/664

[5] Mahon, L. &amp; Lukasiewicz, T.. (2023). Efficient Deep Clustering of Human Activities and How to Improve Evaluation. <i>Proceedings of The 14th Asian Conference on Machine Learning, Proceedings of Machine Learning Research 189:722-737

[6] Kim, K., Kim, K., Lee, Y., & Cho, Y. (2019). Early Prediction of Learning Disabilities Using Machine Learning Techniques. International Journal of Medical Informatics, 128, 47-53. doi: 10.1016/j.ijmedinf.2019.04.006.

[7] https://doi.org/10.1016/j.drudis.2022.01.016.Sarica, A., Quattrone, A., Quattrone, A. (2021). Explainable Boosting Machine for Predicting Alzheimer's Disease from MRI Hippocampal Subfields. In: Mahmud, M., Kaiser, M.S., Vassanelli, S., Dai, Q., Zhong, N. (eds) Brain Informatics. BI 2021. Lecture Notes in Computer Science(), vol 12960. Springer, Cham. https://doi.org/10.1007/978-3-030-86993-9_31.

[8] Yousefnezhad, M., & Ghazanfari, M. (2020). Early Prediction of Learning Disabilities in Children Based on k- Nearest Neighbor Algorithm. Journal of Ambient Intelligence and Humanized Computing, 11, 1765-1777. doi: 10.1007/s12652-019-01318-8.

[9] Gupta, N., & Das, A. (2021). Prediction of Learning Disabilities in Children Using Machine Learning Algorithms. Proceedings of the 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 1-5. doi: 10.1109/ICICCS51215.2021.9403471.

[10] Alpaydin, E. (2010). Introduction to machine learning (2nd ed.). Cambridge, MA: MIT Press. (Chapter 6).

[11] R. Paul, T. F. M Andlauer, D. Czamara, D. Hoehn, S. Lucae, B. Pütz, P.G Sämann, (2019). Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Translational Psychiatry, 9(1). doi:10.1038/s41398-019- 0524-4.

[12] M. S Srividya, Mohanavalli, & N. Bhalaji,Behavioral Modeling for Mental Health using Machine Learning Algorithms. J Med Syst 42, 88 (2018). https://doi.org/10.1007/s10916-018-0934-5.

[13] M. Elgendi, & C. Menon, (2020). Machine Learning Ranks ECG as an Optimal Wearable Biosignal for Assessing Driving Stress. IEEE Access, 8, 34362–34374. doi:10.1109/access.2020.2974933.

[14] J. Kashihara, Y. Takebayashi, Y. Kunisato, M. Ito (2021) Classifying patients with depressive and anxiety disorders according to symptom network structures: A Gaussian graphical mixture model-based clustering. PLoS ONE 16(9): e0256902. https://doi.org/10.1371/journal.pone.0256902.

[15] Z. Yang, Li C. Chen, L. Yao, & X. Zhao, (2020). Unsupervised Classifications of Depression Levels Based on Machine Learning Algorithms Perform Well as Compared to Traditional Norm-Based Classifications. Frontiers in Psychiatry, 11. doi:10.3389/fpsyt.2020.00045