

Mental Health Diagnosis (Chronic Fatigue Syndrome and Depression) using Decision Tree Algorithm

Ach. Zubairi ¹, Ahmad Homaidi ^{2✉}, Irma Yunita ³, Jarot Dwi Prasetyo ⁴, Hermanto ⁵

¹ Sharia Business Management, Faculty of Sharia and Islamic Economics, Universitas Ibrahimy, Indonesia

^{2,3} Information Technology, Faculty of Science and Technology, Universitas Ibrahimy, Indonesia

^{4,5} Computer Science, Faculty of Science and Technology, Universitas Ibrahimy, Indonesia

✉ **Corresponding Author** : Ahmad Homaidi (e-mail: ahmadhomaidi@ibrahimyy.ac.id)

Article Information	ABSTRACT
Article History Received : July 02, 2025 Revised : July 11, 2025 Published : July 26, 2025	Mental health is an important aspect that affects an individual's life, impacting productivity, social relationships and overall quality of life. The World Health Organization (WHO) states that one in four people worldwide will face mental health challenges. With the increasing incidence of conditions such as depression and Chronic Fatigue Syndrome (CFS), effective detection and intervention methods are urgently needed. Data mining, specifically using Decision Tree algorithms, presents a promising approach to address this challenge. This study utilizes a quantitative methodology to classify depression and CFS patients using a public dataset. The data collection from Kaggle included variables such as demographics and clinical evaluations, consisting of 1,000 records and 15 predictive attributes. Data preprocessing addressed noise, specifically missing values, to ensure model accuracy above 80%. A Decision Tree was implemented, displaying the interpretability of the method by partitioning the data based on the selected attributes. Evaluation metrics such as accuracy, precision, recall, and F1 score showed accuracy of 99% and precision and recall of 100%. The results emphasize the potential of the Decision Tree in distinguishing between depression and CFS, enabling early intervention through accurate patient identification. This study advocates the integration of such machine learning models into clinical practice to improve mental health diagnostics and management, by addressing an important aspect of public health.
Keywords: <i>Mental Health; Chronic Fatigue Syndrome; Depression; Decision Tree</i>	

INTRODUCTION

Mental health is an important aspect of an individual's life that affects various aspects, including productivity, social relationships, and overall quality of life. According to the World Health Organization (WHO), one in four people in the world will experience a mental health problem at some point in their lives (WHO, 2022). With the increasing prevalence of mental health disorders, such as depression and anxiety, there is an urgent need to develop effective methods for the detection and treatment of these issues. In today's digital age, data mining is emerging as one of the potential solutions to overcome the challenges of determining mental health status.

Depression and Chronic Fatigue Syndrome (CFS) are two mental health conditions that are increasingly common in modern society. According to data from the World Health Organization (WHO), more than 264 million people worldwide experience depression, with prevalence increasing every year (WHO, 2023). On the other hand, CFS, characterized by extreme

and persistent fatigue, is estimated to affect an estimated 836,000 to 2.5 million people in the United States alone (Centers for Disease Control and Prevention, 2024; Fukuda et al., 1994). These two conditions not only affect the quality of life of individuals, but also place a significant burden on the global health system.

Given the diagnostic challenges associated with these conditions, researchers have turned to computational approaches such as machine learning to support mental health assessment. The use of algorithms in the diagnosis and treatment of mental health conditions has been the focus of intense research. One promising approach is the use of Decision Tree algorithm, which is a popular machine learning method for classification and prediction. This method can assist in identifying patterns from complex data and provide easy-to-understand interpretations [3]. Decision Tree works by dividing data into smaller subsets based on specific attribute values, making it easier to identify patterns that can be used to determine a person's mental health status (Breiman et al., 2017). However, despite the many studies that have been conducted, there are still gaps in the application of these algorithms specifically for depression and CFS. To the best of the authors knowledge, existing studies have not specifically focused on differentiating between CFS and Depression using interpretable machine learning algorithms. The novelty of this research lies in the implementation of the Decision Tree algorithm to classify and distinguish overlapping symptoms between the two conditions using a structured dataset. This approach aims to provide clinicians with a transparent, explainable, and efficient diagnostic tool for early mental health intervention

One of the main challenges in the diagnosis of depression and CFS is the complexity of symptoms that often overlap. For example, symptoms of depression such as fatigue, loss of interest, and difficulty sleeping may also be present in individuals with CFS. This makes proper diagnosis difficult, and it often takes a long time to reach an accurate conclusion (Nieciecka et al., 2023). Therefore, a data-driven approach that can utilize Decision Tree algorithms to analyze and interpret symptoms can be an effective solution. In this context, it is important to explore how the Decision Tree algorithm can be applied to differentiate between depression and CFS. Previous research has shown that the use of machine learning models in mental health diagnosis can improve accuracy and efficiency (Thieme et al., 2020). However, further research is needed to determine the effectiveness of this algorithm in the context of CFS.

RESEARCH METHODS

This research uses a quantitative research approach to classify people with depression and CFS available in the public dataset. The quantitative approach was chosen due to its ability to analyze large data. In addition, the use of the Decision tree algorithm was chosen with the aim of obtaining model accuracy results above 80%.

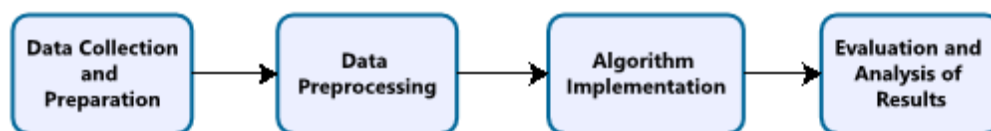


Figure 1. Proposed Model Overview

Data Collection and Preparation

Data collection is a crucial first step in this research, where the data used is taken from the Kaggle platform (Aliyev, 2025). Kaggle is one of the largest data sources that provides various datasets for analysis and research. In this context, the dataset includes information on patients with depression and Chronic Fatigue Syndrome (CFS), including demographic variables, symptoms, and clinical evaluation results. The data is synthetic but modeled based on real-world clinical features such as fatigue severity, depression scores, sleep quality, cognitive symptoms, and lifestyle factors. The data includes realistic impairments and missing values to simulate conditions found in actual medical datasets. This dataset contains 1000 data records and has fifteen predictive attributes including: age, gender, sleep_quality_index, brain_fog_level, physical_pain_score,

stress_level, depression_phq9_score, fatigue_severity_scale_score, pem_duration_hours, hours_of_sleep_per_night, pem_present, work_status, social_activity_level, exercise_frequency, meditation_or_mindfulness, and diagnosis as its labels.

Data Preprocessing

Data preprocessing is an important step that aims to improve data quality before it is applied in machine learning algorithms. At this stage, handling data noise is the main focus, as unclean or irrelevant data can significantly affect the analysis results. Noise in data can come from various sources, including measurement errors, inconsistent data entry, and incomplete data. Noise in data can reduce model accuracy by up to 30% if not handled properly (Nazari et al., 2018; Qinghao et al., 2020; Saseendran et al., 2019; Tiwari et al., 2020). Therefore, it is important to apply proper techniques to clean the data before further analysis.

Algorithm Implementation

The implementation of the Decision Tree algorithm is a key step in this research, where a model is built to classify depression and Chronic Fatigue Syndrome (CFS) status based on pre-selected features. Decision trees work by recursively dividing the data based on the most informative features. This process continues until a certain concentration criterion is reached, such as the maximum tree depth or the minimum number of samples per leaf. Decision Tree is one of the most popular machine learning algorithms due to its ability to provide easy-to-understand interpretations. It works by dividing a dataset into smaller subsets based on feature values, resulting in a tree structure that can be used to make decisions. Decision trees have advantages in terms of transparency and interpretability (Breiman et al., 2017; Sivananda & Kumar, 2024), which is particularly important in the context of mental health. The steps for implementing this algorithm are through the data division, training and testing stages, followed by searching for entropy and information gain which are then used as the basis for forming a decision tree by applying pruning using the RapidMiner Studio.

Entropy is a measure of uncertainty or purity in a dataset. The higher the entropy, the less "pure" the data (various classes). If all data belongs to one class $\rightarrow$ entropy = 0 (pure). The purpose of calculating entropy is to measure how random or unpredictable the dataset is based on its class distribution.

$$Entropy(S) = - \sum_{i=1}^n p_i \cdot \log(p_i) \quad (1)$$

where:

S = Dataset

n = number of classes

p_i = proportion of data in the i -th class

After selecting one attribute A , the dataset is divided based on the unique values of that attribute. The purpose of this stage is to determine how good attribute A is at reducing the uncertainty of the dataset after dividing by its values.

$$Entropy(S, A) = \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \quad (2)$$

where:

$\text{Values}(A)$ = unique values of the attribute A

S_v = subset of S where attribute $A = v$

Next is the Information Gain which is used to show how much "information" is gained when attribute A is used to split the data. The purpose of this stage is to find the best attribute to be used as a node in the decision tree. The attribute with the **highest Gain** value is selected as the **node** in the Decision Tree.

$$Gain(S, A) = Entropy(S) - Entropy(S, A) \quad (3)$$

In this stage, the processed dataset is divided into two parts: training data and testing data. The training data is used to build the model, while the testing data is used to evaluate the performance of the built model. This division is usually done in a ratio of 80:20 or 90:10, depending on the size of the dataset. A proper data split can significantly affect the accuracy of the model (Joseph, 2022). Therefore, it is important to ensure that the data split is done correctly to produce a reliable model.

Evaluation and Analysis of Result

Model evaluation is an important step to assess how well the Decision Tree algorithm has been applied in this study. Once the model is built and tested, evaluation metrics such as accuracy, precision, recall, and F1-score are used to measure the performance of the model. Accuracy is the proportion of correct predictions out of the total predictions made, while precision measures how many of the positive predictions are actually positive. Recall, on the other hand, measures how many of the overall positive cases were successfully identified by the model. F1-score is the harmonic mean of precision and recall, which gives a more comprehensive picture of the model's performance, especially on imbalanced datasets (Chauhan, 2020; Hamner & Frasco, 2018).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (6)$$

$$\text{F1-Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \quad (7)$$

RESULTS AND DISCUSSION

In this research, the data processing and algorithm implementation process uses a laptop with Intel Core i-3-7020U processor specifications with a speed of 2.3 GHz, 8 GB RAM, and 1 TB of storage. For processing tools using RapidMiner and the dataset used uses a public dataset taken from kaggle (Aliyev, 2025). The processing architecture of the algorithm application in this study can be seen in Figure 2.

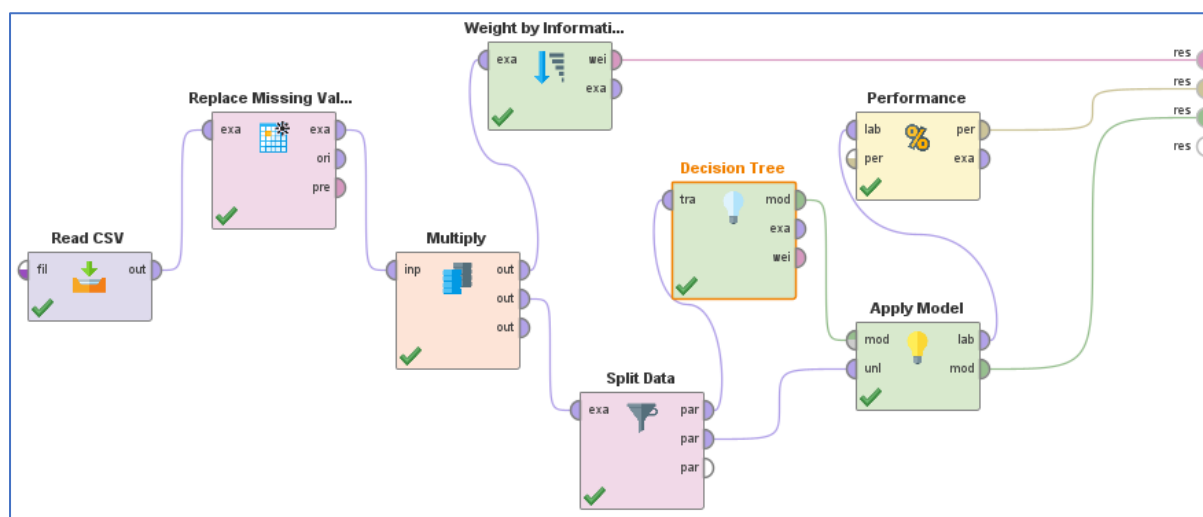


Figure 2. Algorithm Implementation Architecture

Replace Missing Value

From the dataset after data analysis, it is known that there is some noise (missing value) data spread across several attributes. Table 1 below is the distribution of noise data contained in the dataset.

Tabel 1. Attribute with noise data

No	Attribute Name	Attribute Type	Number of Missing Value
1	age	Predictive	-
2	gender	Predictive	-
3	sleep_quality_index	Predictive	47
4	brain_fog_level	Predictive	48
5	physical_pain_score	Predictive	37
6	stress_level	Predictive	48
7	depression_phq9_score	Predictive	22
8	fatigue_severity_scale_score	Predictive	21
9	pem_duration_hours	Predictive	24
10	hour_of_sleep_per_night	Predictive	21
11	pem_present	Predictive	-
12	work_status	Predictive	47
13	social_activity_level	Predictive	40
14	exercise_frequency	Predictive	39
15	meditation_or_mindfulness	Predictive	11
16	Diagnosis	Label	-

Preprocessing the data to handle missing values as indicated in Table 1, here is done by filling in the average value using the replace missing value operator, whose pseudocode can be seen as follows.

```

FOR each attribute (kolom) in dataset:
  IF attribute is numerical:
    Initialize list_of_valid_values = []
    FOR each value in the attribute:
      IF value is NOT noise (misalnya bukan NaN, NULL, kosong, atau outlier):
        Append value to list_of_valid_values
    Compute mean_value = AVERAGE of list_of_valid_values

    FOR each value in the attribute:
      IF value is noise (misalnya NaN, NULL, kosong, atau outlier):
        Replace value with mean_value

Save or return the cleaned dataset

```

Figure 3. Pseudocode for handling noise data

Model Execution and Output

After handling missing values, the next step in this research is the implementation of the Decision Tree algorithm. Decision Tree is one of the popular methods in machine learning due to its ability to handle categorical and numerical data well. This algorithm works by dividing a dataset into smaller subsets based on certain attribute values, thus forming a tree structure that can be used for decision making.

The Decision Tree implementation process starts with the selection of relevant attributes from the dataset. Once the attributes are selected, the algorithm will build a decision tree based on certain criteria. The results of information gain weighting can be seen as table 2.

Tabel 2. Information Gain Weighting Results

Atribut	Weight
pem_present	1,000
depression_phq9_score	0,941
fatigue_severity_scale_score	0,531
sleep_quality_index	0,203

physical_pain_score	0,075
hours_of_sleep_per_night	0,047
stress_level	0,017
brain_fog_level	0,010
social_activity_level	0,005
pem_duration_hours	0,005
age	0,004
work_status	0,004
exercise_frequency	0,003
meditation_or_mindfulness	0,002

Based on the weighting results of table 2, the pem_present attribute is used as a node. Furthermore, it is important to prune the Decision Tree to avoid overfitting (Setiawan et al., 2023). Overfitting occurs when the model is too complex and unable to generalize to new data. Pruning is done by removing insignificant branches, thereby increasing the accuracy of the model on test data (Mehedi Shamrat et al., 2021; Xie & Shang, 2014). The results of attribute pruning based on the implementation of this algorithm can be seen as Figure 4.

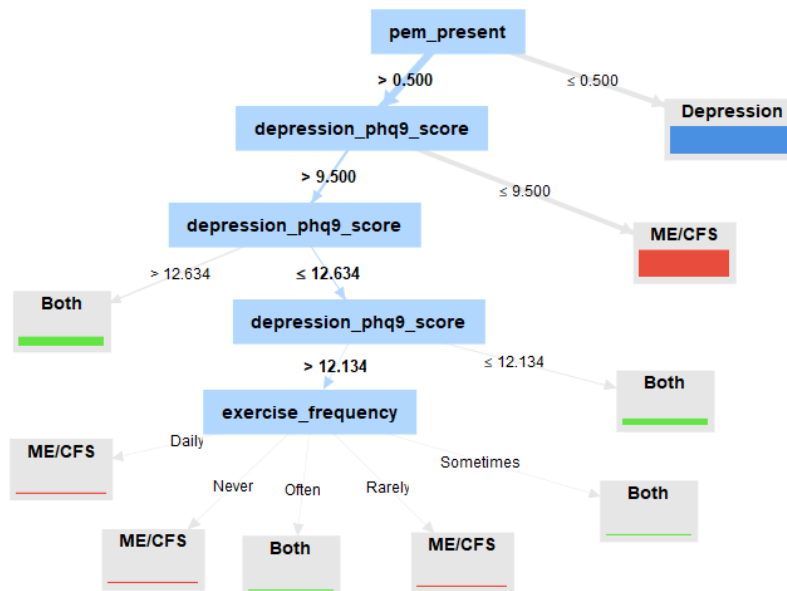


Figure 4. Decision tree graph model

Figure 4 provides a visual view of how the model makes decisions based on pruning the attributes in the data. I am excited about the potential of utilizing this decision tree image to communicate the model results with the team and other stakeholders. Furthermore, the model can be used to predict depression and CFS status in new patients. Using this model, medical personnel can quickly and accurately assess the patient's condition and plan appropriate interventions. This shows the great potential of the Decision Tree algorithm in the field of mental health, especially in the context of diagnosis and management of CFS and depression.

Model Evaluation

After the implementation of the Decision Tree algorithm, the next step is to evaluate the model that has been built. Model evaluation is very important to determine how good the model is in making predictions. Some commonly used metrics in model evaluation are accuracy, precision, recall, and F1-score. These metrics provide a clear picture of the model's performance in classifying the status of depression, CFS or both.

The evaluation is done by dividing the dataset into training and test data (Arunakumari et al., 2023; Wernigg & Wernigg, 2022). The training data is used to build the model, while the test data is used to measure the performance of the model. The evaluation results show that the Decision Tree model built by pruning attributes has an accuracy of 99%, precision of 100%, recall

of 100%, and F1-score of 100%, indicating that this model is very effective in identifying patients with depression, CFS or both.

Tabel 4. Confusion matrix

	true Depression	true Both	true ME/CFS	class precision
pred. Depression	40	0	0	100.00%
pred. Both	0	20	1	95.24%
pred. ME/CFS	0	0	39	100.00%
class recall	100.00%	100.00%	97.50%	

CONCLUSION

In this study, we used data that included various variables, such as physical, emotional, and medical history symptoms. Using the Decision Tree, we were able to identify certain patterns related to depression and CFS symptoms. For example, in one case, a patient who reported extreme fatigue, sleep disturbances, and difficulty concentrating could be easily identified as possibly suffering from CFS. The success of the algorithm in classifying this patient shows the great potential of this approach in clinical practice.

In addition, we also found that Decision Tree can be used to screen patients who are at high risk of depression. By analyzing factors such as family history, stress, and lifestyle, we were able to create a model that predicts the likelihood of someone developing depression. These results are very important, as they allow for early interventions that can reduce the negative impact of this condition. For example, a patient identified as high-risk could be provided with counseling or therapy earlier, which could significantly improve their quality of life.

Overall, this study shows that the Decision Tree algorithm has great potential in the determination of depression and CFS status. With its ability to analyze data efficiently and provide results that can be interpreted easily, Decision Tree can be a valuable tool for healthcare professionals in the diagnosis and treatment of these conditions. With mental health awareness on the rise, it is important that we continue to develop and implement technologies that can help us understand and address these issues more effectively.

ACKNOWLEDGEMENTS

Our gratitude goes to those who have supported and helped until the completion of this research.

REFERENCES

- Aliyev, A. (2025, June 8). *ME/CFS vs Depression Classification Dataset*. <https://www.kaggle.com/datasets/storytellerman/mecfs-vs-depression-classification-dataset>
- Arunakumari, B. N., Patel, T., Patil, S., Roopa, L. S., & Singh, V. (2023). Enhancing the Quality and Efficiency of Mental Health Care using Decision Trees. *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*. <https://doi.org/10.1109/ICCCNT56998.2023.10306470>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). Classification and regression trees. In *Classification and Regression Trees*. <https://doi.org/10.1201/9781315139470>
- Centers for Disease Control and Prevention. (2024, May 10). *Myalgic Encephalomyelitis/Chronic Fatigue Syndrome*. Centers for Disease Control and Prevention. <https://www.cdc.gov/mecfs/about/index.html>

- Chauhan, N. S. (2020). Model Evaluation metrics in Machine learning. *Kdnuggets*.
- Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., Komaroff, A., Schluederberg, A., Jones, J. F., Lloyd, A. R., Wessely, S., Gantz, N. M., Holmes, G. P., Buchwald, D., Abbey, S., Rest, J., Levy, J. A., Jolson, H., Peterson, D. L., Vercoulen, J. H. M. M., ... Reeves, W. C. (1994). The Chronic Fatigue Syndrome: A Comprehensive Approach to Its Definition and Study. *Annals of Internal Medicine*, 121(12). <https://doi.org/10.7326/0003-4819-121-12-199412150-00009>
- Hamner, B., & Frasco, M. (2018). Metrics: Evaluation metrics for machine learning. In *R package version 0.1*.
- Joseph, V. R. (2022). Optimal ratio for data splitting. *Statistical Analysis and Data Mining*, 15(4). <https://doi.org/10.1002/sam.11583>
- Mehedi Shamrat, F. M. J., Chakraborty, S., Billah, M. M., Das, P., Muna, J. N., & Ranjan, R. (2021). A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm. *Proceedings of the 5th International Conference on Trends in Electronics and Informatics, ICOEI 2021*. <https://doi.org/10.1109/ICOEI51242.2021.9452898>
- Nazari, Z., Nazari, M., Sayed, M., Danish, S., & Kang, D. (2018). Evaluation of Class Noise Impact on Performance of Machine Learning Algorithms. In *IJCSNS International Journal of Computer Science and Network Security* (Vol. 18, Issue 8).
- Nieciecka, A., Tomys-Składowska, J., Lamch, M., Jabłońska, M., Błasik, N., Janiszewska, M., & Wójcik-Kula, A. (2023). Chronic fatigue syndrome – challenge in diagnosis and management: a literature review. *Journal of Medical Science*. <https://doi.org/10.20883/medical.e877>
- Qinghao, G., Ligu, S., & Sunying, H. (2020). Impact of data set noise on distributed deep learning. *Journal of China Universities of Posts and Telecommunications*, 27(2). <https://doi.org/10.19682/j.cnki.1005-8885.2020.1005>
- Saseendran, A. T., Setia, L., Chhabria, V., Chakraborty, D., & Roy, A. B. (2019). Impact of Noise in Dataset on Machine Learning Algorithms. *Machine Learning Module CS7CS4/CS4404*, February.
- Setiawan, I., Fina Antika Cahyani, R., & Sadida, I. (2023). Exploring Complex Decision Trees: Unveiling Data Patterns And Optimal Predictive Power. *Journal of Innovation And Future Technology (IFTECH)*, 5(2). <https://doi.org/10.47080/ifttech.v5i2.2829>
- Sivananda, Mr. M., & Kumar, Dr. G. K. (2024). Classification and Regression Based on Decision Tree Algorithm for Machine Learning. *Interantional Journal Of Scientific Research In Engineering And Management*, 08(02). <https://doi.org/10.55041/ijrem28533>
- Thieme, A., Belgrave, D., & Doherty, G. (2020). Machine Learning in Mental Health: A systematic review of the HCI literature to support the development of effective and implementable ML Systems. In *ACM Transactions on Computer-Human Interaction* (Vol. 27, Issue 5). <https://doi.org/10.1145/3398069>
- Tiwari, V., Garg, B., & Sharma, U. P. (2020). Significant Impact of Improved Machine Learning Algorithm in The Processes of Large Data Sets. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. <https://doi.org/10.32628/cseit206133>
- Wernigg, R., & Wernigg, M. (2022). A case study for assessing the utility of a decision tree based learning algorithm in mental health inpatient care quality management. *European Psychiatry*, 65(S1). <https://doi.org/10.1192/j.eurpsy.2022.454>
- WHO. (2022, June 12). *Mental health*. <https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response>

- WHO. (2023, March 31). *Depressive disorder (depression)*. <https://www.who.int/news-room/fact-sheets/detail/depression>
- Xie, H., & Shang, F. (2014). The study of methods for post-pruning decision trees based on comprehensive evaluation standard. *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2014*. <https://doi.org/10.1109/FSKD.2014.6980959>