Proceedings of the 5th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/112/2024.17914

Research on the Adoption of Machine Learning in the Domain of Sleep Disorder Detection

Kaiwen Deng[1,a,*]

[1]School of Software Engineering, Software Engineering Institute of Guangzhou, Guangzhou, 510990, China
a. dkw2266@smail.seig.edu.cn
**corresponding author*

Abstract: Healthy sleep is very important for people's vitality and physical health. However, due to the need for socio-economic development and changes in people's lifestyles, people suffer from sleep disorder problems. Timely detection of sleep disorders is important for people's lives afterwards. This study collected and evaluated Laksika Tharmalingam data using a variety of machine learning techniques, including Random Forests, Decision Trees, and Logistic Regression (LR) to find out the relationship between 11 factors including lifestyle factors and sleep disorders. The research found that the LR algorithm produced the best model with 89.33 per cent accuracy, 90.24 per cent precision, 89.33 per cent recall, and an f1-score of 0.89. Lifestyle-related factors such as occupation, sleep duration, and level of physical activity all have an impact on whether or not a sleep disorder develops. This paper can serve as a valuable resource and provide insightful information for future research endeavors focused on sleep disorders, contributing to a deeper understanding of the complexities surrounding these conditions and potentially guiding the development of more effective interventions and treatments.

Keywords: Sleep disorder, prediction, decision tree, random forest, logistic regression.

1. Introduction

Nowadays, sleep health has become a health topic of concern for most people. With the development of society, people's revenue has gradually increased. Despite rising global social development and economic levels, many people face chronic sleep deprivation, and the resulting health and economic costs are gradually increasing [1,2]. Some researchers have concluded that among the countries participating in the survey, such as Europe, the United States and Japan, the overall prevalence rate of sleep problems exceeds 20%, and the rate in the United States even exceeds half of the total. Widespread sleep problems impose a heavy burden on different countries and will reduce people's ability to perform their daily lives [3]. Therefore, the issue of sleep disorders has become an important research topic.

The development of sleep disorders is very complex, and it is the result of the interaction of several factors. Physiological factors such as gender, heart rate, and age have been found to have a significant effect on the occurrence of sleep disorders [4-6]. In addition to these factors, lifestyle habits are also factors that should not be ignored, including physical activity, quality of sleep, and so on [7,8].

In terms of using machine learning to study the problem of sleep disorders, R Alazaidah et al. have analyzed and predicted the factors contributing to sleep disorders using a variety of learning approaches and models such as Logistic Regression (LR), Random Forest, and Parsimonious Bayes [9]. Current research has led to a deeper understanding of sleep disorders, but there is still little research linking sleep disorders to people's lifestyles.

Therefore, 11 factors such as gender, age, occupation, and sleep duration are researched in this study to determine if these factors have a relevant effect on the occurrence of sleep disorders. Three distinct machine learning algorithms—random forests, decision trees, and LR, are analyzed and compared to determine if sleep disorders and lifestyle are related and to determine which approach produces the best model.

Predicting and assessing sleep disorders using machine learning methods can change the cumbersome consultation process for patients and quickly lead to the conclusion of whether or not there is a sleep disorder problem, which can facilitate the doctor to determine the next step of the treatment plan and accelerate the efficiency of hospital visits.

2. Material and methods

2.1. Data

Laksika Tharmalingam manually gathered the sleep health and lifestyle dataset for this study from the Kaggle website. There are 374 samples in all in the dataset. Individual ID, gender, age, occupation, amount of sleep, quality of sleep, physical activity level, stress level, body mass index category, blood pressure, heart rate, daily steps, and sleep disorders are among the attributes included in the dataset. Specifics of the attributes are shown in Table 1, and due to the large number of occupation options in the table, the ranges will be denoted by *.

2.2. Methodology

Table 1: Attribution table

Attribution	Type	Range
Person ID	Numeric	1-374
Gender	Categorical	Male/Female
Age	Numeric	27-59
Occupation	String	*
Sleep Duration	Numeric	5.8-8.5
Quality of Sleep	Numeric	4-9
Physical Activity Level	Numeric	30-90
Stress Level	Numeric	3-8
BMI Category	Categorical	Normal/Overweight/Obese
Blood Pressure	Numeric	75-95/115-142
Heart Rate	Numeric	65-86
Daily Steps	Numeric	3000-10000
Sleep Disorder	Categorical	None/Insomnia/Sleep Apnea

2.2.1. Decision tree

Decision tree-based machine learning algorithms are a supervised, regularity-based binary tree construction technique [10-12]. It is an inductive learning algorithm that presents models of decision rules and classification results in a tree data structure. The key point of a decision tree is to take known data, which appear to be disordered and cluttered and transform them by some technical means into a tree model that can predict the unknown data, where each path from the root node to the leaf nodes represents a rule for decision making. Figure 1 shows the basic process of a decision tree.

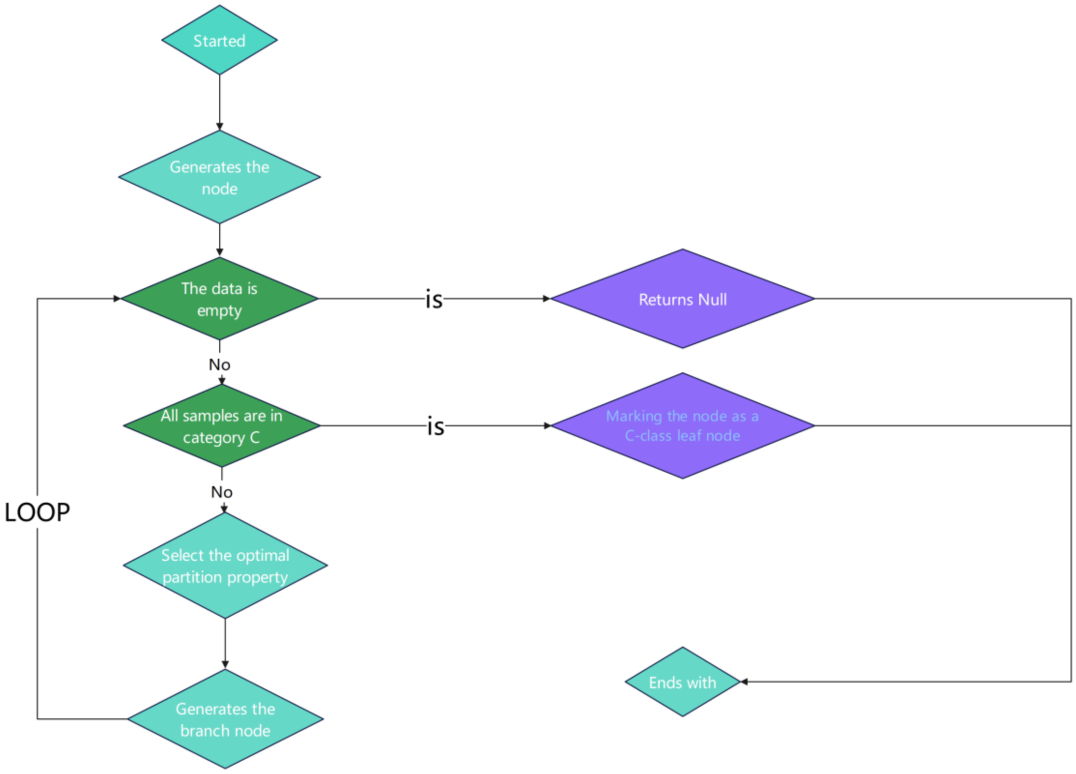
Started
Generates the node
The data is empty
is
Returns Null
No
All samples are in category C
is
Marking the node as a C-class leaf node
No
LOOP
Select the optimal partition property
Ends with
Generates the branch node

Figure 1: Decision Tree Flowchart

This research adopts the Gini index to divide the selection, to generate the optimal decision tree, the smaller the Gini index the better the final formation of the decision tree, the formula is as follows:

$$Gini(D) = \sum_{k=1}^{n} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{\mathrm{n}} p_k^2 \tag{1}$$

In this formula, D denotes the set of samples in the decision tree, n denotes how many different categories of samples are in the decision tree, and the proportion of samples in each category is p_k (where k=1, 2...n).

2.2.2. Random Forest

Random Forest is classified as a supervised machine learning ensemble algorithm, a classifier that contains multiple decision trees. In prediction, each decision tree gives a classification result, and finally, the category with the most votes is chosen as the model prediction. It has more obvious advantages over decision trees and can effectively prevent overfitting. The Random Forest algorithm is a better way to classify large amounts of data and aims to achieve the highest accuracy in both classification and prediction [13]. The random forest algorithm is illustrated in Figure 2:

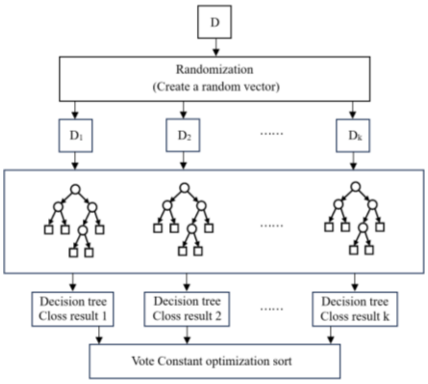

Figure 2: Randon Forest Schematic

2.2.3. LR

LR is a machine learning supervised learning model that is a generalized version of linear regression analysis. It is mostly used to resolve multiple classification or binary classification issues. The model is trained by given n sets of data acting as a training set and at the end of the training the given test set is classified.

Since LR is essentially linear regression, the special feature is that a layer of the logistic function, the sigmoid function, is added to feature-to-result mapping, so logistic regression = linear regression + sigmoid.

Linear regression formula:

$$y = w^{T}x + b \tag{2}$$

For greater simplicity and convenience, the weight vectors and input vectors are expanded and still denoted w and x.

i.e.

$$y = wx \tag{3}$$

The sigmoid formula:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4}$$

Thus, the formula for LR:

$$\pi(x) = \frac{1}{1+e^{-w^{T}x}} \tag{5}$$

3. Result

Table 2: Summary of basic information on data disaggregation

project	options	frequency	percentage
Sleep Disorder	Insomnia	77	20.59%
	None	219	58.56%
	Sleep Apnea	78	20.86%
	sum	374	100.00%
aggregation	effect	374	100.00%
	deficiency	0	0.00%
	total	374	100.00%

To establish the machine learning model, the following independent variables are used: gender, age, occupation, sleep duration, sleep quality, physical activity level, stress level, body mass index category, blood pressure, heart rate, and daily step count. The dependent variable is the result of sleep disorders. As can be seen from Table 2, sleep disorders consist of two types, insomnia and sleep apnea. The proportion of people without sleep disorders is 58.56%, the proportion of people with insomnia symptoms is 20.59% and the proportion of people with sleep apnea is 20.86%. All data in the dataset is valid.

3.1. Decision Tree Result

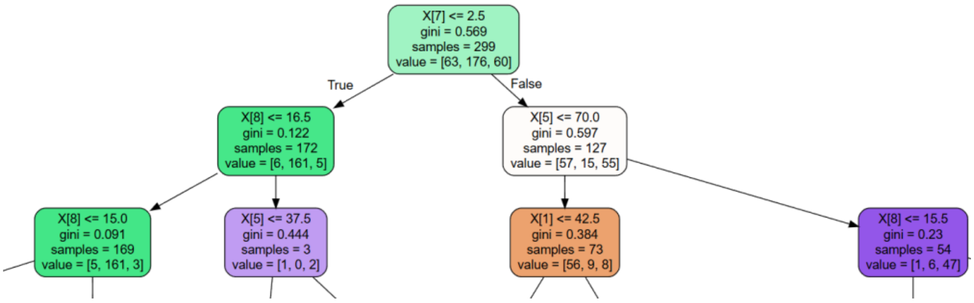

Figure 3: Part of the decision tree structure diagram

Figure 3 shows part of the structure of the decision tree and some of its node indicators (see Figure 7 in the appendix for the complete decision tree), Each node's first line has a variable called X[i] that contains the splitting indicator and the name of the attribute that was used to split the node. i.e., you can judge which independent variable it is through the content of the first line; as this research uses the gini purity judgement indicator on the decision tree, the node The second line (gini) indicates the corresponding Gini coefficient, the smaller the Gini coefficient, the higher the data purity; the sample variable in the third line indicates the number of samples contained in the current node; the value array in the fourth line indicates how many samples there are in different categories, and the one with the largest number of samples is the category of the node; due to the different situations in different nodes, some nodes may lack certain indicators. May be missing some indicators.

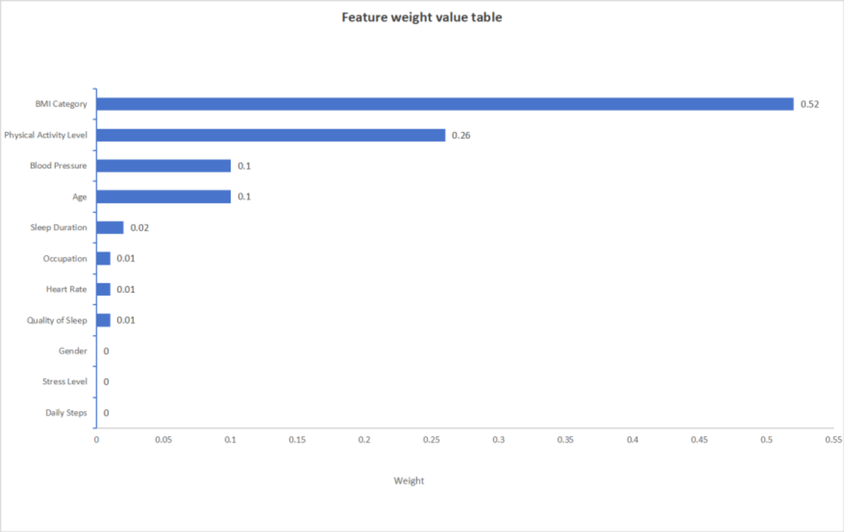

Figure 4: Feature weight value table (decision tree)

Figure 4 shows the importance of each heading in contributing to the model with a summed value of 1. From the above table, it can be seen that: the weight of the BMI Category is 52%, which is the feature with the greatest weight and is essential in the building of the model; the Physical Activity Level is 26%, which is the second most important feature. The weight of Blood Pressure is 10%; the total weight of the above three features accounts for 88%; and the weight of the remaining eight features is not significant.

Table 3: Model summary table (decision tree)

Project	Parameter Name	Value
Model parameter setting	Data preprocessing	None
	Training set Proportion	0.8
	Node splitting Standard	Gini
	Node partitioning method	best
	Node splitting Minimum sample number	2
	Leaf node Minimum sample number	1
	Tree Maximum depth	unlimited
Model evaluation effect	Accuracy	86.667%
	Precision (comprehensive)	86.731%
	Recall rate (comprehensive)	86.667%
	f1-score	0.861

The information in Table 3 makes it clear that the decision tree model building result. The training set ratio is 0.8, the data does not need to be pre-processed, the node splitting criterion is Gini, the node partitioning mode is optimal, and the maximum depth of the tree is unrestricted for decision tree modelling. The final model achieved an accuracy of 86.67%, a combined precision of 86.73%, a combined recall of 86.67%, and a f1-score of 0.86. The model results are acceptable.

3.2. Random Forest Result

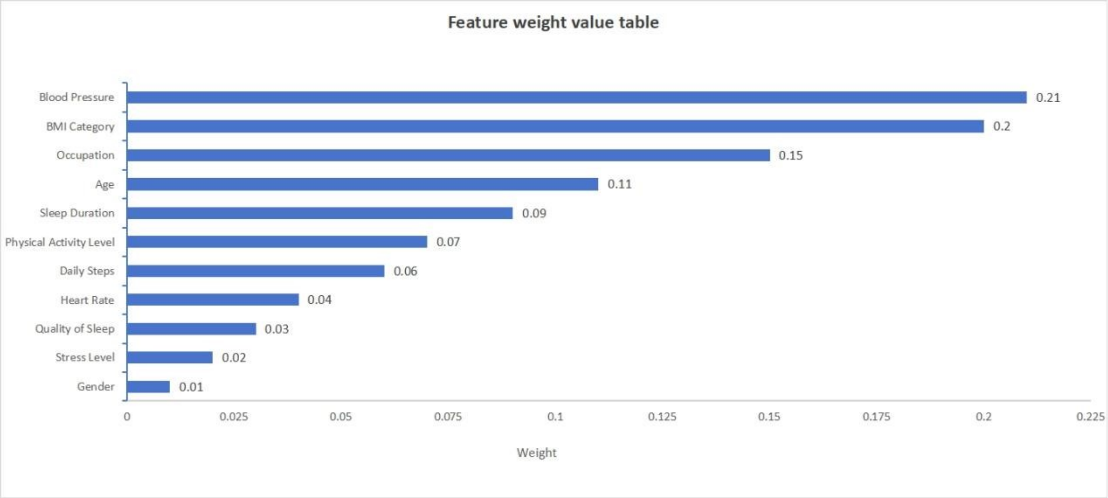

Figure 5: Feature weight value table (random forest)

Combining the results of Figure 5, Blood Pressure accounts for 21%, this aspect is the most significant in the construction of the model.

BMI Category accounts for 20%; Occupation accounts for 15%; Age accounts for 11%; Sleep The weight of Duration is 9%; the total weight of the above five features accounts for 76%; the remaining six features are less important in contributing to the model.

Table 4: Model summary table (random forest)

Project	Parameter Name	Value
Model parameter setting	Data preprocessing	None
	Training set Proportion	0.8
	Decision tree quantity	100
	Node splitting Standard	Gini
	Node splitting Minimum sample number	2
	Leaf node Minimum sample number	1
	Tree Maximum depth	unlimited
	Limit of Maximum number of features	auto
	Put back sampling	Yes
	Perform out-of-pocket data testing	Yes
Model evaluation effect	Accuracy	89.333%
	Precision (comprehensive)	89.358%
	Recall rate (comprehensive)	89.333%
	f1-score	0.889

Through an examination of Table 4's data, it can be obtained that the Random Forest is modelled by keeping the same model conditions as the decision trees except for the additional conditions of having 100 decision trees with put-back sampling and out-of-bag testing. This table also shows that: the final model obtained 89.33% accuracy, 89.36% precision (combined), 89.33% recall (combined) and 0.89 f1-score (combined) on the test set. The model results are acceptable.

3.3. LR result

Table 5: Model summary table (LR)

Project	**Parameter Name**	**Value**
Model parameter setting	Data preprocessing	None
	Training set Proportion	0.8
	optimization algorithm	lbfgs
	regularization	L2
	Setting the intercept	是
	Maximum number of iterations	100
	Model convergence parameters	0.001
Model evaluation effect	Accuracy	89.333%
	Precision (comprehensive)	90.237%
	Recall rate (comprehensive)	89.333%
	f1-score	0.894

In the algorithm using LR, the training set scale is evidently set to 0.8. The lbfgs optimization algorithm is used and LR modelling is carried out using L2 regularization and with a set intercept.

From Table 5: The final model obtained 89.33% accuracy, 90.24% precision (combined), 89.33% recall (combined) and 0.89 f1-score (combined) on the test set. The model results are acceptable.

4. Discussion

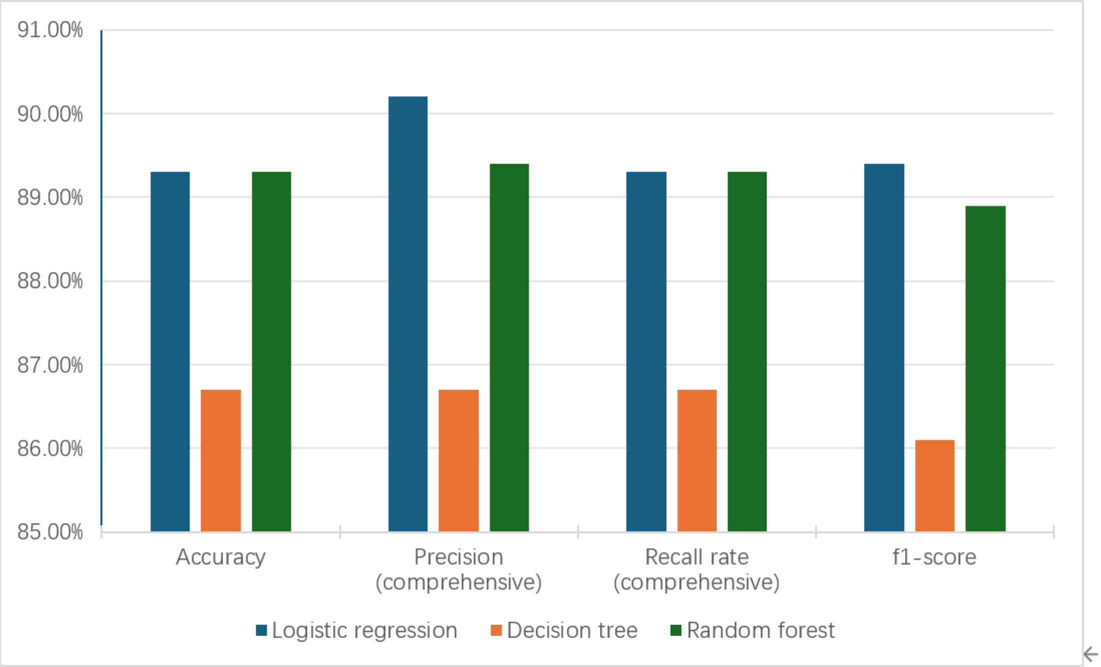

Figure 6: Comparison of different methods

Figure 6 shows the comparison of three basic algorithms for learning from machines: LR, decision trees and random forests, on four performance metrics: accuracy, precision, recall rate, and f1-score. It can be found that LR and random forests outperform decision trees overall, with LR outperforming random forests on two metrics: precision and f1-score. Firstly, compare the Random Forest and Decision Tree, there are multiple decision trees in the Random Forest and the Random Forest can be voted to select the optimal classification result. Due to this reason, it makes a more obvious difference in the result of the model. On the other hand, decision trees are prone to overfitting or underfitting, especially on small datasets (e.g., the 374-article dataset in this paper). This problem limits the generalization ability and robustness of the model. Strong correlations among several possible input factors may lead to the selection of variables that enhance the model statistics without having a causal relationship with the desired output [14]. LR, as a linear model, is specifically used to deal with classification problems, and the research in this paper is also to classify the question of whether different people have sleep disorders, which is in line with the use of LR scenarios, all in the results LR will be better than the other two models.

In this paper, only three machine learning methods are used to study the sleep disorder problem using a small dataset. More comprehensive and complex machine learning algorithms can be used in subsequent research on small datasets.

Lebedev et al. and Cohen et al. have researched random forests and noted that they work well even when data in the dataset is lost and maintain accuracy in the face of rapid data growth [15,16]. This paper does not conduct research in these two areas and should be validated later using the models in this paper with missing data and short periods of rapidly growing data.

5. Conclusion

This paper employs LR, decision tree and random forest algorithms and uses a small dataset to construct a model to determine whether a person is suffering from a sleep disorder. Among the three machine learning algorithms—random forest, decision tree, and LR—it was discovered that LR performed the best. Its model has the best performance with 89.33% accuracy, 90.24% precision, 89.33% recall and an f1-score of 0.89.

This paper's primary contribution is to compare three different machine learning algorithms to find out the superiority of the LR algorithm, among these three algorithms, the LR algorithm is simpler to implement compared to the other two algorithms and is powerful enough to help the doctors to quickly determine whether the person attending the clinic suffers from the problem of sleep disorders or not by using the model, to improve the efficiency of the hospitals and the accuracy of the judgement.

However, due to the small amount of data in the dataset, the evaluation of the model is only acceptable, and it is hoped that more relevant data can be collected, and more varied methods can be used to make the detection of sleep problems as accurate and fast as possible.

References

[1] *Bonnet, M. H., & Arand, D. L. (1995). We are chronically sleep deprived. Sleep, 18(10), 908-911.*

[2] *Niekamp, P. (2019). Economic conditions and sleep. Health Economics, 28(3), 437-442.*

[3] *Léger, D., Bayon, V., Laaban, J. P., & Philip, P. (2008). An international survey of sleeping problems in the general population. Current Medical Research and Opinion, 24(1), 307-317.*

[4] *Zeng, L. N., Chen, L. G., Yang, C. M., et al. (2020). Gender difference in the prevalence of insomnia: A meta-analysis of observational studies. Frontiers in Psychiatry, 11, 577429.*

[5] *Takano, Y., Imai, H., Takahashi, M., et al. (2024). Nonrestorative sleep and its association with insomnia severity, sleep debt, and social jetlag in adults: Variations in relevant factors among age groups. Sleep Medicine, 121, 203-209.*

[6] *Sequeira, V. C. C., Bandeira, P. M., & Azevedo, J. C. M. (2019). Heart rate variability in adults with obstructive sleep apnea: A systematic review. Sleep Science, 12(03), 214-221.*

[7] *Takano, Y., Imai, H., Takahashi, M., et al. (2024). Nonrestorative sleep and its association with insomnia severity, sleep debt, and social jetlag in adults: Variations in relevant factors among age groups. Sleep Medicine, 121, 203-209.*

[8] *Kang, J. M., Kang, S. G., Cho, S. J., & Park, H. J. (2013). Factors associated with poor sleep quality in primary care. Korean Journal of Family Medicine, 34(2), 107.*

[9] *Alazaidah, R., Jaradat, M. A., Saleh, A. I., et al. (2023). The potential of machine learning for predicting sleep disorders: A comprehensive analysis of regression and classification models. Diagnostics, 14(1), 27.*

[10] *Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31, 249-268.*

[11] *Murthy, K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery, 2(4), 345-389.*

[12] *Decision tree learning. (n.d.). In Wikipedia. Retrieved from http://en.wikipedia.org/wiki/Decision_tree_learning*

[13] *Shaik, A. B., & Srinivasan, S. (2019). A brief survey on random forest ensembles in classification model. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2 (pp. 213-222). Springer Singapore.*

[14] *Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: Applications for classification and prediction. Shanghai Archives of Psychiatry, 27(2), 130.*

[15] *Lebedev, A. V., Westman, E., Van Westen, G. J., et al. (2014). Random forest ensembles for detection and prediction of Alzheimer's disease with a good between cohort robustness. NeuroImage: Clinical, 6, 115-125.*

[16] *Cohen, L., Schwing, A. G., & Pollefeys, M. (2014). Efficient structured parsing of facades using dynamic programming. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3206-3213).*

Appendix

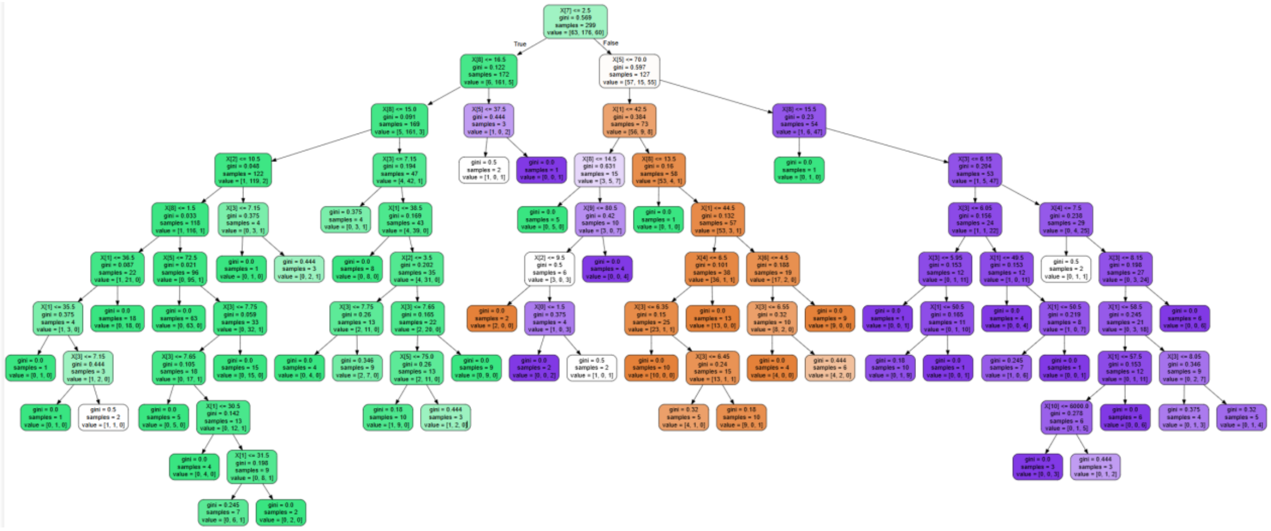
X[7] <= 2.5 gini = 0.569 samples = 299 value = [63, 178, 60]
True
False
X[8] <= 16.5 gini = 0.122 samples = 172 value = [6, 161, 5]
X[5] <= 70.0 gini = 0.597 samples = 127 value = [57, 15, 55]
X[8] <= 15.0 gini = 0.091 samples = 169 value = [5, 161, 3]
X[5] <= 37.5 gini = 0.444 samples = 3 value = [1, 0, 2]
X[1] <= 42.5 gini = 0.384 samples = 73 value = [56, 9, 8]
X[8] <= 15.5 gini = 0.23 samples = 54 value = [1, 6, 47]
X[2] <= 10.5 gini = 0.048 samples = 122 value = [1, 119, 2]
X[3] <= 7.15 gini = 0.194 samples = 47 value = [4, 42, 1]
gini = 0.5 samples = 2 value = [1, 0, 1]
gini = 0.0 samples = 1 value = [0, 0, 1]
X[8] <= 14.5 gini = 0.631 samples = 15 value = [3, 5, 7]
X[8] <= 13.5 gini = 0.16 samples = 58 value = [53, 4, 1]
gini = 0.0 samples = 1 value = [0, 1, 0]
X[3] <= 6.15 gini = 0.204 samples = 53 value = [1, 5, 47]
X[8] <= 1.5 gini = 0.033 samples = 118 value = [1, 116, 1]
X[3] <= 7.15 gini = 0.375 samples = 4 value = [0, 3, 1]
gini = 0.375 samples = 4 value = [0, 3, 1]
X[1] <= 38.5 gini = 0.169 samples = 43 value = [4, 39, 0]
gini = 0.0 samples = 5 value = [0, 5, 0]
X[9] <= 80.5 gini = 0.42 samples = 10 value = [3, 0, 7]
gini = 0.0 samples = 1 value = [0, 1, 0]
X[1] <= 44.5 gini = 0.132 samples = 57 value = [53, 3, 1]
X[3] <= 6.05 gini = 0.156 samples = 24 value = [1, 1, 22]
X[4] <= 7.5 gini = 0.238 samples = 29 value = [0, 4, 25]
X[1] <= 36.5 gini = 0.087 samples = 22 value = [1, 21, 0]
X[5] <= 72.5 gini = 0.021 samples = 96 value = [0, 95, 1]
gini = 0.0 samples = 1 value = [0, 1, 0]
gini = 0.444 samples = 3 value = [0, 2, 1]
gini = 0.0 samples = 8 value = [0, 8, 0]
X[2] <= 3.5 gini = 0.202 samples = 35 value = [4, 31, 0]
X[2] <= 9.5 gini = 0.5 samples = 6 value = [3, 0, 3]
gini = 0.0 samples = 4 value = [0, 0, 4]
X[4] <= 6.5 gini = 0.101 samples = 38 value = [36, 1, 1]
X[6] <= 4.5 gini = 0.188 samples = 19 value = [17, 2, 0]
X[3] <= 5.95 gini = 0.153 samples = 12 value = [0, 1, 11]
X[1] <= 49.5 gini = 0.153 samples = 12 value = [1, 0, 11]
gini = 0.5 samples = 2 value = [0, 1, 1]
X[3] <= 8.15 gini = 0.198 samples = 27 value = [0, 3, 24]
X[1] <= 35.5 gini = 0.375 samples = 4 value = [1, 3, 0]
gini = 0.0 samples = 18 value = [0, 18, 0]
gini = 0.0 samples = 63 value = [0, 63, 0]
X[3] <= 7.75 gini = 0.059 samples = 33 value = [0, 32, 1]
X[3] <= 7.75 gini = 0.26 samples = 13 value = [2, 11, 0]
X[3] <= 7.65 gini = 0.165 samples = 22 value = [2, 20, 0]
gini = 0.0 samples = 2 value = [2, 0, 0]
X[0] <= 1.5 gini = 0.375 samples = 4 value = [1, 0, 3]
X[3] <= 6.35 gini = 0.15 samples = 25 value = [23, 1, 1]
gini = 0.0 samples = 13 value = [13, 0, 0]
X[3] <= 6.55 gini = 0.32 samples = 10 value = [8, 2, 0]
gini = 0.0 samples = 9 value = [9, 0, 0]
gini = 0.0 samples = 1 value = [0, 0, 1]
X[1] <= 50.5 gini = 0.165 samples = 11 value = [0, 1, 10]
gini = 0.0 samples = 4 value = [0, 0, 4]
X[1] <= 50.5 gini = 0.219 samples = 8 value = [1, 0, 7]
X[1] <= 58.5 gini = 0.245 samples = 21 value = [0, 3, 18]
gini = 0.0 samples = 6 value = [0, 0, 6]
gini = 0.0 samples = 1 value = [0, 1, 0]
X[3] <= 7.15 gini = 0.444 samples = 3 value = [1, 2, 0]
X[3] <= 7.65 gini = 0.105 samples = 18 value = [0, 17, 1]
gini = 0.0 samples = 15 value = [0, 15, 0]
gini = 0.0 samples = 4 value = [0, 4, 0]
gini = 0.346 samples = 9 value = [2, 7, 0]
X[5] <= 75.0 gini = 0.26 samples = 13 value = [2, 11, 0]
gini = 0.0 samples = 9 value = [0, 9, 0]
gini = 0.0 samples = 2 value = [0, 0, 2]
gini = 0.5 samples = 2 value = [1, 0, 1]
gini = 0.0 samples = 10 value = [10, 0, 0]
X[3] <= 6.45 gini = 0.24 samples = 15 value = [13, 1, 1]
gini = 0.0 samples = 4 value = [4, 0, 0]
gini = 0.444 samples = 6 value = [4, 2, 0]
gini = 0.18 samples = 10 value = [0, 1, 9]
gini = 0.0 samples = 1 value = [0, 0, 1]
gini = 0.245 samples = 7 value = [1, 0, 6]
gini = 0.0 samples = 1 value = [0, 0, 1]
X[1] <= 57.5 gini = 0.153 samples = 12 value = [0, 1, 11]
X[3] <= 8.05 gini = 0.346 samples = 9 value = [0, 2, 7]
gini = 0.0 samples = 1 value = [0, 1, 0]
gini = 0.5 samples = 2 value = [1, 1, 0]
gini = 0.0 samples = 5 value = [0, 5, 0]
X[1] <= 30.5 gini = 0.142 samples = 13 value = [0, 12, 1]
gini = 0.18 samples = 10 value = [1, 9, 0]
gini = 0.444 samples = 3 value = [1, 2, 0]
gini = 0.32 samples = 5 value = [4, 1, 0]
gini = 0.18 samples = 10 value = [9, 0, 1]
X[10] <= 6000.0 gini = 0.278 samples = 6 value = [0, 1, 5]
gini = 0.0 samples = 6 value = [0, 0, 6]
gini = 0.375 samples = 4 value = [0, 1, 3]
gini = 0.32 samples = 5 value = [0, 1, 4]
gini = 0.0 samples = 4 value = [0, 4, 0]
X[1] <= 31.5 gini = 0.198 samples = 9 value = [0, 8, 1]
gini = 0.0 samples = 3 value = [0, 0, 3]
gini = 0.444 samples = 3 value = [0, 1, 2]
gini = 0.245 samples = 7 value = [0, 6, 1]
gini = 0.0 samples = 2 value = [0, 2, 0]

Figure 7: complete decision tree structure diagram