Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 3 No. 2 (2023) 71 - 76 E-ISSN: 2809-8064

Classification of Sleep Disorders Using Random Forest on SleepHealth and Lifestyle Dataset

Idfian Azhar Hidayat[1*]

[1*]Department of Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

[1*]20104031@ittelkom-pwt.ac.id

Abstract

This study aims to classify sleep disorders using the Random Forest method on the Sleep Health and Lifestyle dataset. This dataset contains information about sleep, lifestyle, and relevant health factors. In this study, the dataset was processed and divided into training and testing subsets. The Random Forest model was trained using the training subset with sleep and health related features. The quality of the split in each decision tree was measured using the Gini Index. The model was evaluated using the testing subset to measure its accuracy and classification performance. The evaluation results showed that the Random Forest model was able to predict sleep disorders with good accuracy. Analysis of class distributions, correlation relationships between features, and visualization by gender provided insights into the factors that influence sleep disorders. This research has thepotential to contribute to the field of health and medicine, especially in the recognition and diagnosis of sleep disorders.

Keywords: Sleep Disorders, Classification, Random Forest, Sleep Health and Lifestyle, Machine Learning

1. Introduction

Sleep disorders are common health problems that can significantly affect a person's quality of life. Accurate identification and classification of sleep disorders is essential for proper diagnosis and provision of appropriate treatment. In this study, a classification of sleep disorders using Random Forest model is performed based on features related to sleep, lifestyle, and other health factors contained in the Sleep Health and Lifestyle dataset.

The Sleep Health and Lifestyle dataset used in this study was obtained from the Kaggle website. This dataset contains information on a few sleep-related features, lifestyle, and health factors of several respondents. It has information that includes individual ID, gender, age, occupation, sleep duration, sleep quality, physical activity level, stress level, BMI category, blood pressure, heart rate, and daily steps.

Random Forest is one of the algorithms in machine learning that is widely used for classification and regression tasks. It uses the concept of ensemble. learning, which consists of many decision trees that work independently and then combine their prediction results to produce a final prediction [1].

Random Forest was chosen for this research based on several reasons. First, this algorithm can overcome the overfitting problem by using bootstrapping techniques and random restrictions on feature selection [2]. This helps reduce the risk of overfitting the model. Secondly, Random Forest has high accuracy in making predictions [3]. In the context of this research, by using features related to sleep, lifestyle, and health factors, Random Forest can learn the patterns in the dataset and produce accurate predictions related to sleep disorders. Thirdly, the algorithm is robust to unbalanced data, which is common in datasets where the class distribution may be unbalanced [4]. Random Forest can handle this problem well because each tree in the ensemble is trained using a random subset of the data, including a subset of the minority classes [5].

This helps in selecting relevant features and understanding the contribution of each feature to the sleep disorder prediction results. Considering Random Forest's ability to classify, ability to overcome overfitting, high accuracy, and ability to evaluate the importance of features, this algorithm was chosen as an appropriate choice for this study.

Received: 05-08-2023 | Revised: 07-08-2023 | Published: 10-08-2023

The purpose of this research is to develop a classification model that can predict sleep disorders based on the features in the dataset. The model used is Random Forest, which is one of the algorithms in machine learning that is effective for classification tasks. By using this model, it is expected that accurate predictions can be obtained and can help in the identification of sleep disorders in individuals.

The method used in this research consists of several steps. First, the Sleep Health and Lifestyle dataset was explored and understood. Information about the distribution of features, relationships between features, and the distribution of sleep disorder classes was analyzed and visualized. Then, data processing was carried out by selecting appropriate features and converting categorical variables into numerical using the one-hot encoding technique.

Next, the dataset is divided into training data and testing data using the train_test_split method. The Random Forest model is then built by setting hyperparameters such as the number of trees (n_estimators) and the maximum depth of trees (max_depth). The model is trained using the training data and then used to predict labels using the testing data.

Model evaluation is done by calculating model accuracy using testing data. In addition, Confusion Matrix calculations are also carried out to evaluate the performance of the model in predicting sleep disorder classes. The results of this evaluation will provide information on the extent to which the developed model can classify sleep disorders accurately.

This research has the potential to contribute to the field of health and medicine, especially in the recognitionand classification of sleep disorders. By using a Random Forest model based on relevant features, this research is expected to be the basis for further development of classification methods in predicting sleep disorders. The results of this study are expected tobe used as a reference in more effective and targeted diagnosis and treatment related to sleep disorders.

2. Research Methods

This study uses the Random Forest method to classify sleep disorders based on the Sleep Health and Lifestyle dataset. This dataset contains information on individual ID, gender, age, occupation, sleep duration, sleep quality, physical activity level, stress level, BMI category, blood pressure, heart rate, and daily steps. In this study, the dataset was processed and divided into training and testing subsets.

The Random Forest model was trained using a training subset consisting of sleep and health-related features such as age, gender, sleep duration, sleep quality, physical activity level, stress level, BMI category, blood pressure, heart rate, and daily step count. The model uses the Random Forest algorithm which is an ensemble learning of decision trees.

In each decision tree in Random Forest, the split quality is measured using the Gini Index. Gini Index measures the level of impurity or inhomogeneity of a node in the dataset. Gini Index values range between 0 and 1, where 0 indicates a homogeneous node (all samples have the same label), and 1 indicates a heterogeneous node (samples have different labels equally).

A Random Forest model with 200 trees (`n_estimators=200`) and a maximum tree depth of 10 (`max_depth=10`) is built using the Gini Index algorithm. The model was trained using subset training to learn the patterns and relationships between the features present in the dataset.

After training the model, the test subset is used to test the performance of the model. The accuracy of the model is calculated using the `score()` method on the model object using the test data. The model's predictionresults are also evaluated using the classification report which provides evaluation metrics such as precision, recall, and f1-score. Confusion matrix is also calculatedto see the distribution of correct and incorrect predictions.

In this study, the main objective of this research is to develop a classification model that can predict sleep disorders based on the features present in the dataset.

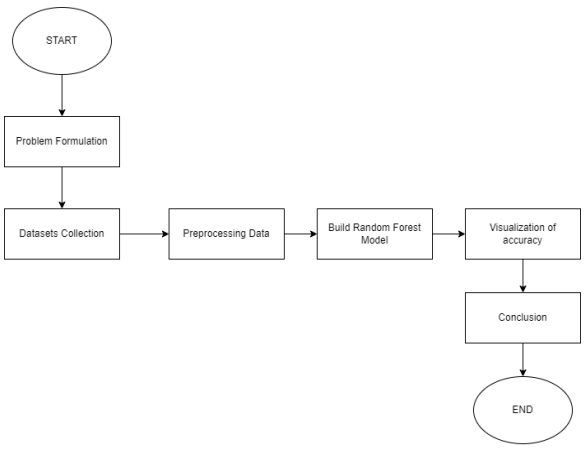

Figure 1. Research Flow

2.1. *Identifying the Problem*

The problem of this study refers to sleep disorders based on the Sleep Health and Lifestyle dataset. Sleep disorders, also known as sleep disorders, refer to a variety of conditions that affect the quality, pattern, and duration of a person's sleep. Sleep disorders can affect adequate sleep, quality sleep, or both. Sleep disorders can have a negative impact on a person's quality of life, physical, cognitive, and emotional health. Based on this

problem, this research develops a method to diagnose and predict sleep disorders.

2.2. *Dataset Collection*

In this study, the classification of sleep disorders using the Random Forest model is carried out based on features related to sleep, lifestyle, and other health factors contained in the Sleep Health and Lifestyle dataset. The Sleep Health and Lifestyle dataset used in this study was obtained from the Kaggle website. This dataset contains information on a few sleep-related features, lifestyle, and health factors of several respondents. It has information that includes individual ID, gender, age, occupation, sleep duration, sleep quality, physical activity level, stress level, BMI category, blood pressure, heart rate, and daily steps.

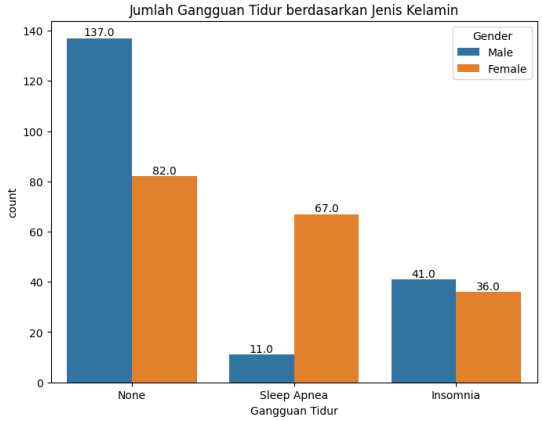

Figure 2. Number of Sleep Disorders by Gender

2.3. *Preprocessing Data*

The preprocessing stage before Random Forest model. building involves steps such as reading the dataset, selecting relevant features, data processing such as one-hot encoding, dividing the dataset into training and testing data, handling missing values, standardizing, or normalizing data, handling unbalanced data, and verifying the dataset [6], [7]. In addition, in the data splitting stage, the Gini Index method is used to determine the optimal split at each node in the decision tree. The GiniIndex helps select the split that optimizes the splitting.
of the target class, resulting in an accurate and efficient decision tree in the Random Forest model.

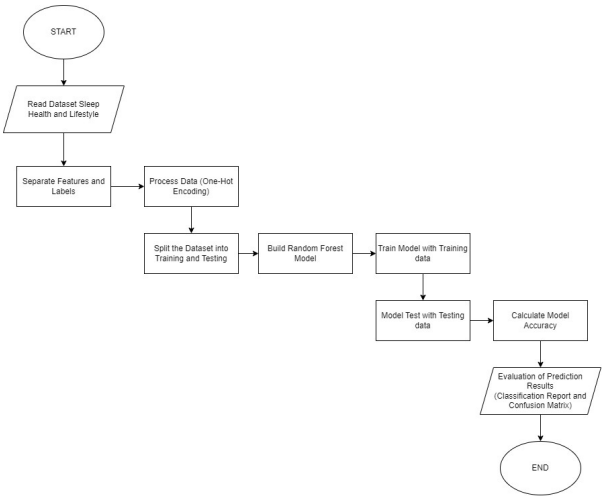

Figure 3. Processing Flow

Gini Index is one of the methods used in Random Forest algorithm to measure the level of impurity or inhomogeneity of a node in the dataset. The Gini Index value ranges between 0 and 1, where 0 indicates a homogeneous node (all samples have the same label), and 1 indicates a heterogeneous node (samples have different labels evenly) [8]. Mathematically Gini indexcan be written as:

$$Gini\ Index = 1 - \sum_{i=1}^{n} \left(P_i\right)^2$$

$$= 1 - [(P_+)^2 + (P_-)^2]$$

Figure 4. Gini Index Formula

Where P is the probability of a positive class and P_ is the probability of a negative class. In this study, the Random Forest algorithm uses the specified hyperparameters (`n_estimators=200` and `max_depth=10`) with the Gini Index as a method to measure the split quality when building the decisiontree in Random Forest

The following is a pseudocode that explains the program steps performed:

Table 1. Pseudocode Program

```
Program Pseudocode
INPUT:path dataset
(Sleep_health_and_lifestyle_dataset.csv)
  IMPORT pandas, matplotlib.pyplot as plt,
  seaborn as sns, RandomForestClassifier from
  sklearn.ensemble, train_test_split from
  sklearn.model_selection,
  classification_report, confusion_matrix from
  sklearn.metrics
```

```
dataset <- pd.read_csv(path_dataset)
DISPLAY dataset.head()
PRINT dataset.info()
X <- SELECT_FEATURES(dataset)
y <- SELECT_LABEL(dataset)
DISPLAY COUNT_PLOT(dataset)
DISPLAY HEATMAPT(dataset)
DISPLAY COUNT_PLOT(dataset)
X <- ONE_HOT_ENCODING(X)
X_train, X_test, y_train, y_test <- SPLIT_DATASET(X, y)
model <- RANDOM_FOREST_MODEL()
TRAIN_MODEL(model, X_train, y_train)
y_prediksi <- PREDICT_LABEL(model, X_test)
akurasi <- CALCULATE_ACCURACY(y_test, y_prediksi)
PRINT "Akurasi Random Forest (Testing):", akurasi
PRINT CLASSIFICATION_REPORT(y_test, y_prediksi)
cm <- CALCULATE_CONFUSION_MATRIX(y_test, y_prediksi)
PRINT "Confusion Matrix:"
PRINT cm
DISPLAY BAR_PLOT(akurasi)
DISPLAY HEATMAP(cm)
END
```

2.4. *Build Random Forest Model*

After preprocessing the data, the Random Forest model was created. This model uses a combination of multiple decision trees to make predictions. The steps include determining the number of trees, building decision trees with random subsets of training data, making predictions by combining the prediction results from each tree, evaluating the model with appropriate evaluation metrics, and saving the model for future use [9], [10]. Random Forest can overcome overfitting, provide stable and accurate predictions, and provide information about the importance of features in prediction.

In this program, there are several types of visualizations used to analyze and understand the dataset and classification results using the Random Forest model, namely: Countplot, used to display the class distribution on the sleep disorder dataset. This helps in understanding the distribution of the number of samplesin each sleep disorder class. Heatmap, used to display the correlation between features in the dataset. Heatmap provides a matrix visualization with different colors to illustrate the level of correlation between features. This helps to see the relationship between features and can provide insight into potentially influential features in classification. Countplot, used to display the number of sleep disorders by gender. This countplot provides information about the distribution of the number of samples in each sleep disorder class by gender. Bar plot, used to display the accuracy of the Random Forest model. This bar plot provides a visualization of the model accuracy in bar form. This helps to understand the extent to which the model performs well. Heatmap, used to display the Confusion Matrix. Confusion matrix provides information about the number of correct and incorrect predictions for each class.

3. Results and Discussion

Based on research using the Random Forest method on the Sleep Health and Lifestyle dataset, it can be concluded that the Random Forest classification model is able to predict sleep disorders with fairly good accuracy. In this study, the Sleep Health and Lifestyle dataset is used as a data source containing information about sleep, and other health factors contained in the dataset. This dataset is processed by separating sleep and health related features as input features (X) and sleep disorders as labels (y). The Random Forest modelwas built using 200 trees and a maximum depth of 10. The Gini Index algorithm was used to measure the split quality within each decision tree. The model was trained using the training subset and evaluated usingthe testing subset. The model evaluation results showed that the Random Forest model achieved a good level of accuracy on the test data.

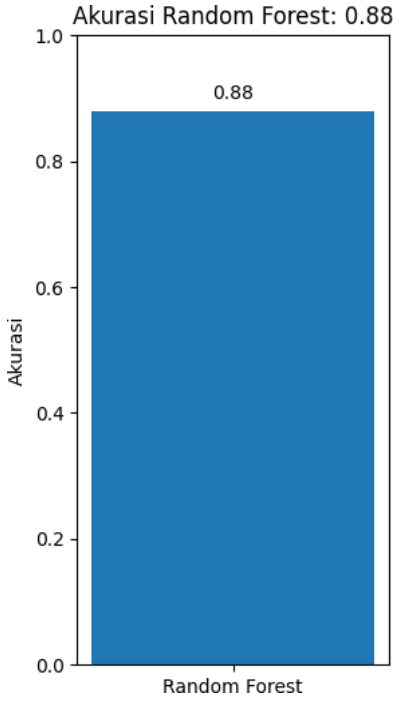

Figure 5. Random Forest Accuracy

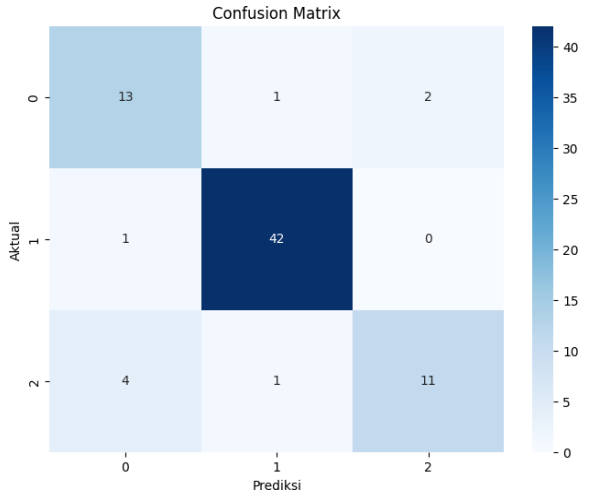

Figure 6. Confusion Matrix

In addition, the classification report that includes precision, recall, and f1-score provides a more detailed picture of the model's performance in predicting sleep disorders.

Table 2. Classification Report

	precision	recall	f1-score	support
Insomnia	0.72	0.81	0.76	16
None	0.95	0.98	0.97	43
Sleep Apnea	0.85	0.69	0.76	16
accuracy			0.88	75
macro avg	0.84	0.83	0.83	75
weighted avg	0.88	0.88	0.88	75

From the visualization of the class distribution, it can be seen that the dataset has a different number of samples for each sleep disorder class. However, the model is able to overcome this imbalance and provide good results for each class.

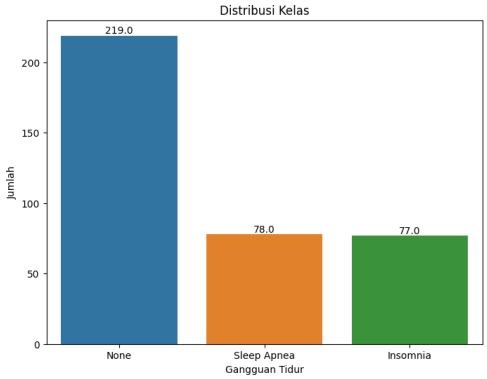

Figure 7. Class Distribution

In addition, correlation analysis of the relationship between features and visualization of the number of sleep disorders by gender also provide insights into factors that may influence sleep disorders.

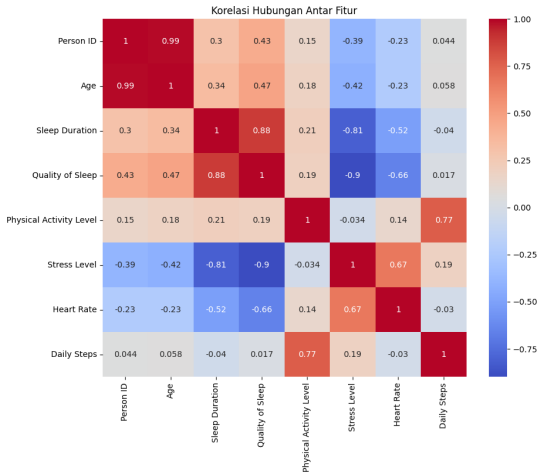

Figure 8. Correlation Between Features

Thus, this study shows that the use of the Random Forest method on the Sleep Health and Lifestyle datasetcan be an effective model for classifying sleep disorders. The results of this study are expected to havethe potential to contribute to the field of health and medicine, especially in the recognition and classification of sleep disorders.

4. Conclusion

This study uses the Random Forest model to classify sleep disorders based on sleep-related features, lifestyle, and health factors in the Sleep Health and Lifestyle dataset. This model is able to overcome overfitting, has high accuracy, is resistant to unbalanced data, and provides information on the importance of each feature in prediction. The evaluation results show that the model is effective in predicting sleep disorders with good accuracy. The results of this study can be used in more effective diagnosis and treatment of sleep disorders.

References

[1] D. Zhuang, I. Rao, Ali K Ibrahim, A Machine Learning Approach to Automatic Classification of Eight Sleep Disorders. Arxiv, 14 April 2022.
[2] Yoga Religia, Agung Nugroho, Wahyu Hadi Kristanto, "Analisis Perbandingan Algoritma Optimasi pada Random Forest Untuk Klasifikasi Data Bank Marketing" Jurnal Resti, Vol. 5, No. 1, pp. 189, 2021.
[3] V. Wanika Siburian, I. Elvina Mulyana, "Prediksi Harga Ponsel Menggunakan Metode Random Forest" Annual Research Seminar, Vol.4, No.1, pp. 144, 2018.

[4] A. Ferdita Nugraha, R. Faticha Alfa Aziza, Y. Pristyanto, "Penerapan metode Stacking dan Random Forest untuk Meningkatkan Kinerja Klasifikasi pada Proses Deteksi Web Phishing" Vol.7, No. 1, pp. 44, 2022.

[5] F. Nugroho Yudianto, Y. Sibaroni, "Klasifikasi Hashtag Buzzer/Bot Menggunakan Algoritma Random Forest dengan Atribut Komunitas untuk Mengurangi Disinformasi Pada Twitter" Vol, 9, No.3, pp.1896, Juni 2022.

[6] Ramadhan, Nur Ghaniaviyanto. "Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus." *Sci. J. Informatics,* vol. 8, no. 2, pp. 276-282, 2021.

[7] Putri, Tiara Khumaira, et al. "Diabetes Diagnostic Expert System using Website-Based Forward Chaining Method." *Journal of Dinda: Data Science, Information Technology, and Data Analytics,* vol. 3, no. 1, pp. 11-17, 2023.

[8] Ebrahem A. Algehyne, Muhammad Lawan Jibril, Naseh A. Algehainy, Osama Abdulaziz Alamri, Abdullah K. Alzahrani, "Fuzzy Neural Network Expert System with an Improved Gini Index Random Forest-Based Feature Importance Measure Algorithm for Early Diagnosis of Breast Cancer in Saudi Arabia" Big Data and Cognitive Computing, Vol 6, No. 1, pp. 5-6, 2022.

[9] Sharfina, Nabilah, and Nur Ghaniaviyanto Ramadhan. "Analisis SMOTE Pada Klasifikasi Hepatitis C Berbasis Random Forest dan Naïve Bayes." *JOINTECS (Journal of Information Technology and Computer Science)*, vol. 8, no. 1, pp. 33-40, 2023.

[10] Ramadhan, Nur Ghaniaviyanto, et al. "Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression." *JURIKOM (Jurnal Riset Komputer),* vol. 9, no. 2, pp. 251-256, 2022.