Predicting Stress Levels Based on Sleep and Lifestyle Habits

Author1:Aslıhan Akan
Author2:Gülçin Yüzgüleç
Author3:Esra Bayrak

December 17, 2025

Abstract

In our daily lives, stress significantly impacts everyone's lives and leads to problems; therefore, developing systems that predict stress levels based on our habits is an attractive topic for software developers. The aim of this study is to predict stress levels by examining sleep patterns and lifestyle habits. The dataset used includes variables such as sleep duration, sleep quality, physical activity level, occupation, body mass index (BMI), and sleep disorders. Within the scope of the study, the dataset was prepared for analysis; data preprocessing steps such as completing missing data, coding necessary variables, and normalization were applied. Exploratory data analysis was performed to examine the relationships between sleep and lifestyle variables and stress levels. Subsequently, machine learning methods were applied to predict stress levels using the adjusted data. The results show that sleep duration, sleep quality, and lifestyle-related variables play a significant role in predicting stress levels. The results demonstrate that healthy sleep habits and a regular lifestyle are crucial in reducing stress.

Keywords:stress level,life style,applied informatics,machine learning

Arel Üniversitesi, Müh. Mim Fak. Bilgisayar Mühendisliği. Türkoba Mahallesi, Erguvan Sokak, No:26, K 34537, Tepekent – Büyükçekmece, İstanbul

Öz

Güncel yaşantımızda stres hepimizin hayatını yoğun şekilde etkileyip yaşantımızda sorunlara yol açıyor dolayısıyla alışkanlıklarımıza bağlı stres oranlarını tahmin eden sistemler geliştirmek yazılımcılar için dikkat çekici bir konudur.Bu çalışmanın amacı, uyku düzeni ve yaşam tarzı alışkanlıklarını inceleyerek stres seviyelerini tahmin etmektir. Kullanılan veri seti; uyku süresi, uyku kalitesi, fiziksel aktivite düzeyi, meslek, vücut kitle indeksi (BMI) ve uyku bozuklukları gibi değişkenleri içermektedir. Çalışma kapsamında veri seti analiz için uygun hale getirilmiş; eksik verilerin tamamlanması,gerekli değişkenlerin kodlanması ve normalizasyon gibi veri ön işleme adımları uygulanmıştır. Uyku ve yaşam tarzı değişkenleri ile stres seviyeleri arasındaki ilişkileri incelemek amacıyla keşifsel veri analizi gerçekleştirilmiştir. Sonrasında, düzenlenmiş veriler kullanılarak stres seviyelerinin tahmini için makine öğrenmesi yöntemleri uygulanmıştır. Elde edilen sonuçlar, uyku süresi, uyku kalitesi ve yaşam tarzına bağlı değişkenlerin stres seviyesinin tahmininde önemli rol oynadığını göstermektedir. Sonuçlar, sağlıklı uyku alışkanlıkları ve düzenli bir yaşam tarzının stresin azaltılmasında oldukça önemli olduğunu göstermektedir.regular lifestyle are crucial in reducing stress.

Anahtar Kelimeler: stres seviyesi, yaşam tarzı, uygulamalı bilişim, makine öğrenmesi

1 Introduction

Stress is one of the most important factors affecting physical and mental health in our current living conditions. Intense work schedules, irregular sleep, and unhealthy lifestyles cause stress levels to rise in people. Long-term stress can lead to serious health problems such as anxiety, depression, and cardiovascular diseases. Therefore, early detection of high stress levels is important for both health and technology-focused research. Thanks to advancements in applied computing, data-driven studies and machine learning methods have become the preferred methods for examining complex health problems. Analyzing large datasets makes it possible to predict certain conditions by looking at individuals' daily habits. In particular, sleep duration, sleep quality, physical activity, and lifestyle factors are considered important elements in stress analysis. Resources have shown a strong relationship between sleep habits and stress levels. However, traditional assessment methods are mostly based on surveys and subjective evaluations. In contrast, machine learning systems offer more objective solutions by utilizing real data. The aim of this study is to predict stress levels based on sleep patterns and lifestyle habits using a synthetic dataset, employing data analysis and machine learning techniques. The benefits of applied computing in stress management and health-focused decision support systems are demonstrated

Arel Üniversitesi, Müh. Mim Fak. Bilgisayar Mühendisliği. Türkoba Mahallesi, Erguvan Sokak, No:26, K 34537, Tepekent – Büyükçekmece, İstanbul

through data preprocessing, data analysis, and prediction models. Such studies show that stress is not only a subject in the health field but can also be a subject in many different areas, such as software. Currently, it is possible to make meaningful inferences from people's daily habits using data analysis and machine learning. Analyzing large datasets, in particular, allows for more accurate and faster prediction of stress levels. The results obtained in this study clearly demonstrate the effect of sleep patterns and lifestyle habits on stress levels. The results of the study can contribute to software systems in the field of stress management. Furthermore, such systems can help individuals evaluate their own habits and lead healthier lives.

2 Procedure

In this study, we attempted to follow a structured process for estimating stress levels.

First, after conducting a general search to find the most suitable dataset for our project topic, a synthetic dataset containing sleep health and lifestyle habits was used. The content of the dataset was examined, and stress level was determined as the target variable. Then, to avoid problems during the model training process, the dataset was analyzed, and missing values and inconsistencies were corrected.

In the second step, data preprocessing stages were applied. In this section, missing data were completed, encoding was performed to make categorical variables numerical, and normalization was applied to reduce scale differences between variables. This step aims to prepare the data for modeling.

In the third step, exploratory data analysis was performed. The relationships between stress level and variables such as sleep duration, sleep quality, occupation, and gender were examined by visualizing them with graphs.

In the final step, our dataset, which was organized by applying the described steps, was used to create machine learning models.

3 Problem Definition

Stress is a complex emotion that negatively affects people's physical and mental health and is influenced by many variables, such as sleep patterns and lifestyle habits. Past methods frequently used to assess stress levels have generally relied on questionnaires and subjective evaluations. This makes it difficult to accurately examine stress levels. The main problem examined in our study is the difficulty of accurately predicting individuals' stress levels based on their daily life habits. Although variables such as sleep duration, sleep quality, physical activity, and lifestyle are known to be related to stress, it is not always clear how these variables affect each other. Therefore, the problem definition of our study is to develop a data-driven approach that can predict stress levels using sleep patterns and lifestyle variables.

3.1 Data

Our study utilized a synthetic, publicly available dataset containing sleep health and lifestyle habits. The dataset included various variables such as gender, age, occupation, sleep duration, sleep quality, physical activity level, BMI, blood pressure, heart rate, daily step count, and sleep disorders. Stress level was identified as the target variable to be predicted in the study.

3.2 Evaluation

The machine learning models we created were evaluated by comparing the predicted values with actual stress levels. Our results show that sleep and lifestyle variables are effective in predicting stress levels.

4 Data and Methodology

4.1 Description and Analysis of the Dataset

In this study, a publicly available synthetic dataset regarding sleep health and lifestyle habits was used to estimate stress levels. The dataset includes information obtained from individuals with different characteristics. The dataset contains variables representing both sleep-related characteristics and daily life habits. The dataset consists

of variables such as Person ID, Gender, Age, Occupation, Sleep Duration, Quality of Sleep, Physical Activity Level, Stress Level, BMI Category, Blood Pressure, Heart Rate, Daily Steps, and Sleep Disorder. Stress level was considered the target to be estimated. These variables are considered important values for stress analysis because they are directly related to physical and mental health. Before the modeling phase, preliminary analyses were performed to better understand the structure of the dataset. In this context, basic examinations were carried out, the distributions of the variables were observed, and missing data were identified. In addition, the relationships between sleep habits, lifestyle factors, and stress levels were examined to create a general overview of the dataset. These analyses contributed to generating ideas for future methods. For example, the graph below, obtained from our data analysis, shows the distribution of stress levels according to different occupational groups. Looking at the graph, it is clear that there are significant differences in stress levels between occupations. It is noteworthy that stress levels are higher and more variable in some occupational groups, while in others they are clustered within a more stable range. This information shows that occupational characteristics have an effect on stress. However, it also shows that stress should be evaluated not only in relation to occupation but also in conjunction with other variables such as sleep patterns, lifestyle, and personal habits. Therefore, the occupational variable was considered as a helpful factor contributing to stress prediction in this study.

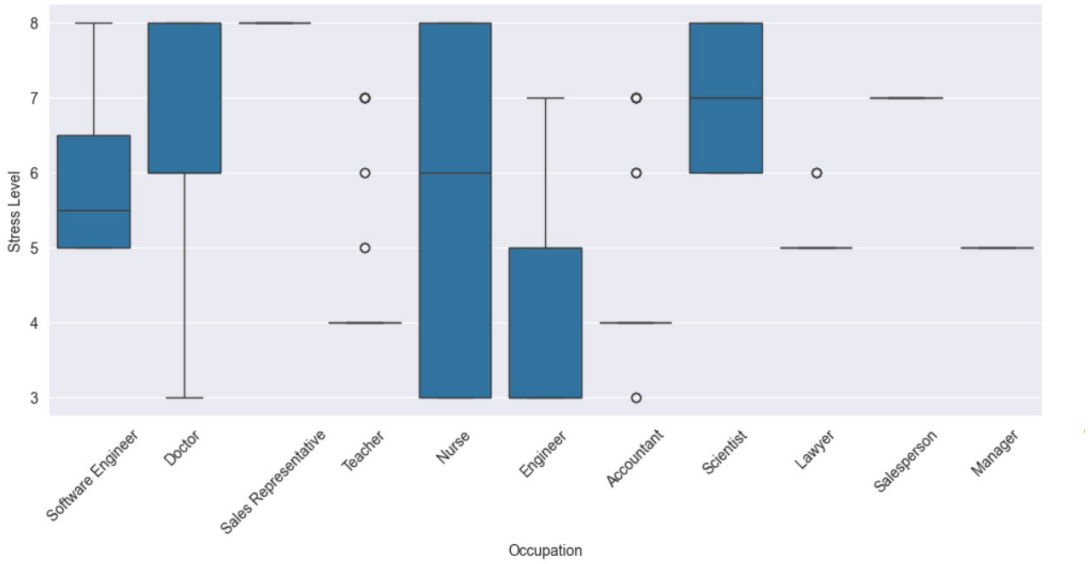

Figure 1: Distribution of stress levels according to occupational groups.

As mentioned above, there are many variables related to stress levels; to give another example, even our sleep duration is closely related to our stress levels. The graph shows the relationship between stress level and sleep duration. Looking at the graph, it can be seen that as stress levels increase, sleep duration tends to decrease. People with low stress levels tend to have longer and more regular sleep durations, while those

with high stress levels experience shorter, more concentrated sleep intervals. This suggests that stress can have a negative impact on sleep duration. These findings, when considered alongside various other variables such as the occupation studied, reveal that stress has a multifaceted structure. Observing varying sleep durations at the same stress level demonstrates that stress is related to different factors such as lifestyle diversity and individual characteristics. Using data analysis, a model was developed by examining the effects of these different characteristics. The results obtained during the study show that sleep plays a significant role in decision support systems developed for stress management. In summary, data on sleep duration can be said to have an important role in predicting stress levels. The fact that individuals with regular and sufficient sleep are less stressed indicates that healthy sleep plays a significant role in stress management. Including the sleep duration variable in the model in this study contributed to achieving more accurate results in stress-related research.Thus, examining various variables was beneficial in determining whether some were more or less related to stress, and it contributed to our ability to develop new ideas for model improvement.

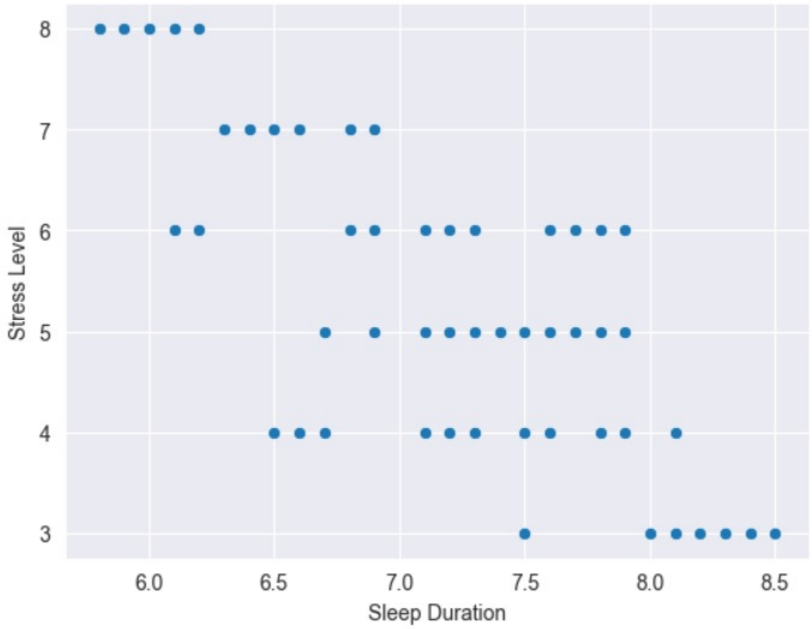

Figure 2: The relationship between stress levels and sleep duration.

5 Model Train-Test Split

In the study, "Stress Level" was defined as the target variable in the dataset. This variable represents the stress levels of individuals. This column, used as the dependent variable in the modeling process, was separated from the dataset to create a label vector (y). All other variables were defined as an independent variables matrix (X). This approach is used to enable machine learning models to more clearly understand the relationship between input data and the target variable. The dataset was divided into two separate subsets: training and test data, to objectively evaluate model performance. In this process, 80 percent of the data was allocated for training and 20 percent for testing. Stratified sampling was employed to maintain a balanced class distribution between the training and test sets, minimizing the potential negative impact of target variable imbalances on model performance. Additionally, the reproducibility of the experiments was ensured by setting the parameter random state=42. Many machine learning algorithms are sensitive to feature scaling. Therefore, feature scaling was applied to the dataset before the modeling process. In this context, the StandardScaler method was used to normalize each feature so that its mean is zero and its standard deviation is one. The scaling process was only learned on the training data, and the same transformation was applied to the test data to prevent data leakage.

6 Machine Learning Models

Four different machine learning algorithms were used for the stress level prediction problem. The aim in selecting the models was to compare the performance of both linear and nonlinear approaches.

6.1 Logistic Regression Model

First, the Logistic Regression model was applied. Logistic Regression is a fundamental algorithm frequently used in classification problems that establishes linear decision bounds. In this study, the maximum number of iterations was set to 1000 to prevent the model from experiencing convergence problems. The Logistic Regression model was considered as a baseline model for comparison with other models.

6.2 Support Vector Machine (SVM) Model

The Support Vector Machine (SVM) algorithm was used as the second model. The SVM model aims to find the decision boundary (hyperplane) that best distinguishes between classes. In this study, a linear kernel was preferred. The regularization parameter was set to C=1. The SVM algorithm yields effective results, especially in high-dimensional datasets.

6.3 K-Nearest Neighbors (KNN) Model

The K-Nearest Neighbors (KNN) algorithm was applied as the third model. The KNN algorithm is a sample-based method that does not involve any learning process. It makes predictions by looking at the classes of the nearest neighbors of the test data. In this study, the number of neighbors was determined as k=5, and the prediction performance of the model was evaluated on the test data.

6.4 Random Forest Classifier Model

Finally, the Random Forest Classifier model was used. Random Forest is one of the ensemble learning methods created by combining multiple decision trees. In this study, 100 decision trees were used. To prevent overfitting, the tree depth was limited to max depth=5. The Random Forest model was evaluated as a powerful classifier due to its ability to detect nonlinear relationships.

6.5 K-Fold Cross Validation

To evaluate whether the selected Random Forest model exhibits consistent performance not only on the test data but also on the overall dataset, a 5-fold K-Fold cross-validation method was applied. In this method, the dataset was divided into five equal parts. Each part was used sequentially as test data, and model training was repeated.

6.6 Feature Importance Analysis

One of the significant advantages offered by the Random Forest model is its ability to calculate feature importance values, which measure the contribution of each feature to model decisions. In this context, the relative importance of features in predicting stress levels in the dataset was calculated using the trained Random Forest model. Examination of the results revealed that some features played a more dominant role than others in determining stress levels. Feature importance values were visualized through a graph, allowing for a clearer analysis of the key factors influencing stress levels. This analysis can also contribute to future model improvement and feature selection studies. When the results obtained for our model are examined, it is seen that the variables Sleep Duration, Quality of Sleep, and Heart Rate have higher significance values compared to other characteristics and play a more dominant role in determining the stress level. This situation reveals that the stress level is more strongly influenced by certain factors in our dataset.

7 Results and Discussion

7.1 Model Performance Comparison

The performance of Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest models was evaluated using both accuracy and weighted F1-score metrics on a consistent test dataset. The results indicate that the Random Forest and SVM models achieved the highest accuracy scores among the evaluated algorithms. SVM and Logistic Regression demonstrated comparable performance across the assessed metrics. In contrast, the KNN and Logistic Regression models exhibited lower accuracy compared to Random Forest and SVM. Based on its ensemble nature and consistently robust performance, the Random Forest model was selected for use in subsequent stages of analysis. This version streamlines the comparisons, clarifies the results, and explains the rationale for selecting the Random Forest model.

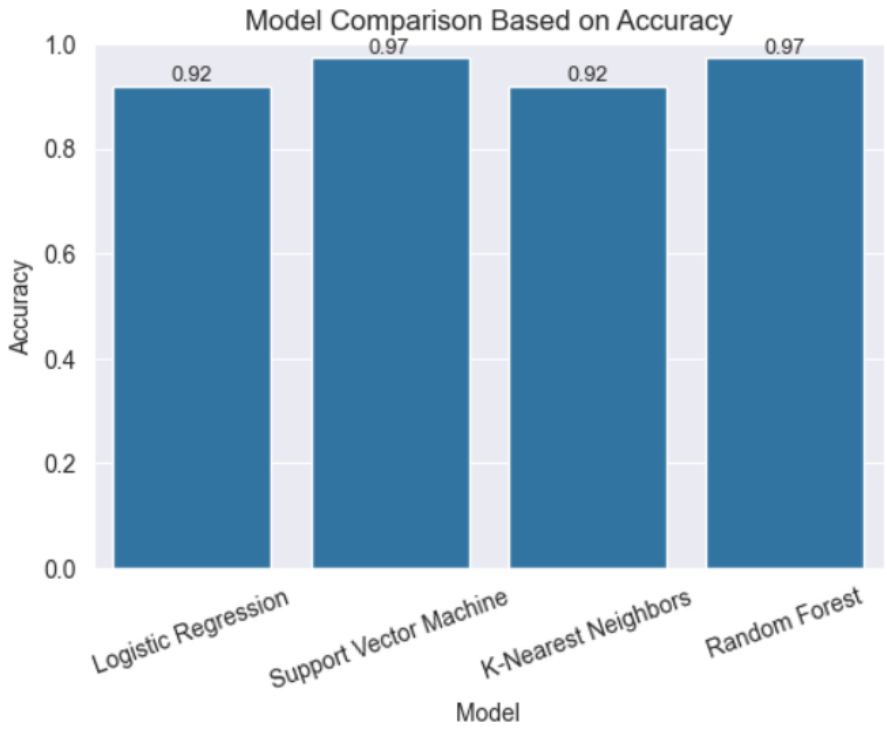

Figure 3: Comparison of the accuracy values of models used for stress level estimation.

7.2 Model Optimization Using GridSearchCV

GridSearchCV was used to improve the performance of the Random Forest model. This method tests different hyperparameter combinations with 5-fold cross-validation. Accuracy was chosen as the evaluation metric. The best parameters were found to be: max depth=None, max features=sqrt, min samplesleaf=1, min samples split=2, n estimators=100. The best cross-validation accuracy is approximately 0.9598. With these parameters, an accuracy of 100.00 percent was obtained in the test set.

7.3 Final Model Evaluation (TEST + Confusion Matrix)

As a result of hyperparameter optimization performed using the GridSearchCV method, the Random Forest model with the best parameter combination was selected as the final model. At this stage, the performance of the optimized Random Forest model was evaluated on a test dataset.

First, the model's accuracy value was calculated, and according to the results ob-

tained, the model correctly classified all samples in the test dataset. This indicates that the model correctly classified all samples in the test data.

To examine the model performance in more detail, a complexity matrix was created. The complexity matrix shows the relationship between the actual stress levels and the stress levels predicted by the model. As seen in the matrix, all classes were correctly predicted, and no misclassification occurred between classes.

A classification report was used to examine the model's precision, recall, and F1-score values. The classification report results reveal that the optimized Random Forest model exhibits a balanced and reliable performance in stress level prediction.

The results show that the Random Forest model, after hyperparameter optimization, is successful in predicting stress levels using sleep pattern and lifestyle-related variables.

Figure 4: Confusion matrix of the optimized Random Forest model

7.4 Error Analysis

This section details the prediction errors of the optimized Random Forest model on the test data. Correct and incorrect classifications were analyzed by comparing the actual stress levels with the stress levels predicted by the model.

First, an error analysis table was created using the predictions made on the test data. This table includes the actual stress level, the predicted stress level, and whether the prediction was correct for each sample. This allows for a clearer observation of which samples the model was successful in and which cases it made errors.

The analysis revealed that the number of incorrectly classified samples in the test dataset was zero. The model correctly classified all samples in the test dataset. This result shows that the model fits the test data very well.

Additionally, the relationship between correct and incorrect predictions and the sleep quality variable was examined. For this purpose, the sleep quality distribution for correctly predicted samples was visualized using a box plot. When the graph is examined, it is seen that in the examples where the predictions were correct, the sleep quality valuesare concentrated within a certain range.

Finally, the actual and predicted stress levels were compared according to the test sample indices. This visualization shows that the model's predictions largely coincide with the actual values. Overall, the results show that the optimized Random Forest model performs quite successfully in the stress level prediction problem.

References

[1] Sanidhya Jadaun, *Stress Level Prediction in Sleep Patterns*, Kaggle Notebook, 2023. Available at: `https://www.kaggle.com/code/sanidhyajadaun/stress-level-prediction-in-sleep-patterns/notebook` (Accessed: 17 December 2025).

[2] GeeksforGeeks, *Data Preprocessing in Machine Learning with Python*, 2023. Available at: `https://www.geeksforgeeks.org/machine-learning/data-preprocessing-machine-learning-python/` (Accessed: 17 December 2025).

[3] GeeksforGeeks, *What is Exploratory Data Analysis?*, GeeksforGeeks Article, 2023. Available at: `https://www.geeksforgeeks.org/data-analysis/what-is-exploratory-data-analysis/` (Accessed: 17 December 2025).

[4] GeeksforGeeks, *Pandas Cheat Sheet*, GeeksforGeeks Article, 2025. Available at: https://www.geeksforgeeks.org/pandas/pandas-cheat-sheet/ (Accessed: 17 December 2025).

[5] Stack Overflow, *Data analyse with pandas*, StackOverflow Q&A, 2022. Available at: https://stackoverflow.com/questions/70665959/data-analyse-with-pandas (Accessed: 17 December 2025).

[6] GeeksforGeeks, *One-Hot Encoding vs Label Encoding in Machine Learning*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/one-hot-encoding-vs-label-encoding/ (Accessed: 17 December 2025).

[7] Emre Yıldız, *Label Encoding ile OneHotEncoder Farkı*, Medium Blog, 2021. Available at: https://medium.com/@iamemreyildiz/label-encoding-ile-onehotencoder-fark%C4%B1-9cf7ad6028b5 (Accessed: 17 December 2025).

[8] Serdar Tafralı, *Veri Biliminde Özellik Ölçeklendirme (Feature Scaling)*, Medium Blog, 2021. Available at: https://serdartafrali.medium.com/veri-biliminde-%C3%B6zellik-%C3%B6l%C3%A7eklendirme-feature-scaling-e05d9f3e96ce (Accessed: 17 December 2025).

[9] GeeksforGeeks, *Understanding Logistic Regression*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/understanding-logistic-regression/ (Accessed: 17 December 2025).

[10] DataCamp, *Understanding Logistic Regression in Python*, DataCamp Tutorial, 2022. Available at: https://www.datacamp.com/tutorial/understanding-logistic-regression-python (Accessed: 17 December 2025).

[11] GeeksforGeeks, *Classifying Data Using Support Vector Machines (SVMs) in Python*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/classifying-data-using-support-vector-machinessvms-in-python/ (Accessed: 17 December 2025).

[12] Towards Data Science, *Support Vector Machines Explained with Python Examples*, Medium Article, 2020. Available at: https://medium.com/data-science/support-vector-machines-explained-with-python-examples-cb65e8172c85 (Accessed: 17 December 2025).

[13] Scikit-learn, *KNeighborsClassifier Documentation*, Scikit-learn Official Documentation, 2024. Available at: https://scikit-learn.org/stable/modules/

generated/sklearn.neighbors.KNeighborsClassifier.html (Accessed: 17 December 2025).

[14] Draj, *K-Nearest Neighbor (KNN) Using Python*, Medium Blog, 2020. Available at: https://medium.com/@draj0718/k-nearest-neighbor-knn-using-python-d0a6bb295e7d (Accessed: 17 December 2025).

[15] GeeksforGeeks, *K-Nearest Neighbors with Python*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/k-nearest-neighbors-with-python-ml/ (Accessed: 17 December 2025).

[16] Scikit-learn, *RandomForestClassifier Documentation*, Scikit-learn Official Documentation, 2024. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (Accessed: 17 December 2025).

[17] Towards Data Science, *Random Forest Explained: A Visual Guide with Code Examples*, Medium Article, 2020. Available at: https://towardsdatascience.com/random-forest-explained-a-visual-guide-with-code-examples-9f736a6e1b3c/ (Accessed: 17 December 2025).

[18] Analytics Vidhya, *Understanding Random Forest*, Analytics Vidhya Blog, 2021. Available at: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/ (Accessed: 17 December 2025).

[19] GeeksforGeeks, *Random Forest Classifier Using Scikit-learn*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/dsa/random-forest-classifier-using-scikit-learn/ (Accessed: 17 December 2025).

[20] Bilişim Hareketi, *Cross Validation Nedir, Nasıl Çalışır?*, Medium Blog, 2021. Available at: https://medium.com/bili%C5%9Fim-hareketi/cross-validation-nedir-nas%C4%B1l-%C3%A7al%C4%B1%C5%9F%C4%B1r-4ec4736e5142 (Accessed: 17 December 2025).

[21] GeeksforGeeks, *Understanding Feature Importance and Visualization of Tree Models*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/understanding-feature-importance-and-visualization-of-tree-models/ (Accessed: 17 December 2025).

[22] GeeksforGeeks, *Machine Learning Model Evaluation*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/machine-learning-model-evaluation/ (Accessed: 17 December 2025).

[23] GeeksforGeeks, *ML Models Score and Error*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/ml-models-score-and-error/ (Accessed: 17 December 2025).

[24] GeeksforGeeks, *Random Forest Hyperparameter Tuning in Python*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/random-forest-hyperparameter-tuning-in-python/ (Accessed: 17 December 2025).

[25] GeeksforGeeks, *Hyperparameter Tuning in Machine Learning*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/hyperparameter-tuning/ (Accessed: 17 December 2025).

[26] GeeksforGeeks, *Random Forest Algorithm in Machine Learning*, GeeksforGeeks Article, 2021. Available at: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/ (Accessed: 17 December 2025).

[27] Amy R. Mahdy, *Model Comparison: A Step-by-Step Guide to Comparing Several Machine Learning Models for Predictive Tasks*, Medium Article, 2021. Available at: https://amyrmahdy.medium.com/model-comparison-a-step-by-step-guide-to-comparing-several-machine-learning-models (Accessed: 17 December 2025).

[28] Analytics Vidhya, *Tune Hyperparameters with GridSearchCV*, Analytics Vidhya Blog, 2021. Available at: https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/ (Accessed: 17 December 2025).

[29] Himani Gulati, *Hyper-Parameter Tuning in Decision Trees and Random Forests*, Medium Article, 2020. Available at: https://medium.com/@himani-gulati/hyper-parameter-tuning-in-decision-trees-and-random-forests-3bdee09ea5af (Accessed: 17 December 2025).

[30] Nerd For Tech, *Predicting Price of Smartphones by Technical Specs Using Random Forest and Logistic Regression*, Medium Article, 2021. Available at: https://medium.com/nerd-for-tech/predicting-price-of-smart-phones-by-technical-specs-random-forest-logistic-regress (Accessed: 17 December 2025).

Dataset: https://data.mendeley.com/datasets/46j8wrc7p7/1 Source code: https://github.com/gulcinyzglc/Predicting-Stress-Level-Machine-Learning