Predicting Stress Level Based On Sleep and Lifestyle Habits

1st Aslıhan Akan
Computer Engineering
Istanbul Arel University
İstanbul, Türkiye
aslihanakan22@arel.edu.tr

2nd Gülçin Yüzgüleç
Computer Engineering
Istanbul Arel University
İstanbul, Türkiye
gulcinyuzgulec22@arel.edu.tr

3rd Esra Bayrak
Computer Engineering
Istanbul Arel University
İstanbul, Türkiye
esrabayrak22@arel.edu.tr

***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 preprocess- ing steps such as completing missing data, coding necessary variables, and normalization were applied. Exploratory data analysis was performed to ex- amine 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.**

***Index Terms*—stress level,life style,applied informatics,machine learning**

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.

I. 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 cer- tain 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 sur- veys 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 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 infer- ences 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 pat- terns 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.

II. 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 numeri- cal, 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

III. 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.

A. Data

Our study utilized a synthetic, publicly available dataset containing sleep health and lifestyle habits. Stress level was identified as the target variable to be predicted. The dataset included variables such as Personal ID, Gender, Age, Occupation, Sleep Duration, Sleep Quality, Physical Activity Level, Stress Level, BMI (Body Mass Index) Category, Blood Pressure, Heart Rate, Daily Step Count, and Sleep Disorder. These variables were included in the study because they are directly or indirectly related to physical health, mental state, and stress. The dataset contains 4862 data points in 374 rows and 13 columns.

B. Evaluation

In our study, machine learning models were evaluated by comparing the predicted stress levels with the actual stress values in the dataset. This evaluation was conducted to determine the extent to which the models successfully learned the relationships between sleep and lifestyle variables and stress levels. In the evaluation phase, appropriate performance comparison results were used to objectively measure the model. These comparison results provide information about the accuracy level of the resulting models by analyzing the distribution between the predicted and actual values. By analyzing the differences between the prediction results and the actual stress levels, the strengths of the models and areas for improvement were identified. The results show that sleep and lifestyle-related variables are effective in predicting stress levels. In particular, it was observed that changes in sleep duration, sleep quality, and some other variables significantly contributed to model performance. In this case, we realize that stress is the result of the interaction of numerous factors, and that considering these processes together improves the prediction. Furthermore, the evaluation results show that machine learning methods can model complex relationships that we cannot easily detect with simple supporting solutions. In conclusion, the results obtained during the evaluation phase indicate that the selected features and the program used are suitable for voltage level estimation. These results suggest that further predictions can be improved through more comprehensive model processing, enhanced feature selection, or the use of additional data.

IV. Data and Methodology

A. Description and Analysis of the Dataset

In this study, a publicly available dataset on sleep health and lifestyle habits was used to estimate stress levels. The dataset was structured to reflect information from individuals with diverse demographic, occupational, and lifestyle characteristics. This allowed for a broad examination of the relationship between stress and both physiological and behavioral factors. Stress is not a condition caused by a single factor; it is a complex emotion resulting from the combined effect of many factors such as sleep patterns, lifestyle, physical activity level, and general health status. Therefore, evaluating stress based on only one variable can lead to incomplete results. The variables used in this study were considered important information as they reflect both the physical and behavioral dimensions of

stress. Before proceeding with modeling, a comprehensive Exploratory Data Analysis (EDA) was conducted to better understand the structure of the dataset and identify potential problems. The main purpose of this stage was to reveal the general characteristics of the data, evaluate data quality, and prepare a solid foundation for subsequent analysis steps. In this context, the basic structure of the dataset was first examined, data types were checked, and general distributions among variables were observed. In the EDA process, the distributions of both numerical and categorical variables were analyzed in detail. The dataset was checked for missing values, and the impact of missing data on the analysis process was evaluated. These steps contributed to determining the methods to be applied in the data preprocessing process. Furthermore, a general overview of the dataset was provided by examining the relationships between sleep habits, lifestyle factors, and stress levels. Through these analyses, preliminary ideas about possible relationships between variables were obtained. This information guided the modeling and feature selection processes to be used in subsequent stages. As an example of these exploratory analyses, the following graph shows the distribution of stress levels in the dataset. As seen in the graph, stress levels are generally distributed around moderate and high values. This indicates that a significant portion of the individuals in the dataset experience moderate or high levels of stress. While the distribution appears generally balanced, it is noteworthy that some stress levels are observed more frequently than others.

Fig. 1. Distribution of stress levels

These results have shown us that the dataset represents various stress levels. The co-occurrence of moderate and high stress levels is important for developing models for stress prediction, as it allows them to distinguish between different stress levels. Furthermore, this indicates that stress is not dependent on a single variable, but rather on the combined effect of many factors such as sleep patterns, physical activity, and lifestyle.

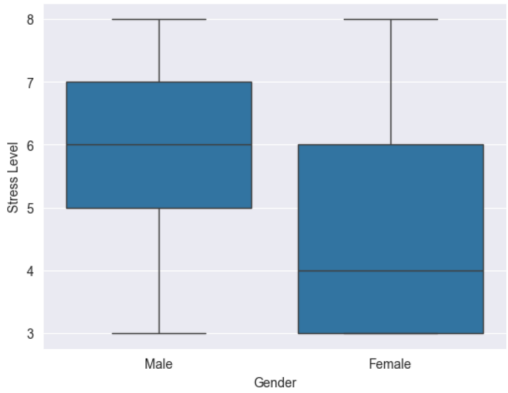

Fig. 2. A box plot showing the distribution of stress levels by gender.

This box plot shows the distribution of stress levels by gender. Examining the graph, we see differences in the distribution of stress levels between the male and female groups. The median stress level is higher in the male group, while the stress levels in the female group are distributed over a wider range. This indicates that stress levels are more variable in female individuals. The box plot reveals the central trends, interquartile ranges, and possible outliers of stress levels for both groups. The results suggest that stress levels may exhibit different distributions depending on the gender variable. In conclusion, this visual analysis demonstrates that gender may have a potential effect on stress levels and supports the consideration of gender as a helpful feature for stress estimation.

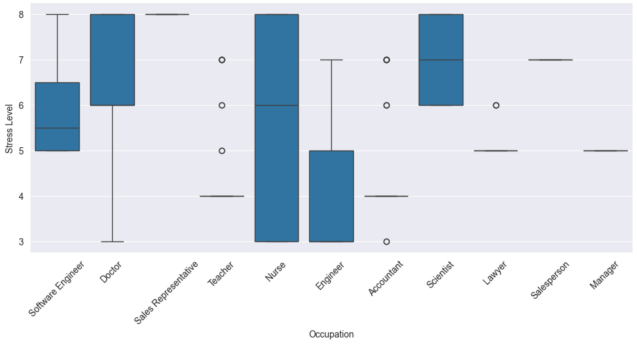

Fig. 3. A box plot showing the distribution of stress levels by occupation.

This box plot shows how stress levels are distributed across different occupational groups. The graph reveals significant differences in stress levels among occupations. Some occupational groups exhibit higher and more variable stress levels, while others show a more narrower concentration of stress levels. Stress levels show a wider distribution, particularly in health and science professions. Conversely, some occupational groups show a more limited concentration of stress levels. Box plots reveal the central trends, interquartile ranges, and potential outliers of stress levels for each occupation. These differences demonstrate that stress can be influenced not only

by personal factors but also by factors such as occupational conditions, workload, and the work environment.

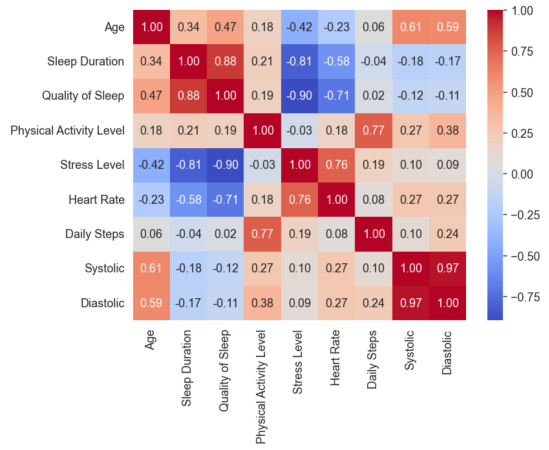

Fig. 4. Stess Level Correlation Matrix

In this section, correlation analysis was performed to examine the relationships between the variables in more detail. The correlation matrix given above shows the linear relationships between the numerical variables. Correlation coefficients range from -1 to +1, and the strength of the relationship between the variables increases as the coefficient approaches positive or negative 1. Examining the graph, we see a strong and negative relationship between stress level and sleep duration and sleep quality. In particular, the high negative correlation between stress level and sleep quality indicates that higher quality sleep is associated with lower stress levels. Similarly, it has been observed that stress levels tend to decrease as sleep duration increases. These results prove that sleep patterns have a significant effect on stress. Furthermore, it is noteworthy that there is a very high positive correlation between systolic and diastolic blood pressure. This is actually expected, confirming the consistency of blood pressure measurements. The age variable shows a positive relationship with blood pressure values and a negative relationship with stress level. A strong and positive correlation is observed between stress level and heart rate. This suggests that increased stress levels may be associated with physiological responses, particularly an increase in heart rate. In contrast, the correlations between stress level and physical activity level and daily step count are weak. This result indicates that physical activity may have an indirect or combined effect on stress.These correlation results indicate that stress levels are significantly influenced by sleep-related variables. Therefore, variables such as sleep duration and sleep quality have been considered important inputs for stress estimation. Correlation analysis contributed to determining the variables to be used in the modeling phase and understanding the necessity of a multivariate approach.

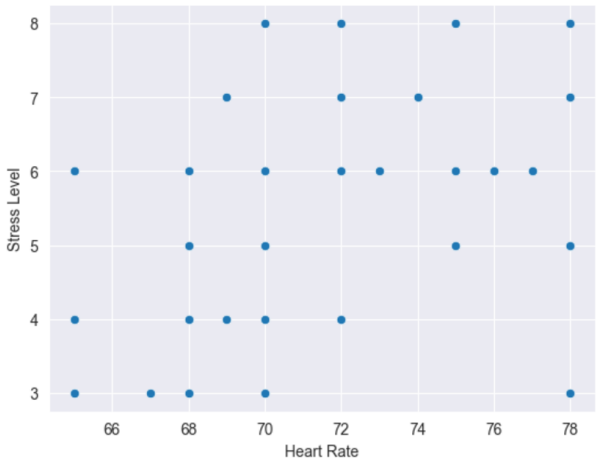

Fig. 5. Scatter plot showing the relationship between stress level and heart rate.

The scatter plot shown above more concretely illustrates the relationship between stress level and heart rate, and it is observed that it supports the findings obtained from the correlation analysis. As seen in the graph, stress levels generally increase as heart rate increases. This trend is consistent with the positive and strong relationship determined in the correlation matrix. Although the points appear regular, the overall trend reveals a positive relationship between heart rate and stress level. This indicates that high stress levels may be associated with physiological responses, particularly an increase in heart rate. However, the distribution observed in the graph reveals that stress cannot be explained solely by heart rate, and that sleep patterns, lifestyle, and other health indicators may also influence this relationship.

As mentionedin this parts, 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.

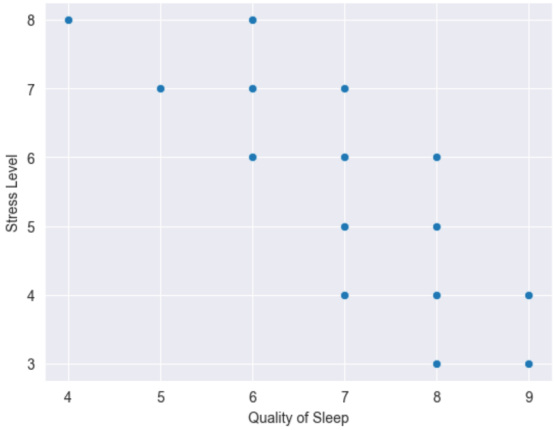

Fig. 6. Scatter plot showing the relationship between stress level and quality of sleep.

As another example, the scatter plot given above visually shows the relationship between stress level and sleep quality. As we can see from the graph, it is observed that as sleep quality increases, stress levels generally decrease. Individuals with low sleep quality have higher and more variable stress levels, while conversely, individuals with high sleep quality tend to have lower stress levels. The observation of a specific distribution in the data points is a clearer example of the negative relationship between sleep quality and stress level that we analyzed in the correlation matrix. This finding shows that sleep quality can have a significant effect on stress. However, the distribution in the graph indicates that stress cannot be explained solely by sleep quality and that other factors may also influence this relationship.

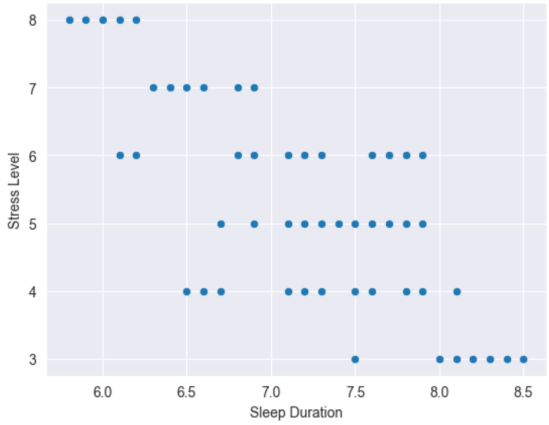

Fig. 7. The relationship between stress levels and sleep duration

In conclusion, the data analysis part of this study played a significant role in understanding the dataset. Examining the variable distributions and stress levels significantly contributed to planning data preprocessing steps, selecting appropriate features, and determining the modeling approach. These findings provide a strong foundation for stress prediction related to sleep health and lifestyle habits.

B. 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 depen- dent 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 un- derstand 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 bal- anced 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. There- fore, 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.

C. Machine Learning Models

In this study, four different machine learning algorithms were used to solve the stress level prediction problem. The main purpose of using multiple models was to compare their performances and observe how different approaches behave on the same dataset.

Both linear and nonlinear machine learning methods were included in the analysis. Linear models were selected because they are simple, easy to understand, and commonly used as baseline models. Nonlinear models were chosen because they can capture more complex patterns and relationships within the data. By applying different models, it became possible to evaluate their strengths and weaknesses in predicting stress levels. All models were trained and tested using the same dataset and evaluation metrics to ensure a fair comparison. The comparison of these machine learning models helps identify which algorithm is more suitable for stress level prediction based on sleep and lifestyle-related variables. This approach also provides a better understanding of how model complexity affects prediction performance.

1) Logistic Regression Model: First, the Logistic Regression model was used in this study. Logistic Regression is a simple and commonly used method for classification problems. It works by creating linear boundaries between different classes. Because of its simple structure, it is usually chosen as a baseline model.

In this study, the maximum number of iterations was set to 1000 so that the model could train properly without stopping early due to convergence problems. The main reason for using Logistic Regression was to see how a basic model performs and to use this result as a comparison for other models.

After training, the model was tested using the test dataset. The Logistic Regression model achieved an accuracy of 0.92, which means that most of the test samples were classified correctly. To better understand the results, a confusion matrix and a classification report were examined.

From the confusion matrix, it can be seen that many stress levels were predicted correctly. However, some mistakes occurred between close stress levels. This shows that Logistic Regression has difficulty handling more complex patterns in the data because of its linear structure.

The classification report also shows that the model achieved generally good precision, recall, and F1-score values, although the performance was lower for some classes. Overall, the Logistic Regression model provides a good starting point, but it is not sufficient on its own for stress level prediction.

	precision	recall	f1-score	support
3	1.00	0.86	0.92	14
4	0.88	1.00	0.93	14
5	0.87	0.93	0.90	14
6	0.78	0.78	0.78	9
7	1.00	0.90	0.95	10
8	1.00	1.00	1.00	14

Fig. 8.

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.

	precision	recall	f1-score	support
3	1.00	1.00	1.00	14
4	0.93	1.00	0.97	14
5	0.93	1.00	0.97	14
6	1.00	0.89	0.94	9
7	1.00	0.90	0.95	10
8	1.00	1.00	1.00	14

Fig. 9.

After the training process, the SVM model was tested using the test dataset. The model achieved an accuracy of about 0.97. This means that almost all test samples were predicted correctly. When compared to the Logistic Regression model, SVM gave better results. When the confusion matrix was checked, it was seen that the model made only a few mistakes. Most stress levels were predicted correctly, and only some close stress levels were confused with each other. According to the classification report, the SVM model shows high performance values. In general, these results show that the SVM model works well for stress level prediction and performs better than the basic model.

3) K-Nearest Neighbors (KNN) Model: To evaluate whether the selected Random Forest model exhibits consistent and reliable performance, not only on a single test set but across the entire dataset, a 5-fold K-Fold cross-validation strategy was employed. Cross-validation is a widely used resampling technique that provides a more robust estimation

of a model's generalization capability by reducing the dependency on a specific train–test split. In the applied 5-fold K-Fold cross-validation approach, the dataset was randomly divided into five equally sized and mutually exclusive subsets, referred to as folds. During each iteration, one fold was reserved as the validation set, while the remaining four folds were used for training the model. This process was repeated five times, ensuring that each fold was used exactly once as validation data. The performance of the Random Forest model was evaluated in each iteration, and the obtained accuracy scores were recorded. The final cross-validation performance was calculated by averaging the results across all folds. This averaging process helps mitigate the effects of data variance and provides a more stable and unbiased estimate of the model's predictive performance.

By employing K-Fold cross-validation, the risk of overfitting is reduced, as the model is exposed to different subsets of data during training and validation. Moreover, this method allows for a more comprehensive assessment of how well the model is expected to perform on unseen data. Therefore, the use of K-Fold cross-validation strengthens the reliability of the experimental results and supports the robustness of the selected Random Forest model

	precision	recall	f1-score	support
3	1.00	0.86	0.92	14
4	0.78	1.00	0.88	14
5	0.93	0.93	0.93	14
6	0.88	0.78	0.82	9
7	1.00	1.00	1.00	10
8	1.00	0.93	0.96	14

Fig. 10.

As the third model, the K-Nearest Neighbors (KNN) method was used. 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 value of k was chosen as 5. This means that the model looks at the five nearest neighbors and selects the stress level that appears the most among them.

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

	precision	recall	f1-score	support
3	1.00	0.93	0.96	14
4	0.93	1.00	0.97	14
5	1.00	1.00	1.00	14
6	0.89	0.89	0.89	9
7	1.00	1.00	1.00	10
8	1.00	1.00	1.00	14

Fig. 11.

due to its ability to detect nonlinear relationships In this study, the Random Forest model was configured with 100 decision trees (n estimators = 100). Using a sufficiently large number of trees enhances the robustness of the ensemble by averaging out individual tree errors. However, increasing the number of trees alone may lead to overly complex models. To mitigate this risk, the maximum depth of each decision tree was constrained to a value of max depth = 5. Limiting the tree depth helps control model complexity and reduces the likelihood of overfitting, especially when working with datasets of limited size.

The Random Forest model was further evaluated for its ability to capture nonlinear relationships between input features and stress levels. Unlike linear models, Random Forest can effectively model complex feature interactions, making it particularly suitable for real-world health-related datasets where relationships between variables are rarely linear.

Overall, the Random Forest Classifier was assessed as a powerful and reliable classification model due to its strong predictive performance, robustness against overfitting, and inherent interpretability through feature importance analysis. These characteristics make Random Forest a suitable choice for stress level prediction based on sleep health and lifestyle attributes .

m

5) K-Fold Cross Validation: To evaluate whether the selected Random Forest model exhibits consistent perfor- mance 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.

```
KFold scores: [0.97333333 0.92       0.94666667 0.98666667 0.93243243]
KFold mean: 0.9518198198198199
```

Fig. 12.

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.

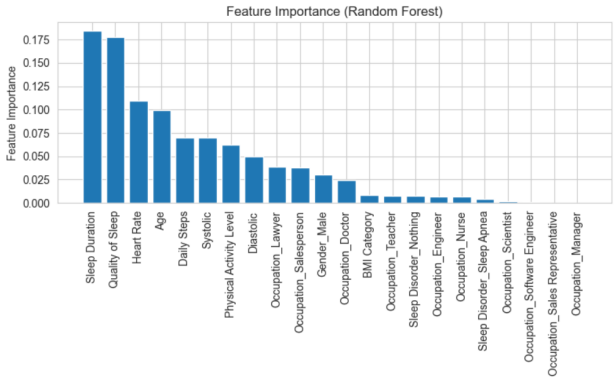

Fig. 13. Feature importance values obtained from the Random Forest model

```
Sleep Duration                0.184209
Quality of Sleep              0.177637
Heart Rate                    0.109802
Age                           0.099497
Daily Steps                   0.070047
Systolic                      0.069727
Physical Activity Level       0.062836
Diastolic                     0.049479
Occupation_Lawyer             0.038450
Occupation_Salesperson        0.038158
Gender_Male                   0.030045
Occupation_Doctor             0.024873
BMI Category                  0.008384
Occupation_Teacher            0.007902
Sleep Disorder_Nothing        0.007545
```

Fig. 14.

```
Occupation_Lawyer                0.038450
Occupation_Salesperson           0.038158
Gender_Male                      0.030045
Occupation_Doctor                0.024873
BMI Category                     0.008384
Occupation_Teacher               0.007902
Sleep Disorder_Nothing           0.007545
Occupation_Engineer              0.006726
Occupation_Nurse                 0.006566
Sleep Disorder_Sleep Apnea       0.004729
Occupation_Scientist             0.002057
Occupation_Software Engineer     0.001272
Occupation_Sales Representative  0.000059
Occupation_Manager               0.000000
dtype: float64
```

Fig. 15.

When the results are examined, it can be seen that some features are more important than others. In particular, Sleep Duration, Quality of Sleep, and Heart Rate have higher importance values compared to the remaining features. This shows that these variables play a more significant role in determining stress levels in the dataset. Other features such as age, daily steps, physical activity level, and BMI category have lower importance values. This means that although these features still contribute to the prediction process, their effect on stress level prediction is smaller compared to sleep-related variables. Overall, this analysis shows that stress levels are more strongly influenced by sleep-related factors in this dataset. In addition, feature importance analysis provides useful information for future studies, as it can help researchers focus on the most

important variables and simplify the model by removing less important features.

V. Results and Discussion

A. 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. Using the same test set for all models ensured a fair and reliable comparison.

	Model	Accuracy	F1-score (weighted)
1	Support Vector Machine	0.973333	0.973050
3	Random Forest	0.973333	0.973316
0	Logistic Regression	0.920000	0.920202
2	K-Nearest Neighbors	0.920000	0.920884

Fig. 16. Comparison of classification models based on accuracy and weighted F1-score

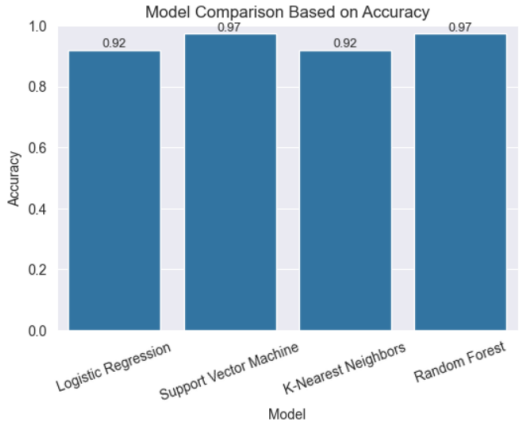

Fig. 17. Comparison of model accuracy values for stress level prediction

As presented in Fig. 4 and Fig. 5, Random Forest and Support Vector Machine models achieved the highest accuracy values among the evaluated models. While Fig.4 provides a numerical comparison based on accuracy and weighted F1-score, Fig.5 visually illustrates the differences in accuracy across models. This result indicates that these two models were more successful in correctly classifying stress levels compared to the other approaches.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 per- formance 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 For- est model was selected for use in subsequent stages of analysis.

Its superior performance and stability make it a suitable choice for stress level prediction using sleep and lifestyle-related features.

This version streamlines the comparisons, clarifies the results, and explains the rationale for selecting the Random Forest model.

B. Model Optimization Using GridSearchCV

GridSearchCV was used to improve the performance of the Random Forest model by systematically searching for the most suitable hyperparameter combination. This approach evaluates multiple parameter settings using 5-fold cross-validation, which helps reduce the risk of overfitting and provides a more reliable performance estimation.

In this study, accuracy was selected as the evaluation metric during the optimization process. Several hyperparameters, including the number of trees, tree depth, and node splitting criteria, were tested. As a result of this process, the best parameters were found to be as follows: max depth=None, max features=sqrt, min samples leaf=1, min samples split=2, and n estimators=100.

The best cross-validation accuracy is approximately 0.9598. This demonstrates strong and consistent model performance across different validation layers. When the optimized Random Forest model was evaluated on the independent test dataset, it achieved an accuracy of 100.00 percent. This result suggests that the selected hyperparameters significantly improved the model's ability to correctly classify stress levels based on sleep and lifestyle-related features.

C. 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 obtained, 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 confusion matrix was created, as shown in Fig.6. The confusion 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.

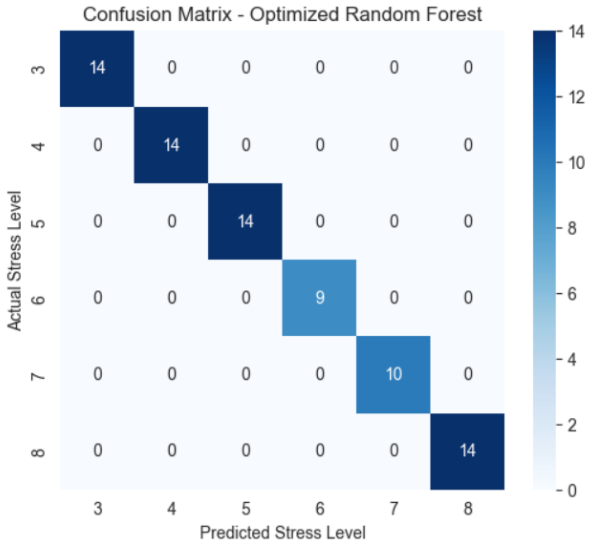

Fig. 18. Confusion matrix of the optimized Random Forest model

In addition, a classification report was generated to further evaluate the model in terms of precision, recall, and F1-score values for each stress level class, as presented in Fig.6. The results of the classification report indicate that the optimized Random Forest model achieved perfect scores across all evaluation metrics, demonstrating a balanced and reliable performance in stress level prediction.

Classification Report - Optimized Random Forest:

	precision	recall	f1-score	support
3	1.00	1.00	1.00	14
4	1.00	1.00	1.00	14
5	1.00	1.00	1.00	14
6	1.00	1.00	1.00	9
7	1.00	1.00	1.00	10
8	1.00	1.00	1.00	14
accuracy			1.00	75
macro avg	1.00	1.00	1.00	75
weighted avg	1.00	1.00	1.00	75

Fig. 19. Classification report of the optimized Random Forest model on the test dataset

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

D. 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.

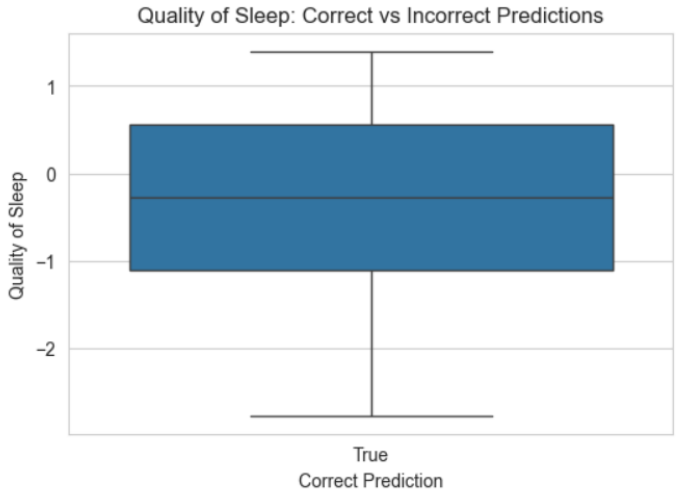

Fig. 20. Distribution of sleep quality values for correctly predicted samples

Additionally, the relationship between correct and incorrect predictions and the sleep quality variable was examined. In addition, the relationship between prediction correctness and the sleep quality variable was analyzed. For this purpose, the distribution of sleep quality values for correctly predicted samples was visualized using a box plot, as shown in Fig. 8. When the figure is examined, it can be observed that the sleep quality values of correctly classified samples are concentrated within a specific range. This observation suggests that sleep quality plays an important role in the model's ability to accurately predict stress levels.

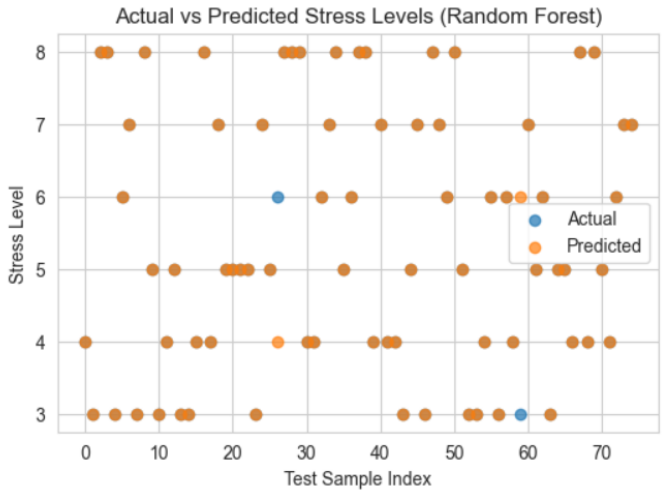

Fig. 21. Comparison of actual and predicted stress levels on the test dataset using the optimized Random Forest model

Finally, the actual and predicted stress levels were compared based on the test sample indices, as illustrated in Fig.9. This visualization shows that the model's predictions largely coincide with the actual values. The strong overlap between actual and predicted stress levels further confirms the reliability of the optimized Random Forest model. Overall, the results show that the optimized Random Forest model performs quite successfully in the stress level prediction problem.

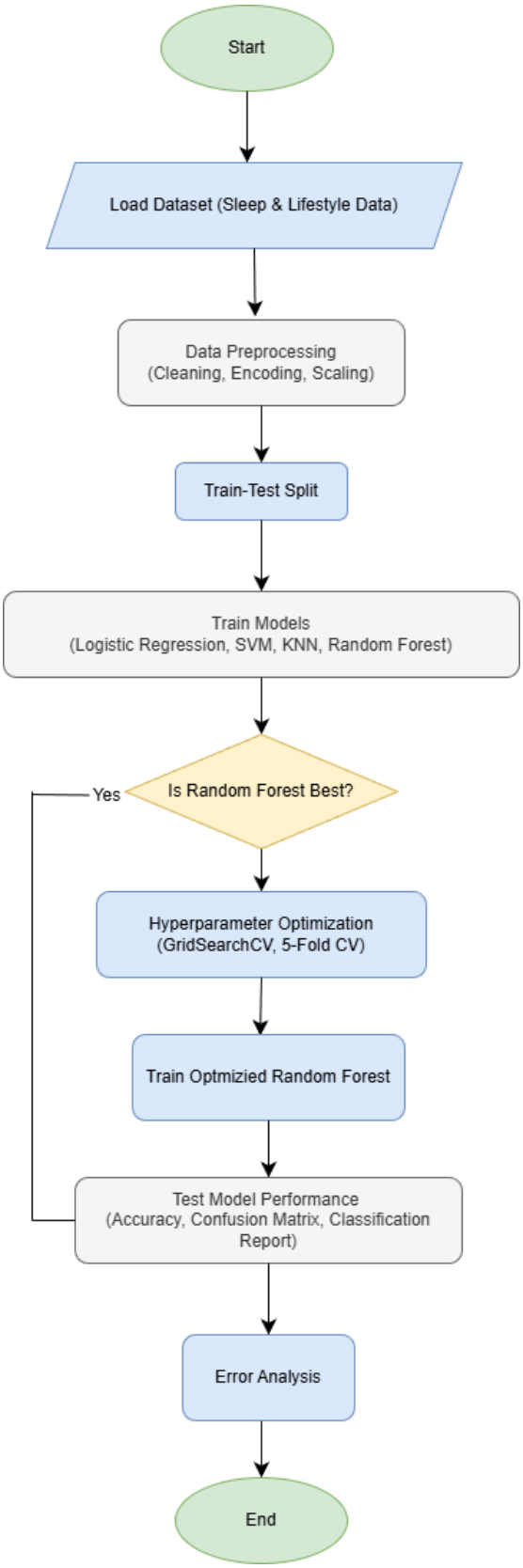

Fig. 22. Flowchart of the proposed machine learning pipeline for stress level prediction

-

VI. Future Work and Project Improvements

In this study, various machine learning models were developed and compared for the purpose of stress level prediction using sleep health and lifestyle data. The obtained results indicate that the Random Forest model provides high predictive performance as well as interpretability. However, the current version of the study has certain limitations and can be further improved in several directions.

First, the dataset used in this study contains a limited number of samples and consists of static features. In future studies, utilizing larger datasets that cover diverse demographic groups would enhance the generalizability of the model. Moreover, considering that stress is a dynamic phenomenon that changes over time, incorporating temporal data into the analysis could lead to more realistic and accurate predictions. In this context, time-series data collected from wearable devices, such as heart rate, sleep duration, and physical activity measurements, may be used.

Another important direction for improvement is the integration of deep learning-based approaches. Recurrent neural network architectures, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), are particularly effective in modeling temporal dependencies. Employing such models could enable the prediction of stress levels based not only on current data but also on historical patterns.

In addition, this study provides a basic level of model interpretability through feature importance analysis derived from the Random Forest model. Future work may incorporate Explainable Artificial Intelligence (XAI) techniques to further enhance transparency and trust. Methods such as SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) allow both global and local explanations by revealing the contribution of individual features to model predictions. Integrating XAI techniques would transform the model from a black-box system into an interpretable and trustworthy decision-support tool.

Furthermore, while the current study formulates stress level prediction as a classification problem, future research may explore regression-based approaches to model stress as a continuous variable. This could allow for more fine-grained and sensitive stress level estimation.

Finally, to increase the practical applicability of the proposed approach, the developed model could be integrated into a decision support system or a user-friendly interface. Such a system would enable individuals to input their own sleep and lifestyle data and receive personalized feedback regarding their stress levels. In this way, the proposed study could extend beyond an academic analysis and evolve into a real-world application with societal benefits.

Overall, these future directions demonstrate that the proposed study provides a solid foundation for more comprehensive and realistic stress level prediction systems in subsequent research.

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.

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(Accessed: 17 December 2025).

[8] Serdar Tafralı, *Veri Biliminde Özellik Ölçeklendirme (Feature Scaling)*, Medium Blog, 2021. Available at: https://serdartafrali.medium.com/veri-biliminde-(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(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-for-predictive-a8fc57af45b (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-regression-48ddc0cdeb0c (Accessed: 17 December 2025).

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