

1. Code Summary

The notebook "Arda_dsa210_new-2.ipynb" serves as a data analysis and modeling framework for examining relationships between lifestyle metrics (resting heart rate, coffee consumption, sleep hours, walking activity, and screen time) for two individuals, Arda and Ayşe. Below is a summary of the major components of the notebook:

- **Data Loading and Preparation:**

The dataset is loaded and processed.

The Date column is converted to datetime format for proper time-series analysis.

Descriptive statistics are computed for key variables, including resting heart rate, coffee consumption, and walking activity.

- **Exploratory Data Analysis (EDA):**

Visualizations include:

- Time-series plots of sleep hours for Arda and Ayşe.
- Trends and distributions of resting heart rate.
- Scatter plots analyzing the relationship between coffee consumption and resting heart rate.

Correlation analysis is performed to understand relationships between variables.

- **Feature Engineering:**

Key features such as screen time, coffee consumption, and walking steps are used as predictors for modeling sleep hours.

- **Machine Learning:**

A Random Forest Regressor is trained to predict Arda's sleep hours.

The model evaluates performance using metrics like Mean Squared Error (MSE) and R^2 score.

Feature importance analysis highlights the impact of different variables on sleep predictions.

2. Results and Findings

Exploratory Insights:

- **Resting Heart Rate:**

Arda has a higher average resting heart rate (74.43 bpm) compared to Ayşe (59.14 bpm).

The distributions reveal consistent heart rate ranges for both individuals, with occasional spikes.

- **Sleep Hours:**

Time-series plots indicate variability in sleep hours, with occasional extended sleep periods.

Ayşe's sleep hours appear more consistent compared to Arda's.

- **Coffee Consumption:**

Scatter plots reveal no strong linear relationship between coffee consumption and resting heart rate for either individual.

Modeling Results:

- The Random Forest model achieved:

MSE: A measure of the average squared difference between predicted and actual sleep hours.

R² Score: Indicates the proportion of variance in the target variable explained by the model.

- **Feature Importance:**

Features like walking steps and screen time have higher predictive power for Arda's sleep hours compared to coffee consumption.

3. Documentation

Project Context: This notebook aligns with the goals of the DSA210 project by providing a comprehensive framework for analyzing personal health metrics and their interactions. The key aspects of the analysis include:

- **Techniques and Tools:**

Libraries: pandas for data manipulation, matplotlib and seaborn for visualizations, sklearn for modeling.

Random Forest Regressor as the primary model for prediction tasks.

- **Limitations:**

The dataset size may limit the generalizability of findings.

External factors, such as diet or stress levels, are not included in the analysis.

- **Future Work:**

Expand the dataset to include additional individuals and metrics.

Test advanced models such as Gradient Boosting or Neural Networks.

Reproducibility:

1. Upload the dataset `NIHAI_TABLO_full_reconstructed.csv`.
2. Install the required Python libraries (pandas, matplotlib, seaborn, sklearn).
3. Run the notebook step-by-step to replicate the analysis.

4. Presentation Summary

This project showcases the practical application of data science techniques to uncover patterns and insights in personal health metrics. Key contributions include:

- Developing a data pipeline for processing and visualizing daily health data.
- Using machine learning models to predict sleep hours based on lifestyle factors.
- Highlighting actionable insights, such as the impact of screen time and walking activity on sleep quality.

The analysis provides a foundation for further exploration into personalized health data, emphasizing the importance of integrating multiple metrics for a holistic understanding of well-being.