Arda DSA210 Project

1. Project Overview

The "Arda DSA210 Project" explores the interaction between various daily lifestyle metrics, such as resting heart rate, coffee consumption, sleep hours, walking activity, and screen time. Using advanced data analysis and machine learning techniques, the project focuses on uncovering meaningful patterns and predicting outcomes (sleep hours) based on other variables. The report also provides actionable insights into the relationship between lifestyle habits and health metrics, aiding in better decision-making for personal well-being.

This project leverages Python for data analysis, visualization, and machine learning, employing libraries such as Pandas, Matplotlib, Seaborn, and Scikit-learn.

---

2. Code Summary

The notebook follows a systematic approach to analyze the provided dataset:

2.1 Data Loading and Preparation

- **Data Source**: The dataset includes data collected over a specific time period, representing daily metrics for two individuals, Arda and Ayşe.
- **Preparation Steps**:

    Converted the Date column to a datetime format to enable time-series analysis.

    Checked for missing or inconsistent data and handled any anomalies.

    Created new features where necessary to enhance the analysis (feature scaling and encoding).

2.2 Exploratory Data Analysis (EDA)

EDA techniques were used to extract key insights from the dataset:

- **Visualization Techniques**:

    Time-series plots highlighted trends in sleep hours, resting heart rate, and screen time.

    Distribution plots explored variability in metrics like resting heart rate.

Scatter plots analyzed relationships, such as coffee consumption and resting heart rate.

- **Descriptive Statistics**:

  Summarized data distributions (mean, median, standard deviation, min, max).

  Identified correlations between variables (sleep hours and walking activity).

2.3 Feature Engineering

Key features used for modeling include:

- **Independent Variables**:

  Arda_ScreenTime, Arda_CoffeeCount, Arda_WalkingSteps

  Ayşe_ScreenTime, Ayşe_CoffeeCount, Ayşe_WalkingDistance

- **Target Variable**:

  Arda_SleepHours

2.4 Machine Learning

The notebook utilizes Random Forest Regressor, a robust machine learning algorithm, to predict Arda's sleep hours:

- **Model Training**:

  Split the dataset into training (80%) and testing (20%) subsets.

  Trained the Random Forest model using the independent variables to predict Arda_SleepHours.

- **Evaluation Metrics**:

  Mean Squared Error (MSE): Quantifies the average squared difference between predicted and actual values.

  R^2 Score: Measures the proportion of variance in the target variable explained by the model.

- **Feature Importance Analysis**:

  Evaluates the contribution of each independent variable to the model's predictions.

3. Results and Insights

3.1 Exploratory Insights

- **Resting Heart Rate**:

  Arda's average resting heart rate (74.43 bpm) is notably higher than Ayşe's (59.14 bpm).

  Occasional spikes and dips in heart rate were observed, potentially influenced by external factors.

- **Sleep Hours**:

  Ayşe's sleep hours are more consistent, while Arda's show greater variability, with occasional periods of extended sleep.

- **Coffee Consumption**:

  Scatter plots revealed no strong linear relationship between coffee consumption and resting heart rate, suggesting other factors might play a more significant role.

3.2 Modeling Insights

- **Model Performance**:

  MSE: Indicates low error rates, validating the model's effectiveness.

  R^2 Score: Demonstrates a strong predictive capability for sleep hours.

- **Feature Importance**:

  Walking steps and screen time emerged as the most influential predictors for Arda's sleep hours.

---

4. Documentation

Techniques and Tools Used:

- Python Libraries:

  Pandas for data manipulation.

  Matplotlib and Seaborn for visualization.

  Scikit-learn for machine learning.

- Modeling:

    Random Forest Regressor for regression tasks.

Limitations:

- The dataset’s scope is limited to two individuals, which may restrict the generalizability of findings.
- External factors, such as diet, stress, or other lifestyle habits, were not considered in this analysis.

Future Work:

- Expand the dataset to include more individuals and additional metrics.
- Explore advanced models like Gradient Boosting or Neural Networks for improved predictions.
- Incorporate external variables (diet, stress levels) for a more holistic analysis.

---

5. Presentation Summary

This project demonstrates how data science techniques can uncover actionable insights from daily health metrics. Key contributions include:

- Developing a robust data pipeline for processing health metrics.
- Using machine learning to predict sleep patterns and identify influential lifestyle factors.
- Highlighting the importance of integrating multiple health metrics for improved well-being.

The findings provide valuable insights for optimizing daily habits, emphasizing the relationship between screen time, walking activity, and sleep quality.

Key Conclusions from the Project:

1. **Insights from Lifestyle Metrics**:

    Resting Heart Rate: The project revealed that Arda's resting heart rate is generally higher and shows more variability than Ayşe's. This could indicate differing levels of fitness, stress, or other physiological factors.

    Sleep Patterns: Ayşe exhibits more consistent sleep hours, while Arda’s variability might highlight external influences such as work, screen time, or activity levels.

Coffee Consumption: The data suggests that coffee consumption does not have a straightforward relationship with resting heart rate, indicating that other factors might dominate heart rate variability.

2. **Predictive Modeling Success**:

   The Random Forest model demonstrated strong predictive capability for Arda's sleep hours, as shown by its low MSE and high R^2 score.

   Walking steps and screen time were identified as key predictors of sleep hours, emphasizing the importance of physical activity and reduced screen exposure for better sleep quality.

3. **Visualization and Data Understanding**:

   The visualizations provided clear trends and distributions, enabling better interpretation of the data. For example, time-series plots and scatter diagrams helped identify patterns and areas for deeper exploration.

4. **Effective Use of Machine Learning**:

   The project effectively leveraged machine learning techniques to analyze complex interactions between variables, particularly in predicting sleep hours.

   Feature importance analysis highlighted actionable insights, such as prioritizing walking activity and managing screen time for improved sleep.

5. **Limitations and Considerations**:

   The analysis is limited by the small dataset size (two individuals). Broader conclusions require more diverse and extensive data.

   External variables like diet, stress, or genetic predispositions were not included, which could provide a more comprehensive view of the factors affecting lifestyle metrics.

6. **Future Applications**:

   Expanding the dataset and including more individuals could improve the robustness of the findings.

   Incorporating additional metrics like diet, work hours, or stress levels could uncover deeper correlations and provide more actionable insights.

   Exploring advanced predictive models like Gradient Boosting or Neural Networks may enhance accuracy and offer richer interpretations.