**Sabancı University**

**DSA210 – Introduction to Data Science**

**Term Project Report**

**Zeynep Su Yaşar – 30694**

1. **Introduction**

This is the report for my project of the course, DSA210 - Introduction to Data Science. The codes and the data files can be found in the GitHub page.

<https://github.com/zeynepsuyasar/DSA210_Term_Project>

1. **Motivation**

This project investigates how various stages of the menstrual cycle affect physical activity levels. The main aim is to identify trends and relationships by using data science methods, which can help us gain insights into personal health and optimize activity scheduling. By exploring this link, the project highlights the significance of tailored health approaches and improved health monitoring through predictive analytics. Moreover, it acts as a hands-on application of data science skills in a real-world scenario.

1. **Objectives**

* Utilize data science methodologies on practical health data.
* Examine the correlation between phases of the menstrual cycle and daily step counts.
* Perform exploratory data analysis (EDA) to uncover patterns.
* Develop machine learning models to forecast menstrual cycle days based on step count information.

1. **Data Collection**

* **Menstrual Cycle Data: Daily records from Apple’s Health Application, featuring a binary indicator (Cycle Yes/No) for cycle days.**
* **Step Count Data:**Daily step totals from Apple’s Health Application, recorded from 2020 to the present.
* The data was organized to fix any missing or repeated entries.
* The "Dates" column was aligned to combine the datasets properly.

1. **Data Analysis**

* Descriptive statistics, such as averages and ranges, were calculated.

**A white sheet with black text

Description automatically generated**

**A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a computer

Description automatically generated**

* Visualization

A histogram displaying my daily step counts helps to reveal patterns in my activity levels. The majority of my days are between 0 and 5000 steps, which suggests I have a moderate level of activity, while only a few days go beyond 10,000 steps:

**A diagram of a step count

Description automatically generated**

I created a box plot to look at my step counts on days when I cycle versus days when I don’t. It shows that my step counts are usually lower on cycling days, and the numbers are more consistent. In contrast, on non-cycling days, my step counts vary a lot more:

**A graph of a step count

Description automatically generated**

I created a line plot to monitor my daily step counts and see how my activity levels change over time. It shows that my activity varies a lot, with some days where I hit over 20,000 steps, while on other days, the numbers are significantly lower:

**A graph of a step

Description automatically generated**

A line graph featuring a moving average helps me monitor my daily step counts over time and reduces the ups and downs in my activity levels. It reveals steady patterns, with the red 7-day moving average indicating times of higher or lower activity, while some days occasionally exceed 20,000 steps:

A graph showing a step count

Description automatically generated

A pie chart illustrates the breakdown of my total step counts on days I cycled versus days I didn't. It clearly shows that 82.1% of my steps were recorded on non-cycle days, compared to just 17.9% on cycle days, indicating a big difference in how active I was:

**A blue circle with a pink triangle and black text

Description automatically generated**

A violin plot was created to compare how my step counts vary on cycle days versus non-cycle days. The results indicate that non-cycle days have a broader range of step counts and higher activity levels, whereas the step counts on cycle days are more stable and typically lower:

**A diagram of a step count distribution

Description automatically generated**

A bar graph illustrates the average number of steps taken each day of the week. It reveals that I tend to be most active on Fridays and Saturdays, whereas Sundays have the fewest steps on average:

**A graph of a step count

Description automatically generated with medium confidence**

A line graph displays my step count patterns over time, with cycle days marked in red. It shows that my step counts usually drop on those cycle days, highlighting how the menstrual cycle affects my activity levels:

A graph of blue and red dots

Description automatically generated

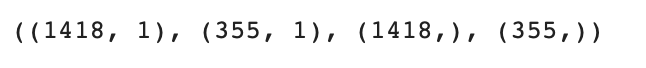
The correlation matrix:

**A screenshot of a graph

Description automatically generated**

1. **Machine Learning**

The dataset is split into 80% training and 20% testing, using 'Step Count' as the feature and 'Cycle Yes/No' as the target. The resulting shapes are (1418, 1) for training and (355, 1) for testing:



The dataset includes three new columns: Day of Week, Month, and Is Weekend. For example, in the first row, the date corresponds to a weekday (Day of Week = 3, Is Weekend = 0), and the month is January (Month = 1):

**A screenshot of a number

Description automatically generated**

Null Hypothesis (H₀): There is no significant relationship between step count and menstrual cycle days. Step counts are not predictive of whether a day is part of the menstrual cycle or not.

Alternative Hypothesis (H₁): There is a significant relationship between step count and menstrual cycle days. Step counts can be used to predict whether a day is part of the menstrual cycle.

**A screenshot of a number of numbers

Description automatically generatedA graph of a curve

Description automatically generated**

Null Hypothesis (H₀): Step count data does not significantly contribute to predicting menstrual cycle days. The XGBoost classifier will perform no better than random guessing.

Alternative Hypothesis (H₁): Step count data significantly contributes to predicting menstrual cycle days. The XGBoost classifier will perform better than random guessing and show strong predictive performance.

**A screenshot of a computer

Description automatically generatedA graph of a curve

Description automatically generated**

1. **Findings**

Patterns Observed: There are noticeable differences in step counts on days of the menstrual cycle compared to non-menstrual days.

Model Effectiveness: XGBoost outperformed Random Forest in terms of prediction accuracy.

Significance of Features: Enhancing features, especially the Is Weekend variable, led to better results in the model.

1. **Limitations and Future Work**

A . Limitations

* The study is based only on step count data, which might not fully reflect the intricate effects of the menstrual cycle.
* Other influences, like stress, sleep quality, or diet, were not taken into account.
* The models could be too tailored to the data because of the small number of features used.

B. Future Work

* Broaden Features: Add more health indicators such as heart rate and stress levels.
* Gather More Data: Utilize long-term datasets to observe trends over longer time frames.
* Try Deep Learning: Investigate more advanced algorithms for better accuracy and scalability.
* Test Models: Use the models on different groups of people to make sure they work well for everyone.