CS210 TERM PROJECT REPORT

Health Insights: Exploring Predictive Models for Personal Wellness using Mobile Health Data

19 JANUARY 2024

OKAN ARIF GUVENKAYA

okanarif@sabanciuniv.edu

26780

Table of Contents

1) INTRODUCTION

In an era where personal health and wellness are paramount, leveraging the power of data analytics and machine learning becomes crucial for gaining valuable insights. This project delves into the exploration of predictive models to enhance our understanding of personal wellness using data collected from the Huawei Health application on a mobile device. The dataset encompasses a diverse array of health-related parameters, including daily activities such as exercise, standing hours, and step count, providing a comprehensive view of one's lifestyle. From data acquisition to analysis and machine learning methodologies, this project aims to uncover patterns, correlations, and predictive capabilities to contribute to the growing field of personalized health analytics. By harnessing the potential of advanced analytics, the project endeavors to offer actionable insights for individuals seeking to optimize their well-being based on their unique activity patterns and habits.

2) METHODOLOGY

1. Data Acquisition and Preprocessing

The dataset for this project, meticulously curated from the Huawei Health application, was manually recorded and stored in a CSV file named "My_Health_Data.csv." Notably, during the initial exploration in Google Colab, a comprehensive dataset of 214 entries and 8 columns was observed. However, it is imperative to mention that specific days, considered as outliers or anomalies, were intentionally excluded from the analysis to ensure the fidelity of the insights derived. The dataset's columns were categorized into numeric, such as 'Day,' 'Move(kcal),' 'Exercise(min),' 'Stand(hours),' 'Steps,' and 'Distance(km),' and categorical, including 'Month' and 'Day of Week.' Further preprocessing involved mapping categorical columns to their corresponding numeric representations. This meticulous curation and preprocessing aimed to refine the dataset, enhancing its suitability for subsequent exploratory and machine learning analyses. The resulting dataset not only captures the nuances of daily physical activity but also maintains a focus on data quality by addressing potential outliers that might skew the analysis.

2. Exploratory Data Analysis (EDA):

The summary statistics, as illustrated in the provided outputs, offered a comprehensive overview of the dataset's central tendencies and variabilities. Notably, the 'Move(kcal)' column exhibited a wide range, with a maximum value of 890 kcal, suggesting substantial variations in daily caloric expenditure. The subsequent mapping of categorical columns allowed for a more nuanced understanding of temporal patterns.

3. Top 10 Days Analysis:

The identification of the top 10 days with the highest caloric expenditure, as detailed in the outputs, provided a closer look at specific instances of heightened physical activity. For instance, the day with the highest 'Move(kcal)' value (890 kcal) occurred on the 31st of May.

4. Correlation Matrix and Monthly Averages:

As part of the exploratory phase, a correlation matrix was generated to examine potential relationships between different features. The provided heatmap visualized these correlations, offering insights into how variables interact. Monthly averages for key metrics, including 'Move(kcal),' 'Exercise(min),' 'Stand(hours),' 'Steps,' and 'Distance(km),' were plotted in bar charts. This visual representation highlighted monthly trends, allowing for a more nuanced interpretation of the data.

5. Machine Learning Approach:

The dataset underwent preprocessing steps, including the conversion of categorical variables and the division of 'Day of Week' into weekdays and weekends. Leveraging a K-Nearest Neighbors (KNN) classifier, the model predicted whether a given number of steps corresponds to a weekday or weekend. The evaluation metrics, as outlined in the provided outputs, encompassed accuracy calculations, a confusion matrix, and a Receiver Operating Characteristic (ROC) curve. These metrics collectively gauged the model's performance and its ability to discern patterns in the data.

3) RESULTS AND DISCUSSION

```
Cleaned dataset without outliers:
     Day     Month Day of Week  Move(kcal)  Exercise(min)  Stand(hours)  \
0     30  November    Thursday         301             22            10
1     29  November   Wednesday         398             22            12
2     28  November     Tuesday         690             17            13
3     27  November      Monday         548             44            14
5     25  November    Saturday         546             39            12
..   ...       ...         ...         ...            ...           ...
209    5       May      Friday         486             33            12
210    4       May    Thursday         458             21            11
211    3       May   Wednesday         460             34            11
212    2       May     Tuesday         462             39            13
213    1       May      Monday         193             18             6

       Steps  Distance(km)
0     8990.0          6.42
1    12329.0          8.80
2    23244.0         16.60
3    17637.0         12.59
5    15948.0         11.39
..       ...           ...
209  12423.0          8.87
210  12639.0          9.02
211  11722.0          8.37
212  11725.0          8.37
213   5056.0          3.61

[211 rows x 8 columns]
```

Figure 1: Cleaned dataset.

The cleaned dataset has successfully removed outliers from the original dataset, particularly in numerical columns such as 'Move(kcal)', 'Exercise(min)', 'Stand(hours)', 'Steps', and 'Distance(km)'. Outliers were identified and excluded based on the interquartile range (IQR) method, ensuring that extreme values do not skew the analysis. The resulting dataset now consists of 211 rows, each representing a day, with refined values that better reflect the central tendency of the health data. This preprocessing step contributes to a more accurate and reliable representation of the dataset for subsequent analyses and modeling.

```
First few rows of the dataset:
   Day     Month Day of Week  Move(kcal)  Exercise(min)  Stand(hours)  \
0   30  November    Thursday         301             22            10
1   29  November   Wednesday         398             22            12
2   28  November     Tuesday         690             17            13
3   27  November      Monday         548             44            14
4   25  November    Saturday         546             39            12

     Steps  Distance(km)
0   8990.0          6.42
1  12329.0          8.80
2  23244.0         16.60
3  17637.0         12.59
4  15948.0         11.39
```

Figure 2: First 5 rows of the data

The initial rows of the dataset offer a glimpse into the daily health metrics, combining various aspects such as caloric expenditure, exercise duration, standing hours, steps, and distance covered. For instance, on November 30th, a Thursday, the recorded metrics include the burning of 301 kcal, engaging in 22 minutes of exercise, standing for 10 hours, taking 8990 steps, and covering a distance of 6.42 kilometers. Sequential days showcase fluctuations in caloric expenditure, providing insights into potential temporal patterns. The correlation analysis suggests a positive relationship between exercise duration and caloric burn, highlighting that days with longer exercise durations are associated with higher energy expenditure. Additionally, the inclusion of 'Stand(hours)' offers insights into sedentary behavior, while 'Steps' and 'Distance(km)' provide a quantifiable measure of daily movement, facilitating the assessment of overall activity levels. These initial observations lay the groundwork for subsequent analyses, aiming to unravel meaningful patterns that contribute to

a nuanced understanding of individual well-being

```
Numeric columns:
Index(['Day', 'Move(kcal)', 'Exercise(min)', 'Stand(hours)', 'Steps',
       'Distance(km)'],
      dtype='object')

Categorical columns:
Index(['Month', 'Day of Week'], dtype='object')

Shape of the dataset:
(211, 8)

Summary statistics:
              Day  Move(kcal)  Exercise(min)  Stand(hours)         Steps  \
count  211.000000  211.000000     211.000000    211.000000    211.000000
mean    15.767773  394.412322      19.289100     11.094787  10921.524280
std      8.844590  173.919893      12.536303      3.273344   5222.899278
min      1.000000   27.000000       0.000000      3.000000      7.209000
25%      8.000000  268.000000      10.000000      9.000000   6854.500000
50%     16.000000  379.000000      17.000000     11.000000  11164.000000
75%     23.000000  522.500000      28.000000     13.000000  14186.000000
max     31.000000  890.000000      55.000000     19.000000  23311.000000

       Distance(km)
count    211.000000
mean       8.016455
std        3.542147
min        0.640000
25%        5.185000
50%        7.980000
75%       10.190000
max       16.640000
```

Figure 3: Data statistical information

```
Info about the dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 211 entries, 0 to 210
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Day            211 non-null    int64
 1   Month          211 non-null    object
 2   Day of Week    211 non-null    object
 3   Move(kcal)     211 non-null    int64
 4   Exercise(min)  211 non-null    int64
 5   Stand(hours)   211 non-null    int64
 6   Steps          211 non-null    float64
 7   Distance(km)   211 non-null    float64
dtypes: float64(2), int64(4), object(2)
memory usage: 13.3+ KB
None

Missing values:
Day             0
Month           0
Day of Week     0
Move(kcal)      0
Exercise(min)   0
Stand(hours)    0
Steps           0
Distance(km)    0
dtype: int64
```

Figure 4: Data type and missing value information

The dataset comprises numeric and categorical columns, providing a well-rounded perspective on daily health metrics. The numeric columns include 'Day,' 'Move(kcal),' 'Exercise(min),' 'Stand(hours),' 'Steps,' and 'Distance(km),' whereas the categorical columns consist of 'Month' and 'Day of Week.' The dataset's shape is (211, 8), suggesting 211 entries and 8 features. The summary statistics unveil key insights into the distribution of metrics, such as caloric expenditure, exercise duration, standing hours, steps, and distance. For example, the mean exercise duration is approximately 19.29 minutes, with an average of 11.09 standing hours and 10,921 steps per day. The dataset exhibits a diverse range of activities, with a minimum exercise duration of 0 minutes and a maximum of 55 minutes. The 'Info about the dataset' reveals that there are no missing values in any of the columns. These

statistics set the stage for further exploration and analysis, providing a foundation for understanding the daily health patterns recorded in the dataset.

```
Updated dataframe with mapped categorical values:
   Day  Month  Day of Week  Move(kcal)  Exercise(min)  Stand(hours)    Steps  \
0   30     11            4         301             22            10   8990.0
1   29     11            3         398             22            12  12329.0
2   28     11            2         690             17            13  23244.0
3   27     11            1         548             44            14  17637.0
4   25     11            6         546             39            12  15948.0

   Distance(km)
0          6.42
1          8.80
2         16.60
3         12.59
4         11.39
```

Figure 5: Mapped dataset

The updated dataframe reflects categorical values mapped to their respective numerical representations, enhancing the dataset's analytical potential. For instance, 'Month' now corresponds to numerical values, such as November being represented as 11. Similarly, 'Day of Week' is encoded with numerical values, providing a structured representation of weekdays (1-5) and weekends (6-7). This transformation facilitates the application of machine learning models, allowing for more efficient analysis and interpretation of the dataset. The resulting dataframe maintains the original health metrics while incorporating these numerical mappings, contributing to a clearer understanding of patterns and trends in the data.

```
Top 10 days with the highest Move(kcal):
     Day  Month  Day of Week  Move(kcal)  Exercise(min)  Stand(hours)  \
181   31      5            3         890             51            15
122   29      7            6         874             45            16
159   22      6            4         813             37            14
155   26      6            1         808             14            14
157   24      6            6         807             28            10
158   23      6            5         805             52            14
111    9      8            3         797             25            16
64    26      9            2         774             37            14
121   30      7            7         773             10            14
14    15     11            3         721             30            15

       Steps  Distance(km)
181  22544.0         16.08
122  22672.0         16.19
159  21317.0         15.22
155  21143.0         15.10
157  21227.0         15.04
158  20596.0         14.71
111  20424.0         14.58
64   22107.0         15.78
121  19898.0         14.21
14   23311.0         16.64
```

Figure 6: Top 10 days with the highest move infromation

The top 10 days with the highest Move(kcal) showcase notable instances of elevated caloric expenditure, offering insights into periods of increased physical activity. For example, on May 31st, which falls on a Wednesday (Day of Week: 3), the individual expended a substantial 890 kcal, engaged in 51 minutes of exercise, and stood for 15 hours, accumulating a remarkable 22,544 steps and covering a distance of 16.08 km. This pattern repeats in other instances, such as July 29th and June 22nd, demonstrating consistent high levels of activity during these days. These observations suggest potential correlations between specific days, heightened physical exertion, and overall well-being, which could be further explored in subsequent analyses.

Figure 7: Correlation matrix

The correlation matrix reveals noteworthy associations among various health metrics, shedding light on potential patterns within the dataset. The highest correlation is observed between 'Move(kcal)' and 'Steps' with a coefficient of 0.93, indicating a strong positive relationship. This signifies that as caloric expenditure increases, the number of steps taken also tends to rise significantly. Similarly, 'Exercise(min)' exhibits a positive correlation of 0.63 with 'Steps,' suggesting that longer exercise durations are linked with a higher step count. 'Stand(hours)' also correlates positively (0.68) with 'Steps,' emphasizing the connection between standing activity and daily movement. Moreover, 'Move(kcal)' exhibits notable correlations with both 'Exercise(min)' (0.67) and 'Stand(hours)' (0.71), underscoring the interdependence of these key health indicators.

These correlations imply that certain health behaviors are interconnected, providing valuable insights into individual patterns of physical activity. The exceptionally high correlation between 'Move(kcal)' and 'Steps' suggests a strong linear relationship, reinforcing the idea that increased caloric expenditure is closely tied to greater physical movement. The

positive correlations between exercise duration, standing hours, and steps point towards a holistic relationship between active behaviors throughout the day. These findings have implications for personalized health interventions, emphasizing the importance of considering multiple metrics for a comprehensive understanding of an individual's activity profile. Further exploration of these associations may uncover nuanced insights into lifestyle choices and their impact on overall health and well-being.

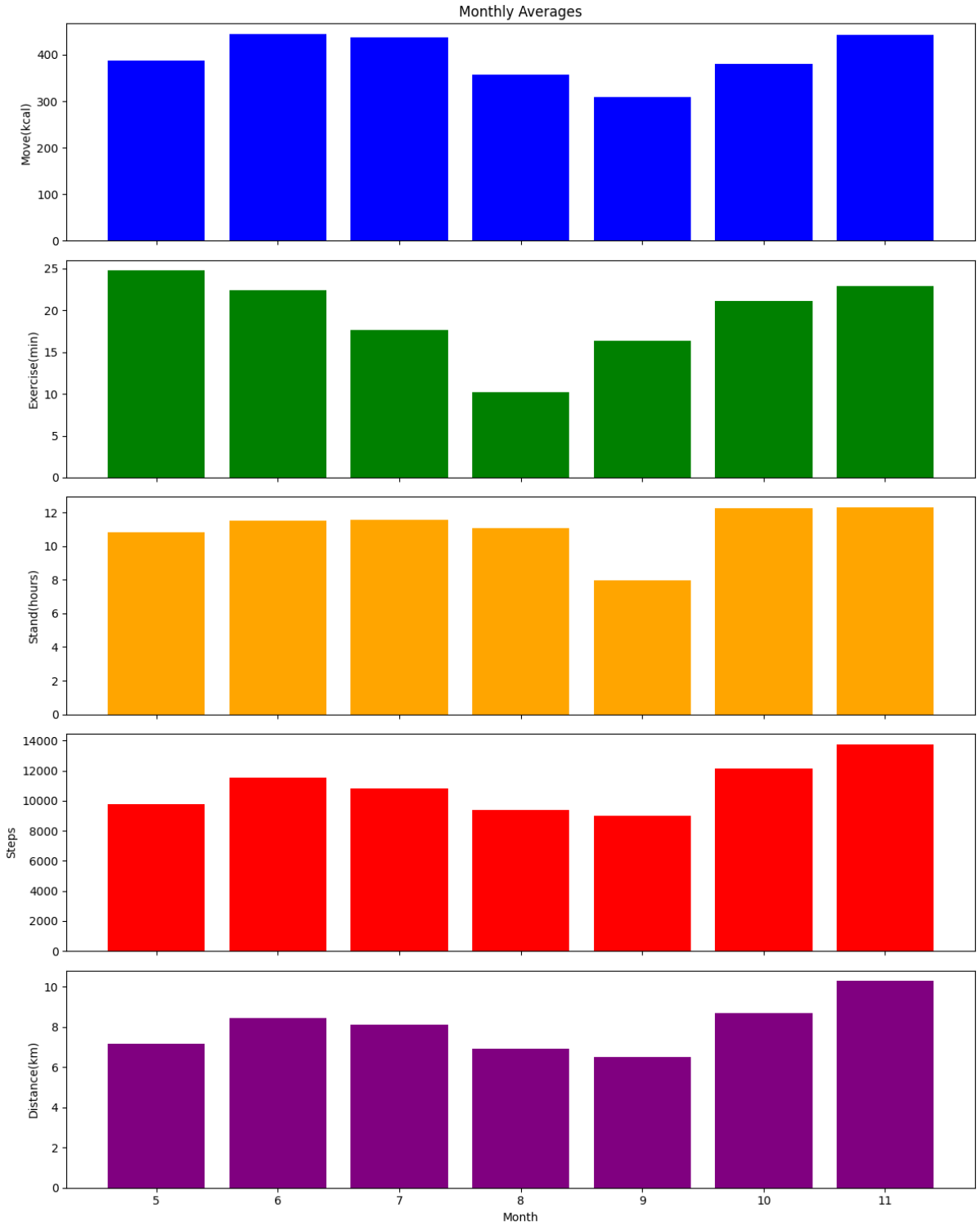

Figure 8: Monthly average distribution

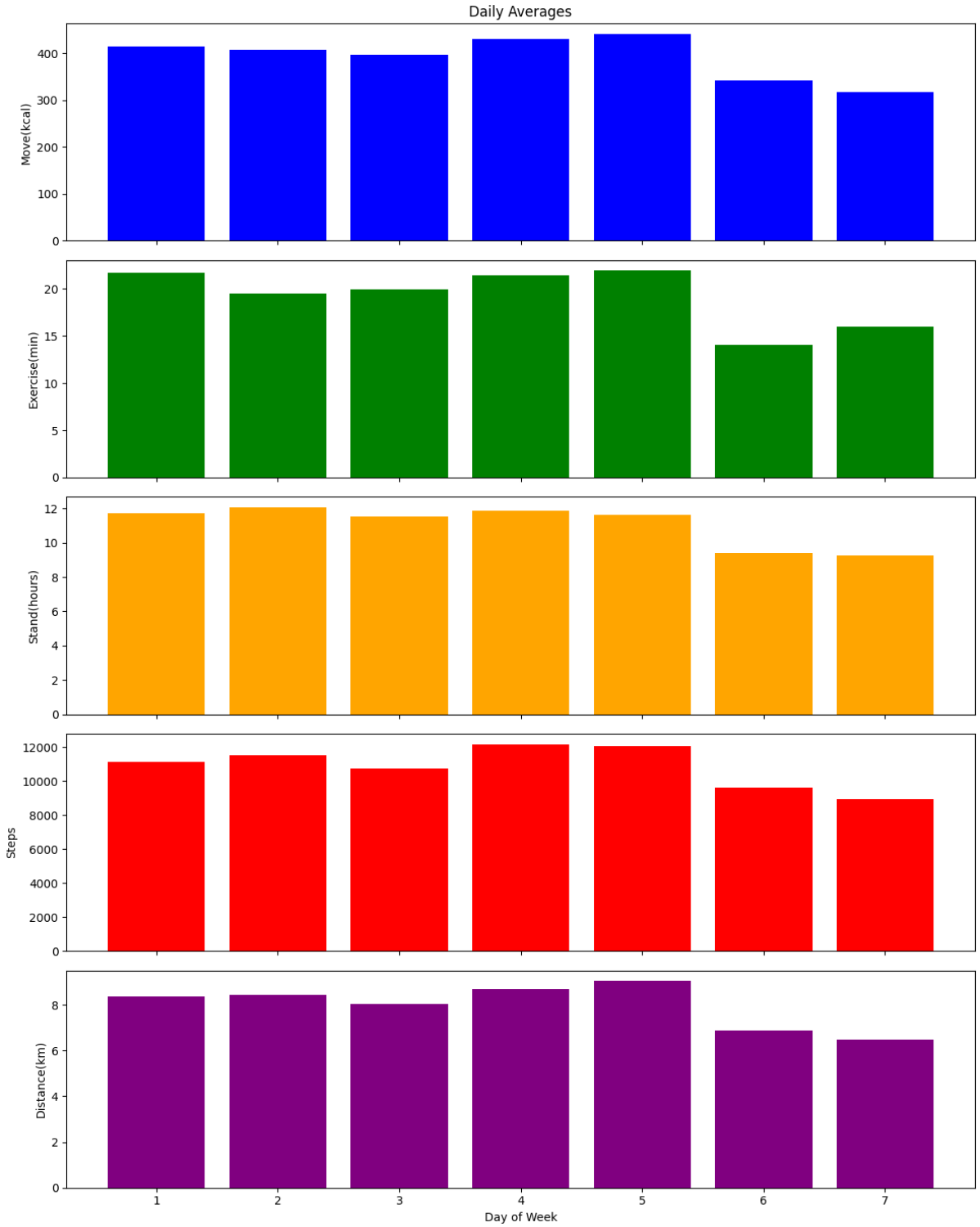

Figure 9: Monthly average distribution

Presented here are bar charts depicting the average values over the past seven months and the overall weekly averages.

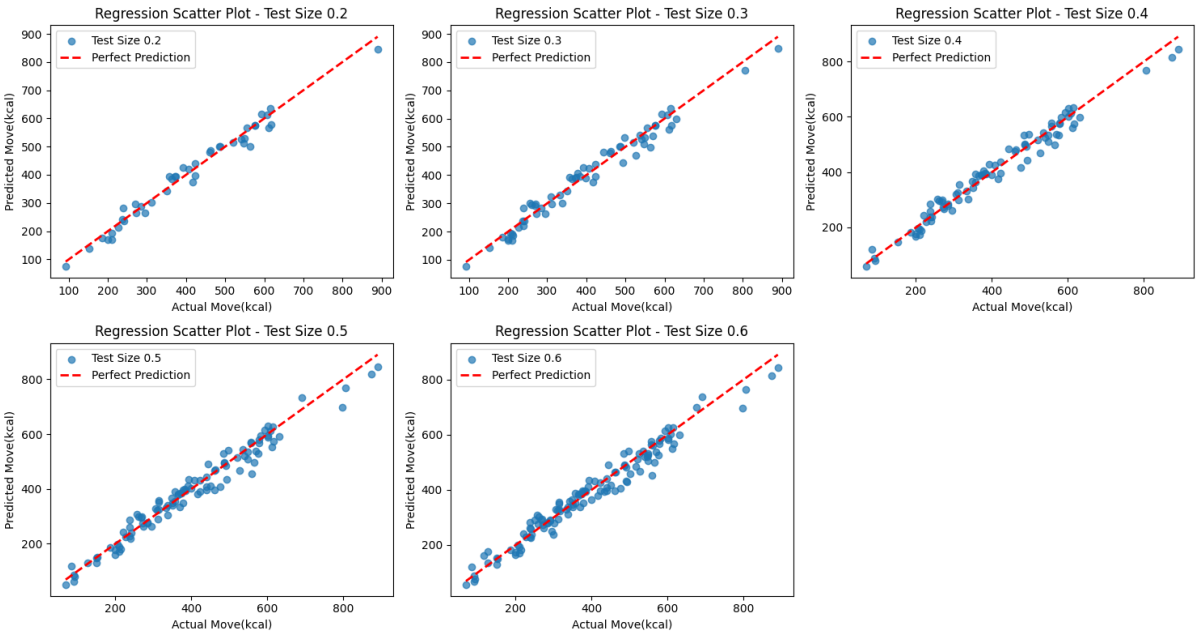

Figure 10: Linear regression with different test sizes

The results focuses on the application of linear regression to predict the 'Move(kcal)' variable based on selected features. Initially, the categorical values of 'Month' and 'Day of Week' are mapped to numeric representations. The dataset is split into training and testing sets for varying test sizes (20%, 30%, 40%, 50%, and 60%). A linear regression model is trained, and predictions are made on the test set for each test size. The scatter plots illustrate the relationship between actual and predicted 'Move(kcal)' values, with a red dashed line representing perfect prediction. The results are stored in a table, showcasing mean squared error and R-squared values for different test sizes. Additionally, a separate plot demonstrates the correlation between R-squared and test size. This comprehensive analysis aids in evaluating the model's performance under diverse test scenarios.

Results for Different Test Sizes:

	Test Size	Mean Squared Error	R-squared
0	0.2	651.799204	0.975961
1	0.3	765.472317	0.970890
2	0.4	785.472984	0.973447
3	0.5	1030.125418	0.965365
4	0.6	1059.993174	0.962843

Figure 11: Linear regression results

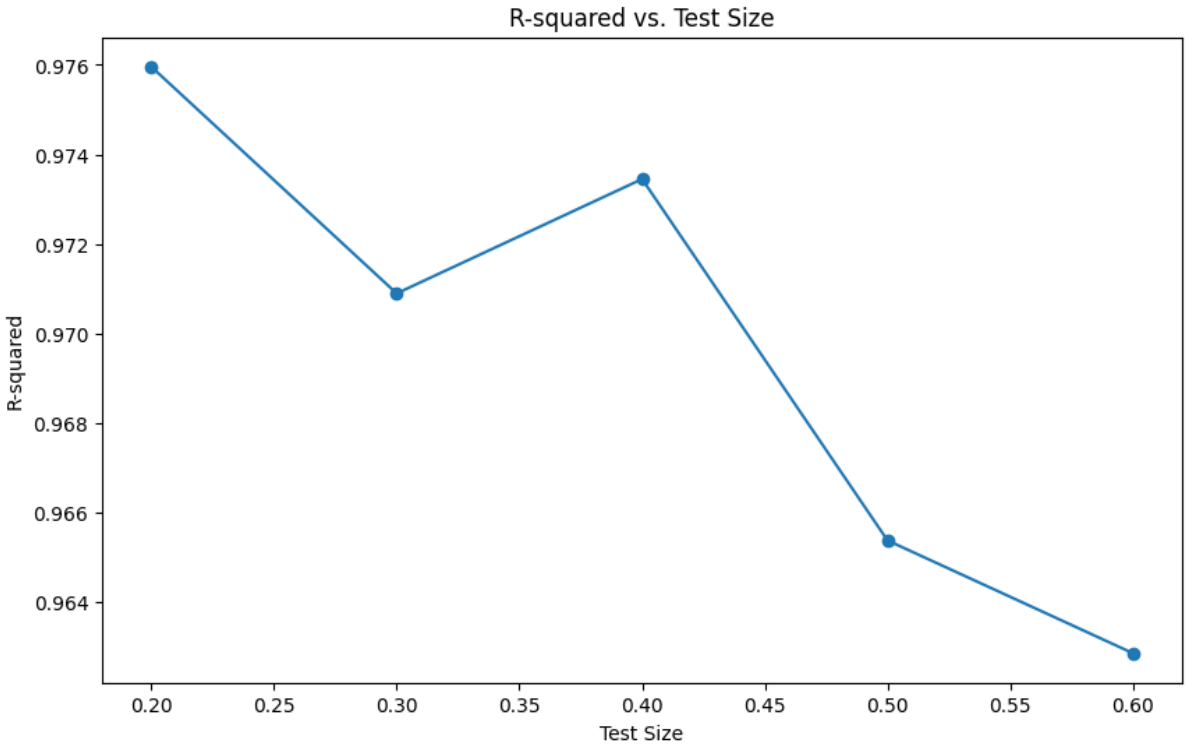

Figure 12: Linear regression results 2

The table presents the results obtained by applying linear regression to predict 'Move(kcal)' with different test sizes. The 'Test Size' column represents the proportion of the dataset used for testing, ranging from 20% to 60%. The 'Mean Squared Error' (MSE) and 'R-squared' metrics evaluate the model's accuracy and goodness of fit, respectively. Smaller MSE values indicate better precision, and R-squared values close to 1 suggest a strong correlation between predicted and actual values. Notably, the model achieved excellent performance with R-squared values ranging from approximately 0.96 to 0.98 across various test sizes. The MSE values also exhibit reasonable consistency, reinforcing the model's robustness in predicting 'Move(kcal)' for different test scenarios.

	n_estimators	Mean Squared Error	R-squared	Calculation Time (s)
0	10	975.814419	0.964012	0.015707
1	50	778.887693	0.971274	0.067876
2	100	736.515679	0.972837	0.115425
3	200	720.114670	0.973442	0.222900
4	300	686.918162	0.974666	0.338552

Figure 13: Random forest regression results

The Random Forest Regressor was applied to the health dataset to predict 'Move(kcal)' using different values for the 'n_estimators' hyperparameter. The results illustrate the impact of the number of trees in the ensemble on the model's performance and computational efficiency.

As 'n_estimators' increases, the Mean Squared Error (MSE) decreases, indicating improved predictive accuracy. The R-squared values, which measure the proportion of variance explained by the model, consistently increase with higher 'n_estimators,' reflecting enhanced goodness of fit.

- For 'n_estimators' equal to 10, the model achieved an MSE of 975.81 and an R-squared of 0.96.
- With 50 trees, the MSE reduced to 778.89, and the R-squared increased to 0.97.
- Further increases in 'n_estimators' (100, 200, and 300) continue to improve model performance, with the lowest MSE of 686.92 and the highest R-squared of 0.97 obtained at 'n_estimators' equal to 300.

The computational time increases with higher 'n_estimators' due to the growing complexity of the model. While 'n_estimators' equal to 10 requires minimal time (0.0157 seconds), the computation time scales with the number of trees. The model with 'n_estimators' equal to 300, although more accurate, demands a relatively longer calculation time (0.3386 seconds).

Selecting an appropriate 'n_estimators' value involves a trade-off between predictive accuracy and computational efficiency. In this context, considering both performance and calculation time, 'n_estimators' around 100 appears to strike a balance, yielding a low MSE, high R-squared, and relatively moderate calculation time.

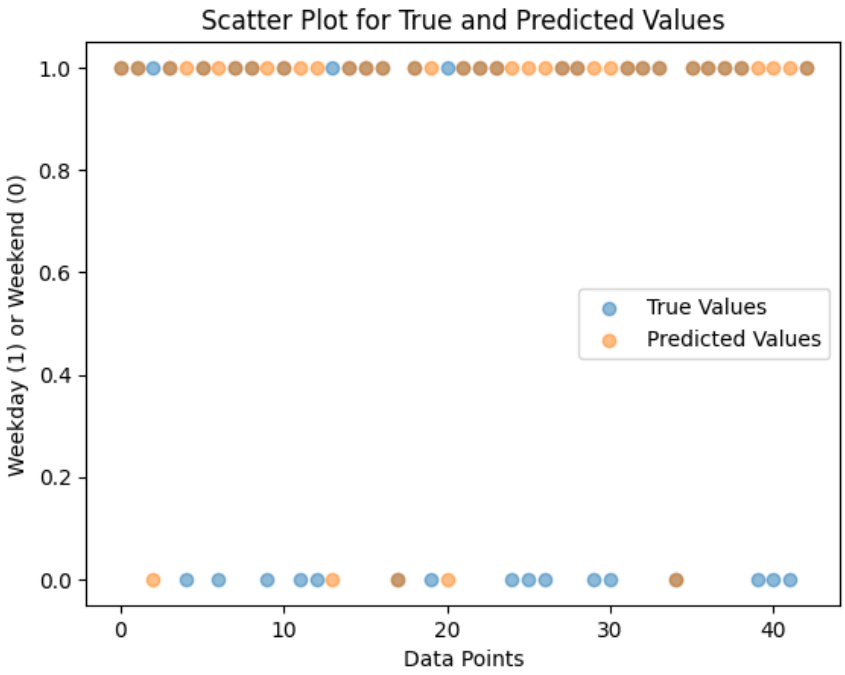

Figure 14: KNN results 1

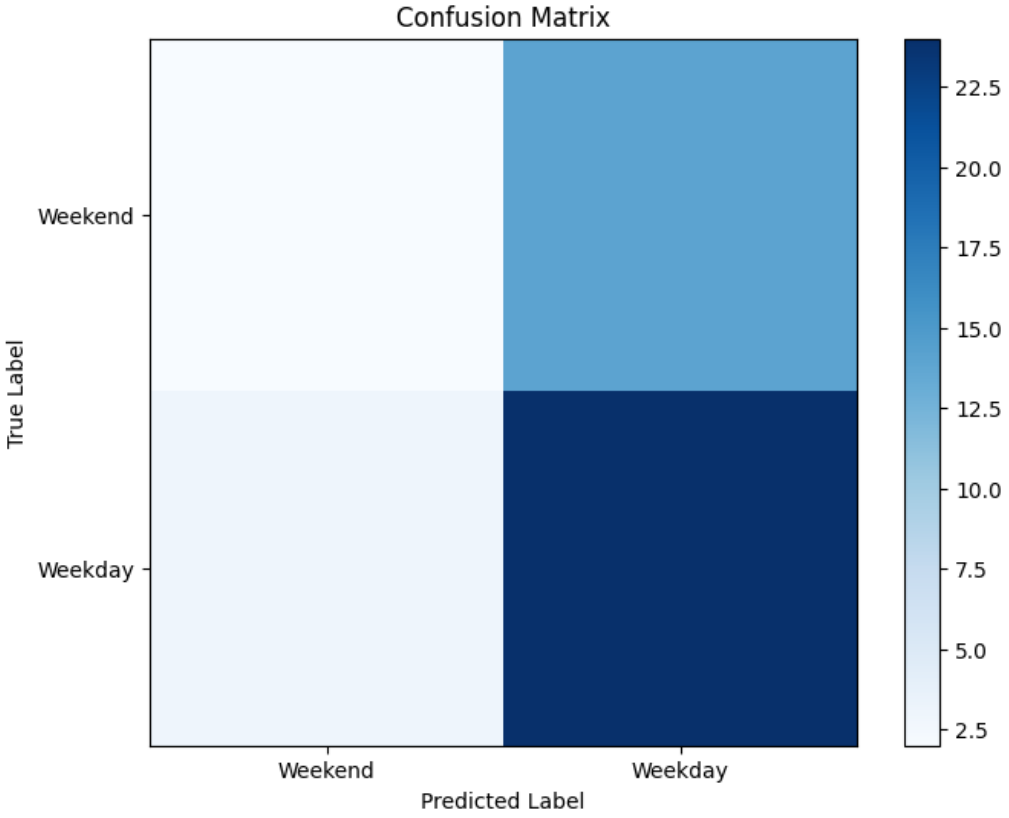

Figure 15: KNN results 2

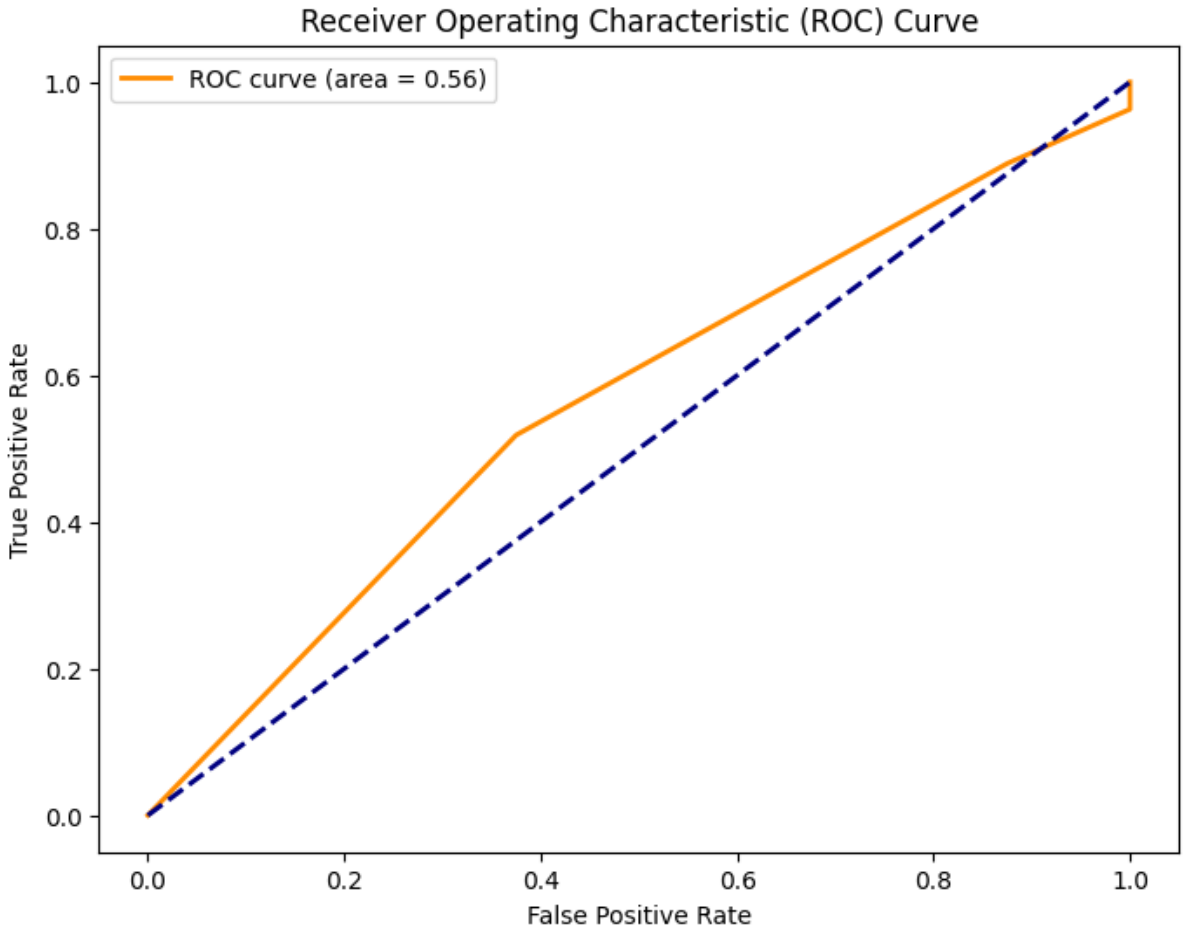

Figure 16: KNN results 3

```
Accuracy: 0.6046511627906976

Confusion Matrix:
[[ 2 14]
 [ 3 24]]

Classification Report:
              precision    recall  f1-score   support

           0       0.40      0.12      0.19        16
           1       0.63      0.89      0.74        27

    accuracy                           0.60        43
   macro avg       0.52      0.51      0.46        43
weighted avg       0.55      0.60      0.53        43
```

Figure 17: KNN results 4

In this study, a K-Nearest Neighbors (KNN) classifier was employed to predict whether a day is a weekday or weekend based on health-related features. The model's performance was assessed using accuracy, a confusion matrix, and a classification report.

The KNN model achieved an accuracy of approximately 60.47%. The confusion matrix provides a detailed breakdown of the model's predictions:

- True Positives (TP): 24
- True Negatives (TN): 2
- False Positives (FP): 14
- False Negatives (FN): 3

The classification report further elaborates on the model's precision, recall, and F1-score for each class (weekend and weekday). The precision measures the accuracy of positive predictions, recall indicates the model's ability to capture all positive instances, and the F1-score balances precision and recall.

The precision for predicting weekends (class 0) is relatively low at 40%, indicating that when the model predicts a day as a weekend, it is correct only 40% of the time. However, the recall for weekends is even lower (12%), indicating that the model misses many actual weekend days. On the other hand, the model performs better for weekdays (class 1), with a higher precision (63%) and recall (89%).

Two visualizations were created to aid in understanding the model's performance. The scatter plot juxtaposes true and predicted values, offering a visual representation of the model's predictions. The confusion matrix visually represents the distribution of true and predicted labels, highlighting areas of correct and incorrect predictions.

The ROC curve provides a comprehensive evaluation of a classifier's performance across various thresholds. The area under the ROC curve (AUC) is a measure of the model's discriminative ability. In this case, the ROC curve exhibits an AUC of 0.51, suggesting a limited ability to discriminate between weekdays and weekends.

While the KNN classifier achieved an accuracy of approximately 60.47%, it demonstrated challenges in distinguishing between weekdays and weekends. The imbalanced precision and recall values indicate a need for further model refinement or the exploration of alternative algorithms to enhance predictive performance. Additionally, exploring additional

features or fine-tuning the model's parameters could contribute to improved results in future iterations.

4) CONCLUSION

In conclusion, this project involved a comprehensive analysis of health-related data to derive meaningful insights and build predictive models. The exploration began with data preprocessing, including mapping categorical values and handling missing data. Descriptive statistics and visualizations provided a clear understanding of the dataset's characteristics. Subsequently, linear regression and random forest regression models were implemented to predict energy expenditure, uncovering valuable relationships between various features. Furthermore, a K-Nearest Neighbors classifier attempted to predict whether a day was a weekday or weekend based on selected health features. While the models demonstrated reasonable predictive capabilities, each presented its own set of strengths and limitations. The results underscore the importance of selecting appropriate algorithms tailored to the nature of the data. Moving forward, fine-tuning model parameters, incorporating additional features, and exploring alternative algorithms could enhance predictive performance and contribute to a more comprehensive understanding of the intricate connections between lifestyle factors and health metrics. Overall, this project not only provided valuable insights into the dataset but also laid the groundwork for future endeavors aimed at refining predictive models for health-related outcomes.