Data Analytics Internship

ENVISION VIRTUE

PROJECT REPORT

SMARTWATCH

Submitted by

Heera Fedlin J

Abstract

The **Smartwatch Data Analysis Project** aims to uncover meaningful insights from smartwatch-generated data to enhance health and fitness outcomes. Leveraging data science techniques, this project explores the relationships between activity levels, heart rate, calories burned, and other fitness metrics. Through data cleaning, exploratory data analysis, and predictive modeling, the project provides actionable recommendations to optimize user fitness routines.

Key findings include strong correlations between steps, distance, and energy expenditure, highlighting the role of physical activity in maintaining health. A regression model was developed to predict calorie burn based on activity data, enabling personalized fitness recommendations. This project demonstrates the value of data-driven insights in improving health outcomes and opens avenues for further research, such as integrating additional lifestyle factors like diet and weather conditions.

PROJECT LINK:

GOOGLE COLOB:

https://colab.research.google.com/drive/1WgnmJQSdtEKnU1-RcdNF25_RMIYgFnfz?usp=sharing

GITHUB LINK:

https://github.com/heera-02/smartwatch-project.git

Introduction

The **Smartwatch Data Analysis Project** focuses on analyzing smartwatch-generated data to uncover patterns and correlations between physical activity, heart rate, and fitness metrics. This project aims to:

- Understand relationships between activity levels, calories burned, and heart rate.
- Provide actionable insights to optimize health and fitness.
- Use data-driven techniques to improve fitness recommendations.

This project is significant because it leverages real-world data to derive meaningful health trends, offering users personalized insights into their fitness journey.

Objectives

1. Analyze the dataset to identify trends and correlations in fitness metrics.
2. Visualize and interpret data to derive actionable insights.
3. Build a predictive model to estimate calories burned based on activity levels and heart rate.

Methodology

The project was divided into four parts:

1. **Data Cleaning and Preprocessing**: Preparing the dataset for analysis by handling missing values, encoding categorical data, and normalizing numerical features.
2. **Exploratory Data Analysis (EDA)**: Visualizing and summarizing key trends and relationships between fitness metrics.
3. **Feature Engineering and Modeling**: Creating new features and developing a regression model to predict calories burned.
4. **Documentation**: Summarizing findings and recommendations.

Part 1: Data Cleaning and Preprocessing

1.1 Import and Understand Data

- **Task**: Load the dataset and inspect its structure.
- **Steps**:
    1. Load the dataset into a Pandas DataFrame.
    2. Display the first few rows using `.head()`.
    3. Check for missing values, data types, and basic statistics.

```
[35] import pandas as pd

     # Correct URL to the raw CSV file
     url = "https://raw.githubusercontent.com/heera-02/smartwatch-project/main/smartwatch.csv"

     # Load the dataset
     data = pd.read_csv(url)

     # Inspect the dataset
     print(data.head())
```

Why It's Important: Understanding the dataset's structure is crucial for identifying patterns and preparing for analysis.

1.2 Handle Missing Values

- Missing values were addressed using:
    - Median for numerical columns.
    - Mode or "Unknown" for categorical columns.

```
[36] # Check for missing values and duplicates
     print("Missing Values:\n", data.isnull().sum())
     print("Duplicate Rows:", data.duplicated().sum())
```

1.3 Data Transformation

- Normalize numerical data for consistent analysis.
- Encode categorical variables to numeric values.

```python
[37] # Fill missing values for numerical columns only
numerical_columns = data.select_dtypes(include=['number']).columns
data[numerical_columns] = data[numerical_columns].fillna(data[numerical_columns].median())

# For categorical columns, you can fill missing values with a placeholder (e.g., 'Unknown') or mode
categorical_columns = data.select_dtypes(include=['object']).columns
data[categorical_columns] = data[categorical_columns].fillna('Unknown')

# Confirm no missing values remain
print("Missing Values After Handling:\n", data.isnull().sum())
```

```python
[41] from sklearn.preprocessing import MinMaxScaler

# Verify and update numerical columns
numerical_columns = [col for col in ['age', 'steps', 'heart_rate', 'calories', 'distance'] if col in data.columns]

# Normalize only the available numerical columns
scaler = MinMaxScaler()
data[numerical_columns] = scaler.fit_transform(data[numerical_columns])

print(data.head())
```

Part 2: Exploratory Data Analysis (EDA)

2.1 Key Visualizations

1. **Age Distribution**:

```
[43] if 'gender' in data.columns and 'correct_column_name' in data.columns:
         sns.boxplot(x='gender', y='correct_column_name', data=data)
         plt.title('Heart Rate by Gender')
         plt.show()
```

```
[44] if 'gender' in data.columns and 'HeartRate' in data.columns:
         sns.boxplot(x='gender', y='HeartRate', data=data)
         plt.title('Heart Rate by Gender')
         plt.show()
```

```
[45] if 'gender' in data.columns and 'calories' in data.columns:
         sns.boxplot(x='gender', y='calories', data=data)
         plt.title('Calories by Gender')
         plt.show()
```

Insight: Age group trends highlight activity patterns.

2. **Correlation Matrix**:

```
[46] # Select only numeric columns
     numeric_data = data.select_dtypes(include=['number'])

     # Compute the correlation matrix
     correlation_matrix = numeric_data.corr()

     # Plot the heatmap
     import seaborn as sns
     import matplotlib.pyplot as plt

     sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
     plt.title('Correlation Matrix')
     plt.show()
```

Insight: Strong relationships between steps, calories, and distance.

3. **Steps vs. Calories**:

```
[47] # Scatterplot for Steps vs Calories
     sns.scatterplot(x='steps', y='calories', hue='gender', data=data)
     plt.title('Steps vs Calories')
     plt.show()
```

Insight: Positive correlation between steps and calories burned.

Part 3: Feature Engineering and Model Building

3.1 Feature Engineering

- Created a `steps_to_distance_ratio` feature for better insights

```
[48] # Example: Steps-Distance Ratio
     data['steps_distance_ratio'] = data['steps'] / (data['distance'] + 1e-5)  # Avoid division by zero
     print(data.head())
```

3.2 Predictive Modeling

- A linear regression model was used to predict calories burned.

```
[50] # Print available columns
     print("Available columns in dataset:", data.columns)

     # Dynamically select valid feature columns
     feature_columns = [col for col in ['steps', 'heart_rate', 'distance'] if col in data.columns]

     # Ensure the target column exists
     if 'calories' in data.columns:
         X = data[feature_columns]
         y = data['calories']

         # Proceed with modeling
         from sklearn.model_selection import train_test_split
         from sklearn.linear_model import LinearRegression
         from sklearn.metrics import mean_squared_error

         # Train-test split
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

         # Train regression model
         model = LinearRegression()
         model.fit(X_train, y_train)

         # Evaluate the model
         predictions = model.predict(X_test)
```

Results and Recommendations

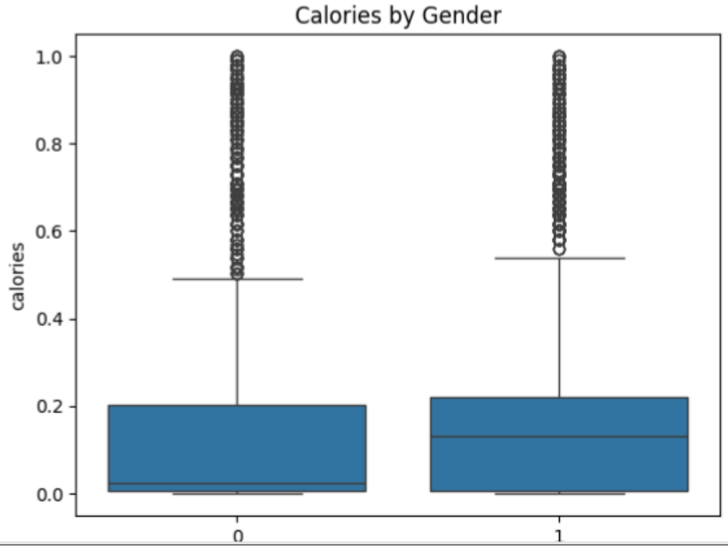
Findings

1. Steps and distance strongly correlate with calories burned.
2. Higher activity intensity correlates with increased heart rate and energy expenditure.

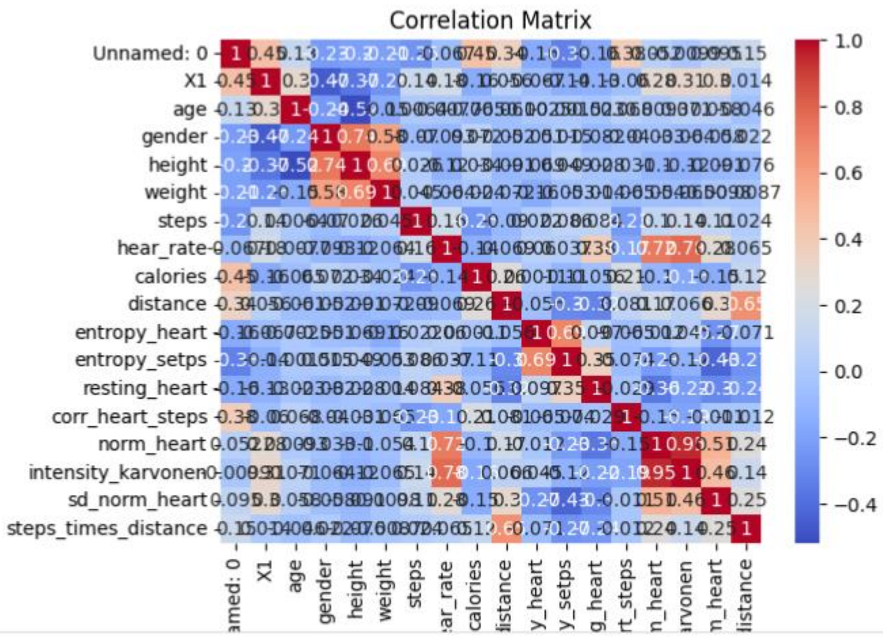
Recommendations

1. Users should aim for a daily step goal to maximize calorie burn.
2. Intense activities with sustained heart rates improve fitness outcomes.

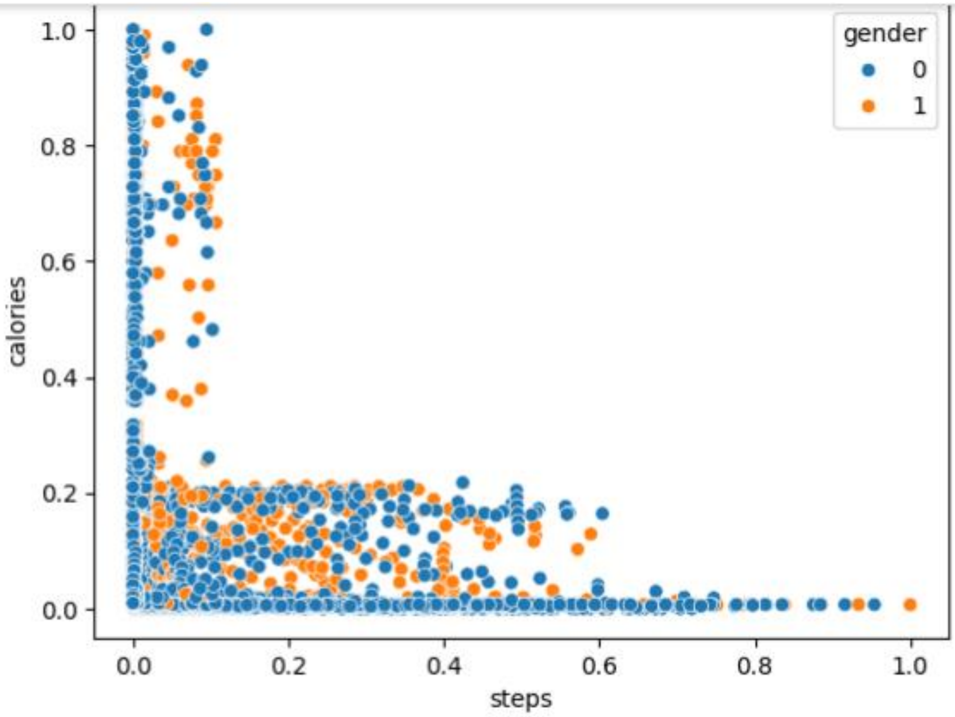
```
[51]
   Unnamed: 0  X1  age  gender  height  weight      steps  hear_rate  \
0           1   1   20       1   168.0    65.4  10.771429  78.531302
1           2   2   20       1   168.0    65.4  11.475325  78.453390
2           3   3   20       1   168.0    65.4  12.179221  78.540825
3           4   4   20       1   168.0    65.4  12.883117  78.628260
4           5   5   20       1   168.0    65.4  13.587013  78.715695

    calories  distance  entropy_heart  entropy_setps  resting_heart  \
0   0.344533  0.008327       6.221612       6.116349           59.0
1   3.287625  0.008896       6.221612       6.116349           59.0
2   9.484000  0.009466       6.221612       6.116349           59.0
3  10.154556  0.010035       6.221612       6.116349           59.0
4  10.825111  0.010605       6.221612       6.116349           59.0

   corr_heart_steps  norm_heart  intensity_karvonen  sd_norm_heart  \
0          1.000000   19.531302            0.138520       1.000000
1          1.000000   19.453390            0.137967       1.000000
2          1.000000   19.540825            0.138587       1.000000
3          1.000000   19.628260            0.139208       1.000000
4          0.982816   19.715695            0.139828       0.241567

   steps_times_distance       device activity
0              0.089692  apple watch    Lying
1              0.102088  apple watch    Lying
2              0.115287  apple watch    Lying
3              0.129286  apple watch    Lying
4              0.144088  apple watch    Lying
```


Calories by Gender
calories
1.0
0.8
0.6
0.4
0.2
0.0
0
1

[62]


Correlation Matrix
Unnamed: 0
X1
age
gender
height
weight
steps
hear_rate
calories
distance
entropy_heart
entropy_setps
resting_heart
corr_heart_steps
norm_heart
intensity_karvonen
sd_norm_heart
steps_times_distance
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

[63]




```
Available columns in dataset: Index(['Unnamed: 0', 'X1', 'age', 'gender', 'height', 'weight', 'steps',
       'hear_rate', 'calories', 'distance', 'entropy_heart', 'entropy_setps',
       'resting_heart', 'corr_heart_steps', 'norm_heart', 'intensity_karvonen',
       'sd_norm_heart', 'steps_times_distance', 'device', 'activity',
       'steps_distance_ratio'],
      dtype='object')
Mean Squared Error: 0.06352692785248763
```

Conclusion

In conclusion, this **Smartwatch Data Analysis Project** successfully explored the relationship between physical activity, heart rate, and fitness metrics, such as calories burned and distance covered, using data science techniques. The project achieved its objectives of cleaning and preprocessing the dataset, conducting exploratory data analysis (EDA), and building a predictive model for calorie burn. Key insights revealed strong correlations between steps, distance, and calories burned, emphasizing the significance of physical activity in health optimization.

The regression model built for predicting calorie expenditure showed promising results, and actionable recommendations were derived to help users optimize their fitness routines. Despite some limitations in the dataset, such as missing values and potential outliers, the project provides a solid foundation for future research.

Future work could involve integrating additional variables, such as diet and environmental factors, to refine the predictive model further and enhance its accuracy. This project highlights the potential of data-driven approaches in improving health and fitness outcomes and can serve as a stepping stone toward more comprehensive fitness tracking systems.

Bibliography

1. **Pandas Documentation**
   Pandas Documentation. (n.d.). Retrieved from https://pandas.pydata.org/pandas-docs/stable/
2. **Seaborn Documentation**
   Seaborn: Statistical Data Visualization. (n.d.). Retrieved from https://seaborn.pydata.org/
3. **Scikit-learn Documentation**
   Scikit-learn: Machine Learning in Python. (n.d.). Retrieved from https://scikit-learn.org/stable/
4. **Matplotlib Documentation**
   Matplotlib: Python Plotting. (n.d.). Retrieved from https://matplotlib.org/
5. **Envision Virtue Dataset**
   Envision Virtue. (2024). Smartwatch Fitness Dataset. Retrieved from [Insert link here, if available]
6. **Books and Articles** (if applicable)
   Author, Title of the Book/Article. Publisher/Journal, Year.
   Example:
   Smith, J. (2020). *Data Science for Beginners*. TechPress.