Human Activity Prediction based on sensor Data

SONGA AKHIL, SAI VASAVI HARSHAVARDHNA GUPTA SOMISETTY, SRI TEJA KUMAR REDDY TETALI
College of Computing & Informatics, Drexel University, Philadelphia, PA 19104, USA

Abstract - *Data science and machine learning researchers are working to solve the important problem of human activity recognition (HAR), which has numerous applications in the fields of health, fitness, and wellness. In order to develop a precise classification model using machine learning algorithms, this project aims to investigate the use of smartphone sensor data for HAR. In this project, raw sensor data from a smartphone's accelerometer and gyroscope are processed to create the Human Activity Recognition with Smartphones database. Data preparation, feature selection, model optimization, model validation and testing, and deployment of the finished model are all included in the project. This project's findings demonstrate that machine learning algorithms can categorize human activity from smartphone sensor data, making it a promising method for HAR in practical applications. The project has important ramifications for healthcare, fitness, and wellness, where precise tracking of human activity can yield individualized advice and enhance health outcomes.*

1.INTRODUCTION

The dataset titled "Physical Activity Prediction," available on the UCI Machine Learning Repository, provides a collection of physiological and movement data of individuals recorded through wearable sensors. The dataset aims to make it possible to create predictive models for each person's level of physical activity. Accurate prediction of activity levels is crucial for tailoring health interventions because physical activity has been shown to have a significant impact on overall health and well-being. The dataset includes data from different sensors, including accelerometer and gyroscope measurements, as well as demographic data. The target variable in this dataset is the level of physical activity, which is divided into five categories ranging from very active to sedentary. The dataset can be used in a number of ways, such as to make activity tracking systems, tailor-made fitness plans, and find risk factors for different illnesses.

2. Dataset

The mHealth (Mobile Health) dataset is a publicly available dataset from the Machine Learning Repository of the University of California, Irvine (UCI). The dataset was collected as part of a research project aimed at developing machine learning algorithms for human activity recognition using mobile sensors. The dataset contains sensor readings from 10 different sensors on a smartphone and a smartwatch worn by the participants, while they performed different physical activities. The mHealth dataset contains data from 10 participants, who were asked to perform 12 different physical activities, including walking, standing, jumping, and cycling. The dataset includes a total of 27,000 instances, with each instance consisting of 23 sensor readings and a label that indicates the corresponding activity. The sensor data was collected at a sampling rate of 50 Hz, resulting in a total of 23 sensor features for each instance. The sensors included in the dataset are the accelerometer, gyroscope, and magnetometer of the smartphone and smartwatch. The data was preprocessed to remove noise and the effect of gravity, resulting in a clean and standardized dataset for machine learning analysis. The targeted variable in the dataset holds the values of the general activities like standing still, lying down, walking, climbing stairs, cycling, jogging or running and many more.

Below are the all available features that are present in the dataset

Feature	Description
alx	acceleration from the left-ankle sensor (X axis)
aly	acceleration from the left-ankle sensor (Y axis)
alz	acceleration from the left-ankle sensor (Z axis)
glx	gyro from the left-ankle sensor (X axis)
gly	gyro from the left-ankle sensor (Y axis)
glz	gyro from the left-ankle sensor (Z axis)
arx	acceleration from the right-lower-arm sensor (X axis)
ary	acceleration from the right-lower-arm sensor (Y axis)
arz	acceleration from the right-lower-arm sensor (Z axis)
grx	gyro from the right-lower-arm sensor (X axis)
gry	gyro from the right-lower-arm sensor (Y axis)
grz	gyro from the right-lower-arm sensor (Z axis)
Activity	Activity that the person is currently engaged in

The Activity feature in the dataset holds the below values.

Value	Meaning
0	None
1	Standing still
2	Sitting and relaxing
3	Lying down
4	Walking
5	Climbing stairs
6	Waist bends forward
7	Frontal elevation of arms
8	Knees bending
9	Cycling
10	Jogging
11	Running
12	Jump front & back

3. Targeted Audience

Researchers and academics working in the fields of data science, machine learning, and human-computer interaction may find the project useful. The outcomes and learnings from this project can be used to develop and enhance machine learning algorithms in this field, as well as to better understand how to recognize human activity from smartphone sensor data. Mobile application developers: The project may be useful for those who create fitness- and health-related mobile applications. The outcomes of this project can be used to create apps that monitor a user's physical activity and make tailored suggestions for workout plans and lifestyle adjustments. Healthcare Providers: Healthcare providers looking to create digital health solutions may find the project useful. The results of this project can be used to make apps that track patients' physical activity and give them personalized tips for becoming more active and making other changes to their lives. Individuals: The project can be useful for people who want to keep track of their fitness objectives and monitor their physical activity. The outcomes of this project can be used to create apps that, based on a person's physical activity data, recommend customized workout plans and lifestyle changes.

4. Exploratory Data Analysis

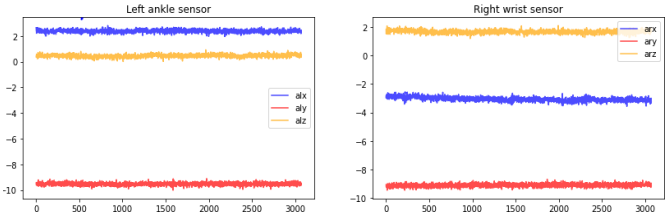

Fig 1.1 Line Graph of Sitting and relaxing, Accelerometer

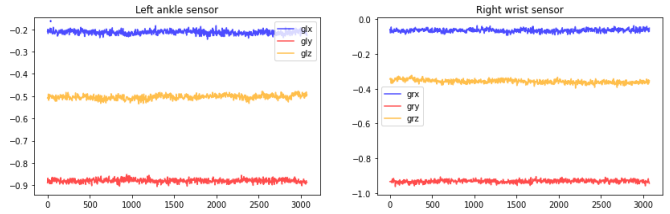

Fig 1.2 Line Graph of Sitting and relaxing, Gyroscope

When we looked at the line graph (Fig 1.1 and 1.2) of the activities for multiple subjects, we saw that the accelerometer and gyroscope data for the "Sitting and relaxing" activity showed little to no movement or activity. However, when the subjects were walking, we noticed that the accelerometer data from the left ankle sensor showed significantly more movement recorded on the z and y-axis compared to the x-axes. When we observed the right wrist sensor, its y axis recorded the most among others. On the other hand, the gyroscope data showed oscillatory motion, with most of the movement recorded on the z-axis, the next peak at the y-axis, and a little bit of oscillation found on the x-axis as well. Interestingly, the z-axis remained constant throughout the walking activity.

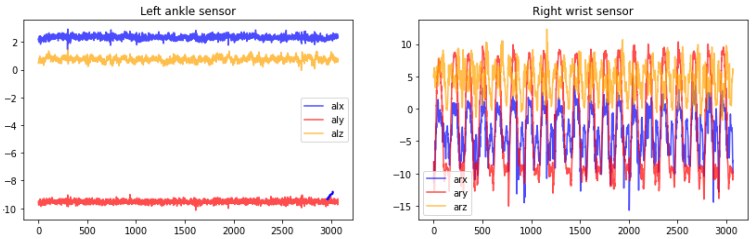

Fig 2.1 Line Graph Frontal Elevation of Arms, Accelerometer

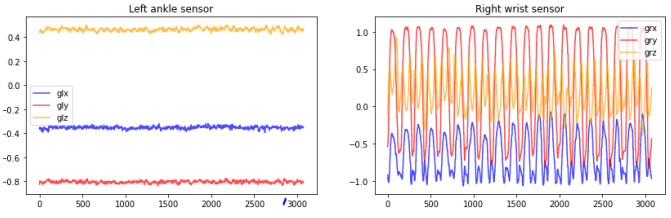

Fig 2.2 Line Graph of Frontal Elevation of Arms, Gyroscope

We observed interesting patterns in the plot for the frontal elevation of arm activity (Fig 2.1, 2.2). The plot showed little to no movement along the y-axis, but significant movement was observed along the x and z-axes. When we looked at the data from the right wrist sensor, we found that a lot more data was recorded here compared to the leg sensor. Specifically, the z-axis had more peaks, the y-axis had fewer peaks, and the x-axis was in the middle. We also analyzed the gyroscope data and found that there was a two-phase oscillatory motion on the x, y and z-axes for the leg sensor. For the right wrist sensor, we observed huge changes in the data with oscillatory patterns, indicating that the subject was performing the frontal elevation of arms activity.

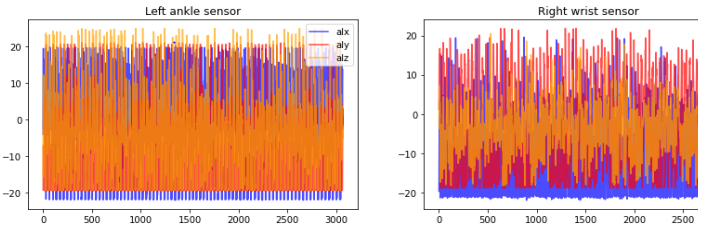

Fig 3.1 Line Graph of Running, Accelerometer

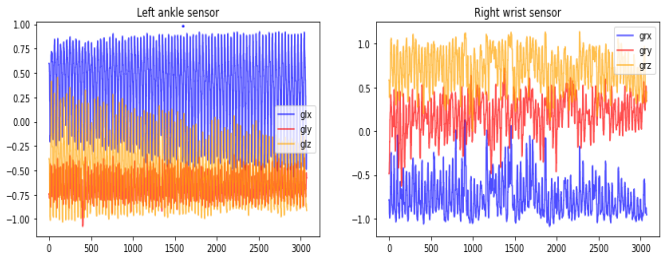

Fig 3.2 Line Graph of Running, Gyroscope

When we analyzed the graph for the running activity, we observed interesting patterns. The left leg sensor recorded significant changes in the data, with all three axes (x, y, and z) showing large variations. On the other hand, the right wrist sensor showed oscillatory motion, indicating the subject was running. We also analyzed the gyroscope data and found that there was significant motion recorded. The z-axis had the most variation, with short valleys and higher peaks, which further confirmed that the subject was performing the running activity.

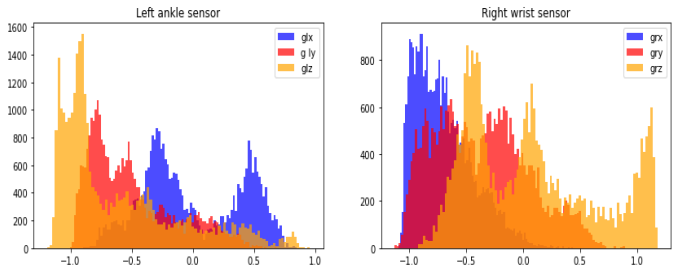

Fig 4.1 Distribution Graph of Running

By analyzing the distribution of all the activities across multiple subjects, we gain insights into the possible values and how frequently they occur. This information is useful in understanding the patterns and trends in the data and can help us make informed decisions during the modeling process. And Most of the data seems to have a near to normal distribution (bell-curve), which is really advantageous for developing an efficient model to predict the human activity

5. Methodology

5.1 Data Cleaning and Resampling:

Before conducting exploratory data analysis (EDA), we analyzed the data and found some imbalances. For example, there were 80,000 observations for Activity 0, while all other activities had an average of only 30,000 observations. To address this imbalance, we resampled Activity 0 to 40,000 observations to ensure that our model isn't skewed toward Activity 0.

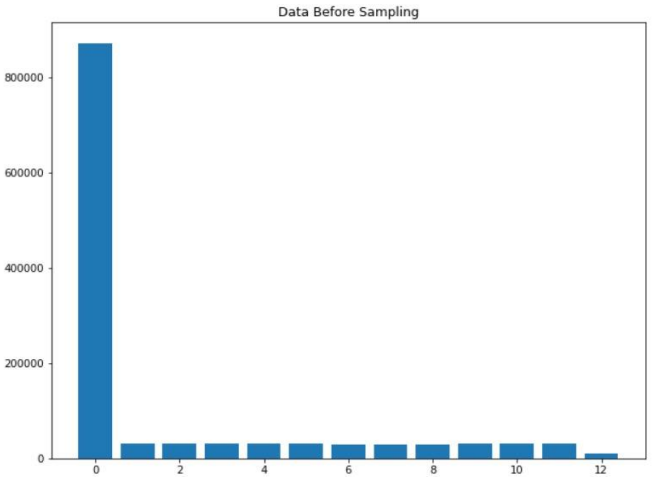

Fig 5.1 Bar Graph representation of data before sampling

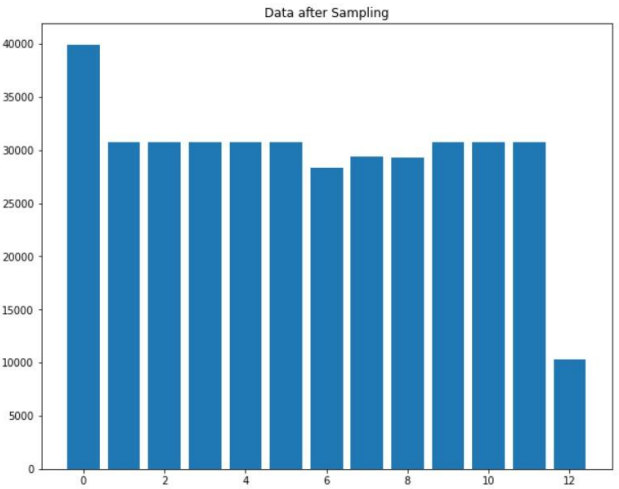

Fig 5.2 Bar Graph representation of data after sampling

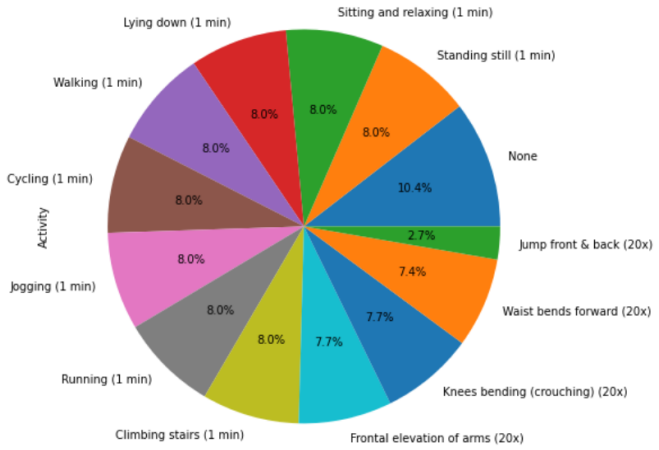

Fig 5.3 Pie chart representation of data after sampling

5.2 Feature Selection:

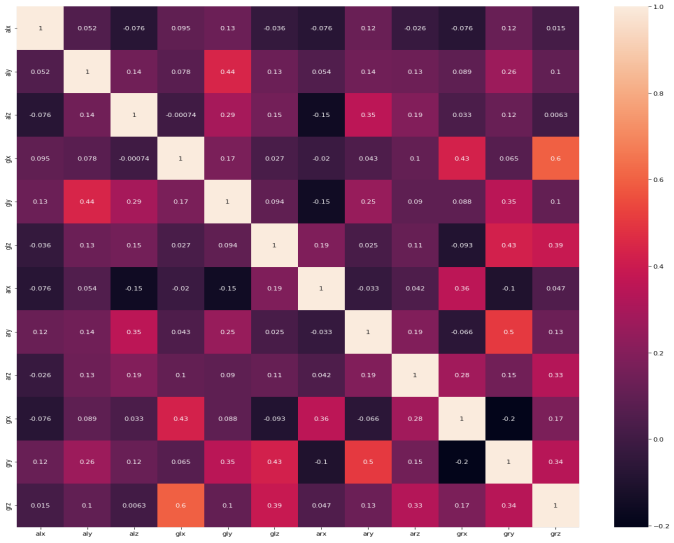

Fig 6 Heat map representing correlation among features

To visualize the correlation between features and activities, we used a heatmap. Lastly, to improve the quality of the data, we got rid of features that had data that was outside of the 98% confidence interval.

5.3 Machine Learning Modeling

To enhance the models, we implemented PySpark pipelines
that included a vector assembler, a standard scaler, and the chosen model. We selected two models for our analysis: Logistic Regression, which uses a linear approach to model the relationship between the dependent variable and independent variables. Decision Tree, which employs a tree-like model of decisions and their possible consequences to classify data. By using these pipelines, we were able to streamline the data processing and model building process, resulting in more accurate and efficient models.

6. Results

Both models performed well, but the decision tree algorithm was slightly more accurate than the logistic regression algorithm when predicting human activity based on these specific features. This suggests that tree-based algorithms may be more effective for this particular task than logistic regression.

The logistic regression model has a train accuracy of 0.7258 and a test accuracy of 0.6915. In comparison, the decision tree model has a train accuracy of 0.7330 and a test accuracy of 0.7591, with a test error of 0.26699.

Model	Precision	recall	Accuracy	F1 Score
Logistic Regression	0.7056	0.7595	0.6915	0.7363
Decision tree	0.7250	0.7966	0.7330	0.7591

7. Conclusion

The study's aim is to develop prediction models for the identification of physical activity. A model might be used to predict the activity of new individuals based on their sensor readings by training it on the activity labels and sensor readings in this dataset. This might be used for many other things, including fitness trackers and assistive equipment for people with mobility issues. The information from this study may also be utilized in parallel to comprehend physical activity patterns and how they differ across individuals and population groupings. The data can be further broken down by gender, age, or body mass index, for instance, to reveal trends or disparities in patterns or levels of physical activity. This could influence recommendations for physical exercise or public health initiatives. The information might also be utilized to comprehend the connection between various sensor readings and degrees of physical activity. Researchers might determine which sensor readings are most useful for activity detection and use this information to guide the design of future physical activity monitoring devices by examining the link between various sensor readings and activity labels.