Group 14

Project Report

Human Activity Recognition Using Smartphones

Group 14:

Raaga Sindhu Mangalagiri
mangalagiri.r@northeastern.edu
Shriram Vijaykumar
vijaykumar.s@northeastern.edu
PrajwalSrinivas
srinivas.pra@northeastern.edu
Varun Kumar Kumaravel
Kumaravel.v@northeastern.edu
Siva Vasanta Harika Mangu
mangu.s@northeastern.edu

Abstract:

To predict the human activities such as (walking,standing,sitting,Laying) by analysing the historical time-series data which was collected by the sensors present in the samsung device.The readings were taken from 30 volunteers on a daily basis.This analysis will help us determine the activity performed by a person by looking at the sensor gyroscopic and accelerometric readings.

Introduction:

The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

Project Objective:

The goal of this project is to develop a robust and accurate model for better prediction for human activity recognition using data collected from smartphones. This model will classify activities such as walking, sitting, standing, etc., based on the sensor data.

Proposal:

Implementing a predictive analysis on the time series data to analyze the activity performed by observing various accelerometer and gyroscopic readings of a person and predicting the activity performed by the person using various statistical techniques and methods which will be applicable on our multivariate response variable.

Data description:

The dataset was taken from the UCI machine learning repository.each and every observations consists of 561 readings/features.Our X(X_train,X_test) consists of 561 features Among all 561 some of the readings are the acclerometeric readings,gyroscopic readings along XYZ axis.The predicted value y(y_train,y_test) consists the labels of our activity (standing,sitting,lying,walking). There are 10,299 instances in which we divided 70% of our data to be our training data and 30% to be our testing data.The link of our dataset to the UCI repository is provided below.
Link: UCI HAR Dataset

Methods:

The 3 methods we carried out our model training are:

1.Gaussian Naive Bayes:

The Gaussian Naive Bayes algorithm is well-suited for Human Activity Recognition using Smartphones because it efficiently handles high-dimensional data like sensor readings. This algorithm assumes each feature is independent, simplifying calculations and making it ideal for large datasets. Gaussian Naive Bayes also works well with continuous data, typical in smartphone sensors. It provides a solid baseline for performance comparison with more complex models, making it a practical choice for initial classification tasks in this project.

2.SoftMax Regression

Applying softmax regression (also known as multinomial logistic regression) to the Human Activity Recognition project using smartphones is suitable because it is effective for multi-class classification problems. This algorithm can handle multiple classes seamlessly, which is crucial for differentiating various human activities like walking, sitting, and standing. Softmax regression is particularly adept at managing probabilities for each class, providing a more nuanced understanding than simple binary classification. Moreover, it works well with high-dimensional datasets, like those derived from smartphone sensors, ensuring robust performance even with complex input data.

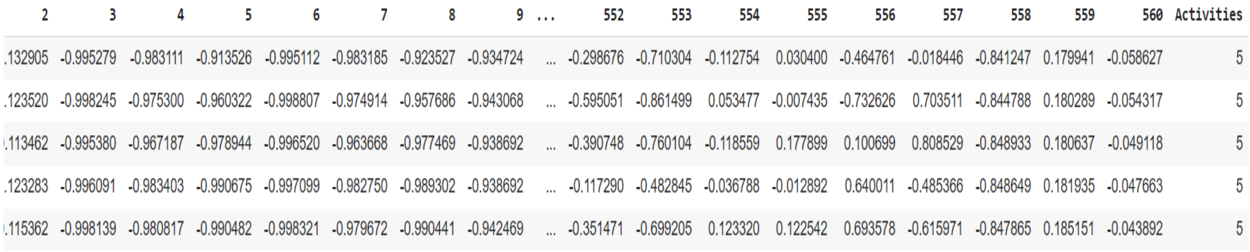
3.Neural Networks

Applying a neural network algorithm to the Human Activity Recognition project using smartphones is suitable due to its ability to learn complex patterns in high-dimensional data. Neural networks excel in feature extraction, meaning they can automatically detect and utilize intricate patterns from raw sensor data, which is crucial for accurately identifying diverse human activities. Furthermore, their flexible architecture allows for customization and optimization to improve performance on the specific types of data generated by smartphone sensors. This makes neural networks a powerful tool for handling the complexity and variability inherent in human activity recognition tasks.

Exploratory Data Analysis:

1.Data Collection:

The data is collected from the UCI repository .There are 561 features/observations for each and every activity and the activity being our response variable.



2	3	4	5	6	7	8	9	...	552	553	554	555	556	557	558	559	560	Activities
.132905	-0.995279	-0.983111	-0.913526	-0.995112	-0.983185	-0.923527	-0.934724	...	-0.298676	-0.710304	-0.112754	0.030400	-0.464761	-0.018446	-0.841247	0.179941	-0.058627	5
.123520	-0.998245	-0.975300	-0.960322	-0.998807	-0.974914	-0.957686	-0.943068	...	-0.595051	-0.861499	0.053477	-0.007435	-0.732626	0.703511	-0.844788	0.180289	-0.054317	5
.113462	-0.995380	-0.967187	-0.978944	-0.996520	-0.963668	-0.977469	-0.938692	...	-0.390748	-0.760104	-0.118559	0.177899	0.100699	0.808529	-0.848933	0.180637	-0.049118	5
.123283	-0.996091	-0.983403	-0.990675	-0.997099	-0.982750	-0.989302	-0.938692	...	-0.117290	-0.482845	-0.036788	-0.012892	0.640011	-0.485366	-0.848649	0.181935	-0.047663	5
.115362	-0.998139	-0.980817	-0.990482	-0.998321	-0.979672	-0.990441	-0.942469	...	-0.351471	-0.699205	0.123320	0.122542	0.693578	-0.615971	-0.847865	0.185151	-0.043892	5

Fig 1 Dataset

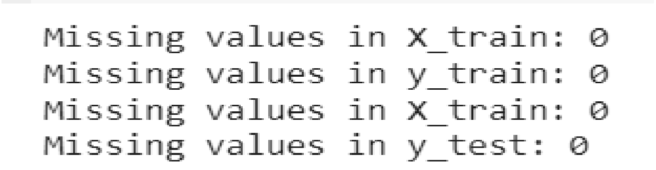
2.Data Cleaning :

The dataset that we have choosen is in a text format and response varible is encoded from 0 to 6 by mapping label_file with our y_train and y_test.

```
activity_labels = {0: 'WALKING', 1: 'WALKING_UPSTAIRS', 2: 'WALKING_DOWNSTAIRS', 3: 'SITTING', 4: 'STANDING', 5: 'LAYING'}
```

2.1 Missing values

- Checking for missing values within my data and found there are 0 missing values in our traning and testing datasets.There are no rows or columns we had to remove for our analysis.As all of our features mostly contribute to our analysis.We didn’t find any columns with the missing values more than 15%.So we are not dropping any features for our analysis



```
Missing values in X_train: 0
Missing values in y_train: 0
Missing values in X_train: 0
Missing values in y_test: 0
```

Fig 2:Checking for missing values

2.2 Visulalizations:

- Plotted a bar plot of all activities present in our data set .This bar plot illustrates response variable (activity) nearly equally distributed.

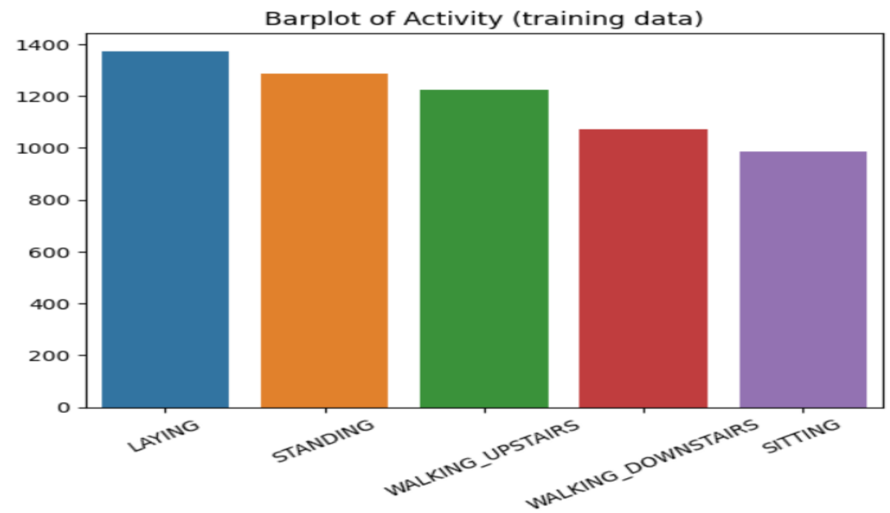


Fig 3: Barplot of Activities(y_train)

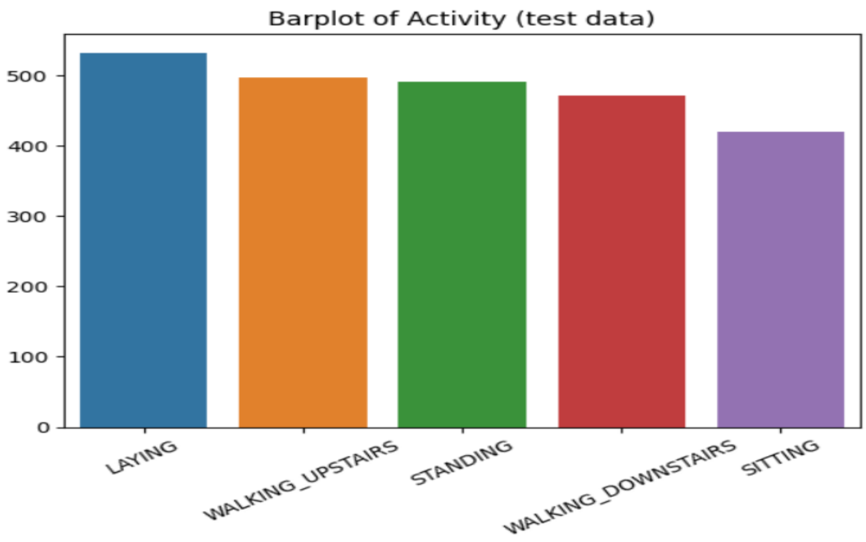


Fig 4: Distribution of activities (y_test)

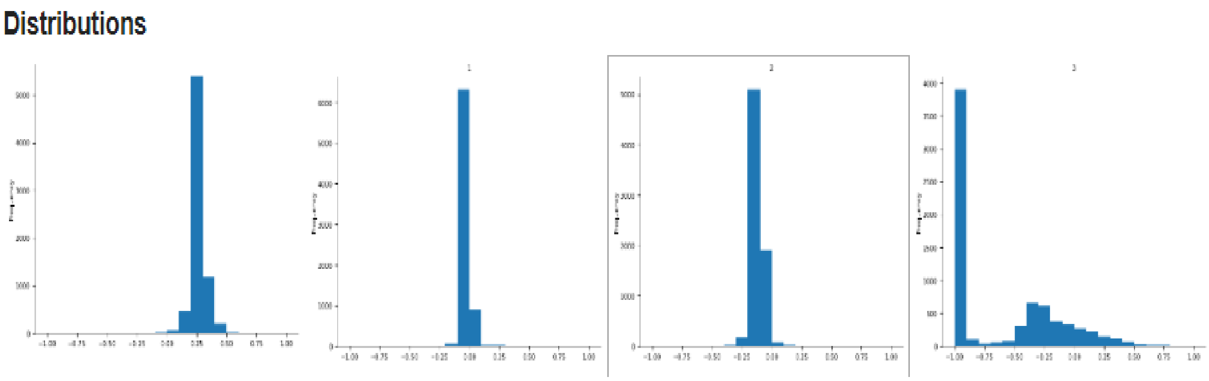


Fig 5: Distributions of features (X_train)

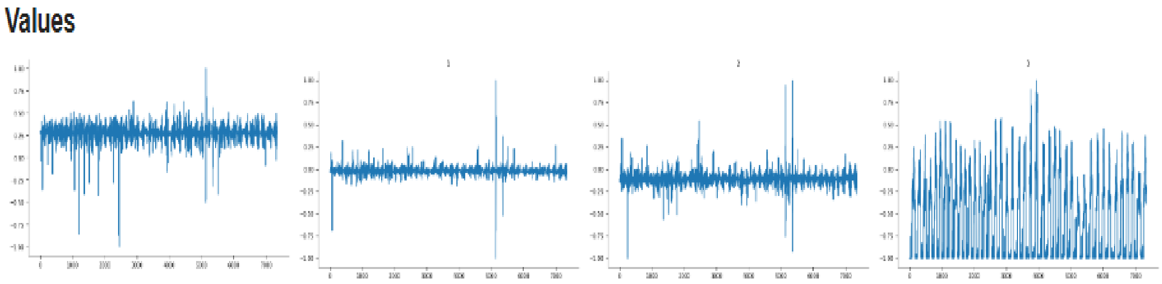


Fig 6: Distributions across various values

3. Standardizing the data:

- We standardized our data of all 561 features for a better analysis and interpretation and few reasons why we standardized our data is described below:
  Standardizing data normalizes the distribution of the variables, making them more consistent and easier to work with. Normalized data often follows a standard distribution (mean of 0 and standard deviation of 1), which can simplify analysis and modeling.Many machine learning algorithms perform better when working with standardized data. Algorithms like K-means clustering, support vector machines, and principal component analysis (PCA) assume that the data is centered around zero and has a standard deviation of 1. Standardizing the data helps these algorithms converge faster and prevents features with larger scales from dominating those with smaller scales.Srandardization helps to make certain methods sensitive to outliers,leading to more robust and accurate results in various analytical processes.

4. Perfoming statistical analysis

- To understand the overview of the distribution and the presence of missing values or potential issues with the data,indicate the count of non-null entry values in our data.
- From the values we observed that the column 4 has the highest standard deviation so the spread of data is across the column 4.Quaratile of 25%,50%,75% indicates the distribution of data across the columns.From this we understood the distribution tendency of our data.

5. Dimensionality reduction and feature extraction

- PCA is used to our dataset for dimensionality reduction and feature extraction
  The purpose of Dimensionality reduction PCA reduces number of variables to components , while preserving the most critical information. It transforms the original variables into a new set of variables (principal components) that are linear combinations of the originals.

- Number of principal components=2

.

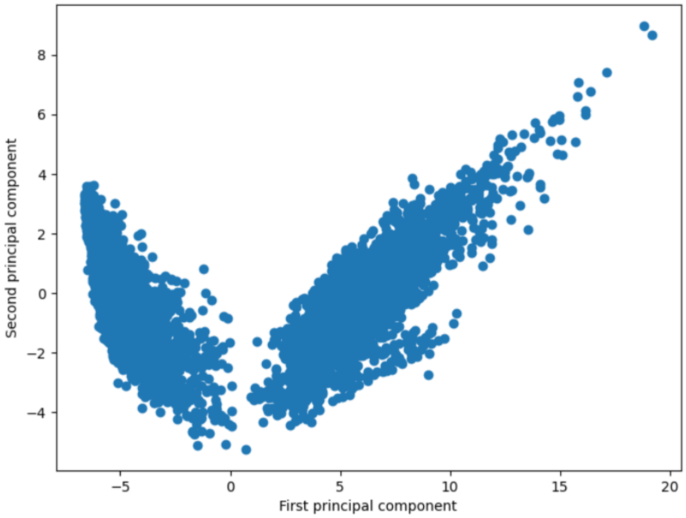


Fig 7: PCA(n=2) correlation analsis

Observation:

The PCA graph shows a scatter plot of the dataset projected onto the first two principal components. These components are linear combinations of the original variables that capture the maximum variance within the data. The shape of the distribution suggests two or more subgroups within the dataset. The spread along the principal component axes indicates that these components explain a significant portion of the data's variability. The gap in the center hints at possible clustering, although the absence of clear separation between clusters may suggest complex underlying relationships that PCA, as a linear method, cannot fully disentangle. The presence of outliers, particularly towards the right side of the plot, could represent extreme cases or anomalies in the data. Overall, the PCA graph provides a visual summary of the data, highlighting potential patterns and areas for further investigation.

Results :

1. Gaussian Naive Bayes:

The Gaussian Naive Bayes algorithm is well-suited for Human Activity Recognition using Smartphones because it efficiently handles high-dimensional data like sensor readings. This algorithm assumes each feature is independent, simplifying calculations and making it ideal for large datasets. Gaussian Naive Bayes also works well with continuous data, typical in smartphone sensors. It provides a solid baseline for performance comparison with more complex models, making it a practical choice for initial classification tasks in this project.

Metrics:
Accuracy: 0.8018
Macro Precision: 0.8089
Macro Recall: 0.7960
Macro F1 Score: 0.7974

Gaussian Naive Bayes often serves as a good baseline model. Due to its simplicity and speed, it's useful to initially apply it to the dataset and gauge its performance before moving on to more complex algorithms.

2. Softmax logistic Regression:

Applying softmax regression (also known as multinomial logistic regression) to the Human Activity Recognition project using smartphones is suitable because it is effective for multi-class classification problems. This algorithm can handle multiple classes seamlessly, which is crucial for differentiating various human activities like walking, sitting, and standing. Softmax regression is particularly adept at managing probabilities for each class, providing a more nuanced understanding than simple binary classification. Moreover, it works well with high-dimensional datasets, like those derived from smartphone sensors, ensuring robust performance even with complex input data.

Metrics:
Accuracy: 0.9426535459789617
Macro Precision: 0.9424782590110062
Macro Recall: 0.942143792737112
Macro F1 Score: 0.9422141985901069

3.Neural Network:

Applying a neural network algorithm to the Human Activity Recognition project using smartphones is suitable due to its ability to learn complex patterns in high-dimensional data. Neural networks excel in feature extraction, meaning they can automatically detect and utilize intricate patterns from raw sensor data, which is crucial for accurately identifying diverse human activities. Furthermore, their flexible architecture allows for customization and optimization to improve performance on the specific types of data generated by smartphone sensors. Number of epochs=10

```
 Metrics :
Accuracy: 95.32%
Precision: 95.30%
Recall: 95.25%
F1 Score: 95.26%

  precision   recall  f1-score   support

0       0.98      0.97      0.97       496
1       0.95      0.96      0.95       471
2       0.94      0.95      0.95       420
3       0.94      0.89      0.92       491
4       0.91      0.95      0.93       532
5       1.00      1.00      1.00       537


Confusion Matrix:
[[480   3  13   0   0   0]
 [  7 451  12   0   1   0]
 [  2  19 399   0   0   0]
 [  0   1   0 439  49   2]
 [  0   0   0  29 503   0]
 [  0   0   0   0   0 537]]
```

From above observation the loss decreases indicating that model is learning and reducing error.

The precision,recall and F-1 scores indicating good performance and the confusion matrix is plotted to check the true positives on diagonal and misclassifications off-diagnal.Overall the model is performing well with increasing accuracy and decreasing loss.

Overall Observations:

Overall Observation of various metrics on our dataset are as follows:

Metrics for Gaussian Naive Bayes:
Accuracy: 80.18%
Precision: 80.89%
Recall: 79.60%
F1 Score: 79.74%

Metrics for Softmax Regression:

Accuracy: 94.27%
Precision: 94.25%
Recall: 94.21%
F1 Score: 94.22%

Metrics for Neural Network:
Accuracy: 95.32%
Precision: 95.30%
Recall: 95.25%
F1 Score: 95.26%

- Model neural networks is having highest accuracy compared to other two models.

Discussions:

Gaussian Naive Bayes: 80.89% of the instances predicted as positive are actually positive.

Softmax Regression: 94.25% precision indicates a high level of reliability in the positive predictions.

Neural Network: 93.82% precision also indicates high reliability.

Cost Analysis:

Cost Implications: High precision is crucial in scenarios where false positives have high costs (e.g., medical diagnosis, spam detection). Lower precision might lead to higher costs due to unnecessary follow-up actions or treatments.while the metrics provide a quantitative measure of model performance, the cost analysis requires a qualitative understanding of the specific application domain and the implications of different types of errors.

Total Cost for Gaussian Naive Bayes: $29200
Total Cost for Softmax Regression: $8450
Total Cost for Neural Network: $7650

Observations of cost functions:

When integrating machine learning models into business operations, the choice of the model significantly impacts both performance and cost. Here's a concise comparison of Gaussian Naive Bayes, Softmax Regression, and Neural Networks in terms of their implementation and maintenance costs:

Model 1: Gaussian Naive Bayes:

Cost: High, Complexity: Low, Efficiency: High

Model 1 Explanation: Gaussian Naive Bayes models are computationally efficient and easy to implement, making them a cost-effective choice for smaller datasets and less complex problem domains. Their simplicity also ensures low maintenance costs.But as the cost is higher Gaussian naive bayes is not suggested for our approach.

Model 2: Softmax Regression

Cost: Moderate, Efficiency: Moderate to High, Complexity: Moderate

Model 2 Explanation: Softmax Regression, an extension of logistic regression for multiclass classification, offers a balance between performance and computational efficiency. It is more resource-intensive than Naive Bayes but still maintains moderate complexity, keeping implementation and maintenance costs reasonable.

Model 3: Neural Networks

Cost: Moderate , Efficiency: Low (considering computational resources), Complexity: High

Model 3 Explanation: Neural Networks, particularly deep learning models, require substantial computational resources and expertise for implementation and maintenance, leading to higher costs. They are well-suited for complex tasks and large datasets, where their higher cost is justified by significant gains in accuracy and performance.But ,for our problem neural network is preferable model.

Summary

Based on the provided metrics and cost functions, both "Softmax Regression" and "Neural Network" models outperform the "Gaussian Naive Bayes" model in terms of prediction quality and cost-effectiveness.