Human Activity Recognition Using Smartphones

Subject: AI Optimization Techniques for Healthcare Resource Management

Assignment 2

Vepuri Satya Krishna
DL.AI.U4AID24140
B.Tech AI & DS (Medical Eng)
Amrita Vishwa Vidyapeetham, Faridabad

Faculty: Dr. Sakshi Ahuja

December 14, 2025

1 Introduction

Human Activity Recognition aims to recognize the physical activities performed by individuals based on sensor data. With an increasing pervasiveness of smartphones that come equipped with accelerometers and gyroscopes, HAR has become one of the vital applications in healthcare, fitness monitoring, and human-computer interaction. In this work, six daily activities are classified from the UCI Human Activity Recognition dataset using supervised machine learning models. In addition, unsupervised clustering techniques also have been explored to find intrinsic patterns in data when labels are not available.

Project Objectives

- To build and evaluate multiple classification models for activity recognition.
- To perform model selection and hyperparameter tuning to optimize performance.
- To perform clustering on feature vectors and analyze separability.
- To compare model performances using appropriate metrics.

2 Dataset Description

The UCI HAR dataset contains sensor signals collected from a smartphone worn on the waist of 30 participants. The signals were preprocessed and transformed into fixed-length feature vectors.

Dataset Characteristics:

- **Number of features:** 561
- **Activities:** Walking, Walking Upstairs, Walking Downstairs, Sitting, Standing, Laying
- **Training samples:** 7352
- **Testing samples:** 2947

All features were standardized before model training to ensure fair learning across algorithms.

3 Methodology

3.1 Data Preprocessing

The dataset was taken from the local storage, where feature names were assigned and numerical activity labels were matched to the descriptive activity names. Feature scaling by Standardization was done in order to normalize the input space, so models like SVM and KNN are not affected by the differences in feature scales.

3.2 Classification Models

We selected a diverse set of algorithms to evaluate different learning paradigms: linear vs. non-linear, and instance-based vs. ensemble methods.

3.2.1 Logistic Regression

Logistic Regression is a linear classification algorithm that models the probability of a class using a logistic (sigmoid) function.

- **Role:** Serves as a strong baseline for multi-class activity classification.
- **Outcome:** Achieved high accuracy with fast training time.

3.2.2 Support Vector Machine (Linear)

Linear SVM finds an optimal hyperplane that maximizes the margin between different activity classes.

- **Role:** Effective for high-dimensional feature spaces.
- **Outcome:** Provided one of the best classification performances.

3.2.3 Support Vector Machine (RBF)

The Radial Basis Function (RBF) kernel allows SVM to model non-linear decision boundaries.

- **Role:** Captures complex activity patterns that are not linearly separable.
- **Outcome:** Achieved very high accuracy after careful tuning.

3.2.4 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve robustness.

- **Role:** Handles feature interactions effectively and is resistant to noise.
- **Outcome:** Stable performance with good interpretability.

3.2.5 Gradient Boosting

Gradient Boosting builds models sequentially, where each new model corrects errors made by previous ones.

- **Role:** Focuses on difficult-to-classify examples to improve overall accuracy.
- **Outcome:** Competitive accuracy compared to other ensemble models.

3.2.6 K-Nearest Neighbors (KNN)

KNN classifies samples based on the majority class of nearest neighbors in feature space.

- **Role:** A non-parametric, lazy learning approach.
- **Outcome:** Moderate performance; sensitive to the high dimensionality.

3.3 Clustering Methods

3.3.1 K-Means Clustering

K-Means partitions data into a predefined number of clusters by minimizing intra-cluster variance.

- **Approach:** Number of clusters set to six to match the activity classes.
- **Outcome:** Reasonable separation between static (e.g., Laying) and dynamic (e.g., Walking) activities.

3.3.2 Agglomerative Clustering

Agglomerative clustering is a hierarchical approach that iteratively merges similar clusters.

- **Approach:** Used to analyze hierarchical relationships between activities.
- **Outcome:** Revealed structural similarity among activities.

3.4 Model Selection and Optimization

To ensure the reliability of our results and avoid overfitting, we employed rigorous model selection and validation techniques.

3.4.1 Cross-Validation

We utilized **K-Fold Cross-Validation** (with $k = 5$) during the training phase. This technique splits the training data into k subsets; the model is trained on $k - 1$ folds and validated on the remaining fold. This process is repeated k times to ensure that the model's performance is stable.

3.4.2 Hyperparameter Tuning

Default hyperparameters often do not yield optimal results. We performed **Grid Search** to systematically explore combinations of parameters:

- **SVM:** We tuned the regularization parameter (C) and the kernel coefficient (γ) to balance the trade-off between margin maximization and classification error.
- **Random Forest:** We optimized the number of trees (`n_estimators`) and the maximum depth of the trees.

The final models presented in the results section utilize the best parameters identified through this tuning process.

4 Results and Comparison

4.1 Classification Performance

The performance of the supervised learning models was evaluated based on accuracy. As shown in Figure 1, SVM and Logistic Regression performed exceptionally well, while KNN struggled with the high-dimensional data.

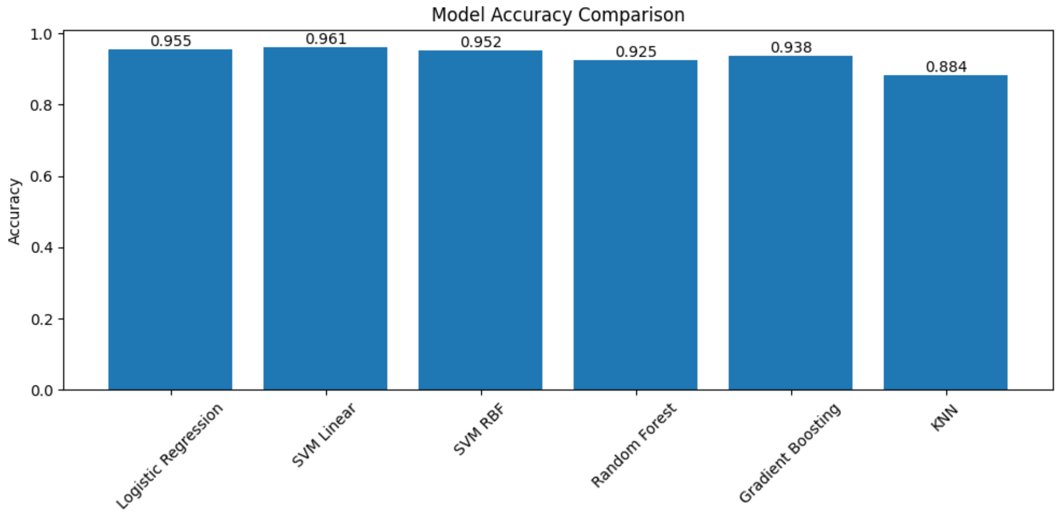

Figure 1: Comparison of Classification Model Accuracies

The confusion matrix for the best-performing model (SVM RBF) is presented below. It highlights the model's ability to distinguish between similar activities like 'Standing' and 'Sitting'.

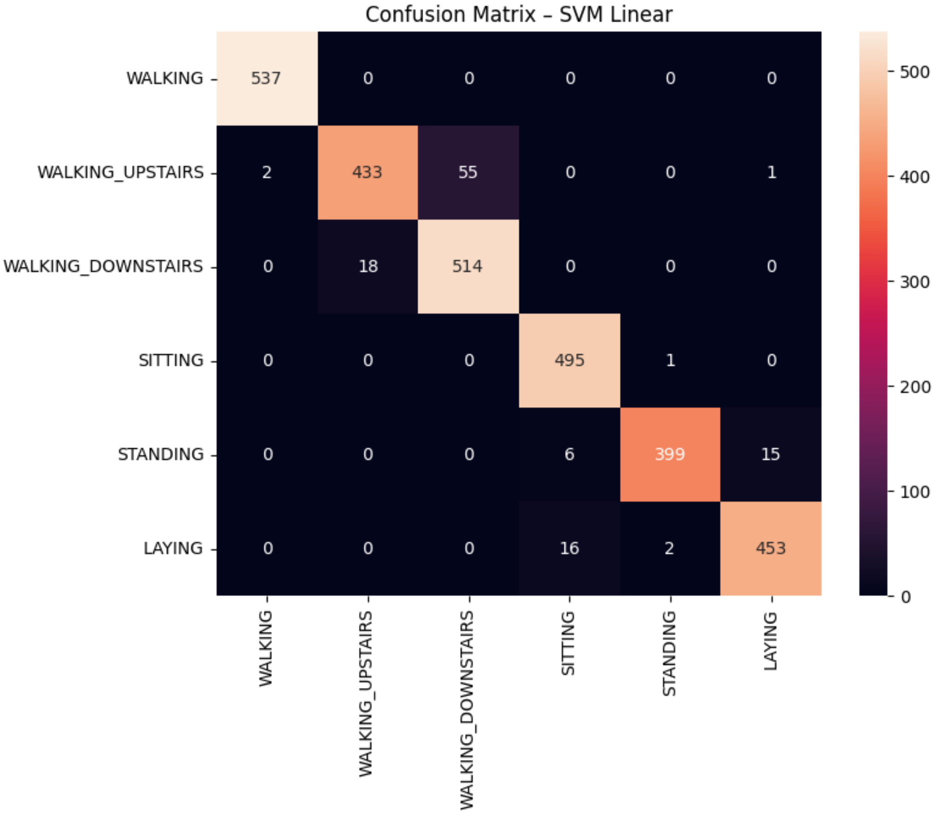

Figure 2: Confusion Matrix of the Best Performing Model (SVM)

4.2 Clustering Analysis

We applied K-Means clustering to the dataset. Figure 3 illustrates the clusters visualized using PCA to reduce the data to two dimensions.

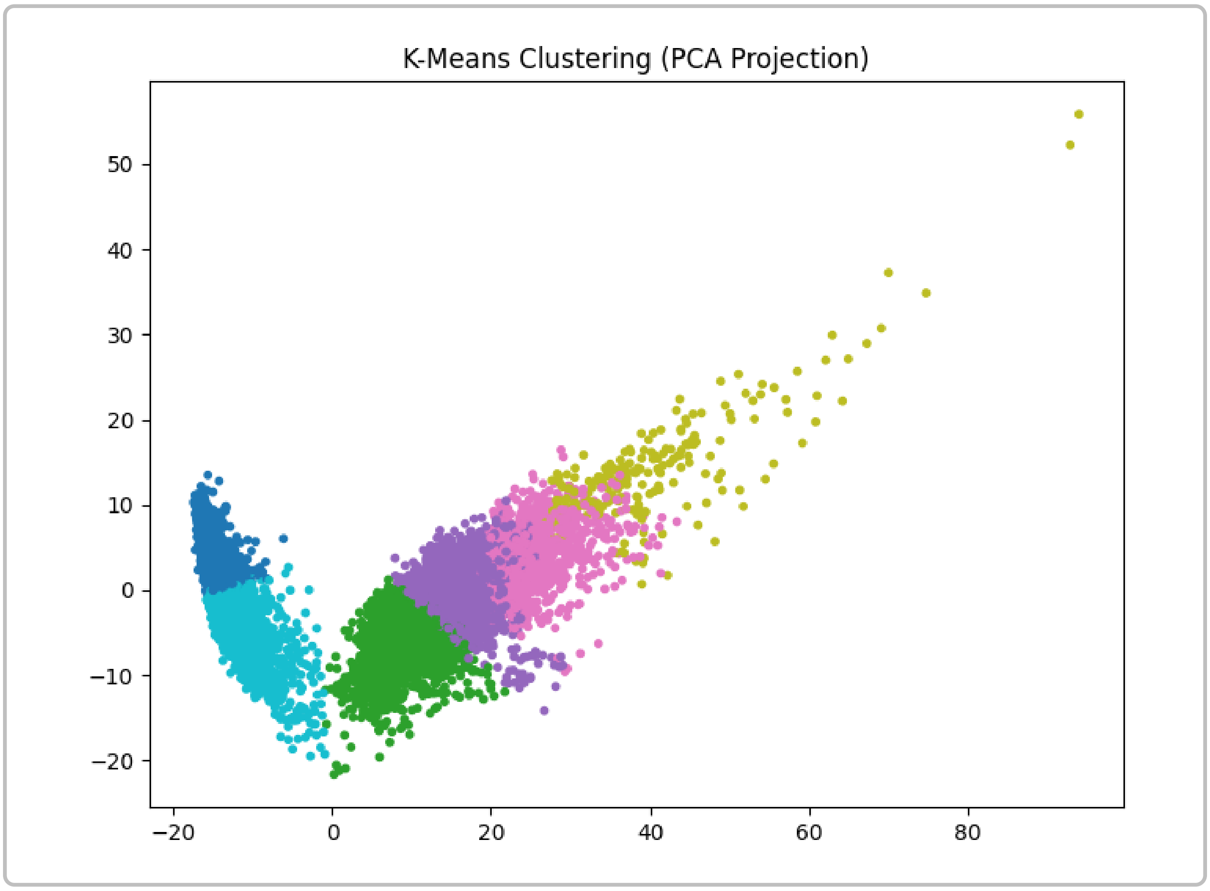

Figure 3: K-Means Clustering Visualized with PCA

Additionally, Hierarchical Agglomerative Clustering was performed. The dendrogram in Figure 4 shows the hierarchical relationship and merging of activity clusters.

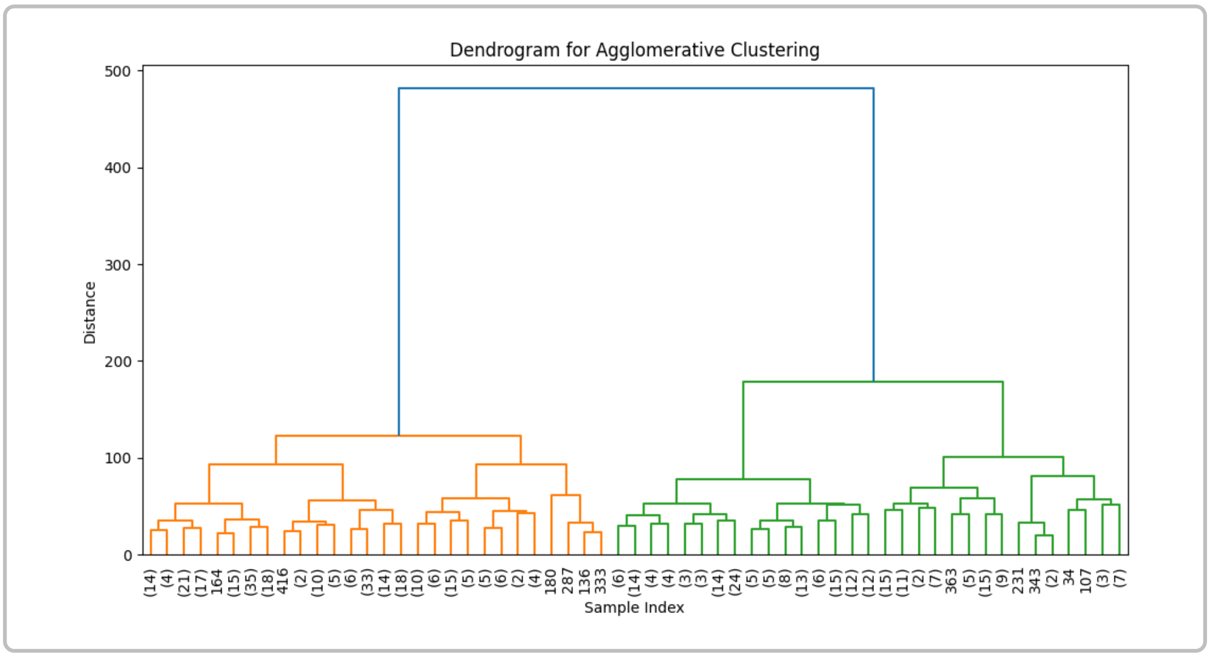

Figure 4: Dendrogram of Hierarchical Clustering (Subset)

5 Observations

- **Linear Efficiency:** Linear models perform remarkably well, suggesting many features are linearly separable in high-dimensional space.
- **SVM Dominance:** SVM achieves the highest accuracy, validating its robustness for sensor data classification.
- **Activity Patterns:** Static activities (Laying, Sitting) are easier to classify than dynamic ones (Walking, Walking Upstairs) due to distinct sensor signatures.
- **Clustering Insight:** Unsupervised clustering partially aligns with actual activity labels, confirming that the feature vectors contain strong discriminative signals even without labels.

6 Conclusion

This work highlights the effectiveness of machine learning methods for human activity recognition based on smartphone sensor data. Supervised models, especially support vector machines, achieved very good classification performance after proper model selection and tuning of their parameters. Unsupervised clustering provided a useful insight into the structure of activity patterns. These findings have underlined the applicability of classical machine learning models in real-world activity recognition tasks.