Summary Report

by Ammar Hassan

Report on Human Activity Recognition on the WISDM Dataset

Overview of the Data:

The **WISDM dataset** (Wireless Sensor Data Mining) is widely used for human activity recognition, containing accelerometer and gyroscope readings collected from mobile devices during various physical activities like walking, jogging, sitting, and standing. The dataset comprises labelled sequences of sensor data, making it ideal for both supervised classification and unsupervised clustering.

Key Characteristics:

- **Features**: Sensor readings like acceleration along x, y, and z axes, time stamps, and activity labels.
- **Challenges**: High dimensionality, overlapping activity patterns, and noise/missing data.
- **Pre-processing Steps**: Data cleaning, handling missing values by dropping or replacing with mean, scaling, and splitting into training and test sets.

Key Findings from Each Model:

1. **K-Nearest Neighbours (KNN)**:

   KNN achieved moderate accuracy (50-65%) on the test set, with its performance dependent on the choice of K. Lower K values resulted in overfitting, while higher K values provided better results. However, it was an easier approach to implement. It is computationally expensive when it comes to work with larger datasets.

2. **K-Means/K-Medoids Clustering**:

   Clustering accuracy was below expectations, as it relied purely on feature similarity without labels. Being sensitive to initialization and higher dimensionality it lacked accuracy in making clusters.

3. **Support Vector Machine (SVM)**:

   SVM achieved the higher accuracy as compared to other models, particularly with the RBF kernel. It had a great ability to handle complex features data. However, SVM is comparatively so much computationally expensive.

Comparison of Model Performances

Model	Training Time	Strengths	Weaknesses
KNN	Fast	Simple	Slow predictions, sensitive to K
K-Means	Moderate	Unsupervised insights	Sensitive to initialization
SVM (RBF)	Slow	has a capability of handling complex boundaries.	High computational cost

Overall Best Model: In my opinion after properly analyzing due to robustness and capability of handling complex boundaries SVM is overall the best model.

Final Reflections on the Best Approach

The **SVM model** is the best approach for the WISDM dataset, offering the highest accuracy and the ability to handle complex activity patterns. While it has higher training costs, but these computational costs are always neglected when we see the accuracy of SVM.

The **KNN model**, though straightforward and easy to implement, but it is not that much suitable for large scale due to its inefficiencies during predictions.

Similarly, **K-Means clustering** was not effective for this dataset, as unsupervised learning struggled with the intricate patterns in human activity data.

Challenges Faced

- **Missing Data**: The dataset contained missing values, such values were either dropped or replaced with the mean, which stumble the model performance.
- **Class Overlap**: Activities like walking and jogging showed significant feature overlap and this thing provided challenges in clustering.
- **Computational Intensity**: SVM required extensive computational resources, especially during hyperparameter tuning.