Ali Saleem (23L-2638)

Mini-Project: Human Activity Recognition Using WISDM Dataset

Comparison and Reflection:

- **Summarize the performance of KNN, K-Means/K-Medoids, and SVM. Which model had the best accuracy? Which model was fastest to train?**

After running KNN, K-Means, and SVM on the dataset, I found that SVM gave the best accuracy. It performed well at classifying activities, especially when the classes were distinct from one another. The model consistently delivered high accuracy, which was impressive, but it took longer to train compared to the other models. This is because SVM involves a more complex optimization process, which can be slow.

On the other hand, KNN had good accuracy, around 75%, and was faster to train because it doesn't require an explicit training phase. Instead, it stores the training data and works by calculating distances during prediction, making it quick to set up. K-Means/K-Medoids, though the fastest to train, had the lowest accuracy for classification tasks. It's more suited for clustering rather than classification, and because of the overlapping activity patterns, it struggled to accurately separate classes. While it was the quickest, it didn't perform as well in terms of accuracy.

- **Which approach do you think is best suited for this dataset and why?**

For this dataset, I believe SVM is the best approach because it performs well with high-dimensional data, which is common with accelerometer readings. It creates clear decision boundaries between classes, making it ideal for activity classification, especially for closely related activities like walking and jogging. Additionally, SVM has consistently delivered high accuracy in my experiments, even though it's slightly slower to train compared to other models.

If computational efficiency is more important, I'd consider KNN, as it is straightforward to implement and effective for smaller datasets.

- **Did the clustering results match your expectations? What challenges did you face?**

The clustering results were somewhat aligned with my expectations, but they weren't as accurate as I hoped. Clustering methods like K-Means struggled to differentiate between similar activities because the accelerometer data had overlapping features. For example, jogging and walking often had overlapping patterns, which made it hard for clustering to separate them effectively.

The main challenges I faced was Class imbalance; Some activities were underrepresented in the dataset, which impacted the model's ability to generalize well.

- **How could you improve these models or try other techniques to get better results?**

To improve the models, I could focus on **feature engineering** by deriving more meaningful metrics like mean, variance, or frequency-based features from the accelerometer data. Normalizing the dataset would also help models like SVM and KNN perform better. Additionally, tuning hyperparameters like the number of clusters for K-Means or the neighbors for KNN could significantly boost performance.

Lastly, combining clustering and classification might be an effective approach, where clustering can act as a preprocessing step to group similar activities before applying SVM or KNN for classification.