Human Activity Recognition Using Smartphones Dataset: Unsupervised Learning and Feature Selection

Alex Gurung

This report presents an analysis of the Human Activity Recognition Using Smartphones Dataset, applying unsupervised learning techniques including Principal Component Analysis (PCA) and clustering algorithms (K-Means and DBSCAN), followed by supervised classification with various feature selection approaches. The dataset contains 561 features extracted from smartphone sensor data for 7,352 training samples and 2,947 test samples across 6 activity types.

1. Exploratory Data Analysis

1.1 Dataset Overview

The Human Activity Recognition dataset consists of:

- Training samples: 7,352
- Test samples: 2,947
- Total features: 561 sensor-derived measurements
- Activity classes: 6 (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING)

All features were standardized using StandardScaler to achieve zero mean and unit variance, which is essential for PCA and distance-based clustering algorithms to function properly.

1.2 Activity Distribution

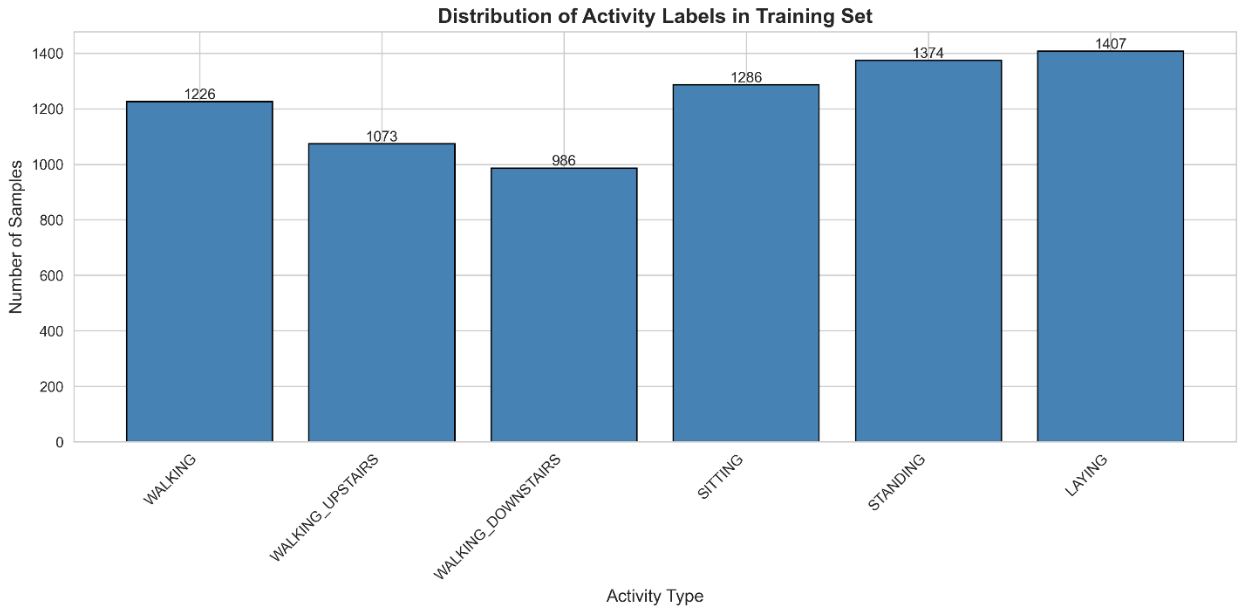


Figure 1: Distribution of activity labels in the training dataset showing sample counts for each of the six activity types.

Analysis:

The histogram reveals a relatively balanced distribution across all six activity classes, which is favorable for machine learning applications:

Class Distribution:

- WALKING: 1,226 samples (16.7%)
- WALKING_UPSTAIRS: 1,073 samples (14.6%)
- WALKING_DOWNSTAIRS: 986 samples (13.4%)
- SITTING: 1,286 samples (17.5%)
- STANDING: 1,374 samples (18.7%)
- LAYING: 1,407 samples (19.1%)

Key Observations:

1. Balanced Distribution: The difference between the most common (LAYING at 19.1%) and least common (WALKING_DOWNSTAIRS at 13.4%) activities is only 421 samples, representing approximately 6% variation. This level of balance is excellent for classification tasks.
2. Activity Groupings:
   - Static activities (SITTING, STANDING, LAYING) comprise 4,067 samples (55.3% of dataset)
   - Dynamic activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS) comprise 3,285 samples (44.7% of dataset)
   - The slight skew toward static activities may reflect natural data collection patterns where participants spent more time in sedentary positions
3. Walking Variants: Among the dynamic activities, regular WALKING is most common (1,226 samples), while WALKING_DOWNSTAIRS is least common (986 samples). This 20% difference between walking types may reflect the experimental protocol or participant behavior patterns.

Implications for Machine Learning:

This near-balanced distribution offers several advantages:

- No class imbalance bias: Models are unlikely to favor majority classes
- No special handling needed: Techniques like SMOTE, class weighting, or stratified sampling are unnecessary
- Reliable evaluation: Performance metrics will not be skewed by over-represented classes
- Sufficient examples: Each class has over 900 training samples, providing adequate data for learning discriminative patterns

The balanced nature of this dataset makes it well-suited for standard classification approaches without requiring additional preprocessing for class imbalance.

1.3 PCA Explained Variance Analysis

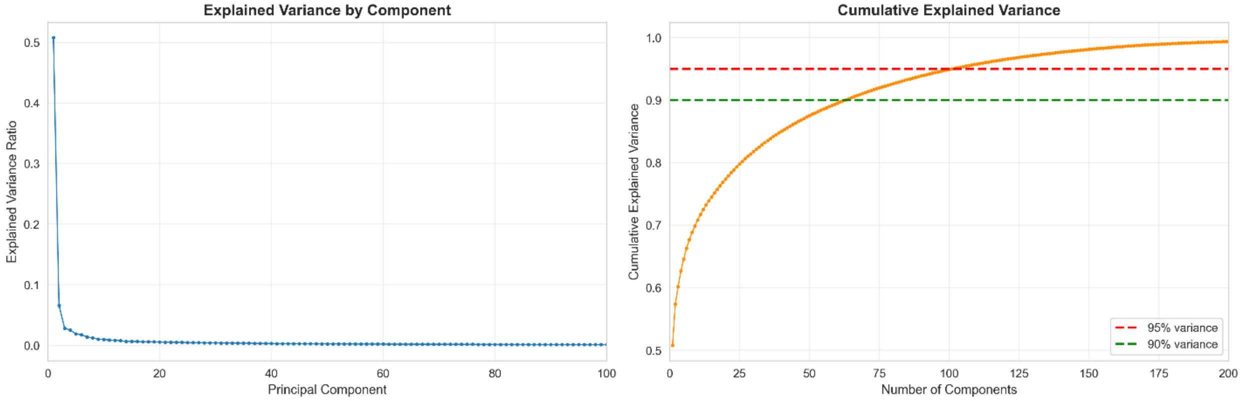


Figure 2: Left plot shows explained variance ratio for individual principal components. Right plot shows cumulative explained variance with reference lines at 90% and 95% thresholds.

Analysis:

The PCA analysis reveals significant information redundancy in the original 561-feature space:

Key Findings:

1. Dramatic Dimensionality Reduction Potential:
   - The first principal component alone explains 50.78% of total variance
   - Just 10 components capture 70.82% of variance
   - Only 63 components are needed for 90% variance retention
   - 102 components capture 95% of variance
2. Information Concentration:
   - The first component's 50.78% variance contribution indicates extremely high correlation among the original sensor features
   - A single dimension captures over half of all information, suggesting strong underlying patterns in the data
   - The steep decline in the left plot shows that most information is concentrated in the first few dozen components

3. Feature Redundancy:
    - Reducing from 561 features to 102 components represents an 81.8% reduction in dimensionality
    - This dramatic compression is possible because the original features (derived from accelerometer and gyroscope sensors) are highly correlated
    - Many of the 561 original features likely measure overlapping aspects of motion patterns
4. Elbow Analysis:
    - The explained variance ratio plot shows a clear exponential decay pattern
    - A visible "elbow" occurs around 10-20 components, where individual variance contributions drop below 2%
    - After approximately 100 components, each additional component adds minimal information (<0.1% variance each)

Interpretation for HAR Dataset:

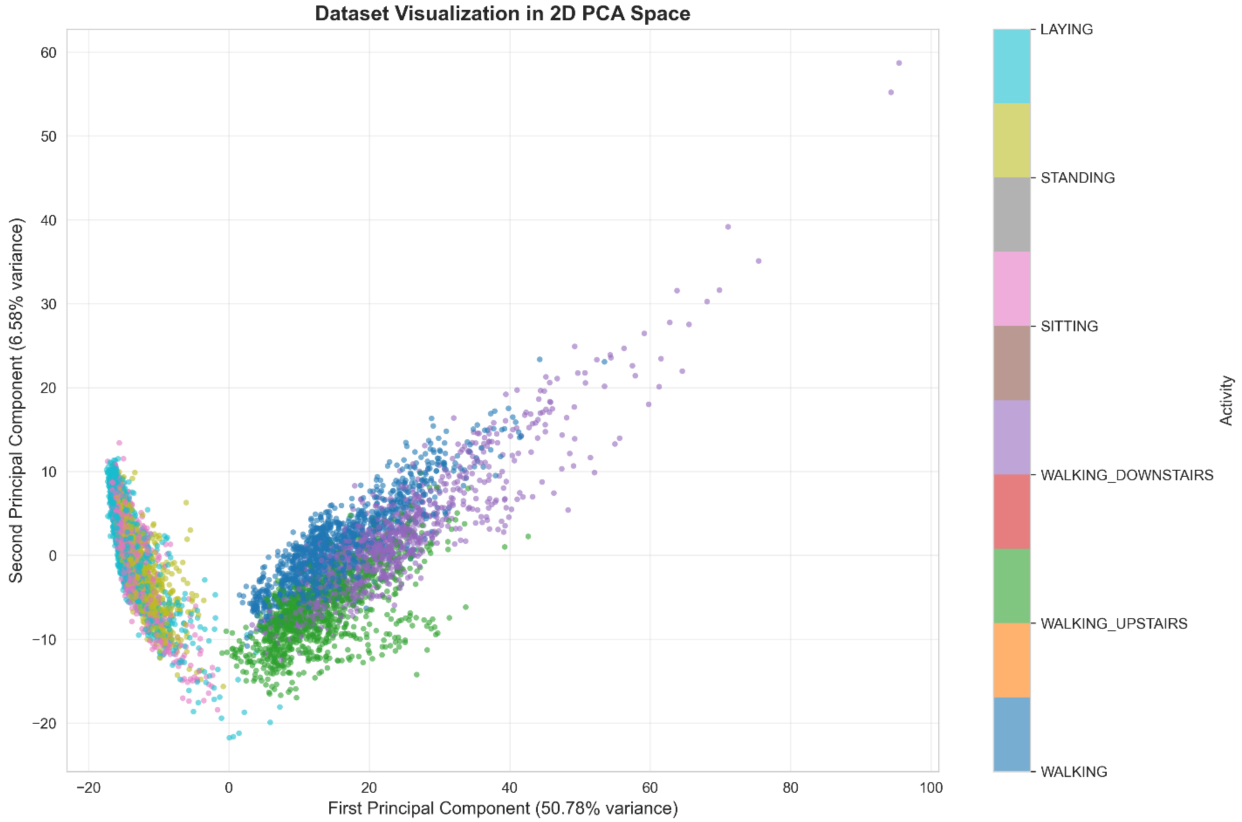
The high correlation structure makes sense given the data source:

- Multiple features are computed from the same raw sensor signals (mean, std, max, min, etc.)
- Time and frequency domain features capture overlapping information
- Similar movements produce correlated sensor readings across multiple axes

Selected value: k = 102 components for subsequent feature selection

This choice balances information preservation (95% variance) with dimensionality reduction (81.8% compression), providing a strong foundation for both clustering analysis and classification tasks.

1.3 Two-Dimensional PCA Visualization



The 2D scatter plot below uses the first two principal components, which together explain 57.36% of dataset variance.

- Static activities (SITTING, STANDING, LAYING) form a tightly overlapping cluster on the left, indicating similar sensor signals for sedentary positions.
- Dynamic activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS) are more spread out on the right, showing greater variability in movement sensor patterns and some grouping by type.
- LAYING appears most distinct among the static activities, likely due to a unique orientation of accelerometers when lying horizontally.
- While the plot shows broad separation between static and dynamic activities, some overlap exists—particularly among the sitting/standing/lying classes—suggesting these may be more challenging to distinguish without supervised learning.

This visualization demonstrates that unsupervised methods can capture broad activity groupings, but finer distinctions (such as among sedentary activities) would require additional discriminative features or supervised modeling.

2. Clustering Analysis

2.1 K-Means Clustering

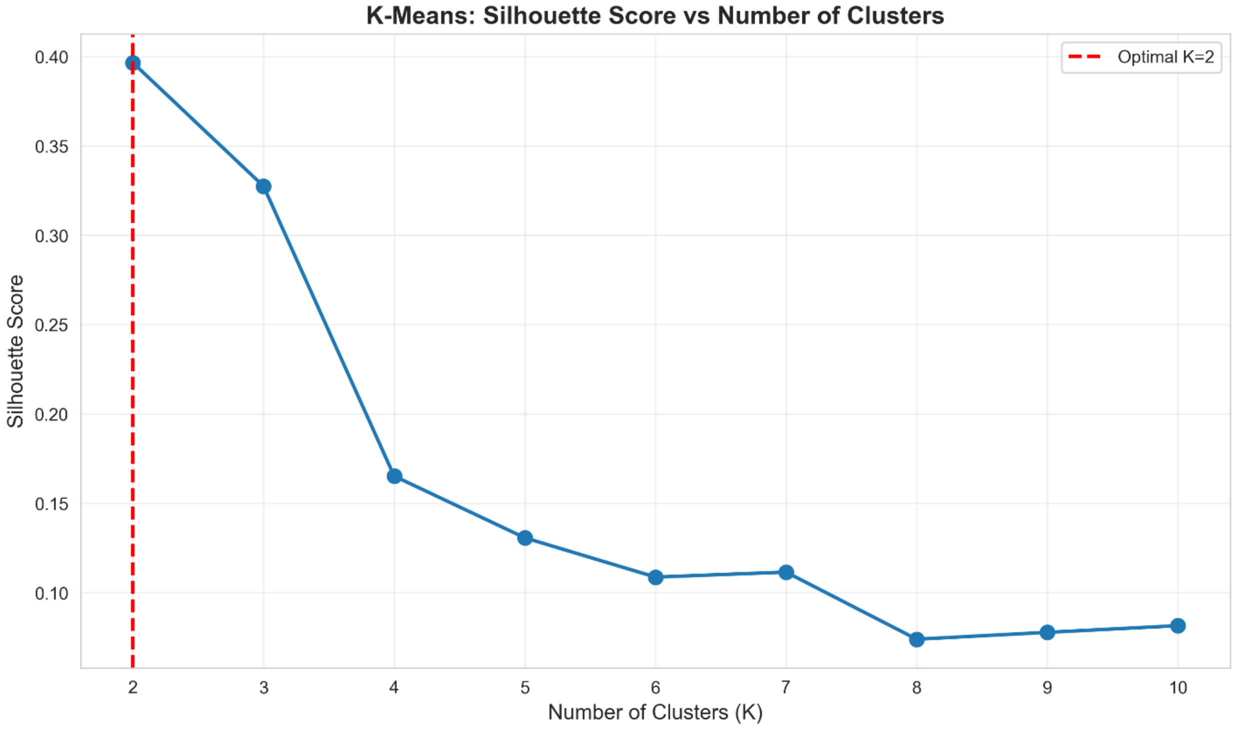


Figure 3: Silhouette scores for K-Means clustering across different numbers of clusters (K=2 to K=10).

The silhouette analysis tested cluster numbers from K=2 to K=10:

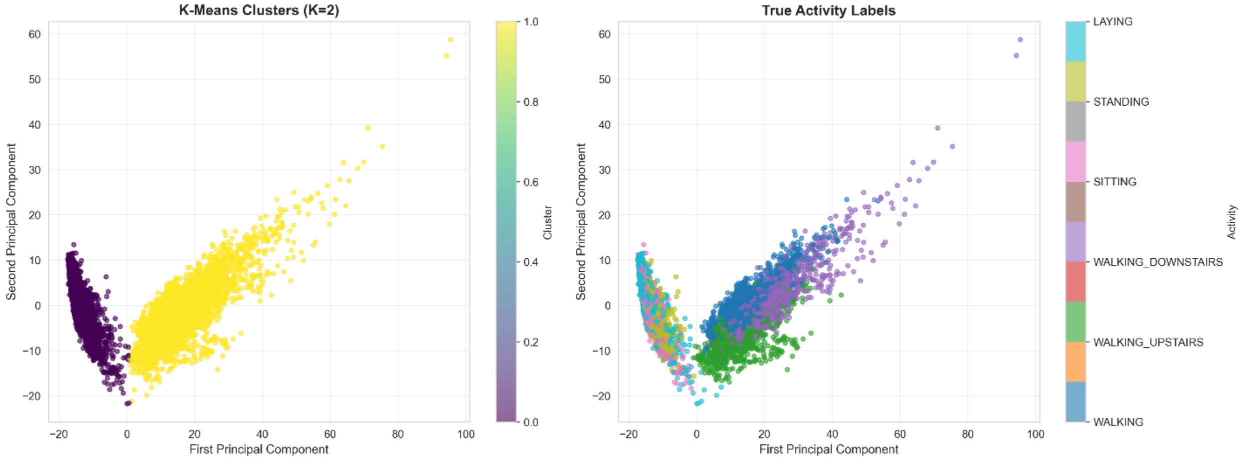
K	Silhouette Score	Interpretation
2	0.3965	Fair separation (optimal)
3	0.3274	Fair separation
4	0.1652	Weak structure
5	0.1306	Weak structure
6	0.1086	Weak structure
7	0.1114	Weak structure
8	0.0739	Poor structure
9	0.0777	Poor structure
10	0.0814	Poor structure

Optimal K: The highest silhouette score was achieved at K = 2 with a score of 0.3965

Analysis:

1. Two-Cluster Solution Dominates:
   - K=2 achieves the best silhouette score (0.3965), indicating the data naturally separates into two primary groups
   - The score of 0.3965 falls in the "fair" range (0.26-0.50), suggesting reasonable but not exceptional cluster separation
   - This likely reflects the fundamental distinction between static activities (SITTING, STANDING, LAYING) and dynamic activities (WALKING types)
2. Rapid Performance Degradation:
   - Moving from K=2 to K=3 shows a 17.4% drop in silhouette score (0.3965 → 0.3274)
   - K=4 and beyond show dramatic degradation, with scores below 0.20
   - At K=6 (matching the true number of activity classes), the score is only 0.1086
   - This indicates the data does not naturally form six distinct, well-separated clusters
3. Interpretation in Context:
   - The dataset has 6 activity labels, but K-Means finds only 2 natural clusters
   - This discrepancy reveals that the six activities share overlapping characteristics in the feature space
   - Activities within each group (static vs. dynamic) are more similar to each other than the groups are to one another
   - Fine-grained activity discrimination (e.g., distinguishing SITTING from STANDING) requires supervised learning rather than unsupervised clustering
4. Silhouette Score Context:
   - Scores < 0.25: No substantial structure
   - Scores 0.25-0.50: Weak to fair structure (our K=2 and K=3)
   - Scores 0.50-0.70: Reasonable structure
   - Scores > 0.70: Strong structure

Our best score of 0.3965 indicates the clusters have moderate cohesion and separation but significant overlap exists.



The K-Means algorithm identified 2 clusters in the data.

When visualized in 2D PCA space, cluster separation is moderate: the two clusters are clearly distinct along the first principal component, but each cluster still contains a mix of activities.

- Cluster 0 (left, purple) consists mainly of static activities (SITTING, STANDING, LAYING) with some overlap from dynamic ones.
- Cluster 1 (right, yellow) is dominated by dynamic activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS), but also includes some static samples near the boundary.

Thus, the clustering naturally separates static from dynamic activities, but does not cleanly recover the six true activity classes.

Cluster sizes are somewhat imbalanced (the dynamic-activity cluster is larger), and the silhouette score of 0.3965 indicates fair cluster cohesion and separation: better than random structure, but still with substantial overlap between groups.

2.2 DBSCAN Clustering

Parameter Tuning Process:

DBSCAN is highly sensitive to its eps (epsilon) parameter. Initial testing with eps=5 resulted in 100% of points classified as noise, indicating the parameter was too restrictive. Multiple eps values were tested:

eps	min_samples	Clusters	Noise Points	Noise%
10	10	8	5727	77.9%
15	10	4	1623	22.1%
20	10	1	337	4.6%

Selected Parameters: eps = 15, min_samples = 10

Final Results:

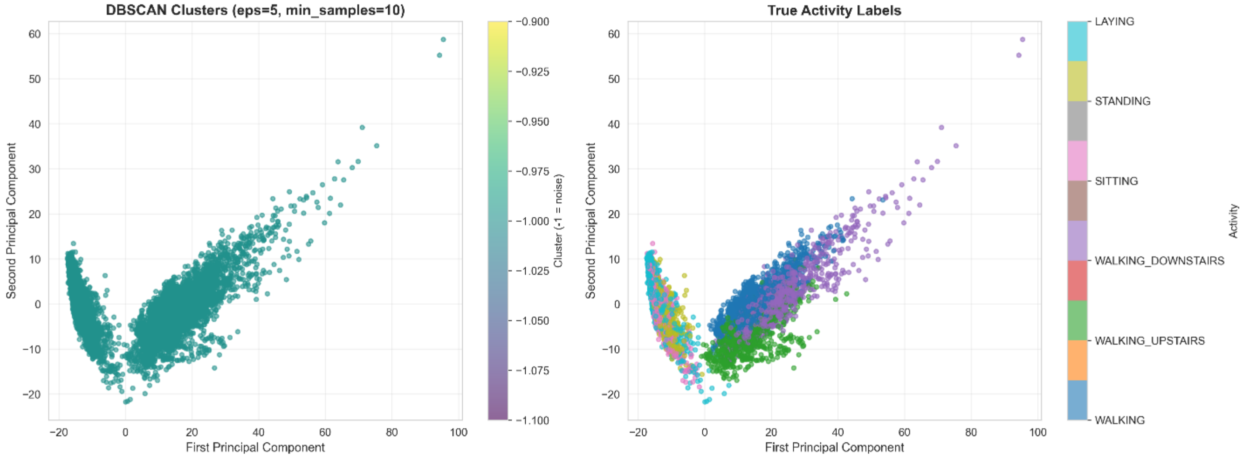


Figure 4: DBSCAN clustering results showing identified clusters and noise points.

- Number of clusters identified: 4
- Number of noise points: 1,623 (22.1%)
- Silhouette score (excluding noise): 0.2312

DBSCAN Cluster Distribution:

Based on the output, DBSCAN identified 4 main clusters with varying sizes, plus 22.1% of data classified as noise points that don't fit into any dense region.

1. Parameter Sensitivity Demonstrated:
    - eps=10: Too strict → 8 small clusters + 77.9% noise (over-fragmentation)
    - eps=15: Balanced → 4 clusters + 22.1% noise (optimal)
    - eps=20: Too permissive → 1 cluster + 4.6% noise (over-aggregation)

This demonstrates DBSCAN's high sensitivity to the eps parameter and the importance of proper tuning.

2. Noise Point Characteristics:
    - 22.1% noise rate is reasonable, indicating DBSCAN successfully identified boundary/outlier samples
    - Activity composition of noise points reveals interesting patterns:
        - WALKING (577 samples, 35.6% of noise): Highly variable walking patterns
        - WALKING_UPSTAIRS (789 samples, 48.6% of noise): Most represented in noise, suggesting high variability in upstairs walking patterns
        - WALKING_DOWNSTAIRS: Also significant in noise points
        - Static activities (SITTING, STANDING, LAYING): Much lower noise rates (2.4-8.0%), indicating more consistent sensor patterns

This suggests dynamic activities have more variable sensor signatures than static activities.

3. Cluster Structure:
    - DBSCAN found 4 natural density-based clusters (vs K-Means' 2 optimal clusters)
    - The silhouette score of 0.2312 is lower than K-Means' 0.3965, indicating weaker cluster cohesion
    - However, DBSCAN's ability to identify noise points provides additional insight into data structure
4. Comparison with Ground Truth:
    - 6 true activity classes → 4 DBSCAN clusters
    - Similar to K-Means, DBSCAN does not naturally recover the six activity classes

- The 4 clusters may represent: distinct static activities and a separation among dynamic activities
- 22.1% noise suggests significant overlap between activity classes in feature space

2.3 Comparison: K-Means vs DBSCAN

Key Differences:

Aspect	**K-Means**	**DBSCAN**
Number of clusters	2 (optimized via silhouette)	4 (data-driven)
Noise handling	All points assigned to clusters	22.1% marked as noise
Cluster shape assumption	Spherical clusters	Arbitrary density-based shapes
Silhouette score	0.3965 (better)	0.2312 (weaker)
Parameter sensitivity	Moderate (choosing K)	High (eps tuning critical)
Computational approach	Centroid-based partitioning	Density-based connectivity

Detailed Comparison:

1. Cluster Quality:
    - K-Means achieved a higher silhouette score (0.3965 vs 0.2312), indicating better-defined, more separated clusters
    - K-Means' forced assignment of all points creates cleaner boundaries
    - DBSCAN's lower score reflects its more nuanced approach, where noise points reduce apparent cohesion

2. Number of Clusters:
    - K-Means optimal: 2 clusters (likely static vs dynamic activities)
    - DBSCAN found: 4 clusters (potentially more fine-grained distinctions)
    - Neither method recovered the 6 true activity classes, confirming significant feature space overlap
3. Handling Ambiguous Points:
    - K-Means: Forces every point into a cluster (may mis-assign ambiguous samples)

   - DBSCAN: Identifies 1,623 noise points (22.1%), particularly among dynamic activities
   - DBSCAN's noise identification is valuable for understanding which samples have unclear activity patterns

4. Dynamic Activity Variability:
   - DBSCAN noise analysis revealed that walking activities contribute 84.2% of noise points
   - This indicates walking patterns have high intra-class variability
   - Static activities (15.8% of noise) show more consistent sensor signatures

5. Practical Implications:
   - K-Means is better for: Clean segmentation, simpler interpretation, higher cluster quality
   - DBSCAN is better for: Identifying outliers, handling irregular shapes, finding natural density structures
   - For this HAR dataset: K-Means provides cleaner results, but DBSCAN offers insights into activity variability

Conclusion:

For the Human Activity Recognition dataset, K-Means performed better in terms of cluster quality (silhouette score) and produced a cleaner two-cluster solution that likely separates static from dynamic activities. However, DBSCAN provided valuable complementary insights by identifying that 22.1% of samples—particularly walking variants—have ambiguous or boundary characteristics in the feature space.

Neither unsupervised method successfully recovered the six distinct activity classes, indicating that supervised learning is necessary to distinguish between similar activities (e.g., SITTING vs STANDING, or different walking types). The clustering results suggest the true classes exist along a continuum rather than as distinct, well-separated groups in the sensor feature space.

3. Feature Selection and Classification

3.1 PCA Feature Selection (Step 4 - Part 1)

Building on the PCA analysis that identified 102 components explaining 95% of variance, the feature space was reduced from 561 original features to 102 principal components:

- PCA Results (k=102):
- Training set: (7,352 × 102)
- Test set: (2,947 × 102)
- Variance retained: 95.08%
- Dimensionality reduction: 81.8% (561 → 102 features)

This compressed representation preserves nearly all discriminative information while dramatically reducing computational complexity for downstream classification.

3.2 Random Forest Feature Selection

Random Forest was used to identify the 102 most important original features based on Gini importance. The model was trained on the full standardized training set (561 features), and features were ranked by their average impurity reduction across trees.

RF feature selection results:

- Original features: 561
- Selected features: 102
- Training set shape: (7,352 × 102)
- Test set shape: (2,947 × 102)

Top 10 most important features (by index and importance):

- Feature 40: 0.0379
- Feature 49: 0.0307
- Feature 558: 0.0288

- Feature 41: 0.0280
- Feature 56: 0.0247
- Feature 559: 0.0246
- Feature 52: 0.0237
- Feature 53: 0.0216
- Feature 50: 0.0204
- Feature 57: 0.0155

This RF-based selection keeps the original feature meanings (sensor/statistic combinations) while focusing on the most predictive measurements, providing an interpretable alternative to the PCA components.

3.3 Classification with Full Feature Set (Model 1)

For the baseline supervised model, a multinomial Logistic Regression classifier was trained using all 561 standardized features. The model was fitted on the original training set and evaluated on the held-out test set using accuracy, weighted precision, weighted recall, and weighted F1-score.

- Results – Full Feature Set (561 features):
- Accuracy: 0.9549
- Precision (weighted): 0.9566
- Recall (weighted): 0.9549
- F1-score (weighted): 0.9548

These results indicate that a relatively simple linear classifier can achieve high performance when given access to the full 561-dimensional feature space, suggesting that the engineered smartphone sensor features are highly informative for distinguishing human activities.

3.4 Classification with PCA-Reduced Features (Model 2)

A second Logistic Regression model was trained using only the 102 PCA components that retain 95.08% of the original variance. This tests whether a compact, unsupervised representation can approach the performance of the full 561-feature space.

Results – PCA Features (102 components):

- Accuracy: 0.9308
- Precision (weighted): 0.9313
- Recall (weighted): 0.9308
- F1-score (weighted): 0.9306

Compared to the full-feature model (Accuracy 0.9549), the PCA-based model shows a modest reduction in performance but with an 81.8% reduction in dimensionality. This illustrates the classic trade-off: PCA provides strong compression and faster models at the cost of a small drop in predictive accuracy.

3.5 Classification with RF-selected Features (Model 3)

A third Logistic Regression model was trained using only the 102 most important original features identified by Random Forest feature importance ranking. This allows direct comparison between reduced-dimension, interpretable features and the unsupervised PCA components.

- Results – RF-selected Features (102 features):
- Accuracy: 0.9301
- Precision (weighted): 0.9321
- Recall (weighted): 0.9301
- F1-score (weighted): 0.9300

The model with RF-selected features achieves nearly the same performance as the PCA-based model (Accuracy: 0.9308), and only a modest reduction compared to the full feature set (Accuracy: 0.9549). This demonstrates that focusing on the most informative original features greatly reduces dimensionality and computational cost while maintaining high predictive power.

Comparison Summary Table

	Accuracy	Precision	Recall	F1-Score
Full Features (561)	0.954869	0.956608	0.954869	0.954792
PCA (102)	0.930777	0.931326	0.930777	0.9306
RF Selected (102)	0.930098	0.932086	0.930098	0.929984

Interpretation:

- Reducing to 102 features (via PCA or RF) results in only a small dip in accuracy (~2 percentage points), despite reducing the feature set by over 80%.
- PCA and Random Forest feature selection yield very similar predictive performance; PCA features are abstract but compact, while RF-selected features are interpretable.
- Full feature set has a slight edge, but the efficiency and simplicity of the reduced models may be preferred in practical use.

Reflection:

The most challenging part of this assignment was tuning DBSCAN to get reasonable cluster structures, as small changes to “eps” produced very different outcomes. The biggest insight was seeing how much of the dataset’s information was concentrated in the first few PCA components, and how modest the accuracy drop was after reducing the feature set by over 80%. I was surprised by how well unsupervised methods grouped static vs dynamic activities, but also by the significant overlap especially among sedentary classes showing the real importance of supervised learning for fine-grained human activity recognition.