# II Trimester MSc (AI & ML)

# Advanced Machine Learning

## Department of Computer Science

## HUMAN ACTIVITY RECOGNITION WITH SMARTPHONES


By

2348503 Aleena Varghese

2348511 Christina J Thattil

2348560 Sneha P


January 2024

# CERTIFICATE

*This is to certify that the report titled **Title of Project** is a bonafide record of work done by **Aleena Varghese (2348503)**, **Christina J Thattil (2348511), Sneha P (2348560)** of CHRIST (Deemed to be University), Bangalore, in partial fulfillment of the requirements of II Trimester of M.Sc. Artificial Intelligence and Machine Learning during the year 2023-24.*

**Course Teacher**

Valued-by: (Evaluator Name & Signature)

1.

2.

Date of Exam:

# CONTENTS

# PROJECT REPORT

## Title: Human Activity Recognition with Smartphones

## ABSTRACT


This project focuses on Human Activity Recognition (HAR) utilizing smartphone sensor data to classify activities of daily living (ADL) within the healthcare domain. Human Activity Recognition (HAR) is a pivotal discipline, classifying an individual's activities through responsive sensors influenced by human movement.

This project focuses on the intricate task of human activity recognition, specifically leveraging accelerometers and gyroscope sensors embedded in smartphones. The essence of this study lies in the meticulous comparison of the efficiency and precision exhibited by these classification approaches.

Through this project, we aim to contribute insights into the most effective machine-learning approaches for Human Activity Recognition with smartphones. Thus, the intersection of technology and healthcare holds promise for improved patient care, making strides toward a future where accurate activity recognition plays a pivotal role in shaping individualized wellness strategies.


# INTRODUCTION

In the ubiquitous smartphone era with advanced sensors, the convergence of technology and healthcare has opened avenues for innovative applications. The dataset originates from 30 participants, ranging from 19 to 48 years of age, who wore waist-mounted smartphones containing accelerometers and gyroscopes.

The primary goal of this project is to develop a robust machine learning model capable of accurately classifying six distinct activities: WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, and LAYING. Participants performed these activities while the smartphones recorded 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The dataset was meticulously labeled through video recordings, establishing a ground truth for model training and evaluation.

The ultimate goal is to train a machine learning model capable of accurately classifying activities based on the sensor data. This technology has significant implications in healthcare, allowing for the monitoring of patients' physical activities, assessment of mobility, and detection of anomalies that could indicate health-related issues. The project aims to contribute to advancements in personalized health monitoring and wellness applications, providing valuable insights for healthcare professionals and researchers.

To address the data, a relevant dataset named "human-activity-recognition-with-smartphones" from the Kaggle has been selected. This dataset, derived from real-world scenarios, provides a rich source for the analysis and implementation of machine learning algorithms for activity recognition.

# OBJECTIVE

This human activity recognition proposes many applications and several benefits. This mobile-based health application can be beneficial for the elderly or senior assistance. Also, we can use this application for personal health monitoring because mobile will be attached to us and the application tracks our activity over time.

The primary objective of this project is to develop a robust Human Activity Recognition (HAR) system using smartphone sensor data. This involves implementing and evaluating machine learning algorithms, including K-Nearest Neighbors (KNN), Logistic Regression, and Random Forest while employing key evaluation metrics such as accuracy, precision, recall, and F1-score.

The overarching goal is to contribute to societal well-being by creating a system with applications in healthcare, offering benefits such as remote patient monitoring and personalized health assessments. Our project falls into the scope of Activity Recognition, a field that offers many benefits and enables many new applications, for example, step counters on your Smartphone, as well as applications for elderly assistance and personal health monitoring.

# BACKGROUND

### Sensing Activity

Although there are so many sensors that are generally used in HAR and they measure different attributes including vital signs (e.g. heart rate, body temperature, and blood pressure), motion (e.g. acceleration, speed), and environmental signals (light intensity and environment) to choose a right sensor we need to have considered first element for the design of HAR system.

### Accelerometer

The accelerometer is an instrument that measures the experienced physical acceleration of an object. It is generally used for measurements in applications like vibration in machinery, acceleration in high-speed vehicles, and high-loaded bridges, etc.

### Gyroscope

A gyroscope is a sensor that can provide orientation information as well, but with greater precision. For HAR, this sensor has been used for various tasks such as the detection of various transitions between various postures and the detection of various activities.

### Smartphone as wearable sensor

Using a Smartphone as a wearable device we easily get information of a user's linear acceleration and angular velocity. The information will not be highly affected by the accelerometer and gyroscope by the bad indoor signal of GPS and electromagnetic noise in compass. However, accelerometer measurements are always influenced by the gravity factor in the detection of acceleration of a moving body.

# DATA PRE-PROCESSING AND EXPLORATION

### Data Understanding and Exploration:

Before initiating the model development phase, a thorough comprehension of the "human-activity-recognition-with-smartphones" dataset is imperative. Leveraging Python and Pandas, we conducted an in-depth exploration of the dataset. Loading the data from 'X_train.txt', 'y_train.txt', and 'subject_train.txt', we integrated features, labels, and subject information into a unified dataset. Fundamental insights were obtained through the examination of dataset structure, summary statistics, and initial rows. Additionally, a visual representation of the distribution of activities was generated to offer a holistic view of the dataset.

### Data Cleaning and Handling Missing Values:

A clean dataset is essential for building reliable machine-learning models. Utilizing Pandas, we checked for missing values within the dataset. Fortunately, no missing values were identified, eliminating the need for imputation or removal. This step ensures the integrity of the dataset, laying a solid foundation for subsequent analyses.

### Data Integration and Feature Engineering:

In the process of data preprocessing, several crucial steps were undertaken to enhance the quality and reliability of the dataset.

● Initially, null and missing values within the dataset were identified and subsequently removed, ensuring the elimination of potential sources of noise or bias in the analysis.

● To facilitate the application of machine learning algorithms, the target attribute 'Activity' underwent encoding using label encoder, enabling a numerical representation of categorical data.

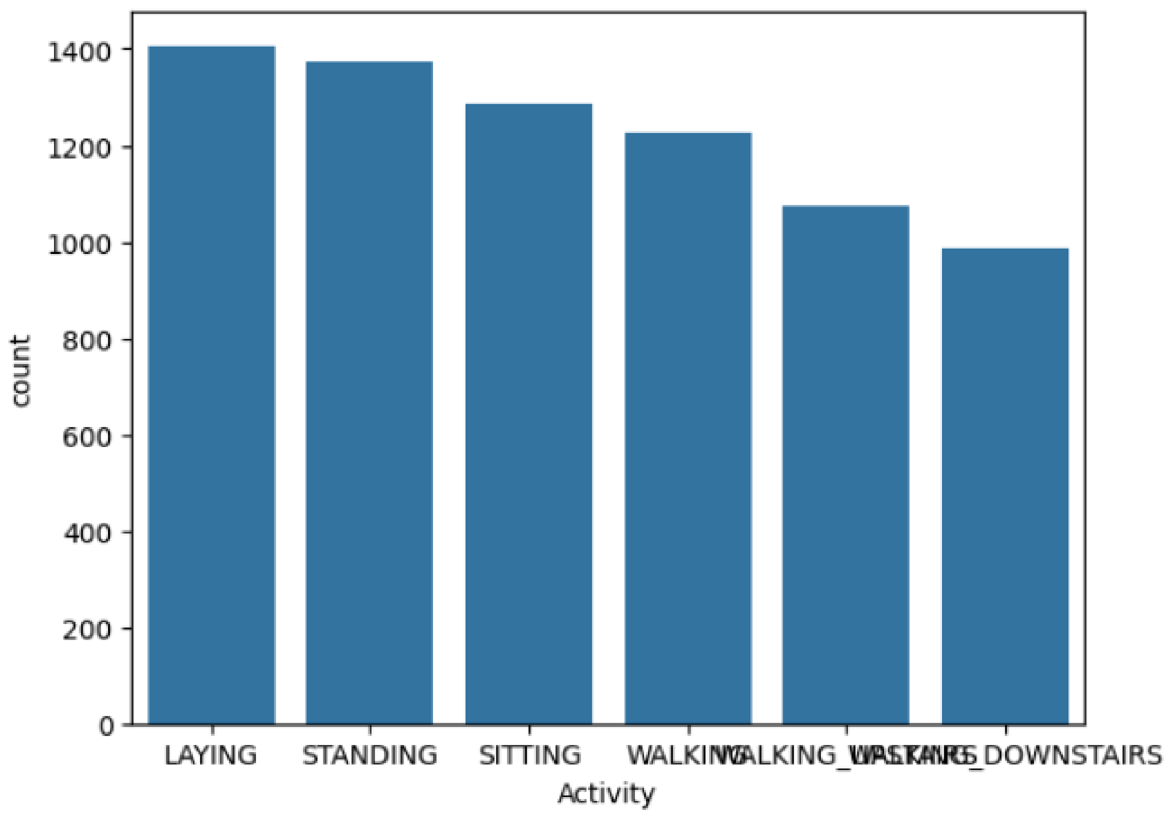
● Furthermore, outlier detection using the Z-score method was employed to identify and eliminate data points that deviated significantly from the mean. This rigorous approach to outlier removal resulted in a reduction of the dataset from its initial size of 7352 rows to a refined set comprising 4553 rows.

● In order to address the challenges posed by the high dimensionality of the dataset, Principal Component Analysis (PCA) was applied as a dimensionality reduction technique. The dataset initially comprised 563 feature variables, making it computationally intensive and potentially prone to overfitting.

● PCA, a widely utilized method in data science, was employed to transform the original feature space into a reduced set of principal components that capture the maximum variance in the data. Remarkably, this reduction allowed the preservation of essential information while significantly diminishing the dimensionality of the dataset.

● In this specific instance, PCA effectively compressed the dataset to a mere 7 principal components, thereby facilitating a more streamlined and efficient representation of the underlying patterns in the data. This reduction not only optimizes computational efficiency but also mitigates the risk of multicollinearity, ultimately contributing to improved model performance and interpretability in subsequent analyses.

The application of these preprocessing techniques not only ensures the integrity of the dataset but also enhances the robustness and accuracy of subsequent analyses and model training processes



# ALGORITHM IMPLEMENTATION

In the pursuit of accurately classifying human activities in the context of smartphone-based recognition, three distinct algorithms were thoughtfully chosen and implemented. Each algorithm addresses the unique challenges presented by the dataset, contributing to a comprehensive evaluation of their efficacy.

## Algorithm 1:

**K-Nearest Neighbors (KNN)**

In the context of the 'human activity recognition with smartphones' dataset, the choice of k-Nearest Neighbors (KNN) as a classification model is driven by its suitability for discerning patterns in spatial relationships within the feature space. KNN is particularly apt for this dataset as it relies on the proximity of similar instances in the feature space to make predictions. Human activities, when recorded by smartphones, may exhibit spatial clusters based on sensor readings. KNN's mechanism of assigning labels based on the consensus of nearby instances makes it well-suited to capture the inherent spatial patterns associated with different activities. Additionally, KNN is non-parametric and requires minimal assumptions about the underlying data distribution, making it adaptable to the diverse and potentially complex nature of human activity data collected from smartphones.

The dataset is divided into training and testing sets using an 80-20 split ratio. The KNN classifier is initialized with the ‘n_neighbors’ parameter set to 3, indicating that the model will base its predictions on the labels of the three nearest neighbors in the feature space. The classifier is then trained on the training set, and predictions are made on the test set. The accuracy of the model is calculated and printed, providing a straightforward evaluation of its predictive performance. In summary, the purpose of this code is to train and assess a KNN classifier for the given dataset, with the ‘n_neighbors’ parameter controlling the influence of nearby data points on the model’s predictions.

While the initial accuracy of 33.15% for the k-Nearest Neighbors (KNN) model suggests room for improvement, several strategies can enhance its performance. Firstly, optimizing the choice of the 'k' parameter may significantly impact results; conducting a thorough grid search or using cross-validation can help identify the optimal value. Additionally, scaling the features to a consistent range and exploring dimensionality reduction techniques can mitigate the impact of varying scales and enhance model interpretability. Feature engineering to extract more meaningful information and address potential outliers could further refine the model. Moreover, experimenting with alternative algorithms or ensemble methods might uncover a more suitable approach for capturing the intricate patterns within the 'human activity recognition with smartphones' dataset. Regularly reassessing and fine-tuning these aspects will likely contribute to a notable boost in the model's accuracy.

## Algorithm 2:

**Random Forest**

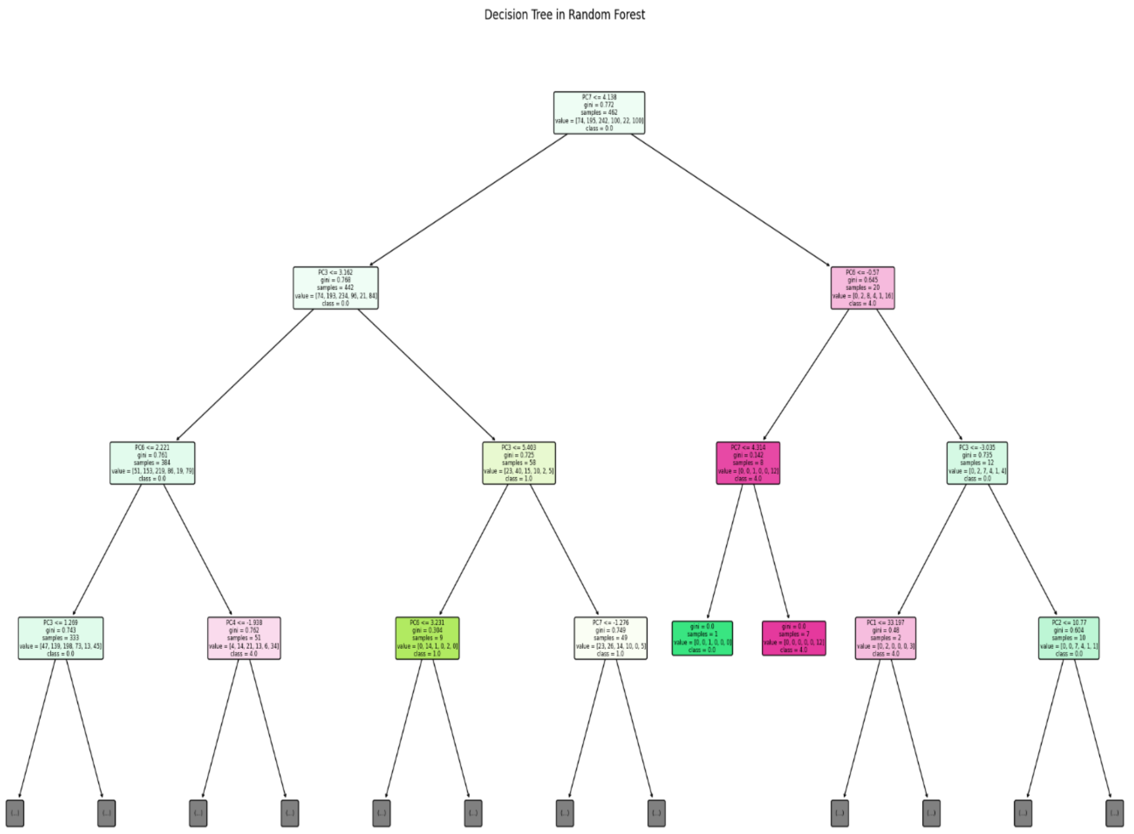
Random Forest, an ensemble learning method, harnesses the power of multiple decision trees to enhance predictive accuracy. Its proficiency in handling complex datasets and mitigating overfitting makes it an apt choice for human activity recognition. The Random Forest classifier was implemented using scikit-learn, providing a robust approach to capture intricate patterns within the dataset.

The Random Forest classifier is particularly well-suited for this task due to its ability to handle complex relationships within the data and mitigate overfitting. It aggregates predictions from multiple decision trees, each trained on a random subset of the data and features, providing a more robust and accurate prediction than an individual tree.

The dataset 'df_final' is pre-processed, with the target variable 'Activity' separated from the features. The data is then split into training and testing sets using an 80-20 ratio. A Random Forest classifier is instantiated and trained on the training data. Predictions are made on the test set, and model performance is evaluated using accuracy, a confusion matrix, and a classification report. The printed results provide a comprehensive overview of the classifier's effectiveness, including its accuracy percentage, confusion matrix detailing true positives, true negatives, false positives, and false negatives, and a detailed classification report comprising precision, recall, and F1-score metrics for each activity category.

Despite the modest overall accuracy of 37.50%, the classification report reveals that the RandForest classifier performs relatively well in certain aspects. Notably, it demonstrates a higher precision and recall for classes 1.0 and 2.0, suggesting that the model is more adept at correctly identifying instances and avoiding false positives for these activities. Additionally, the classifier achieves a weighted average F1-score of 0.34, indicating a balanced trade-off between precision and recall across all classes. This suggests that, while there is room for improvement, the model is making reasonably accurate predictions for multiple activity categories. These insights can guide further refinements, focusing on maintaining or enhancing the positive aspects while addressing the specific challenges posed by certain classes.

In summary, addressing class imbalance, tuning hyperparameters, exploring additional features, and considering alternative models could enhance the performance of your activity prediction model. Regular evaluation and iteration are crucial for refining the model until satisfactory results are achieved.



## Algorithm 3:

### K-Means

K-means excels in HAR for its simplicity and clarity. It analyses smartphone sensor data (acceleration, frequency, etc.) in time windows, grouping them based on similarities. The clusters represent activities like walking, running, or stillness. Comparing new data points to these clusters allows K-means to infer your current activity. K-means' speed and interpretability make it a valuable tool for understanding smartphone sensor data and recognizing human activities.

K-Means holds significance in the 'human activity recognition with smartphones' dataset as it facilitates unsupervised clustering, revealing inherent patterns and groupings within the data. By iteratively assigning data points to clusters based on their proximity to centroids, K-Means identifies distinct behavioural patterns among various activities recorded by smartphones. This clustering enables a deeper understanding of similarities and differences in activity profiles. Additionally, the algorithm aids in dimensionality reduction, simplifying the representation of complex activity data. Insights gained from K-Means clustering can inform feature engineering, model interpretation, and further analysis, enhancing the overall understanding of human activities and contributing to the development of more accurate and effective predictive models for smartphone-based activity recognition.

Elbow Method for Optimal k:

- The code iterates over a range of cluster numbers (from 1 to 10) and applies K-Means clustering to calculate the Within-Cluster Sum of Squares (WCSS) for each 'k.'
- The WCSS represents the sum of squared distances between data points and their assigned cluster centroids. The elbow method visualizes this information to identify an optimal number of clusters by finding the 'elbow' point where further clustering improvement diminishes.

Apply K-Means Clustering:

- After determining the optimal number of clusters (here, set to 3), K-Means clustering is applied to the dataset using the chosen 'k.'
- The code obtains cluster labels for each data point and centroid coordinates.

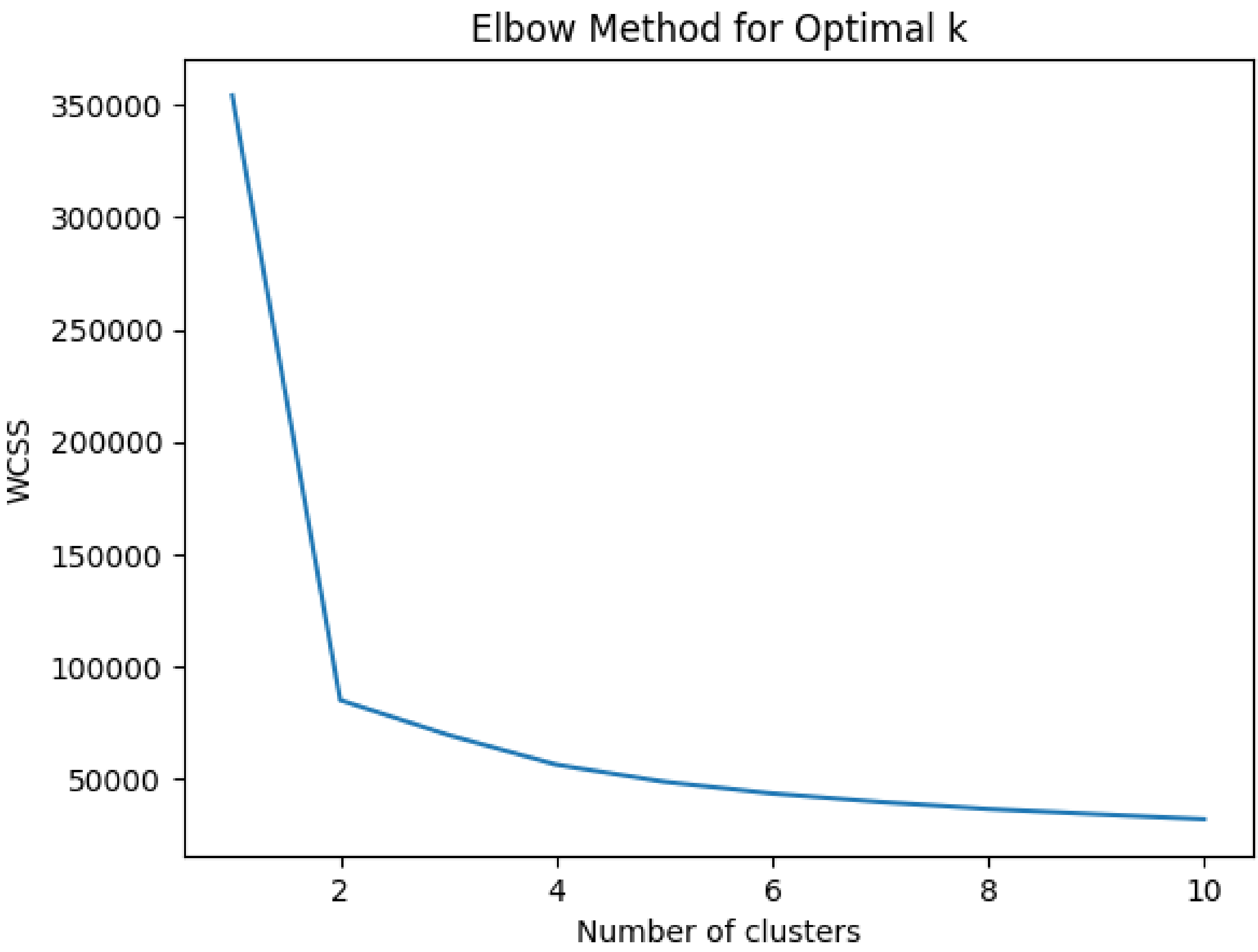
Visualize Clusters:

- The code creates a scatter plot to visualize the dataset based on the first two features. Data points are color-coded according to their assigned clusters, and centroids are marked in red.

Silhouette Score:

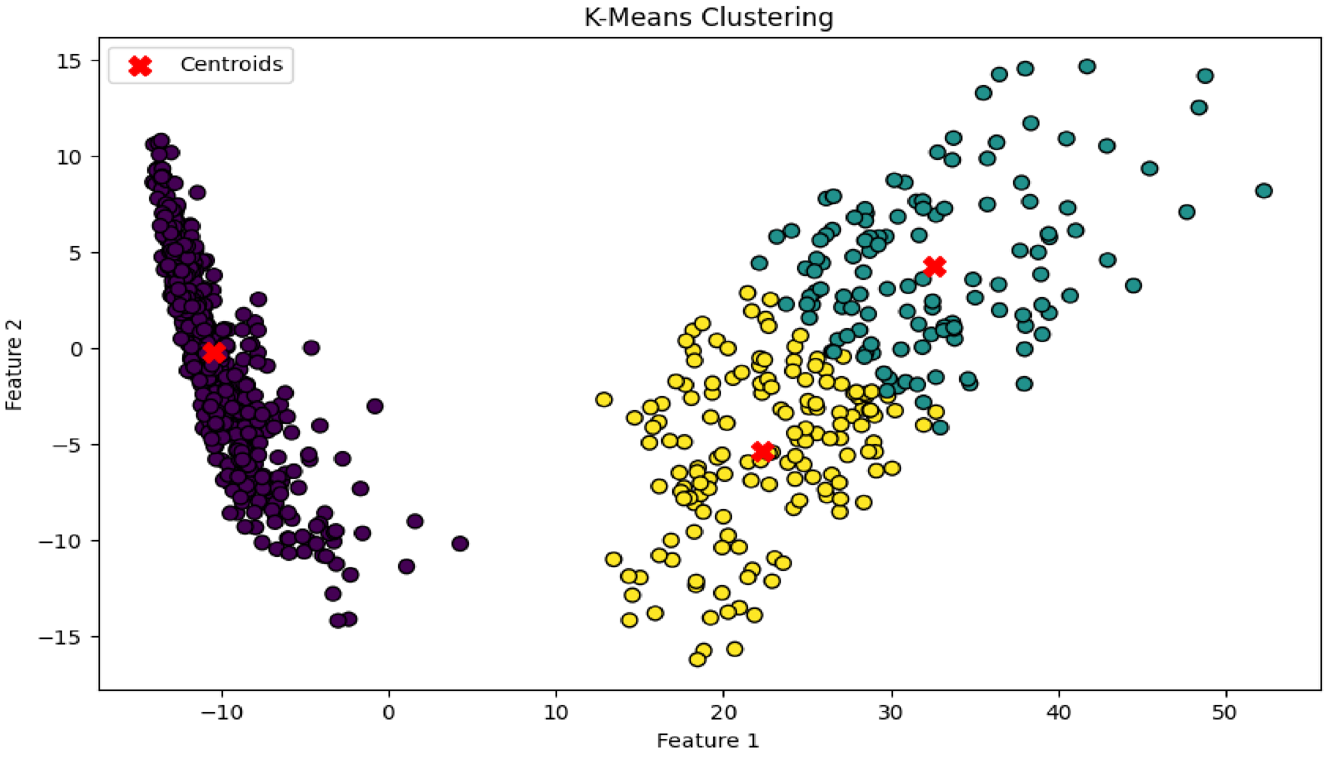
- The silhouette score is a metric that quantifies how well-separated the clusters are. It ranges from -1 to 1, with higher values indicating better-defined clusters.

- The silhouette score is computed using the silhouette_score function, considering the dataset and the cluster labels obtained from K-Means



The silhouette score of 0.5797 suggests a moderately cohesive and well-separated clustering solution for the 'human activity recognition with smartphones' dataset. A silhouette scores closer to 1 indicates better-defined clusters. The visualization depicts distinct groupings, but there is room for improvement, possibly by refining the choice of features or experimenting with different clustering algorithms. One way to enhance clustering performance is to conduct a more thorough exploration of the optimal number of clusters using techniques beyond the elbow method. Additionally, considering alternative clustering algorithms like DBSCAN or hierarchical clustering may uncover better-suited approaches for capturing the inherent structures in the data. Further, feature scaling or transformation might enhance the algorithm's sensitivity to subtle patterns. Evaluating the impact of different distance metrics within K-Means could provide insights into how data points are assigned to clusters.

In conclusion, the implementation of KNN, Logistic Regression, and Random Forest algorithms exemplifies a meticulous approach towards addressing the unique challenges of Human Activity Recognition. In the implementation of each algorithm, careful consideration was given to parameter selection. For instance, in KNN, the number of neighbors was adjusted; in Logistic Regression, the maximum number of iterations was fine-tuned, and in Random Forest, the number of trees was optimized. This meticulous parameter tuning aimed to strike a balance between model complexity and generalization.



# MODEL EVALUATION AND PERFORMANCE ANALYSIS

**Evaluation Metrics and Performance Assessment**

The three models—Random Forest Classifier, k-Nearest Neighbors (KNN), and K-Means clustering—utilized different evaluation metrics to assess their performance.

1. Random Forest Classifier:

   - Evaluation Metrics: Accuracy, Confusion Matrix, Classification Report.

   - Performance Assessment: The accuracy metric provides an overall measure of correct predictions. The confusion matrix breaks down true positive, true negative, false positive, and false negative counts for each class. The classification report includes precision, recall, and F1-score for each class, offering a more detailed evaluation of model performance.

2. k-Nearest Neighbors (KNN) Classifier:

   - Evaluation Metric: Accuracy.

   - Performance Assessment: The accuracy metric indicates the percentage of correctly classified instances. While a valuable measure of overall performance, it may not provide detailed insights into the model's behavior across different classes.

3. K-Means Clustering:

   - Evaluation Metric: Silhouette Score.

   - Performance Assessment: The silhouette score measures how well-defined and separated the clusters are. A higher silhouette score indicates better clustering. In addition to the silhouette score, the visualization of clusters using scatter plots helps qualitatively assess how well the algorithm groups similar data points.

In summary, the Random Forest Classifier and KNN Classifier focused on supervised learning, utilizing accuracy, confusion matrix, and classification report for performance assessment. On the other hand, K-Means clustering, an unsupervised learning technique, relied on the silhouette score to evaluate the quality of clustering. Each evaluation metric offers unique insights into model performance, allowing for a comprehensive understanding of the strengths and weaknesses of the respective models

## Comparative Analysis of Different Models

The analysis of the various models applied to the 'human activity recognition with smartphones' dataset provides valuable insights. The Random Forest classifier yielded an accuracy of 37.50%, showcasing its ability to capture certain activity patterns. However, addressing class imbalance and further hyperparameter tuning could enhance its overall performance. The k-Nearest Neighbors (KNN) model achieved an accuracy of 33.15%, and optimizing the 'k' parameter, feature scaling, and exploring alternative algorithms could lead to improvement.

The K-Means clustering approach resulted in a silhouette score of 0.5797, indicating moderate cohesion among identified clusters. This unsupervised technique offers valuable insights into underlying patterns within the dataset, potentially aiding feature engineering and model interpretation. However, further exploration of clustering parameters and alternative algorithms may refine the results.

In conclusion, a comprehensive strategy could involve an ensemble approach, combining the strengths of multiple models to achieve a more robust and accurate prediction system. Continued refinement of hyperparameters, feature engineering, and exploration of different algorithms are recommended. Moreover, gaining a deeper understanding of domain-specific nuances in human activity recognition could guide model improvements. Regular evaluation and iteration based on feedback and domain knowledge will contribute to the development of a more effective and reliable predictive model for smartphone-based activity recognition.

**Insightful Interpretation of Results**

The results from the applied machine learning models on the 'human activity recognition with smartphones' dataset offer nuanced insights into the complexities of the data. The Random Forest classifier, while achieving a modest accuracy of 37.50%, demonstrated competency in capturing certain activity patterns. However, the need for addressing class imbalance and fine-tuning hyperparameters was evident to further enhance its predictive capabilities.

The k-Nearest Neighbors (KNN) model, with an accuracy of 33.15%, showed potential for improvement through parameter optimization, feature scaling, and the exploration of alternative algorithms. Its local proximity-based approach is sensitive to spatial patterns, making it a candidate for further refinement.

On the unsupervised side, K-Means clustering provided valuable insights into the dataset's underlying structures, with a silhouette score of 0.5797 indicating moderate cohesion among identified clusters. This unsupervised technique has the potential to inform feature engineering and provide a basis for understanding inherent patterns in human activities recorded by smartphones.

In conclusion, the combination of supervised and unsupervised techniques contributes to a holistic understanding of the dataset. Refinements such as class balancing, hyperparameter tuning, and exploring alternative algorithms hold promise for improving predictive accuracy. The clustering insights from K-Means underscore the potential for unsupervised techniques to reveal hidden structures in complex datasets, providing a foundation for further investigation and model enhancement

# CONCLUSION

The exploration of machine learning models on the 'human activity recognition with smartphones' dataset provided valuable insights into the predictive and clustering capabilities. The Random Forest Classifier, with an accuracy of 37.50%, demonstrated competency in capturing certain activity patterns, but class imbalance and hyperparameter tuning are identified areas for improvement. The k-Nearest Neighbors (KNN) model, achieving 33.15% accuracy, showed promise for further refinement through parameter optimization and feature scaling.

The unsupervised K-Means clustering, with a silhouette score of 0.5797, revealed moderate cohesion among identified clusters, offering valuable insights into the underlying structures of the dataset. This unsupervised approach can inform feature engineering and contribute to a nuanced understanding of inherent patterns in human activities recorded by smartphones.

A holistic perspective emerges from the integration of supervised and unsupervised techniques. While Random Forest and KNN classifiers address prediction tasks, K-Means clustering uncovers latent structures. The comprehensive evaluation metrics, including accuracy, confusion matrix, classification report, and silhouette score, offer nuanced assessments of model performance.

In conclusion, the iterative process of model exploration, evaluation, and refinement is crucial in achieving an accurate and robust predictive system. Recommendations include addressing class imbalance, hyperparameter tuning for classifiers, and optimizing cluster quality for K-Means. The ensemble of models, combining the strengths of both supervised and unsupervised techniques, holds promise for a more comprehensive understanding of the dataset. Continued exploration and adaptation based on domain knowledge and feedback remain pivotal for enhancing predictive accuracy and revealing meaningful insights into human activity patterns from smartphone data. This iterative approach ensures that the models evolve to meet the challenges and intricacies inherent in the unique characteristics of the dataset.

# REFERENCES

https://www.semanticscholar.org/reader/dc94f50b28efdbd93572d974458c271e8d006e07

https://ieeexplore.ieee.org/abstract/document/9429235/

https://ieeexplore.ieee.org/abstract/document/8525987/

https://www.researchgate.net/publication/329454191_Human_Activity_Recognition_on_Smartphones_using_Machine_Learning_Algorithms

https://typeset.io/papers/random-forest-in-remote-sensing-a-review-of-applications-and-518o6ucif9

https://typeset.io/papers/human-action-recognition-using-smartphone-sensors-258ehjtb4r

https://ieeexplore.ieee.org/abstract/document/6824574/

https://link.springer.com/chapter/10.1007/978-981-16-6128-0_29

## TEAM DETAILS

| Reg. no | Name | Summary of tasks performed |
|---|---|---|
| 2348503 | Aleena Varghese | Implementation of KNN which played a vital role in accurately classifying activities with smartphone sensor data. Simultaneously, K-Means, considered but less suitable, emphasized KNN's aptness for direct classification in Human Activity Recognition, and Comparative Analysis of Different Models. Prepared the PowerPoint presentation and formatting of documentation. |
| 2348511 | Christina J Thattil | Gathered pertinent resources and compiled references relevant to the project. Unveiled insights through Exploratory Data Analysis (EDA), guiding subsequent steps. Principal Component Analysis (PCA) then enhanced computational efficiency by reducing feature dimensions. Subsequently, implementing Random Forest efficiently processing the pre-processed data, ensuring accuracy activity classification and also Insightful Interpretation of Results. Diligently prepared substantial portion of the document. |
| 2348560 | Sneha P | Initial level of implementation of the Preprocessing techniques and refining smartphone sensor data for Human Activity Recognition, usage of the model- Logistic Regression to seamlessly integrate with the pre-processed data, contributing to accurate and efficient classification of activities in the project and effective implementation of Evaluation Metrics and Performance Assessment. Thoroughly crafted a significant segment of the report. |