# ActivitySense: Classification of Human Activities Using Smartphone Sensor Data

## Master of Computer Engineering for Robotics and Smart Industry

## Course: Machine Learning & Artificial Intelligence

**Shakiba Sharifi, Arash Gholami**

**Academic Year: 2023**

## Contents

## Abstract:

This project aims to develop a machine learning model for accurately classifying human activities using smartphone sensor data. The dataset consists of recordings from individuals performing various activities of daily living, captured through wearable smartphones [1]. The objective is to categorize activities such as walking, sitting, standing, and more. The project applies data preprocessing, exploratory data analysis, and feature engineering techniques. Several supervised learning algorithms are implemented to train the model, enabling it to accurately classify human activities. The successful classification of activities has significant implications in industries such as fitness tracking and healthcare, where it can be used for behavior monitoring and remote patient care. The project contributes to the field of human activity recognition and provides a foundation for further research and advancements in this domain.

**Keywords:** human activity recognition, smartphone sensor data, classification, machine learning, data preprocessing, feature engineering, supervised learning.

## 1. Motivation:

The increasing prevalence of smartphones equipped with inertial sensors has opened up new possibilities for human activity recognition. Accurately classifying human activities from smartphone sensor data holds great importance in various industries and healthcare domains. In industries such as fitness tracking, understanding user activities can provide personalized recommendations and behavior monitoring. In healthcare, activity recognition can enhance remote patient care and monitoring by analyzing activity patterns and detecting anomalies. This project aims to develop a machine learning model that can effectively classify human activities based on smartphone sensor data. By doing so, it aims to contribute to advancements in behavior analysis and monitoring, leading to the development of innovative applications and services that rely on activity recognition. The outcomes of this project have the potential to significantly impact individuals' daily lives and support healthcare professionals in delivering improved care.

## 2. Objectives

The objectives of this project are centered around developing a machine-learning model to accurately classify human activities using smartphone sensor data. The dataset used in this project was built from the recordings of 30 study participants performing activities of daily living (ADL) while wearing a waist-mounted smartphone with embedded inertial sensors. The primary goal is to classify activities into one of the six performed activities: WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, and LAYING.

The dataset captures 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz using the smartphone's embedded accelerometer and gyroscope. The sensor signals were pre-processed by applying noise filters and sampled in fixed-width sliding windows. Each window has a duration of 2.56 seconds and a 50% overlap, resulting in 128 readings per window. To separate the body motion components, a Butterworth low-pass filter was used to extract body acceleration and gravity from the sensor acceleration signal [2][3].

The dataset contains a total of 561 features, which are derived from variables calculated from the time and frequency domain of the sensor signals. These features provide information about the triaxial acceleration, estimated body acceleration, and triaxial angular velocity.

It is worth noting that the dataset has been randomly partitioned into two sets: a training set, comprising 70% of the volunteers, and a test set, comprising the remaining 30%. This partitioning allows for the evaluation of the model's performance on unseen data.

Given the characteristics of the dataset, including the presence of both time and frequency domain features, as well as the combination of accelerometer and gyroscope data, the objective is to apply various machine learning techniques to accurately classify the activities performed by individuals. By exploring different supervised learning algorithms, performing feature engineering, and evaluating the model's performance using appropriate metrics, the project aims to determine the most effective approach for classifying human activities based on smartphone sensor data.

## 3. Methodology

The methodology employed in this project encompasses a series of distinct steps aimed at achieving the objectives outlined. Firstly, the data preprocessing phase involves loading the training and test datasets, meticulously handling missing values, checking for duplicates, and ensuring data quality. Through exploratory data analysis, an in-depth examination of the distribution of activities is conducted, visualizing the class distribution and identifying potential patterns or correlations present within the sensor data.

### 3.1. Class distribution in dataset

The distribution of different classes in a dataset plays a crucial role in machine learning and statistical modeling. An evenly distributed dataset, where each class has a similar number of samples, is generally considered desirable for several reasons:

- **Balanced Training:** When training a machine learning model, having an even distribution of samples across classes helps ensure that the model receives sufficient exposure to each class. This balance allows the model to learn and generalize well across all classes, leading to better overall performance.

- **Avoiding Bias:** Imbalanced class distributions can introduce bias in the model's learning process. If one or more classes are significantly underrepresented, the model may not receive enough examples to learn their patterns effectively. This can result in the model being biased towards the majority class, leading to poor performance on minority classes.

- **Evaluation Accuracy:** An imbalanced dataset can mislead the evaluation of a model's performance. Accuracy alone may not be a reliable measure when classes are imbalanced. For example, in a binary classification problem with 95% of samples belonging to class A and only 5% belonging to class B, a model that predicts class A for all samples would achieve a high accuracy of 95%. However, this model would fail to capture the patterns of class B. By having a balanced dataset, we can obtain a more accurate evaluation of the model's performance across all classes.

Now, referring to the following pie chart (figure 1), the chart shows a reasonably distributed representation of the different classes, it indicates that the dataset already has a close-to-balanced distribution. In such case, there is not a strong need to change or augment the samples to achieve a more balanced distribution.

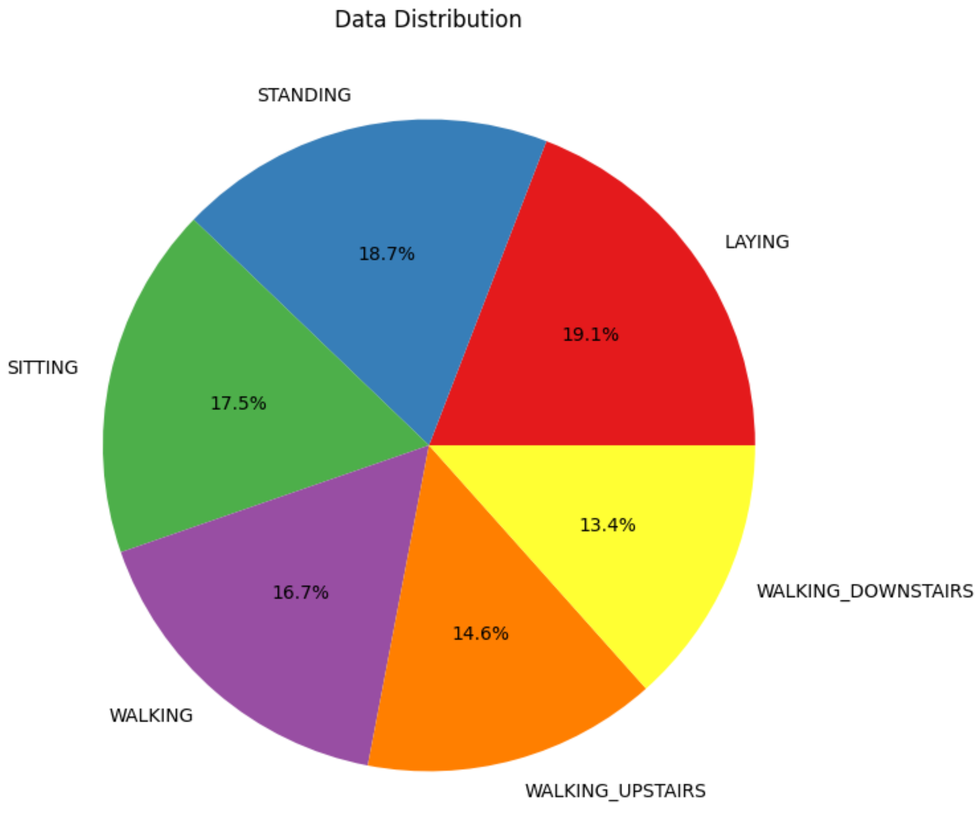

Figure 1 : The distribution of different classes in the dataset

## 3.2. Feature engineering

Feature engineering techniques are then applied to extract pertinent features from the sensor data, taking into consideration both time and frequency domain variables. Dimensionality reduction methods, such as Principal Component Analysis (PCA), are employed to effectively reduce the dimensionality of the data, facilitating more efficient processing and analysis.

Principal Component Analysis (PCA) is a widely used technique in this context. By projecting the data onto a lower-dimensional subspace, PCA effectively reduces the dimensionality of the dataset while preserving as much variance as possible. This reduction not only simplifies the data representation but also facilitates more efficient processing and analysis, particularly for machine learning algorithms.

By performing feature engineering and applying dimensionality reduction techniques like PCA, the sensor data can be transformed into a more compact and informative representation. This process enables the machine learning models to focus on the most relevant and discriminative aspects of the data, leading to improved accuracy, reduced computational complexity, and better interpretability of the results.

## 3.3. Machine learning algorithms

In this project, we have employed several machine learning algorithms to address the task of classifying human activities using smartphone sensor data. These algorithms were selected based on their suitability for supervised learning problems and their ability to handle the complexity and non-linearity of the dataset. The rationale behind using different algorithms and varying their settings is to explore different approaches and evaluate their performance in accurately classifying activities.

### 3.3.1. Gaussian Naive Bayes

We have chosen Gaussian Naive Bayes as one of our algorithms due to its simplicity and effectiveness in handling high-dimensional data. The algorithm operates on the assumption of feature independence, which simplifies the calculation of class probabilities. Gaussian Naive Bayes assumes that the features follow a Gaussian distribution and utilizes Bayes' theorem to calculate the posterior probabilities of different classes given the observed features. Despite its assumption of independence, Naive Bayes can still provide competitive performance in many real-world scenarios, particularly when the feature independence assumption holds reasonably well.

### 3.3.2. Logistic Regression

Logistic Regression is a widely used linear classification algorithm that models the probability of an instance belonging to a specific class. We have implemented logistic regression with the 'saga' solver, which is suitable for large datasets. Logistic regression is well-suited for binary classification problems, but it can also be extended to handle multi-class problems using techniques such as one-vs-rest (OvR). OVR (One-Vs-Rest) is a strategy used in multi-class classification to extend binary classifiers to handle multiple classes. In OVR, a separate binary classifier is trained for each class, treating it as the positive class and the remaining classes as the negative class. During prediction, the class with the highest probability from the binary classifiers is assigned to the instance. OVR is a simple and intuitive approach that allows us to leverage the strengths of binary classifiers for multi-class problems. The algorithm learns the coefficients of the linear decision boundary that separates different classes by optimizing a logistic loss function. By applying appropriate regularization techniques, logistic regression can effectively handle overfitting and provide interpretable results.

### 3.3.3. Support Vector Machines (SVM)

SVMs are powerful algorithms for both binary and multi-class classification tasks. In our project, we have utilized two variants of SVM: LinearSVC and SVC with different kernels. LinearSVC employs a linear kernel and is suitable for datasets with linearly separable classes. It maximizes the margin between classes while minimizing classification errors. On the other hand, SVC with kernels such as sigmoid, polynomial, and radial basis function (RBF) allows for non-linear separation boundaries. The decision function shape parameter is set to OvR to apply the one-vs-rest strategy for multi-class classification. The C parameter controls the regularization strength, where smaller values impose stronger regularization. SVMs can handle complex decision boundaries and are effective when dealing with datasets with a large number of features.

### 3.3.4 K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm that classifies instances based on their proximity to labeled instances in the training set. We have implemented KNN with two different distance metrics: Euclidean distance and Manhattan distance. KNN operates on the principle that instances belonging to the same class tend to be close to each other in the feature space. By considering the majority class among the k nearest neighbors, KNN assigns a class label to the test instance. The choice of distance metric, such as Euclidean or Manhattan, allows for flexibility in capturing different patterns in the data. KNN is intuitive and simple to implement, making it a popular choice for classification tasks.

By incorporating these diverse algorithms with different settings and hyperparameters, we aim to explore their performance characteristics and identify the most effective approach for accurately classifying human activities. The selection of various algorithms allows us to consider their strengths and limitations, leading to a more comprehensive evaluation. By evaluating the results using metrics such as accuracy, precision, and recall, we can gain insights into the algorithms' performance on our specific dataset. This comparative analysis enables us to make informed decisions about selecting the algorithm(s) that offer the highest classification accuracy and align with the project's objectives.

## 4. Experiments and Results

In the evaluation phase, the performance of the implemented machine learning algorithms was assessed using various evaluation metrics. These metrics provide insights into the accuracy and effectiveness of the models in classifying human activities based on smartphone sensor data.

The evaluation process involved splitting the dataset into training and testing subsets using stratified sampling to ensure representative class distributions in both sets. The algorithms were then trained on the training data and evaluated on the testing data.

This process is applied on both original data and the data which PCA is applied on it to compare the difference of the result.

- **Accuracy:** It measures the overall correctness of the model's predictions by calculating the ratio of correctly classified instances to the total number of instances. A higher accuracy score indicates better performance.

- **Precision:** It measures the ability of the model to correctly identify instances of a particular class. Precision is the ratio of true positives to the sum of true positives and false positives. It gives an indication of the model's ability to minimize false positives.

- **Recall:** Also known as sensitivity or true positive rate, it measures the ability of the model to correctly identify instances of a particular class among all the instances that belong to that class. Recall is the ratio of true positives to the sum of true positives and false negatives. It gives an indication of the model's ability to minimize false negatives.

By considering these evaluation metrics, the project assessed the performance of each algorithm in terms of accuracy, precision, and recall. These metrics provide a comprehensive understanding of the strengths and weaknesses of the models and their ability to accurately classify human activities.

## 4.1. Comparison of Execution Time: PCA Reduced Data vs. Original Data

The initial dataset consists of 561 features, while the PCA-reduced dataset contains only 157 features. Figure 2 demonstrates that there is a significant difference in execution time for logistic regression between the original data and the PCA-reduced data. However, for the remaining algorithms, the use of PCA does not have a significant effect on the execution time.

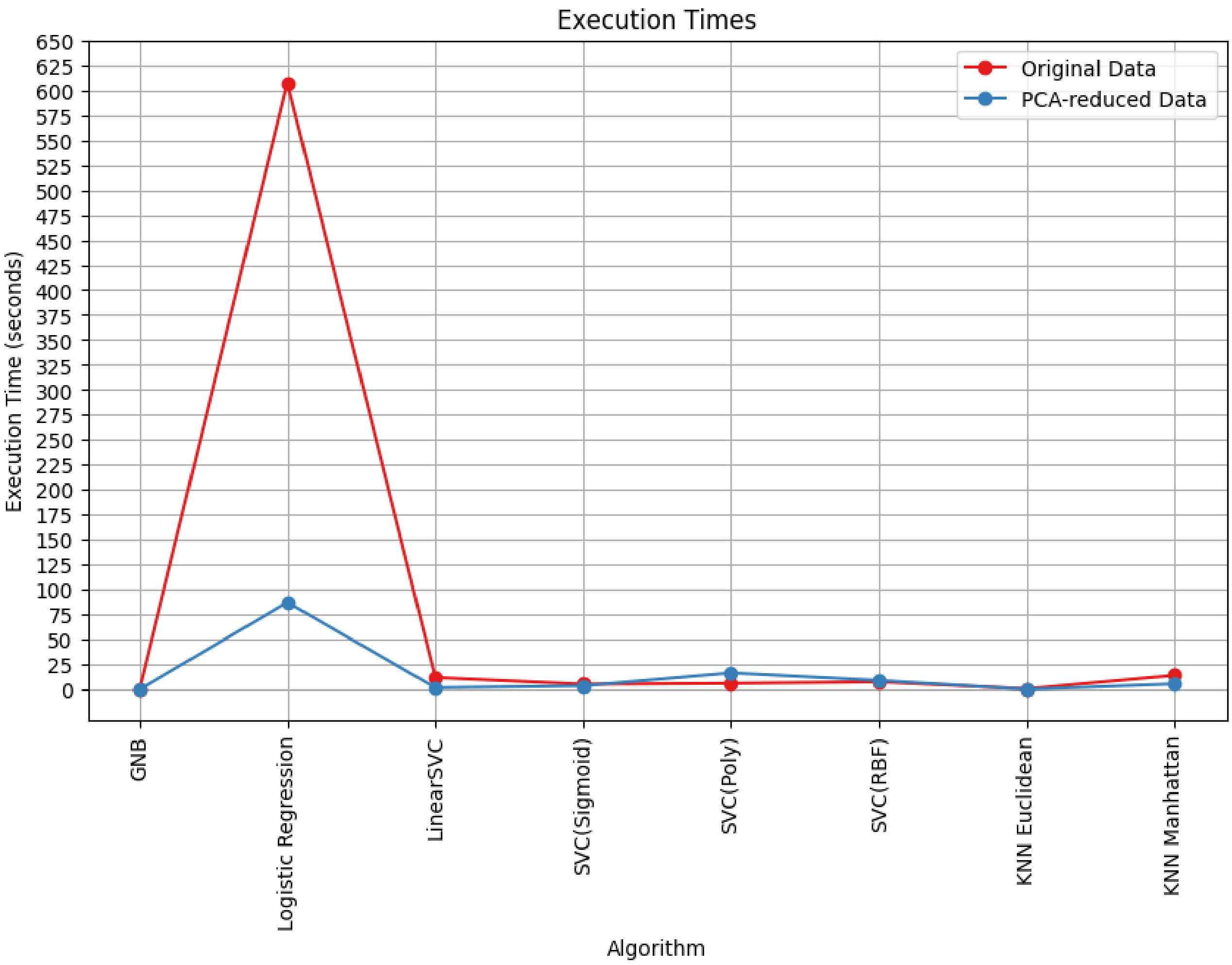

Figure 2: Comparison of Execution Time

## 4.2. Evaluation Metrics for Original Data

The evaluation of the implemented algorithms yielded the following results. The accuracy values ranged from 73.01% (GNB) to 98.20% (Logistic Regression), indicating the overall correctness of the predictions made by the models. Precision scores varied between 78.41% (GNB) and 98.40% (LinearSVC), reflecting the algorithms' ability to accurately identify instances of specific classes. The recall scores, which measure the algorithms' capability to correctly identify instances of a particular class, ranged from 73.01% (GNB) to 98.20% (Logistic Regression).

Examining the results, it is evident that Logistic Regression and LinearSVC achieved the highest accuracy, precision, and recall scores, surpassing other algorithms in performance. GNB, although displaying lower scores, still demonstrated reasonable accuracy, precision, and recall. The SVC algorithms with different kernels (Sigmoid, Poly, and RBF) also exhibited satisfactory performance, consistently maintaining high scores across the evaluation metrics. The KNN algorithms, employing Euclidean and Manhattan distance measures, achieved commendable accuracy, precision, and recall, placing them among the top-performing models.

The evaluation outcomes highlight the strengths and weaknesses of each algorithm in accurately classifying human activities using smartphone sensor data. The superior performance of Logistic Regression and LinearSVC makes them viable choices for this classification task. The SVC algorithms and KNN methods, despite slightly lower scores, still exhibit promising results, further reinforcing their suitability for the task at hand.

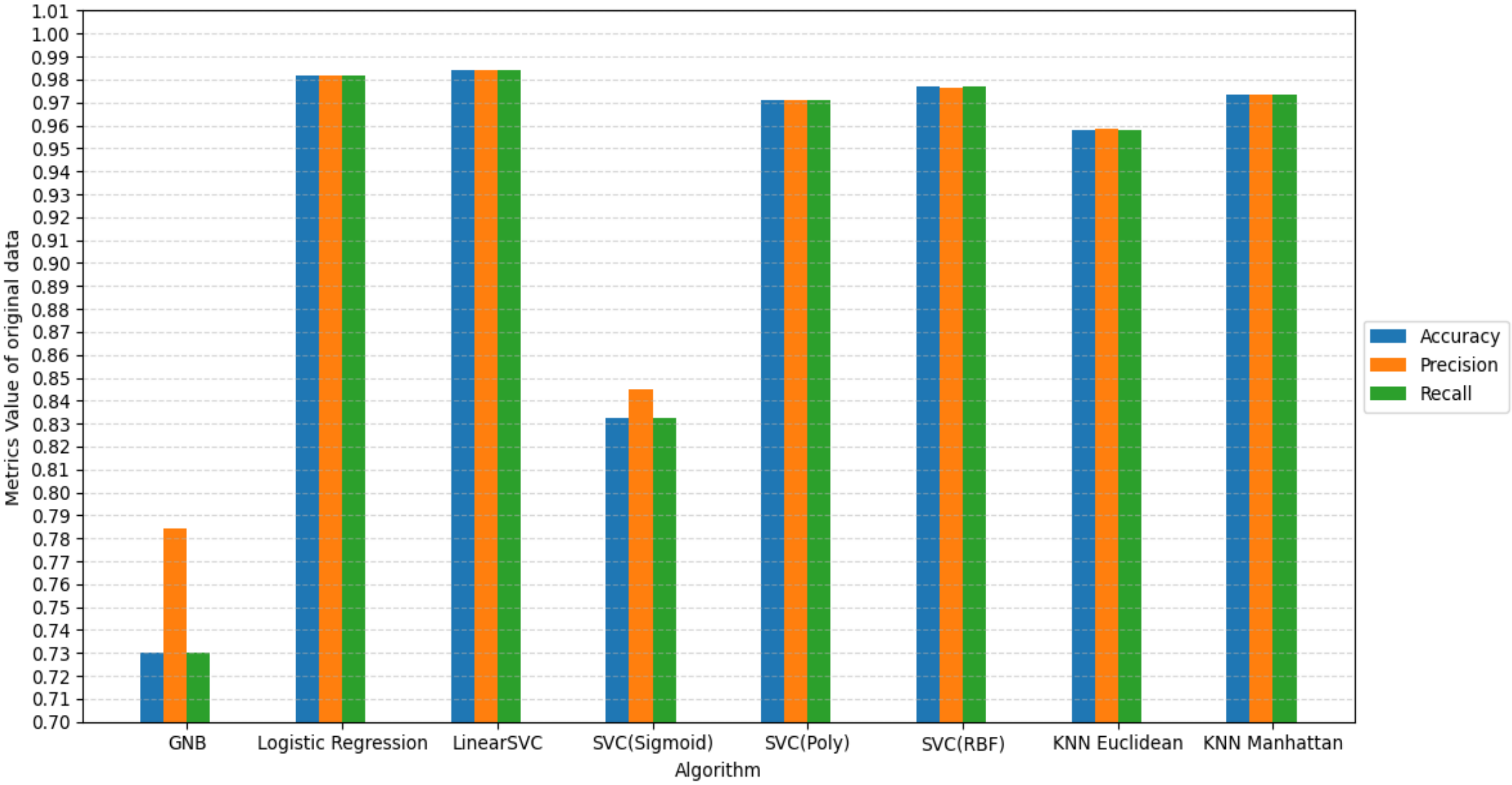

Figure 3: Evaluation Metrics for Original Data

It is important to note that these results are based on the specific dataset and evaluation metrics used in this project. In practice, the choice of algorithm may depend on the specific requirements and characteristics of the problem domain. Therefore, it is crucial to consider the context and make informed decisions when selecting an appropriate algorithm for activity classification using smartphone sensor data.

### 4.3. Evaluation metrics for PCA reduced Data

In this section, we present the evaluation metrics for the PCA reduced data, which underwent dimensionality reduction using Principal Component Analysis. The performance of the implemented algorithms was assessed using various metrics, including accuracy, precision, and recall, to gain insights into their effectiveness in classifying human activities based on smartphone sensor data.

The evaluation results for the PCA reduced data revealed interesting findings. Notably, the Gaussian Naive Bayes (GNB) algorithm exhibited improved performance compared to the original data, achieving an accuracy of 83.69%, precision of 84.12%, and recall of 83.69%. This demonstrates a substantial increase in accuracy by approximately 10.68% when compared to the results obtained using the original data. Similarly, other metrics such as precision and recall also showed considerable improvements.

Analyzing the Logistic Regression algorithm, we observe a marginal improvement in performance with the PCA reduced data. The accuracy, precision, and recall increased by approximately 0.05%, indicating a slight enhancement in classification accuracy. These modest improvements may be attributed to the reduced dimensionality of the data, which helped the algorithm better capture the underlying patterns and make more accurate predictions.

Regarding the Linear Support Vector Classifier (LinearSVC), the performance with the PCA reduced data remained consistent with the original data. The accuracy, precision, and recall values remained almost unchanged, differing by a mere 0.10%. This suggests that dimensionality reduction did not significantly impact the performance of the LinearSVC algorithm, indicating its robustness to the reduction in feature space.

In contrast, the Support Vector Classifier with sigmoid kernel (SVC(Sigmoid)) demonstrated noticeable improvements when evaluated using the PCA reduced data.

The accuracy, precision, and recall increased by approximately 14.08%, showcasing the effectiveness of dimensionality reduction in enhancing the performance of this particular algorithm. The sigmoid kernel, which captures non-linear relationships, seems to benefit from the reduced feature space, resulting in more accurate classification outcomes.

For the Support Vector Classifier with polynomial kernel (SVC(Poly)) and Support Vector Classifier with RBF kernel (SVC(RBF)), the PCA reduced data led to a slight decrease in accuracy, precision, and recall when compared to the original data. The differences ranged from approximately 2.51% to 4.03%, indicating that the dimensionality reduction might have removed some valuable information necessary for these algorithms to achieve optimal performance.

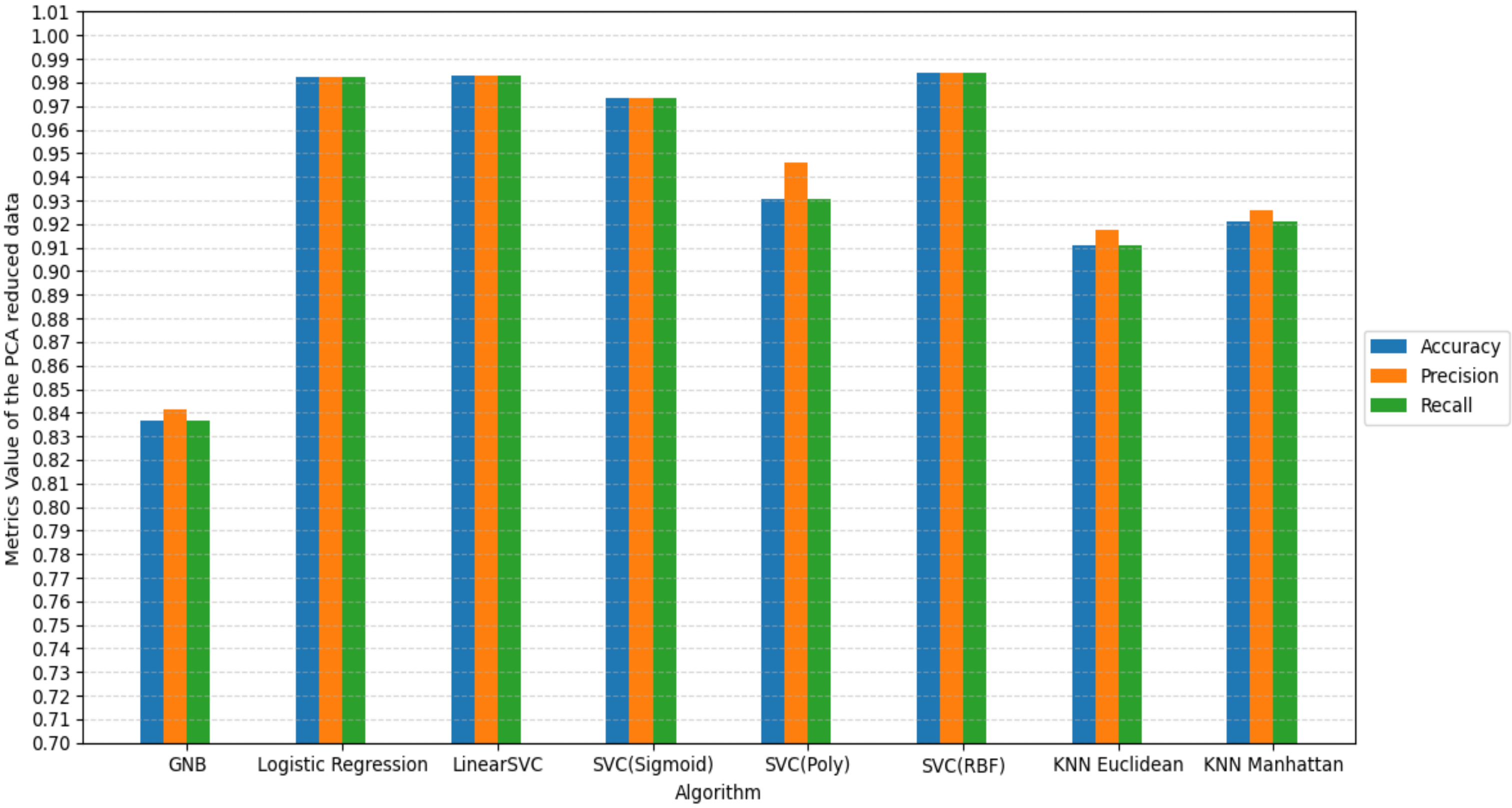

Figure 4: Evaluation metrics for PCA reduced Data

Lastly, the K Nearest Neighbor (KNN) algorithm with Euclidean and Manhattan distances exhibited mixed results. The KNN algorithm with Euclidean distance showed improved performance with the PCA reduced data, achieving higher accuracy, precision, and recall compared to the original data. On the other hand, the KNN algorithm with Manhattan distance displayed slightly decreased accuracy.

In conclusion, the evaluation metrics for the PCA reduced data revealed both improvements and slight performance variations across the implemented algorithms. The Gaussian Naive Bayes algorithm showcased significant enhancements in accuracy, precision, and recall. Additionally, the Logistic Regression algorithm exhibited marginal improvements, while the Linear Support Vector Classifier remained robust to dimensionality reduction. The Support Vector Classifier with sigmoid kernel demonstrated substantial performance gains, while the SVC(Poly) and SVC(RBF) algorithms experienced minor declines. The KNN algorithm demonstrated mixed outcomes, with the Euclidean distance variant benefiting from dimensionality reduction, and the Manhattan distance variant experiencing slight performance variations.

## 5. Conclusions

In conclusion, this project has endeavored to develop a machine-learning model for the precise classification of human activities utilizing smartphone sensor data. By employing advanced techniques and evaluating several supervised algorithms, our objective has been to achieve a high level of accuracy in activity classification. The initial evaluation results indicate that logistic regression shows promise in accurately classifying activities. However, further analysis and comparison of the implemented algorithms are required to determine the most effective approach.

The findings of this project contribute to the field of human activity recognition and lay the groundwork for future research and advancements in this area. By enhancing the accuracy of activity classification, we can facilitate the development of various applications and services that can leverage the understanding of human behavior and activity patterns. For instance, this technology can have implications in fields such as healthcare, sports performance analysis, and personalized user experiences.

The evaluation of the implemented algorithms has provided insights into their performance and strengths. Algorithms such as logistic regression, linear support vector classifier, support vector classifier with polynomial kernel and RBF, as well as K nearest neighbor with Manhattan distance, have demonstrated commendable results with accuracy, precision, and recall values exceeding 97%. Notably, the linear support vector classifier has consistently shown the highest performance, with an accuracy of nearly 98% across all metrics.

To further improve the accuracy and robustness of the classification model, future work can focus on exploring additional feature engineering techniques and fine-tuning the hyperparameters of the algorithms. Additionally, investigating ensemble methods that combine the strengths of multiple algorithms may lead to even better results.

Overall, this project provides a comprehensive exploration of machine-learning techniques for activity classification using smartphone sensor data. The insights gained from this research contribute to the growing body of knowledge in the field of human activity recognition. By continuing to refine and enhance the algorithms, we can unlock the full potential of activity classification, enabling a wide range of applications and services that benefit from a deeper understanding of human behavior and activity patterns.

## 6. References

1. Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra and Jorge L. Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2013. Bruges, Belgium 24-26 April 2013.

2. https://www.kaggle.com/code/essammohamed4320/human-activity-recognition-scientific-prespective#Human-Activity-Recognition

3. https://www.kaggle.com/code/morrisb/what-does-your-smartphone-know-about-you