Data Science Project Report
Title: Activity Estimation

Manshaa Kapoor (2021540) Mayank Pandey (2021264) Niharika Singh (2021545)

Ria Malhotra (2021412) Vishal Singh (2021575)

1. Problem Statement

The primary challenge in this task is to develop robust methods for accurately estimating activities from PPG signals despite the presence of motion artifacts during various daily activities. Existing public datasets present significant limitations, including short recording durations (typically 5 minutes per subject), limited activity scenarios (mainly treadmill walking or running), and controlled laboratory settings that fail to capture real-world conditions. The core technical objective is to process 8-second windows of PPG signal data sampled at 64 Hz, account for motion artifacts using accelerometer data sampled at 32 Hz, and estimate activities that align with the ground truth. Additionally, the solution must effectively handle transitions between activities, varying motion intensities, and maintain accuracy across diverse subjects and activity types.

2. Dataset Description [2]

Photoplethysmography (PPG) is a widely adopted technology for continuous heart rate monitoring in wearable devices. While PPG offers non-invasive heart rate measurement, its accuracy is significantly compromised by motion artifacts during daily activities, presenting a major challenge in real-world applications.

Key Components

1. Activities: 8 daily activities including
   (a) Stationary: sitting, working
   (b) Dynamic: walking, cycling, climbing stairs
   (c) Real-world: driving, lunch break, table soccer
   (d) Plus transition periods between activities
2. Sensor Data:
   (a) Wrist device (Empatica E4): PPG (64 Hz), Accelerometer (32 Hz)
   (b) Chest device (RespiBAN): ECG (700 Hz), Accelerometer (700 Hz)
3. Data Format:
   (a) Raw sensor data files per subject
   (b) Synchronized and labeled data in pickle format
   (c) Heart rate from ECG using 8-second windows with 2-second shifts

3. Challenges and Solutions - Github [1]

- Data Imbalance Across Activities
  - **Challenge:** Stationary activities (e.g., sitting, working) were overrepresented, while dynamic ones (e.g., table soccer, climbing stairs) were underrepresented, causing biased model predictions.
  - **Solution:** Used stratified sampling in train-test splits to ensure balanced activity representation.
- Scaling Feature Distributions
  - **Challenge:** The dataset included features with varying scales (e.g., heart rate, accelerometer values), which could bias models like Logistic Regression and MLPClassifier.
  - **Solution:** Implemented various scaling techniques, such as Standard Scaling, Min-Max Scaling, Robust Scaling, and Quantile Scaling, to normalize feature distributions.
- High Dimensionality of Raw Sensor Data
  - **Challenge:** The large number of raw features from PPG and accelerometer signals increased model complexity, leading to longer training times and overfitting.
  - **Solution:** Applied Principal Component Analysis (PCA) for dimensionality reduction, retaining components with maximum variance.

4. Hypothesis Tests

4.1. ANOVA Test

The F-statistic was 2374.0640, and the p-value was 0.0000.

ANOVA Results Interpretation:

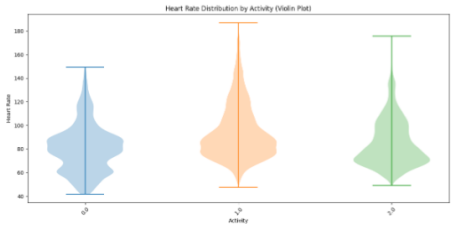

Figure 1.

- Since the p-value is less than the common significance level, we rejected the null hypothesis.
- This indicates that there is statistically significant evidence that the mean heart rates differ across the activity groups.

4.2. Chi-Square Test

A chi-square test was conducted on the dataset to check whether there exists an association between the activities performed and the heart rate.

Based on the chi-square test results:

Chi-square statistic: 6078.63 p-value: 0.0000000000 (extremely small, ¡ 0.05) Since the p-value is less than the common significance level, we rejected the null hypothesis. This suggests that there exists an association between the heart rate and activity type.

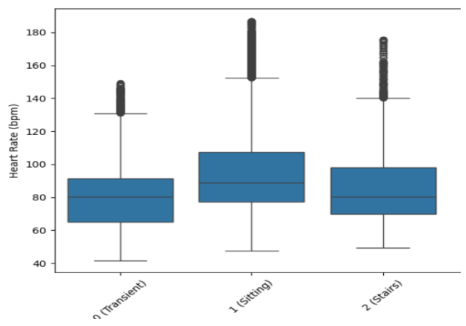

Figure 2.

4.3. Z Test

Two-tailed Z-tests were performed to compare mean heart rates across activity pairs, testing the null hypothesis that the means are equal against the alternative of significant differences.

Based on the Z test results:

All p-values were below 0.05, indicating significant differences in mean heart rates across activity pairs.

VALIDATION EXPERIMENT : Z Test

- The experiment validates the Z-test's accuracy by analyzing Type I (false positives) and Type II (false negatives) errors through simulations.
- Results show the Z-test reliably detects true differences (low Type II error) but occasionally exceeds the 0.05 threshold for false positives (Type I error).

5. Scaling Methods

- **Standard Scaling**
  Standard Scaling is a data preprocessing technique used to standardize the features of a dataset so that they have a mean of 0 and a standard deviation of 1. For each feature xi, the scaled value is computed as:

$$z = \frac{x_i - \mu}{\sigma}$$

- **Min-Max Scaling**
  A scaling technique that normalizes features to a fixed range, typically [0,1]. For each feature xi, the scaled value is computed as:

$$x^{'} = \frac{x_i - min(x)}{max(x) - min(x)}$$

- **Quantile Scaling**
  A non-linear scaling method transforms features to a specific distribution (e.g., uniform or normal) by first sorting the data and assigning percentile ranks, which are then mapped to the target distribution.
- **Robust Scaling**
  A technique that scales features using statistics that are robust to outliers, such as the median and the interquartile range (IQR). For each feature xi, the scaled value is computed as:

$$x^{'} = \frac{x_i - median(x)}{IQR}$$

5.1. PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique used to simplify large datasets while preserving as much variance as possible. It transforms the data into a new coordinate system, where the axes (principal components) are linear combinations of the original features and are ordered by the amount of variance they capture. PCA helps reduce complexity, remove redundancy, and improve visualization, while retaining the most critical information in the data.

6. Models

6.1. Logistic Regression

Logistic Regression is a statistical and machine learning algorithm used for binary and multi-class classification tasks. Despite its name, it is not used for regression tasks but to predict probabilities and classify data points into discrete classes. Logistic regression applies a logistic function (sigmoid) to the weighted sum of input features to predict a probability (between 0 and 1). Based on a threshold (e.g., 0.5), it classifies the data points into classes.

Observations:

- The No Scaling approach has the lowest accuracy, macro average, and weighted average scores, indicating that the model performs better with some form of scaling applied.
- Quantile Scaling, Standard Scaling, Min-Max Scaling, and Robust Scaling all show improvements in the accuracy, macro average, and weighted average scores compared to No Scaling.
- Among the scaling techniques, Quantile Scaling and Robust Scaling appear to have the highest accuracy, macro average, and weighted average scores, suggesting they may be the most effective scaling methods for this dataset and problem.

The Results for each scaling method and no scaling approach for Logistic Regression are: In summary, the results demonstrate that applying scaling

Metrics	No Scaling	Standard Scaling	Min - Max Scaling	Robust Scaling	Quantile Scaling
Accuracy	0.56	0.59	0.59	0.60	0.60
Precision	0.60	0.61	0.61	0.61	0.61
Recall	0.49	0.55	0.55	0.55	0.56
F1-Score	0.50	0.57	0.56	0.57	0.57

Figure 3.

techniques, such as Quantile Scaling, Standard Scaling, Min-Max Scaling, and Robust Scaling, can significantly improve the model's performance compared to the No Scaling approach.

6.2. Naive Bayes

Naive Bayes is a fast, simple probabilistic algorithm based on Bayes' theorem with feature independence, widely used for text classification, spam filtering, and sentiment analysis.

Why Used:

- Low computational complexity
- Works well with high-dimensional data
- Requires small training datasets
- Handles missing values effectively

Without PCA and Scaling

Scaling Method	Accuracy	Precision	Recall	F1-Score
No Scaling	0.54967	0.56	0.54	0.55
Standard Scaling	0.54967	0.56	0.54	0.55
Min-Max Scaling	0.54967	0.56	0.54	0.55
Robust Scaling	0.54967	0.56	0.54	0.55
Quantile Scaling	0.55858	0.56	0.56	0.55

Figure 4.

- Accuracy is identical (0.54967) for No Scaling, Standard, Min-Max, and Robust Scaling, with Quantile Scaling slightly higher (0.55858), showing minimal impact of scaling methods.

- Quantile Scaling achieves the highest recall (0.56), outperforming other methods (0.54), indicating better identification of positive instances.
- F1-scores (0.55) are consistent across all scaling methods, showing no significant improvement in precision-recall balance.

With PCA and Scaling techniques

PCA is used in Naive Bayes to reduce the dimensionality of the data before classification. By projecting high-dimensional data into a lower-dimensional space while preserving variance, PCA helps simplify the feature set, making the Naive Bayes model more efficient and less prone to overfitting. The reduced features, which capture the most important information, are then used as input to the Naive Bayes classifier, improving computational efficiency and handling multicollinearity effectively.

Observations:

Metrics	Standard Scaling	Min-Max Scaling	Robust Scaling	Quantile Scaling	No scaling
Accuracy	0.28	0.27	0.28	0.27	0.32
Precision	0.18	0.03	0.15	0.09	0.22
Recall	0.28	0.11	0.15	0.27	0,24
F-1 Score	0.18	0.05	0.12	0.12	0.20

Figure 5.

- The No Scaling approach achieves the highest accuracy (0.32), precision (0.22), and F1-score (0.20), suggesting that scaling may not always improve performance for this model and dataset. However, recall (0.24) is relatively lower compared to other methods.
- Standard Scaling and Robust Scaling produce similar accuracy scores (0.28), but Standard Scaling shows slightly better recall (0.28) and precision (0.18) compared to Robust Scaling (0.15 precision, 0.15 recall).
- Min-Max Scaling demonstrates the poorest performance across all metrics, with the lowest accuracy (0.27), precision (0.03), recall (0.11), and F1-score (0.05), indicating it may not be suitable for this problem.
- Quantile Scaling shows a balanced performance with accuracy (0.27) and the second-highest recall (0.27), though precision (0.09) and F1-score (0.12) remain low.

6.3. MLP Classifier

The MLPClassifier is a supervised learning algorithm from the neural network family in scikit-learn. It consists of an input layer, one or more hidden layers, and an output layer, with a fully connected architecture. The model is trained using backpropagation, optimizing weights with solvers like Adam, SGD, or LBFGS.

Why Used:

- Captures Nonlinear Patterns: Ideal for complex data relationships.
- Versatile: Works across various classification tasks.
- Boosted by Scaling: Performs better with scaled features.

Without PCA and Scaling

	precision	recall	f1-score	support
0.0	0.58	0.22	0.32	2949
1.0	0.60	0.97	0.74	7342
2.0	0.00	0.00	0.00	2649
accuracy			0.60	12940
macro avg	0.39	0.40	0.36	12940
weighted avg	0.47	0.60	0.50	12940

Figure 6.

With PCA and Scaling techniques

Observations:

Metrics	No Scaling	Standard Scaling	Min - Max Scaling	Robust Scaling	Quantile Scaling
Accuracy	0.82	0.90	0.85	0.90	0.87
Precision	0.83	0.90	0.85	0.90	0.86
Recall	0.81	0.91	0.87	0.91	0.88
F1-Score	0.82	0.90	0.86	0.90	0.87

Figure 7.

- Accuracy is lowest at 0.82 with no scaling, indicating that unscaled data leads to suboptimal model performance due to varied input feature distributions, which affects the model's ability to learn effectively.
- Standard and Robust Scaling achieve the highest accuracy of 0.90, with macro and weighted F1-Scores at 0.90, enhancing model generalization by standardizing feature distributions, which allows the model to learn more consistently from all features.
- Min-Max Scaling shows an accuracy of 0.85 with macro and weighted F1-Scores at 0.86 and 0.85, offering better performance than no scaling but not as effective as Standard and Robust Scaling, as it compresses data within a range that may not handle outliers well.
- Quantile Scaling achieves an accuracy of 0.87 with macro and weighted F1-Scores at 0.87, improving on no scaling but less effective than Standard and Robust Scaling due to its distribution-based approach, which can distort data if not uniformly distributed.

Overall, scaling improves model performance, with Standard and Robust Scaling providing the best results by ensuring uniform feature distributions, which helps in more stable and reliable training of the MLP classifier.

6.4. Decision Tree Classifier

Decision Tree Classifier is a supervised learning algorithm used for classification tasks. It builds a tree-like model where each internal node represents a decision based on a feature, and each leaf node represents a class label.

Why Used:

- Interpretability
- Handles Non-linear Data
- No Feature Scaling Required

Without PCA and Scaling

	precision	recall	f1-score	support
0.0	0.31	0.31	0.31	2949
1.0	0.61	0.61	0.61	7342
2.0	0.25	0.26	0.26	2649
accuracy			0.47	12940
macro avg	0.39	0.39	0.39	12940
weighted avg	0.47	0.47	0.47	12940

Figure 8.

With PCA and Scaling techniques Observations:

Metrics	No Scaling	Standard Scaling	Min - Max Scaling	Robust Scaling	Quantile Scaling
Accuracy	0.9487	0.9486	0.9486	0.9486	0.9484
Precision	0.95	0.95	0.95	0.95	0.95
Recall	0.95	0.95	0.95	0.95	0.95
F1-Score	0.95	0.95	0.95	0.95	0.95

Figure 9.

- Similar Accuracy Across Scaling Methods: All scaling techniques (Standard, Min-Max, Robust, Quantile) yield similar accuracy around 94.8%, indicating that scaling has minimal impact on model performance.

- Consistent Precision and Recall: Precision, recall, and F1-scores for all classes remain stable (0.92–0.97) across scaling methods, with minimal variation in model performance.
- High Performance on Majority Class: Class '1' (likely the majority class) shows excellent precision and recall (0.99), indicating strong performance on the most frequent class.
- Negligible Impact of Scaling on Class 2 and 3: Minor recall differences (e.g., for class 2 with Quantile Scaling) do not significantly affect overall model performance, and both macro and weighted averages are nearly identical across scaling techniques.

6.5. Random Forest Classifier

Random Forest Classifier is a supervised learning algorithm from the ensemble learning family in scikit-learn. It is composed of multiple decision trees, typically trained using the bagging method. Each tree is built on a random subset of the dataset and features, introducing diversity and reducing overfitting.

Why Used:

- Robustness to overfitting
- Handles nonlinear data by capturing intricate patterns
- Provides insights into feature contributions for interpretation

Metrics	No Scaling	Standard Scaling	Min - Max Scaling	Robust Scaling	Quantile Scaling
Accuracy	0.97	0.97	0.967	0.967	0.96
Precision	0.97	0.97	0.97	0.97	0.97
Recall	0.96	0.96	0.96	0.96	0.96
F1-Score	0.96	0.96	0.96	0.96	0.97

Figure 10.

Observations:

- Accuracy: The model achieves the highest accuracy (0.97) with No Scaling and Standard Scaling, while other scaling methods result in a slightly reduced accuracy (0.967 for Min-Max and Robust Scaling, 0.96 for Quantile Scaling). This indicates that the model is generally robust across scaling methods but performs marginally better without scaling or with Standard Scaling.
- Precision: Precision remains consistently high at 0.97 across all scaling methods, suggesting that the model effectively identifies true positives irrespective of feature scaling.
- Recall: Recall is consistent at 0.96 across all scaling techniques, indicating the model's ability to capture true positives remains unaffected by scaling.

The model demonstrates robustness across scaling methods, with minimal performance differences. However, No Scaling and Standard Scaling yield slightly better accuracy, making them preferred choices. Quantile Scaling improves F1-Score but doesn't enhance accuracy.

7. Small Dataset vs Our Dataset

7.1. Logistic Regression

Scaling Technique	Small Dataset Accuracy	Running Time	Large Dataset Accuracy
No Scaling	0.833333	0.0132432	0.56
Standard Scaler	0.833333	0.007303	0.59
Robust Scaling	0.833333	0.00522995	0.60
Min Max Scaling	0.733333	0.0145316	0.59
Quantile Scaling	0.733333	0.0164938	0.60

Figure 11.

Scaling Technique	Small Dataset Accuracy	Running Time	Large Dataset Accuracy
No Scaling	0.866667	0.374468	0.82
Min Max Scaling	0.866667	0.448871	0.84
Standard Scaler	0.833333	0.586755	0.84
Robust Scaling	0.833333	0.423964	0.84
Quantile Scaling	0.766667	0.151891	0.85

Figure 12.

7.2. MLP Classifier

Observations:

- Accuracy Comparison: The MLP Classifier shows a notable improvement in accuracy on larger datasets (up to 0.85 with Quantile Scaling), while Logistic Regression sees a drop in accuracy on large datasets (best at 0.60 with Robust and Quantile Scaling).
- Running Time: Logistic Regression is faster with Robust Scaling and Standard Scaler, whereas MLP Classifier has the shortest running time with Quantile Scaling.

Overall, the MLP Classifier performs better on larger datasets than Logistic Regression, especially with Quantile Scaling. However, Logistic Regression is generally more efficient in terms of running time for smaller datasets.

8. Conclusion

- **Scaling Enhances Performance for Most Models**
  - Logistic Regression, MLPClassifier, and Naive Bayes show improved performance with scaling methods, as scaling helps normalize feature distributions and ensures that all features contribute equally to the model.
  - Models like Decision Trees and Random Forests are robust to scaling, with little to no improvement observed.
- **Quantile Scaling and Robust Scaling**
  - Logistic Regression: Quantile Scaling and Robust Scaling stand out as the most effective, offering the highest accuracy, macro average, and weighted average scores.
  - Naive Bayes: Quantile Scaling slightly improves recall and accuracy, making it marginally better than other methods.
  - MLPClassifier: Quantile Scaling improves accuracy and F1-Scores but is outperformed by Standard and Robust Scaling.
- **Standard Scaling and Robust Scaling Are Generally Reliable**
  - MLPClassifier: These scaling methods achieve the best accuracy and F1-Scores, making them ideal for datasets with diverse feature ranges.
  - Logistic Regression: Standard Scaling is consistently effective, offering significant improvement over no scaling.

9. Future Work

- **Evaluation and Benchmarking**
  - Cross-Dataset Validation: Test models on other public PPG datasets to evaluate generalizability and robustness across datasets with different demographics and conditions.
  - Explainability and Trust: Incorporate interpretable AI techniques to understand model decisions, particularly in cases of false positives or false negatives.
- **Personalized Estimation**
  - Subject-Specific Modeling: Develop personalized models that adapt to individual physiological variations, leveraging transfer learning or fine-tuning.

References

[1] https://github.com/mayank2021264/Activity$_{E}stimation_{B}asedOnHeartRate$.

[2] REISS, ATTILA, I. I., AND SCHMIDT, P. PPG-DaLiA. UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C53890.