Statistical Theory - Final Project: Gym Members Exercise Tracking

Roni Shine Zilberberg 215229212
Hoshen Maimon 330941154

August 2025

Abstract

In this project, we applied several supervised models to predict gender from gym exercise data. Although we tried multiple algorithms with proper tuning, gender could not be accurately predicted from behavioral features, suggesting a high degree of gender equality in exercise patterns. We then applied dimensionality reduction (PCA and UMAP) and several clustering algorithms to explore latent group structures in the data. Clustering quality was evaluated using silhouette scores, and we tested for statistical association with gender using ANOVA and Kruskal-Wallis tests. The results showed that the clusters did not align strongly with gender, confirming the earlier finding. The goal of the study is to compare supervised and unsupervised models, and to assess whether gender-related structure emerges naturally from exercise behavior. The code is available at: https://github.com/ronizil/Statistical-Theory.git.

1 Introduction

Understanding patterns of physical exercise behavior can offer meaningful insights into population-wide fitness habits, behavioral variability, and underlying structure in health-related data. In this study, we analyze a dataset containing behavioral and body composition measurements from gym members, including variables such as resting heart rate, calories burned, body fat percentage, and BMI. The dataset [1] includes self-reported gender, which is used only for evaluation in the unsupervised phase and not as input during modeling.

We begin with a supervised learning analysis, training multiple classification models [10] to predict gender based solely on exercise-related features. Despite proper preprocessing and algorithm tuning, the models exhibited limited predictive accuracy. This result suggests that gender is not strongly encoded in the observed behavioral patterns, consistent with the hypothesis of behavioral parity between male and female members in this context.

To further explore the data's latent structure, we applied unsupervised learning methods [27], including dimensionality reduction [9] (PCA [11] and UMAP [12]), followed by clustering algorithms (K-Means [21], Gaussian Mixture and Hierarchical Clustering). Clustering quality was assessed using internal validation via the silhouette score [24] and the Elbow method using Loss function score [25]. In addition, we evaluated whether the resulting cluster structures exhibit implicit alignment with gender using statistical tests (One-way ANOVA [6] and Kruskal-Wallis [7]).

We had several goals in this study: the first was to compare the performance of supervised and unsupervised models in identifying gender-related patterns; the second was to examine whether gender-based structure could naturally emerge from behavioral patterns, even when gender was not explicitly included in the models. Throughout the project, we conducted exploratory data analysis, applied supervised and unsupervised learning methods, performed dimensionality reduction, ran statistical tests, and generated visual representations.

2 Methods

2.1 Gym Members Exercise Tracking Data

The dataset comprises $N = 973$ records of gym members, each containing behavioral and body composition measurements such as resting heart rate, calories burned, body fat percentage, and BMI. A self-reported gender attribute was included but intentionally excluded from the feature set during the unsupervised learning

phase to avoid biasing the models. This attribute was only reintroduced for evaluation purposes, enabling assessment of whether discovered patterns aligned with gender differences [1].

Data preprocessing: Rows containing missing values were removed to ensure model stability and consistency across algorithms. Numerical variables were standardized using the *StandardScaler* [2] to equalize feature contributions and prevent variables with larger ranges from dominating distance-based algorithms. Categorical variables with fewer than 10 unique categories were one-hot encoded using the *pandas* library [3], enabling algorithms to handle them without introducing artificial ordinality. This preprocessing pipeline ensured that all features were on a comparable scale and suitable for algorithms sensitive to both scale and data representation.

Normality testing: Many statistical procedures assume normally distributed data. To ensure the validity of hypothesis testing, the distributions of continuous variables were evaluated using both the *Shapiro-Wilk* [4] and *Kolmogorov-Smirnov* [5] tests. These tests guided the selection of appropriate inferential methods: parametric tests such as *ANOVA* [6] when normality was satisfied, and non-parametric tests such as *Kruskal-Wallis* [7] when it was not. This step prevented the application of statistical procedures under violated assumptions, reducing the risk of *Type I and Type II errors* [8].

2.2 Dimensionality Reduction

Dimensionality reduction [9] was applied prior to *clustering* [10] to uncover latent structure in the data, improve cluster separation, mitigate the curse of dimensionality, and enable meaningful 2D visualizations. Two complementary techniques were employed:

- **Principal Component Analysis (PCA)**: A linear projection method that identifies orthogonal components capturing maximal variance, useful for noise reduction and detecting global patterns [11].
- **Uniform Manifold Approximation and Projection (UMAP)**: A non-linear embedding method preserving both local neighborhoods and global relationships, often better suited for complex, non-linear manifolds [12].

Using multiple methods ensured robustness, as each has different strengths in capturing structure.

2.3 Supervised Gender Prediction

In the supervised phase, the goal was to predict gender from the behavioral and physiological features. Six diverse classification algorithms were used to cover a range of model complexities and inductive biases:

- **Logistic Regression**: A linear model providing interpretable coefficients and a probabilistic output [13].
- **Support Vector Machine (SVM)**: A maximum-margin classifier effective in high-dimensional spaces [14].
- **k-Nearest Neighbors (KNN)**: An instance-based method relying on majority voting among the k closest samples [15].
- **Naïve Bayes**: A probabilistic classifier based on Bayes' theorem with independence assumptions [16].
- **Multi-Layer Perceptron (MLP)**: A feedforward neural network capable of modeling complex non-linear boundaries [17].
- **Random Forest**: An ensemble of decision trees leveraging randomness in feature selection and sampling to improve generalization [18].

Hyperparameters were optimized via grid search, and performance was assessed with accuracy, *Precision, recall* [19], F1-score, and area under the *ROC curve* (AUC) [20]. Precision and recall were included to address potential class imbalance. Confusion matrices were examined to identify misclassification patterns.

2.4 Clustering Analysis

In the unsupervised phase, clustering was performed without gender labels to examine whether group structures emerged naturally. Four algorithms were applied to capture different notions of "clusters" [10]:
K-Means: Partitions data into k clusters by minimizing within-cluster variance, assuming spherical clusters [21].
Gaussian Mixture Models (GMM): Fits data as a probabilistic mixture of Gaussians, enabling soft cluster membership [22].
Agglomerative Hierarchical Clustering: Iteratively merges the closest clusters, producing a dendrogram for hierarchical structure analysis [23].
The optimal number of clusters was determined using complementary internal validation indices: the Elbow method [25], Silhouette Score [24], *One-Way ANOVA* [6] and Kruskal-Wallis Test [7].

2.5 Statistical Testing of Cluster–Gender Association

To assess whether identified clusters corresponded to gender differences, statistical tests were conducted:
One-Way ANOVA: Parametric comparison of means across clusters under normality [6].
Kruskal-Wallis Test: Non-parametric alternative for median comparison when normality was violated [7].
Mann–Whitney U Test: Pairwise post-hoc comparisons after significant Kruskal–Wallis results, to localize the source of differences [26].
All tests were conducted at $\alpha = 0.05$, with p-values reported. This ensured rigorous control over Type I error rates.

3 Results

3.1 Normality Test and Outliers

To verify statistical assumptions, we first examined the normality of each numeric feature separately for male and female participants using the Shapiro-Wilk test, and tested equality of variances across genders using Levene's test. For the Shapiro-Wilk procedure, safeguards were implemented to handle small sample sizes, constant values, or very large groups by returning `NaN` in such cases. Similarly, Levene's test was applied only when both groups contained at least two valid observations, using the median as the center for robustness.
In parallel, outlier detection and removal were performed using the Interquartile Range (IQR) method. For each feature, the 25th percentile (Q1), 75th percentile (Q3), and IQR (IQR $= Q3 - Q1$) were calculated. Lower and upper bounds were defined as $Q1 - 1.5 \times \text{IQR}$ and $Q3 + 1.5 \times \text{IQR}$, respectively. Data points falling outside these bounds were considered outliers and removed. When `group_by_label` was enabled, the IQR bounds were computed separately within each gender group to avoid bias caused by distributional differences. The cleaning process produced a reduced dataset with all outlier indices recorded, as well as a detailed summary table reporting the number and percentage of removed points per feature and per group.

3.2 Gender Prediction Using Fair Features

To assess whether gender could be predicted from neutral, non-biological features, we trained two Random Forest classifiers. The first model relied solely on behavioral attributes, including workout frequency, session duration, experience level, and workout type, while the second combined these with physiological metrics such as Resting BPM, Avg BPM, and Max BPM.
Both models performed poorly, with accuracy scores of 46.2% and 50.7%, and ROC-AUC values of 0.457 (95% CI: 0.396–0.524) and 0.517 (95% CI: 0.457–0.583), respectively, indicating near-random classification. The confusion matrices revealed patterns of symmetric misclassifications, and no single characteristic exhibited strong discriminative power (Figures 1A, 1B, 1C).
These findings suggest that neither behavioral nor physiological patterns sufficiently differentiate male from female participants in this dataset. This result supports the fairness of using such neutral features, implies that participant privacy is preserved, and promotes a shift toward personalized fitness guidance based on individual characteristics rather than demographic labels.
To further investigate feature contributions to gender classification, we examined the importance scores from

the combined model. As shown in Figure 2, `Session_Duration` was by far the most influential variable, followed by physiological indicators like `Max_BPM`, `Avg_BPM`, and `Resting_BPM`. Behavioral features such as workout frequency and type contributed only marginally. Despite the limited predictive success, the consistent dominance of session duration may point to subtle behavioral differences. Nonetheless, the overall low importance values support the conclusion that gender cannot be reliably inferred from these features, aligning with the fairness assumptions underlying this analysis.

Overall, the results provide strong evidence that the gym environment, as reflected in this dataset, does not exhibit gendered behavioral or physiological patterns, emphasizing a context of gender equality in workout habits and physiological state.

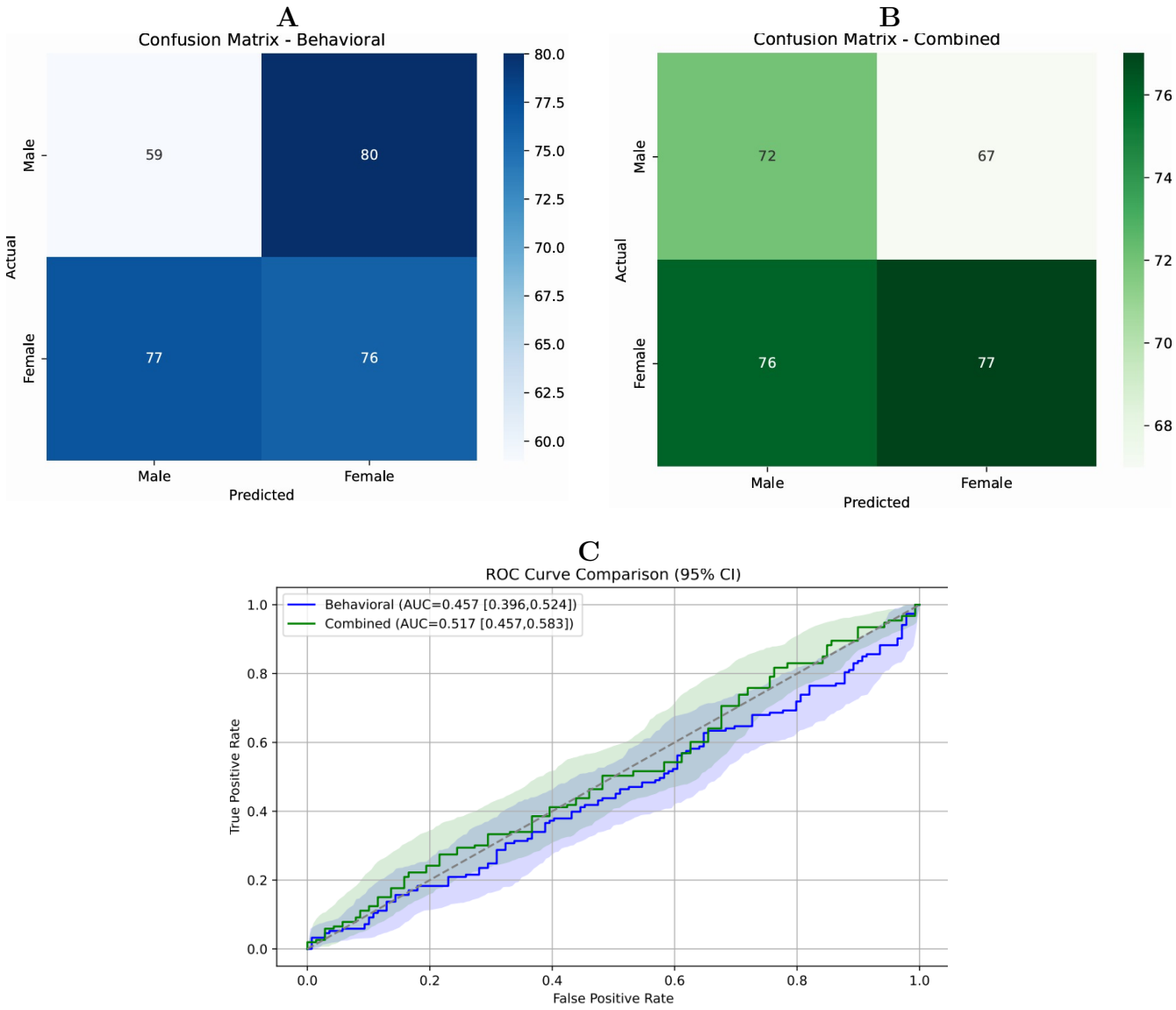


Figure 1: Gender prediction evaluation. (A) Confusion matrices for behavioral and combined models. (B) ROC curves comparing both models. (C) ROC curve comparison (zoomed view).

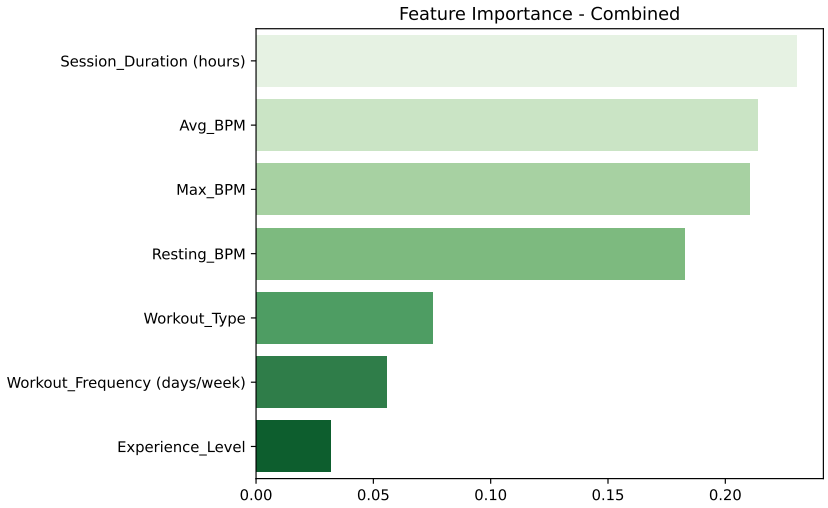


Figure 2: Feature importance in the combined model.

3.3 Best Clustering Algorithm

We selected the One-way ANOVA and the Kruskal-Wallis tests to assess the relationship between the external variable `gender` and the clustering results for the optimal dimension-cluster combinations of GMM, K-Means, and hierarchical clustering. Although One-way ANOVA is a parametric test that assumes normally distributed data, it is commonly used as a baseline to compare mean differences between groups and is robust to mild deviations from normality, especially with large sample sizes. Since our normality checks indicated that the data were not normally distributed, we also applied the Kruskal-Wallis test, a non-parametric alternative that does not require the normality assumption and instead compares the rank distributions between groups. This combination ensures both comparability with standard methods and validity under our data's distributional properties.
According to the silhouette measure, K-Means is the best algorithm (see GitHub for the added code). For the statistical tests, the ANOVA p-values were ($p = 7.86 \times 10^{-83}$) for K-Means, ($p = 5.37 \times 10^{-69}$) for GMM, and ($p = 1.30 \times 10^{-61}$) for Hierarchical. The Kruskal-Wallis p-values were ($p = 2.62 \times 10^{-68}$) for K-Means, ($p = 9.28 \times 10^{-59}$) for GMM, and ($p = 2.18 \times 10^{-53}$) for Hierarchical (See Table 1).

	ANOVA p-value	Kruskal p-value
KMeans	7.86e-83	2.62e-68
GMM	5.37e-69	9.28e-59
Hierarchical	1.30e-61	2.18e-53

Table 1: ANOVA and Kruskal-Wallis p-values for each chosen clustering algorithm with reference to the external variable (Gender).

As explained, based on the highest silhouette score among all algorithms, as well as the results of the two statistical tests, the optimal clustering algorithm was determined to be K-Means.

3.4 Heatmap of Feature Correlations

To further explore the relationships between physiological and behavioral variables, we generated a correlation heatmap (Figure 3). The heatmap provides a visual overview of pairwise correlations, highlighting both positive and negative associations between features. Strong positive correlations were observed among body-size indicators such as `Weight (kg)`, `Height (m)`, and `BMI`, as expected. Training-related variables, including `Experience_Level`, `Workout_Frequency`, and `Session_Duration`, also exhibited moderate to strong positive correlations with each other. Conversely, `Fat_Percentage` showed a negative correlation with most training-related variables and body-size indicators, suggesting an inverse relationship between adiposity and both training exposure and certain anthropometric measures. This visualization helps contextualize the patterns identified in clustering analysis by illustrating how groups of features tend to vary together.

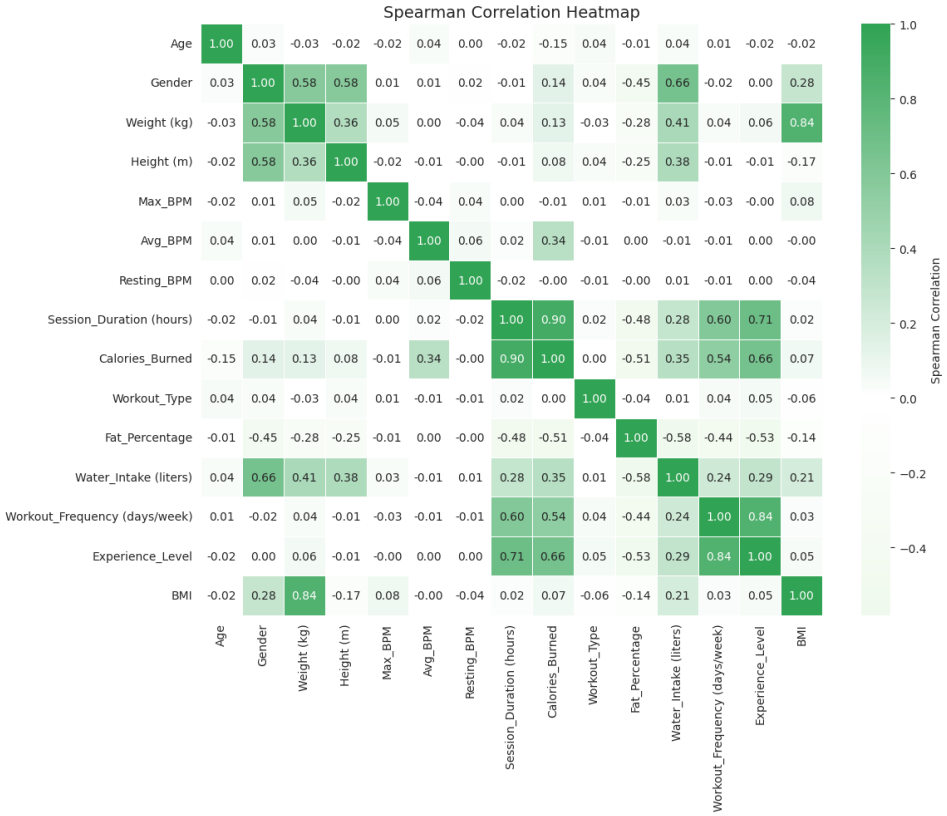


Figure 3: Correlation heatmap showing pairwise relationships between physiological and behavioral features. Darker colors indicate stronger positive or negative correlations.

3.5 Visualizations

To illustrate the clustering structure, we applied K-Means with $k = 3$ clusters using two different embedding spaces.

UMAP Visualization: First, the scaled feature space was reduced to two dimensions using PCA to remove noise and preserve the main directions of variance. K-Means clustering was then performed in this PCA space, and the resulting clusters were projected using the UMAP algorithm. UMAP is a non-linear dimensionality reduction method that preserves local neighborhood relationships, making it well-suited for visualizing high-dimensional clustering results (Figures 4A).

PCA Visualization: For comparison, we also visualized the clusters directly in the two-dimensional PCA space used to fit the K-Means model. PCA is a linear method that projects the data onto orthogonal components maximizing global variance. This representation provides a direct view of how the clusters align with the principal axes of variation in the dataset and helps identify which features contribute most to the separation between clusters. (Figures 4B)

The silhouette score was computed for each visualization to quantify the cohesion and separation of the clusters. In both visualizations, clusters are color-coded, and the axes represent the reduced feature dimensions (UMAP1/UMAP2 or PC1/PC2).

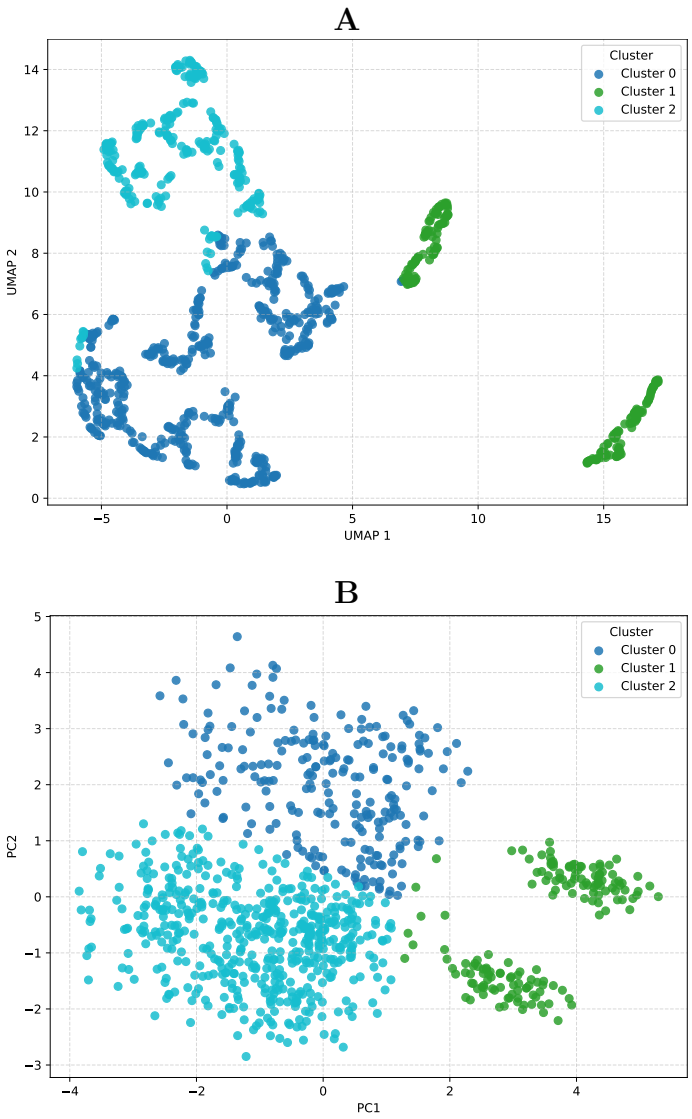


Figure 4: Clustering visualization using K-Means with $k = 3$ in two embedding spaces. (A) UMAP projection of PCA-reduced features, preserving local neighborhood structure. (B) PCA projection highlighting global variance. Clusters are color-coded consistently across both views.

3.6 Cluster Profiles: Gender Distribution and Key Features

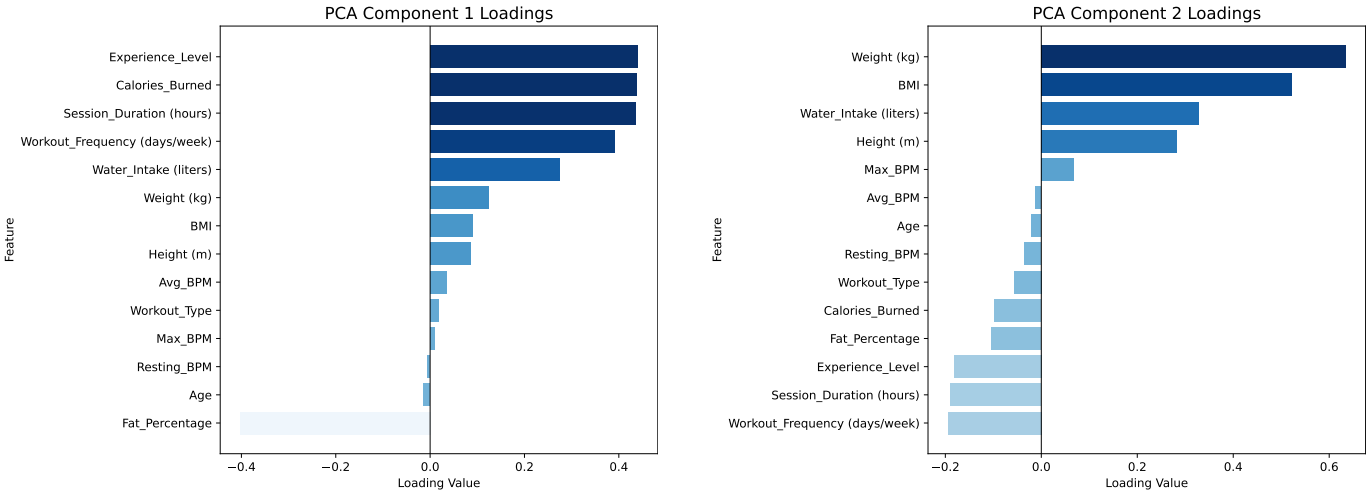
Using the final clustering solution (PCA embedding followed by hierarchical clustering, $k = 3$), we examined gender composition in each cluster and interpreted the defining attributes via the PCA loading plots (Figures 5A, 5B).

Cluster 0 was overwhelmingly female: 99.6% female (235 participants) and a single male. Cluster 1 was mixed, with 55.3% female (88 participants) and 44.7% male (71 participants). Cluster 2 was predominantly male: 69.2% male (371 participants) and 30.8% female (235 participants).

Feature patterns aligned with these compositions. Cluster 0 showed the highest `Fat_Percentage` relative to the other clusters, whereas Cluster 1 was characterized by higher training exposure and behaviorExperience_Level, `Session_Duration` (hours), and `Workout_Frequency` (days/week) were all elevated. Cluster 2 was defined by larger body-size indicators: `BMI`, `Weight (kg)`, and `Height (m)`, consistent with the PCA loadings, where the first component loaded positively on activity/volume measures (e.g., `Experience_Level`, `Calories_Burned`, `Session_Duration`, `Workout_Frequency`) and the second component on anthropometrics (e.g., `Weight`, `BMI`, `Height`).

Thus, while clusters were formed without using gender labels, the resulting groups display different gender proportions that appear linked to distinct behavioral and anthropometric profiles.

In conclusion, unlike the supervised algorithms that failed to predict gender, the unsupervised method revealed prominent features that can characterize males and females based on training data.



A: PCA Component 1 Loadings

B: PCA Component 2 Loadings

Figure 5: PCA loading plots for the first two principal components. PC1 is primarily associated with training activity and engagement measures (e.g., `Experience_Level`, `Calories_Burned`, `Session_Duration`), while PC2 is dominated by anthropometric indicators (e.g., `Weight`, `BMI`, `Height`).

4 Discussion

This study combined supervised and unsupervised machine learning approaches to examine physiological and behavioral data, with the goal of uncovering latent population structures and assessing the feasibility of gender prediction. The central finding is a clear divergence between the two methodological paradigms: while supervised models failed to accurately predict gender, indicating gender parity in the measured variables unsupervised clustering revealed a natural separation aligned with physiological and behavioral patterns. This contrast suggests that gender, although not directly predictable from the available features, remains indirectly embedded in the latent structure of the population.

Three distinct clusters were identified. Cluster 0 was almost entirely female, characterized by the highest `Fat_Percentage`. Cluster 1 had a balanced gender distribution and was marked by greater training exposure, including higher `Experience_Level`, longer `Session_Duration`, and more frequent `Workout_Frequency`. Cluster 2 was predominantly male, displaying larger body-size indicators such as `BMI`, `Weight (kg)`, and `Height (m)`. These profiles indicate that physiological and behavioral patterns partly aligned with gender can emerge spontaneously in latent space, even without explicit gender data.

The observed divergence between supervised and unsupervised outcomes aligns with the broader understanding that complex behavioral and physiological traits may correlate with demographic variables in subtle ways. By capturing these patterns in latent structures, clustering methods can reveal demographic signals that remain hidden to direct predictive models. This finding underscores the complementary value of unsupervised approaches in demographic and behavioral analysis.

From an applied perspective, these insights could inform targeted interventions in sports science, health management, and personalized fitness technology. Potential uses include tailoring training loads for profiles with high fat percentage, designing nutrition plans according to body composition patterns, and identifying early signs of health risks through combined physical and behavioral metrics. Integration of such strategies into fitness platforms could enable adaptive, cluster-specific recommendations.

Future research should investigate the influence of environmental and social factors such as seasonal variation, cultural context, and access to training facilities on cluster membership; incorporate real-time wearable sensor data to track shifts in cluster affiliation; explore the role of psychological and motivational factors in shaping latent groupings; and evaluate the robustness of these patterns in populations differing by culture, age, and physical activity level. Additionally, longitudinal studies could assess whether cluster-based interventions yield measurable improvements in health and performance over time.

Ultimately, this work demonstrates that combining supervised and unsupervised approaches provides a richer understanding of population structure, uncovering subtle demographic and behavioral differences that may be invisible to direct prediction but evident in latent representations. Such insights bridge the gap between statistical modeling and actionable, personalized health strategies.

References

[1] *Gym Members Exercise Tracking.* Kaggle. https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset

[2] *StandardScaler.* https://www.i-mri.org/Synapse/Data/PDFData/1040IMRI/imri-28-61.pdf

[3] *pandas.* https://d1wqtxts1xzle7.cloudfront.net/117768488/pyhpc2011_submission_9-libre.pdf

[4] *Shapiro-Wilk.* https://www.researchgate.net/publication/272353940_Wafer_Level_Molding_of_3D_TSV_Stack_Module

[5] *Kolmogorov Smirnov.* https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat06558

[6] *ANOVA.* https://link.springer.com/chapter/10.1007/978-1-4614-3725-3_8

[7] *Kruskal-Wallis.* https://link.springer.com/chapter/10.1007/978-3-319-30634-6_6

[8] *Type I and Type II errors.* https://archive.ymsc.tsinghua.edu.cn/pacm_download/21/194-2015A_Hypothesis_Testing_Method_Based_on_Sample_and_Two_Types_of_Errors.pdf

[9] *Dimensionality Reduction.* https://www.scirp.org/journal/paperinformation?paperid=111638

[10] *Clustering.* https://ieeexplore.ieee.org/abstract/document/6558109?casa_token=ArqO7D6lPsgAAAAA:1BSAiB2uNvHz7w5h9Z8JJqhtRpVrUmDQBqkyukBAaO6La-aboFqtVrBL4Ybml1B_aDbX0fhu27I

[11] *PCA.* https://doi.org/10.1098/rsta.2015.0202

[12] *UMAP.* https://arxiv.org/pdf/1802.03426.pdf

[13] *Logistic Regression.* https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.106.682658

[14] *SVM.* https://www.sciencedirect.com/science/article/pii/S0360835205000100?casa_token=MJcFcKI5UK0AAAAA:EzlP5x3AjxhENZzEW5R1zzfD2Mx4kaIAsnsYwZi1vfGbR698545zItF7_shmK1jH1etS32CPlOE

[15] *KNN.* https://www.mdpi.com/2073-8994/12/7/1167

[16] *Naïve Bayes.* http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf

[17] *MLP.* https://www.academia.edu/97527970/Constru%C3%A7%C3%B5es_concessivas_intensivas

[18] *Random Forest.* https://www.academia.edu/97527970/Constru%C3%A7%C3%B5es_concessivas_intensivas

[19] *Precision and recall.* https://hal.science/hal-01532152/document

[20] *ROC curve.* https://people.inf.elte.hu/kiss/11dwhdm/roc.pdf

[21] *K-Means.* https://inria.hal.science/inria-00321515/file/verbeek01tr.pdf

[22] *GMM.* https://revistas.ucc.edu.co/index.php/in/article/download/3074/2813

[23] *Agglomerative Hierarchical Clustering.* https://arxiv.org/pdf/1109.2378

[24] *Silhouette Score.* https://ieeexplore.ieee.org/abstract/document/9260048?casa_token=QMj2RGy52S0AAAAA:XpjuJeIbIaDgSMBdJ0_qu6AeDUO0mYod-mqQA1aqoXGGPx4mPIkDhtxmOS7k3ATUPzDj0Sd-zY0

[25] *Loss Function.* https://link.springer.com/article/10.1007/s40745-020-00253-5

[26] *Mann–Whitney U Test.* https://www.tqmp.org/RegularArticles/vol04-1/p013/p013.pdf

[27] *Unsupervised Learning.* https://www.ajodo.org/article/S0889-5406(23)00193-2/fulltext

[28] *Anomaly Detection.* https://ieeexplore.ieee.org/abstract/document/4138201casa_token=LGAxetiUIqYAAAAA:5AETl070saOgeA8LLafh5-U0erAJaE6yn-k7wAy-9XdNM8s7DCOBInPYeJltiNix0Q0Pq8ck6Y