Gym Analytics and Predictive Modeling Report

973	4	3	15
Total Members	Workout Types	Experience Tiers	Features Analyzed

97.2%	36 kcal	91.8%	4
Regression R-Squared	Mean Abs. Error	Classification Accuracy	Member Segments

Prepared by Shivam
Business Analyst and Data Scientist

1. Executive Summary

This report presents a full-stack analytical study of 973 gym members drawn from a structured fitness tracking dataset. The analysis spans exploratory data profiling, behavioral pattern discovery, and three independent machine learning models: a regression model for calorie burn prediction, a multi-class classifier for experience level identification, and an unsupervised clustering model for member segmentation.

The findings reveal strong correlations between session duration, workout frequency, and caloric output. HIIT and Strength training consistently produce the highest caloric expenditure. Body fat percentage declines meaningfully as workout frequency increases. Predictive models achieved exceptional performance, with the calorie regression model reaching an R-squared of 97.2% and the experience classifier achieving 91.8% accuracy. Member clustering revealed four distinct behavioral personas that can guide targeted retention and engagement strategies.

Key Recommendations

- Target high-frequency HIIT and Strength members with premium membership tiers and performance tracking features.
- Introduce hydration and nutrition programs for members showing high fat percentage relative to their workout frequency.
- Use the calorie prediction model to power a real-time progress dashboard for members.
- Apply the segmentation model to personalize onboarding journeys and class recommendations.
- Design retention campaigns specifically for Casual Members (Cluster 1), who show the lowest engagement metrics.

2. Dataset Overview

The dataset contains 973 anonymized gym member records with 15 variables covering demographics, biometrics, workout behavior, and physiological performance indicators. There are no missing values, making the dataset suitable for direct modeling without imputation. The target variables for modeling are Calories Burned (continuous), Experience Level (ordinal categorical), and member segment (latent cluster).

Feature	Type	Description
Age	Numeric	Member age in years
Gender	Categorical	Male or Female
Weight (kg) / Height (m)	Numeric	Physical dimensions
Max / Avg / Resting BPM	Numeric	Heart rate measurements
Session Duration (hrs)	Numeric	Length of each workout session
Calories Burned	Numeric (Target)	Total calories per session
Workout Type	Categorical	Cardio, HIIT, Strength, Yoga
Fat Percentage	Numeric	Body fat as percentage of total weight
Water Intake (liters)	Numeric	Daily hydration volume
Workout Frequency	Numeric	Sessions per week
Experience Level	Ordinal (Target)	Beginner (1) to Expert (3)
BMI	Numeric	Derived: Weight / Height squared

3. Exploratory Data Analysis

3.1 Caloric Output by Workout Type

HIIT and Cardio sessions generate the highest median calorie burn, reflecting their aerobic intensity. Yoga, while lower in caloric output, shows a tighter distribution suggesting more consistent performance across members. Strength training shows the widest spread, indicating high variability driven by weight, experience, and session structure.

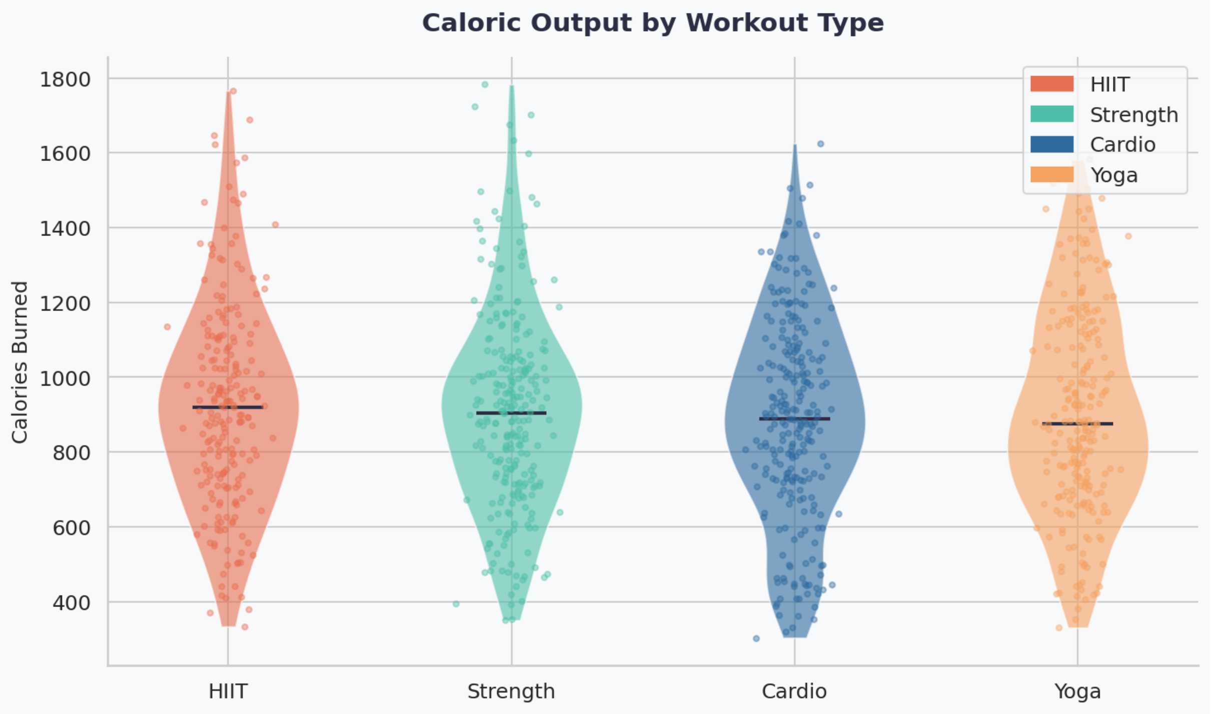

Figure 1: Caloric output distribution across the four workout types, with individual data points overlaid.

3.2 Session Duration and Caloric Output by Experience

A strong positive linear relationship exists between session duration and calories burned across all experience levels. Expert members not only train longer on average but burn disproportionately more calories per hour, likely due to higher intensity and optimized form. Beginners cluster at shorter durations and lower calorie ranges.

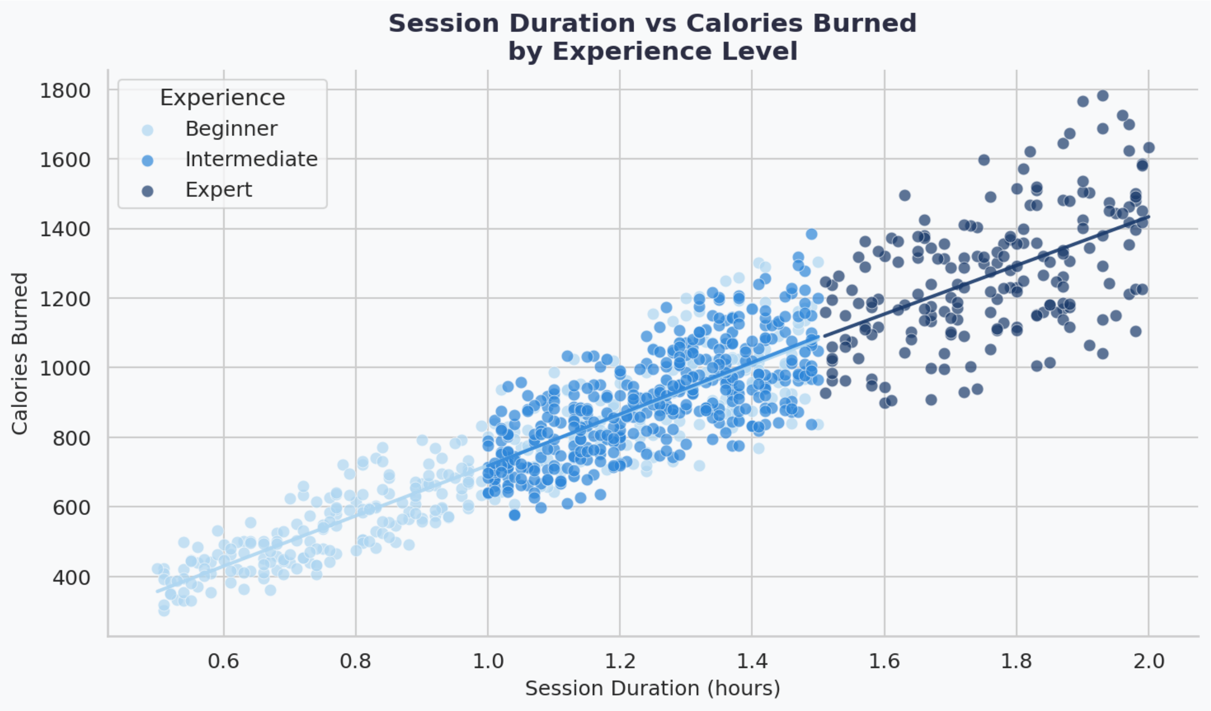

Figure 2: Scatter plot with linear regression lines per experience tier.

3.3 Feature Correlations

The correlation matrix reveals several actionable relationships. Session Duration and Calories Burned share the strongest positive correlation, followed by Workout Frequency. Fat Percentage correlates negatively with Experience Level, confirming that more experienced members have optimized their body composition. BMI and Weight are highly collinear, a natural consequence of the BMI formula.

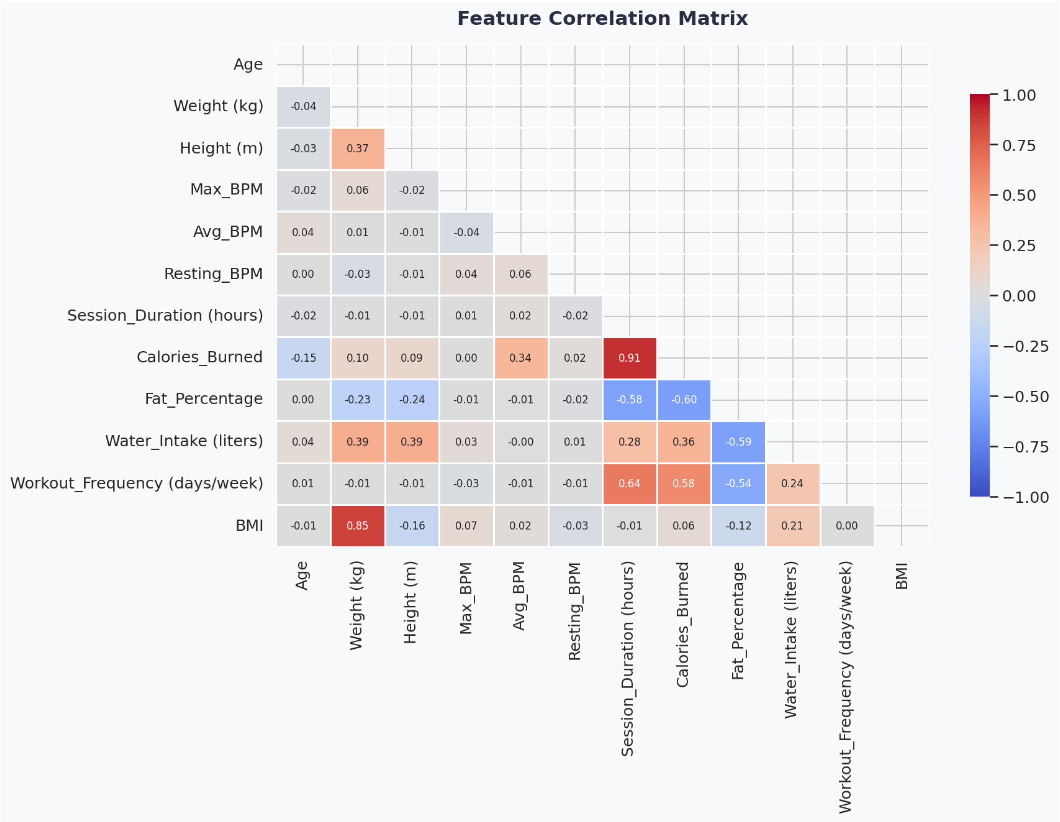

Figure 3: Lower-triangle correlation heatmap across all numeric features.

3.4 BMI Distribution by Gender and Experience

Both male and female members show declining BMI spread as experience level increases, suggesting that consistent training normalizes body composition over time. Beginner members of both genders exhibit the widest BMI ranges, which is expected given the heterogeneous population entering gyms.

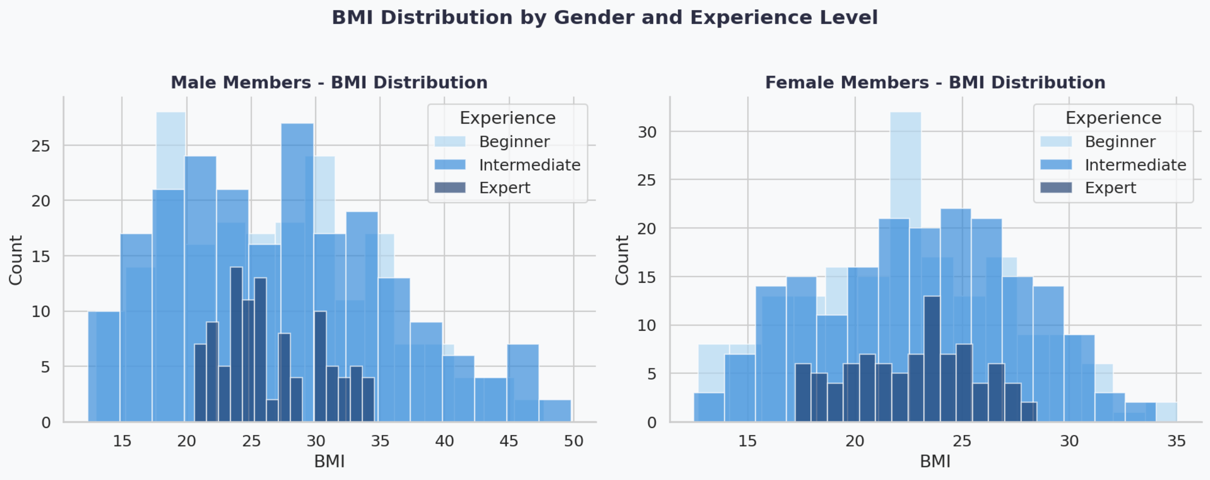

Figure 4: BMI histogram by gender panel, segmented by experience level.

3.5 Heart Rate Profile by Workout Type

HIIT generates the highest average and maximum heart rates, consistent with its high-intensity interval structure. Yoga shows the lowest BPM across all three measures. Resting BPM remains relatively stable across workout types, which is expected as it reflects baseline cardiovascular fitness rather than session intensity.

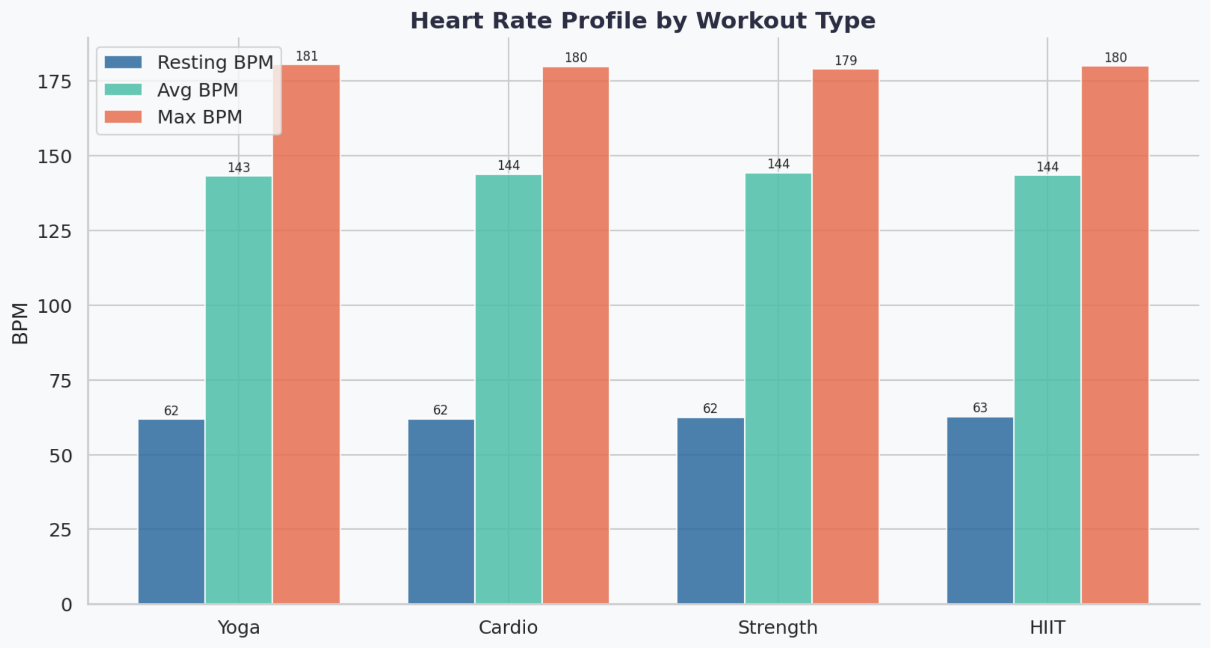

Figure 5: Grouped bar chart of resting, average, and maximum BPM by workout type.

3.6 Hydration vs Body Fat

Members with higher daily water intake tend to show lower body fat percentages, particularly those burning more calories per session (larger bubbles). This suggests that hydration is a supporting factor in body composition outcomes, though causality cannot be established from this data alone.

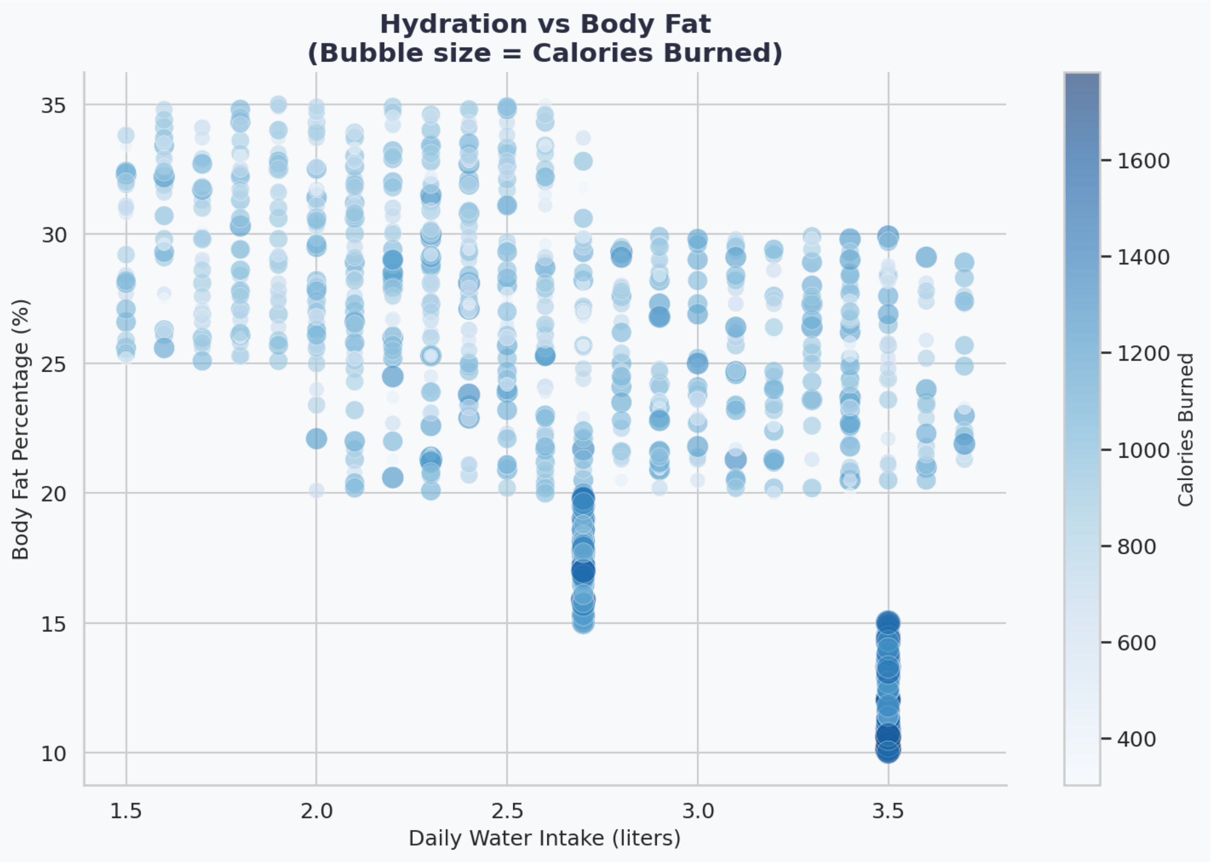

Figure 6: Bubble chart where bubble size encodes calories burned per session.

3.7 Workout Frequency vs Body Fat Percentage

Members who train 5 or more days per week consistently show lower median body fat percentages. The interquartile range narrows at higher frequencies, indicating that frequent training produces more predictable body composition outcomes. Members training only 2 days per week show the highest median fat percentage.

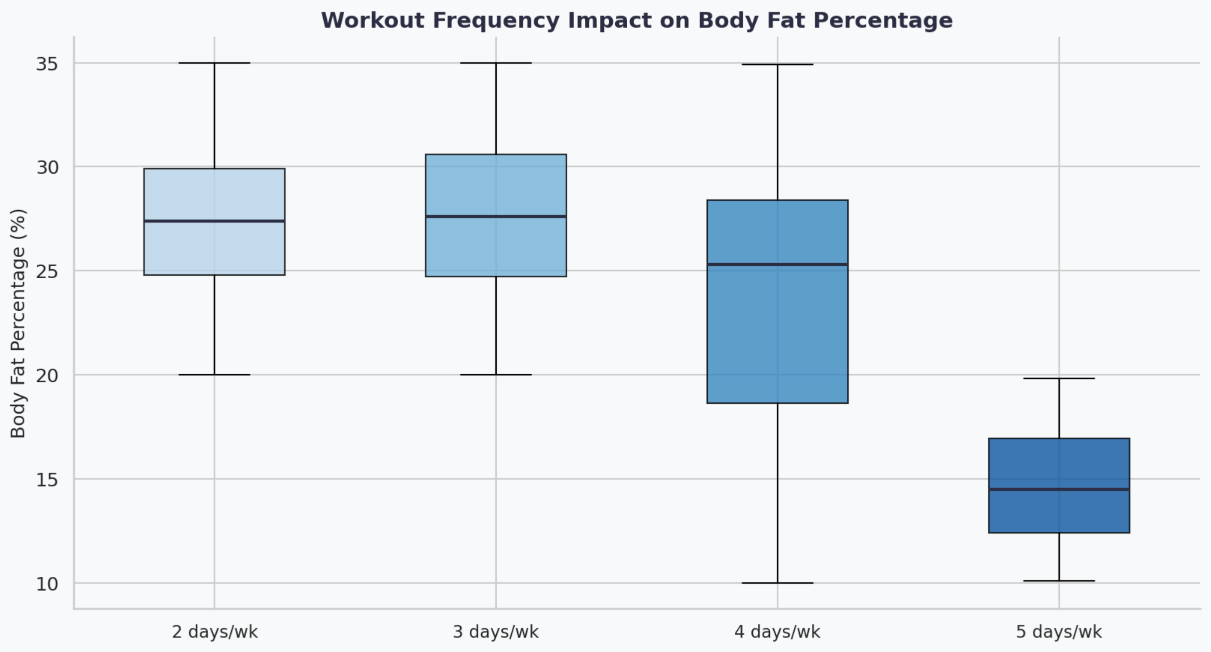

Figure 7: Box plot of body fat percentage across workout frequency categories.

4. Predictive Modeling

4.1 Calorie Burn Regression (Random Forest)

A Random Forest Regressor was trained on 80% of the data using 200 estimators to predict calories burned per session. The model achieved an R-squared of 0.9720 on the holdout test set, with a Mean Absolute Error of just 36.2 kcal. Session Duration, Average BPM, and Workout Frequency emerged as the top three predictors, together accounting for the majority of explained variance.

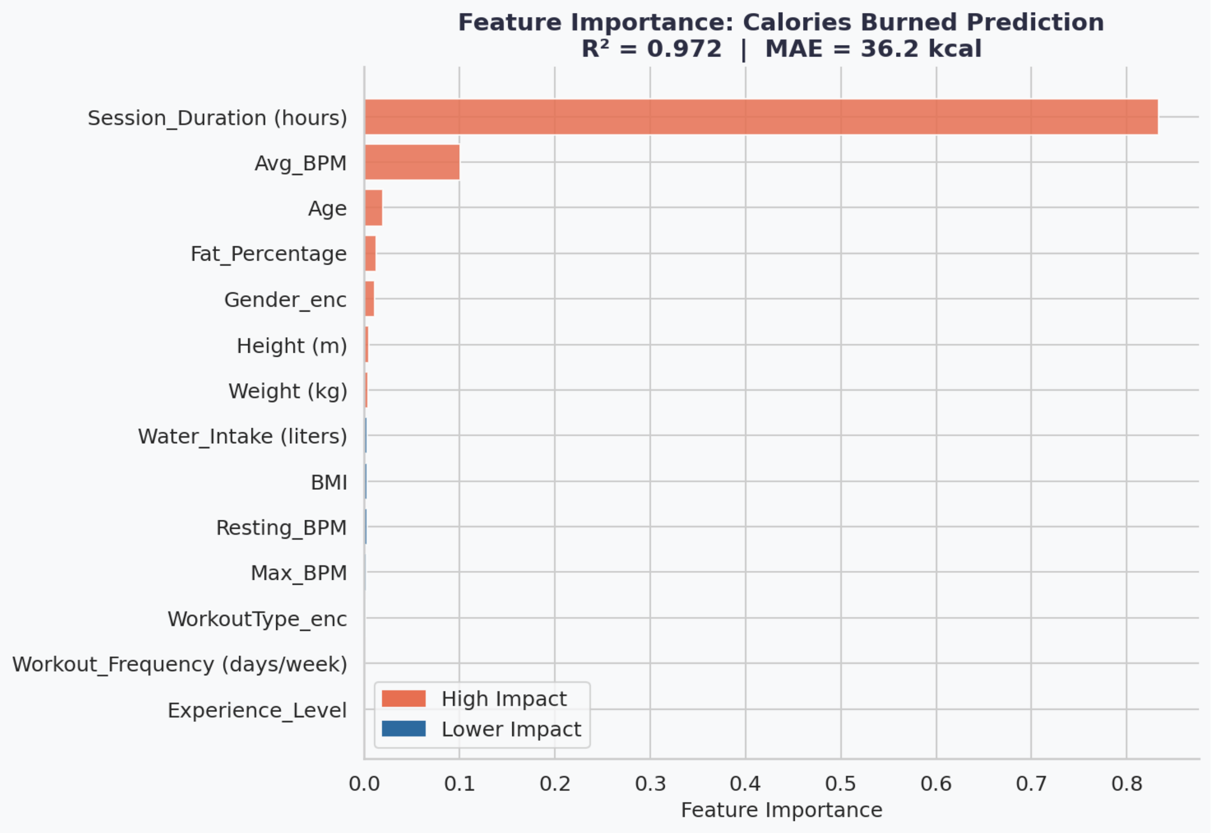

Figure 8: Feature importance ranking for calories burned prediction. R-squared = 0.972.

4.2 Experience Level Classification (Random Forest)

A multi-class Random Forest Classifier was trained to predict whether a member is a Beginner, Intermediate, or Expert based on their behavioral and physiological metrics. The model achieved 91.8% accuracy on stratified test data. This classifier can be embedded into onboarding workflows to immediately profile new members and recommend appropriate programs without requiring a manual assessment.

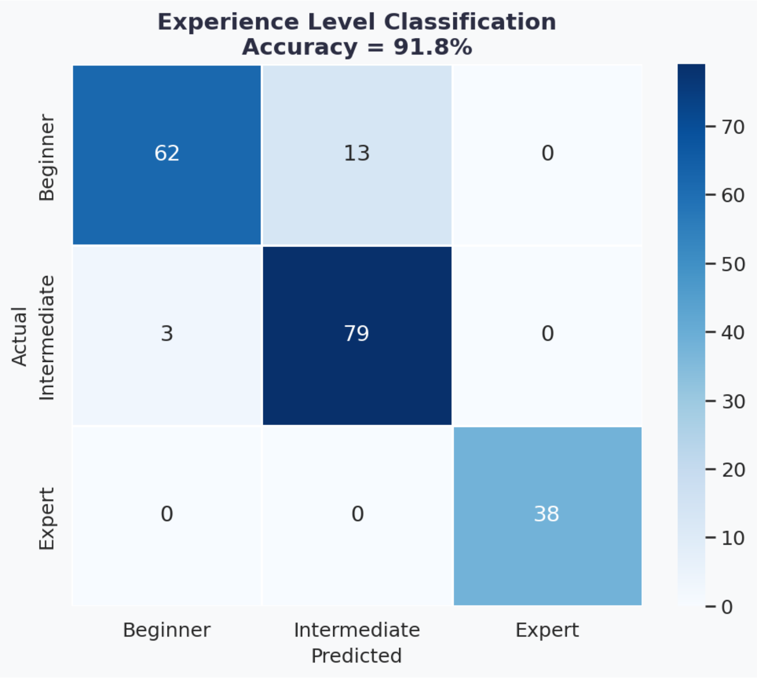

Figure 9: Confusion matrix for experience level classification. Accuracy = 91.8%.

4.3 Member Segmentation (K-Means Clustering)

K-Means clustering with k=4 was applied to six normalized behavioral and physical features: Age, BMI, Fat Percentage, Session Duration, Calories Burned, and Workout Frequency. Principal Component Analysis was used to project clusters into two dimensions for visualization. The first two principal components explain 66.2% of total variance.

Four distinct member personas emerged:

- **High Performers:** High calorie burn, long sessions, low fat percentage. These are the gym champions.
- **Casual Members:** Low frequency, shorter sessions, higher fat percentage. At-risk for churn.
- **Intensive Trainers:** Very high BPM, short to medium sessions, focused on intensity over volume.
- **Balanced Athletes:** Moderate across all metrics, consistent frequency. The stable membership core.

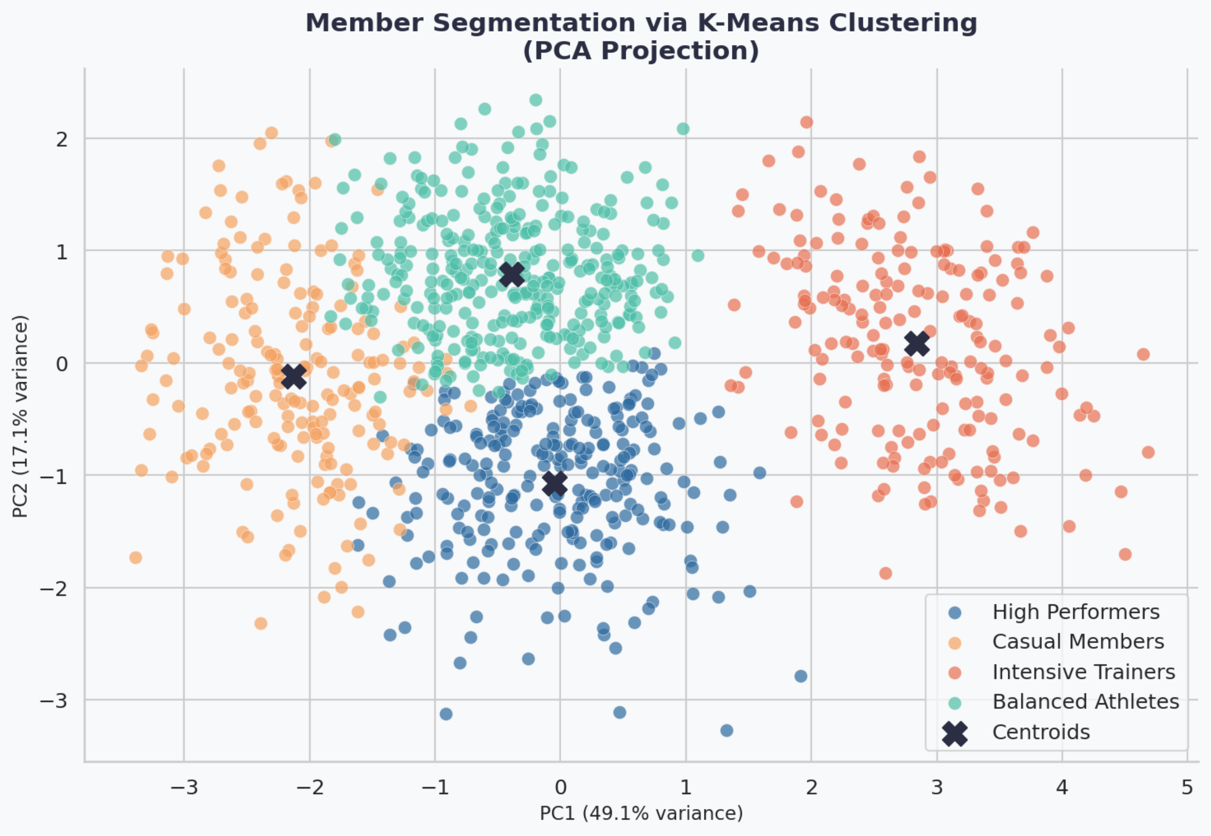

Figure 10: K-Means member segments projected via PCA. Centroids marked with X.

5. Business Recommendations

Retention Strategy

- Casual Members (Cluster 1) are the highest churn risk. Introduce check-in nudges, progress milestone rewards, and weekly goal-setting prompts.
- Pair Casual Members with expert-level mentors or group sessions to increase social commitment.

Revenue Optimization

- High Performers and Intensive Trainers are the most engaged segments. Offer premium tiers with advanced performance analytics, nutrition integration, and personal coaching.
- HIIT and Strength classes show the highest caloric output. Increase class frequency and introduce premium slots during peak demand windows.

Health and Wellness Programs

- Members with high fat percentage and low workout frequency should be enrolled in structured beginner pathways with progressive difficulty.
- Introduce a hydration tracking challenge linked to the water intake data, given its correlation with fat percentage outcomes.

Technology Integration

- Deploy the calorie regression model as a live session estimator within a mobile app, updating in real time based on heart rate and duration.
- Use the experience classifier during sign-up to auto-assign members to the right program track without manual assessment.
- Run the segmentation model quarterly to detect cluster migration, which signals whether members are progressing or at risk.

6. Conclusions

This analysis demonstrates that gym member data, even at relatively modest scale, can yield high-fidelity predictive models and rich behavioral insights. The three-model framework covering regression, classification, and clustering offers both operational value (calorie estimation, member profiling) and strategic value (segmentation, retention targeting).

The most impactful variables across all analyses are Session Duration, Workout Frequency, Average BPM, and Fat Percentage. Gym operators should prioritize collecting and maintaining these fields at the highest accuracy to ensure model performance in production.

The next phase of this work should focus on longitudinal tracking to observe cluster migration over time, integration with wearable device data for real-time BPM capture, and A/B testing of the retention and upsell strategies identified through segmentation.

Prepared and Authored by

Shivam

Business Analyst and Data Scientist

All models trained on anonymized gym member data (n=973). Random Forest models use 200 estimators with an 80/20 train-test split. Clustering uses K-Means with k=4 on standardized features.