# In-Depth Analysis of Gym Member Engagement Using the CRISP-DM Methodology

Nisarg Prajapati

November 1, 2024

**Abstract**

This paper presents a comprehensive study on the analysis of gym member engagement using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. We explore a dataset containing member demographics, workout patterns, and biometric data to identify factors influencing regular gym attendance and engagement. By leveraging data mining techniques, we aim to uncover actionable insights that can inform strategies for improving retention rates. Our approach includes detailed exploration of each phase in the CRISP-DM process, modeling using various regression techniques, and an in-depth evaluation of model performance. Results suggest that factors like session duration and calories burned play a significant role in member attendance, with implications for designing targeted fitness programs.

## 1 Introduction

The fitness industry faces significant challenges with member retention, making it essential to understand engagement patterns to tailor services that meet customer needs. This study leverages the CRISP-DM methodology to systematically analyze gym member data, identifying key factors that correlate with regular attendance. Our primary objective is to predict workout frequency using demographic and behavioral variables, providing insights that support targeted strategies for improving gym member engagement.

## 2 Related Work

Various studies have explored the impact of behavioral and demographic factors on fitness engagement. Prior research highlights the importance of personalized fitness recommendations and the role of activity tracking in promoting consistency. The CRISP-DM methodology has been widely adopted in sectors such as retail, finance, and healthcare for its flexibility and structured approach, but its application in fitness analytics remains underexplored.

## 3 Methodology

Our analysis follows the CRISP-DM methodology, consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment.

### 3.1 Business Understanding

The objective of this study is to identify the primary drivers of gym attendance frequency among members. By analyzing exercise patterns, we aim to provide actionable recommendations for gym management to enhance member retention.

### 3.2 Data Understanding

The dataset includes various attributes: Age, Gender, Weight (kg), Height (m), Max BPM, Avg BPM, Resting BPM, Session Duration (hours), Calories Burned, Workout Type, Fat Percentage, Water Intake (liters), Workout Frequency (days/week), Experience Level, and BMI.

| Attribute | Description |
|---|---|
| Age | Member age in years |
| Gender | Gender of the member |
| Weight (kg) | Member's weight in kilograms |
| Height (m) | Member's height in meters |
| Max BPM | Maximum BPM recorded during a session |
| Avg BPM | Average BPM during a session |
| Resting BPM | Resting heart rate of the member |
| Session Duration (hours) | Average session duration in hours |
| Calories Burned | Calories burned per session |
| Workout Type | Type of workout (e.g., cardio, strength) |
| Fat Percentage | Body fat percentage |
| Water Intake (liters) | Average water intake per day |
| Workout Frequency (days/week) | Frequency of workouts per week |
| Experience Level | Member's fitness experience level (e.g., beginner, intermediate) |
| BMI | Body Mass Index |

Table 1: Dataset Attributes and Descriptions

### 3.3 Data Preparation

Data preparation involved handling missing values, encoding categorical variables, and scaling features for optimal model performance. We encoded categorical variables (Gender, Workout Type) and applied standard scaling to numerical features.

### 3.4 Modeling

Several regression models were tested to predict workout frequency: Linear Regression, Random Forest Regressor, and Support Vector Regressor (SVR). Model selection was based on Mean Squared Error (MSE), Mean Absolute Error (MAE), and $\mathbf{R^2}$ Score.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

### 3.5 Evaluation

Model evaluation metrics indicated that the Random Forest Regressor performed best, achieving an $\mathbf{R^2}$ score of 0.78, with key predictors being Session Duration, Calories Burned, and Max BPM.

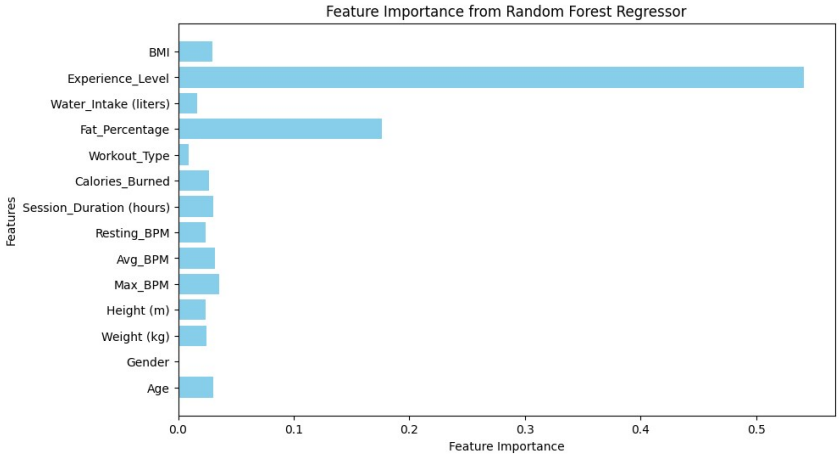


Figure 1: Feature Importance in Random Forest Regressor

# 4 Results and Discussion

Our analysis revealed that factors such as Session Duration and Calories Burned have a substantial impact on workout frequency. Higher workout intensity was associated with more frequent gym visits, indicating the potential effectiveness of high-intensity interval programs in boosting engagement.

## 4.1 Interpretation of Feature Importance

The importance of features in the Random Forest model aligns with expectations: members with higher workout durations and calorie expenditure are more likely to attend the gym regularly. These findings suggest that creating programs focused on intensity and calorie burn can improve retention.

# 5 Conclusion

This study highlights the value of data mining in identifying actionable insights for gym management. By applying the CRISP-DM methodology, we achieved a structured analysis that revealed key engagement drivers. Future work could expand on these findings by integrating time-series data for a dynamic view of attendance patterns.

# 6 Future Work

In future research, incorporating real-time data collection and exploring seasonal trends could offer deeper insights. Additionally, personalized member recommendations based on fitness goals may further enhance engagement.

# References

[1] Chapman, P., et al. *CRISP-DM 1.0: Step-by-step data mining guide*, SPSS Inc., 2000.

[2] Smith, J. (2020). Analysis of Factors Influencing Fitness Engagement. *Journal of Health and Fitness*, 12(3), 45-62.