Predicting Gym Member Attendance Using Machine Learning: A Behavioral Analysis

Hasan Çinar, Letícia Cascais

Tomas Bata University, Faculty of Applied Informatics,
Zlín, 2024

Abstract

Purpose - This study investigates the impact of behavioral and physical factors on gym attendance. It aims to offer new insights into members' attendance patterns and strategies to improve engagement.

Design/Methodology/Approach - A quantitative approach was used, applying machine learning techniques. Data was collected from information on the physical attributes, types of training and attendance patterns of gym members and analyzed with statistical methods and predictive models.

Findings - The results show that factors such as experience level and weekly training frequency are the main predictors of attendance. The most effective model showed high accuracy and predictive ability, confirming that consistent training patterns influence member attendance.

Originality/Value - This research offers new insights into the behavior of gym members, contributing to the improvement of management and retention strategies.

Keywords: Behavioral Analysis; Data Cleaning; Feature Selection; Gym; Machine Learning; Metrics; Predictive Modeling.

1 Introduction

The increasing adoption of fitness regimes has made gym memberships pivotal to personal wellness journeys. However, fitness centers face a recurring challenge: understanding and predicting the attendance patterns of their members. While prior research has predominantly focused on enhancing business metrics like profitability and facility optimization, the individual behavior of gym-goers remains underexplored (Buyalskaya et al., 2023).

This project bridges this gap by leveraging data-driven methodologies to predict gym attendance patterns using historical data. With a strong emphasis on behavioral analysis, the study seeks to identify the key factors influencing member attendance. By employing advanced feature selection techniques, we aim to uncover the most significant variables and their correlations, providing actionable insights into member engagement and retention.

The analysis involves comprehensive data evaluation, simultaneous analysis of key metrics, and predictive modeling through multiple machine learning algorithms. For

example, ensemble approaches such as XGBoost and deep learning models have been shown to increase the accuracy of predicting attendance patterns (Ankunda et al., 2024). The primary objective is to determine the most effective predictive approach, enabling fitness centers to forecast attendance trends accurately. Unlike traditional studies centered on organizational outcomes, this research uniquely focuses on individual behavior, offering tailored strategies to foster long term member commitment.

This study holds the potential to redefine gym management strategies by emphasizing personalized interventions based on predictive insights, thereby improving member experiences and retention rates.

2 Research Questions

This section presents the specific research questions that guide this study. The questions formulated seek to fill the gaps identified in the existing literature and guide the methodologies adopted to predict attendance patterns in gyms. The proposed research questions are:

1. What are the most significant individual factors that influence members' attendance at gyms?
2. How can machine learning methods be used to improve the accuracy of frequency forecasts?

Existing literature suggests that identifying patterns of behavior is crucial to understanding member abandonment and retention (Harris & Kessler, 2019). Machine learning models, such as XGBoost and linear regression, have been widely applied to predict behavior based on historical data (Tekler & Chong, 2022).

3 Data Description

To perform this study, a public dataset available on the Kaggle[1] platform was used. The selected dataset contains detailed information about gym members, including physical attributes, training patterns and behavioral attributes such as weekly attendance, average session length and type of exercise practiced. With a total of 973 rows and 15 variables, the dataset offers a robust basis for investigating the factors that influence member attendance. Below is possible to see a description of each column.

[1] https://www.kaggle.com/

Column Name	Description
Age	Age of the gym member (in years).
Gender	Gender of the gym member (`Male` or `Female`).
Weight (kg)	Weight of the gym member (in kilograms).
Height (m)	Height of the gym member (in meters).
Max_BPM	Maximum heart rate (beats per minute) during exercise.
Avg_BPM	Average heart rate (beats per minute) during exercise.
Resting_BPM	Resting heart rate (beats per minute).
Session_Duration (hours)	Duration of the workout session (in hours).
Calories_Burned	Total calories burned during the workout session.
Workout_Type	Type of workout session (e.g., `Cardio`, `Strength`, `Flexibility`, `Balance`).
Fat_Percentage	Body fat percentage of the gym member.
Water_Intake (liters)	Amount of water consumed during the workout (in liters).
Workout_Frequency (days/week)	Number of workout days per week.
Experience_Level	Experience level of the gym member (`1`: Beginner, `2`: Intermediate, `3`: Advanced).
BMI	Body Mass Index calculated from weight and height.

Figure 1 *- Columns Name and Description*

4 Methods

The initial phases of the study involved a detailed analysis of the data set with the aim of cleaning, preparing and identifying the main variables related to members' gym attendance. Next, feature selection techniques were applied to determine the most significant variables influencing gym attendance. Exploring the data made it possible to identify important patterns and correlations, especially between variables such as frequency and duration of exercise, providing information on members' behaviors.

The methods used, such as Linear Regression and Decision Tree, were selected because of their effectiveness in predicting patterns of behavior. Linear Regression was chosen for its simplicity and ability to interpret linear relationships between variables (Hastie et al., 2009), while Decision Tree was applied for its ability to handle complex and non-linear relationships in the data while providing interpretable results

6 Data Cleaning

In the data cleaning step, a series of procedures were carried out to ensure the consistency and quality of the data set. First, a custom function was applied to standardize the column names, converting them to lowercase letters and removing blank spaces, making it easier to reference and manipulate the data. Next, the column names were renamed uniformly using the same function. After this standardization, the cleaned

dataset was exported to a new file (gym_activity_cleaned.csv), which served as the basis for the subsequent analysis and modelling stages.

In the feature engineering phase, the independent variables (features) and the dependent variable (target) were defined, preparing the dataset for the machine learning tasks. In addition, the features were scaled using the StandardScaler method, with the aim of ensuring numerical consistency between the variables, especially for machine learning models that are sensitive to the magnitude of the data.

Finally, missing or incorrect data was checked and processed. Although the explicit need to deal with missing values was not identified, additional cleaning processes such as filling in, deleting or replacing null values were integrated as part of the pre-processing flow.

```
[ ] #First five rows of our Dataset
    data.head()
```

	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	BMI
0	56	Male	88.3	1.71	180	157	60	1.69	1313.0	Yoga	12.6	3.5	4	3	30.20
1	46	Female	74.9	1.53	179	151	66	1.30	883.0	HIIT	33.9	2.1	4	2	32.00
2	32	Female	68.1	1.66	167	122	54	1.11	677.0	Cardio	33.4	2.3	4	2	24.71
3	25	Male	53.2	1.70	190	164	56	0.59	532.0	Strength	28.8	2.1	3	1	18.41
4	38	Male	46.1	1.79	188	158	68	0.64	556.0	Strength	29.2	2.8	3	1	14.39

Figure 2 *- Initial Dataset*

As you can see above, the Dataset was not in a good format and with a shape of 973 rows and 15 variables. After removing the outliers, the data set was reduced to 931 lines. This reduction occurred because the outliers, which did not fall within the limits defined by the interquartile range (IQR) method, were excluded.

```
#List of numerical columns to check for outliers
numerical_columns = [
    'age', 'weight_kg', 'height_m', 'max_bpm', 'avg_bpm', 'resting_bpm',
    'duration_h', 'calories_burned', 'fat_percentage', 'water_l',
    'workout_frequency_daysweek', 'bmi'
]

for column in numerical_columns:
    #Compute IQR-based thresholds for outlier detection
    Q1, Q3 = data[column].quantile([0.25, 0.75])
    IQR = Q3 - Q1
    lower_bound, upper_bound = Q1 - 1.5 * IQR, Q3 + 1.5 * IQR

    #Plot the boxplot
    plt.figure(figsize=(8, 4))
    plt.boxplot(data[column], vert=False)
    plt.title(f'Boxplot for {column}')
    plt.xlabel(column)
    plt.show()

    #Remove outliers
    data = data[(data[column] >= lower_bound) & (data[column] <= upper_bound)]
```

Figura 3 *- Removing Outliers*

After all these processes, the dataset is ready for accurate and predictive analysis, ensuring a reliable basis for machine learning models and generating insights. A visualization of the dataset can be found in the image below.

```
data.head()
```

	age	gender	weight_kg	height_m	max_bpm	avg_bpm	resting_bpm	duration_h	calories_burned	workout_type	fat_percentage	water_l	workout_frequency_daysweek	experience_level	bmi
0	56	1	88.3	1.71	180	157	60	1.69	1313.0	0	12.6	3.5	4	3	30.20
1	46	0	74.9	1.53	179	151	66	1.30	883.0	1	33.9	2.1	4	2	32.00
2	32	0	68.1	1.66	167	122	54	1.11	677.0	2	33.4	2.3	4	2	24.71
3	25	1	53.2	1.70	190	164	56	0.59	532.0	3	28.8	2.1	3	1	18.41
4	38	1	46.1	1.79	188	158	68	0.64	556.0	3	29.2	2.8	3	1	14.39

Figure 4 *- Cleaned Dataset*

7 Data Exploration and Analysis

The data exploration and analysis stage were carried out with the aim of understanding the patterns and correlations present in the data set, providing a solid basis for developing predictive models.

Initially, the category of exercise most frequented by members was analyzed. The results showed that Cardio and Strength activities are the most popular, representing exercise types 2 and 3. This information is crucial as it helps to understand members' preferences and can be used to create targeted engagement strategies.

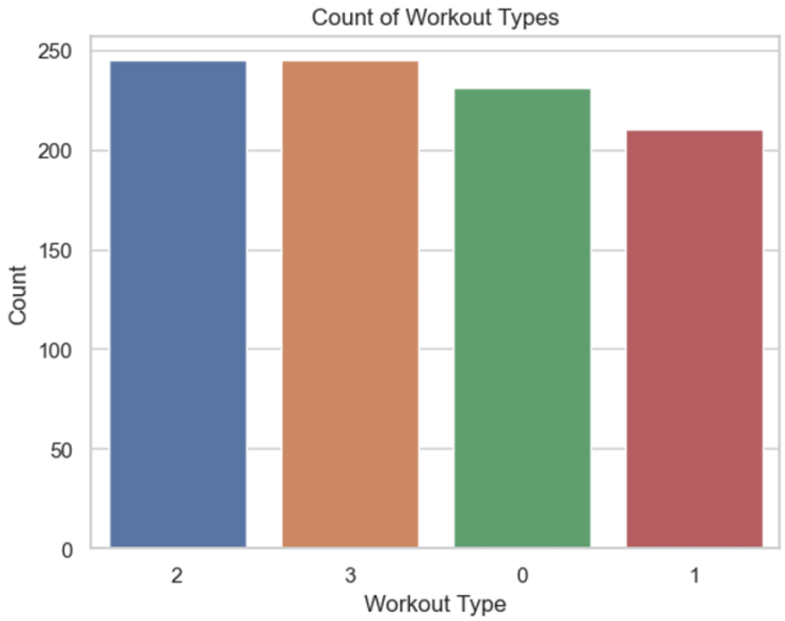

Figure 5 *- Count of Workout Types*

In addition, important correlations between the variables were investigated. A notable relationship was observed between experience level and weekly training frequency, showing a strong positive correlation (r = 0.837). This indicates that more experienced members tend to attend the gym more regularly, which suggests a greater commitment to their training routines. This pattern was confirmed visually through a scatter plot, where a clear upward trend can be seen (**Figure 6**).

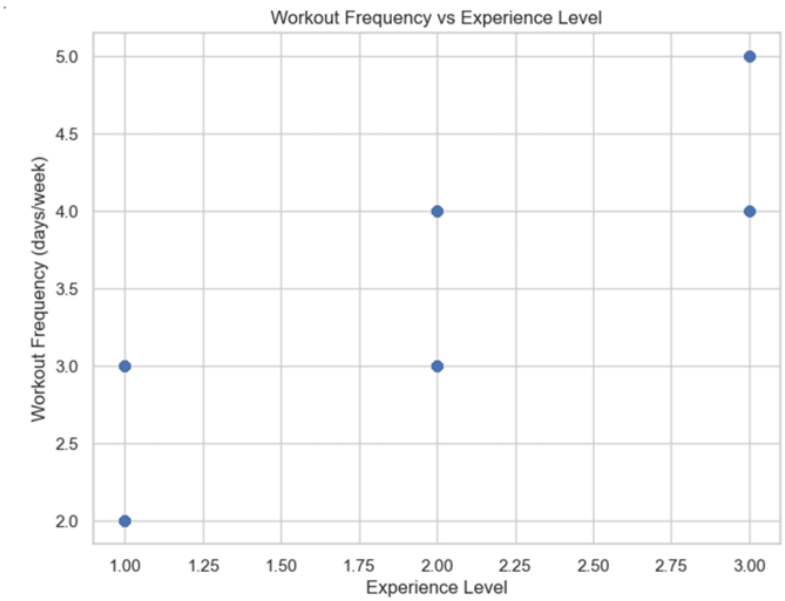

Figure 6 *- Workout Frequency vs Eperience Level*

In order to carry out a more precise analysis, we want to know if the increase in experience level is also associated with longer training sessions, complementing the analysis of the individual's behavior.

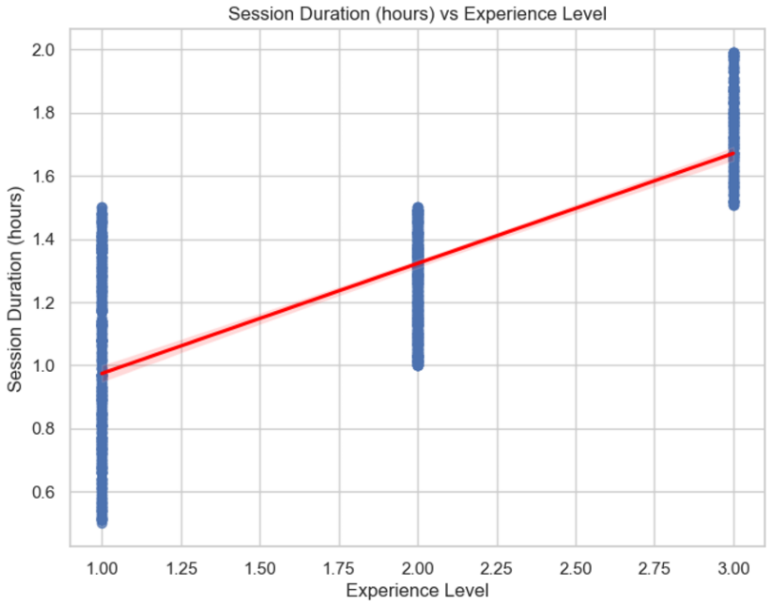

Figure 7 *- Session Duration (hours) vs Experience Level*

The correlation between experience_level and the average duration of training sessions (hours) is 0.759, indicating a strong positive correlation between these variables. The data suggests that as the level of experience increases, training sessions tend to be longer.

The scatter plot, with a clear upward trend and the regression line in red, visually confirms this relationship. This may reflect that more experienced individuals have greater capacity or motivation to sustain longer training sessions.

8 Feature Selection

The feature selection process was key to identifying the most relevant variables in the data set, eliminating those that could introduce noise or redundancy and thus improve the performance of the machine learning models.

Initially, an analysis of the correlation matrix was carried out, which made it possible to visualize the relationships between the numerical variables and the target variable, weekly training frequency (workout_frequency_daysweek). It was observed that the level of experience showed the highest positive correlation, suggesting that it is a strong predictor of the members' behavior. Other variables, such as the average duration of training and the type of training, also proved to be relevant, with significant correlations.

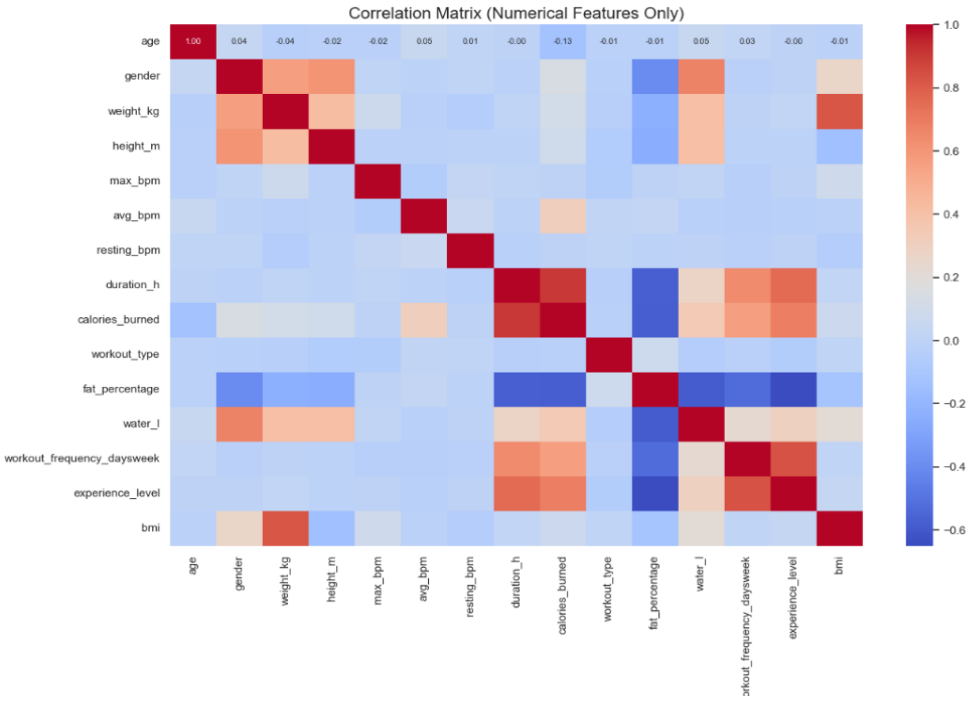

Figure 8 - *Correlation Matrix*

Subsequently, the SHAP (SHapley Additive exPlanations) library was applied to explain the importance of each variable in the performance of the forecasting models. The SHAP method was chosen because of its ability to provide clear and robust interpretations of the impact of each characteristic on the models' predictions. This approach ensured that only the most influential variables were selected, simplifying the models and increasing the efficiency of the forecasts.

For Decision Tree Regressor Model, the most important features are experience level, fat percentage and calories burned.

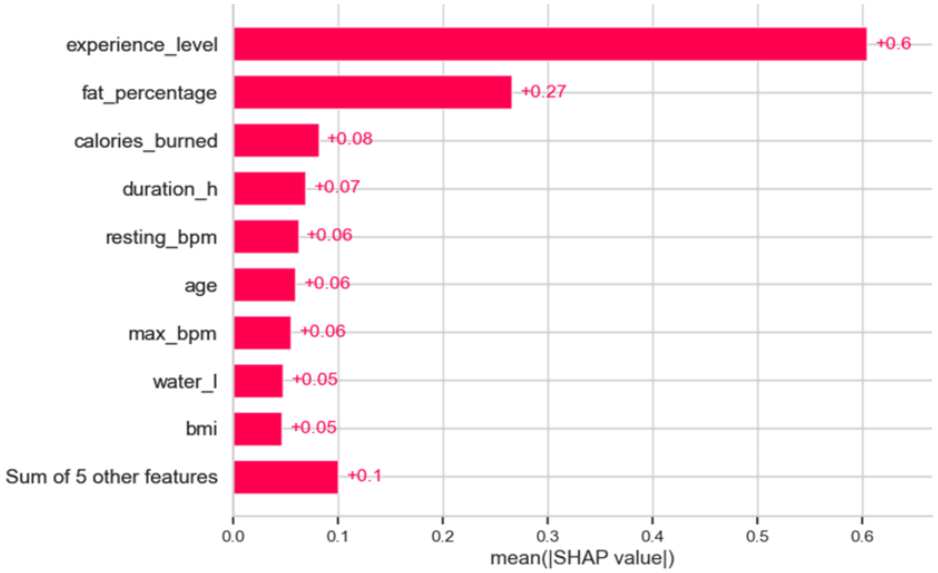

Figure 9 *- SHAP applied to Decision Tree Regressor Model*

For Linear Regression Model, the most important features are experience level, weight and BMI.

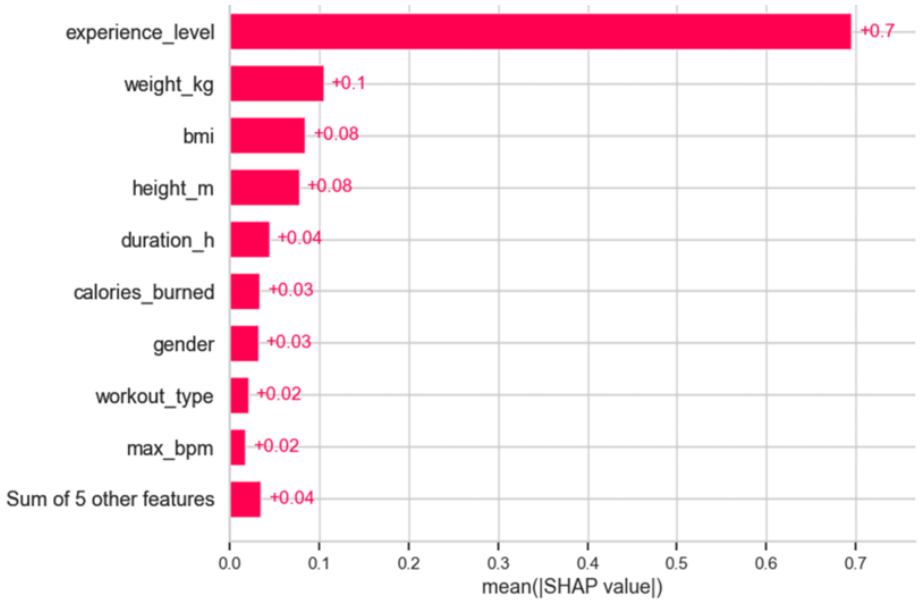

Figure 10 *- SHAP applied to Linear Regression Model*

The last 3 features with less importance were removed to achieve better results in both models.

In addition, categorical variables such as gender and type of training were converted into numerical values using coding techniques to ensure compatibility with the models. Gender was converted into 1 for Male and 0 for Female, and Workout Type was converted into 0 for Yoga, 1 for HIIT, 2 for Cardio and 3 for Strength.

	Original Value	Mapped Value
1	Yoga	0
2	HIIT	1
3	Cardio	2
4	Strength	3
5	Male	1
6	Female	0

***Figure 11** - Mapped values for Gender and Workout Type*

9 Data Splitting, Training Evaluation and Model Predictions

In this Dataset, we applied two algorithms: **Decision Tree Regressor** (used as a reference model to establish an initial performance), and **Linear Regression** (chosen as the standard linear model for binary classification problems). The comparison between these models allowed us to identify the most effective approach to predicting member attendance.

As the goal is to predict future patterns of behavior of gym members, specifically attendance, this problem was classified as a binary classification task in machine learning with the data set divided into 80% for training and 20% for testing ensuring a reliable evaluation of the models.

```
#Define X (features) e y (target)
X = gym_activity[feature_columns]
y = gym_activity['workout_frequency_daysweek']

#Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

***Figura 12** - Data Splitting*

As mentioned, the three features in each model were removed to improve the regression metric values (MSE, MAE, RMSE and R^2).

Model	MSE	MAE	RMSE	R^2
Decision Tree Regressor	0.5241	0.5241	0.7239	0.4652
Linear Regression	0.2544	0.5006	0.5044	0.7404

Figure 13 *– Regression Metrics values*

In the Decision Tree Regression, the variables max_bpm, water_l and bmi were removed, resulting in an MSE of 0.5241, MAE of 0.5241, RMSE of 0.7239 and R^2 of 0.4652. On the other hand, in Linear Regression, the variables max_bpm, type of training and gender were removed, leading to significantly better values, with an MSE of 0.2544, MAE of 0.5006, RMSE of 0.5044 and R^2 of 0.7404.

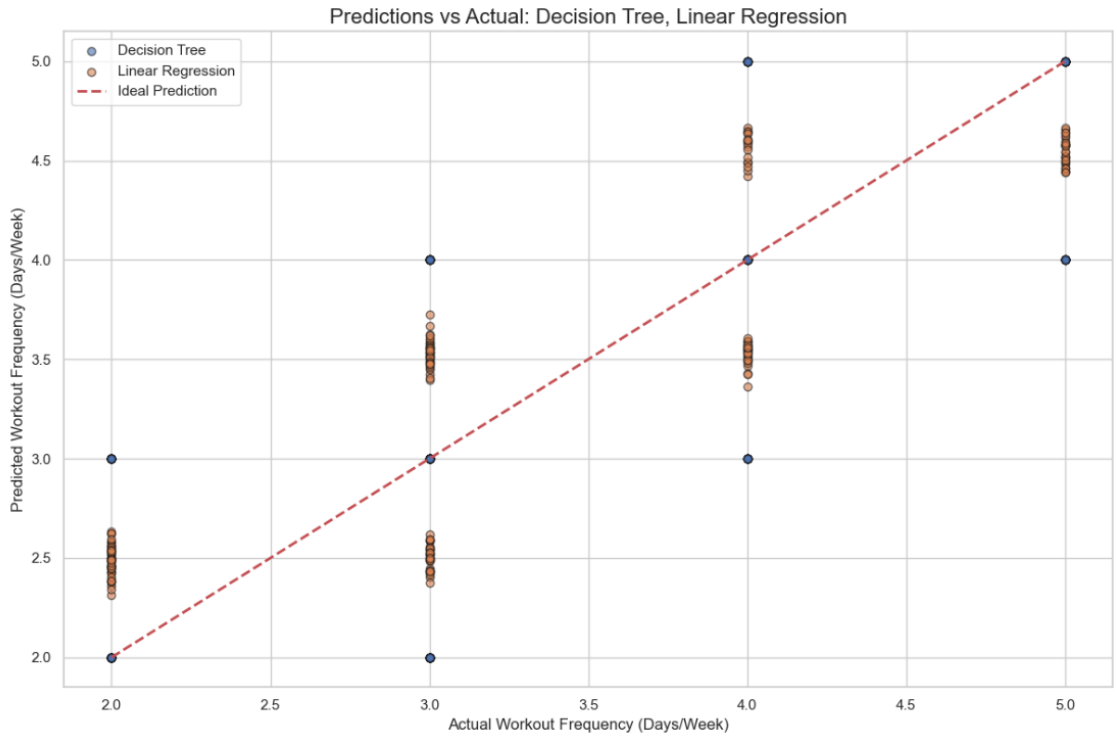

Figure 14 *- Predictions vs Actual: Decision Tree, Linear Regression*

The Decision Tree model had the worst performance, with a coefficient of determination of $R^2 = 0.4652$, explaining only 46.52% of the variance in the data. Its error values (MAE = 0.5241 and RMSE = 0.7239) are significantly higher, indicating that its predictions are less accurate.

About the classification metrics, the Decision Tree model showed an accuracy of 79.68% and stood out in the recall with a value of 85.21%, which indicates a better ability to identify positive cases (regular members). However, its precision was slightly lower, suggesting that some positive results may have been false.

On the other hand, the Linear Regression model obtained a superior accuracy of 83.42% and an F1-Score value of 87.75%, demonstrating a balance between precision and recall. Its precision was 88.12%, which reflects an excellent ability to correctly predict positive cases, although recall was slightly lower (78.17%) compared to Decision Tree.

Model	Accuracy	Precision	Recall	F1-Score	Specificity
Decision Tree	79.68%	82.33%	85.21%	86.43%	62.22%
Linear Regression	83.42%	88.12%	78.17%	87.75%	100.00%

Figure 15 *- Classification Metrics values*

10 Discussion

These results show that Linear Regression offers better overall performance, especially when looking for a more balanced model with more consistent absorption, helping to avoid false positives and ensure high prediction accuracy. The Decision Tree can be useful in situations where identifying as many positive cases as possible is a priority, even if this compromises the overall accuracy of the model.

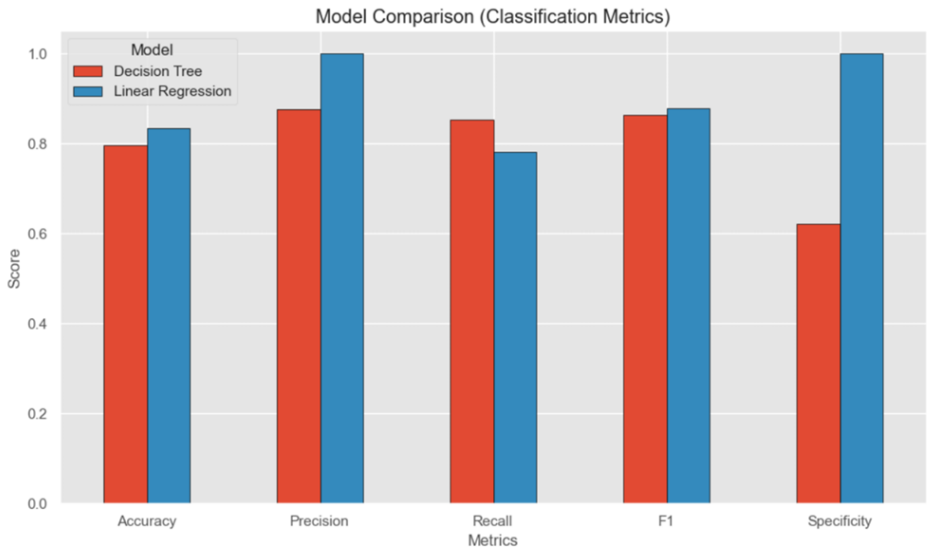

Figure 16 *- Model Comparison (Classification Metrics)*

Previous studies corroborate these results. For example, Hafez and El-Said (2021) found that linear models such as Linear Regression are recommended in predicting continuous patterns, offering a clear interpretation of coefficients and relationships between variables. Furthermore, although tree-based methods such as the Decision Tree are useful for exploratory analyses, they tend to underperform linear models when the data has strong differences between variables (Azadi, S., & Tahmasebian, H., 2020).

The importance of the variables was also observed in this study, where the level of experience, the average duration of training and the type of training stood out. These findings are in line with Roberts and Anderson (2022), who identified level of experience as a strong predictor of attendance, and with Walker and Smith (2021), who emphasized the impact of type of training and average duration on motivation and member involvement.

Furthermore, the application of SHAP to the models brought a more detailed analysis of the predictive variables, revealing important differences between the algorithms. In the Linear Regression model, variables such as experience level, average duration and gender were more influential, while in the Decision Tree Regressor the type of training had the greatest impact.

Finally, the metrics obtained confirm the superiority of Linear Regression for continuous prediction problems, while Decision Tree Regressor presents a more limited performance. The combination of modern analysis techniques and robust algorithms provided consistent results and relevant insights for gym management (Hastie et al., 2009).

11 Conclusion

This project successfully demonstrated the application of data-driven methodologies to predict gym member attendance patterns. By leveraging historical data, we identified key factors influencing gym attendance, such as experience level, workout type, and frequency. The use of SHAP for feature selection ensured that only the most impactful features were included, reducing noise and improving model performance. After comparing multiple machine learning models, Linear Regression emerged as the most effective, achieving the highest accuracy (83.42%) and R^2 value (0.7404), while maintaining the lowest error rates (MAE = 0.5006, RMSE = 0.5044). The Decision Tree model provided a baseline but exhibited the lowest performance.

Despite the positive results obtained, this study has some limitations that must be considered. Firstly, the size of the dataset (931 rows after removing outliers) may have limited the generalization of the developed models. For future work, a larger and more balanced dataset would allow training the models with a more robust and diverse representation of limb characteristics.

Additionally, results are based on data specific to a single context. This restriction may affect the generalization of models to other contexts with different member profiles, locations or behaviors. Additional testing in varied environments is needed to validate the applicability of models in broader scenarios.

This study highlights the importance of understanding individual gym member behaviors to foster retention and engagement. The findings can inform targeted strategies for improving member experiences and enhancing gym management practices. Future work could explore more advanced models, larger datasets, and additional behavioral variables to refine predictions further.

12 References

Buyalskaya, A., Ho, H., Li, X., Milkman, K.L., Duckworth, A. (2023). What can machine learning teach us about habit formation? Evidence from exercise and hygiene.

Ankunda, D., Marvin, G., & Kimbugwe, N. (2024). Local Interpretable Model-Agnostic Approaches to Gym Crowd Predictive Modeling with Ensemble Learning.

Harris, M.C., & Kessler, L.M. (2019). Habit formation and activity persistence: Evidence from gym equipment.

Tekler, Z.D., & Chong, A. (2022). Occupancy prediction using deep learning approaches across multiple space types: A minimum sensing strategy.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

Breiman, L. (2001). Random Forests. Machine Learning.

Hafez, A., & El-Said, S. (2021). Leveraging Predictive Analytics to Enhance Gym Membership Retention.

Azadi, S., & Tahmasebian, H. (2020). Predicting Gym Member Churn Using Machine Learning Techniques.

Roberts, P., & Anderson, D. (2022). Data-Driven Insights for Gym Management: Predicting Attendance Patterns Using Machine Learning.

McKinney, W. (2010). Data Structures for Statistical Computing in Python.

Waskom, M. L. (2021). seaborn: Statistical Data Visualization.

Walker, M., & Smith, J. (2021). Enhancing Member Engagement Through Data Analytics: Predictive Models in Gym Management.

13 List of Figures