REPORT

Analyzing Fitness Patterns & Building Predictive Models using Gym Members Exercise Dataset (Link to Dataset)

Team 27- tc3117, em3907, ap4613, tz2634, jy3344

INTRODUCTION

In an era where sedentary lifestyles are increasingly prevalent, understanding and promoting effective fitness behaviors has become crucial for public health.
Given the comprehensive dataset of gym members' exercise routines, physical attributes, and fitness metrics, we seek to explore the following questions by leveraging machine learning techniques:

1. What are the most significant factors influencing fitness outcomes across different demographic groups?
2. How can we develop predictive models for personalized exercise recommendations that are tailored to individual characteristics and fitness goals?
3. What patterns in workout adherence and progression can be uncovered to inform strategies for improving long-term engagement with fitness routines?

DATASET INTERPRETATION

The dataset comprises 973 data points, each characterized by 15 features spanning demographic, physiological, and workout data. Importantly, no missing values were identified, ensuring data completeness for analysis. The features were categorized into three groups for clarity: demographic data includes \`Age\`, \`Gender\`, \`Weight\`, and \`Height\`; physiological data includes variables such as \`Max BPM\`, \`Avg BPM\`, and \`BMI\`; and workout data captures key exercise metrics like \`Workout Session Duration\`, \`Workout Frequency\`, and \`Calories Burned\`. The target variable, \`Calories_Burned\`, was chosen due to its practical relevance in guiding health and fitness decisions, making it a valuable outcome for real-world applications.

EXPLORATORY DATA ANALYSIS AND DATA PREPROCESSING

A total of 44 outliers were identified across several features but were retained to preserve the dataset's diversity and inclusivity. We will mitigate their impact using robust scaling techniques.

Bivariate analyses revealed meaningful patterns across the dataset. Males exhibited slightly higher median calorie burns compared to females, although the overall distributions were similar. Among workout types, calorie burns were largely comparable, with slightly greater variability in Strength and Yoga workouts, possibly reflecting differing intensity levels. The \`Experience_Level\` feature demonstrated a clear upward trend, with more experienced individuals burning significantly more calories, indicating its potential as a strong predictor.

Correlation analysis further emphasized the importance of certain features. \`Session_Duration (hours)\` showed the strongest positive correlation with \`Calories_Burned\` (r = 0.91), confirming its role as a key predictor. Other features, such as \`Workout_Frequency\` (r = 0.64) and \`Avg BPM\` (r = 0.34), also demonstrated moderate correlations with the target variable. However, multicollinearity was observed between some features, such as \`Weight\` and \`BMI\` (r = 0.85), which could introduce redundancy. To address this, regularization techniques like Lasso will be employed during model training to penalize less informative features and improve interpretability. Standard Scaling, Robust Scaling for outliers and one-hot encoding have been selected for feature preprocessing to ensure data readiness for predictive modeling.

MACHINE LEARNING

In order to answer the first part of the problem statement that determines the most significant factors that influence fitness outcomes, we have chosen the target as the most important fitness outcome everyone is interested in real life: **Calories_Burned.** Post setting the target, we have used 5 regression models to determine the performance, implemented hyperparameter tuning to improve the model performance and based on that chosen the best model and determined the most significant features using feature importance. The **regression models** that have been implemented are as follows: **Linear Regression, Ridge Regression, Lasso Regression, Support Vector Machines, RandomForest Regressor and FeedForward Neural Network.**

Secondly, to tackle the second part of the problem statement where the idea is to develop predictive models for personalized exercise recommendations that are tailored to individual characteristics and fitness goals we have performed the following tasks. The idea is to recommend the users what workout type they could follow to maximize fitness results. To implement this we have analyzed the formation of clusters given the target variable **Workout_Type**.

- We divided the dataset into **6 groups** based on **gender** and **experience level** to explore clustering patterns within these subgroups to identify correlations between **workout type** and the group a user belongs to.
- Applied **DBSCAN** to identify density-based clusters, using **Euclidean distance** and adjusting parameters epsilon and min_samples
- Used **PCA** for dimensionality reduction, selecting the first 6 components that explained **60-70%** of the variance.
- Implemented **K-Means clustering** to partition users into clusters but found no correlation between workout type and clusters.
- Applied **Neural Collaborative Filtering (NCF)** to model **user-item interactions** and recommend top 2 workouts to new users using **deep learning**
- We try to explore the User-Item interaction by including information about other user features too. We hope to identify complex user-item interactions through this method and recommend top 2 workouts to the users.We measure the performance of the recommender system with MAP@K

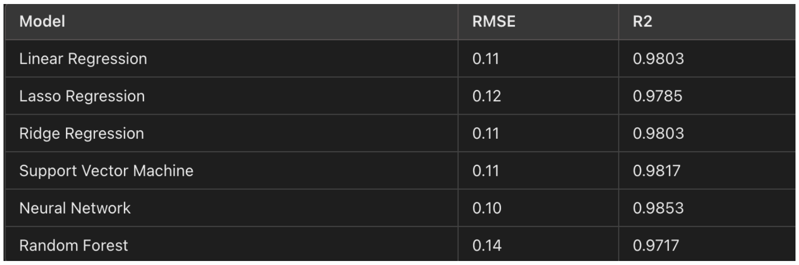
On getting the top 2 recommenders, we have built on top of it a suggestion based model which recommends that given the workout type as an input feature itself, for the given workout type how many calories can the user burn in the suggested session duration. To implement this we have used a **RandomForest Regressor** as a multi output regressor for the targets **Calorie_Burned and Session Duration** together. We have further performed hyperparameter tuning using n_estimators, min_samples_split, min_samples_leaf and max_depth with different values to maximize the model performance and then choose the best set of hyperparameters to train the model.

Lastly, to address the problem of recognizing patterns in workout adherence and progression, we chose Experience_Level as the target and performed a classification using RandomForest Classifier, Logistic Regression Classifier and Support Vector Machine and we have then tuned the hyperparameters.. Post this implementation, we have performed progression analysis by grouping the various factors like workout frequency, session duration, water intake and calorie burned with respect to the Experience Level and identified results that would help with the business perspective.

PERFORMANCE AND INFERENCE

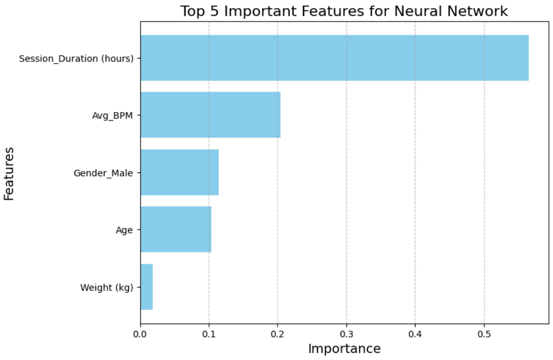
Performance Analysis for problem statement 1:

Based on our results, **Neural Network with small relu** provides the best results and here are the metric values and the five most significant features that influence fitness performance.



Model	RMSE	R2
Linear Regression	0.11	0.9803
Lasso Regression	0.12	0.9785
Ridge Regression	0.11	0.9803
Support Vector Machine	0.11	0.9817
Neural Network	0.10	0.9853
Random Forest	0.14	0.9717

Performance Results



Feature Importance

Performance Analysis for problem statement 2 :

DBSCAN- Even after tuning epsilon and minimum points, the majority of the points were classified as noise. This means the Data is uniformly distributed. DBSCAN works well for dense clusters. So we try other clustering algorithms like k-means clustering. We also perform PCA on the data picking the first 6 principal components to facilitate dimensionality reduction. We hope that we might find clusters in lower dimensional space.

PCA- On our data, we perform PCA and pick the first 6 principal components. When we look at the first 6 principal components for each subgroup of our dataset, adding them up, we get roughly 60-70%. This is sufficient for dimensionality reduction for better visualization.

K-means- We do not observe any correlation between the clusters and type of work-out even after adjusting the cluster centroids. This is to say that, even if we manage to group similar users together k-means, we do not find a common tendency to their choice of work-out types within this group.

NCF- Constructing a ranking matrix is not possible with our dataset, and no clustering patterns were observed based on which ranking matrix can be constructed. So, the concept of dot product similarity and filling the missing matrix with factorization of lower-rank matrices would not be feasible. So, we try to explore the User-Item interaction by including information about other user features too. We hope to identify complex user-item interactions through this method and recommend top 2 workouts to the users.We measure the performance of the recommender system with MAP@K.

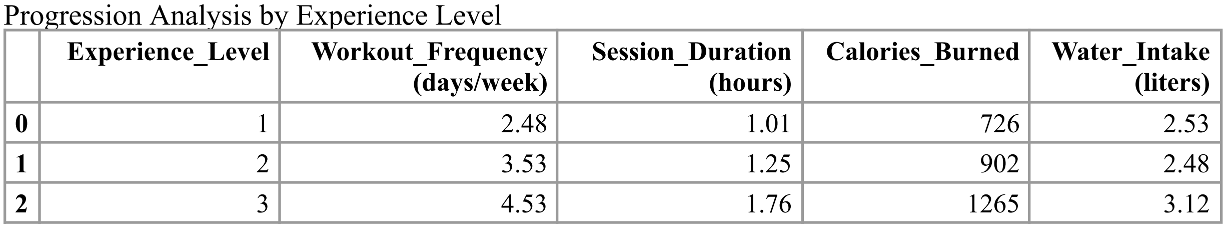
It computes the average precision for each query up to the top **K** recommended items and then takes the mean across all queries. It considers both the relevance of recommended items and their ranking order. Higher MAP@K values indicate better alignment between predicted rankings and user preferences.

We get 0.535 MAP@2 Score. From a 20% accuracy in classification, 0.535 MAP@2 is an improvement. The model is severely underfitting due to very little data ~ 1000 users only. If we provide more user data to the model, the underfitting problem can be addressed.

RandomForest Regressor for Calorie Burned and Session Durations give a decent R^2 value: 0.65 and 0.62 respectively. The RMSE score for Calorie Burned is ~ 173 and given the wide range of calorie values in our dataset, this value seems acceptable. For Session Duration, the RMSE score ~ 0.2.

Performance Analysis for problem statement 3:

The RandomForest Classifier model performs well in classifying the three levels of experience and achieves an accuracy ~ 92%, with perfect recall and precision values for level 3.



Progression Analysis by Experience Level

	Experience_Level	Workout_Frequency (days/week)	Session_Duration (hours)	Calories_Burned	Water_Intake (liters)
0	1	2.48	1.01	726	2.53
1	2	3.53	1.25	902	2.48
2	3	4.53	1.76	1265	3.12

This analysis suggests that people who have achieved level 3 are already performing to the best of their capacities to achieve their fitness goals. We can see that people at level 3 have almost twice the values that people at level 1.Therefore, from a business standpoint this analysis suggests that we can add more incentives for level 1 people so that they can progress faster and achieve experience level 3 and could get closer to achieving their fitness goals.

CONCLUSION

This project successfully identified **Session Duration, Avg_BPM, Gender, Age and Weight** as the primary factors influencing calorie burn, providing actionable insights for fitness optimization. Using Neural Collaborative Filtering, we developed a personalized recommendation system, achieving a MAP@2 score of 0.535 and demonstrating the potential for tailoring exercise plans to individual goals. Progression analysis highlighted significant differences in workout adherence and performance across experience levels, suggesting opportunities to enhance engagement, particularly for beginners. These results emphasize the value of leveraging machine learning to drive data-driven fitness solutions, with future work focusing on scaling models and improving predictive accuracy through larger datasets.