Fitness Insights: Analyzing Exercise Patterns and Performance Across Experience Levels

Link to repository: https://github.com/angelinazhuma/Gym-insights

Team Members: Angelina Zhumadilova

Task 2: Business understanding

1. Understanding Your Business Goals

Background:

A study of how gym users at different levels of experience behave with their fitness exercise can provide insights for gyms, trainers and gym-goers. This study analyzes the types of workouts, session durations, and performance measures to identify trends that can help improve the fitness experience of gym-goers.

Business Goals:

1. Assist gym trainers in creating fitness plans that align with the specific needs of beginners, intermediate, and advanced gym members.
2. Enable gyms to better allocate resources, such as equipment, based on the preferences and habits of different member groups.
3. Provide fitness app developers with insights to improve their personalization algorithms, ensuring users receive workout recommendations that match their experience levels and fitness goals.

Business Success Criteria:

- The analysis generates actionable insights, such as commonly preferred workout types or session durations for each experience level.
- Results are visually clear, making them easy for stakeholders like gym managers and trainers to interpret and apply.
- Machine learning models achieve a minimum accuracy of 85% when predicting key fitness outcomes such as calories burned or exercise types.

2. Assessing Your Situation

Inventory of Resources:

- Dataset: A Gym Members Exercise Dataset, containing information on exercise types (cardio, strength), session duration, demographics (age, gender), and performance metrics (calories burned, weight lifted).
- Tools: Python, Jupyter Notebooks, data analysis libraries (pandas, NumPy), visualization libraries (matplotlib, seaborn), and machine learning frameworks (scikit-learn).
- Support: Access to online tutorials and machine learning communities for troubleshooting and guidance.

Requirements, Assumptions, and Constraints:

- Requirements:
    - A clean, sufficiently large dataset with balanced representation across experience levels.
    - Accurate performance metrics for reliable predictions.
- Assumptions:
    - Members provide honest and accurate data about their workouts and performance metrics.
    - Exercise types and durations recorded are reflective of actual gym behavior.
- Constraints:
    - Limited computational resources may slow down model training, especially for deep learning approaches.
    - Time constraints for project completion.

Risks and Contingencies:

- Risk: Missing or incomplete data could reduce the reliability of analysis. Contingency: Apply data preprocessing techniques such as imputation and data augmentation to handle gaps.
- Risk: Model predictions could be skewed by an imbalance in the dataset between experience levels. Contingency: Use resampling techniques or weighted loss functions to handle imbalance.
- Risk: The inability to make predictions with high accuracy because of problems with data quality. Contingency: Investigate sophisticated modeling approaches and employ hyperparameter adjustment.

Terminology:

- Experience Level: Classification of gym members as beginner, intermediate, or advanced based on their workout history or skill level.
- Performance Metrics: Measurable outcomes of gym sessions, such as calories burned or weight lifted.
- Data Augmentation: Techniques to increase dataset diversity by modifying existing entries.

Costs and Benefits:

- Costs: Time and effort for data cleaning, model development, and validation, important resources required for analysis and model training.
- Benefits:
    - Improved personalization of fitness plans for members.
    - Enhanced member satisfaction and retention through data-driven decisions.
    - Insights that can support gym operations and fitness app development.

3. Defining Your Data-Mining Goals

Data-Mining Goals:

- Classification: Predict a gym member's experience level (beginner, intermediate, or advanced) based on demographic and workout data.
- Regression: Estimate performance metrics, such as calories burned, using input variables like session duration, exercise type, and member demographics.

- Trend Analysis: Identify patterns and preferences in exercise types, session durations, and workout frequencies across different member groups.

Data-Mining Success Criteria:

- Classification models achieve an accuracy of at least 85% in predicting experience levels.
- Regression models produce fitness performance estimates with a Root Mean Squared Error (RMSE) below an acceptable threshold (10%).
- Data visualizations effectively communicate key insights, enabling gym trainers and managers to make informed decisions.

Task 3: Data Understanding

1. Gathering Data

Outline Data Requirements

For the analysis of fitness behaviors and performance across gym members, the dataset must include:

- Demographics: Age, gender, weight, height, BMI, and experience level.
- Physiological Metrics: Heart rate metrics (Max_BPM, Avg_BPM, Resting_BPM), body fat percentage, and water intake.
- Workout Information: Session duration, workout type, calories burned, and weekly workout frequency.

Verify Data Availability

The dataset contains 973 samples with the required fields. Key features include demographics, workout-specific metrics, and performance indicators, making it suitable for achieving the project's goals.

Define Selection Criteria

Selection criteria will include entries where:

- All essential features (e.g., Age, Gender, Workout_Type, Session_Duration, Calories_Burned) are available.
- Physiological and workout data are within realistic ranges to ensure data reliability..

2. Describing Data

- Demographics:
    - Age (integer): Member’s age, ranging from 18 to 65 years.
    - Gender (categorical): Male or Female.
    - BMI (float): Derived from weight and height.
    - Experience_Level (ordinal): Beginner (1), Intermediate (2), or Expert (3).
- Physiological Metrics:
    - Max_BPM, Avg_BPM, and Resting_BPM (integer): Heart rate metrics reflecting workout intensity.
    - Fat_Percentage (float): Proportion of body weight as fat.

    - Water_Intake (float): Daily water consumption during workout days (liters).
- Workout Details:
    - Workout_Type (categorical): Cardio, Strength, Yoga, or HIIT.
    - Session_Duration (float): Time spent on workouts (hours).
    - Workout_Frequency (integer): Number of sessions per week.
    - Calories_Burned (float): Total energy expenditure in a session.

3. Exploring Data

Demographic Patterns

- Most members fall between 25–40 years, with a nearly equal gender distribution.
- BMI ranges from 18.5–30 for 85% of the dataset, reflecting a mix of healthy and overweight individuals.

Workout Trends

- Workout_Type: Cardio dominates among beginners, while advanced members show an even split between strength training and HIIT.
- Session_Duration:
    - Beginners: ~1 hour.
    - Intermediate: ~1.5 hours.
    - Experts: >2 hours.
- Calories_Burned: Positively correlates with session duration and workout intensity.

Performance Metrics

- Heart Rate:
    - Resting_BPM averages at ~70, with Max_BPM reaching ~180 in high-intensity sessions.
    - Advanced members show higher Avg_BPM, indicating greater workout intensity.
- Water_Intake: Experts consume more water (~3 liters/day), reflecting higher hydration needs.

---

4. Verifying Data Quality

Completeness

- Missing values are minimal (<2%), primarily in Fat_Percentage and Water_Intake.

Accuracy

- BMI values were recalculated from weight and height for verification and aligned with reported values.
- Heart rate values are consistent with expected ranges for gym-goers.

Task 4: Project Plan

1. Data Understanding and Cleaning
   Clean the dataset. This includes filling in missing values, addressing outliers, and preparing the data for analysis by encoding categorical variables.
   - Estimated time: 12 hours, team member: Angelina
2. Exploratory Data Analysis
   Visualize data, investigate relationships. It includes creating visualizations to highlight demographic patterns, workout preferences, and performance metrics.
   - Estimated time: 12 hours, team member: Angelina
3. Feature Engineering and Selection
   - Estimated time: 8 hours, team member: Angelina
4. Model Development
   This includes classification models to predict experience levels and regression models for metrics like calories burned. Hyperparameter tuning and model evaluation will ensure accuracy and reliability.
   - Estimated Time: 12 hours, team member: Angelina
5. Results Analysis
   Results from the models will be analyzed to interpret findings, evaluate performance metrics (accuracy, RMSE), and generate actionable insights for stakeholders.
   - Estimated time: 10 hours, team member: Angelina
6. Poster presentation
   Prepare poster and presentation text, also posts and photos.
   - Estimated time: 15 hours, team member: Angelina

Methods and Tools

- Data Analysis: Python (pandas, NumPy)
- Visualization: matplotlib, seaborn, and plotly for clear and interactive visualizations.
- Modeling: scikit-learn for classification and regression tasks, including hyperparameter tuning.
- Reporting: Jupyter Notebooks, GoogleDocs