F5 Report

Magnus Traks, Hannes Arumäe, Reimo Kaabel

Task 2:

1. Identifying your business goals

Background

Fitness and health are becoming more important. This is because people try to maintain a healthy lifestyle. Gyms collect enormous amounts of data about their members' activities and demographics. Analyzing these datasets can extract valuable insights to improve individual fitness plans. Increase efficiency in resource allocation and improve member experience This project uses exercise routines. Physical features and gym member fitness metrics to understand exercise patterns and progress.

Business goals

1. Identify fitness patterns: Explore relationships between demographic factors (age, gender) and fitness outcomes to inform personalized training programs.
2. Evaluate workout effectiveness: Determine which workout types are most effective for specific goals, such as calorie burning, fat reduction, or heart rate improvement.
3. Assess recovery needs: Understand how different factors, such as heart rate metrics and workout intensity, influence recovery requirements.

Business success criteria

1. Provide insights for fitness trainers to create data-supported personalized workout plans.
2. Based on the user's physical characteristics such as age, gender, weight and height, as well as the experience they have with training and the fitness aims they have, help fitness lovers understand how to optimize their workout plans.
3. Develop a detailed report summarizing key discoveries and their relevance to the field of health and fitness.

2. Assessing your situation

Inventory of resources

1. **Data:** The dataset contains 973 samples with attributes like age, gender, weight, height, heart rate metrics, calories burned, and workout types.
2. **Tools:** Access to data analytics software (Python, Jupyter Notebook) and visualization libraries (e.g., Matplotlib, Seaborn).
3. **Expertise:** Knowledge in machine learning, data preprocessing, and fitness science.

Requirements, assumptions, and constraints

Requirements:
Access to relevant tools and resources.
Familiarity with exercise-related indicators and their effects

Assumptions:
Data sets are considered complete and accurate. with no significant missing data.
The simulated dataset reflects the fitness model well in reality.
Variables like workout type and session duration are recorded consistently.

Constraints:
Data sets may lack real-world noise or unexpected anomalies. This limits generalizability and limits reliance on stable updates.

Risks and contingencies
Risk: Misinterpretation of results due to the synthetic nature of the dataset.
Contingency: Cross-reference findings with fitness studies.
Risk: Misinterpretation of results due to domain complexity.
Contingency: Collaborate within the team for result validation

Terminology

1. Max_BPM: Maximum heart rate (beats per minute) during workout sessions.
2. Avg_BPM: Average heart rate during workout sessions.
3. BMI: Body Mass Index, calculated from height and weight, an indicator of body fat based on weight and height.
4. Water_Intake (liters): Daily water intake during workouts.
5. Fat_Percentage: Body fat percentage of the member.
6. Experience_Level: A measure of gym proficiency, ranging from 1 (beginner) to 3 (expert).
7. Workout_Type: The type of workout performed (e.g., Cardio, Strength, Yoga, HIIT)

Costs and benefits

Costs: Time investment, hardware and software access

Benefits: Improved understanding of fitness optimization, enhanced training programs, and better-informed gym management strategies.

3. Defining your data-mining goals

Data-mining goals:

1. Predict trends in fitness metrics based on demographic and physiological attributes to identify distinct fitness patterns.
2. Identify the most effective workout types for specific demographic groups or fitness objectives.

3. Model recovery needs based on factors like resting BPM, calories burned, and workout frequency.

Data-mining success criteria:

1. Model performance should achieve over 80% accuracy in predicting fitness patterns.
2. Accurately predict the best suitable workout type based on the input features.
3. Present findings in a clear, actionable format with supporting visualizations.

Task 3:

1. Gathering data

Outline data requirements

The dataset must provide detailed gym member metrics, including demographics, workout metrics, and physiological indicators. Specific requirements include:

Time range: Data generated should represent a broad spectrum of realistic gym activity patterns.
Format: Tabular data in CSV

Field requirements:

1. **Demographic data:** Age, gender
2. **Physical metrics:** Weight, height, BMI, fat percentage
3. **Workout details:** Session duration, workout type, workout frequency, calories burned
4. **Health indicators:** Heart rate metrics (max BPM, avg BPM, resting BPM), water intake

Verify data availability:

The dataset includes all required variables for the analysis, with no immediate evidence of missing critical fields. The database is available and accessible to work on.

Define selection criteria:

The dataset (gym_members_exercise_tracking.csv) will initially include all 973 rows and 15 columns for exploration.

2. Describing data

Source: gym_members_exercise_tracking.csv
Format: Tabular data, suitable for analysis in Jupyter Notebook

The dataset includes the following fields:

1. **Age:** Numeric, continuous; range from 18 to 59.
2. **Gender:** Categorical; values are "Male" or "Female."
3. **Weight (kg):** Numeric, continuous; member's weight in kilograms.
4. **Height (m):** Numeric, continuous; member's height in meters.
5. **Max_BPM:** Numeric, continuous; maximum heart rate during workouts.
6. **Avg_BPM:** Numeric, continuous; average heart rate during workouts.
7. **Resting_BPM:** Numeric, continuous; heart rate at rest.
8. **Session_Duration (hours):** Numeric, continuous; time spent per workout.
9. **Calories_Burned:** Numeric, continuous; energy expenditure per session.
10. **Workout_Type:** Categorical; values include "Cardio," "Strength," "Yoga," "HIIT."
11. **Fat_Percentage:** Numeric, continuous; body fat percentage.
12. **Water_Intake (liters):** Numeric, continuous; daily water intake.
13. **Workout_Frequency (days/week):** Numeric, discrete; number of sessions per week.
14. **Experience_Level:** Numeric, discrete; ranges from Beginner (1) to Expert (3).
15. **BMI:** Numeric, continuous; calculated from height and weight.

Suitability:

The dataset meets the requirements for initial analysis and meets the project's data-mining goals. All required fields are present, and there is sufficient diversity in the variables to explore relationships and trends.

3. Exploring Data

1. **Demographics:**
    - Age**:** Ranges from 18 to 59 years (mean: 38.7). The dataset has a wide age span, covering young adults to middle-aged individuals.
2. **Physical attributes:**
    - Weight (kg)**:** Ranges from 40 to 129.9 kg (mean: 73.9 kg), with a standard deviation of 21.2 kg, indicating significant diversity in the sample.
    - Height (m): Ranges from 1.5 to 2.0 meters (mean: 1.72 m), with a narrow standard deviation (0.13 m).
    - BMI**:** Body Mass Index varies from 12.3 to 49.8 (mean: 24.9). The high maximum indicates potential outliers or individuals with high body mass.

3. **Health indicators:**
    - Max BPM: Maximum heart rates range between 160 and 199 bpm (mean: 179.9).
    - Avg BPM: Average heart rates range from 120 to 169 bpm (mean: 143.8).
    - Resting BPM: Ranges from 50 to 74 bpm (mean: 62.2), reflecting healthy resting heart rates.
    - Water intake (liters): Ranges from 1.5 to 3.7 liters (mean: 2.63 liters).

    - Experience level: A 3-level ordinal variable (1=Beginner, 2=Intermediate, 3=Expert). Mean: 1.81.
4. **Workout details:**
    - Session duration (hours): Ranges from 0.5 to 2 hours (mean: 1.26 hours).
    - Calories burned: Varies from 303 to 1783 calories (mean: 905.4). The high variability aligns with diverse workout intensities.
    - Workout frequency (days/week): Values span from 2 to 5 days (mean: 3.3).

Distributions:

- Most variables appear to have plausible ranges and distributions.
- Outliers might exist in BMI and Calories burned.

Data quality:

- No missing values were detected.
- Categorical variables have a manageable number of unique values.

4. Verifying data quality

Completeness:
Preliminary checks indicate no missing values in key fields.

Accuracy:

- **BMI validation:** Cross-checked BMI values with height and weight; no discrepancies identified.
- **Heart rate metrics:** Verified that Avg_BPM < Max_BPM for all records; no anomalies found.

Relevance:
All data fields are relevant to the goals, and no redundant or unrelated fields are present.

Consistency:
Categorical fields such as Gender and Workout_Type are standardized. Numeric fields are consistently formatted.

Identified issues:
Apart from the possible outliers, we identified no other issues.

Task 4:

Tasks	**Description**	**Hours allocation(per member)**	**Total hours**	**Comments**
Data understanding	Explore the dataset. Verify data quality and identify outliers/issues.	Magnus: 3 Hannes: 2 Reimo: 3	8	
Data preparation	Clean data if necessary, handle outliers, and create new features if necessary.	Magnus: 6 Hannes: 6 Reimo: 5	17	
Exploratory data analysis	Analyze useful patterns, using statistical methods and visualizations.	Magnus: 7 Hannes: 6 Reimo: 6	19	Focus on regression, clustering
Model development	Build models to train and test the data.	Magnus: 9 Hannes: 8 Reimo: 8	25	
Input field implementation	Design and implement an input field for users to provide their physical attributes to get their personalized results.	Magnus: 2 Hannes: 2 Reimo: 2	6	Users can input their physical attributes like age, gender, height, and weight
Reporting and presentation	Summarize findings and prepare a poster.	Magnus: 3 Hannes: 6 Reimo: 6	15	Ensure clarity and emphasis on actionable insights.

Tools and methods

1. **Tools:**
    a. Jupyter Notebook
    b. Pandas
    c. NumPy
    d. Python
    e. Matplotlib

f. Plotly
g. Scikit-learn
h. Canva
i. HTML
j. CSS

2. **Methods:**
   a. Heatmaps
   b. Histograms
   c. Boxplots
   d. Scatterplots
   e. Website for input field
   f. Decision trees
   g. Clustering
   h. Regression analysis

github link: https://github.com/ReimoK/KAGGLE-GYM-DATA