Discovery of Training Patterns and Exercise Progression Dynamics in Personal Strength Training Data

Carlos Anthony Cruz
CS 4412: Data Mining
Kennesaw State University
Ccruz31@students.kennesaw.edu

***Abstract*—This project proposes a comprehensive data mining analysis of 2 years of personal strength training logs to discover patterns in exercise progression, workout structure, and performance. Using association rule mining and clustering techniques, the aim is to uncover hidden relationships between exercises and identify natural training phases. The dataset consists of structured workout sessions with exercise names, weights, repetitions, and sets, providing a rich source for temporal pattern discovery in athletic performance.**

I. Dataset Description

A. Overview and Source

The dataset comprises personal strength training logs spanning January 2026 and extending back 2 years (exact range to be confirmed during data collection phase). The data is recorded in a structured text format, documenting each training session with exercise names, load (weight in pounds with plate notation), and repetitions performed.

Dataset name: Personal Strength Training Log Dataset (PSTLD)

Availability: Self-collected dataset; not publicly hosted due to its personal nature.

Dataset characteristics:

- **Source:** Personal training logs (self-collected)
- **Time span:** 2 years of consistent training data
- **Estimated size:** 400–800 workout sessions, 8,000–15,000 individual exercise sets
- **Format:** Structured text files organized by month

Approximate dimensionality:

- **Rows:** 8,000–15,000 exercise-level records
- **Columns:** 8–12 attributes per record

B. Data Structure and Key Features

Each workout session is organized by training split (e.g., "Torso 1", "Limbs 2") and contains multiple exercises with the following attributes:

Core features:

- **Exercise name:** E.g., "Legend shoulder", "DB bench", "RDLs"
- **Weight/Load:** Recorded in plate notation (e.g., "3PPs+30" = 3 plates per side + 30lbs) or direct weight for dumbbells
- **Repetitions:** Number of reps performed per set
- **Sets:** Multiple working sets per exercise
- **Date:** Session timestamp
- **Workout type:** Training split category
- **Exercise modifiers:** Special notations like "+" for paused sets or set extenders depending on the exercise

Derived features for analysis:

- Total volume (weight × reps × sets)
- Relative intensity (percentage of historical max)
- Exercise frequency and ordering
- Rest periods between sessions
- Strength fluctuations across weeks
- Progressive overload rate

Table 1 shows a representative sample of the raw data structure.

TABLE I
Sample Raw Data Structure

Exercise	Weight	Reps
Legend shoulder	3PPs+30	3
Legend shoulder	3PPs+2.5	7.5
DB bench	90	8
DB bench	80	12
Dips	2PPs	6.5
Dips	1PPs+10	11

C. Data Quality Considerations

Known issues and preprocessing needs:

- **Notation inconsistencies:** Weight notation varies (plate notation vs. direct weight), requiring standardization
- **Exercise name variations:** Same exercise may have slight naming differences (e.g., "DB OHP" vs. "DB overhead press")
- **Incomplete sessions:** Some sessions may have missing exercises or abbreviated notation
- **Fractional reps:** Half reps occasionally noted (e.g., "7.5"), requiring handling decisions
- **Equipment changes:** Same exercise on different machines (e.g., "Med X Leg curls" vs. "Strive Leg ext")
- **Set extenders:** Some isolation movements have set extenders included in the set (e.g., "Tricep pushdown: 150/13+3")

- **Range loading within machines:** Strive machines have different pegs with loaded ranges of motion (short vs. lengthened) (e.g., “(4) SL Leg Extension: 150/5”)

Preprocessing strategy:

1) Parse and normalize weight notation to absolute values
2) Standardize exercise names using fuzzy matching and manual mapping
3) Handle missing values and incomplete sessions
4) Create temporal features (day of week, session gap, training phase)
5) Calculate derived metrics (volume, intensity, frequency)

II. Discovery Questions

This project focuses on three discovery questions that explore patterns in training behavior, exercise relationships, and performance dynamics.

A. Question 1: Exercise Association Patterns

What exercises are frequently performed together within the same workout session, and what compound movement patterns emerge across different training splits?

Rationale: Understanding which exercises naturally cluster together can reveal training principles, muscle group synergies, and programming structure.

B. Question 2: Training Phase Segmentation

Can we identify distinct training phases based on volume, intensity, and exercise selection patterns without prior labeling?

Rationale: Discovering these phases through unsupervised clustering can validate intuitive training decisions and reveal unrecognized patterns.

C. Question 3: Performance Variation Patterns

What characterizes sessions with unusually high or low performance compared to typical training sessions?

Rationale: Identifying anomalous sessions can provide insights into recovery, fatigue, or breakthrough performances.

III. Planned Techniques

I plan to apply data mining techniques from multiple categories to address my discovery questions.

A. Association Rule Mining

The Apriori algorithm will be used to discover exercises frequently performed together within the same workout session. Each workout session will be treated as a transaction with exercises as items.

B. Clustering

K-Means clustering will be used to identify natural groupings in training sessions based on features such as total volume, average intensity, and exercise selection. The elbow method will be used to determine the optimal number of clusters.

C. Anomaly Detection

To address Discovery Question 3, anomaly detection techniques such as z-score analysis and Isolation Forests will be applied to identify training sessions with unusually high or low performance metrics. These methods will help characterize outlier sessions related to recovery, fatigue, or performance fluctuations.

D. Implementation

These techniques will be implemented using Python with the `mlxtend` library for association rule mining and `scikit-learn` for clustering and anomaly detection. Data preprocessing will be handled using `pandas`, and results will be visualized using `matplotlib` and `seaborn`.

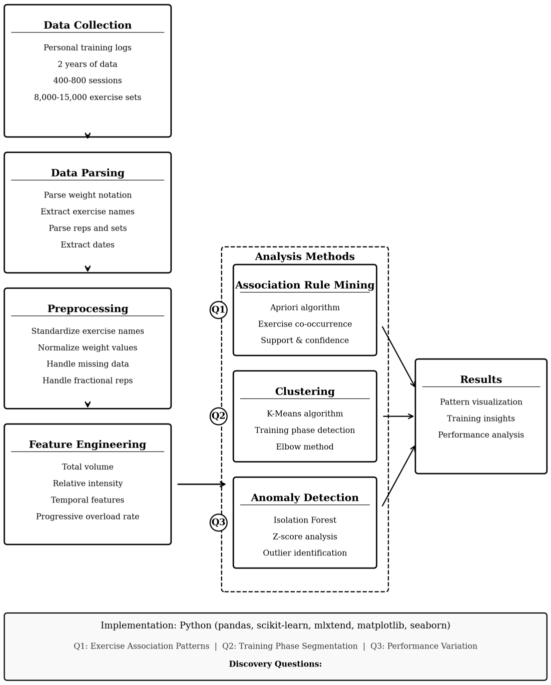

Fig. 1. Planned data mining workflow from raw training logs to pattern discovery

IV. Preliminary Timeline

A. Milestone 2: Data Preparation and EDA

Target completion: Week 6–7

Deliverables:

- Complete data collection
- Data parsing and cleaning pipeline
- Exploratory data analysis
- Feature engineering

Anticipated challenges:

- Parsing inconsistent notation formats
- Handling exercise name variations
- Determining appropriate feature representations

B. Milestone 3: Pattern Mining and Analysis

Target completion: Week 10–11
Deliverables:

- Association rule mining results
- Clustering and anomaly detection analysis
- Pattern interpretation

C. Milestone 4: Final Report and Presentation

Target completion: Final week
Deliverables:

- Final report
- GitHub repository with code and documentation
- Presentation materials

V. Conclusion

This project leverages a unique personal dataset to explore strength training patterns using data mining techniques. By applying association rule mining, clustering, and anomaly detection, the project aims to uncover meaningful insights into exercise relationships and training periodization. All code and analysis artifacts will be maintained in a GitHub repository following course submission guidelines.