Powerlifting Performance Analysis

Data Mining on the OpenPowerlifting Dataset

---

1. Introduction

This project analyzes the OpenPowerlifting dataset, which contains detailed records of strength athletes' performances from global competitions. The primary objectives are to:

- Explore and preprocess the data,
- Uncover patterns and relationships among lifter characteristics and outcomes,
- Apply feature selection and dimensionality reduction,
- Build predictive models for classification and regression,
- Perform clustering to identify natural groupings in the data.

---

2. Data Selection and Preprocessing

Data Selection:

- The original dataset (`openpowerlifting.csv`) was filtered to retain essential columns:

`Best3SquatKg`, `Best3BenchKg`, `Best3DeadliftKg`, `Age`, `Country`, `MeetState`, `Tested`.

- Rows with missing values in these columns were dropped.
- Individual lift attempt columns were removed.
- A random sample of 2,000 rows was selected for analysis and saved as `sample_openpowerlifting.csv`.

Preprocessing:

- Numeric and categorical summaries were generated.
- Missing values were identified and handled.
- Distributions of key numeric features were visualized (see `distribution_plots.png`).

Summary Statistics (from `summary_report.txt`):

- **Sample Size:** 2,000 lifters.
- **Mean Age:** 34.1 years.
- **Mean Bodyweight:** 81.8 kg.
- **Mean Total Lifted:** 515.2 kg.
- **Minimal missing data** after preprocessing.

---

3. Exploratory Data Analysis (EDA)

- **Correlation Heatmap:**

Numeric features show strong correlations, especially among lift totals and scoring metrics (`correlation_heatmap.png`).

- **Boxplots:**
  - Males lift more on average than females.
  - Equipped lifters (vs. raw) have higher totals (`boxplots_totalkg.png`).
- **Federation & Country Trends:**
  - USA and a few federations dominate participation (`barplots_federation_country.png`).
- **Distributions:**
  - Age, bodyweight, and lift totals visualized for normality and outliers (`distribution_plots.png`).

---

4. Feature Selection

- **Methods Used:**
  - Mutual Information and Random Forest feature importance (see `feature_selection_analysis.py`).
- **Top Features (Both Methods):**
  - Age, BodyweightKg, WeightClassKg, Best3SquatKg, Best3BenchKg, Best3DeadliftKg, TotalKg, Wilks.
- **Comparison:**
  - Both methods agree on key predictors.
  - Visual comparison saved as `feature_selection_comparison.png`.
- **Model Evaluation:**
  - Random Forest classifier with top 5 features from each method.
  - Cross-validation used to assess performance.

---

5. Dimensionality Reduction (PCA)

- **PCA performed on numeric features** (see `pca_analysis.py`):
  - Features standardized before PCA.
  - First 5 principal components explain ~97.5% of variance.

- Feature contributions visualized (`pca_feature_contributions.png`).

- **Modeling with PCA:**
  - Random Forest classifier trained on first 5 PCs.
  - Performance compared to feature selection methods.
  - Explained variance visualized (`pca_explained_variance.png`).

---

6. Classification Analysis

- **Target:**
  - Binary classification: StrongLifter (Wilks above median).
- **Models Evaluated:**
  - Logistic Regression, KNN, Decision Tree, Random Forest (see `classification.py`).
- **Results:**
  - Random Forest and Logistic Regression performed best.
  - 10-fold cross-validation and overfitting analysis included.
  - Class balancing performed by downsampling.

Example Results (from code output):

- Random Forest:
  - Accuracy, Precision, Recall all high (exact values in code output).
  - 10-fold CV and overfitting analysis confirm robustness.

---

7. Regression Analysis

- **Target:**
  - Predicting TotalKg (total weight lifted).
- **Models Evaluated:**
  - Linear Regression, KNN, Decision Tree, Random Forest (see `regression.py`).
- **Results:**
  - Random Forest Regression achieved highest R^2.
  - KNN optimized for best n_neighbors.
  - Model performances compared by R^2 and RMSE.

Example Results (from code output):

- Random Forest Regression:
  - Highest R^2, lowest RMSE among models.

---

8. Clustering

- **Hierarchical Clustering:**
  - Performed on raw and standardized numeric features (`clustering_hierarchical.py`).
  - Dendrograms visualized for both (`dendrogram_raw.png`, `dendrogram_standardized.png`).

---

9. Key Visualizations

- Distribution plots (`distribution_plots.png`)
- Correlation heatmap (`correlation_heatmap.png`)
- Boxplots by sex and equipment (`boxplots_totalkg.png`)
- Federation and country barplots (`barplots_federation_country.png`)
- Feature selection comparison (`feature_selection_comparison.png`)
- PCA explained variance and contributions (`pca_explained_variance.png`, `pca_feature_contributions.png`)
- Clustering dendrograms (`dendrogram_raw.png`, `dendrogram_standardized.png`)

---

10. Conclusion

- Data mining reveals key predictors of powerlifting performance.
- Random Forest feature selection and PCA both provide robust models.
- Random Forest models generally outperform others in both classification and regression.
- Clustering and EDA provide additional insights into the structure and trends in the data.
- The analysis can inform athlete training and competition strategies.

Appendix:

- All code files and generated plots are included in the project folder for reference and reproducibility.
- Original Dataset: https://www.kaggle.com/datasets/open-powerlifting/powerlifting-database?resource=download
- GitHub: https://github.com/HCanKayim/Powerlifting-Performance-Analysis