Predicting Physical Activity Levels and Intervention Success of the PE4MOVE study in Adolescents Using Machine Learning

Anna-Lena Klöckner, Jakob Werner

Application Domain

This project lies at the intersection of sports science, youth physical activity promotion, and applied machine learning. It uses real-world data from the PE4MOVE school-based intervention study, which integrates motivational modules into physical education (PE) lessons to promote healthier and more active lifestyles among adolescents.

Problem Description

Low levels of physical activity (PA) in youth represent an increasing public health concern, influencing long-term health, fitness, and psychosocial well-being. School-based interventions attempt to increase PA, yet individual responses differ widely: some adolescents substantially increase activity levels, while others do not. Identifying which factors predict baseline PA and which individuals benefit the most from interventions would support more personalized program design and allow greater intervention success.
This project addresses two central predictive tasks:

1. Predict intervention success, operationalized as change in PA or fitness outcomes from pre-intervention (T0) to post-intervention (T1).

2. Identify which features have the greatest impact on intervention success.

These tasks were chosen as they address key challenges in school-based physical activity promotion. Predicting intervention success helps identify factors that influence individual responsiveness to PE-based interventions, supporting more effective and targeted program design. Feature importance analysis ensures model interpretability and provides insight into which demographic, motivational, psychological, or fitness-related variables most strongly contribute to behavioral change. Together, the tasks combine predictive accuracy with explainability, enhancing their practical relevance for research and practice.

Data Sources

- **PE4MOVE Study Dataset**: Includes demographic variables, daily and weekly indicators of moderate to vigorous physical activity (MVPA), psychosocial constructs (motivation, self-efficacy, social influences) and physical fitness measures (e.g. standing long jump, handgrip strength).

- **PE4MOVE Trial Protocol**: Provides measurement design, variable definitions, and intervention components.

Methodology and System Architecture

The project follows a standard machine learning workflow. First, data is preprocessed by handling missing-values, normalization, and categorical encoding. MVPA across days is aggregated into weekly totals, and the success of the intervention is calculated as a change in scores (e.g. $\Delta\text{MVPA} = \text{MVPA}_{T1} - \text{MVPA}_{T0} > 0$). Feature engineering includes scale-building for motivational constructs, combining fitness metrics (e.g physical fitness test score) and obtaining the most influencial features.

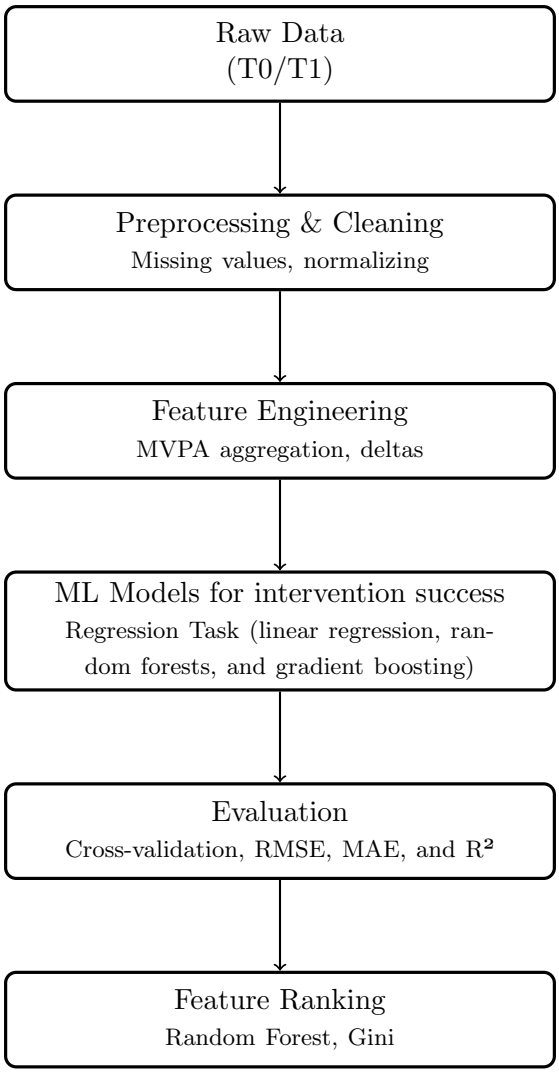

Figure 1: System architecture for MVPA prediction and intervention success modelling.

Multiple prediction models will be evaluated: linear regression, random forests, and gradient boosting for continuous outcomes and feature ranking. Evaluation uses cross-validation and metrics such as RMSE, MAE, and R^2.
Linear regression is used as an interpretable baseline model, while random forests and gradient boosting are chosen because they can capture non-linear relationships and interactions between demographic, psychological, and fitness-related features. Cross-validation is applied to ensure reliable performance estimates and to reduce overfitting. Model performance is evaluated using regression metrics such as RMSE, MAE, and R^2, which quantify prediction error and explained variance for continuous intervention outcomes.

Rationale

This proposal ensures feasibility by defining clear tasks, available data sources, and a robust methodological plan. The focus on feature importance and intervention responsiveness – aligns with the main research questions in promoting physical activity in youth.
Given the complexity of the dataset, the preprocessing and cleaning required – particularly handling missing values, normalization, and feature construction – the project benefits from the collaborative effort of two students. Dividing responsibilities across data preparation, modeling, evaluation, and interpretability ensures a thorough analysis and sufficient methodological depth within the project timeframe.