Predicting Physical Activity Levels and Intervention Success in Adolescents Using Machine Learning

Anna-Lena Klöckner, Jakob Werner

Abstract

Low levels of physical activity (PA) among adolescents represent a major public health concern, with implications for physical health, mental well-being, and long-term disease risk. Although school-based interventions aim to increase PA, their effectiveness varies considerably between individuals. This project applies machine learning techniques to data from the PE4MOVE school-based intervention study to predict physical activity levels and intervention success in adolescents.

Using demographic, psychosocial, physical activity, and fitness-related variables measured before and after the intervention, multiple regression models were trained to predict changes in PA outcomes. Linear regression served as an interpretable baseline, while random forests and gradient boosting models were used to capture non-linear relationships and interactions. Model performance was evaluated using cross-validation and regression metrics, including RMSE, MAE, and R^2. Feature importance analysis was conducted to improve interpretability and identify key predictors of intervention response.

The results demonstrate that intervention success is moderately predictable from baseline characteristics. The findings highlight the role of existing behavior and physical capacity in shaping intervention outcomes and illustrate the potential of machine learning to support more targeted physical activity promotion strategies.

Code repository:
`https://github.com/annalena13403/PE4Move_Project`

Contents

1 Application Domain

1.1 Physical activity and health in adolescence

Regular physical activity during adolescence is associated with numerous health benefits, including improved cardiorespiratory fitness, healthier body composition, and positive effects on mental health. Adolescence is a critical period for establishing long-term physical activity habits, yet a large proportion of young people fail to meet current physical activity guidelines. Insufficient physical activity during this stage of life increases the risk of chronic disease and sedentary behavior in adulthood.

1.2 School-based physical activity interventions

Schools provide an ideal setting for physical activity promotion, as they reach nearly all children and adolescents regardless of socioeconomic background. Physical education (PE) plays a central role in these efforts, often combining physical training with educational and motivational components. However, previous research has shown that school-based interventions typically produce only small-to-moderate average effects, with considerable inter-individual variability in outcomes.

1.3 Motivation-focused approaches

Modern intervention designs increasingly target psychosocial determinants of physical activity, such as motivation, self-efficacy, and social support. The PE4MOVE intervention integrates motivational modules into PE lessons to promote autonomous motivation and long-term participation in physical activity. Despite this theoretically informed approach, not all participants benefit equally, motivating the use of predictive modeling techniques.

2 Description of Data Sources

2.1 PE4MOVE study overview

This project is based on data from the PE4MOVE school-based intervention study, which was designed to promote physical activity through motivationally enriched physical education lessons. The study includes pre-intervention (T0) and post-intervention (T1) measurements. The participants were adolescents who attended regular physical education classes. The intervention integrated motivational components grounded in behavioral theory into standard PE curricula, with the aim of increasing moderate-to-vigorous physical activity (MVPA) as well as related psychosocial and fitness outcomes.

The analysis focused on the intervention group only, with baseline characteristics used to predict individual changes in physical activity outcomes following the intervention. This setup reflects a realistic application scenario in which predictions must be made prior to intervention delivery.

2.2 Dataset structure

The dataset consists of a heterogeneous set of variables collected at baseline and follow-up. The variables were grouped into several conceptual domains to support structured preprocessing and modeling.

- **Demographics:** age, sex, and school-related information
- **Physical activity:** daily and weekly MVPA indicators
- **Psychosocial variables:** motivation, self-efficacy, and social influences
- **Physical fitness:** performance-based fitness tests (e.g. standing long jump, handgrip strength)

2.3 Data complexity

The PE4MOVE dataset reflects the characteristics of a real-world health intervention study rather than a controlled benchmark dataset. It contains a mixture of continuous, ordinal, and categorical variables measured on different scales and units. Several variables are self-reported and therefore subject to measurement error. The relatively high number of features with respect to the sample size further increases the analytical complexity.

3 Methodology and System Architecture

3.1 Problem formulation

The central objective of this project is to predict individual responses to a school-based physical activity intervention using only baseline data. Rather than predicting absolute physical activity levels, the task focuses on modeling change in physical activity after the intervention, thereby directly addressing intervention effectiveness at the individual level.

Intervention success was operationalized as the difference between the post-intervention (T1) and baseline (T0) measurements of physical activity outcomes, such as moderate-to-vigorous physical activity (MVPA). Using change scores allows the models to capture improvements or declines over time while implicitly accounting for baseline differences between participants.

The prediction task was formulated as a supervised regression problem, where continuous outcome variables (e.g. ΔMVPA) were modeled as functions of baseline characteristics. In addition to predictive performance, the project aimed to identify which baseline factors – demographic, behavioral, fitness-related, or psychosocial – are most strongly associated with positive changes in physical activity, supporting both methodological evaluation and practical relevance for intervention design.

3.2 Data preprocessing

Data preparation and preprocessing represented the most time-consuming and methodologically demanding component of the project. The dataset included a considerable amount of missing data resulting from incomplete questionnaire responses, absent fitness test results, and partial follow-up participation. Missing data patterns varied considerably across variables, requiring variable-specific handling strategies rather than a single global approach.

The heterogeneous nature of the dataset further increased complexity. Continuous variables differed widely in scale and distribution, while categorical variables required encoding compatible with all machine learning models. Several psychosocial constructs were measured using multiple questionnaire items, necessitating careful aggregation prior to inclusion in the feature set.

Physical activity measures required aggregation and transformation to ensure interpretability and comparability. Daily or frequency-based indicators were consolidated into weekly mea-

sures to align with the modeling objectives. Additional preprocessing steps included detecting and resolving implausible values, standardizing fitness test scores, and ensuring consistent units across measurements.

For selected variables with structured missingness, imputation was performed using the strongest available correlating variables, allowing missing values to be estimated while preserving sample size and limiting distortion of the underlying data structure.

All preprocessing steps were implemented using reproducible pipelines to ensure consistent transformation of training and test data and to prevent information leakage during model evaluation.

3.3 Feature engineering

Feature engineering was performed to improve model performance, interpretability, and theoretical coherence. Given the high dimensionality and conceptual overlap among raw variables, domain-informed feature construction was necessary.

Psychosocial questionnaire items were aggregated into scale scores that represent motivational and self-regulation constructs. This aggregation reduced noise associated with individual questionnaire items and ensured alignment with established behavioral theory. Scale construction followed predefined scoring rules and was applied consistently across participants.

Physical fitness measures were standardized and, where appropriate, combined to capture broader aspects of physical capacity rather than isolated test performance. Baseline physical activity variables were retained in multiple forms to represent different behavioral dimensions, including structured exercise, leisure activity, and sedentary behavior. Contextual variables such as extracurricular sports participation and physical education exposure were retained to reflect environmental influences on behavior change.

All engineered features were derived exclusively from baseline (T0) data to maintain a realistic prediction setting. Feature construction balanced domain knowledge with practical considerations, ensuring robustness while maintaining interpretability.

In general, the complexity of the PE4MOVE dataset required extensive preprocessing and feature construction before machine learning models could be applied. This effort was essential to ensure the validity of later analysis and to enable a meaningful interpretation of the model results.

3.4 Machine Learning

3.4.1 Objective and Prediction Targets

The goal of the machine learning analysis was to examine whether changes in physical activity following the intervention could be predicted from baseline participant characteristics. Rather than relying on a single outcome, separate models were trained for different physical activity change measures to capture distinct behavioral dimensions.

The following outcome variables were modeled independently as changes between baseline (T0) and follow-up (T1):

- Δ MVPA Frequency
- Δ MVPA during a Usual Week
- Δ Leisure Exercise
- Δ Leisure Physical Activity

3.4.2 Input Features

All models used the same set of baseline attributes as predictors. These included demographic variables, psychological measures, fitness-related indicators, and baseline physical activity measures. Using an identical feature set across outcomes ensured that differences in predictive performance were attributable to the outcome variables rather than changes in model inputs. Prior to model training, features were standardized and strings encoded to allow a fair comparison between models and to support regularized regression methods.

3.4.3 Data Splitting

To evaluate generalization performance, the dataset was divided into a training set and a held-out test set:

- Training set: 557 participants, used for model fitting and hyperparameter tuning
- Test set: 140 participants, used exclusively for final evaluation

The test set was not involved in any stage of model selection.

3.4.4 Model Selection and Training

A range of regression models was evaluated to compare linear and non-linear approaches. Linear models with regularization were included for their interpretability and robustness to correlated predictors, while non-linear models were tested to assess whether more complex relationships could improve prediction performance.

The following models were evaluated:

- Ridge regression
- Lasso regression
- Elastic Net
- Random Forest regression
- Gradient Boosting regression
- Support Vector Regression
- k-Nearest Neighbors regression

For each outcome variable, models were compared using cross-validated R^2 scores on the training set. Although model selection was performed separately for each outcome, Ridge regression, Support Vector Regression, and Random Forests consistently ranked among the top-performing models across outcomes. These three models were therefore selected for further tuning and evaluation on the held-out test set.

3.4.5 Hyperparameter Tuning

Hyperparameter tuning was conducted using GridSearchCV with five-fold cross-validation on the training set. For each outcome, model-specific hyperparameter grids were explored and performance was assessed using cross-validated R^2.

3.4.6 Evaluation Metrics

Model performance was evaluated using several complementary metrics in order to capture both explained variance and the absolute prediction error. The coefficient of determination (R^2) was used to assess how much of the variance in the outcome variable is explained by the model:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

An R^2 value of 1 indicates perfect prediction, while a value of 0 means that the model performs no better than predicting the mean of the outcome. Negative values are possible and indicate that the model performs worse than a mean-based prediction. Therefore, higher R^2 values reflect better explanatory power.

To quantify the prediction error on the original outcome scale, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) were computed:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \qquad \text{MAE} = \frac{1}{n}\sum|y_i - \hat{y}_i|$$

Both RMSE and MAE range from 0 to positive infinity, with lower values indicating more accurate predictions. RMSE penalizes larger errors more strongly due to squaring, while MAE provides a more direct measure of the average absolute prediction error.

All reported metrics were calculated on the held-out test set to ensure that performance estimates reflect true out-of-sample generalization.

3.5 System architecture

Figure 1 illustrates the overall system architecture, which follows a modular pipeline design that separates data handling, modeling, and evaluation into clearly defined stages. This structure ensures transparency, reproducibility, and flexibility when testing different modeling approaches.

The architecture allows individual components, such as preprocessing strategies or learning algorithms, to be modified without affecting the overall pipeline. This was particularly important given the exploratory nature of the analysis and the need to iteratively refine preprocessing and feature construction decisions.

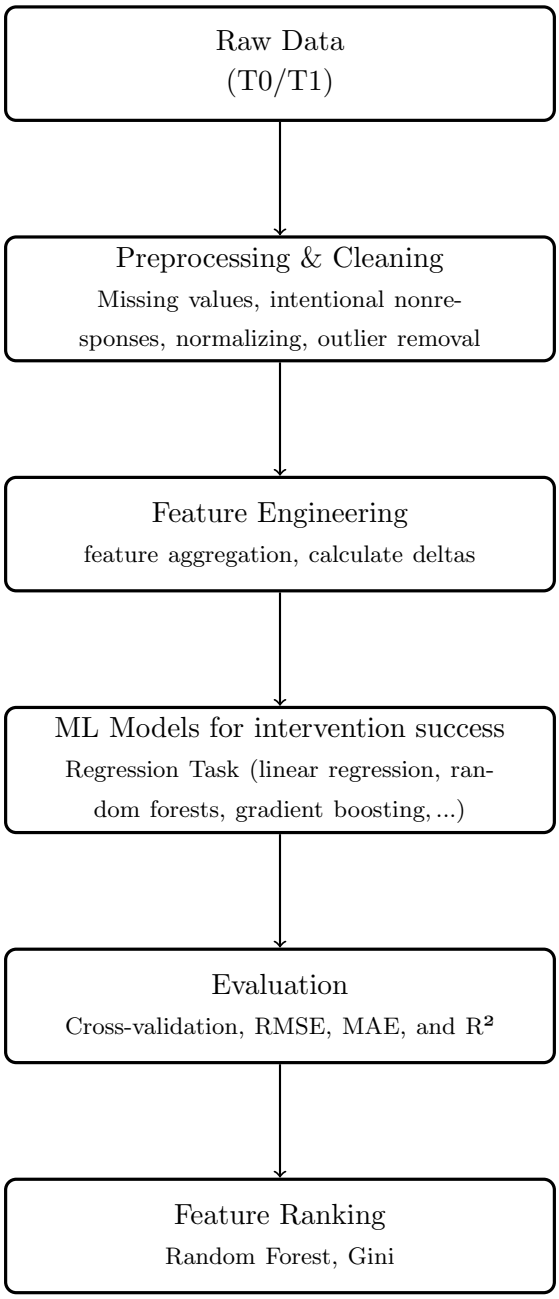

Figure 1: System architecture for predicting physical activity outcomes and intervention success.

4 Results

4.1 Predictive Models

Figure 2 shows that hyperparameter tuning consistently improved model performance across outcomes and model types. In nearly all cases, tuned models achieved higher cross-validated R^2 scores compared to their default versions, with particularly strong improvements observed for Support Vector Regression and Random Forest models. This demonstrates that systematic hyperparameter optimization was an important step in achieving reliable predictive performance.

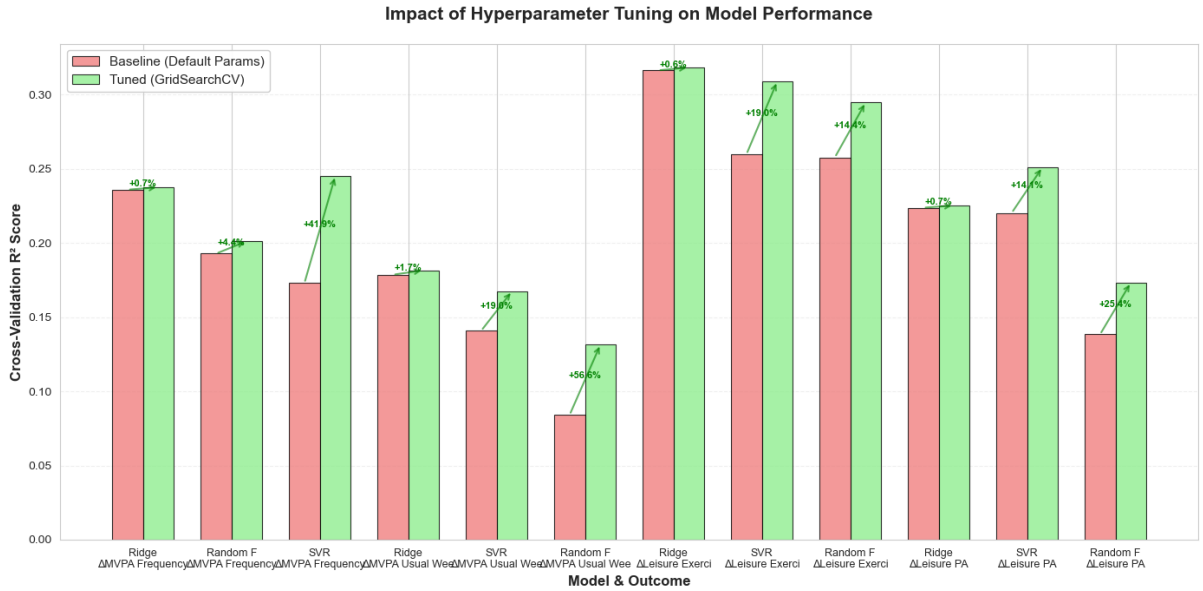

Figure 2: Impact of hyperparameter tuning on cross-validated R^2 performance. Bars compare default model settings with tuned models using GridSearchCV. Percentage annotations indicate relative performance improvements.

Across the full set of seven evaluated models, Ridge regression, Support Vector Regression, and Random Forests consistently emerged as the best-performing approaches (Figure 3). Other models, such as k-Nearest Neighbors and Gradient Boosting, did not achieve comparable performance and showed weaker generalization. This pattern suggests that a combination of regularized linear models and selected non-linear approaches provided the best balance between flexibility and robustness for the present data.

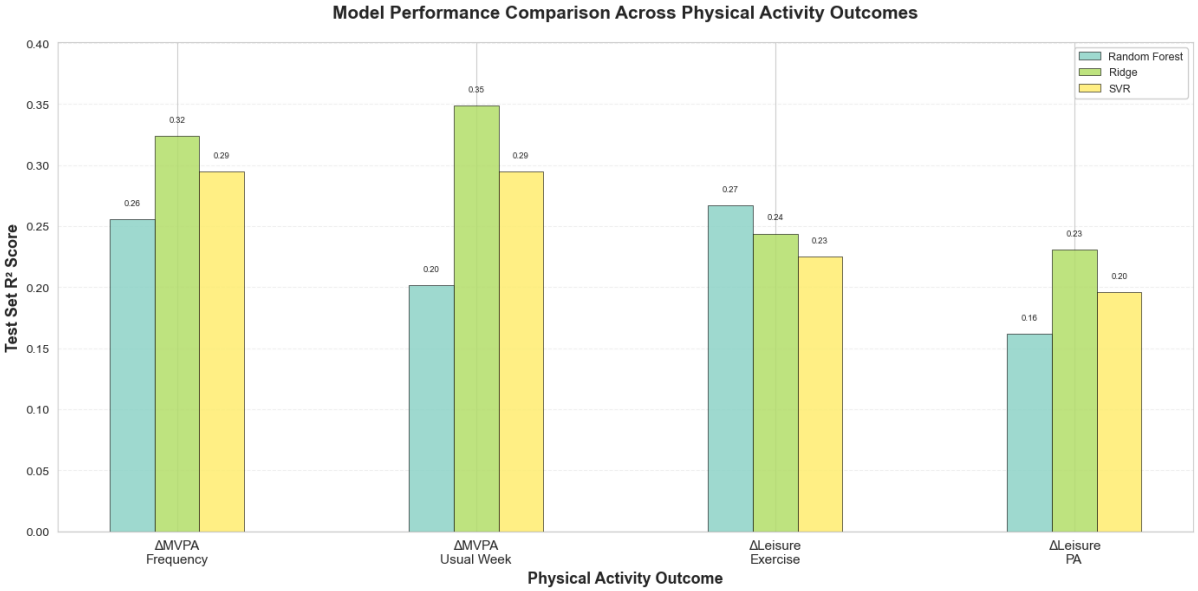

Figure 3: Comparison of test set R^2 performance for the three best-performing models (Ridge, Support Vector Regression, and Random Forest) across all physical activity outcomes.

Final performance results on the held-out test set are summarized in Figure 4. Ridge regression achieved the highest predictive performance for changes in MVPA frequency, MVPA during a usual week, and leisure physical activity, with test set R^2 values ranging from approximately 0.23 to 0.35. For changes in leisure exercise, Random Forest models performed best, indicating that non-linear relationships may play a slightly larger role for this outcome.
Overall, the achieved R^2 values indicate modest but meaningful predictive power, which is expected for behavioral change outcomes influenced by many unobserved factors. Error metrics (MAE and RMSE) were relatively similar across outcomes and models. This consistency suggests that the models did not perform very well on one outcome while failing on others.

Machine Learning Performance Summary: Best Models per Outcome

Outcome	Best Model	R^2	MAE	RMSE	N (train)
ΔMVPA Frequency	Ridge	0.324	1.14	1.41	557
ΔMVPA Usual Week	Ridge	0.349	1.00	1.25	557
ΔLeisure Exercise	Random Forest	0.267	0.88	1.20	557
ΔLeisure PA	Ridge	0.231	0.59	0.77	557

Figure 4: Summary of best-performing models for each outcome evaluated on the held-out test set, including R^2, MAE, and RMSE.

Taken together, these results show that hyperparameter tuning considerably improved model performance and that Ridge regression, Support Vector Regression, and Random Forests were the most suitable models for predicting intervention-related changes in physical activity.

4.2 Feature Importance

To better understand which baseline variables contributed most to predicting changes in physical activity, feature importance analyses were conducted using Random Forest models. Although Random Forests were not the best-performing models in terms of prediction accuracy, they were used here because they provide intuitive importance scores that help explore which variables are most influential across outcomes.

Figure 5 shows the top-ranked features for each outcome separately. Across all outcomes, a clear and consistent pattern emerges: baseline measures of the same or closely related physical activity domain are the strongest predictors of change. For example, baseline leisure exercise is the dominant predictor of changes in leisure exercise, and baseline MVPA measures strongly predict changes in MVPA-related outcomes. This indicates that initial behavior plays a central role in shaping how much change occurs during the intervention.

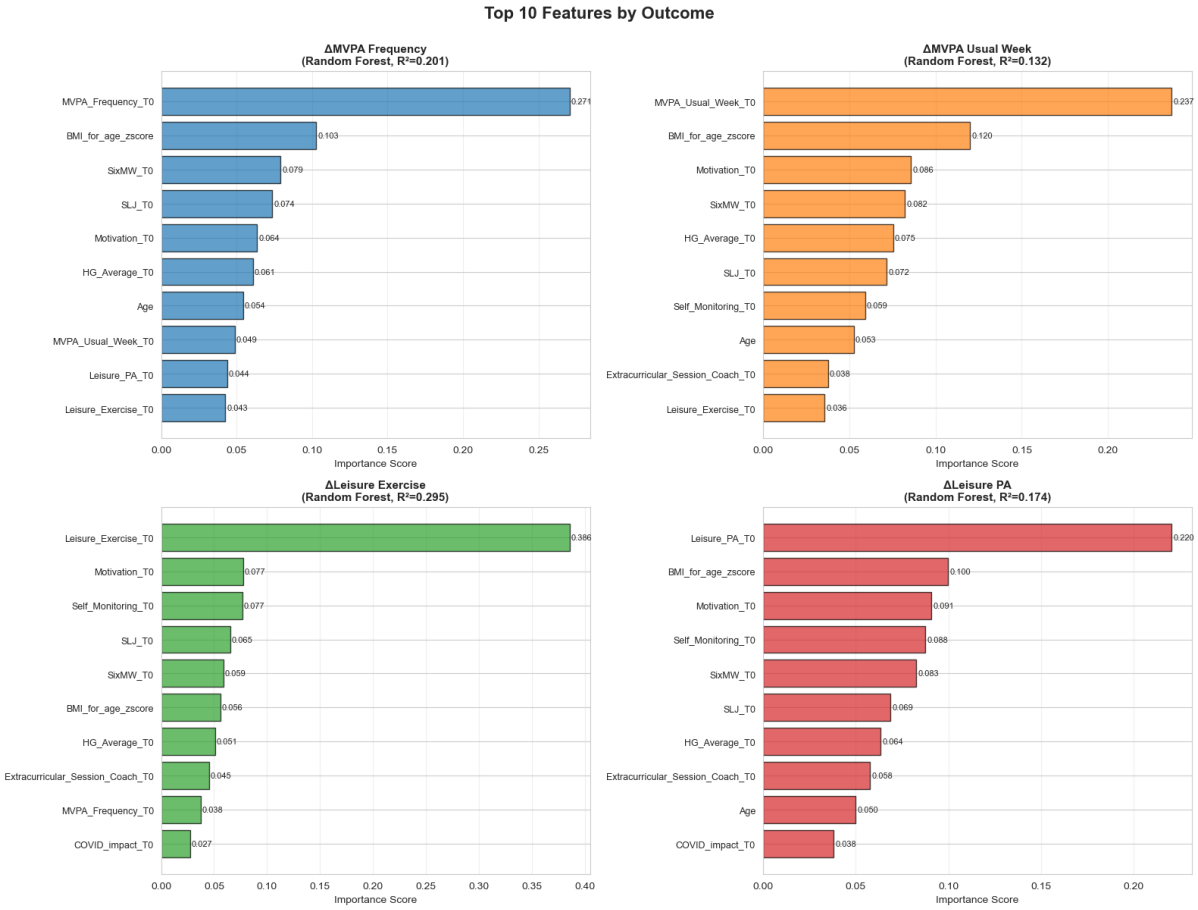

Figure 5: Feature importance patterns across all outcomes based on Random Forest models. Each panel shows the top-ranked predictors for one intervention outcome.

Figure 6 highlights features that are consistently ranked among the top predictors across multiple outcomes. Several fitness- and psychosocial-related variables, such as motivation, self-monitoring, fitness test scores, and BMI-for-age z-scores, appear repeatedly. This suggests that, beyond baseline activity levels, both physical capacity and motivational factors play a meaningful supporting role in physical activity change.

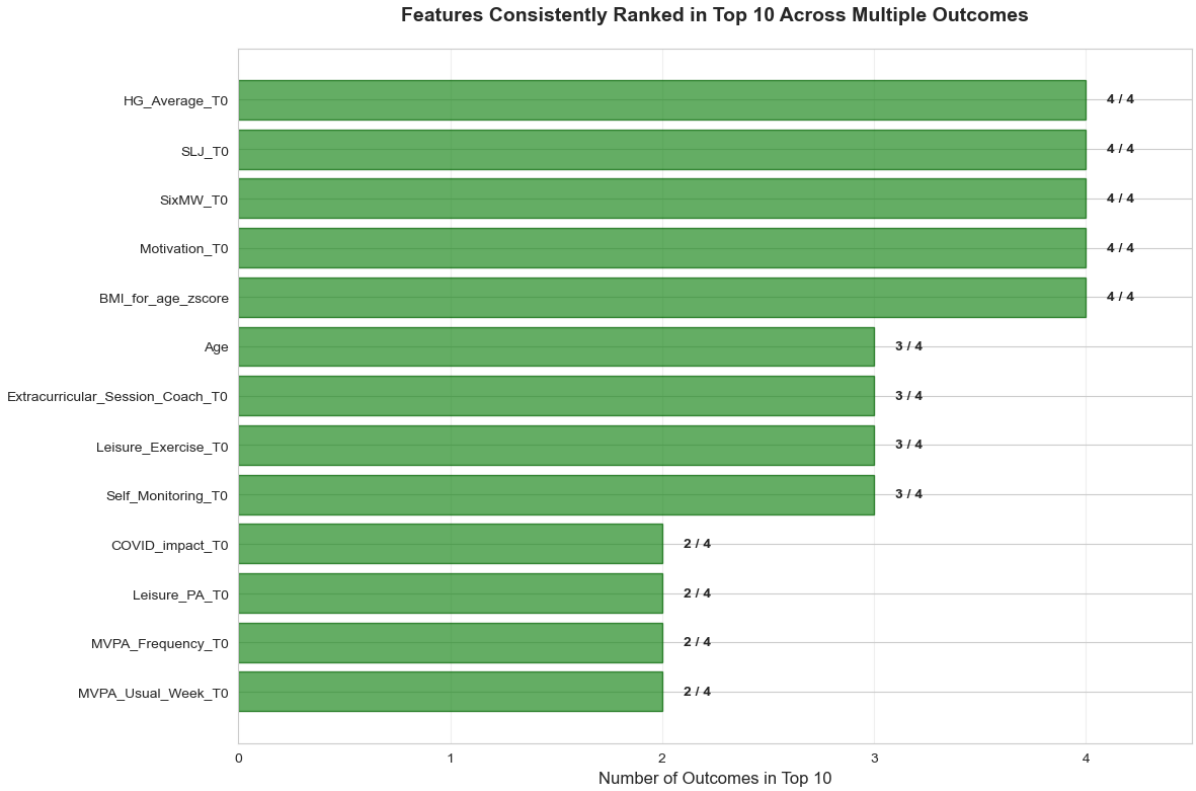

Figure 6: Features consistently ranked among the top predictors across multiple outcomes. Bars indicate how often a feature appears in the top 10 across outcomes.

To provide a higher-level overview, features were grouped into conceptually meaningful categories (Figure 7). The categories were defined as follows:

- **Demographic/Anthropometric**: age, sex, BMI-related measures
- **Baseline Physical Activity**: MVPA, leisure physical activity, leisure exercise, sedentary behavior
- **Fitness/Motor Competence**: six-minute walk test, standing long jump, handgrip strength
- **Psychosocial**: motivation and self-monitoring
- **Contextual/Environmental**: PE hours, extracurricular sessions, COVID-related impact

Baseline physical activity shows the highest average importance, followed by psychosocial and fitness-related categories. Demographic and contextual variables contribute less overall, suggesting that behavioral and motivational factors are more informative for predicting individual changes than static background characteristics.

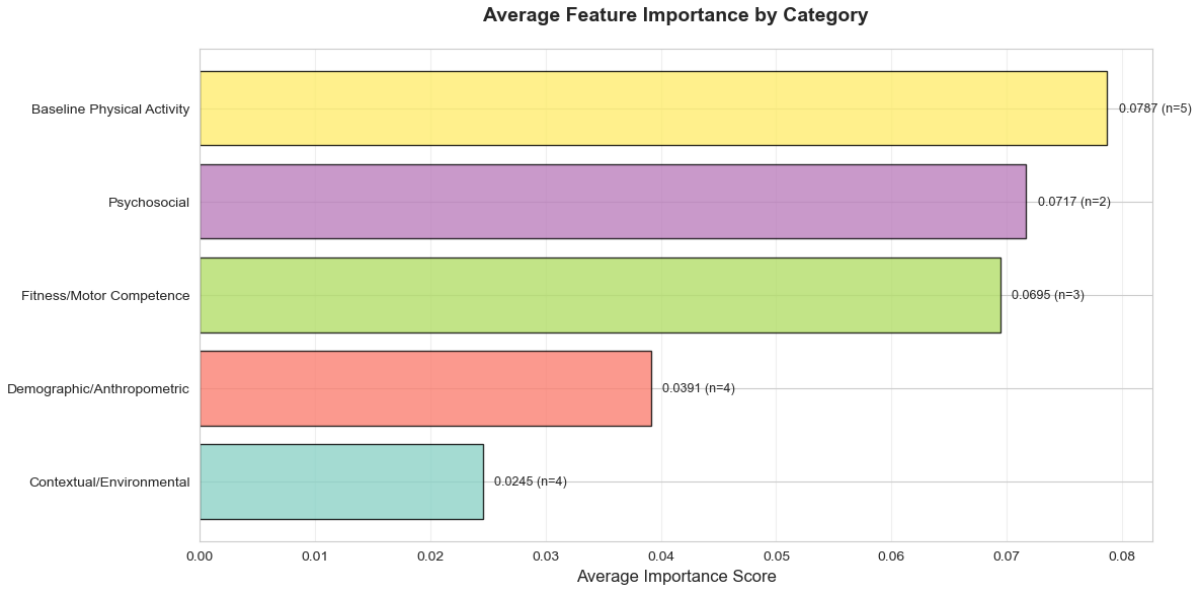

Figure 7: Average feature importance grouped by category, based on Random Forest models. Categories are defined according to Self-Determination Theory and the intervention design.

4.3 Evaluation and Discussion

The feature importance results show clear and understandable patterns that relate directly to the project goal. The main objective was to predict individual responses to the intervention and to identify which baseline characteristics are linked to positive changes in physical activity. The results show that this is possible to a meaningful extent, even though prediction remains limited.

Baseline physical activity measures are the strongest predictors across all outcomes. This is expected, because starting behavior strongly influences how much change is possible. Adolescents who are already very active have less room to improve, while those with lower baseline activity levels show greater potential for change. This supports the decision to model change scores rather than absolute activity levels.

In addition to baseline behavior, psychosocial variables such as motivation and self-monitoring consistently appear among the relevant predictors. This fits well with Self-Determination Theory, which highlights the role of motivation and self-regulation in behavior change. The repeated importance of these variables suggests that motivated adolescents may engage more effectively with the intervention, even if motivation alone does not fully explain the outcomes.

Physical fitness and motor performance measures also show stable importance. This indicates that physical capacity may support increases in physical activity, especially for more demanding activities. In contrast, demographic and contextual variables such as age, sex, and PE hours show low importance, suggesting that they contribute little to explaining individual changes in this dataset.

The feature importance results should be interpreted as descriptive and not causal. Random Forest importance shows how useful a variable is for prediction, not whether it directly causes change. However, the observed patterns help explain why regularized linear models such as Ridge regression performed well. These models capture strong baseline effects while reducing the influence of less informative variables.

Overall, the results indicate that changes in physical activity are mainly driven by baseline behavior, with additional contributions from motivation, self-regulation, and physical fitness. At the same time, the moderate prediction performance highlights that physical activity change is complex and influenced by factors that are not fully captured by baseline data.

5 Lessons Learned

5.1 What worked well

Combining machine learning methods with domain knowledge from physical activity research worked well in this project. Regularized regression models provided a good balance between interpretability and predictive performance and were well suited to the structure of the data.

The modular preprocessing and modeling pipeline allowed iterative development and ensured reproducibility across different outcomes. Using cross-validation before hyperparameter tuning helped identify suitable models early and avoided focusing on poorly performing approaches.

Overall, it was possible to select, train, and tune models that were able to predict changes in physical activity reasonably well, despite the complexity of the dataset.

5.2 Limitations

A major limitation was the amount of missing data. Due to the data structure and missing values, more than half of the original observations had to be excluded during preprocessing. This great reduction in sample size likely limited model performance and generalizability.

In addition, the dataset was high-dimensional and conceptually complex, with many variables from different domains. Considerable time was required to clean, reduce, and organize the data before modeling, which limited the time available for further model experimentation.

Finally, physical activity outcomes were based on self-reported measures, which are prone to measurement error and reporting bias. This likely introduced additional noise and reduced the maximum achievable prediction accuracy.

5.3 Future improvements

Future work could explore data augmentation or imputation strategies to reduce information loss caused by missing values. Preserving a larger portion of the dataset may improve model stability and predictive performance.

Further improvements could also include the use of objective physical activity measurements, such as data from wearable devices, to reduce measurement error and improve outcome reliability.