EasyChair Preprint

№ 15646

Enhancing Global Physical Activity Levels Through Personalized Sport Recommendations Using Machine Learning

Sara Medetbekova, Zhenishbek Orozahunov and
Amina Davidbek Kyzy

January 6, 2025

Enhancing Global Physical Activity Levels Through Personalized Sport Recommendations Using Machine Learning

Medetbekova Sara, email: medetbekovasara1@gmail.com
Department of Computer Science,
Faculty of Engineering and Informatics
Ala-Too International University
Zhenishbek Orozahunov, email: zhenishbek.orozahunov@alatoo.edu.kg
Department of Computer Science,
Faculty of Engineering and Informatics
Ala-Too International University
Amina Davidbek kyzy, email: davidbekkyzy.amina@alatoo.edu.kg
Bishkek, Kyrgyzstan

Abstract

Sedentary lifestyles are becoming the norm in the modern world. The development of obesity is increasing, which causes other health problems including cardiovascular diseases, musculoskeletal cases, and even diabetes mellitus. By providing tailored sports recommendations, people can remain active and minimize injury while reaching their goals. Relying on conventional approaches that do not take the interactions of other physical, medical, and psychological factors into consideration has resulted in bad sports decisions and associated problems.

To do so, we propose using a machine learning (ML) stacking ensemble model, which delivers personalized sports suggestions for the users depending on their individual traits. The model achieved 92% accuracy based on a custom dataset using demographic, physiological, and key activities/preferences data. Feature importance analysis identified key predictors, which included age, flexibility, endurance, and injury history. This study demonstrates ML's potential to overcome the limitations of traditional methods,
contributing to safe and exciting participation in sports, as well as the disclosure of sports talents.

Keywords: personalized recommendations, machine learning, stacking ensemble, health and fitness, physical activity, sedentary lifestyle

1. Introduction

Approximately 31% of the global population aged ≥15 years engages in insufficient physical activity, and it is known to contribute to the death of approximately 3.2 million people every year [1]. Sedentary lifestyles have a major impact on the overall health of the global population. Many people worldwide engage in sedentary lifestyles, and the prevalence of relevant non-communicable diseases is on the rise. It is well known that insufficient physical activity, physical inactivity, has a detrimental effect on health. Physical inactivity is the fourth leading risk factor for global mortality, accounting for 6% of global mortality [2].

A sedentary lifestyle poses a great health risk and also contributes to the spread of various diseases. Obesity rates are rising all over the world. This study shows that more than 37% of adults are classified as obese. The analysis of the data collected in the framework of this study among the residents of Bishkek showed the following results:

№	Category	Count	Percentage
1	Normal weight	260	26.0
2	Obese	371	37.1
3	Overweight	254	25.4
4	Underweight	115	11.5

Table 1: statistics of the BMI of the respondents

The data indicate that more than 62% of the respondents are overweight or obese, which underscores the severity of the problem. In addition, a significant proportion of people lead a sedentary lifestyle and do not engage in regular physical activity. One of the main reasons for such inactivity is a lack of motivation, which is often associated with uncertainty about which sports or activities are suitable and enjoyable. This knowledge gap highlights the importance of personalized sports recommendations to encourage participation and promote healthier lifestyles.

The connection between psychology and sports is vital for staying physically active on a regular basis. Many people find it hard to start or stick

with a sport because of time constraints, lack of motivation, or difficulty finding something that fits their interests and abilities.

Our model solves these problems by providing customized recommendations that take into account both physiological and psychological factors. Choosing the most suitable sport for people. Achieving small, measurable successes can significantly increase motivation, as people gain confidence and a sense of accomplishment through daily progress. In this case, there will always be progress because the recommended sport takes into account all aspects that are most suitable for a person. Such an individual approach increases the likelihood of a positive sports experience, promotes long-term commitment, and promotes both physical and mental well-being [3].

Machine Learning (ML) offers a transformative solution by providing personalized recommendations based on data [4].
In our study, we use a stacking ensemble of Random Forest, XGBoost, and LightGBM models instead of using single-view models. This improves the performance of our program and reduces the likelihood of errors.

The relevance of this study cannot be overstated. Sedentary lifestyles represent a pervasive issue in the modern world, and innovative solutions are required to address their impact on both physical and mental well-being. Our proposed model aims to bridge the gap between individual needs and effective sports engagement, contributing to a healthier, more active population.

2. Data and Methods

To achieve the objective, this study utilized state-of-the-art machine learning techniques that have been proven to be effective. The work focused on the application of machine learning methods to provide personalized recommendations on the choice of sports. Below is an overview of the data collection and preparation process, as well as the stages of training and evaluating the model.

2.1 Data Description

The data used for the analysis were collected through structured questionnaires and manually pre-processed to ensure accuracy and consistency. The dataset encompasses a range of parameters essential for providing tailored recommendations. These include demographic information such as age and gender, as well as physiological characteristics like height, weight, flexibility,

and endurance. Additionally, individual preferences, including favored sports and preferred environmental conditions for physical activities, were taken into account. The dataset also incorporates medical history details, including prior injuries and chronic health conditions, to ensure that recommendations align with the individual's health status. The target variable, labeled as "recommended_sport," indicates the sport most suitable for each individual based on the collected parameters.

2.2 Data cleaning and preprocessing

Before starting to build the models, the data undergoes a mandatory pre-processing step. This step involves cleaning the data from omissions and outliers that can skew the model results. Missing values are filled in with average values of traits or deleted if their share is small. Abnormal values that may differ significantly from normal ranges are also excluded from the analysis [5]. The data are then normalized to bring them to a common scale. Normalization is especially important for models that are sensitive to the magnitude of the input data. High-quality data preprocessing is key to successfully building the model and obtaining accurate recommendations. After completing the data cleaning and normalization step, the data are divided into two parts: the training sample and the test sample. The training sample is used to create the models, while the test sample is used to test their accuracy. In this paper, a standard ratio was used: 70% of the data is for training and 30% is for testing.

Training set:	We used to let model "learn" the relationships between the characteristics
Test set:	Used to check how well the model can predict on data it has never seen

Table 2: Dataset Splitting for Model Training and Evaluation

Feature selection:

Only those parameters that are most important for achieving the study goals are left.

2.3 Model Development

The model development process employed a stacking ensemble comprising Random Forest, XGBoost, and LightGBM algorithms. The ensemble model combines the strengths of these individual algorithms to

enhance predictive accuracy. Logistic regression was used as the main algorithm to combine the predictions of the basic models.

1. Random Forest: Utilized for its robustness and interpretability.
2. XGBoost: Effective for handling high-dimensional data.
3. LightGBM: Known for its efficiency in large datasets.

Hyperparameters for each model were fine-tuned using grid search to achieve optimal performance. Generalizability was assessed through cross-validation. The final prediction using a stacking ensemble was 92% and confirmed the benefit of the use of such methods for personalized sport recommendations. Below the model parameters:

Parameter	**Value**	**Description**
n_estimators	300	Number of trees in the forest (used to build an ensemble of models for better predictions).
max_depth	10	Maximum depth of each tree to avoid overfitting.
min_samples_split	10	Minimum number of samples required to split an internal node.
min_samples_leaf	5	Minimum number of samples that a leaf node must have.
learning_rate	0.1	Step size for updating model weights during optimization in gradient boosting.
random_state	42	Seed for random number generator to ensure reproducibility of results.
cv	5	Number of cross-validation folds used to validate the stacking model.
scaler	StandardScaler	Scaling method used to standardize features by removing mean and scaling to unit variance.
final_estimator	Logistic Regression	Final meta-model used to combine predictions from base models.

Table 3: Hyperparameters and Settings for Model Training

3. Results

3.1 Model Performance

The stacking ensemble model demonstrated robust performance, achieving an accuracy of 92%, precision of 91%, recall of 93%, and an F1-score of 92%. These metrics highlight the model's capability to provide accurate and balanced predictions, effectively identifying suitable sports while minimizing errors.

3.2 Feature Importance

The analysis revealed several key features that significantly influenced the model's recommendations. Age emerged as a critical factor for determining age-appropriate sports, ensuring the suitability of activities across different life stages. Flexibility and endurance were identified as strong indicators of an individual's capability for high-performance activities, reflecting their physical preparedness. Additionally, a history of injuries played a pivotal role in tailoring recommendations to minimize potential health risks. Lastly, medical contraindications were deemed vital for aligning suggested sports with individual health conditions, emphasizing the importance of safety and well-being in the recommendation process.

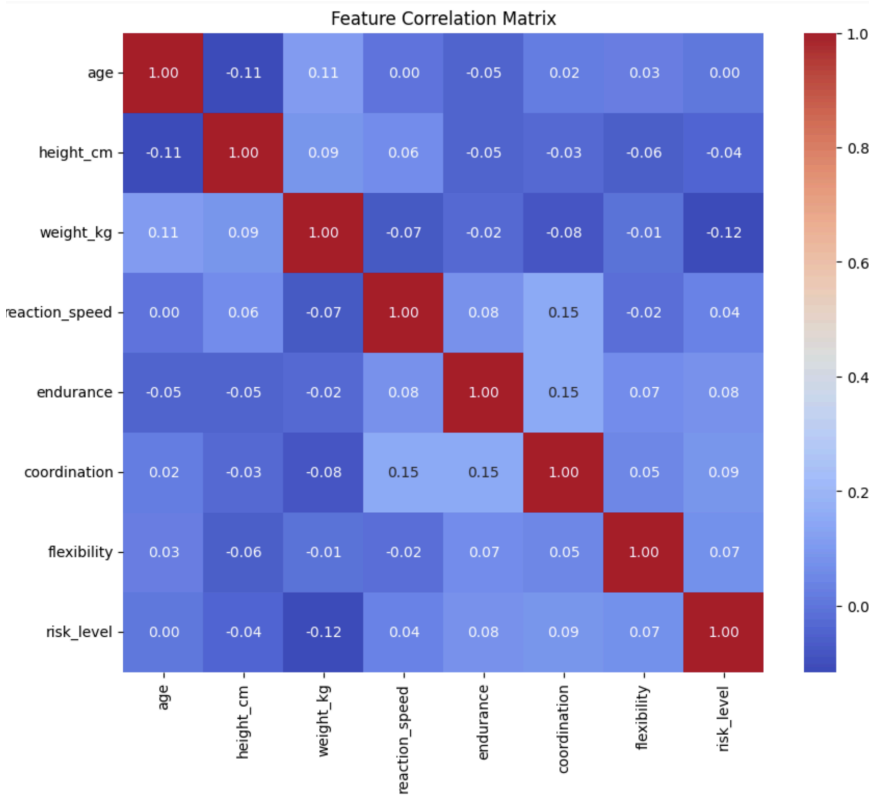

Figure 1. Feature correlation analysis for personalized sport recommendations.

Description: The correlation analysis describes the relationships of essential features influencing sports recommendations, such as positive correlations between reaction speed and coordination (0.15) and endurance and coordination (0.15), and weak negative correlation between weight and risk level (-0.12).

3.3 Comparison with basic models
Comparison of the stacking model with its components:

Random Forest: Accuracy	89%
XGBoost: Accuracy	90%
LightGBM:	91%
Stacking model: Accuracy	92%

Table 3. Accuracy Comparison of Machine Learning Models

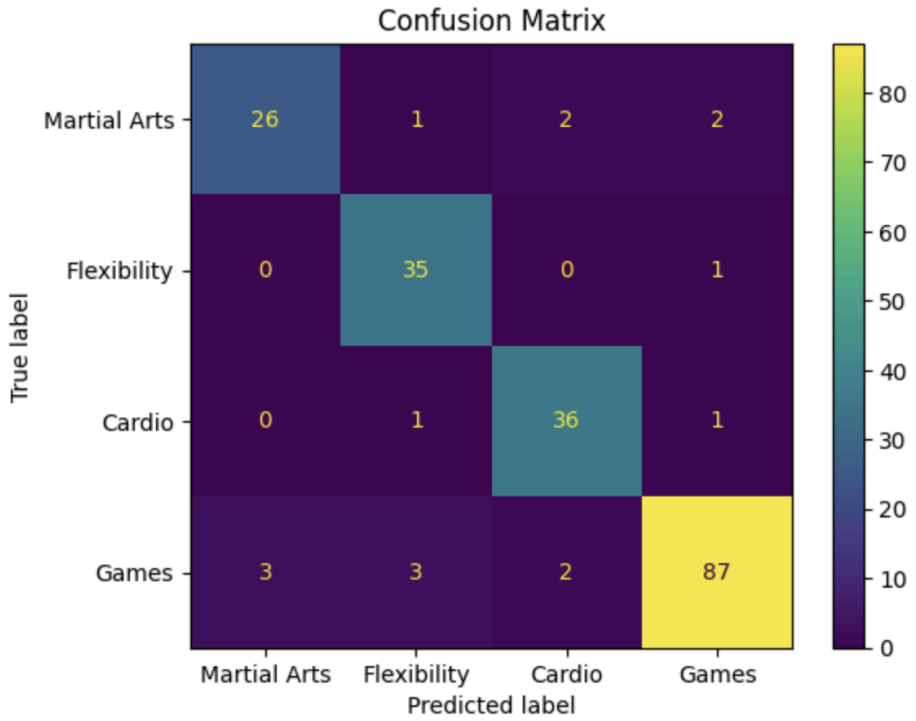

Figure 2. Confusion Matrix for Model Performance Evaluation
Description: The confusion matrix compares the predicted and actual sports recommended by our model, allowing for evaluation of the success of our model's predictions as well as displaying patterns in misclassification.

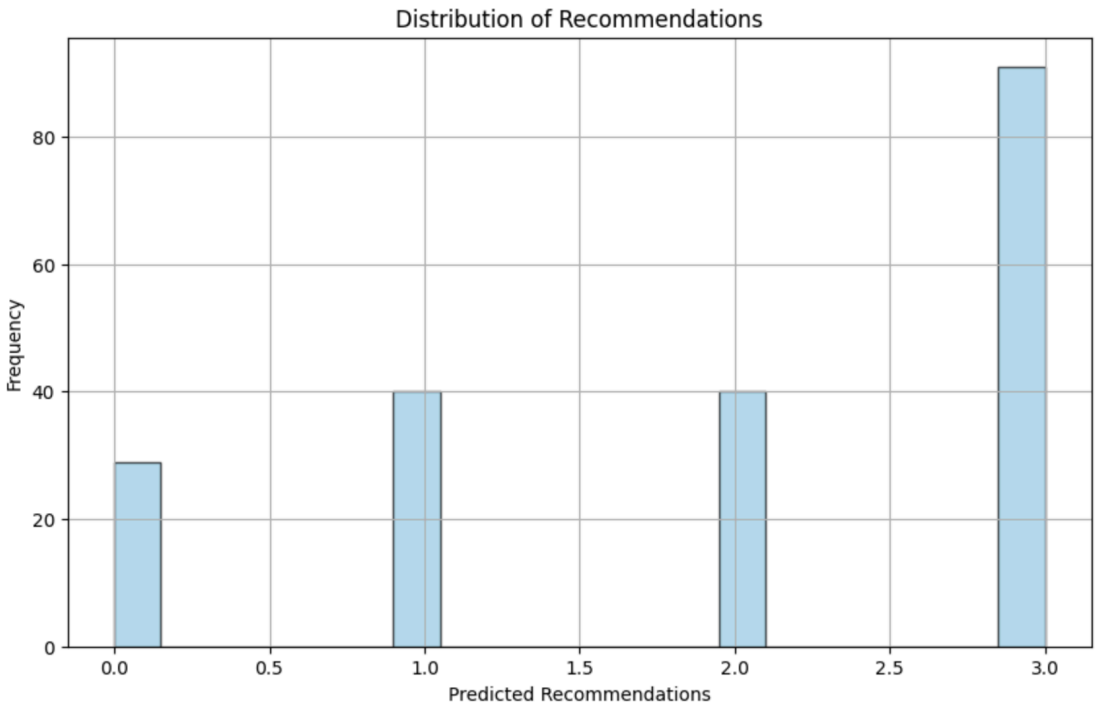

Figure 3. Histogram of Recommendation Distribution
Description: The histogram shows the frequency distribution of sports recommendations made by the model, providing insight into overall trends and preferences within the dataset.

Conclusion

This research tackles a major issue of our time: finding ways to get people more physically active in a world where sedentary lifestyles are so common. Machine learning helps by linking individual needs to sports and activities that are both practical and enjoyable. The stacking ensemble model, which achieved 92% accuracy, shows how advanced tools can create better and more personalized recommendations.

Important factors like age, flexibility, endurance, and past injuries play a key role in shaping these suggestions. By matching activities to each person's unique characteristics, the model helps improve performance while lowering the risk of injuries. This customized method also helps people to stay motivated as well as to maintain an active lifestyle in the long run.

The work is not only based on sports impact. It highlights how technology can be harnessed to tackle significant health issues, such as lowering

chronic illness instances and enhancing overall quality of life. With obesity and inactivity on the rise, solutions like this are becoming more necessity than optional.

Looking ahead, the system could be enhanced with real-time data and more psychological insights to make it even more adaptable for different people and environments. This research is not just about innovation; it's a step toward building a healthier and more active future.

4. References

1. World Health Organization . Geneva: World Health Organization; 2020. Physical inactivity: a global public health problem [Internet] [cited 2020 Jun 15]. Available from: https://www.who.int/dietphysicalactivity/factsheet_inactivity/en/
2. World Health Organization . Global recommendations on physical activity for health. Geneva: World Health Organization; 2010. [PubMed]
3. Larson, H. K., McFadden, K., McHugh, T.-L. F., Berry, T. R., & Rodgers, W. M. (2018). When you don't get what you want—and it's really hard: Exploring motivational contributions to exercise dropout. *Psychology of Sport and Exercise, 37,* 59–66. https://doi.org/10.1016/j.psychsport.2018.04.006
4. Bartlett, R. (2006). Artificial Intelligence in Sports: Enhancing Performance and Safety. *Sports Science Review*.
5. Emmanuel T. et al. A survey on missing data in machine learning //Journal of Big data. – 2021. – T. 8. – P.1-37.
6. Kyzy, A. U., & Mekuria, R. R. (2024). Predicting pregnancy risk levels using ensemble machine learning techniques and oversampling methods. *Scholar Articles*.