A Clustering Approach for Personalized Coaching Applications

Annika Van Buren[1], Audrey Kwan[1], Harald. H. Rietdijk[2](✉),
Talko B. Dijkhuis[2], Patricia Conde-Cespedes[3], Hilbrand Oldenhuis[2],
and Maria Trocan[3]

[1] Stanford University, Stanford, CA 94305, USA
[2] Hanze University of Applied Sciences, 9747AS Groningen, Netherlands
h.h.rietdijk@pl.hanze.nl
[3] ISEP, 92130 Issy-les-Moulineaux, France

Abstract. Insufficient physical activity presents a significant hazard to overall health, with sedentary lifestyles linked to a variety of health issues. Monitoring physical activity levels allows the recognition of patterns of sedentary behavior and the provision of coaching to meet the recommended physical activity standards. In this paper, we aim to address the problem of reducing the time consuming process of fitting classifiers when generating personalized models for a coaching application. The proposed approach consists of evaluating the effects of clustering participants based on their walking patterns and then recommending a unique model for each group. Each model consists of a random forest classifier with a different number of estimators each. The resulting approach reduces the fitting time considerably while keeping nearly the same classification performance as personalized models.

Keywords: Personalized coaching · sedentary lifestyle · fitting time optimization · clustering · Random Forests · estimators · variability

1 Introduction

Insufficient physical activity poses a significant risk to health and well-being. Sedentary lifestyles are associated with a variety of health problems, such as heart disease, cancer, stroke, and diabetes [1–3]. Through the use of wearable devices, such as smartwatches, it has become easier to gain insight into actual physical activity levels. Monitoring physical activity levels allows individuals to identify patterns of sedentary behavior and assess whether they meet the recommended activity guidelines. Furthermore, by tracking activity levels, individuals would be able to take proactive steps to mitigate potential health risks and make the necessary adjustments to incorporate more movement into their daily routines [4,5]. In addition to its physical benefits, monitoring physical activity also supports mental health and cognitive function. Regular exercise has been shown to alleviate symptoms of depression and anxiety, improve mood, and improve cognitive abilities, such as memory and concentration [6–8].

N.-T. Nguyen et al. (Eds.): ICCCI 2024, CCIS 2166, pp. 351–363, 2024.
https://doi.org/10.1007/978-3-031-70259-4_27

In recent years, there has been a large number of innovations in the field of personalization and individualization in healthcare and coaching applications. In [9] an overview of physical activity coaching applications can be found, and in [10] a review of behavior-changing therapy and rehabilitation applications is presented. A more specific example of an application that uses wearable devices to monitor physical activity can be found in the works of Dijkhuis *et al.*[11,12]. The authors propose the use of a machine learning-based procedure to train a digital activity coach that can provide information on daily fitness of a person. Specifically, the digital coach will monitor the probabilities throughout the day of the likelihood that the user will reach their goal of physical activity.

The paper [11] evaluates the performance of eight different machine learning algorithms and finds that tree algorithms and tree-based ensemble algorithms are the most promising for training a digital activity coach with personalized models. The results presented in the paper show a significant improvement in performance when one model is fitted per participant instead of using generalized models for the entire group (the approach also suggested in [13]).

Further improvements are expected to be obtained by generating more models that are fitted to specific sections of the data. However, as noted by the authors, *"it is problematic to provide a generalized recommendation for specific algorithms, parameters, or parameter settings"*. The problem could be solved by *"investigating the underlying mechanisms to be able to choose the best algorithm beforehand"*. However, fitting one model per person is too resource consuming. That is why in this paper we approach the problem differently. First, we perform clustering on the participants based on the hypothesis that people inside a cluster share similar patterns, and we fit only one model per cluster. When these insights are obtained, algorithm selection and model fitting can be optimized. The remainder of this paper is organized as follows; first, in Sect. 2, we present an overview of the relevant results and data used in [11], and explain the result we want to achieve. In Sect. 3 we present an overview of our approach which consists of two main parts, clustering and classification. In Sect. 4 we detail the clustering procedure. In Sect. 5 the classification results are presented. Finally, in Sect. 6 a discussion of the results is presented.

2 Preliminary Results

In the original study by Dijkhuis *et al.* [11] the step data of 48 participants were collected over a period of 33 weeks, using Fitbit activity trackers. These participants were involved in a health program conducted at the Hanze University of Applied Sciences Groningen. Data from 43 of these participants could be used for the study. These data were used to evaluate the performance achieved when predicting the probability that the participant will reach the step goal at 6 p.m. each day. To compensate for divergent behavior outside working hours, only step data collected between 7 a.m. and 6 p.m. on workdays was considered. Furthermore, to be able to use the data as input for regular machine learning methods, it was necessary to add the cumulative step count. In this way, instead of having to treat each data point as part of a time series, it was possible to use the data as input for a classification process. The resulting features that were used to train the different models were weekday, hour, step count per hour, and cumulative step count up to the given hour.

The open source Scikit-Learn library was used for the study. This package offers the choice of 41 classifiers. Using the flow chart provided by Scikit-Learn [14] and a cheat sheet found on the Microsoft Azure Machine Learning platform [15], a selection of eight classifiers was made; AdaBoost (ADA), Decision Trees (DT), KNeighborsClassifier (KNN), Logistic Regression (LR), Neural Networking(NN), Stochastic Gradient Descent (SGD), Random Forest (RF) and Support Vector Classification (SVC). These classifiers were first used to evaluate the performance of a generalized model across the entire data set, and subsequently a model was generated for each individual participant. In this last step, a selected set of hyperparameters for each classifier were 'fitted' to obtain the best performing model per participant. The comparison of results showed that the performance levels were significantly higher when personalized models were used. The maximum accuracy for the generalized model ranged from 71% to 78%, but reached 94% when using personalized models (considered data split is 70% for training and 30% for validation).

As stated in the discussion of the original study, and can also be seen, for example, in the work of Sarker [16], making general recommendations for classifier selection and hyperparameter settings is problematic. Using the Scikit-Learn grid search functionality provides a tool to determine optimal parameter settings per classifier. However, when this approach is used for each personalized model, the fitting times can become very large, depending on the number of parameters and their range of possible values. Therefore, the objective of this study is to investigate whether it is possible to improve the fitting times while maintaining the performance levels of the personalized approach in the original study. We use the source code of the original research as a baseline for this analysis, as well as the data set that contains the steps per hour of each participant. In the proposed solution, we seek to streamline and improve the training process by applying clustering of the participants based on the results obtained by analyzing the walking patterns of the participants and examining the hyperparameter behavior of each machine learning algorithm used. This analysis should lead to a recommendation of selected parameter settings per cluster, which will reduce the fitting time.

3 Description of the Proposed Solution

In Fig. 1 a schematic overview is given of the steps that were taken in the process to elaborate the proposed solution. Our approach is separated in two parts : clustering (first four steps in diagram) and classification (last step).

The clustering procedure was decomposed in the following steps :

1. The first step was to perform a thorough analysis of the preliminary results. The goal of this analysis was to limit the scope of the experimental setup to focus on one single classifier.

2. The second step was to formulate a working hypothesis and define the relevant metrics needed to perform the clustering process.
3. In step three, the available data was analyzed using the metric defined in step two. The results of this analysis form the input for step four, in which the clusters were defined.
4. Step four consisted of applying the clustering to the available data.

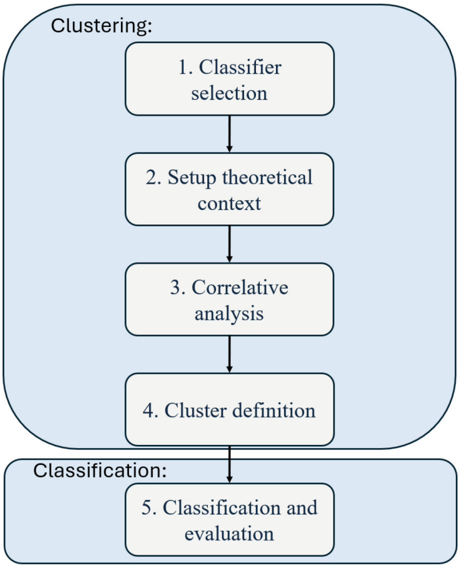


Fig. 1. Clustering steps.

Finally, the last step consists of performing a specific classification method for each cluster obtained from the previous steps. All the details are described in Sect. 5.

4 Clustering Procedure

In this section we describe all the steps that allowed us to partition the data into clusters, that is, steps 1 to 4 of the diagram in Fig. 1.

4.1 Experimental Setup

Our initial step in this research project was to evaluate the performance of various machine learning algorithms to identify the candidate most suitable for optimization. We considered the algorithm performance metrics F1-score, accuracy, and the operational performance metric given by the runtime of the fitting

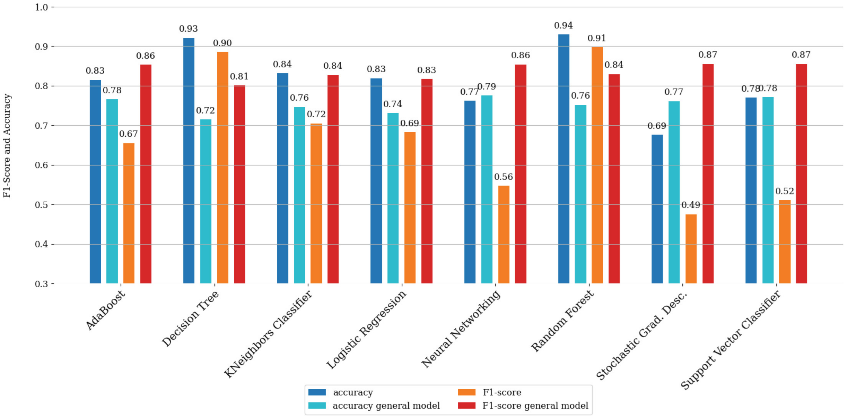


Fig. 2. Average F1-scores and accuracy per classifier.

procedure, to fully evaluate each algorithm. The results of the algorithm performance metrics are shown in Fig. 2. The operational metrics are shown in Fig. 3.

The results of the performance metrics show that the Random Forest and Decision Tree algorithms have the highest accuracy and F1-score for the models generated per participant. The proposed solution to cluster participants for model generation should give a better result than the generalized approach. Both classifiers also show a significant improvement in performance when the results of personalized modeling are compared with general modeling, making them suitable candidates for the experiment.

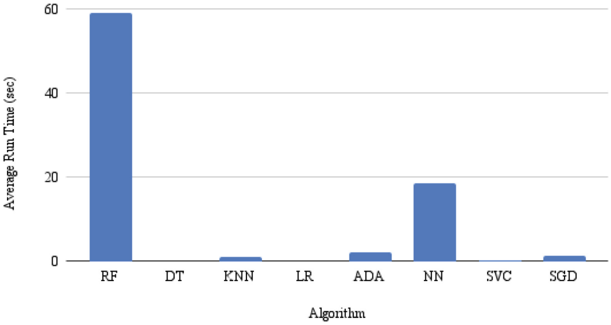


Fig. 3. Average run time per model of the fitting procedure for each classifier.

For further selection, we next considered the operational performance of the algorithms. Although the RF algorithm had the highest accuracy and F1-score for most of the participants, it also demonstrated significant computational demands, reflected in its extended runtime. On the other hand, the fitting times for the DT algorithm were already minimal. Since our objective was to enhance the model generation efficiency without compromising the algorithm's predictive capability, taking these results on performance and computational efficiency into consideration led us to select Random Forest for further optimization.

Since the fitting time depends on the number of values supplied for each hyperparameter, the generation efficiency can be further improved by presetting one or several of these values. Taking this into account, the clustering should focus on finding criteria that group participants with identical values for the hyperparameter setting. For the Random Forest algorithm, the parameters used in the original fitting process are the number of trees in the forest (`n_estimators`), the number of features to consider when looking for the best split (`max_features`), and the function to measure the quality of the split (`criterion`). The RF classifier has more parameters [17], but these three were considered to be the most likely to have a significant impact on the performance of the resulting model in the context of the original study.

4.2 Theoretical Context Setup

The preliminary clustering approach came from the hypothesis that the greater variability in the distribution of the features of the data set may correlate with higher optimal hyperparameter values for the Random Forest Algorithm, especially `n_estimators`. The logic underpinning this hypothesis is that a higher day-to-day step count variability might necessitate more estimators to accurately capture fluctuations, thereby reducing model variance and enhancing robustness. Tailoring an `n_estimators` value for each cluster would allow for a model that adapts to the characteristics of each group, with the aim of reducing over-fitting in low-variability clusters and enhancing robustness in high-variability clusters. To be able to focus on the evaluation of the impact of the characteristics of the data set on the algorithm performance, this approach was taken instead of using more traditional clustering approaches.

The features of the original data set consist of weekday, hour, step count per hour, and cumulative step count. Since weekday and hour should not show any relevant differences per participant, we have to look at the other two features. To investigate the relationship between participant step count variability and optimal `n_estimators` value, variability was quantified using the following methodology.

- *Data aggregation by interval:* For each participant, steps were aggregated to calculate the average sum steps per hour. The average cumulative sum steps up to each hour was also calculated. Data for the step count for each participant were taken from the training set used in the fitting process.
- *Standard Deviation (SD):* The standard deviation of the sum steps per hour was calculated. This metric quantifies the degree of variation from the average step count within each hour, providing insight into the consistency of the level of activity of the participants. The standard deviation of the cumulative steps for each hour was also calculated.

- *Coefficient of Variation (CV):* To facilitate a normalized comparison of variability that accounts for differences in average activity levels among participants, the coefficient of variation is calculated. This value is calculated by dividing the standard deviation of the aggregated sum by the average step count for the respective interval. The CV provides a relative measure of variability, allowing for an understanding of fluctuation in activity levels in proportion to the mean steps.
- *Hyperparameter values:* The values of `n_estimators` of the personalized models generated in the original study were used. These models were fitted with possible values 10, 50, 100, or 500 for this parameter.

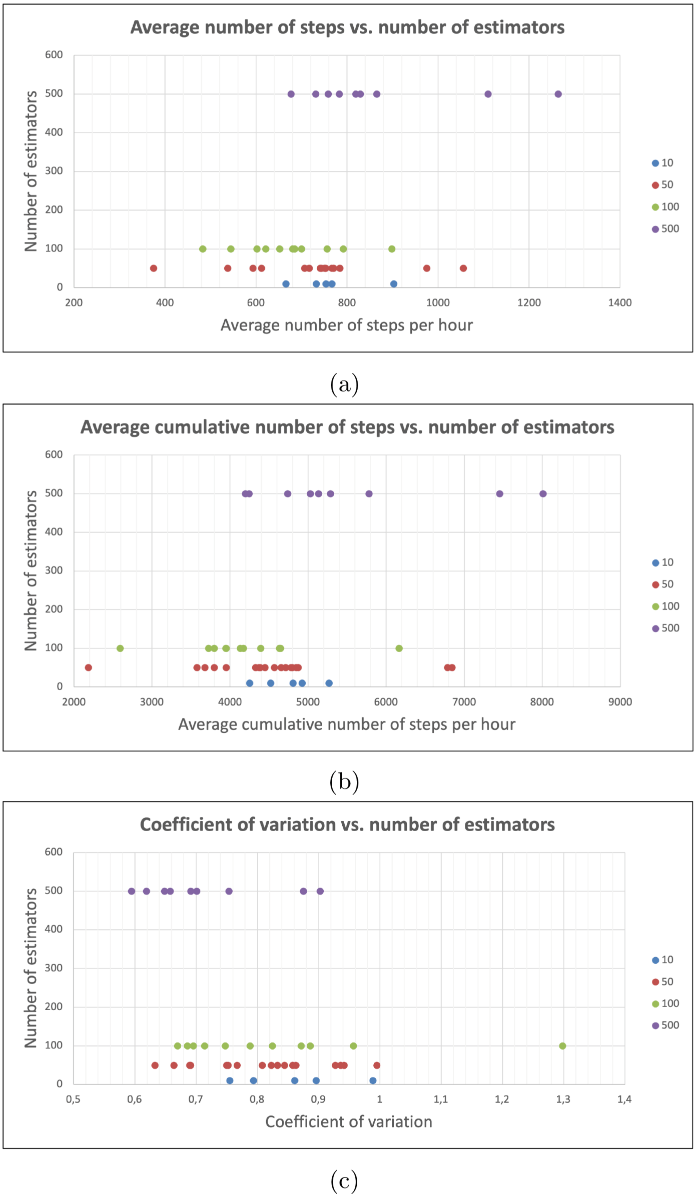
In Fig. 4 the averages calculated over the aggregated values, cumulative step counts and the coefficient of variation are plotted against the corresponding `n_estimators` value. The results of the variability measures and averages for each participant were compared to the optimal `n_estimators` values designated for their respective RF models. The analysis revealed a consistent trend between the variability indicators, highlighting a correlation in which the increase in the variability in step counts was correlated with a preference for higher optimal `n_estimators` values. The increase in the average step count was also correlated with a preference for higher optimal `n_estimators` values. Moreover, those with both higher average step counts and notable variability tend to require a higher number of trees to effectively model their data.

Optimal `n_estimators` values correlated most closely with the average step count per hour. Average cumulative sum steps per hour showed a weaker positive correlation with `n_estimators`. The correlation coefficient of the average cumulative sum steps per hour showed a weaker negative correlation with `n_estimators`. When compared to the correlation between average step count per hour and optimal `n_estimators` values, participants with an unexpectedly high `n_estimators` were often associated with a low correlation of average cumulative sum steps per hour or high average cumulative sum steps per hour.

4.3 Clustering Definition

On the basis of these observations, the following approach for the clustering process was defined. The first step in constructing a structured framework for selecting the value of the `n_estimators` parameter in Random Forest models is determined by the participant step count data. Therefore, the initial step in the selection process is partitioning based on `Average Sum Steps`. Then adjustments are made according to `Average Cumulative Sum Steps per Hour` and `CV Average Cumulative Sum Steps per Hour`.

In the analysis of the relationship between similar ranges in participant average steps per hour and the value of `n_estimators` three clusters were determined, each with a specified value of `n_estimators` of 50, 100, or 500. This classification is shown in Table 1.



(a)

(b)

(c)

Fig. 4. Variability Metrics by `n_estimators`. (**a**) Average sum steps.(**b**) Average cumulative sum steps. (**c**) Hour coefficient of variation.

Further refinement of the setting of `n_estimators` is carried out based on the `Average Cumulative Sum Steps per Hour` and the `CV Average Cumulative Sum Steps per Hour`. Adjustments are only applied to the first two groups and considering the upper limit for the increment of `n_estimators`.

Table 1. Initial partitioning based on average sum steps.

`Average Sum Steps` value	`n_estimators` value
Up to 657.048	50
Between 657.048 and 808.093	100
Above 808.093	500

For the first cluster of participants with steps up to 657.048 that were initially assigned a value of 50 to `n_estimators`, an increase to 100 is applied if both of the following criteria are satisfied:

- `Average Cumulative Sum Steps per Hour` ≥ 4569.44 and
- `CV Average Cumulative Sum Steps per Hour` ≤ 0.788.

This adjustment is based on the assumption that increasing `n_estimators` is justified by the increase in activity levels, provided the variability remains within a reasonable range.

For the second cluster, with steps ranging from 657.048 to 808.093, the initial recommendation of 100 `n_estimators` is maintained unless activity and variability metrics meet the following criteria:

- `Average Cumulative Sum Steps per Hour` ≥ 5000 and
- `CV Average Cumulative Sum Steps per Hour` ≤ 0.829.

In this case an elevation to a value of 500 for `n_estimators` is applied. This is aimed at accommodating significantly elevated activity levels while ensuring that variability is sufficiently managed to optimize the model's performance.

4.4 Resulting Clusters

With the clustering framework in place, a data frame was created for each of the three clusters. Each data frame was composed by collecting training data from the original personalized fitting run of the participants in each specific cluster. Figure 5 shows the number of participants in each cluster. From left to right, there are 9 participants in the first cluster (`n_estimators` $= 50$), 23 in the second cluster (`n_estimators` $= 100$), and 11 in the third cluster (`n_estimators` $= 500$). Therefore, each group has at least 20% of the total number of participants in the experiment.

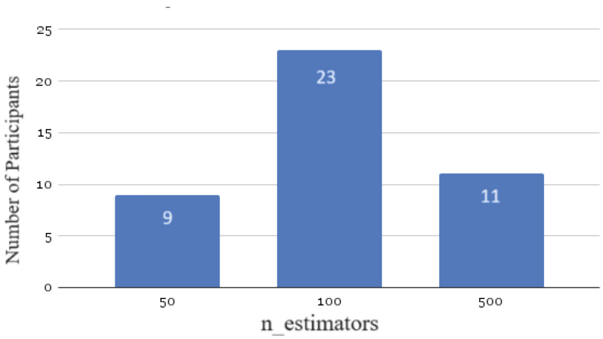


Fig. 5. The number of participants in each cluster.

5 Classification Results

The RF classifier fitting process was executed with a unique `n_estimators` value for each data frame, and several performance metrics (i.e., accuracy, F1-score, fitting time) were evaluated, comparing them to the original results of Dijkhuis *et al.* [11]. As in the original research, five-fold cross-validation was applied.

5.1 Performance Evaluation : Accuracy and F1 Score

Figure 6a and Fig. 6b show the total average accuracy and F1 score results of the original personalized run and the run using our clustering framework. Specifically, "with clustering" indicates the results obtained when one model was generated per clustered group of participants, using the specific `n_estimators` value for that cluster. "Without clustering" represents the original results obtained using a personalized model per participant. The results show that Total average accuracy was extremely similar between the clustering and no clustering trials. However, the clustering trial performed better than the no clustering trial in regards to total average F1-score. This indicates that on a global scale, the clustering method does not adversely affect the performance of the Algorithm.

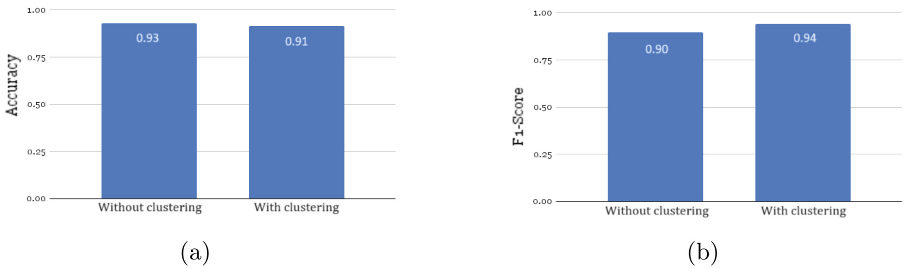
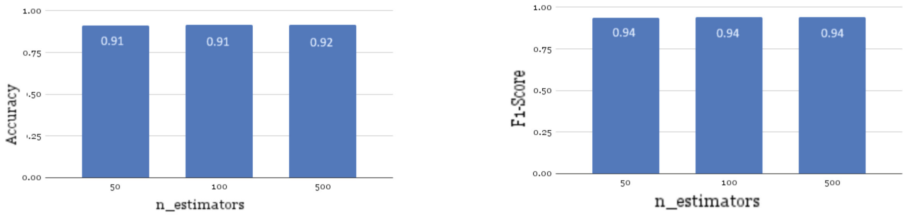


Fig. 6. Total average performance for all participants. (**a**) Accuracy. (**b**)F1-score.

Figure 7a and Fig. 7b illustrate the accuracy and F1-performance for each cluster. This view provides a more in-depth analysys of the Random Forest algorithm's performance within each cluster. Accuracy is around the same for all of the clusters, with the lowest accuracy being 0.912683 (`n_estimators` = 50) and the highest accuracy being 0.915141 (`n_estimators` = 500). Like accuracy, F1-scores were about the same for all of the clusters, with the lowest F1-score being 0.938219 (`n_estimators` = 50) and the highest F1-score being 0.941708 (`n_estimators` = 500).



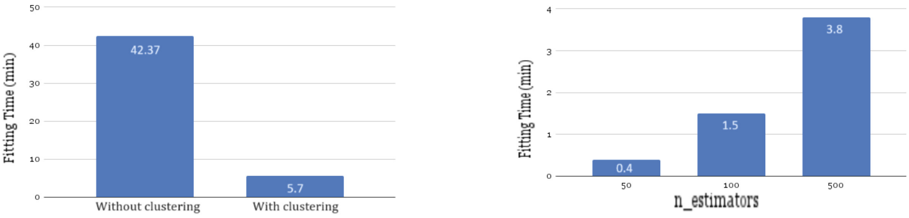
(a) (b)

Fig. 7. Per cluster average performance. (**a**) Accuracy. (**b**)F1-score.

5.2 Fitting Time

As expected clustering led to a substantial decrease in total fitting time, as can be seen in Fig. 8a. Without clustering, total fitting time was 42.37 min, while with clustering, total fitting time was 5.70 min. Thus, clustering led to an 86.55% decrease in total fitting time, which is a highly significant optimization that will greatly reduce overall computational time.

Looking at fitting time per cluster, presented in Fig. 8b, there is a positive correlation between the number of `n_estimators` and the amount of the fitting time. This is to be expected since the resulting model is more complex. Therefore, there is a trade-off between performance and computational cost that must be taken into account when determining the `n_estimators` value for each cluster.



(a) (b)

Fig. 8. Fitting times. (**a**) Totals. (**b**)Time per cluster.

6 Conslusions and Discussion

As shown in the results, the total accuracy and total F1-scores were about the same or better with than without clustering. This proves that clustering the data by walking variability does not lead to detriments in the performance of the RF algorithm, and instead leads to equally reliable predictions. Total fitting time, though, was significantly reduced by clustering. This is a particularly significant result, since RF was the most computationally expensive algorithm out of all the machine learning algorithms initially considered in [11]. Thus, decreasing fitting time while maintaining RF's high accuracy and F1-scores, through using preselected hyperparameters based on the characteristics of the dataset, would significantly improve the practicality of using machine learning algorithms in a digital physical activity coach.

Several considerations were taken into account to accommodate the size of the available data. For example, at all times the same random seed was used to ensure that the used selections were identical. Furthermore, the clustering train set was constructed using the train sets generated in the personalized run, instead of generating a new train set using all the data per cluster.

References

1. World Health Organization: Physical activity (2024). https://www.who.int/health-topics/physical-activity
2. Losina, E., Yang, H.Y., Deshpande, B.R., Katz, J.N., Collins, J.E.: Physical activity and unplanned illness-related work absenteeism: data from an employee wellness program. PloS one **12**(5), e0176872 (2017). https://doi.org/10.1371/JOURNAL.PONE.0176872
3. Lee, I.M., Shiroma, E.J., Lobelo, F., Puska, P., Blair, S.N., Katzmarzyk, P.T.: Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. Lancet **380**(9838), 219–229 (2012). https://doi.org/10.1016/S0140-6736(12)61031-9
4. Carter, D.D., Robinson, K., Forbes, J., Hayes, S.: Experiences of mobile health in promoting physical activity: a qualitative systematic review and meta-ethnography. PLoS ONE **13**(12), e0208759 (2018). https://doi.org/10.1371/JOURNAL.PONE.0208759
5. Chatterjee, A., Prinz, A., Gerdes, M., Martinez, S.: Digital interventions on healthy lifestyle management: systematic review. J Med Internet Res **23**(11), e26931 (2021). https://doi.org/10.2196/26931. http://www.ncbi.nlm.nih.gov/pubmed/34787575
6. Ekelund, U., et al.: Does physical activity attenuate, or even eliminate, the detrimental association of sitting time with mortality? A harmonised meta-analysis of data from more than 1 million men and women. Lancet **388**(10051), 1302–1310 (2016). https://doi.org/10.1016/S0140-6736(16)30370-1
7. Sharma, A., Madaan, V.: Exercise for mental health. Primary Care Companion J. Clin. Psychiatry **8**(2), 106–107 (2006). https://doi.org/10.4088/pcc.v08n0208a
8. Kamali, M.E., et al.: Virtual coaches for older adults' wellbeing: a systematic review. IEEE Access **8**, 101884–101902 (2020). https://doi.org/10.1109/ACCESS.2020.2996404

9. Gámez Díaz, R., Yu, Q., Ding, Y., Laamarti, F., El Saddik, A.: Digital twin coaching for physical activities: a survey. Sensors **20**(20), 5936 (2020). https://doi.org/10.3390/s20205936, https://www.mdpi.com/1424-8220/20/20/5936
10. Lauer-Schmaltz, M.W., Cash, P., Hansen, J.P., Maier, A.: Designing human digital twins for behaviour-changing therapy and rehabilitation: a systematic review. In: Proceedings of the Design Society, vol. 2, pp. 1303–1312. Cambridge University Press (2022). https://doi.org/10.1017/pds.2022.132
11. Dijkhuis, T.B., Blaauw, F.J., van Ittersum, M.W., Velthuijsen, H., Aiello, M.: Personalized physical activity coaching: a machine learning approach. Sensors (Switzerland) **18**(2), 623 (2018). https://doi.org/10.3390/s18020623
12. Blok, J., Dol, A., Dijkhuis, T.: Toward a generic personalized virtual coach for self-management: a proposal for an architecture. In: 9th International Conference on eHealth, Telemedicine, and Social Medicine 2017. Hanze University of Applied Sciences, Nice (2017)
13. Chatterjee, A., Pahari, N., Prinz, A., Riegler, M.: Machine learning and ontology in eCoaching for personalized activity level monitoring and recommendation generation. Scientific Reports **12**(1) (2022). https://doi.org/10.1038/s41598-022-24118-4
14. Scikit-Learn: Choosing the right estimator - scikit-learn 1.4.1 documentation. https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
15. Microsoft: Machine Learning Algorithm Cheat Sheet - designer - Azure Machine Learning — Microsoft Learn. https://learn.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet?view=azureml-api-1
16. Sarker, I.H.: Machine learning: algorithms, real-world applications and research directions. SN Comput. Sci. **2**(3) (2021). https://doi.org/10.1007/s42979-021-00592-x
17. Scikit-Learn: Scikit-Learn documentation on RandomForestClassifier. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html