Impact of Dataset Characteristics on Optimal Model Selection: A Comparative Analysis of Simulated and Real-World Data

Harald H. Rietdijk
Lectoraat Digital Transformation
Hanze University of Applied Sciences
Groningen, The Netherlands
h.h.rietdijk@pl.hanze.nl

Olayemi Shola Alabi
Isep, LISITE
28 rue Notre Dame des Champs,
Paris, France
shola-alabi.olayemi@eleve.isep.fr

Patricia Conde-Cespedes
Isep, LISITE
28 rue Notre Dame des Champs,
Paris, France
patricia.conde-cespedes@isep.fr

Talko B. Dijkhuis
Lectoraat Digital Transformation
Hanze University of Applied Sciences
Groningen, The Netherlands
t.b.dijkhuis@pl.hanze.nl

Hilbrand K.E. Oldenhuis
Lectoraat Digital Transformation
Hanze University of Applied Sciences
Groningen, The Netherlands
h.k.e.oldenhuis@pl.hanze.nl

Maria Trocan
Isep, LISITE
28 rue Notre Dame des Champs,
Paris, France
maria.trocan@isep.fr

***Abstract*—In the rapidly evolving field of Machine Learning , selecting the most appropriate model for a given dataset is crucial. Understanding the characteristics of a dataset can significantly influence the outcomes of predictive modeling efforts, making the study of the properties of the dataset an essential component of data science. This study investigates the possibilities of using simulated human data for personalized applications, specifically for testing clustering approaches. In particular, the study focuses on the relationship between dataset characteristics and the selection of the optimal classification model for clusters of datasets. The results of this study provide critical insights for researchers and practitioners in machine learning, emphasizing the importance of dataset characteristics and variability in building and selecting robust models for diverse data conditions. The use of human simulation data provide valuable insights but requires further refinement to capture the full variability of real-world conditions.**

***Index Terms*—machine learning, simulated data, data characteristics, model selection**

I. Introduction

In the rapidly evolving field of Machine Learning (ML), selecting the most appropriate model for a given dataset is crucial. Model performance is closely tied to the underlying characteristics of the data, such as distribution, variance within the features, and dimensionality [1]. Understanding these characteristics can significantly influence the outcomes of predictive modelling efforts, making the study of dataset properties an essential component of data science.

When data is used where the object of study is a human being and the data contains personal data, such as, for example, medical data for clinical research or activity data, it is important to understand that applying ML techniques to the dataset as a whole will not have the same results as applying these techniques to subsets of the data containing the data of a single individual. The optimal classification model, for example, will not be the same for both the whole set and the individual subset, where only the data generated by a single participant are considered [2]. Therefore, analysis of human datasets' characteristics should also consider individual subsets. The results of such an analysis will help improve the applicability of ML techniques on a personal level, for example, by applying clustering based on similar characteristics [3].

When working with such human datasets, it can be complicated to collect a large number of samples. In clinical research, for example, the number of patients who meet the research criteria can be limited [4], or it can be expensive and complicated to monitor a sufficiently large group of participants. In such cases, simulated data could be a solution. In research and industry, simulated datasets are commonly used to test and refine models before applying them to real-world data. However, the degree to which these simulated datasets accurately represent real-world scenarios is a critical factor in their utility [5].

Our study aims to investigate the possibilities of using simulated data when using human datasets for personalized applications, specifically for testing clustering approaches. The rest of the paper is organized as follows. In Section II a number of related works are presented. Next, in Section III, the results are presented in Section IV and discussed in Section V, and the final conclusions are presented in Section VI.

II. Related Works

The exploration of dataset characteristics and their influence on model selection has been a subject of considerable interest within machine learning. Several studies have focused on the impact of dataset characteristics, such as class distribution, feature relevance, and noise levels, on model performance. For instance, Sheykhmousa *et al.* [6] conducted a comprehensive study on the role of feature distribution in model accuracy, concluding that models like Random Forest (RF) and Support Vector Machines (SVM) are particularly sensitive to the distribution of input features. Similarly, Wang *et al.* [1] explored how class imbalance affects the performance of classification algorithms, highlighting the need for careful preprocessing to ensure balanced datasets for optimal model performance.

The use of simulated datasets as proxies for real-world data has been another area of significant research. The work of Liu and Demosthenes [5] examined the fidelity of simulated datasets in replicating the characteristics of real-world data. Their findings indicated that while simulated datasets can be effective in controlled experiments, they often lack the variability and complexity of real-world data, leading to discrepancies in model performance. This underscores the importance of validating model results on original datasets to ensure their applicability in practical scenarios.

Clustering techniques have been widely used to group datasets with similar characteristics to study model predictability. Ikotun *et al.* [7] utilized K-means clustering to categorize datasets and then examined the model performances within each cluster. Their research suggested that while clustering can provide insights into which models might perform well on similar datasets, the predictability of the optimal model is not guaranteed, as other factors, such as dataset complexity, also play a crucial role. The work of Van Buren *et al.* [3] shows that clustering of individual datasets, based on the characteristics of the individual subsets can be applied without significant loss of performance.

III. Methodology

To investigate the usability of simulated datasets in personalized applications, the dataset used in the personalized coaching approach presented by Dijkhuis *et al.* [2] is used as the real-world data reference. This data set consists of step count data from 43 participants, collected using wearable devices, which is used to predict whether the participant who produces these data will reach the set step goal for the day. The prediction is made by taking the step count of the past hour, the cumulative step count up to that hour, the hour, and the weekday as input for a binary classification.

This research aims to establish whether simulated data can be used to explore the relationship between the datasets' characteristics and the selection of the optimal classification model for a cluster of datasets. The process used to achieve this goal can be decomposed into the following steps:

A) Data generation;
B) Data characteristics calculation;
C) Classification model selection;
D) Clustering and Predictability Analysis;
E) Evaluation.

A. Data generation

A generator developed in Dijkhuis et al.'s original research[8] was used to generate the simulated data. This generator takes 16 parameters as input to generate step data with different characteristics for a given period of time. Among these 16 parameters, only two are relevant for this paper. The other parameters concern some technical settings, such as simulation time and settings for the drift patterns and intervention moments to be used. The latter two groups were not used because they are not present in the original data.

TABLE I
The 14 Characteristics.

ID	Description
1	threshold
2	train_set_size
3	number_of_observations
4	class_balance
5	average_sum_steps
6	average_sum_steps_hour
7	variance_sum_steps
8	variance_sum_steps_hour
9	standard_deviation_sum_steps
10	standard_deviation_sum_steps_hour
11	Pearson_correlation_hour_sum_steps
12	Pearson_correlation_hour_sum_steps_hour
13	sparsity_hour_sum_steps
14	sparsity_hour_sum_steps_hour

The two parameters used are movement pattern and movement intensity. The pattern can be set to six different values; *day_one*, *morning_two*, *morning_three*, *afternoon_three*, *evening_two* or *evening_three*. Each setting resulting in different peak moment(s), morning, around noon or end of the day, of activity and different number of peaks, one, two or three. The intensity can be set to three values, *high*, *average* and *low*, influencing the height and width of the peak moments.

Combining the values of these two parameters results in 18 different settings. The simulator was started ten times for each possible combination, resulting in 180 datasets for the simulated data.

B. Data characteristics calculation

For each dataset fourteen values were calculated to describe its characteristics. The id's and descriptions of these values are given in Table I. To establish usability of simulated data, it is necessary to verify that for each setting of ten datasets the within-group variability, that is the intra-setting consistency, is low in comparison to between-groups variability, that is, the inter-setting variability. Furthermore, by comparing the values of the simulated data with those of the real-world data, we can evaluate whether the simulated data resemble the original data. Characteristics 1,2 and 3 are determined by the technical settings of the simulator and are therefore not considered to be key characteristics.

C. Classification model selection

The process to determine the optimal classification model for each setting consists of determining the optimal hyperparameter setting from a selection of different models, the fitting, and then evaluating the performance of the fitted model. Since the fitting process is time-consuming, we consider only three different models, the Bayesian Classifier (BAC), K-Nearest

Neighbors (KNN) and Random Forest (RF). These models were chosen because of their differing algorithms and their potential to perform well on various types of data sets[9].

The models were trained using a 30% - 70% split for the train and the test set. The primary metric for evaluating model performance was accuracy of the classification on the test set, although other metrics such as precision, recall, and F1-score were also collected to validate the results.

D. Clustering and Predictability Analysis

To explore the relationship between dataset characteristics and model performance, K-means clustering was applied to obtain clusters of datasets with similar characteristics. Within each cluster the optimal model is determined for the individual datasets, to evaluate if the given characteristics set tends to favor a specific model.

E. Evaluation

The final step of the process involves analyzing the results obtained from steps 2 to 4. The following points should be evaluated.

- Usability of the data generated by the simulator. For the simulated data to be useful, there should be intra-setting consistency and inter-setting variability, and resemblance of the simulated data to the real-world data.
- Optimal Model Identification: The distribution of optimal models across different datasets and settings was analyzed to identify patterns or trends.
- Clustering and Predictability Analysis: The predictability of the optimal model based on the characteristics of the dataset was examined by analyzing the results within each cluster of datasets. The consistency of model selection within clusters was evaluated to determine if similar datasets consistently yield the same optimal model.

IV. Results

A. Usability of the generated data

To evaluate the intra-setting consistency and the inter-setting variability, we used the coefficient of variability (COV) per characteristic. Table II shows the intra- and inter- values for the relevant characteristics. The intra value is the average value of the COV of the characteristic per group. For a given characteristic, the inter-value is COV of the averages of that characteristic among the groups. The differences show that in most cases the inter-COV is significantly higher than the intra-COV, indicating that datasets generated using the same settings exhibit similar characteristics, essential for controlled experiments and model evaluation. Applying different settings allows for variation in properties, which is crucial for testing model robustness across various scenarios.

To evaluate how well the simulated data replicated real-world data the variance of characteristics of the simulated dataset was compared to the corresponding values of the original dataset. From the values in Table III, we see that the original datasets generally exhibited higher variance. This higher variance suggests greater natural variability in the real-world data compared to the more controlled simulated datasets. However, certain characteristics, such as Characteristics 11 and 12, showed similar variance levels between the original and simulated datasets.

TABLE II
Coefficients of variability for the relevant characteristics.

ID	Intra	Inter	Difference
5	0.06	0.21	0.14
6	0.07	0.22	0.14
7	0.31	0.65	0.34
8	0.21	0.50	0.29
9	0.15	0.35	0.20
10	0.1	0.26	0.16
11	0.69	36.45	35.76
12	0.06	0.07	0.01
13	0.46	0.82	0.37
14	1.32	1.22	-0.11

TABLE III
Variances of original and simulated datasets.

ID	mean variance original	mean variance simulated
4	0.012	0.006
5	24249.2	915.8
6	1164409.9	43061.0
7	1.3E+11	7.7E+8
8	4.3E+13	7.8E+10
9	36554.9	4665.7
10	675294.6	36766.1
11	0.010	0.007
12	0.0045	0.0043
13	0.0018	0.0006
14	0.00087	.00001

B. Model Performance and Optimal Model Identification

As shown in Figure 1, the evaluation of model performance across the datasets revealed that no single model consistently outperformed others. Instead, the optimal model varied depending on the dataset characteristics, underscoring the importance of considering dataset-specific properties when selecting a model. Table IV summarizes the performance metrics and frequency of selection for each optimal model across all datasets. The Random Forest (RF) model was the most frequently selected, with a high average accuracy and F1-Score, suggesting its robustness across different data conditions. The Bayesian Classifier (BAC) and K-Nearest Neighbors (KNN) models, while optimal in fewer cases, also demonstrated strong performance in specific scenarios. In comparison, the optimal model performance and frequencies on the original dataset are given in brackets.

C. Clustering and Predictability Analysis

Figure 2 shows the resulting distribution of optimal models within groups with similar characteristics selected by the K-means clustering. Although some clusters exhibited the predominance of a single optimal model, others showed a mix of models. Only two clusters showed a single selected model, seven showed a predominant choice, and six showed two

TABLE IV

Average performance metrics of optimal models across simulated datasets and original dataset (in brackets).

algorithm	Accuracy	F1-Score	Count
BAC	0.95 (0.96)	0.94 (0.94)	73 (16)
KNN	0.95 (0.94)	0.95 (0.94)	27 (1)
RF	0.95 (0.96)	0.94 (0.93)	79 (26)

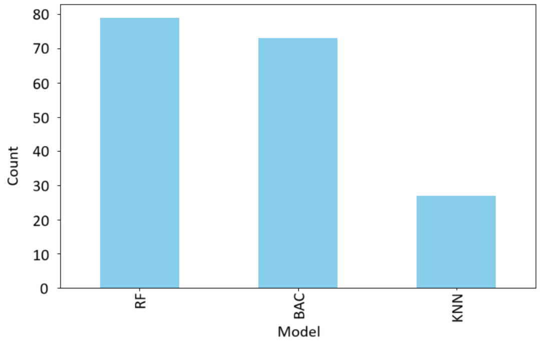

Fig. 1. Bar plot showing the distribution of optimal models across all datasets.

models as optimal choices. This indicates that the unsupervised K-means method applied for clustering is a good first step, but more factors should be considered for model selection.

V. Discussion

The results of this study provide several key insights into the relationship between dataset characteristics and model performance:

Consistency within Settings: The high consistency of characteristics within each setting confirms the reliability of the simulation process, ensuring that datasets generated under the same conditions are comparable.

Variability Across Settings: The significant differences in characteristics across settings show that simulated data can be used to generate different data scenarios when evaluating model performance.

Model Selection and Predictability: The variation in optimal model performance across datasets and the mixed results within clusters of datasets suggest that while dataset characteristics are important, they must be considered alongside other factors, such as model complexity and feature interactions, for effective model selection.

Random Forest (RF) shows strong performance across multiple clusters, often being the dominant or a close second model in many clusters. This suggests that RF is generally a robust model across various dataset characteristics. The same results were found with the original dataset in [2, 3].

Bayesian Classifier (BAC) performs well in several clusters, particularly where RF does not dominate. It shows strong performance in clusters where RF is either less effective or where dataset characteristics uniquely favor BAC.

K-Nearest Neighbors (KNN) appears less frequently as the optimal model, indicating that it may be more specialized or sensitive to specific dataset characteristics. However, when it does perform well, it competes closely with the other models.

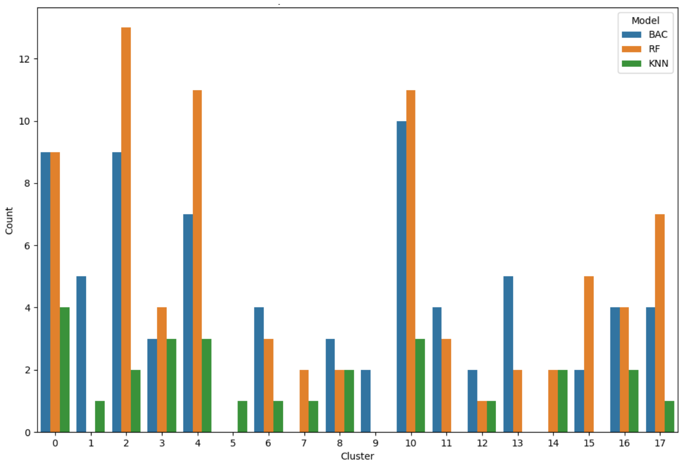

Fig. 2. Bar plot showing the distribution of optimal models within clusters of datasets with similar characteristics.

VI. Conclusion

This study enhances our understanding of the critical role that dataset characteristics play in model selection and performance. The analysis demonstrated that both intra-setting consistency and inter-setting variability are essential for optimizing model performance in practical machine learning applications. The consistency within settings ensures reliable comparisons, while the variability across settings allows for the evaluation of model robustness under diverse scenarios.

This study highlights the importance of the use of simulators when the process of obtaining new data becomes difficult. By comparing simulated datasets to real-world data, it became clear that simulations provide valuable insights but require further refinement to capture the total variability of real-world conditions. Addressing this gap in future work will improve the possibility of generalizing machine learning models, making them more applicable to complex, real-world datasets.

The results also highlight the necessity of carefully considering dataset properties when selecting models. While Random Forest often performed robustly across various conditions, the Bayesian Classifier and K-Nearest Neighbors also showed competitive performance in specific scenarios. This underscores the importance of a dataset-specific approach to model selection. In future work, more insight in model selection can be obtained by considering datasets form other fields. A more detailed analysis of the impact of higher number of features, or different spread in characteristics values will certainly improve our understanding of model performance results.

In conclusion, this study provides critical insights for researchers and practitioners in machine learning, emphasizing the importance of dataset characteristics and variability in building and selecting robust models for diverse data conditions. Future efforts should aim to improve the realism of simulations, leading to better-aligned machine learning outcomes in real-world applications.

References

[1] Le Wang et al. “Review of Classification Methods on Unbalanced Data Sets”. In: *IEEE Access* 9 (2021), pp. 64606–64628. ISSN: 21693536. DOI: 10 . 1109 / ACCESS.2021.3074243.

[2] Talko B. Dijkhuis et al. “Personalized physical activity coaching: A machine learning approach”. In: *Sensors (Switzerland)* 18.2 (Feb. 2018), p. 623. ISSN: 14248220. DOI: https://doi.org/10.3390/s18020623.

[3] Annika Van Buren et al. “A Clustering Approach for Personalized Coaching Applications”. In: *Advances in Computational Collective Intelligence*. Ed. by Ngoc-Thanh Nguyen et al. Leipzig: Springer, Cham, Sept. 2024, pp. 351–363. ISBN: 9783031702587. DOI: https://doi.org/10.1007/978-3-031-70259-4_27.

[4] Harald H. Rietdijk et al. “Feature Selection with Small Data Sets: Identifying Feature Importance for Predictive Classification of Return-to-Work Date after Knee Arthroplasty”. In: *Applied Sciences 2024, Vol. 14, Page 9389* 14.20 (Oct. 2024), p. 9389. ISSN: 2076-3417. DOI: https://doi.org/10.3390/app14209389.

[5] Fang Liu and Panagiotakos Demosthenes. “Real-world data: a brief review of the methods, applications, challenges and opportunities”. In: *BMC Medical Research Methodology 2022 22:1* 22.1 (Nov. 2022), pp. 1–10. ISSN: 1471-2288. DOI: https://doi.org/10.1186/S12874-022-01768-6.

[6] Mohammadreza Sheykhmousa et al. “Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), pp. 6308–6325. ISSN: 21511535. DOI: https://doi.org/10.1109/JSTARS.2020.3026724.

[7] Abiodun M. Ikotun et al. “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data”. In: *Information Sciences* 622 (Apr. 2023), pp. 178–210. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2022.11.139.

[8] H. Kruize and T.B. Dijkhuis. *Git Reps for VFC Simulator*. 2022. URL: https://github.com/HaraldRietdijk/VFCSimulator.

[9] R Devika, Sai Vaishnavi Avilala, and V Subramaniyaswamy. “Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest”. In: *2019 3rd International Conference on Computing Methodologies and Communication (ICCMC)*. 2019, pp. 679–684. DOI: 10.1109/ICCMC.2019.8819654.