Explainable Selection of Machine Learning Algorithms in Social Sciences

Dijana Oreski*
University of Zagreb Faculty of Organization and Informatics
Varazdin, Croatia
dijana.oreski@foi.hr

Luka Katava
University of Zagreb Faculty of Organization and Informatics
Varazdin, Croatia
lkatava@foi.hr

Alen Kisic
VERN University
Zagreb, Croatia
alkisic1@vernnet.hr

Abstract

The increasing availability of machine learning algorithms has posed the challenge of selecting appropriate algorithms for specific data analysis tasks. In domains such as education and business, where many practitioners are not specialists in artificial intelligence, algorithm selection is often performed through trial-and-error experimentation or guided by limited methodological knowledge. Meta-learning has emerged as a promising approach for addressing this challenge by recommending algorithms based on characteristics of previously analysed datasets. However, many meta-learning approaches rely on complex models whose decision processes remain difficult to interpret, limiting their suitability in contexts where transparency and accountability are required. This paper investigates the use of explainable meta-learning models for machine learning algorithm selection in social science domains. Using datasets originating from education and business contexts, we construct a meta-dataset based on dataset characteristics represented as meta-features. These meta-features serve as inputs to interpretable meta-models designed to recommend suitable algorithms for new datasets. We analyse the contribution of individual meta-features to the meta-model decisions, thereby identifying dataset characteristics that drive algorithm recommendations. The results demonstrate that a subset of meta-features plays a key role in determining the predictive power of the meta-model and forms the basis for explainable algorithm selection. By making these relationships explicit, the proposed approach enables transparent and interpretable recommendations that can support non-expert users in selecting appropriate analytical methods. The study contributes to discussions on trustworthy and responsible AI, particularly relevant in the context of emerging AI governance frameworks and certification initiatives that emphasise explainability, accountability, and user trust in AI systems.

1 Introduction

The rapid development of machine learning algorithms has created a fundamental challenge: selecting the most appropriate algorithm for a given dataset and analytical task. This challenge, known as the algorithm selection problem, was first formalized by Rice [1] in 1976 and has become increasingly important as the number and diversity of machine learning algorithms continue to grow. In practice,

*https://louise.foi.hr/members/dijana-oreski/

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

selecting an appropriate algorithm often requires substantial expertise and extensive experimentation, which can represent a significant barrier for practitioners working outside the field of artificial intelligence.

Meta-learning has emerged as a systematic approach to addressing this challenge. By learning from previous experiments conducted on multiple datasets, meta-learning systems aim to recommend suitable algorithms for new data analysis tasks. These systems typically rely on meta-features, which describe structural and statistical characteristics of datasets, enabling models to identify patterns linking dataset properties to algorithm performance. However, many meta-learning approaches rely on complex models whose decision processes remain difficult to interpret, limiting their adoption in domains where transparency and explainability are essential. This issue is particularly relevant in the context of social science research, including domains such as education and business analytics. Researchers and practitioners working with data in these areas often lack deep expertise in machine learning and therefore require decision-support tools that provide not only recommendations but also understandable explanations of the reasoning behind them. Explainable Artificial Intelligence (XAI) has emerged as a key paradigm for addressing the transparency limitations of machine learning systems [2, 3]. When applied to meta-learning, XAI techniques enable the identification of dataset characteristics - represented through meta-features - that influence algorithm selection decisions. By making these relationships explicit, explainable meta-learning models can support more transparent and trustworthy algorithm recommendations.

In this paper, we investigate the use of explainable meta-models for machine learning algorithm selection in datasets originating from social science domains. Using meta-features extracted from a collection of datasets from education and business contexts, we develop and analyse interpretable meta-models capable of recommending suitable algorithms while providing insights into the factors driving these recommendations. The goal is to support data analysts in social sciences by providing algorithm selection mechanisms that are both effective and transparent.

This paper is structured as follows. Section 2 provides a review of the relevant literature on algorithm selection, meta-learning, and explainable artificial intelligence. Section 3 describes the research methodology, including the dataset collection, meta-feature extraction process, and the development of the meta-model used for algorithm recommendation. Section 4 presents the experimental results and evaluates the predictive performance of the proposed approach. Section 5 discusses the implications of the findings for explainability in social science research, particularly in domains such as education and business analytics where interpretability is essential for non-expert users. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Related work

This section presents an overview of the related literature relevant to this study. It focuses on previous research on the algorithm selection problem, meta-learning methods for recommending machine learning algorithms, and explainable AI approaches aimed at improving transparency and interpretability of such systems.

2.1 Meta feature importance

Relevant studies on meta-feature importance revealed patterns alongside domain-specific variations. Meta-features related to dimensionality and sparsity frequently emerge as important across diverse algorithm selection tasks [4], [5], [6]. The mean sparsity of attributes, identified as the most important meta-feature for distance metric recommendation, exemplifies how structural properties fundamentally influence algorithm performance [2]. Statistical meta-features show variable importance depending on the algorithm selection context[5]. In contrast, for clustering tasks, statistical measures of attribute distributions play a more prominent role [4], [5]. Information-theoretic meta-features provide valuable characterization of data complexity and feature relationships [4],[5],[6]. Entropy-based measures and mutual information metrics help predict which algorithms will effectively handle complex feature interactions and class structures. However, the computational cost of extracting some information-theoretic meta-features may limit their practical utility in large-scale meta-learning systems. Domain-specific meta-feature importance patterns have been documented for multi-label classification across text mining, multimedia, and bioinformatics domains [1]. For social

science data encompassing business and education domains, certain meta-features show consistency across domains while others show significant variation [88].

2.2 Performance of Explainable Meta-Learning Systems

Explainable meta-learning systems demonstrate good performance across multiple algorithm selection tasks. For multi-label classification (as case in this research), the automated algorithm selector outperformed all individual algorithms across six different performance metrics, demonstrating the value of meta-learning for this complex task [7]. Distance metric recommendation for k-means clustering achieved about 70% accuracy with the full meta-feature set, improving to 72% when using only the top 25 most important meta-features [4]. This improvement demonstrates that explainability can directly enhance performance by enabling principled feature selection. The reduction in meta-features also decreased computational overhead, providing practical benefits beyond interpretability. The performance gains from meta-learning vary depending on the evaluation metric and dataset characteristics [7]. This variability underscores the importance of explainability: understanding when and why meta-learning provides benefits enables more informed deployment decisions. AutoML systems incorporating explainable meta-learning show promise for democratizing machine learning [6]. By automating algorithm selection while providing transparency through explanation techniques, these systems make machine learning accessible to users without deep technical expertise. However, the effectiveness of these systems depends on the quality of explanations and their alignment with user needs and mental models.

2.3 Domain-Specific Meta-Learning Frameworks

The development of domain-specific meta-learning frameworks for social sciences, namely business and education represents an important research direction. Such frameworks would incorporate domain-specific meta-features, evaluation metrics, and explanation strategies tailored to the unique characteristics and requirements of these domains [6]. For business applications, this might include meta-features related to data quality along with explanations that align with business logic and regulatory requirements. For education applications, domain-specific frameworks should address fairness and the potential for bias in algorithm selection [9]. Explainability techniques could help identify when algorithm selection decisions may have disparate impacts across student populations, enabling proactive mitigation of bias.

Cross-domain meta-learning presents opportunities for leveraging knowledge across related domains. The finding that certain meta-features exhibit consistency across business and education domains suggests potential for transfer learning approaches [8]. A meta-learning model trained on diverse business and education datasets might generalize effectively to new problems in these domains. The development of domain-specific meta-feature models could facilitate more effective meta-learning in business and education contexts.

3 Research methods

Methodology of this research encompasses (i) meta-learning, (ii) explainable AI and ML models, (iii) meta-features role in explainablity of ML algorithms selection meta-models.

3.1 Meta-Learning and Algorithm Selection

Meta-learning, often described as "learning to learn," addresses the algorithm selection problem by leveraging knowledge gained from previous learning experiences. The fundamental premise is that datasets sharing similar characteristics tend to benefit from similar algorithms. Meta-learning systems extract meta-features - general, statistical, clustering, and information-theoretic properties,...-from datasets and use these features to train meta-models that predict algorithm performance or recommend optimal algorithms [1]. The algorithm selection problem manifests across various machine learning tasks. For multi-label classification, where instances can belong to multiple classes simultaneously, selecting appropriate algorithms is particularly challenging due to the diversity of available approaches and the complexity of evaluation metrics [1]. Recent advances have extended meta-learning to AutoML (Automated Machine Learning) systems, which automate the entire machine learning

pipeline including algorithm selection, hyperparameter tuning, and feature engineering [10]. These systems promise to make machine learning more available by reducing the need for expert knowledge.

3.2 Explainable AI: Principles and Techniques

Explainable AI encompasses methods and techniques that make machine learning models interpretable to humans. Post-hoc explanation methods, which generate explanations after model training, have gained prominence due to their model-agnostic nature and applicability to complex black-box models [11]. Among these approaches, feature importance analysis represents one of the most widely used techniques for understanding how input variables influence model predictions. After the development and training of a predictive model, feature importance methods can be used to quantify the contribution of each feature to the model's decision-making process. Such analyses allow researchers to identify which variables have the greatest impact on model predictions and to better understand the relationships captured by the model.

3.3 Meta-Features and Their Role in Algorithm Selection

Meta-features serve as the foundation for meta-learning systems, characterizing datasets in ways that correlate with algorithm performance. These features typically fall into several categories: general measures (number of fetaures, number of instances, number of categorical features, number of numerical features), statistical measures (mean, variance, skewness, kurtosis), information-theoretic properties (entropy, mutual information), and complexity measures (class separability, feature correlation) [7], [4], [5], [10]. The selection and engineering of meta-features significantly impact meta-learning performance. Research has shown that different meta-features give varying importance across domains and tasks [7], [6]. For instance, in multi-label classification, meta-features related to label distribution and label relationships prove particularly influential [7], while clustering tasks prioritize structural properties like attribute sparsity [4]. Understanding meta-feature importance is crucial for several reasons. First, it enables feature selection to reduce computational overhead and improve meta-model generalization [4], [5]. Second, it provides insights into the underlying mechanisms of algorithm performance, potentially guiding algorithm design [6]. Third, it facilitates domain-specific customization of meta-learning systems by identifying which dataset characteristics matter most in particular application contexts [8].

4 Research results

This section presents results of the research divided into three parts. First, we discuss extractuon of meta-features. Next, we explain meta-model development followed by interpretation and explanation of meta-model.

4.1 Meta-Feature Extraction and Characterization

The foundation of explainable meta-learning lies in comprehensive meta-feature extraction that captures relevant data set characteristics. Hereinafter, we have extracted a total of 45 datasets. Datasets were collected from publicly available repositories and analysed in order to describe their properties through meta-features. For each dataset, 179 numerical meta-features were computed, capturing different aspects of the data such as distributional properties, statistical characteristics, and structural complexity. These meta-features represent the input space used to analyse relationships between dataset characteristics and the target classes considered in the meta-learning task. To enable consistent comparison across datasets with different scales and distributions, all numerical meta-features were discretised into ordinal categories. Specifically, each feature was partitioned into eight ordered bins, ranging from extremely low to extremely high. The discretisation was performed using a quantile-based binning procedure, which ensures approximately balanced distributions of instances across bins. In cases where quantile boundaries produced duplicate thresholds due to limited variability in the data, a fallback binning chain was applied to guarantee a valid ordinal categorisation. This discretisation step allowed meta-features to be treated as interpretable categorical descriptors of dataset characteristics. Following discretisation, a feature selection procedure was applied to identify the most informative meta-features. First, mutual information scores were computed for all 179 meta-features with respect to the target variable. Mutual information was used as a model-agnostic

measure of dependency, capturing both linear and non-linear relationships between meta-features and the target classes. To further analyse the relationship between meta-feature categories and the target classes, Kendall's rank correlation coefficient was calculated. For each meta-feature, the correlation between its ordinal category value and each target class was computed. Kendall's was selected because it is particularly suitable for ordinal variables and monotonic relationships, which aligns with the discretised nature of the meta-feature representation. Based on the obtained correlation values, the top 15 meta-features for each class were selected according to the highest absolute Kendall's scores. These selected features represent the dataset characteristics most strongly associated with each class and were subsequently used in the rule extraction and analysis stages of the study. This process enabled the identification of interpretable relationships between dataset properties and model selection outcomes within the meta-learning framework.

4.2 Meta-Learning Model Construction

To construct the meta-learning model for algorithm recommendation, a scoring-based approach was employed. The goal of the model is to estimate the suitability of each candidate machine learning algorithm for a given dataset based on its meta-feature representation.

For each dataset, meta-features describing dataset characteristics were first transformed into normalized values in order to enable consistent comparison across different datasets. First, categorical meta-features were transformed to numerical. Second, normalization was performed on scale [-1,1] Specifically, for each feature value v, the value was calculated as: $(v-3.5)/3.5$. The value 3.5 was chosen as the centering constant because the discretisation procedure maps each meta-feature to one of eight ordered bins, indexed from 1 to 8. Within this scheme, 3.5 represents the midpoint of the bin index range, effectively acting as a neutral reference point. Dividing by 3.5 further normalises the centered values such that the extreme bins (1 and 7) map approximately to the interval [1, +1], ensuring that the magnitude of contributions remains comparable across meta-features regardless of their individual distributions. Consequently, meta-feature values falling below the midpoint produce negative contributions to the algorithm score, while values above the midpoint produce positive contributions, allowing the scoring function to capture the directional influence of each dataset characteristic on algorithm suitability.

This transformation ensures that values below the central category produce negative contributions, while values above the central category produce positive contributions, allowing the model to capture directional influence of meta-features.

For each candidate algorithm class, a weighted scoring function was then applied. Each meta-feature contributes to the overall score proportionally to its associated weight parameter , which reflects the importance of that feature for predicting algorithm suitability. The score for a given algorithm class is computed as the weighted sum of the centered feature values across all meta-features. To ensure comparability of scores across classes, the resulting value was normalized by the magnitude of the weight vector . This normalization prevents classes with larger cumulative weights from being systematically favoured. Finally, the recommendation for each dataset was determined by selecting the algorithm class with the highest normalized score. In this way, the meta-model aggregates the contributions of all 15 meta-features to estimate which algorithm is most suitable for the dataset under consideration.

An important advantage of this scoring-based formulation is its inherent interpretability. Because each meta-feature contributes linearly to the final score through an explicitly defined weight, it is possible to analyse how individual dataset characteristics influence the recommendation of specific machine learning algorithms. This property makes the model particularly suitable for explainable algorithm selection in domains such as education and business analytics, where transparency and interpretability are essential.

It is important to note, that we have compared proposed meta-model against several different approaches, such as: ML based meta-models and multi-crietira decision making approaches . The comparison ensures a fair evaluation using identical meta-feature inputs and evaluation metrics. Figure 1 provides deeper insight into the behavior of the meta-model across algorithm classes. The model demonstrates strong performance in identifying Ridge regression, achieving the highest number of correct predictions (8), indicating that its associated meta-feature patterns are well captured.

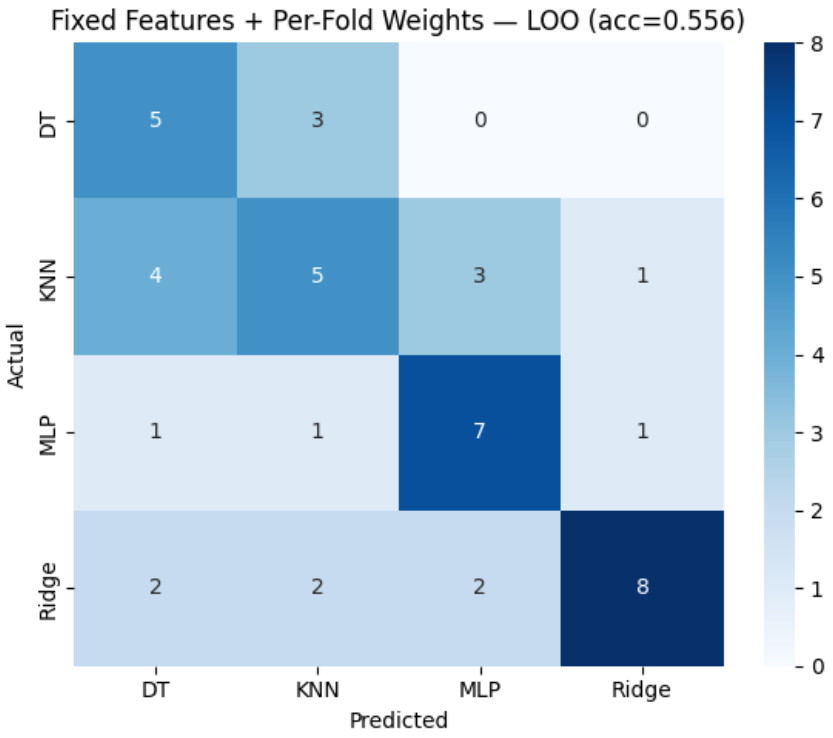

Figure 1: Confusion matrix of meta-model.

In contrast, confusion is more pronounced between DT and KNN, as well as between KNN and MLP, suggesting that these algorithms share overlapping meta-feature characteristics. This is particularly visible in the misclassification of KNN instances, which are frequently predicted as DT or MLP.

These results indicate that while the meta-model effectively distinguishes algorithms with more distinct statistical or structural signatures (e.g., Ridge), it encounters challenges when decision boundaries between algorithms are less clearly separable in the meta-feature space. This pattern suggests that the meta-learning task is inherently complex and influenced by subtle interactions between dataset characteristics, rather than dominated by a single discriminative feature. The feature importance analysis based on Kendall's tau (Figure 2) reveals several notable patterns. Features related to correlation structure, such as cor.sd and cor.mean, exhibit strong associations with specific algorithms, particularly KNN, indicating that distance-based methods benefit from datasets with distinct correlation patterns.

4.3 Explanation Generation and Interpretation

To support explainability of the meta-learning model, a sensitivity analysis was conducted in order to assess the contribution of individual meta-features to the algorithm recommendation process.The analysis was performed using a leave-one-out (LOO) evaluation setup. The baseline predictive performance of the model in this configuration was 0.556. The results of the sensitivity analysis highlight *cor.sd* as the most influential meta-feature in the model. When this feature was removed from the model, the performance decreased by 0.089, indicating a contribution to the predictive capability of the meta-model. This finding is consistent with the results obtained from the leave-one-out (LOO) analysis, confirming the stability of the importance ranking across different validation strategies. In contrast, the remaining meta-features demonstrated smaller or negligible individual effects when analysed independently. While their removal did not produce substantial decreases in predictive performance, this does not imply that they are irrelevant for the model. Instead, the results suggest that the meta-model partially relies on the complementary interaction of multiple meta-features, where the joint presence of several dataset characteristics contributes to the algorithm recommendation process.

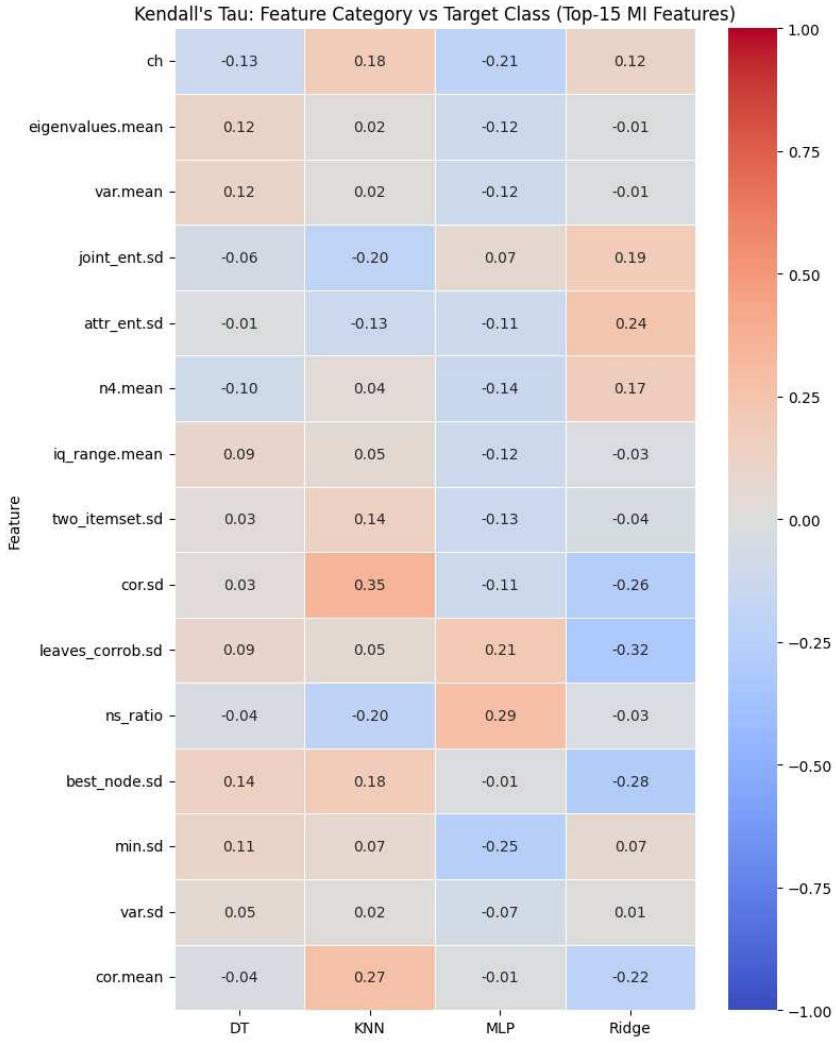

Figure 2: Meta-feature importance.

From the perspective of explainability, these findings provide valuable insights into how dataset characteristics influence the behaviour of the meta-learning model. The prominence of cor.sd indicates that measures describing the variability of attribute correlations play an important role in determining suitable machine learning algorithms for the analysed datasets. At the same time, the relatively distributed influence of the remaining meta-features suggests that the model captures more complex patterns arising from the combined properties of datasets rather than relying on a small number of dominant variables. Such interpretability is particularly important in social science domains such as education and business analytics, where researchers and practitioners require transparent justification for algorithm recommendations. By identifying which meta-features drive the model's decisions, the proposed approach enables users to better understand the relationship between dataset characteristics and algorithm suitability, thereby increasing the transparency and trustworthiness of the recommendation system.

Model-based meta-features such as leavesCorrob.sd and bestNnode.sd show strong influence, especially in distinguishing between MLP and Ridge, highlighting the importance of structural dataset properties in algorithm selection.

5 Discussion and implications

The results obtained from the feature importance (Figure 2) analysis based on scoring meta-model provide insights into how different groups of meta-features contribute to the explainability of the algorithm recommendation process. The meta-features used in this study originate from several established categories in meta-learning literature, including general, statistical, information-theoretic, itemset, model-based, clustering, complexity, landmarking, relative, and concept-based descriptors. Together, these groups capture different aspects of dataset structure and learning difficulty, enabling the meta-model to infer relationships between dataset characteristics and the suitability of particular machine learning algorithms.

An analysis reveals that certain meta-features demonstrate stronger relationships with specific algorithm classes, which provides a basis for interpretable algorithm recommendations. In particular, correlation-based statistical descriptors, such as *cor.sd* and *cor.mean*, show notable influence, especially for the KNN algorithm. The relatively strong positive correlation between cor.sd and KNN suggests that variability in attribute correlations may favour algorithms that rely on distance-based similarity measures. This finding is consistent with the intuition that neighbourhood-based methods can benefit from structured relationships between variables.

Another notable pattern appears in the general meta-features, such as *ns.ratio* and *leaves.corrob.sd*, which show stronger associations with the MLP algorithm. This suggests that datasets with more complex decision boundaries or structural variability may favour algorithms with higher representational capacity.

In contrast, several meta-features demonstrate relatively weak individual correlations with algorithm classes. Rather than indicating irrelevance, this pattern suggests that the meta-model partially relies on the combined contribution of multiple meta-features. In other words, algorithm recommendations are not driven by a single dominant dataset characteristic in most cases, but emerge from the interaction of several complementary descriptors describing dataset structure, distribution, and complexity.

From the perspective of explainable AI, these findings are particularly important. Because the meta-model is constructed using explicit weights derived from correlations between meta-features and algorithm classes, it becomes possible to trace how specific dataset properties contribute to the final recommendation score. This transparency enables researchers and practitioners to understand the reasoning behind algorithm selection decisions.

Such explainability is valuable in social science domains such as education and business analytics, where analysts often require understandable justification for automated recommendations. By linking algorithm choices to interpretable dataset characteristics, the proposed approach supports transparent and trustworthy decision support in data analysis workflows.

While the proposed scoring-based meta-model provides interpretability by explicitly linking meta-features to algorithm selection, this transparency may come at the cost of predictive accuracy compared to more complex models such as ensemble methods or AutoML systems. However, in domains such as social sciences, where explainability and trust are critical, this trade-off is often acceptable. The results suggest that the model achieves competitive performance while offering substantially higher transparency.

6 Conclusion

This paper addressed the problem of selecting appropriate machine learning algorithms for datasets in social science domains by focusing on the explainability of meta-learning approaches. Using datasets originating from education and business contexts, we developed and analysed an interpretable meta-model that recommends suitable machine learning algorithms based on dataset meta-features.

The results show that a set of relevant meta-features can effectively capture dataset characteristics that influence algorithm performance and can therefore serve as a foundation for explainable algorithm selection. Identifying these influential meta-features enables the interpretation of meta-model decisions and provides insights into why specific algorithms are recommended for particular types of datasets. In this way, meta-features not only support predictive performance of the meta-model but also act as the basis for explainability in the algorithm recommendation process.

Explainability is particularly important in domains such as education and business analytics, where data analysts and researchers are often not specialists in artificial intelligence or machine learning. Providing interpretable recommendations can therefore improve trust, transparency, and usability of decision-support systems that assist users in selecting appropriate analytical methods.

This study also has several limitations. The meta-model was trained and evaluated on a relatively limited subset of datasets and machine learning algorithms, which may restrict the generalizability of the findings. The current study considers a limited set of candidate algorithms (DT, KNN, MLP, Ridge), which may restrict the generalizability of the findings. Expanding the pool of algorithms to include ensemble methods (e.g., Random Forest, Gradient Boosting) and modern approaches would improve the practical applicability of the framework. Future work will also focus on expanding the dataset repository and further exploring explainability techniques that can enhance the transparency of algorithm recommendation systems.

Overall, the results suggest that explainable meta-learning approaches represent a promising direction for supporting data analysts in social science domains, enabling more transparent and informed selection of machine learning algorithms.

Acknowledgments and Disclosure of Funding

This research was supported by Croatian Science Foundation under the project SIMON: Intelligent system for automatic selection of machine learning algorithms and Strategic Partnership for Inovation programme under the project OptiSolarAI: Autonomus system for optimal storage and distribution of electric energy based on artificial intelligence.

References

[1] J. R. Rice, “The algorithm selection problem,” in Advances in Computers, vol. 15, pp. 65–118, 1976.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi and F. Giannotti, “A survey of methods for explaining black box models,” ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.

[3] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018.

[4] R. Gonzales et al., "Distance Metric Recommendation for k-Means Clustering: A Meta-Learning Approach," TENCON2022 - 2022 IEEE Region 10 Conference (TENCON), 2022. DOI: 10.1109/TENCON55691.2022.9978037

[5] M. E. M. Gonzales, L. C. Uy, J. A. L. Sy and M. O. Cordel, "Distance Metric Recommendation for k-Means Clustering: A Meta-Learning Approach," TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON), Hong Kong, Hong Kong, 2022, pp. 1-6, doi: 10.1109/TENCON55691.2022.9978037.

[6] M. Garouani, A. Ahmad, and M. Bouneffa, “Explaining meta-features importance in meta-learning through Shapley values,” in Proc. ICEIS, vol. 1, pp. 591–598, Apr. 2023.

[7] A. Kostovska, C. Doerr, S. Džeroski, D. Kocev, P. Panov and T. Eftimov, "Explainable Model-specific Algorithm Selection for Multi-Label Classification," 2022 IEEE Symposium Series on Computational Intelligence (SSCI), Singapore, Singapore, 2022, pp. 39-46, doi: 10.1109/SSCI51031.2022.10022177.

[8] D. Oreški, D. Višnjić, and N. Kadoić, “Unlocking automated machine learning efficiency: Meta-learning dynamics in social sciences for education and business data,” TEM Journal, vol. 13, no. 1, pp. 797–808, Feb. 2024, doi: 10.18421/TEM131-82.

[9] A. Barhrhouj, B. Ananou, and M. Ouladsine, “Exploring explainable machine learning for enhanced ship performance monitoring,” in Machine Learning, Optimization, and Data Science, G. Nicosia, V. Ojha, S. Giesselbach, M. P. Pardalos, and R. Umeton, Eds., Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2024

[10] S. Manna and N. Sett, “Need of AI in modern education: In the eyes of explainable AI (XAI),” in Blockchain and AI in Shaping the Modern Education System, Boca Raton, FL, USA: CRC Press, 2025, pp. 89–115.

[11] T. Han, S. Srinivas, and H. Lakkaraju, “Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations,” in Advances in Neural Information Processing Systems, vol. 35, pp. 5256–5268, 2022.