Explainable Data Mining for High-Dimensional Datasets Based on Hybrid Machine Learning Models

[1] **Dr. Kamran Yousaf**
[2] **Amina Siddiqui**

How to cite this article:

Yousaf, K., & Siddiqui, A. (2024). Explainable Data Mining for High-Dimensional Datasets Based on Hybrid Machine Learning Models. *Journal of Technological Information, Management & Engineering Sciences, 5(2), 31-44.*

Received	Accepted	Published
05 July 2024	10 September 2024	20 November 2024

Abstract

There are several issues with high-dimensional datasets: feature redundancies, overfitting, computational costs, and the lack of interpretability.While high-dimensional datasets provide rich information for prediction and knowledge discovery, they also pose challenges related to feature redundancy, overfitting, computational costs, and limited interpretability. The hybrid machine learning models combine complementary techniques, including feature selection, dimensionality reduction, ensemble learning, and post hoc explanation, in order to enhance the predictive performance as well as transparency. This paper discusses an integrative review of explainable data mining with hybrid machine learning models in high dimensional datasets. It covers feature selection, random forests, gradient-boosted trees, local surrogate explanations and Shapley-based feature-attribution. The review reveals a major unfilled need: hybrid pipelines tend to focus on the accuracy of their predictions, and less systematic attention is given to the validity, stability, and usefulness of their explanations for human use. The paper brings forward a modular approach which integrates filter and embedded feature selection, ensemble modeling, global and local explanations through SHAP, stability analysis and human-centered validation. Results are qualitative analytical results that are not new experimental results but literature based. The results of the analysis suggest a hybrid approach for interpretable high-dimensional data mining, in which explanations are interpreted as model-dependent evidence—an interpretation that does not mean that explanations are causal evidence. The proposed procedure focuses on minimising leakage, replicability, fidelity to the explanation, and reporting uncertainty transparently.

Keywords: *explainable data mining, high-dimensional data, hybrid machine learning, feature selection, SHAP, LIME, ensemble learning.*

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Introduction

High dimensional datasets have many variables compared to the number of observations, or many constructed features. These include: genomic data, text corpora, sensor streams, financial transaction records, and image-based measurements. These datasets can show subtle patterns, but can also contain spurious associations, be multicollinear, overfit and can be computationally expensive to analyze. A model can have high apparent performance in a high dimensional space by taking advantage of noise or leakage instead of stable structures.

The goal of data mining is to find patterns that are useful in the data, and the goal of machine learning is to provide algorithms for prediction, classification, clustering, and representation learning. These activities should not be divorced from explanation in high dimensional applications. It may be hard to validate, debug, regulate, or use responsibly if the prediction can't be tested. Explainable data mining aims not only to produce accurate models, but to provide understandable explanations for the evidence that led to an individual prediction and to provide understand explanations for the patterns.

Hybrid machine learning models are a combination of several components to solve different parts of a problem. A typical pipeline can involve a filter-based feature selection, an embedded selection as part of a model training process, an ensemble learner as a prediction tool and a post hoc explanation method to explain the predictions based on features. This can be a tradeoff among the interpretability, capacity to scale, and the number of models that can be used to predict. Note that each additional component carries design implications and potential sources of error.

This paper has three contributions to make. First, it summarizes the fundamental and representative techniques for feature selection, ensemble learning and explanation. Second, it points to the shortcomings of existing hybrid methods, particularly the lack of quality and stability assessment of explanations. Third, it suggests a modular and replicable approach for explainable data mining in high-dimensional data.

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Research Questions

The review is based on three questions. First, what are the most important hybrid components for high-dimensional data mining? Second, how to render explanations in complex ensemble models valid, stable and useful? Third, what kind of evaluation safeguards are needed to separate meaningful patterns from artifacts of pre-processing, sampling, and/or complexity of the model? In the paper they are not claiming anything about the optimality of any specific hybrid configuration.

Literature Review

Feature Selection and Dimensionality Reduction.

The objective of feature selection is to remove informative features without losing information from the data to the learning model. Guyon and Elisseeff (2003) classify the variables into raw and constructed variables and discuss the three methods of variable selection: the filter method, the wrapper method and the embedded method. The filter methods are based on features that operate without reference to a particular learner, the wrapper methods are based on how well the subsets perform when predicting features, and the embedded methods make selection as part of model training.

There is a difference between Feature Selection and Dimensionality Reduction. While techniques like principal component analysis will produce new components, which may be less directly interpretable, selection will retain original variables. The best option for high-dimensional data mining will depend on whether the user would like to analyse original variables, reduce the cost of measurements or reduce the space used. While feature selection can eliminate noise and reduce computational costs, it can also eliminate rare but important signals, provided that they are not selected.

Data leakage is one of the methodological issues to consider. Feature selection should be done per training fold and/or based on a training-specific preprocessing pipeline. Pre-selection of features can introduce information from observations used for validation or test sets to affect the set of features, leading to an optimistic performance estimate. This concern is even more relevant in high

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

dimensional data sets, with many potential features to which there can be more chance associations.

Ensemble Learning

Ensemble methods aim to achieve good generalization and robustness by using multiple learners. The random forests of Breiman (2001) use many decision trees trained on bootstrap samples and random subsets of features. The method is suitable for high dimensional tabular data, as it can handle nonlinear relationship, interactions, and different types of variables, and offers measures of the importance of variables.

The XGBoost (Chen and Guestrin, 2016) is a scalable gradient-boosted tree system. Gradient boosting adds trees in turn, correcting the errors made by the previous learners. XGBoost supports sparse data, approximate tree training, cache-aware data access, and distributed computing, which are useful for large-scale data-mining applications. However, its predictive power cannot obviate the necessity of a careful regularization, validation, and explanation.

The ensemble models can be called black boxes since there are numerous constituent trees and interactions that make the model hard to examine individually. It certainly doesn't mean that their predictions aren't explainable. Instead, it is the need for methods of explanation that relate to summarizing the effects of some features of the input on a specific prediction or the overall operation of the model. Explanations should be judged not for their validity, but for their fidelity, stability and usefulness – not on whether they seem true based on their visual plausibility.

Literature Review: Explainability

Local Surrogate Explanations

Local Interpretable Model-Agnostic Explanations (LIME) are proposed by Ribeiro, Singh and Guestrin (2016). LIME creates "noisy" examples of an instance, gets predictions from the original model, and learns a local interpretable surrogate model (e.g. a sparse linear model). The weights produced show which features were most important in the prediction in the neighborhood of that instance.

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

The model agnostic nature of LIME is achieved by avoiding any access to the internal structure of the underlying learner. This flexibility allows it to be used with ensembles of trees, neural networks, and other complicated models. This has the drawbacks of being sensitive to the perturbation process, kernel width, sampling design, and selection of local surrogate model. A LIME explanation describes the local behaviors and shouldn't be generalized to the dataset without further evidence.

Shapley-Based Attribution

Lundberg and Lee (2017) introduced a general framework for additive feature attribution, called SHAP (SHapley Additive exPlanations). SHAP calculates the importance value of each feature for a single prediction, inspired by the concepts of Shapley value in cooperative game theory. The resulting framework is designed to meet properties such as local accuracy, missingness and consistency.

SHAP values are used to understand how a model changes the expected prediction to the prediction given for an instance. They can be used to support local explanations (one prediction) and global summaries (aggregating attribution patterns over observations). SHAP does not remove model assumptions, however, or the computational cost. Computation of exact Shapley values can be costly for models with a large number of features, while approximate methods can add to the variability.

Hybrid Model Design

The feature selection, ensemble prediction, and post hoc explanation can be performed in a hybrid explainable pipeline. For instance, variables with very low variance or variables that are very similar to one another can be removed by a filter method, an embedded method might keep certain variables that are important to a tree ensemble and SHAP can provide information about the prediction of a final ensemble. This design can minimize the dimensionality of the system without losing out on the nonlinear effects and interactions between the dimensions themselves.

Each part needs to have a clear function; this is the central design strategy. Noise and dimensionality is reduced by feature selection, the ensemble captures the

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

nonlinear structure, and the explanation method provides evidence for predictions. None of these components shows causality. A high attribution value is not a measure of the impact of changing the feature on the actual outcome, but rather of whether the feature affected a model's prediction under a model and data distribution.

Research Gap

The literature surveyed offers good basis for feature selection, ensemble learning and post hoc explanation. However, there is a lack of integration in evaluating hybrid pipelines in high dimensional data mining. Many studies have focused on predictive accuracy, and explanations have been regarded as auxiliary visualizations to be added to the predictions. This can cause wrong interpretation especially if feature correlations, preprocessing decisions or sampling variations influence the attribution values.

A second gap is with regard to the stability of explanations. The rankings of features in two reasonable implementations of hybrid pipeline could differ due to the random seed, resampling, hyperparameter selection, or the correlation between features. Users without stability analysis can see one rank as a reliable discovery. A third gap is in the area of human utility – explanations could be mathematically sound but not useful for some aspect of a decision, a domain audit or understanding of a domain.

This paper aims to fill these gaps by suggesting an evaluation framework in which the quality of the explanation is a first-class criterion. The framework takes into account explanation fidelity, the predictiveness of the performance, stability of results across resample, cost of computation, and alignment to domain knowledge. It does not state that these criteria have been tested in practice in this review.

Objectives

The general goal is to analyze the performance of hybrid machine learning approaches to provide explanation for data mining in high dimensional data. The goals are:

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Learn feature-selection, ensemble-learning and explanation techniques that are relevant to high-dimensional data.

Review SHAP & LIME as explanations of complex model predictions.

- Know the constraints on existing hybrid pipelines particularly validity and stability of explanations.

Suggest a semi-modular methodology for interpretability in high dimensional data mining that is leak resistant.

Specify evaluation criteria that address prediction, explanation and reproducibility and human usefulness.

Working Proposition

The working hypothesis is that, in order to make explainable data mining possible, the feature selection, modeling and explainability processes should be designed and evaluated in tandem in an integrated pipeline. The literature reviewed here supports this proposition in concept although it is not empirically tested in this paper. It is important that it be evaluated in controlled comparisons and user-centred evaluations in the future.

Methodology

Review Design

The study adopts integrative review research design based on purposively selected set of influential and verifiable scholarly sources. The corpus consists of papers by Guyon andelisseeff (2003) on variable and feature selection, Breiman (2001) on random forests, Chen and Guestrin (2016) on XGBoost, Ribeiro et al. (2016) on LIME, and Lundberg and Lee (2017) on SHAP. The sources are chosen because they directly cover important aspects of hybrid explainable data-mining systems.

The task, algorithmic mechanism, assumptions, scalability implications, and limitations were explored for each source. The review is not meant to be comprehensive of all explainable AI or high dimensional machine learning studies. The quantitative methodik results of the original papers are not pooled because

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

of the differences in the datasets, evaluation protocols, and computational settings.

Proposed Hybrid Pipeline

The pipeline includes the following steps: problem formulation, data auditing, feature engineering, feature selection, hybrid modeling, and explanation. The first step is a "predictive task design" phase in which prediction task, acceptable error costs, fairness constraints and decisions that will be supported by explanations are defined. Data quality is evaluated in the second phase using missingness patterns, duplicate patterns, leakage risk, class imbalance and temporal or group-based dependency.

The focus on feature engineering ensures all the variables are transformed into features ready for modeling, without losing track of them. Transformations are fitted only on training data, as are encodings, imputations, and derived variables. The pipeline then selects the filters to eliminate uninformative or redundant variables, and then it performs embedded filter selection during the model training process. The criterion and threshold for selection, as well as the set of features selected, are documented.

The modeling step is a comparison between a simple interpretable model (regularized logistic regression, shallow decision tree) and a hybrid ensemble model (random forest, gradient-boosted tree ensemble). Nesting of cross-validation is utilized to tune the hyperparameters. This final model is chosen based on a set of predictive and operational criteria, not necessarily based on the maximum score from the validation.

Methodology: Explanation and Evaluation

Explanation Protocol

There are two levels of explanation generation. Global explanations provide an overview of feature importance and interaction patterns across the data, whereas local explanations offer insights into individual predictions. SHAP is proposed as the key attribution method since its framework for attribution is additive and has desirable theoretical properties (Lundberg & Lee, 2017). LIME is also

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

recommended as a local surrogate approach to supplement other approaches, such as rule-based, where the stakeholders require a simple rule-like description of the individual predictions (Ribeiro et al., 2016).

Explain inputs need to be exactly the same as the model's own preprocessing pipeline. When scaling, imputation or encoding or feature selection is done, explanations should be mapped back to meaningful original variables if possible. Special attention to correlated features, since they may be attributed by several related features. The protocol thus describes correlated feature-groups and does not assume any individual attributions as causal effects.

Evaluation Criteria

The predictive performance is assessed using task specific performance measures, such as area under the receiver operating characteristic curve, precision, recall, F1 score, calibration and root-mean-square error. Metrics are calculated on out-of-sample data which were not used in feature selection, hyperparameter tuning, or model selection. Repeated cross-validation and multiple random seeds are used to assess variability.

Explanation quality is assessed based on fidelity, stability and usefulness. Fidelity checks if the explanation is similar to the behavior of the model in the local context. Stability is an evaluation of how stable the feature rankings are for varying resampling, seeds and slight perturbations. The usefulness of the product/service is measured on the basis of the domain expert review, task-based evaluation or user studies, depending on the application. In this paper, no explanation-quality scores are reported and the following criteria are proposed for future empirical work.

Reproducibility and Governance

To ensure reproducibility, datasets need to be versioned, the preprocessing scripts, the definitions of features, model configurations, random seeds, software versions and explanation settings. Each result reported must specify the data split, evaluation protocol and computational environment. For sensitive high dimensional datasets, access controls and privacy review shall be implemented

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

and access shall be documented and reviewed to ensure that aggregation and anonymization are not performed.

Negative result reporting is also required for the methodology. If a variable is excluded from the model due to feature selection, then there is an explanation method that results in unstable rankings or a complicated model that does not perform better than a simpler baseline, then this should be explained. These results do not overestimate the value of hybrid complexity and facilitate the more responsible use of models.

Results

Qualitative Analytical Findings

The first analytical finding is that high dimensional data mining task must explicitly control dimensionality and information leakage. When variables are numerous and observations are few, Guyon and Elisseeff (2003) put forward that feature selection is a key issue. The synthesis suggests that feature selection needs to be integrated into the validation process and not a part of the data splitting process.

The second discovery is that ensemble methods lie on the solid ground for the development of hybrid explainable systems. Random forests are a collection of many decorrelated trees and are able to process a high dimensional tabular data source (Breiman, 2001), while XGBoost is an implementation of a gradient-boosting system that supports sparse and large-scale data (Chen & Guestrin, 2016). These models can help to represent nonlinear effects and interactions that might not be modelled using simple linear relationships.

The third is that there are varying explanations methods with different scopes and interpretation. LIME generates local surrogate explanations by fitting an interpretable model around a particular prediction (Ribeiro et al., 2016). SHAP can be used to give additive attribution values to features for each prediction, and can also be used to provide global summaries (Lundberg & Lee, 2017). Neither approaches imply causal relationships nor automatically imposes independence on the data's distribution.

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.
Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

The fourth discovery is hybrid models bring in new evaluation needs. While a combination of feature selection, ensemble and post hoc explanation will improve the usability of the pipeline, each stage can impact the final explanation. This means that validity of explanations cannot be determined solely on the basis of predictive accuracy. It will be evaluated by fidelity, stability and stakeholder-based evaluation.

The proposed Outputs are not measurements.

The methodological result is a specification for a clear hybrid pipeline, rather than an improvement in performance measured. The deliverables proposed are a pre-processing pipeline that is leak proof, a validated feature-selection record, a tuned ensemble model, global and local explanation reports, stability analyses, and a reproducibility manifest. The following are suggested outputs for subsequent implementation, and not artifacts of this review.

No accuracy values, feature-importance rankings, runtime measurements and user-study results are reported since no experiments were conducted. The literature synthesis is used to validate the design of the proposed workflow, but any claim that the design affects the predictive or explanatory power of the workflow would need to be evaluated independently, empirically.

Discussion

Explain the meaning and implications of the results.

The review suggests that explainability should be incorporated into the data-mining process, and not implemented after the model is deployed. Feature selection can help to minimize noise and enhance interpretability, and ensemble models can account for complex relationships. The post hoc explanation methods then change the behavior of the model into evidence of features. But the value of this sequence is contingent upon whether or not each element is validated with respect to the others.

Practitioners need to differentiate between the local and global importance. A feature can be significant overall (because of its effect on many predictions) or it can have a large effect on just a few predictions (because of its high attribution

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

value). On the other hand, a feature that is important worldwide might not explain a certain decision. Both levels are reported with the necessary uncertainty and stability information that enable more responsible interpretations.

Limitations and Future Research

There are several limitations of this paper. First, it is a selective overview and not a comprehensive look at explainable artificial intelligence, dimensionality reduction, or hybrid modeling. Secondly, it does not perform new experiments, therefore it can't measure the relative performance of different pipelines. Thirdly, the proposed methodology has yet to be applied and tested with real datasets.

The proposed workflow needs to be assessed in other high dimensional domains such as text, genomics, sensor, business data, etc. in the future. Research should be conducted that compares hybrid pipelines to explainable baselines in scenarios where data access is similar, and evaluate the various metrics of predictive quality, fidelity, stability, computational complexity, and user understanding. The impact of correlated features, class imbalance and distribution shift should also be studied for its reliability of explanation.

Conclusion

For high-dimensional datasets, the goal of "explainable data mining" is more than just accuracy in predicting the results. It needs a hybrid work process that is dimensional, pattern learning, and explanations that are valid, stable, and useful to the stakeholders. Feature selection, ensemble learning and post hoc attribution methods are complementary approaches towards accomplishing these objectives. This paper does not present fabricated experimental results, but rather an integrative review and a methodological proposal in the form of modules. The main idea of the text is that an explanation should be assessed within the data-mining process. The proposed workflow needs to be empirically evaluated in the future with clear baselines, leakage-resistant evaluation and human evaluation of the quality of the explanations.

References

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Chen, T., & Guestrin, C. (2016). XGBoost is a scalable boosting system of trees. In the ACM SIGKDD 22nd International Conference on Knowledge Discovery and Data Mining (KDD 2006), proceedings (pp.Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006), pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Guyon, I., & Elisseeff, A. (2003). Introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182.

Lundberg, S. M., & Lee, S.-I. (2017). A consistent interpretation of model predictions. Advances in Neural Information Processing Systems 30 (pp. 1027–1032), 2017.2017 Advances in Neural Information Processing Systems 30 (pp. 1027–1032). 4765–4774) .

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "How do I know that you're trustworthy?" Describing any classifier's predictions. In the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007), pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Atzmueller, M., Fürnkranz, J., Kliegr, T., & Schmid, U. (2024). Explainable and interpretable machine learning and data mining. Data Mining and Knowledge Discovery, 38, 2571–2595. https://doi.org/10.1007/s10618-024-01041-y

Schwalbe, G., & Finzel, B. (2024). A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. Data Mining and Knowledge Discovery, 38, 3043–3101. https://doi.org/10.1007/s10618-022-00867-8

Marcinkevičs, R., & Vogt, J. E. (2023). Interpretable and explainable machine learning: A methods-centric overview with concrete examples. WIREs Data Mining and Knowledge Discovery, 13(3), e1493. https://doi.org/10.1002/widm.1493

Villa-Blanco, C., Bielza, C., & Larrañaga, P. (2023). Feature subset selection for data and feature streams: A review. Artificial Intelligence Review, 56, 1011–1062. https://doi.org/10.1007/s10462-023-10546-9

Tsai, C.-F., & Sung, Y.-T. (2020). Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. Knowledge-Based Systems, 203, Article 106097. https://doi.org/10.1016/j.knosys.2020.106097

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper/7062-a

Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Brahim, A. B., & Limam, M. (2018). Ensemble feature selection for high dimensional data: A new method and a comparative study. Advances in Data Analysis and Classification, 12, 937–952. https://doi.org/10.1007/s11634-017-0285-y

Xiong, H., Li, X., Zhang, X., Chen, J., Sun, X., Li, Y., Sun, Z., & Du, M. (2024). Towards explainable artificial intelligence (XAI): A data mining perspective. arXiv. https://arxiv.org/abs/2401.04374

Azevedo, R. C., Araújo, R. A., & Oliveira, A. L. I. (2024). Hybrid approaches to optimization and machine learning methods: A systematic literature review. Machine Learning, 113, 4055–4097. https://doi.org/10.1007/s10994-023-06467-x

[1] Department of Computer Science, FAST National University of Computer and Emerging Sciences, Islamabad, Pakistan.
[2] Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan.

Copyright (c) 2024 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.