Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28250

Comparative Analysis of SelectKBest and RFE-CV for Feature Selection Based on Random Forest: A Case Study of Student Performance Prediction

Yifei Zhou

School of Software Engineering, Sichuan University, Chengdu, China
zyf122809690@gmail.com

Abstract. Machine learning has been widely employed in education to predict students' learning-related challenges. Meanwhile, feature selection plays a critical role in enhancing model efficiency. During dataset training, irrelevant features often contribute minimally to outcomes and may introduce noise. Eliminating such features reduces data storage redundancy and accelerates model training. This study investigates scientific methods to partition features when prior experience cannot intuitively determine feature importance. Using a dataset of student study habits—including academic behaviors, lifestyle factors, grades, parental information, and socioeconomic status (12 features total), the study compares SelectKBest and RFE-CV feature selection methods within a Random Forest model. Performance is evaluated using R^2 and MSE metrics. Results demonstrate that RFE-CV outperforms SelectKBest, significantly reducing feature dimensionality while preserving prediction accuracy.

Keywords: Random Forest, RFE-CV, MSE

1. Introduction

In the field of machine learning, feature selection is critical for enhancing model efficiency and prediction accuracy. In high-dimensional data, irrelevant or noisy features not only increase computational overhead (e.g., memory usage, training time) but may also introduce redundant information, leading to model overfitting. Particularly in educational and social science research, datasets often contain mixed-type features (e.g., numerical grades, categorical parental education levels, continuous sleep frequency), where nonlinear relationships and interactions between features are more complex [1]. Traditional experience-driven feature selection methods are no longer sufficient [2].

Recent advancements have been made in feature selection methodologies. Early research primarily focused on Filter methods (e.g., SelectKBest), which use statistical tests (e.g., chi-square, F-score) to to identify features highly correlated with the target variable quickly. Despite the continued development of feature selection methods, applied research in educational data analysis has primarily focused on traditional machine learning tasks (such as image recognition and natural language processing). The applicability and effectiveness of different feature selection methods for

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28250

educational data, which has mixed feature types, remains unclear. In particular, in the task of predicting student academic performance, there is a lack of systematic comparative research on how to balance reducing the number of features while maintaining prediction accuracy.

Therefore, this study uses a student academic dataset to compare two feature selection methods (SelectKBest and RFE-CV) in a Random Forest model, aiming to minimize the number of features while maximizing prediction accuracy [3]. The findings are expected to provide methodological guidance for educational data preprocessing, helping researchers identify key influencing factors more efficiently.

2. Literature review and theoretical framework

2.1. Filter methods

Filter feature selection is a model-independent feature evaluation 4method. Its core principle involves directly assessing the relevance or importance of each feature to the target variable through statistical metrics or information-theoretic measures, independent of subsequent machine learning models. This approach offers advantages, including high computational efficiency, reduced risk of overfitting, and suitability for high-dimensional data preprocessing stages.

As a classic implementation of filter methods in the Scikit-learn library, SelectKBest selects the K features with the highest scores via scoring functions and ranking. Recognized as one of the most prominent feature selection techniques, SelectKBest has been widely applied across multiple domains.

The chi-square test is used to quickly screen CT imaging features with statistical distinguishing power to assist in the differential diagnosis of common disease [4]. Wang et al. used feature selection methods to accurately screen out the key factors affecting the balanced allocation of educational resources [5]. Dimensionality reduction of high-dimensional gene expression data based on statistical tests [6]. In agriculture, scholars select oil microelement features with higher relevance using F-score evaluations [7]. They also filter abnormal features from transaction data (e.g., transaction amount deviation, user behavior statistics) for fraud detection [8]. These studies collectively demonstrate the fundamental and practical nature of statistical feature selection methods in interdisciplinary data-driven research.

2.2. Wrapper methods

Wrapper methods directly link feature selection to model performance, using prediction accuracy as the evaluation criterion for feature subsets. The core principle involves iteratively generating feature subsets, training models, and evaluating performance to to select the optimal subset ultimately [9].

For example, Recursive Feature Elimination (RFE) optimizes model performance by progressively removing unimportant features. Its logical workflow comprises three steps: Baseline Training, Feature Pruning, and Iterative Refinement.

2.3. Random Forest

Random Forest (RF), formally proposed by Leo Breiman and Adele Cutler in 2001, constructs multiple decision trees through Bootstrap sampling and random feature selection, leveraging Ensemble Learning to enhance the model's generalization capability [10]. Its dual randomness (in data and features) effectively mitigates the overfitting risk inherent in individual decision trees.

DOI: 10.54254/2755-2721/2025.LD28250

Random forests can be used not only as a prediction model but also as an embedded feature selection framework. They output feature importance scores, assessing the contribution of each variable to the prediction. Therefore, random forests are often used as an effective tool for identifying key features in practical applications, and are particularly suitable for feature selection and explanatory analysis in high-dimensional data.

3. Methodology

3.1. Data collection and preparation

This experiment utilizes the publicly available dataset 'student study habits', which contains academic performance records and 12 features (including study_hours_per_week, attendance_percentage, participation_level_Low, etc.) [3].

Data analysis shows that the dataset comprises information from a total of 500 students, with student grades following a normal distribution. Furthermore, analysis reveals that no high correlation or redundancy exists among the other 12 features. The dataset is divided into training and testing sets, with 70% used for training and 30% for testing, to ensure the model's generalization capability.

3.2. Model design

The random forest model, as a predictive model, can predict potential target values based on input features after training. In this study, the random forest model is employed as the baseline modes. The feature selection method used in this experiment is as follows:

SelectKBest: Uses chi-square as the criterion for feature relevance to rank features (applicable to continuous target variables) and selects the highest scores.

RFE-CV: Through recursive feature elimination and cross-validation, determine the optimal feature subset by gradually eliminating the least important features.

The top 6 ranked features are retained as the optimal number of features.

4. Experiment

4.1. Experimental procedure

A random forest model is employed as the baseline model, with the corresponding R^2 and MSE values recorded as baseline metrics.

The SelectKBest method is applied, using chi-square as the evaluation criterion to select the top 6 significant features. A random forest model is then trained using these 6 features (referred to as the SelectKBest model), and the performance is evaluated.

Subsequently, the RFE-CV method is implemented, which recursively combines with cross-validation to identify the optimal feature combination (retaining 6 features) for model performance. The selected features are used to train another random forest model (the RFE-CV model),.

Through these steps, the effectiveness of the two feature selection methods is quantitatively evaluated and compared with the baseline model.

DOI: 10.54254/2755-2721/2025.LD28250

4.2. Evaluation metrics

This study employs two critical metrics to assess the performance of random forest models under different feature selection methods:

R^2 (Coefficient of Determination): R^2 indicates the proportion of variance in the dependent variable that the independent variables can explain. A value closer to 1 signifies stronger model explanatory power.

MSE (Mean Squared Error): MSE quantifies the average squared difference between predicted and actual values. A smaller MSE indicates reduced prediction errors.

4.3. Results

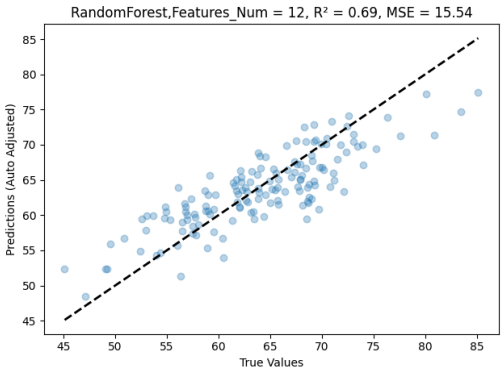


Figure 1. Baseline model

As illustrated in Figure1, the baseline Random Forest model using all 12 features achieved an R^2 of 0.69 and MSE of 15.54. The scatter plot shows the relationship between predicted and actual values. The data points are distributed around the diagonal line, indicating reasonable prediction accuracy for the baseline model.

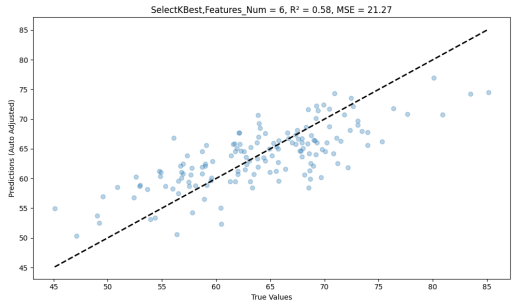


Figure 2. SelectKBest selection model

Figure 2 shows the performance of the SelectKBest feature selection method. The divergence around the prediction line has increased ($R^2$of 0.58 and MSE of 21.27), indicating that the prediction accuracy of the SelectKBest feature subset has decreased.

DOI: 10.54254/2755-2721/2025.LD28250

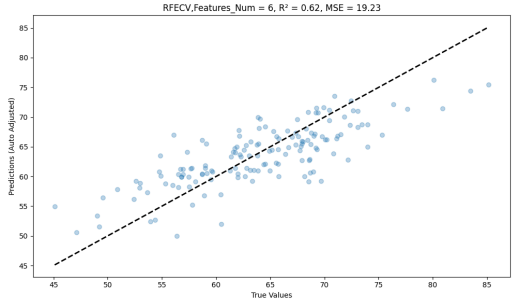


Figure 3. RFE-CV selection model

The results of the RFE-CV method are shown in Figure 3. The final result has an R^2 of 0.62 and an MSE of 19.22. Its predictions are better aligned than those in Figure 2, and the data points are less scattered around the diagonal.

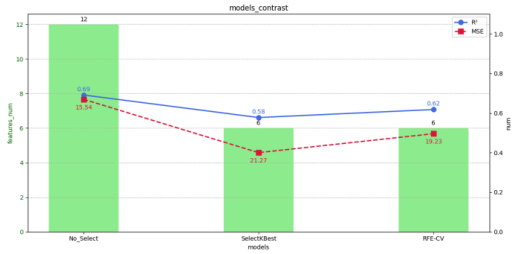


Figure 4. Comparison of three different data segmentation methods

As shown in Figure 4 both the SelectKBest and RFE-CV selection methods reduce the dimensionality from 12 to 6 features, with minimal impact on model performance. Compared to the baseline model, the RFE-CV method exhibits a smaller influence than SelectKBest, with differences of only 0.07 (R^2) and 3.69 (MSE). This demonstrates that the RFE-CV has superior feature subset selection capabilities.

5. Discussion

5.1. Summary

The RFE-CV feature selection method demonstrated optimal performance on the student information dataset, attributable to the interplay between data characteristics and methodological properties. The dataset's mixed-type features (e.g., categorical "parental_education" and continuous "study_hours_per_week") enabled RFE-CV's recursive elimination to effectively capture nonlinear relationships (e.g., "internet access" and "sleep duration" synergies). Concurrently, cross-validation mitigated noise interference, which was particularly critical for handling complex, confounding features like "extracurricular activities" and "work experience" in educational research. This dual advantage allowed RFE-CV to align with real-world multifactorial interactions, reducing redundancy while preserving key information. The proposed framework provides a reproducible methodology for educational data preprocessing, particularly valuable in scenarios requiring balanced consideration of model performance and interpretability, such as educational policy impact assessments.

DOI: 10.54254/2755-2721/2025.LD28250

5.2. Limitations

This research also has several limitations. First, the original dataset is constrained by a relatively small sample size of students, which may lead to overfitting and limited generalizability of the model. Additionally, the target variable (academic performance) is based on a single subject, potentially introducing larger errors. It ignores key contextual factors such as parental information and socioeconomic status, which may affect the prediction results. Future research should employ larger datasets or utilize augmentation techniques to expand the sample size [11].

Regarding the model, the random forest algorithm may be less effective than deep learning models in handling certain complex data patterns, and its performance is dependent on the assumptions and constraints of decision trees, which may not adapt to all types of data distributions. The feature selection methods also have limitations. Student achievement usually depends on a complex interaction among all characteristics. SelectKBest fails to capture inter-feature interactions, by comparison, RFE-CV incorporates recursive feature elimination but at the expense of higher computational complexity.

6. Conclusion

To examine the effectiveness of feature selection methods in education, this study selected two methods (SelectKBest and RFE-CV) and validated their performance in a random forest model. The experimental results showed that RFE-CV achieved significantly better performance when retaining six key features (R^2=0.62, MSE=19.23) compared to alternative approaches. The method's distinctive advantage lies in its ability to handle mixed feature types through cross-validated iterative optimization, which simultaneously captures nonlinear feature interactions while reducing bias from single-model dependency in feature ranking.

However, this study has certain limitations. It only shows two feature selection methods, which cannot fully reflect the complexity of student feature relationships. When selecting sample data, the variable sample setting is relatively simple. Random forests are not as effective as deep learning models (e.g., RNN, Transformer) in capturing complex nonlinear patterns, especially for sequential data such as students' long-term learning trajectories. Future research directions should focus on integrating RFE-CV with deep learning-based feature importance evaluation methods (e.g., PAC). .

References

[1] Arian, H., Norouzi Mobarekeh, D., & Seco, L. (2024). Backtest overfitting in the machine learning era: A comparison of out-of-sample testing methods in a synthetic controlled environment. Knowledge-Based Systems, 305, 112477

[2] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. Frontiers in Bioinformatics, 2, 927312.

[3] Preksha, D. (2023). Student Study Habits. Kaggle. https: //www.kaggle.com/datasets/prekshad2166/student-study-habits

[4] Zhang, R., Xu, Y., & Li, W. (2024). Feasibility study of CT radiomics in differentiating suppurative and tuberculous lymphadenitis in children. Chinese Journal of Medical Imaging, 32(5), 798–805.

[5] Wang, H., & Li, H. (2023). Research on balanced allocation of regional educational resources based on feature selection. Educational Research, 45(3), 112–120.

[6] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A Review of Feature Selection Techniques in Bioinformatics. Bioinformatics. 23(19), 2507-2517.

[7] Jamei, M., Ali, M., Afzaal, H., Karbasi, M., Malik, A., Farooque, A. A., Haydar, Z., & Zaman, Q. U. (2024). Accurate monitoring of micronutrients in tilled potato soils of eastern Canada: Application of an explainable

DOI: 10.54254/2755-2721/2025.LD28250

inspired-adaptive boosting framework coupled with SelectKbest. Computers and Electronics in Agriculture, 216, 108479

[8] Zhang, W., & Chen, M. (2022). Optimization of a credit card fraud detection model based on feature selection. Journal of Financial Research, 59(4), 89–97.

[9] Wang, J., Cao, D., Lan, X., & Gao, J. (2018). Filter-Wrapper feature variable selection method for yield prediction of catalytic cracking unit. CIESC Journal, 69(1), pp. 464–471,

[10] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. DOI: 10.1023/A: 1010933404324

[11] Lee, N., Wattanawong, T., Kim, S., et al. (2024). (2024) LLM2LLM: Boosting LLMs with Novel Iterative Data Enhancement. https: //arxiv.org/abs/2403.15042.