stiti_Aditya_-_Grade_Decision_tree_Feature_selection_XGBoost.pdf

by Stuedent 1

Submission date: 28-Apr-2024 08:58PM (UTC+0700)

Submission ID: 2364202315

File name: stiti_Aditya_-_Grade_Decision_tree_Feature_selection_XGBoost.pdf (438.78K)

Word count: 4843

Character count: 27429

Scientific Journal of Informatics
Vol. 11, No. 1, Feb 2024

p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Implementation of Feature Selection Strategies to Enhance Classification Using XGBoost and Decision Tree

Fhara Elvina Pingky Nadya[1], M. Firdaus Ibadi Ferdiansyah[2], Vinna Rahmayanti Setyaning Nastiti[3]*, Christian Sri Kusuma Aditya[4]

[1,2,3,4]Departmen of Informatics, Faculty of Engineering, Universitas Muhammadiyah Malang, Indonesia

Abstract.

Purpose: Grades in the world of education are often a benchmark for students to be considered successful or not during the learning period. The facilities and teaching staff provided by schools with the same portion do not make student grades the same, the value gap is still found in every school. The purpose of this research is to produce a better accuracy rate by applying feature selection Information Gain (IG), Recursive Feature Elimination (RFE), Lasso, and Hybrid (RFE + Mutual Information) using XGBoost and Decision Tree models.

Methods: This research was conducted using 649 Portuguese course student data that had been pre-processed according to data requirements, then, feature selection was carried out to select features that affect the target, after that all data can be classified using XGBoost and Decision tree, finally evaluating and displaying the results.

Results: The results showed that feature selection Information Gain combined with the XGBoost algorithm has the best accuracy results compared to others, which is 81.53%.

Novelty: The contribution of this research is to improve the classification accuracy results of previous research by using 2 traditional machine learning algorithms and some feature selection.

Keywords: Grade, Decision tree, Feature selection, XGBoost
Received October 2023 / **Revised** February 2024 / **Accepted** February 2024

INTRODUCTION

The success of a country is often gauged by indicators such as education and the economy. Despite their significance, these factors present challenges, especially in developing nations, where education is crucial for addressing issues like poverty and wielding transformative power on individuals, societies, and nations [1]. Recognizing the pivotal role of education, the Indonesian government introduced the 12-year mandatory education program (WAJAR) to improve equal access to quality education [2]. While these initiatives are commendable, grading remains a universal criterion, for advancing to higher education [3].

Various factors, particularly those originating from students, can influence academic performance [4]. Acknowledging the importance of academic monitoring, educators must actively manage it to improve quality and performance [5]. Early identification of potential academic challenges requires teachers to discern factors contributing to students' struggles [6]. Technological advancements facilitate the swift identification of these factors using data mining [7]. Educational Data Mining (EDM) is a burgeoning field that enables researchers to extract valuable insights or patterns from extensive datasets, minimizing decision-making risks in education [8]–[12]. Although relatively new, EDM has been widely used due to its potential to help educators and institutions utilize data-driven insights for more efficient operational processes and extraction of new knowledge from large student data sets [13], [14].

Previous researchers have conducted various studies on the implementation of Educational Data Mining (EDM). Portuguese dataset student in previous research [15] using the same dataset classifying using the boosting algorithm has resulted in the first scenario using 10-fold cross-validation with RPART is 76.64% on Portuguese data which is higher than C5.0, M1, and SAMME which have values of 69.09%, 74.53%, and 71.97%. Further research conducted by Ferda Ünal using the same dataset on Portuguese data by

[*] Corresponding author.
Email addresses: vinastiti@umm.ac.id (Nastiti)

DOI: [10.15294/sji.v11i1.48145](#)

classifying using Wrapper type feature selection with several algorithms resulted in Random Forest combined feature selection wrapper having the highest accuracy of 77.20% compared to J48 and Naive Bayes with an accuracy of 74.88% and 72.57% [16].

In addition, some researchers only look for which variables are most influential on student academic grades. For example, research conducted by S. Rajendran et al [17] focused on identifying parameters that affect students' academic grades. It was found that health-conscious lifestyle and stress had a positive correlation with academic performance. To determine the influential parameters, the study used feature importance from several machine learning algorithms, such as ANN, random forest, gradient boosting, and stacking algorithms which showed that lifestyle factors (physical activity and optimistic thinking) became the main feature that had a relative feature value of more than 75 in influencing academic performance. Then, stress becomes a significant feature, especially in gradient boosting and stacking with relative features that almost reach 100. Other researchers such as Fernandes et al [18] also conducted similar research to determine the factors that influence student academics using the Gradient Boosting Machine (GBM) algorithm importance features took some features that have an importance scale of more than 0.35 and got the results that grades, attendance, environment, school, and age are indicators that have the potential to determine academic achievement. Both previous studies used 2 different data. However, this informs us that there are quite a lot of features from both inside and outside the individual student that affect academic performance. This shows that feature selection in data processing is an important thing to do, because even though data has many features, in reality, not all of these features are important features. Some features can be removed to simplify the analysis without reducing the accuracy value of a classification.

Feature selection itself is not new in data processing. For example, research conducted by Fathania Firwan F et al [19] on classification approaches for heart disease prediction said that there are several feature selection methods namely Filter, Wrapper, Embedded, and Hybrid with the advantages and disadvantages of each method. Feature selection itself is a knowledge technique to find a subset of the original feature set that efficiently represents the input data while reducing the impact of noise and irrelevant features but still provides relatively excellent results for the task and helps analysts obtain good classification performance [20]. There are also several reasons why feature selection is important, namely reducing the number of parameters, reducing training time, and minimizing over-fitting problems [21].

Although some studies above have made valuable contributions related to the classification of the Portuguese student dataset, there is a significant research gap. Previous research only considered one type of feature selection, namely wrapper, without exploring the potential benefits of other feature selection methods. Therefore, this study attempts to fill this gap by utilizing four types of feature selection - filter, wrapper, embedded, and hybrid - using XGBoost and Decision Tree models, because in tabular data classification Decision Tree remains the best choice in terms of performance and training time. According to S. Fayaz et al and V. Borisov et al [22], [23], in their research, (Gradient Boosted Decision Tree) GBDT still dominates and shows superior performance applied to tabular data, and the XGBoost algorithm is considered a recommended choice for real-life tabular data problems. In this context, the uniqueness of our research lies in combining correlation and importance-based feature selection methods. The most relevant features are selected to improve the accuracy of value classification. By prioritizing features that have a significant impact and selecting the optimal classification model, namely Decision Tree and XGBoost. This approach not only improves prediction accuracy but also results in a more efficient model for handling tabular data. Therefore, this research is expected to provide additional contributions to a practical understanding of how various feature selection methods can impact the performance of EDM classification models, thus enhancing the quality of education through more effective planning [24].

METHODS

The research process is illustrated in Figure 1, encompassing dataset preprocessing, data splitting, feature selection, and modeling. In this section, each step will be elaborated in detail in the next section to provide a comprehensive understanding of the research process.

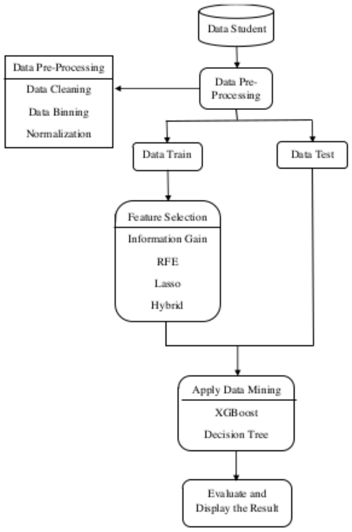

Figure 1. Research flowchart

Participants and datasets

Data taken from https://www.kaggle.com/datasets/uciml/student-alcohol-consumption. It consisted of 649 data with 33 attributes. These attributes included student information such as school (school), sex, age, address, family size (famsize), parental cohabitant status (Pstatus), education of mother (Medu) and father (Fedu), mother's (Mjob) and father's (Fjob) occupation, the reason for choosing school (reason), guardian, travel time from home to school (traveltime), weekly study time (studytime), number of failures in previous grade (failures), extra education support (schoolsup), family education support (famsup), paid extra classes (paid), extracurricular activities (activities), nursery attendance (nursery), desire for higher education (higher), internet access at home (internet), romantic relationships (romantic), quality of family relationships (famrel), free time after school (freetime), activities with friends (goout), alcohol consumption on weekdays (Dalc) and weekends (Walc), current health status (health), and number of school absences (absences). In addition, there are also values associated with the subject of Portuguese, namely G1 (first-period grade), G2 (second-period grade), and G3 (final grade as output target).

Preprocessing data

Preprocessing plays a crucial role in data processing, addressing issues like noise, outliers, and irrelevant attributes in raw data. Data cleaning, particularly handling outliers, is a pivotal step. The Winsorize technique is employed in this dataset for outlier handling, replacing extreme values with the nearest values [25], [26]. Binning of G3 data follows the Portuguese higher education system, categorizing it into ranges like 0-9, 10-13, 14-15, 16-17, 18-19, and 20 [14]. Dataset splitting divides the data into training (80%) and testing (20%) subsets. Data balancing through SMOTE overcomes imbalance issues effectively by generating synthetic minority class samples [27]. Feature selection aims to obtain a minimal, informative subset, excluding irrelevant or highly correlated features [28]. The process involves selecting k-best features, ranging from 5 to 15 features with an increment of 5, denoted as $k \in \{5, 10, 15\}$ [29]. Finally, data normalization, utilizing Z-Score transformation, ensures consistent ranges between data points [27].

Feature selection

Feature selection has several categories, including filtering, wrapping, embedded, and hybrid. In this research, all categories are used to find out which feature selection is most suitable for the dataset. The filtering category has a way of selecting variables based on rank. So, if there is a variable that is below the threshold, it will be eliminated so that in the end the relevant features are obtained. In this case, filtering uses the Information Gain (IG) algorithm which has a working system that ignores features that have little IG value because these features are not very influential on accuracy results or are arguably irrelevant features. In addition, IG also has advantages, such as increasing effectiveness and accuracy, and can also reduce complexity [19], [30], [31].

Then for wrapping, there is Recursive Feature Elimination (RFE) is a feature selection method that aims to identify the most suitable feature subset by utilizing the learned model and classification accuracy. RFE falls under wrapping because it uses a supervised methodology and is wrapped iteratively to remove the worst features based on the target [28], [32]. For embedded in the process of feature selection, choose features Lasso that have non-zero coefficients after applying shrinkage, while those with exactly zero coefficients are excluded. The tuning parameter, also known as the regularization parameter (λ), is used to control the degree of regularization. There are several benefits of using Lasso, such as helping prevent over-fitting problems, resulting in better generalization, and improving interpretability by canceling irrelevant features [33], [34]. Finally, hybrids include mixtures of many feature options. This study employed a mix of Mutual Information and RFE.

XGBoost algorithm

XGBoost stands for Extreme Gradient Boost which completes the learning task by building or combining several weak learning models to become a strong learning model iteratively [35], [36]. This is a simplified group calculation depending on the GBDT (Gradient Boosted Decision Tree). The premise of improving computation is that multiple decision trees perform superior to a single one. Any decision tree can make for a terrible show. When multiple trees are combined, the presentation shows signs of improvement.

In this experiment, the input data is in the form of the final attribute that has been selected above. The XGboost formula can be seen in formulas (1) and (2).

$$L_{xgb} = \sum_{i=1}^{N} L\big(y_i, F(X_i)\big) + \sum_{m=1}^{M} \Omega(h_m) \tag{1}$$

$$\Omega(h) = \gamma T + \frac{1}{2}\gamma ||w||^2 \tag{2}$$

Where T is the number of leaves on the tree and w is the output score of the leaves. A higher γ value will result in a simpler tree. The value γ controls the minimum loss reduction gain required to divide the internal nodes [37].

Decision tree algorithm

Decision tree is one of the popular and effective algorithms in data mining, especially for classification problems. In the context of education, using a Decision Tree classifier can help improve education by identifying patterns and factors that contribute to student grades. With a good understanding of this method, using a decision tree classifier can make an important contribution to improving education through better data analysis and decision-making. The process of making a decision tree is, first the entropy class of each attribute is calculated, and then all the information gained as in the formulas (3) until (5) below.

$$info\ (D) = -\sum_{i=1}^{m} p_i * \log_2(p_i) \tag{3}$$

$$info_A\ (D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * info\big(D_j\big) \tag{4}$$

$$Gain(A) = info\ (D) - info_A\ (D) \tag{5}$$

Evaluation

In the context of assessing model performance, a commonly employed method is the utilization of a confusion matrix. The confusion matrix provides a detailed breakdown of predicted classifications versus actual classifications, forming the foundation for further evaluation. Within the evaluation subsection, key metrics such as precision, recall, and accuracy can be incorporated. Precision is used to find out the true positive predictions for the overall results predicted. The formula for calculating the precision value is in equation (6) [38].

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

Recall is the ratio used to compare true positive predictions with the sum of true positives and false negatives. The formula for calculating the recall value is in equation (7).

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

Accuracy is used to see the ratio that is correctly predicted with all data using the formula in equation (8) [39].

$$Accuracy = \frac{TP+TN}{TP+FP+FN+FP} \tag{8}$$

RESULTS AND DISCUSSIONS

The experimental analysis conducted involved evaluating the performance of XGBoost and Decision Tree models by considering four feature selection methods (IG, RFE, Lasso, and Hybrid) while varying the value of K as the number of top features selected (K=5, 10, 15). The graph below provides visual insights into how the accuracy of the models evolves with changes in the value of K.

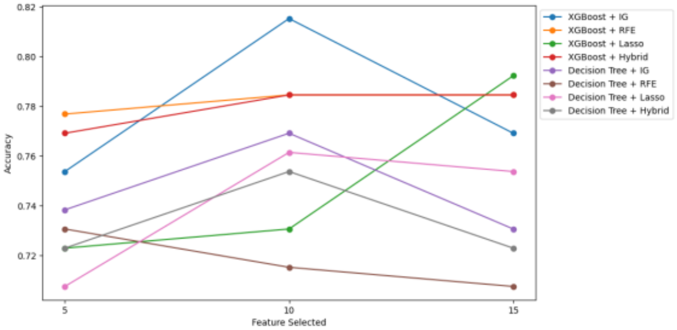

Figure 2. Relationship between the number of features selected and accuracy

Figure 2 reflects differences in the model's performance as the value of K varies. Increases in accuracy are noticeable at specific values of K, indicating that proper feature selection can contribute positively to the model's performance. After thorough analysis, the results indicate that the optimal values of K differ for both models and each feature selection method, leading to unique characteristics and varying key metric values. Consequently, Table 1 is presented to provide a detailed breakdown of the highest precision, recall, and accuracy for each model and feature selection method.

Table 1. The result of XGBoost and Decision Tree before and after using feature selection

Algorithm	Feature selection	Precision (%)	Recall (%)	Accuracy (%)
XGBoost	All features	75.65	76.15	76.15
XGBoost	IG	79.28	81.53	81.53
XGBoost	RFE	78.49	78.46	78.46
XGBoost	Lasso	78.10	79.23	79.23
XGBoost	Hybrid	78.49	78.46	78.46
Decision Tree	All features	66.11	64.61	64.61
Decision Tree	IG	78.01	76.92	76.92
Decision Tree	RFE	74.45	73.07	73.07
Decision Tree	Lasso	76.92	76.15	76.15
Decision Tree	Hybrid	76.39	75.38	75.38

Table 1 provides a comprehensive overview of the highest performance for each model with the selected optimal values of K, aiming to enhance classification performance by choosing attributes aligned with the target and reducing complexity. Notably, the XGBoost algorithm, when coupled with Information Gain, attains the highest accuracy at 81.53%. This success can be attributed to several factors. Firstly, XGBoost is adept at comprehending intricate data relationships through the amalgamation of numerous small learners. Its design prevents overfitting to training data, enhancing its generalization to new data [40], [41]. Information Gain further refines the model by selecting pivotal features and directing its attention to essential components [42]. The ability of XGBoost to handle interrelated features and complex patterns contributes significantly to its robust performance across diverse data types. The synergy between XGBoost and Information Gain renders the model both resilient and accurate. Complementary to this, Table 2 outlines the features selected through the feature selection method, as a basis for reducing redundancy.

Table 2. Feature selection

Feature Selection Method	K-Best	Feature Selected
Information Gain	10	age, Medu, Fedu, Mjob, studytime, failures, freetime, goout, G1, G2
RFE	10	Medu, reason, guardian, studytime, freetime, Walc, health, absences, G1, G2
Lasso	15	school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, failures, higher, G1, G2
Hybrid	10	Medu, reason, guardian, studytime, freetime, Walc, health, absences, G1, G2

The feature selection results in Table 2 indicate that the significant attributes for the classification of final grades involve variables such as age, Medu, Fedu, Mjob, studytime, failures, freetime, goout, as well as the grades G1 and G2. This step is crucial in the context of accuracy comparison with previous research, as demonstrated in Table 3 and Figure 3.

Table 3. Comparison of research results with other studies

Research	Algorithm	Accuracy
Wrapper feature selection [15]	Random Forest	77.20%
Information Gain feature selection	XGBoost	81.53%

In Table 3, the implementation of XGBoost and Information Gain (IG) on the student dataset demonstrated a notable accuracy of 81.53%, surpassing other methods, including Random Forest and Wrapper feature selection. This superiority is attributed to XGBoost's inherent capability to capture complex patterns and the pivotal feature selection by Information Gain, as discussed earlier. It is crucial to note that the Random Forest method in the comparison table also applied to the same dataset.

To reinforce the reliability of our findings, it is worth mentioning that prior research on heart disease datasets employed a consistent methodology XGBoost and IG resulting in an impressive accuracy of 93.44%, despite the distinct nature of the datasets [40]. This consistency underscores the reliability and robustness of the XGBoost with the IG approach. The effectiveness of XGBoost with IG for the classification of student grades offers practical implications for educational support, supported by the method's consistent success across diverse datasets.

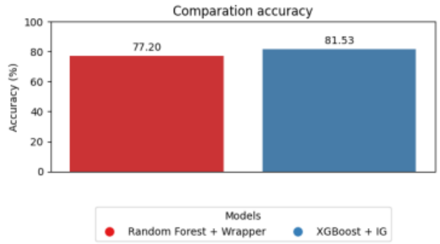

Figure 3. Comparison accuracy

The effectiveness of XGBoost with Information Gain (IG) in the classification of students' grades in Table 3 and Figure 3 presents practical implications for educational support, bolstered by the consistent success of the method across diverse datasets. Moreover, the integration of sophisticated techniques like these into educational practices has the potential to significantly improve the quality of education. By harnessing insights provided by XGBoost and Information Gain, educators can adapt their teaching strategies, identify at-risk students, and implement targeted interventions, thereby fostering a more effective and personalized learning environment. The application of this methodology aligns with the common trend of leveraging data-driven approaches to enhance educational outcomes, underscoring the importance of embracing technological advancements for educational improvement.

CONCLUSION

In response to the challenges of assessment in education, this research focused on the classification of Portuguese final grades (G3). The results confirmed that using the XGBoost algorithm with Information Gain (IG) feature selection provided the best performance with an accuracy of 81.53%. The implication is that this grade classification system can effectively help teachers analyze students who need special attention before exams to achieve optimal results. For future research, it is recommended to consider using the latest datasets and explore deep learning methods, such as neural networks, to improve accuracy with the ability to capture more complex patterns.

REFERENCES

[1] S. Li and T. Liu, “Performance Prediction for Higher Education Students Using Deep Learning,” *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/9958203.

[2] D. O. T. Aritonang, “The Efforts to Improve the Quality of Education in North Tapanuli Regency,” *Int. J. English Lit. Soc. Sci.*, vol. 3, no. 6, pp. 1154–1159, 2018, doi: 10.22161/ijels.3.6.30.

[3] S. G. Sireci and S. Greiff, "Editorial: On the importance of educational tests," *Eur. J. Psychol. Assess.*, vol. 35, no. 3, pp. 297–300, 2019, doi: 10.1027/1015-5759/a000549.

[4] B. K. Francis and S. S. Babu, "Predicting Academic Performance of Students Using a Hybrid Data Mining Approach," 2019.

[5] P. Dabhade, R. Agarwal, K. P. Alameen, A. T. Fathima, R. Sridharan, and G. Gopakumar, "Educational data mining for predicting students' academic performance using machine learning algorithms," *Mater. Today Proc.*, vol. 47, no. xxxx, pp. 5260–5267, 2021, doi: 10.1016/j.matpr.2021.05.646.

[6] J. López-Zambrano, J. A. L. Torralbo, and C. Romero, "Early prediction of student learning performance through data mining: A systematic review," *Psicothema*, vol. 33, no. 3, pp. 456–465, 2021, doi: 10.7334/psicothema2021.62.

[7] A. Khan and S. K. Ghosh, *Student performance analysis and prediction in classroom learning: A review of educational data mining studies*, vol. 26, no. 1. Education and Information Technologies, 2021. doi: 10.1007/s10639-020-10230-3.

[8] S. Khademizadeh, Z. Nematollahi, and F. Danesh, "Analysis of book circulation data and a book recommendation system in academic libraries using data mining techniques," *Libr. Inf. Sci. Res.*, vol. 44, no. 4, p. 101191, 2022, doi: 10.1016/j.lisr.2022.101191.

[9] R. Ordoñez-Avila, N. Salgado Reyes, J. Meza, and S. Ventura, "Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review," *Heliyon*, vol. 9, no. 3, 2023, doi: 10.1016/j.heliyon.2023.e13939.

[10] J. Klimek and J. A. Klimek, "IT and data mining in decision-making in the organization. Education management in the culture of late modernity," *Procedia Comput. Sci.*, vol. 176, pp. 1990–1999, 2020, doi: 10.1016/j.procs.2020.09.235.

[11] D. Hooshyar, M. Pedaste, and Y. Yang, "Mining educational data to predict students' performance through procrastination behavior," *Entropy*, vol. 22, no. 1, p. 12, 2020, doi: 10.3390/e22010012.

[12] Y. S. Mitrofanova, A. A. Sherstobitova, and O. A. Filippova, *Modeling smart learning processes based on educational data mining tools*, vol. 144. Springer Singapore, 2019. doi: 10.1007/978-981-13-8260-4_49.

[13] G. Ramaswami, T. Susnjak, A. Mathrani, J. Lim, and P. Garcia, "Using educational data mining techniques to increase the prediction accuracy of student academic performance," *Inf. Learn. Sci.*, vol. 120, no. 7–8, pp. 451–467, 2019, doi: 10.1108/ILS-03-2019-0017.

[14] A. Abu Saa, M. Al-Emran, and K. Shaalan, *Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques*, vol. 24, no. 4. Springer Netherlands, 2019. doi: 10.1007/s10758-019-09408-7.

[15] F. Jauhari and A. A. Supianto, "Building student's performance decision tree classifier using boosting algorithm," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, pp. 1298–1304, 2019, doi: 10.11591/ijeecs.v14.i3.pp1298-1304.

[16] F. Ünal, *Data Mining - Methods, Applications and Systems*. IntechOpen, 2020.

[17] S. Rajendran, S. Chamundeswari, and A. A. Sinha, "Predicting the academic performance of middle- and high-school students using machine learning algorithms," *Soc. Sci. Humanit. Open*, vol. 6, no. 1, p. 100357, 2022, doi: 10.1016/j.ssaho.2022.100357.

[18] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. Bus. Res.*, vol. 94, no. August 2017, pp. 335–343, 2019, doi: 10.1016/j.jbusres.2018.02.012.

[19] F. F. Firdaus, H. A. Nugroho, and I. Soesanti, "A Review of Feature Selection and Classification Approaches for Heart Disease Prediction," *IJITEE (International J. Inf. Technol. Electr. Eng.*, vol. 4, no. 3, p. 75, 2021, doi: 10.22146/ijitee.59193.

[20] A. K. Shukla, P. Singh, and M. Vardhan, "Predicting Alcohol Consumption Behaviours of the Secondary Level Students," *SSRN Electron. J.*, 2018, doi: 10.2139/ssrn.3170173.

[21] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.

[22] S. A. Fayaz, M. Zaman, S. Kaul, and M. A. Butt, "Is Deep Learning on Tabular Data Enough? An Assessment," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 4, pp. 466–473, 2022, doi: 10.14569/IJACSA.2022.0130454.

[23] V. Borisov, T. Leemann, K. Sessler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–21, 2022,

doi: 10.1109/TNNLS.2022.3229161.

[24] Y. Zhang, Y. Yun, R. An, J. Cui, H. Dai, and X. Shang, “Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis,” *Front. Psychol.*, vol. 12, no. December, pp. 1–19, 2021, doi: 10.3389/fpsyg.2021.698490.

[25] J. H. Sullivan, M. Warkentin, and L. Wallace, “So many ways for assessing outliers: What really works and does it matter?,” *J. Bus. Res.*, vol. 132, no. May, pp. 530–543, 2021, doi: 10.1016/j.jbusres.2021.03.066.

[26] T. Nyitrai and M. Virág, “The effects of handling outliers on the performance of bankruptcy prediction models,” *Socioecon. Plann. Sci.*, vol. 67, no. August, pp. 34–42, 2019, doi: 10.1016/j.seps.2018.08.004.

[27] A. J. Mohammed, “Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 3161–3172, 2020, doi: 10.30534/ijatcse/2020/104932020.

[28] P. Barbiero, G. Squillero, and A. Tonda, “Predictable Features Elimination: An Unsupervised Approach to Feature Selection,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 13163 LNCS, pp. 399–412, 2022, doi: 10.1007/978-3-030-95467-3_29.

[29] D. V. Akman *et al.*, “K-Best Feature Selection and Ranking Via Stochastic Approximation,” *Expert Syst. Appl.*, vol. 213, no. September, 2023, doi: 10.1016/j.eswa.2022.118864.

[30] A. Thakkar and R. Lohiya, “Attack classification using feature selection techniques: a comparative study,” *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 1, pp. 1249–1266, 2021, doi: 10.1007/s12652-020-02167-9.

[31] P. Bhat and K. Dutta, “A multi-tiered feature selection model for android malware detection based on Feature discrimination and Information Gain,” *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9464–9477, 2022, doi: 10.1016/j.jksuci.2021.11.004.

[32] H. Jeon and S. Oh, “Hybrid-recursive feature elimination for efficient feature selection,” *Appl. Sci.*, vol. 10, no. 9, pp. 1–8, 2020, doi: 10.3390/app10093211.

[33] F. Li, L. Lai, and S. Cui, “On the Adversarial Robustness of LASSO Based Feature Selection,” *IEEE Trans. Signal Process.*, vol. 69, pp. 5555–5567, 2021, doi: 10.1109/TSP.2021.3115943.

[34] Y. Bouchlaghem, Y. Akhiat, and S. Amjad, “Feature Selection: A Review and Comparative Study,” *E3S Web Conf.*, vol. 351, pp. 1–6, 2022, doi: 10.1051/e3sconf/202235101046.

[35] K. H. Susheelamma and K. M. Ravikumar, “Student risk identification learning model using machine learning approach,” *Int. J. Electr. Comput. Eng.*, vol. 9, no. 5, pp. 3872–3879, 2019, doi: 10.11591/ijece.v9i5.pp3872-3879.

[36] W. Su *et al.*, “An XGBoost-Based Knowledge Tracing Model,” *Int. J. Comput. Intell. Syst.*, vol. 16, no. 1, 2023, doi: 10.1007/s44196-023-00192-y.

[37] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A Comparative Analysis of XGBoost,” no. February, 2019, doi: 10.1007/s10462-020-09896-5.

[38] Abdullah-All-Tanvir, I. Ali Khandokar, A. K. M. Muzahidul Islam, S. Islam, and S. Shatabda, “A gradient boosting classifier for purchase intention prediction of online shoppers,” *Heliyon*, vol. 9, no. 4, p. e15163, 2023, doi: 10.1016/j.heliyon.2023.e15163.

[39] H. T. T. Nguyen, L. H. Chen, V. S. Saravanarajan, and H. Q. Pham, “Using XG Boost and Random Forest Classifier Algorithms to Predict Student Behavior,” *2021 IEEE Int. Conf. Emerg. Trends Ind. 4.0, ETI 4.0 2021*, 2021, doi: 10.1109/ETI4.051663.2021.9619217.

[40] J. Yang and J. Guan, “A Study of Heart Disease Prediction Model Based on Smote-XGBoost Algorithm,” *Hans J. Data Min.*, vol. 12, no. 03, pp. 220–234, 2022, doi: 10.12677/hjdm.2022.123003.

[41] D. Tarwidi, S. R. Pudjaprasetya, D. Adytia, and M. Apri, “An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach,” *MethodsX*, vol. 10, no. December 2022, 2023, doi: 10.1016/j.mex.2023.102119.

[42] W. T. Astuti, M. A. Muslim, and E. Sugiharti, “The Implementation of The Neuro Fuzzy Method Using Information Gain for Improving Accuracy in Determination of Landslide Prone Areas,” *Sci. J. Informatics*, vol. 6, no. 1, pp. 95–105, 2019, doi: 10.15294/sji.v6i1.16648.

stiti_Aditya_-_Grade_Decision_tree_Feature_selection_XGBoost.pdf

ORIGINALITY REPORT

13%	8%	7%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

MATCH ALL SOURCES (ONLY SELECTED SOURCE PRINTED)

1%

★ Submitted to De Montfort University

Student Paper

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		