TELKOMNIKA Telecommunication, Computing, Electronics and Control
Vol. 19, No. 6, December 2021, pp. 1865~1871
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v19i6.21594

Comparison of the feature selection algorithm in educational data mining

Agung Triayudi, Iskandar Fitri
Department of Informatic, Universitas Nasional, Jakarta, Indonesia

Article Info

Article history:

Received May 11, 2020
Revised Apr 30, 2021
Accepted May 15, 2021

Keywords:

Classification
Decision
Educational data mining
Feature selection algorithm
Student academic

ABSTRACT

Student academic accomplishment is the foremost focus of every educational institution. In developing student achievement in educational institutions, the researchers finally created a new research area, namely educational data mining (EDM). How the feature selection (FS) algorithm works is by removing unrelated data from educational datasets; therefore, this algorithm can improve the classification performance managed in EDM techniques. This research presents an analysis of the performance of the FS algorithm from the student dataset. The results received from other FS algorithms and classifiers will help other researchers to gain some best combination regarding FS algorithms and the classification. Selecting features that are relevant for student forecast models is a sensitive problem to stakeholders in education because they must make decisions based on the results of the prediction models. For the future, our paper seeks to play a decisive part while developing quality concerning education, as well as guiding different researchers in conducting educational interventions.

Corresponding Author:

Agung Triayudi
Department of Informatic
Universitas Nasional
Jakarta, Indonesia
Email: agungtriayudi@civitas.unas.ac.id

1. INTRODUCTION

The most important aspects of building a strong segment of civilization are improvement within the quality of education [1]. Data stored under repositories of educational institutions play a crucial part in extracting deep and unusual trims to help each stakeholder of an educational manner [2]. Several methods were expecting to estimate students' educational accomplishments by creating a bright future for their students [3], [4]. Predicting student performance has continued to a topic that is quite hot within the scope of educational data mining (EDM). Data mining is the best choice used by researchers to analyze student performance [5]. Data mining techniques that mare often used in the processing of educational data today are named EDM [2]. EDM searches educational data to fully recognize student completion problems by adopting a variety of data mining techniques [6]. To assist educational institutions to organise education policies to increase the variety of education, EDM uses educational data manipulation techniques [7].

One of the foremost fields of EDM is foresight. Foresight and analysis of student educational achievement are required to student educational majority. Identification of determinants that affect students' educational accomplishment is a reasonably tricky analysis job [8]. Unique educational data includes a lot of unrelated data, including redundancy. Redundancy data can affect the results of predictions. However, we can decrease some redundancy and increase the relevancy of points without any waste regarding important data with the feature selection (FS) method [9].

The embedded method is a unique method for several learning algorithms given, and this method is also carried out in the training process in classification. The filter method depends on the common features of the practice data, and this method is carried out at the pre-processing step and does not depend on the educational algorithm. The wrapper method uses an educational algorithm to evaluate features [10]. Feature selection (FS) is one of the most productive and very dynamic fields of the analysis field in machine learning and data mining. The primary purpose of this FS is to select a subset through passing variable data. Also, that can improve some efficiency of predictions and reduce the complexity of the decisions acquired. In connection with the feature selection technique, the effectiveness of student achievement forecast models can be improved. FS Techniques can be group into three associations, namely: embedded, filters, and wrapper models [11].

Previously, much work was arranged to divine student achievement using separate FS techniques. Meanwhile, the latest research, the researchers used different feature selection techniques and classification combinations to create more effective forecast models [12]. The analysis is needed to recognize performance reviews in terms of predictive efficiency in conjunction with other feature selection algorithms among different classifications [13]. This paper is a step towards recognizing this forecast efficiency of various feature selection algorithms available in the meaning of the classification adopted in educational data.

2. RESEARCH METHOD

The purpose objective of this analysis is to assess the achievement of other feature selection algorithms on various classification algorithms using educational datasets. The association between various feature selection algorithms gives educational data miners a deep insight into the completion of several feature selection algorithms toward educational data. Therefore, the objectives regarding this analysis can be achieved, the educational dataset is obtained from a credible source; furthermore, another feature selection algorithm is applied to the dataset, which is not used in the dataset. Several classification algorithms are implemented utilizing the chosen feature selection algorithm, then decided to check the most reliable performance amongst all combinations implemented to the educational dataset. The foremost actions of this research will then be explained below.

2.1. Description of the dataset

The dataset used in this study consisted of 439 students and nine attributes in online and distance (ODL) University. In this paper, the primary purpose of utilizing the dataset is to distinguish the most suitable combination regarding the feature selection algorithm and classification to recognize each main special parts concerning educational achievement. In this paper, the primary purpose of utilizing the dataset is to distinguish the most suitable combination regarding the feature selection algorithm and classification to recognize each main special parts concerning educational achievement.

2.2. Experimental setup

Waikato environment for knowledge analysis (WEKA) utilized essentially a tool for data mining techniques. WEKA owns many sources of machine learning algorithms. Weka is an open-source software developed with the JAVA programming language, which provides facilities during improving machine learning techniques for data mining work, produced by the University of Waikato in New Zealand [14].

2.3. Feature selection algorithm and classification

This paper using six feature selection algorithms have been tested before, there are Cfs subset eval [15], Chi squared attribute eval [16], filtered attribute eval [17], gain ratio attribute eval [18], principal components [19], and relief attribute eval [20]. This paper also uses 15 different classification algorithms that have been tested through educational datasets, specifically Bayes net, Naïve Bayes, Naive Bayes updateable, multilayer perceptron, simple logistic, SMO, decision tree, JRip, OneR, PART, decision stump, J48, random forest, random tree, and REP tree [21]-[23].

3. RESULTS AND ANALYSIS

This analysis concentrates about the completion regarding several feature selection algorithms forward with the classification method. The effectiveness of this algorithm is included within the values of F-measure, recall, precision, and forecast efficiency (examples with the correct classification) [24], [25]. The completion of the six feature selection techniques implemented to the 15 classifications is described in Tables 1-6. All the tables are made definitely for the six feature selection techniques, and then every table comprises four columns. The columns present the name of the classification algorithm, the F-measure value, the recall value, and the precision value utilizing the feature selection algorithm.

3.1. Cfs subset eval class

Cfs subset eval class predicts the relevance of a subset of points by considering the unique ominous strength of each point onward by the level of redundancy within them. Table 1 displays the values of F-measure, recall, and precision for every one of the 15 classifications used in Cfs subset eval. Figure 1 is a diagrammatic illustration of Table 1.

The results from Table 1 show that the precision value is always higher than the recall and F-measure values. Besides, there were no significant changes in the results of all classifications used together with Cfs subset eval, but the random tree classification showed the lowest performance in the F-measure, precision, and recall rules utilising the feature selection algorithm. Figure 2 shows the results of each method in graphical form, based on three standards F-measure, precision, and recall rules using the feature selection algorithm.

Table 1. Performance evaluation of Cfs subset eval class

Classification Algorithm	F-Measure	Recall	Precision
Bayes Net	0.706	0.708	0.789
Naive Bayes	0.728	0.731	0.822
Naive Bayes Updateable	0.728	0.731	0.822
Multilayer Perceptron	0.734	0.731	0.742
Simple Logistic	0.721	0.724	0.819
SMO	0.723	0.727	0.827
Decision Tree	0.719	0.722	0.814
JRip	0.72	0.724	0.822
OneR	0.723	0.727	0.827
PART	0.727	0.724	0.744
Decision Stump	0.723	0.727	0.827
J48	0.763	0.761	0.785
Random Forest	0.732	0.731	0.732
Random Tree	0.697	0.697	0.697
REP tree	0.711	0.708	0.74

Figure 1. F-measure, recall, and precision for Cfs subset eval class

3.2. Chi squared attribute eval class

Chi squared attribute eval class determines the attribute by measuring the chi-squared statistical value associated with an existing class. Table 2 presents the results of F-measure, recall, and precision toward 15 classifications accompanying Chi squared attribute eval. Figure 2 is a diagrammatic illustration of Table 2. The results are presented in Table 2, and Figure 2 illustrates the MLP classification that has the lowest performance in educational data sets using Chi squared attribute eval.

Table 2. Performance evaluation of Cfs subset eval class

Classification Algorithm	F-Measure	Recall	Precision
Bayes Net	0.754	0.752	0.77
Naive Bayes	0.729	0.727	0.764
Naive Bayes Updateable	0.729	0.727	0.764
Multilayer Perceptron	0.697	0.695	0.7
Simple Logistic	0.712	0.711	0.76
SMO	0.709	0.711	0.787
Decision Tree	0.759	0.756	0.777
JRip	0.777	0.774	0.797
OneR	0.723	0.727	0.827
PART	0.72	0.72	0.72
Decision Stump	0.723	0.727	0.827
J48	0.754	0.754	0.754
Random Forest	0.751	0.749	0.753
Random Tree	0.705	0.704	0.707
REP tree	0.749	0.747	0.754

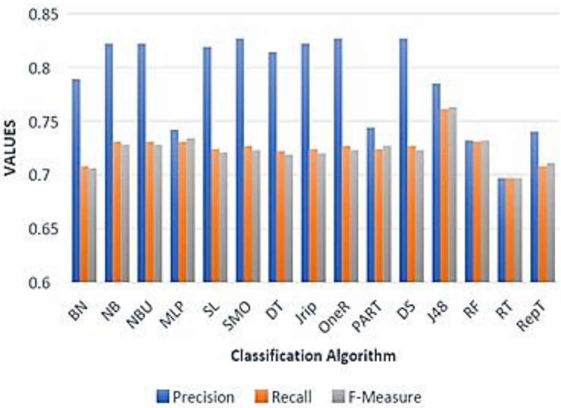

Figure 2. F-measure, recall, and precision of Chi squared attribute eval class

3.3. Filtered attribute eval class

Table 3 and Figure 3 present the results of the classification, which is used in educational data utilising the filtered attribute eval class. The results prove that MLP gives relatively deep values of F-measure, recall, and precision as in the previous method. While JRip's offering is relatively more reliable than other classifications utilising filtered attribute eval class.

Table 3. Performance evaluation of filtered attribute eval class

Classification Algorithm	F-Measure	Recall	Precision
Bayes Net	0.754	0.752	0.77
Naive Bayes	0.729	0.727	0.764
Naive Bayes Updateable	0.729	0.727	0.764
Multilayer Perceptron	0.697	0.695	0.7
Simple Logistic	0.712	0.711	0.76
SMO	0.709	0.711	0.787
Decision Tree	0.759	0.756	0.777
JRip	0.777	0.774	0.797
OneR	0.723	0.727	0.827
PART	0.72	0.72	0.72
Decision Stump	0.723	0.727	0.827
J48	0.754	0.754	0.754
Random Forest	0.751	0.749	0.753
Random Tree	0.707	0.704	0.705
REP tree	0.754	0.747	0.749

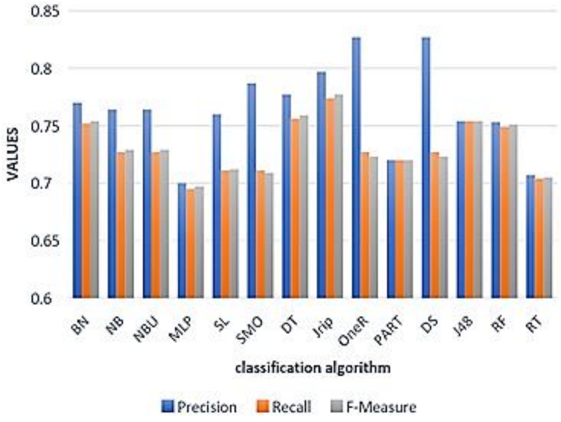

Figure 3. F-measure, recall, and precision of filtered attribute eval class

3.4. Gain ratio attribute eval class

Gain ratio attribute eval class is a non-symmetrical device that was added to recompense for the preference (deviation) of knowledge acquisition [17]. Table 4 and Figure 4 show that using a classification gain ratio attribute eval class performance, which is quite low compared to other classifications.

Table 4. Performance evaluation of gain ratio attribute eval class

Classification Algorithm	F-Measure	Recall	Precision
Bayes Net	0.644	0.708	0.684
Naive Bayes	0.62	0.683	0.631
Naive Bayes Updateable	0.62	0.683	0.631
Multilayer Perceptron	0.632	0.663	0.624
Simple Logistic	0.63	0.695	0.654
SMO	0.578	0.674	0.574
Decision Tree	0.646	0.704	0.673
JRip	0.607	0.667	0.604
OneR	0.57	0.638	0.547
PART	0.614	0.647	0.603
Decision Stump	0.577	0.679	0.58
J48	0.648	0.67	0.641
Random Forest	0.605	0.636	0.593
Random Tree	0.606	0.604	0.608
REP tree	0.621	0.667	0.617

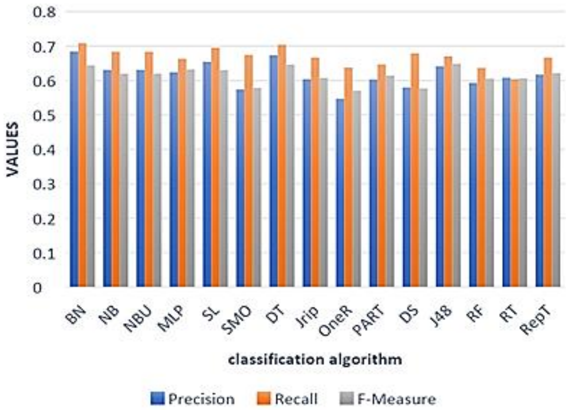

Figure 4. F-measure, recall, and precision of gain ratio attribute eval class

3.5. Principal component class

Table 5 present the appearance of the principal components utilising 15 classifications, which are 15 exist in the WEKA open-source data mining application. Figure 5 is a graph illustration of Table 5. The results show that the Bayes net classification has relatively better performance, while random tree shows low performance.

3.6. Relief attribute eval class

This evaluates the importance of attributes with examples that are taken repeatedly. The outcomes are shown in Table 6 present the results of the relief atribute eval evaluation on the classification shown in the classification on the classification of relief atribute eval on the classification shown in Table 6 differently. Figure 6 is a graph representation of Table 6. The results of the relief attribute eval evaluation have results similar to the gain ratio attribute eval evaluation. The results depict that the Bayes net classification has better performance than the other classifications, but OneR shows the low performance when using relief attribute eval on the student dataset.

Table 5. Performance evaluation of principal components class

Classification Algorithm	F-Measure	Recall	Precision
Bayes Net	0.648	0.711	0.689
Naive Bayes	0.63	0.688	0.642
Naive Bayes Updateable	0.63	0.688	0.642
Multilayer Perceptron	0.598	0.622	0.586
Simple Logistic	0.632	0.697	0.658
SMO	0.57	0.679	0.56
Decision Tree	0.618	0.699	0.668
JRip	0.594	0.677	0.602
OneR	0.57	0.638	0.547
PART	0.624	0.633	0.618
Decision Stump	0.577	0.679	0.58
J48	0.613	0.638	0.602
Random Forest	0.612	0.64	0.6
Random Tree	0.564	0.558	0.57
REP tree	0.614	0.672	0.615

Table 6. Performance evaluation of relief attribute eval class

Classification Algorithm	F-Measure	Recall	Precision
Bayes Net	0.648	0.711	0.689
Naive Bayes	0.64	0.69	0.649
Naive Bayes Updateable	0.64	0.69	0.649
Multilayer Perceptron	0.645	0.677	0.641
Simple Logistic	0.629	0.697	0.658
SMO	0.578	0.674	0.574
Decision Tree	0.618	0.699	0.668
JRip	0.592	0.667	0.588
OneR	0.57	0.638	0.547
PART	0.626	0.642	0.617
Decision Stump	0.577	0.679	0.58
J48	0.618	0.642	0.607
Random Forest	0.628	0.647	0.619
Random Tree	0.621	0.624	0.619
REP tree	0.637	0.674	0.634

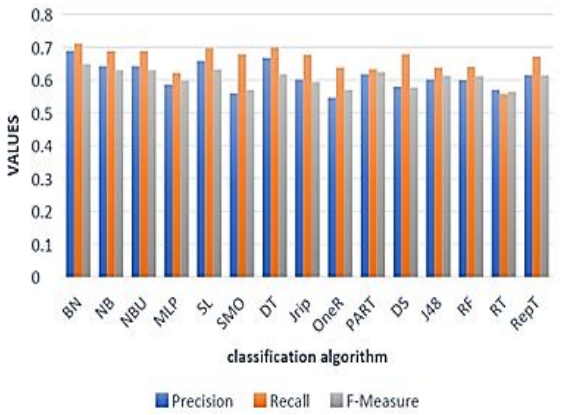

Figure 5. F-measure, recall, and precision of principal components class

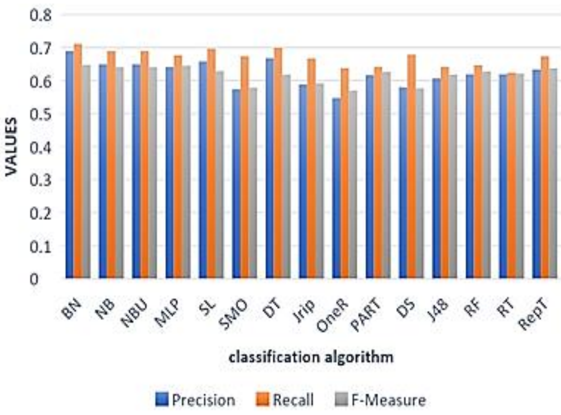

Figure 6. F-measure, recall, and precision of relief attribute eval class

Table 7 presents the values of every feature selection algorithm with various classifications. Finally, the mean and the variance of every feature selection are used to check variations in the appearance of the feature selection algorithm among separate classification methods. The decisiontree (DT) classification has better performance when used on the FS algorithm, and the randomtree (RT) classification has the lowest performance among other classifications.

The results within Figure 7 and Figure 8 present the mean and the variance in the chosen feature selection (FS) algorithm. Cfs subset eval (CSE), Chi squared attribute eval (CSAE), filtered attribute eval (FAE), gain ratio attribute eval (GRAE), principal components (PC), and relief attribute eval (RAE). Bayes net (BN), Naïve Bayes (NB), Naive Bayes updateable (NBU), multilayer perceptron (MP), simple logistic (SL), SMO, decision tree (DT), JRip, OneR, PART, decision stump (DS), J48, random forest (RF), random tree (RT), and REP tree.

Table 7. Evaluation of performance algorithms feature selection in context with correctly classified instances

FS	Correctly Classified Instances (%)								Mean	Variance
	BN	NB	NBU	MP	SL	SMO	DT	JRip		
CSE	70.8	73.1	73.1	73.1	72.43	72.66	72.2	72.43	72.49	0.000186
CSAE	74.4	73.3	73.3	71.2	71.07	71.07	75.62	76.3	73.80	0.000394
FAE	75.1	72.6	72.6	69.4	71.07	71.07	75.62	77.44	73.19	0.000476
GAE	70.8	68.3	68.3	66.2	69.47	67.42	70.38	66.74	66.78	0.000723
PC	71.1	68.7	68.7	62.1	69.7	67.88	69.93	67.65	66.12	0.001474
RAE	71.1	69.0	69.0	67.6	69.7	67.42	69.93	66.74	67.01	0.000629
	OneR	PART	DS	J48	RF	RT	REP tree			
CSE	72.66	72.43	72.66	76.08	73.12	69.7	70.84		72.49	0.000186
CSAE	72.66	74.94	72.66	77.44	75.85	71.75	75.17		73.80	0.000394
FAE	72.66	71.98	72.66	75.39	74.94	70.38	74.71		73.19	0.000476
GAE	63.78	64.69	67.88	66.97	63.55	60.36	66.74		66.78	0.000723
PC	63.78	63.32	67.88	63.78	64	55.8	67.19		66.12	0.001474
RAE	63.78	64.23	67.88	64.23	64.69	62.41	67.42		67.01	0.000629

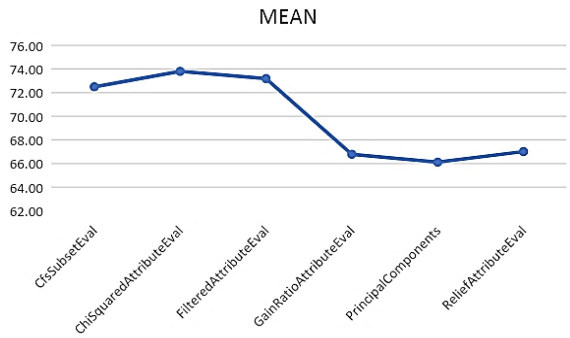

Figure 7. Average FS algorithm

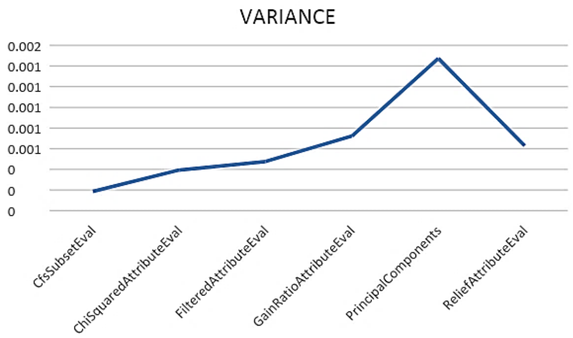

Figure 8. Variance FS algorithm

4. CONCLUSION

In this paper, different algorithms have been assessed and analyzed the FS algorithm. The results in the educational dataset show that there is no important change in the performance of the FS algorithm in the WEKA application. But among all available FS methods, the principal components method shows better results when using FS with Bayes net (BN) classification. This paper also shows that the decision tree (DT) classification performs better than the other classifications in the student dataset, and the random tree (RT) classification is the lowest-performing class among the other classifications. The results represent that there is a need to adjust complex parameters with the FS method, to achieve better performance. For the future FS and its various mixtures, and educational datasets of various areas can also be utilized for evaluation.

ACKNOWLEDGMENT

This research is the result of the basic research scheme of the Indonesian Dikti grant B/112/E3/RA.00/2021.

REFERENCES

[1] A. Triayudi, W. O. Widyarto, and V. Rosalina, "CLG Clustering for Mapping Pattern Analysis of Student Academic Achievement," *ICIC Express Letters*, vol. 14, no. 12, pp. 1225-1234, 2020, doi: 10.24507/icicel.14.12.1225.

[2] E. Osmanbegović, M. Suljić, and H. Agić, *Determining Dominant Factor For Students Performance Prediction By Using Data Mining Classification Algorithms*, Tranzicija, vol. 16, pp. 147-158, 2015. [Online]. Available: https://core.ac.uk/download/pdf/33273916.pdf

[3] A. Triayudi and I. Fitri, "A new agglomerative hierarchical clustering to model student activity in online learning," *TELKOMNIKA Telecommunication Computing Electronics and Control*, vol. 17, no. 3, pp. 1226-1235, 2019, doi: 10.12928/telkomnika.v17i3.9425.

[4] A. Triayudi, Sumiati, S. Dwiyatno, D. Karyaningsih, and Susilawati, "Measure the effectiveness of information systems with the naïve bayes classifier method," *IAES International Journal of Artificial Intelligence*, vol. 10, no. 2, pp. 414-420, 2021, doi: 10.11591/ijai.v10.i2.pp414-420.

[5] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414-422, 2015, doi: 10.1016/j.procs.2015.12.157.

[6] A. Mishra, R. Bansal, and S. N. Singh, "Educational data mining and learning analysis," *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, 2017, pp. 491-494, doi: 10.1109/CONFLUENCE.2017.7943201.

[7] N. Hidayat, R. Wardoyo, and S. Azhari, "Educational Data Mining (EDM) as a Model for Students' Evaluation in Learning Environment," *2018 Third International Conference on Informatics and Computing (ICIC)*, 2018, pp. 1-4, doi: 10.1109/IAC.2018.8780459.

[8] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques," *International Journal of Modern Education and Computer Science*, vol. 8, p. 36, 2016, doi: 10.5815/ijmecs.2016.11.05.

[9] W. Punlumjeak and N. Rachburee, "A comparative study of feature selection techniques for classify student performance," *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2015, pp. 425-429, doi: 10.1109/ICITEED.2015.7408984.

[10] W. Zhang and S. Qin, "A brief analysis of the key technologies and applications of educational data mining on online learning platform," *2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA)*, 2018, pp. 83-86, doi: 10.1109/ICBDA.2018.8367655.

[11] C. Jalota and R. Agrawal, "Analysis of Educational Data Mining using Classification," *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 243-247, doi: 10.1109/COMITCon.2019.8862214.

[12] H. M. Harb and M. A. Moustafa, "Selecting optimal subset of features for student performance model," *IJCSI International Journal Computer Science*, vol. 9, no. 5, pp. 253-262, 2012. [Online]. Available: https://www.ijcsi.org/papers/IJCSI-9-5-1-253-262.pdf

[13] A. Figueira, "Predicting Grades by Principal Component Analysis: A Data Mining Approach to Learning Analyics," *2016 IEEE 16th International Conference on Advanced Learning Technologies (ICALT)*, 2016, pp. 465-467, doi: 10.1109/ICALT.2016.103.

[14] A. K. Pandey, D. S. Rajpoot, and D. S. Rajpoot, "A comparative study of classification techniques by utilizing WEKA," *2016 International Conference on Signal Processing and Communication (ICSC)*, 2016, pp. 219-224, doi: 10.1109/ICSPCom.2016.7980579.

[15] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," *International Journal of Database Theory and Application*, vol. 9, no. 8, pp. 119-136, 2016, doi: 10.14257/ijdta.2016.9.8.13. [Online]. Available: http://article.nadiapub.com/IJDTA/vol9_no8/13.pdf

[16] S. M. Ajibade, N. B. Ahmad, and S. M. Shamsuddin, "An Heuristic Feature Selection Algorithm to Evaluate Academic Performance of Students," *2019 IEEE 10th Control and System Graduate Research Colloquium (ICSGRC)*, 2019, pp. 110-114, doi: 10.1109/ICSGRC.2019.8837067.

[17] T. Velmurugan and C. Anuradha, "Performance Evaluation of Feature Selection Algorithms in Educational Data Mining," *International Journal of Data Mining Techniques and Applications*, vol. 5, no. 2, pp. 131-139, 2016, doi: 10.20894/IJDMTA.102.005.002.007.

[18] M. Peker, A. Arslan, B. Şen, F. V. Çelebi, and A. But, "A novel hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection and random forest algorithm (ReliefF+RF)," *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, 2015, pp. 1-8, doi: 10.1109/INISTA.2015.7276737.

[19] C. Anuradha and T. Velmurugan, "Feature Selection Techniques to Analyse Student Acadamic Performance using Naïve Bayes Classifier," in *The 3rd International Conference on Small & Medium Business*, 2016, pp. 345-350.

[20] J. Novaković, P. Strbac, and D. Bulatović, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav Journal of Operations Research*, vol. 21, no. 1, pp. 119-135, 2016, doi: 10.2298/YJOR1101119N.

[21] S. T. Ahmed, R. Al-Hamdani, and M. S. Croock, "Developed third iterative dichotomizer based on feature decisive values for educational data mining," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 209-217, doi: 10.11591/ijeecs.v18.i1.pp209-217.

[22] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, doi: 10.1002/widm.1355.

[23] Y. S. Mitrofanova, A. A. Sherstobitova, and O. A. Filippova, "Modeling smart learning processes based on educational data mining tools," In *Smart Education and e-Learning 2019*, pp. 561-571, 2019, doi:10.1007/978-981-13-8260-4_49

[24] R. Ahuja, A. Jha, R. Maurya, and R. Srivastava, "Analysis of Educational Data Mining," In: Yadav N., Yadav A., Bansal J., Deep K., Kim J. (eds) Harmony Search and Nature Inspired Optimization Algorithms. Advances in *Intelligent Systems and Computing*, vol. 741, 2019, doi: 10.1007/978-981-13-0761-4_85

[25] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing AutoML in educational data mining for prediction tasks," *Applied Sciences*, vol. 10, no. 1, p. 90, 2020, doi:10.3390/app10010090.