Ermatita Paper 6 1570815512-Dafid

by Ermatita Paper 6

Submission date: 11-Oct-2022 02:38PM (UTC+0700)
Submission ID: 1922393315
File name: Paper2-1570815512-Dafid.pdf (250.78K)
Word count: 4883
Character count: 24992

2022 International Conference on Data Science and Its Applications (ICoDSA)

Filter-Based Feature Selection Method for Predicting Students' Academic Performance

Dafid
Information System
Universitas Multi Data Palembang
Palembang, Indonesia
dafid@mdp.ac.id

Ermatita
Doctoral Program in Engineering Science
University of Sriwijaya
Palembang, Indonesia
ermatita@unsri.ac.id

***Abstract*— Generally, almost all higher education often face the same problem of improving their quality according to students' academic performance. The need to get early information about the poor students' academic performance has forced higher education to find the best solution that the prediction model could achieve. Data mining offers various algorithms for predicting. Therefore, constructing an accurate prediction model becomes a challenging task for higher education. Two factors that drive the accuracy of the prediction model are classifiers and feature selection. Each classifier gives the best result if it meets the appropriate categorized data on a dataset. A few research has provided excellent results in predicting students' academic performance. But, the research only focuses on the classification technique rather than the right feature selection. Vice versa, a few research have reported excellent results increasing the prediction model accuracy. But the research only focuses on feature selection techniques rather than carrying out the right classifier on the right data. Therefore, the prediction model has not given the best accuracy yet. Unlike than existing framework to build a model and select the features ignoring the categorized data on a dataset, this research proposes the right filter-based feature selection methods and the right classifiers based on categorized data. The result will help the researcher find the best combination of filter-based feature selection methods and classifiers. Various classification algorithms and various feature selections that have been tested show classification with appropriate classifiers for specific categorized data and proper feature selection increase the prediction model's accuracy.**

Keywords*— *prediction, academic performance, classifiers, feature selection, accuracy

I. INTRODUCTION

Recently, all higher education concerns improving the quality of learning achieved by generating graduates with better academic performance. The success of improving the quality of learning depends on the availability of early information about students' academic performance raised by prediction. Many prediction techniques are being used to evaluate the students' academic performance[1]. For this reason, data mining is considered to be the best choice for predicting students' academic performance[1,2]. It is defined as a process used to extract data from a larger set of raw data using various techniques from database systems, data visualization, artificial intelligence, knowledge acquisition, pattern recognition, statistics, machine learning, and information theory. Several research has been reported by many researchers and provided great results in the prediction of students' academic performance[1–9]. In most of these studies, classification is the most popular technique to predict students' academic performance. Among the algorithms under classification used are Decision Tree (DT), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naive Bayes (NB,) and Artificial Neural Network (ANN).

Research done by Jishan [10] using Decision Tree found that the final cumulative grade point average (CGPA) achieves the highest accuracy (91%) than other attributes. Meanwhile, researchers [11] using Neural Network achieve the highest accuracy (98%) on Internal assessments and External assessment attributes than others. Osmanbegovic and Suljic [12] have used Naive Bayes algorithms to estimate students' performance. Their research showed that Naive Bayes had achieved the highest accuracy (76%) on CGPA, Student Demographics, High school background, Scholarship, Social network interaction attributes than others. Other researchers Mayilvaganan and Kapalnadevi [13] found that Internal assessment, CGPA, Extracurricular activities attribute under KNN give a good accuracy (83%) than others. The last algorithm is the Support Vector Machine applied by Sembiring [14] gives a better accuracy on Psychometric attribute (83%) than others. Generally speaking, classier will achieve the highest accuracy if it meets the appropriate categorized data on a dataset [1]. However, the attributes on a dataset in the fact still have the problem influencing the accuracy of the prediction model [15–19]. The problems are redundant and irrelevant attributes [20]. To overcome the problem, feature selection is applied to select only relevant attributes and removes the redundant and irrelevant attributes[15-21]. The research above only focuses on the selection of an appropriate classifier rather than carrying out the feature selection on a dataset. Therefore, yet prediction model has not given the best accuracy. To overcome the problem, feature selection is applied for the right classifier on the right categorized data.

Feature selection yields effective results increasing prediction model accuracy[15-21]. But the research only focuses on the removal of redundant and irrelevant attributes rather than carrying out the feature selection on the rig ht classifier on right categorized data. Therefore, still a lot of attention is required to construct a prediction model with the right classifier on the right data and feature selection. Punlum jeak [22] has reported minimum Redundancy and Maximum Relevance (mRMR) method raised the highest accuracy with KNN classifier. Meanwhile Rahman [23] has stated information gain method gives the best result with ANN classifier. Another researcher [20] developed a comparative study and indicated Correlation based Feature Selection (CFS) method with Naïve Bayes classifier and Information Gain with Decision Tree classifier raised the highest accuracy.

This research aims to get the highest accuracy prediction model for predicting students' academic performance under the classification method by combining the right classifiers on the right categorized and right feature selection. All of the prediction models that have been created are next tested in finding the best combination using evaluation measurement which is Accuracy.

978-1-6654-8665-1/22/$31.00 ©2022 IEEE

The next section of this paper is organized as follows: Section II gives the study of literature review. Section III explained the methodology of this research. Section IV discusses the research findings and introduces the study implications. Finally, Section V outlines the conclusion.

II. Literature Review

A. Classification Method

Decision Tree (DT) method has ability to uncover small or large data struture[12], [24]. Its simpilicity and comprehensibility lead Decision Tree be the most popular technique for prediction. Decision tree transform the data table into a tree model, which determines the attributes from the roots, branches to the decision. The determination of attribute is calculated using gain ratio. The value of information gain means how much information is obtained by knowing the value of an attribute while the split information value is used for an attribute that has multiple instances (more than two and varied). The formula is (1):

$$\text{GainRatio(S|A)} = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (1)$$

K-Nearest Neighbor (KNN) is a noise-sensitive classifier and an popular non-parametric classification method which have been successfully implemented in many classification problems[25]. The KNN highly depends on the quality of the training data for its performance. The things affect the accuracy are the noise of data and mislabeled data, outliers, and overlaps regions between the data of different classes or targets lead.

Another popular technique for prediction is Artificial Neural Networks (ANN). ANN has the advantage of doing a complete detection in a nonlinear relationship between dependent and independent variables[26]. ANN also could detect all possible interactions between predictor variables[27].

Naïve Bayes (NB) is used for two class and multiclass classification problem[12]. The formula is (2):

$$\text{P}(h|\text{d}) = \frac{P(d|h)*P(h)}{P(d)} \quad (2)$$

Where:

- P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(d|h) is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(d) is the probability of the data (regardless of the hypothesis)

Select the hypothesis with the highest probability after calculating the posterior probability to get the maximum probable hypothesis (MAP). The formula is (3):

$$\text{MAP}(h) = max\ \frac{(P(d|h)*P(h))}{P(d)} \quad (3)$$

P(d) is a normalizing term which allows us to calculate the probability. If there is an even number of instances in each class in the training data, then the probability of each class (e.g., P(h)) will be equal. Again, this would be a constant term in the equation and drop it so that ends up with (4):

$$\text{MAP}(h) = \max(P(d|h)) \quad (4)$$

One of supervised learning method for classification is Support Vector Machine (SVM). It has advantage in ability to classify the data in small datasets. It also has a good generalization ability and faster than other methods[14].

B. Feature Selection Method

Feature Selection is a preprocessing step utilized to determine relevant attributes before applying a classification model. It aims at selecting a number of relevant features and removing a number of redundant and irrelevant features from dataset [17].

Feature selection method consist of three categories: filter, wrapper, and embedded methods [22]. In filter methods, selected feature identified by using some feature descending order ranking criteria. In wrapper and embedded methods selected feature chosed by classifier. The difference is embedded methods incorporate the feature selection process as part of the construction stage of the prediction model.

Minimum Redundancy and Maximum Relevance (mRMR) select the features by using the relationship between the feature and the target class with the highest relevance [22]. Meanwhile minimum redundancy select the features that they are mutually maximally dissimilar to other features. Let S denote the subset of features and |S| is the number of features in S. The minimal redundancy condition is (5) :

$$\min \text{R(S)},\ \text{R(S)} = \frac{1}{|S|^2} \sum_{x_i x_j \in S} I(xi; xj) \quad (5)$$

The maximal relevance condition is (6) calculated using the mutual information between the individual feature and the target class c as a the relevance feature measurement:

$$\max \text{D(S,c)},\ \text{D(S,c)} = \frac{1}{S} \sum_{x_j \in S} I(xi; c) \quad (6)$$

In conclusion mRMR method is calculated by optimizing by the conditions in equation (5) and (6) simultaneously. Information Gain (IG) select the features by using statistical and entropy-based measurement [22]. In the Information Gain (IG) method, determining the entropy with respect to the class to which they belong is conducted to calculate the value of the features information. Entropy value ranges from 0 to 1, value 0 means that all instances of the variables have the same value and value 1 equals the number of instances of each value. The entropy of N is calculated as (7):

$$\text{Entropy (A)} = -\sum_{i=1}^{k} p_i log_2 p_i \quad (7)$$

In equation (7), p is the probability of class, for which a particular value and this equation calculates the information of all classes.

$$\text{Entropy (Di)} = \sum_{i=1}^{Dj} \frac{Dji}{N} \text{X Entropy (Dji)} \quad (8)$$

Information Gain (IG) is calculating by the difference of equation (7) and equation (8) as follow (9).

$$IG(D)j = Entropy(A) - Entropy(D) \quad (9)$$

Correlation based Feature Selection (CFS) select the features by using correlation coefficients to estimate correlation between subset of attributes and class [28]. Correlation coefficients also used to estimate inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes and decreases with growing inter-correlation. The CFS formula is (10):

$$r_{zc} = \frac{kr_{zi}}{\sqrt{k+k(k-1)r_{ii}}} \quad (10)$$

where r_{zc} is the correlation between the summed feature subsets and the class variable, k is the number of subset features, rzi is the average of the correlations between the subset features and the class variable, and rii is the average inter-correlation between subset features.

III. METHODOLOGY

The aim of this research is to access the different right classifier with the right data and predict the students' academic performance by means of right feature selection. This research would be able to answer the folowing question:

Q: What are the best combinations of feature selection method and classsifier with the right categorized data to predict students' academic performance?

For the purpose the research objective and to get answer the research question above an educational dataset is taken from valid sources and on the dataset various feature selection method is applied.

A. Proposed Framework

The first step to conduct this research is getting educational datasets. Then, identify from them to decide what categorized of the data. The dataset need to be classified to choose the appropriate feature selection method under classification algorithm according the previous research. The data category[29] are *Prior Academic Achievement, Student Demographics, Students' Environment, Psychological and Student E-learning Activity* is showed in Table 1. Fig. 1 describes the framework of this research

TABLE I. DATA CATEGORY

Factor Category	Factor Description
Prior Academic Achievement (AA)	Pre-university data: high school background (i.e., high school results),pre-admission data (e.g. admission test results) University-data: semester GPA or CGPA, individual course letter marks, and individual assessment grades
Student Demographics (SD)	Gender, age, race/ethnicity, socioeconomic status (i.e., parents ' education and occupation, place of residence / traveled distance, family size, and family income).
Students' Environment (SE)	Class type, semester duration, type of program
Psychological (PS)	Student interest, behavior of study, stress, anxiety, time of preoccupation, self-regulation, and motivation.
Student E-learning Activity (EA)	Number of logins times, number of tasks, number of tests, assessment activities, number of discussion board entries, number / total time material viewed

B. Data

This research analyzed a public dataset from Kaggle Repository datasets: Students' Academic Performance. This dataset contains 480 record and 16 attributes. According to Table I, the attributes in this dataset can be categorized into 5 data categories as shown in Table II.

C. Feature Selection and Classifier

The proposed prediction model used three feature selection method: Minimum Redundancy and Maximum Relevance (mRMR), Information Gain (IG) and Correlation based Feature Selection (CFS) and used 5 classifiers: Decision Tree, K-Nearest Neighbor, Artificial Neural Networks, Naïve Bayes, Support Vector Machine

D. Experimental Setup

For the model generation, this research used the Rapid Miner Studio version 9.9 software package. This software is very powerful to build predictive analytic models because it's a suitable platform for data preparation, machine learning and model deployment as well. This experiment run on Intel Processor i7 10th generation, 12 GB RAM and Win10 operating system.

TABLE II. ATTRIBUTES AND ITS DATA CATEGORY

Attributes	Description	Data category
nationality	student's nationality	SD
gender	student's gender	SD
place of birth	student's place of birth	SD
educational stages	educational level student belongs (mark)	AA

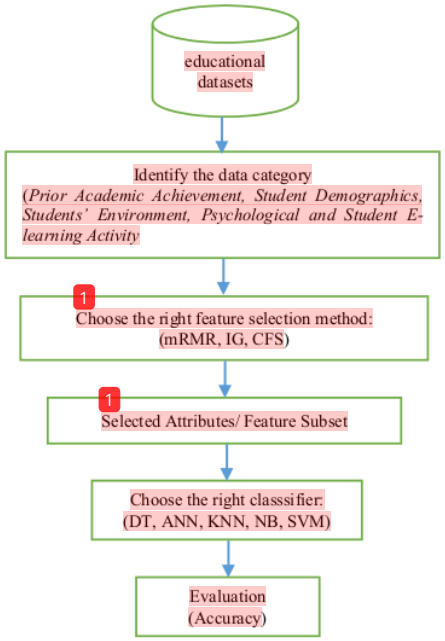

Fig. 1. Proposed framework.

Attributes	Description	Data category
grade levels	grade student belongs	SE
section ID	classroom student belongs	SE
topic	course topic	SE
semester	school year semester	SE
parent responsibility	parent responsible for student	PS
raised hand	how many times the student raises his/her hand on classroom	PS
visited resources	how many times the student visits a course content	EA
viewing announcements	how many times the student checks the new announcements	EA
discussion groups	how many times the student participate on discussion groups	EA
parent answering survey	parent answered the surveys which are provided from school or not	PS
parent school satisfaction	the degree of parent satisfaction from school	PS
student absence days	the number of absence days for each student	PS

IV. RESULT

This section reported the performance evaluation of combination feature selection and classifier for predicting academic performance. The dataset prepared previously then imported to Rapid Miner Studio software. Next the software started to process dataset through following phases: data analysis, data pre-processing, and then design by applying the classification algorithm, training and testing. In the data analysis phase, the target attribute are determined. In the data analysis step, Rapid Miner divides data into two sets:training(90%) and testing(10%) which type of sampling is stratified sampling due to the label is nominal. In data pre-processing, the relevant feature is selected using feature selection. Training process train the model used classification algorithm which requires criterion options like accuracy. During the testing process, it used two operations: The Apply Model on the test dataset and the Performance operation for measuring the model performance. In order to analyze the performance of these students, a prediction model is created based on classification algorithm by selecting appropriate ones which in the end helped to predict which students may passed or failed. The result of various feature selection method are explaind in Table IV by applying right classifier on right data.

A. Experimental Results without Feature Selection

TABLE III. ACCURACY OF VARIOUS MODEL WITHOUT FEATURE SELECTION

Combination of Data Category	Classifier (%)				
	DT	NB	ANN	KNN	SVM
AA+SD+SE	93.04	75.24	94.45	84.91	79.98
AA+SD+PS	92.06	75.75	93.59	83.95	80.48
AA+SD+EA	93.85	75.56	95.12	84.18	80.62
AA+SE+PS	93.72	76.42	94.39	84.14	80.85
AA+SE+EA	92.79	75.96	94.81	83.90	79.76
AA+PS+EA	93.58	74.11	93.77	84.42	80.34
SD+SE+PS	93.05	76.52	94.43	84.11	80.15
SD+SE+EA	93.36	75.43	95.97	83.34	80.44
SD+PS+EA	91.56	77.21	94.05	83.37	80.70
SE+PS+EA	92.84	74.62	93.98	84.54	79.22
AA+SD+SE+PS	93.86	75.16	94.32	84.35	80.68
AA+SD+SE+EA	93.48	77.54	95.44	83.52	81.01
AA+SD+PS+EA	92.47	75.64	94.88	83.23	80.65
SD+SE+PS+EA	93.68	76.29	94.17	84.41	79.45
AA+SE+PS+EA	92.77	74.37	95.13	84.26	80.57
AA+SD+SE+PS+EA	93.24	74.91	95.24	84.18	80.48

B. Experimental Results with Feature Selection

TABLE IV. ACCURACY OF VARIOUS MODEL WITH FEATURE SELECTION

Combination of Data Category	Classifier + Feature Selection				
	DT+ mRMR	NB+ mRMR	ANN+ mRMR	KNN+ mRMR	SVM+ mRMR
	DT+IG	NB+IG	ANN+IG	KNN+IG	SVM+IG
	DT+CFS	NB+CFS	ANN+CFS	KNN+CFS	SVN+CFS
AA+SD+SE	95.05	78.44	95.65	86.41	82.14
	95.12	78.28	95.80	85.57	81.55
	94.85	78.90	95.10	85.98	81.91
AA+SD+PS	95.75	78.02	95.60	85.96	82.93
	95.97	78.43	95.64	85.33	82.26
	95.81	78.64	95.50	85.73	82.47
AA+SD+EA	96.11	78.26	95.49	85.44	81.17
	95.87	79.05	96.53	86.61	82.23
	95.63	78.33	95.71	86.02	81.94
AA+SE+PS	96.07	79.13	96.58	86.16	82.83
	96.66	79.27	96.08	86.25	81.16
	95.39	78.99	95.19	85.77	82.38
AA+SE+EA	96.36	79.76	96.10	86.87	82.83
	96.49	79.34	96.55	86.20	82.91
	96.24	79.41	96.41	86.66	82.50
AA+PS+EA	96.29	78.34	96.48	86.37	83.51
	97.01	79.86	96.38	86.62	82.36
	96.44	79.31	96.35	86.17	83.67
SD+SE+PS	97.05	78.40	96.65	85.61	82.44
	97.12	78.25	95.32	85.75	82.75

Combination of Data Category	Classifier + Feature Selection				
	DT+ mRMR	NB+ mRMR	ANN+ mRMR	KNN+ mRMR	SVM+ mRMR
	DT+IG	NB+IG	ANN+IG	KNN+IG	SVM+IG
	DT+CFS	NB+CFS	ANN+CFS	KNN+CFS	SVN+CFS
	96.28	78.87	95.66	85.83	82.61
SD+SE+EA	94.29	80.34	95.46	86.63	83.21
	94.31	80.46	94.25	86.68	83.53
	94.45	80.32	94.73	86.35	83.99
SD+PS+EA	94.71	80.62	94.79	86.74	82.68
	94.25	80.76	94.56	86.55	82.30
	94.37	80.52	94.22	86.79	82.10
SE+PS+EA	94.33	80.37	94.76	86.66	83.91
	95.91	79.96	94.23	85.61	83.42
	94.79	79.39	94.79	86.39	83.34
AA+SD+SE+PS	96.33	79.27	97.01	85.27	82.42
	96.54	78.34	96.76	85.31	82.47
	96.39	79.41	96.52	85.44	81.85
AA+SD+SE+EA	96.71	78.62	96.89	85.24	81.77
	95.22	78.76	96.63	86.67	82.29
	95.34	78.52	95.91	86.34	81.90
AA+SD+PS+EA	96.97	79.21	96.45	86.26	82.55
	96.62	79.47	96.70	86.57	82.66
	95.33	78.92	95.47	85.94	82.32
SD+SE+PS+EA	95.55	79.33	96.49	85.43	82.18
	95.74	80.32	96.82	85.76	82.24
	96.13	80.46	96.36	85.23	82.50
AA+SE+PS+EA	96.93	79.40	95.11	85.27	82.42
	95.74	78.23	96.65	85.31	82.47
	94.49	79.62	96.72	85.44	81.85
AA+SD+SE+PS+EA	96.83	78.75	96.32	86.39	82.73
	97.43	79.83	96.51	85.71	82.45
	96.09	78.63	96.03	86.44	82.82

TABLE V. SUMMARY OF HIGHEST ACCURACY OF VARIOUS MODEL WITH FEATURE SELECTION + CLASSIFIER

Combination of Data Category	The Best Feature Selection + Classifier
AA+SD+SE	ANN+IG
AA+SD+PS	DT+IG
AA+SD+EA	ANN+IG
AA+SE+PS	DT+IG
AA+SE+EA	ANN+IG
AA+PS+EA	DT+IG
SD+SE+PS	DT+IG
SD+SE+EA	ANN+ mRMR
SD+PS+EA	ANN+ mRMR
SE+PS+EA	DT+IG
AA+SD+SE+PS	ANN+ mRMR
AA+SD+SE+EA	ANN+ mRMR
AA+SD+PS+EA	DT+ mRMR
SD+SE+PS+EA	ANN+IG
AA+SE+PS+EA	DT+ mRMR
AA+SD+SE+PS+EA	DT+ mRMR

The outcome in Tables IV show the result of variety accuracy value for three feature selection and five classifier in predicting students' academic performance. According to Table IV, Table V show the result of the best combination between classifier and feature selection. DT+IG give the highest acuracy (increase the accuracy) for AA+SD+PS, AA+SE+PS, AA+PS+EA, SD+SE+PS and SE+PS+EA data. ANN+IG give the highest acuracy (increase the accuracy) for AA+SD+SE,AA+SD+EA,AA+SE+EA and SD+SE+PS+EA. ANN+mRMR give the highest acuracy (increase the accuracy) for SD+SE+EA, SD+PS+EA, AA+SD+SE+PS and AA+SD+SE+EA. DT+mRMR give the highest acuracy (increase the accuracy) for AA+SD+PS+EA,AA+SE+PS+EA and AA+SD+SE+PS+EA.

V. CONCLUSION

This research succesfully apply feature selection method (mRMR, IG, CFS) and classifier (DT, ANN, KNN, NB, SVM based on data category in predicting students' academic performance. This research succesfully gets the highest accuracy if the data meet the right feature selection and the right classifier as shown in Table V. The performance of prediction model reach the highest prediction result that can be effectively used to predict students' academic performance.

REFERENCES

[1] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414–422, 2015, doi: 10.1016/j.procs.2015.12.157.

[2] A. Hellas *et al.*, "Predicting academic performance: A systematic literature review," *Annu. Conf. Innov. Technol. Comput. Sci. Educ. ITiCSE*, pp. 175–199, 2018, doi: 10.1145/3293881.3295783.

[3] M. A. Al-barrak and M. Al-razgan, "Predicting Students Final GPA Using Decision Trees : A Case Study," *Int. J. Inf. Educ. Technol.*, vol. 6, no. July 2016, pp. 528–533, 2016, doi: 10.7763/IJIET.2016.V6.745.

[4] M. Christina, "Predicting Student Performance using Data Mining," *Int. J. Comput. Sci. Eng.*, vol. 6, no. 10, pp. 172–177, 2018, doi: 10.26438/ijcse/v6i10.172177.

[5] J. Feng, "Predicting Students' Academic Performance with Decision Tree and Neural Network," pp. 2004–2019, 2019.

[6] J. Mesarić and D. Šebalj, "Decision trees for predicting the academic success of students," vol. 7, pp. 367–388, 2016, doi: 10.17535/crorr.2016.0025.

[7] M. Sivasakthi, "Classification and Prediction based Data Mining Algorithms to Predict Students ' Introductory programming Performance," no. Icici, pp. 0–4, 2017.

[8] P. J. M. Estrera, P. E. Natan, B. G. T. Rivera, and F. B. Colarte, "Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School Abstract :," *Philippine*, vol. 3, no. 5, pp. 147–154, 2017.

[9] R. Umer, T. Susnjak, A. Mathrani, and S. Suriadi, "On predicting academic performance with process mining in learning analytics," *J. Res. Innov. Teach. Learn.*, vol. 10, no. 2, pp. 160–176, 2017, doi: 10.1108/jrit-09-2017-0022.

[10] S. T. Jishan, R. I. Rashu, N. Haque, and R. M. Rahman, "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique," *Decis. Anal.*, vol. 2, no. 1, pp. 1–25, 2015, doi: 10.1186/s40165-014-0010-2.

[11] S. A. Kumar and M. N. Vijayalakshmi, "Appraising the Significance of Self Regulated Learning in Higher Education Using Neural Networks," *Int. J. Eng. Res. Dev.*, vol. Volume 1, no. Issue 1, pp. 9–15, 2012.

[12] E. Osmanbegovic and M. Suljic, "Data Mining Approach for Predicting Student Performance," 2012.

[13] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the cognitive skill of students in education environment," *2014 IEEE Int. Conf. Comput. Intell. Comput. Res. IEEE ICCIC 2014*, pp. 113–118, 2015, doi: 10.1109/ICCIC.2014.7238346.

[14] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of Student Academic Performance By an Application of Data Mining Techniques," *Manag. Artif. Intell.*, vol. 6, no. January 2017, pp. 110–114, 2011.

[15] W. Zheng, "A Comparative Study of Feature Selection Methods," *Int. J. Nat. Lang. Comput.*, vol. 7, no. 5, pp. 01–09, 2018, doi: 10.5121/ijnlc.2018.7501.

[16] M. Malekipirbazari, V. Aksakalli, W. Shafqat, and A. Eberhard, "Performance comparison of feature selection and extraction methods with random instance selection," *Expert Syst. Appl.*, vol. 179, no. February 2020, p. 115072, 2021, doi: 10.1016/j.eswa.2021.115072.

[17] S. García, J. Luengo, and F. Herrera, "Feature selection," *Intell. Syst. Ref. Libr.*, vol. 72, pp. 163–193, 2015, doi: 10.1007/978-3-319-10247-4_7.

[18] G. Manikandan and S. Abirami, "An efficient feature selection framework based on information theory for high dimensional data," *Appl. Soft Comput.*, vol. 111, 2021.

[19] S. Asim, A. Shah, H. M. Shabbir, S. U. Rehman, and M. Waqas, "A Comparative Study of Feature Selection Approaches: 2016-2020," *Int. J. Sci. Eng. Res.*, vol. 11, no. February, pp. 469–478, 2020.

[20] M. Zaffar, S. Iskander, and M. A. Hashmani, "A Study of Feature Selection Algorithms for Predicting Students Academic Performance," vol. 9, no. 5, pp. 541–549, 2018.

[21] M. S. Srivastava, M. N. Joshi, and M. M. Gaur, "A Review Paper on Feature Selection Methodologies and Their Applications," vol. 7, no. 6, pp. 57–61, 2013.

[22] W. Punlumjeak and N. Rachburee, "A comparative study of feature selection techniques for classify student performance," *Proc. - 2015 7th Int. Conf. Inf. Technol. Electr. Eng. Envisioning Trend Comput. Inf. Eng. ICITEE 2015*, pp. 425–429, 2015, doi: 10.1109/ICITEED.2015.7408984.

[23] L. Rahman, N. A. Setiawan, and A. E. Permanasari, "Feature selection methods in improving accuracy of classifying students' academic performance," *Proc. - 2017 2nd Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2017*, vol. 2018-Janua, no. 1, pp. 267–271, 2018, doi: 10.1109/ICITISEE.2017.8285509.

[24] M. Quadri and D. Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Glob. J. Comput.*, vol. 10, no. 2, pp. 2–5, 2010.

[25] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, "Predicting student performance: An application of data mining methods with an educational web-based system," *Proc. - Front. Educ. Conf.*, vol. 1, pp. 1–6, 2003.

[26] P. M. Arsad, N. Buniyamin, and J. L. A. Manan, "A neural network students' performance prediction model (NNSPPM)," *2013 IEEE Int. Conf. Smart Instrumentation, Meas. Appl. ICSIMA 2013*, no. November, 2013, doi: 10.1109/ICSIMA.2013.6717966.

[27] G. Gray, C. McGuinness, and P. Owende, "An application of classification models to predict learner progression in tertiary education," *Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014*, no. December 2015, pp. 549–554, 2014, doi: 10.1109/IAdCC.2014.6779384.

[28] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[29] E. Alyahyan and D. Düştegör, "Predicting academic success in higher education: literature review and best practices," *Int. J. Educ. Technol. High. Educ.*, vol. 17, no. 1, 2020, doi: 10.1186/s41239-020-0177-7.

Ermatita Paper 6 1570815512-Dafid

ORIGINALITY REPORT

84%	25%	84%	19%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Dafid, Ermatita. "Filter-Based Feature Selection Method for Predicting Students' Academic Performance", 2022 International Conference on Data Science and Its Applications (ICoDSA), 2022 Publication	78%
2	educationaltechnologyjournal.springeropen.com Internet Source	2%
3	Moohanad Jawthari, Veronika Stoffová. "Predicting students' academic performance using a modified kNN algorithm", Pollack Periodica, 2021 Publication	2%
4	Submitted to Telkom University Student Paper	1%
5	Eyman Alyahyan, Dilek Düştegör. "Predicting academic success in higher education: literature review and best practices", International Journal of Educational Technology in Higher Education, 2020 Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		