An Evaluation of Prediction Method for Educational Data Mining Based on Dimensionality Reduction

B. Vaidehi and K. Arunesh

Abstract In the area of educational data mining (EDM), it is important to develop technologically sophisticated solutions. An exponential growth in educational data raises the possibility that conventional methods could be constrained as well as misinterpreted. Thus, the field of education is becoming increasingly concerned in resurrecting data mining methods. This work thoroughly analyzes and predicts students' academic success using logistic regression, linear discriminant analysis (LDA), and principal component analysis (PCA) to keep track of the students' future performance in ahead. Logistic regression is enhanced by comparing LDA and PCA in a bid to improve precision. The findings demonstrate that LDA improved the accuracy of the logistic regression classifier by 8.86% as compared to PCA's output, which produced 35 more correctly classified data. As a result, it is demonstrated that this model is effective for forecasting students' performance using students' historical data.

Keywords Educational data mining · Linear discriminant analysis · Principal component analysis · Logistic regression · Data mining

1 Introduction

The use of statistics, learning algorithms, and data mining methodologies is the primary emphasis of data mining research in the field of EDM. The importance of data mining technology in the educational setting has grown over the last few decades. It has soared to great prominence in recent years as a result of the accessibility of open datasets and learning algorithms [1]. EDM entails the creation and implementation of data mining techniques that interpret the substantial amounts of

B. Vaidehi (✉) · K. Arunesh
Department of Computer Science, Sri S.Ramasamy Naidu Memorial College (Affiliated to Madurai Kamaraj University, Madurai), Sattur, Tamil Nadu 626203, India
e-mail: vaipri21@gmail.com

K. Arunesh
e-mail: arunesh_naga@yahoo.com

P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*, Lecture Notes in Networks and Systems 789, https://doi.org/10.1007/978-981-99-6586-1_7

data from various educational levels. Anticipating the learning process and evaluating student success are important objectives in the study of EDM [2]. It is a field which discovers underlying relationships and discovers trends in educational data. Heterogeneous data is contributing in the big data paradigm in the sector of education. In order to adaptively extract relevant information from educational datasets, specialized data mining techniques are required [3]. Many educational domains, including learning outcomes, dropout prediction, educational analysis of data, and academic and behavioral analysis, have used data mining methods [4]. EDM has always placed a premium on assessing and forecasting students' academic success. Higher education institutions must examine students based not only on their test results, but they should also consider how they learn, make projections about how they will perform academically in the future, and issue timely academic warnings. This work will assist students in raising their performance, which will enhance the management of educational resources while also assisting higher education in raising the quality of instruction [5]. The challenge of interpreting and making judgments from the enormous amount of information is growing progressively more onerous. The dimensionality is one of the primary challenges, although it can be solved by employing dimensionality reduction techniques. Dimensionality reduction refers the method of converting high-dimensional data into a meaningful less dimensionality. PCA [6] and LDA [7] are two well-liked techniques that have been extensively employed in various classification applications among the various dimensionality reduction approaches that have been developed. LDA employs label information; it can produce better classification results than PCA, which is unsupervised. This study applied the PCA and LDA algorithms for dimensionality reduction. The efficiency and effectiveness of PCA and LDA dimensionality reduction approaches are systematically evaluated in this work [8]. This work focused on evaluating students' academic achievement and to predict future success based on current performance. In order to reduce the dataset's dimensionality, this study suggests PCA and LDA and logistic regression as the dataset's classifier. Section 2 offers an analysis of previous works created by other researchers in the field of academic projection. Section 3 discusses aspects of the experimental methods. The experimental results are described and discussed in Sects. 4 and 5. The conclusion and prospective future approaches are identified in Sect. 6.

2 Related Study

Academic performance prediction has been one of the key goals of academic practitioners. Collaboration research has shown that effective procedures can be created for academic prediction using computational methods (such as data mining). For academic prediction, numerous researchers have created a variety of prediction models incorporating data mining.

Karthikeyan et al. [9] developed a novel method known as a hybrid educational data mining framework to evaluate academic achievement and effectively enhance

the educational experience. Crivei et al. [10] examined the applicability of unsupervised machine learning methods, particularly PCA and association rule mining, to assess student academic performance. EDM incorporates data mining techniques with educational data, according to Javier et al. [11]. In this, the well-known data mining methods are listed, including correlation mining factor analysis, and regression. Zuva et al. [12] provided a model which compares four classifiers in order to identify the best method for forecasting a learner's performance.

A key objective of the research will be to improve the current prediction algorithm in light of the requirement for an efficient prediction method. As a result, a model must be put out to improve the classification process.

3 Methodology

In this research work, the methodology was implemented by integrating the benefits of dimensionality reduction and classification. PCA and LDA are utilized in this work to lower the dimension, and also they are compared. PCA helps to eliminate features that are not essential to the model's goals, which reduces training time and expense and improves model performance [13]. LDA transforms a high-dimensional data into a low-dimensional by increasing between-class scatter and decreasing the within-class scatter. Logistic regression is employed in order to create our supervised classification for the dataset after doing dimensionality reduction. Figure 1 depicts the implemented methodology.

3.1 Dataset Description

The UCI machine learning repository's student dataset is used for this work. The dataset has 400 instances. The dataset consists of one target class and a total of 30 attributes. The dataset contains a total of 266 positive and 130 negative instances. The dataset's attributes are outlined below.

- Mother's Education
- Father's Education
- Home to School Travel Time
- Weekly Study Time
- Number of Past Class Failures
- Free Time After School
- Current Health Status
- Number of School Absences
- First Period Grade
- Second Period Grade
- Final Grade.

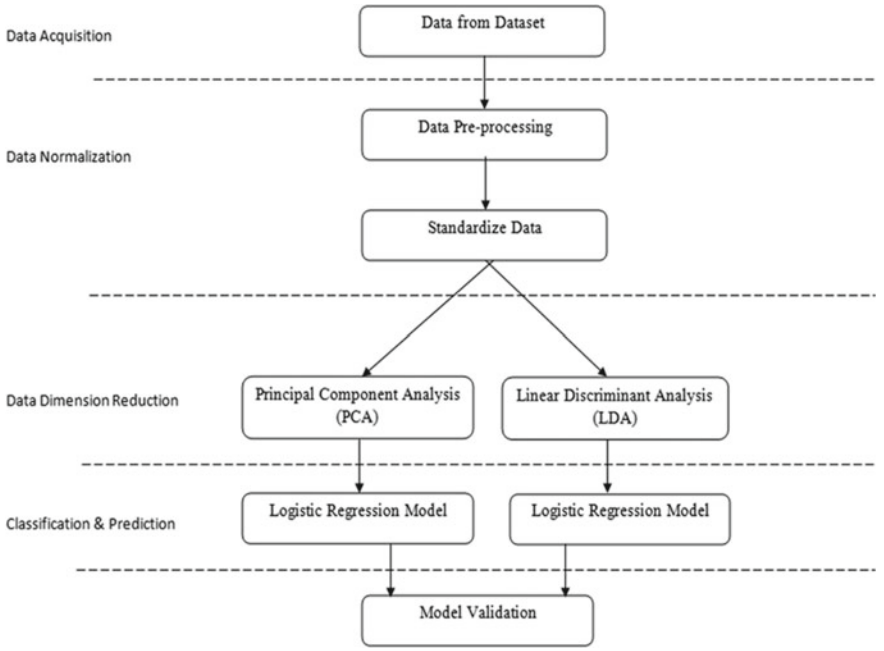

Fig. 1 Model implementation

3.2 Data Preprocessing

Due to enormous volumes and likely origin from diverse sources, real-world databases of today are especially prone to noisy, missing, and inconsistent data [14]. In the data mining process, data quality is crucial since poor data might produce predictions that are erroneous [15]. Data preprocessing overarching goal is to eliminate undesirable variability or impacts for effective modeling [16]. By doing normalization on the dataset, the existing data elements are scaled as part of data preprocessing so that they fall inside a narrow predetermined range of [0, 1] values. Speed will increase, and complexity will go down. Dataset V is normalized using the Z-score method to create a normalized value V' using the following equation:

$$V' = \frac{V - Y}{Z} \tag{1}$$

V' Normalized value,
V Value,
Y Mean,
Z SD.

3.3 Implemented Model

The research work consists of two phases. For the processed dataset, dimensionality reduction was done in the first stage. Supervised classification was employed in the second stage. The well-known dimensionality reduction methods PCA and LDA are investigated in this work. High-dimensional datasets are used for performance analysis. Logistic regression was used to classify data in order to compare how well the dimensionality reduction method is performed. These data were used to infer the differences between the supervised and unsupervised dimensionality reduction methods.

3.4 Principal Component Analysis

Data analysis and machine learning frequently employ the dimensionality reduction method known as PCA. Its primary function is to maintain the majority of the original data while downscaling a high-dimensional dataset into a lower dimensional space. This is accomplished by locating the principal components, which are linear combinations of the original characteristics that encompass the broadest range of data variance.

PCA discovers a significant subset of the estimated parameters with the maximum variance, known as the principle components PCs, that is, how PCA attempts to lower the dimension of the data. The initial PCs were accounted for the majority of the variance, making it possible to ignore with less information loss [17]. PCA is used to keep as much of the given dataset's information as feasible while also reducing the dimensionality of the enormous data [18]. The goal is to convert the dataset X, which has p dimensions, and Y, which has L ($L < p$) dimensions. Y is the PC of X, i.e.,

$$Y = PC(X) \tag{2}$$

(1) Configure Dataset

In X, there are n vectors ($x1$, $x2$, …, xn), which contain dataset instance.

(2) Determine Mean

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} \frac{x_i}{N} \tag{3}$$

(3) Determine the Covariance

$$C = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})^T \tag{4}$$

(4) Find Eigenvalues and Eigenvectors

The directions and magnitude of the new feature space will be determined by the eigenvectors and eigenvalues, respectively.

$$C = \lambda 1 > \lambda 2 > \cdots > \lambda N \quad \text{(Eigenvalues)} \tag{5}$$

$$C = u1, u2 \ldots uN \quad \text{(Eigenvectors)} \tag{6}$$

Creating a feature vector: According to eigenvalue, eigenvectors are ranked from the highest to lowest. This lists the elements in ascending order of importance. The primary element of the data collection is the eigenvector with the highest eigenvalue. The greatest eigenvalue is employed to create the feature vector [19–21]. Creating a new dataset involves selecting the principal components to keep in the data, creating a feature vector, and multiplying the vector by its transposition [19, 22–25].

3.5 Linear Discriminant Analysis

By maximizing between-class scatter and decreasing the within-class scatter, the LDA method reduces the dimensions. It allows dimensionality reduction without information loss and is mostly used prior to classification [18].

(1) Within-class scatter matrix

$$s_w = \sum_{j=1}^{c}\sum_{i=1}^{N_j}\left(x_i^j - \mu_j\right)\left(x_i^j - \mu_j\right)^T \tag{7}$$

c Number of classes
x_i^j ith sample of class j,
μ_j Mean of class j,
N_j Number of samples in class j.

Table 1 Comparison of accuracy with other studies

Paper	Methodology	Accuracy (%)
Proposed method	**LDA + logistic regression**	**97**
Jawad et al. [26]	Random forest classifier with SMOTE	96
Li et al. [12]	Deep neural network	78
Sassirekha et al. [27]	SLASAFP algorithm	96
Musso et al. [25]	ANN	80.7
Karalar et al. [17]	Optimal ensemble model	90.34
Imaran et al. [13]	J48 and MLP	95.78
Pujianto et al. [28]	KNN, C4.5	71.09
Tarbes et al. [29]	Bayesian network models	85
Echegaray et al. [30]	Genetic algorithm with an artificial neural network	84.86
Waheed et al. [31]	ANN, SVM, LR	93
Xu et al. [28]	DT, NN, SVM	76

Note: Bold represent better result

(2) Between-class scatter matrix

$$s_b = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T \tag{8}$$

μ Mean of all classes.

The between-class scatter determinant and within-class scatter determinants of the projected samples are optimized by LDA approaches [18] (Table 1).

3.6 Logistic Regression

Logistic regression is used when classifying data components. In logistic regression, the target variable is binary, which means that it only contains data that can be classified into two distinct groups: 1 or 0, which corresponds to a student who will be passed or failed in the academies. The aim of the logistic regression technique is to find the diagnostically reasonable model that best describes the relationship between the target variable and the predictor variable [15]. The Sigmoid equation below serves as the foundation for the logistic regression model [15]. Figure 2 depicts the Sigmoid function graph.

$$h_\theta(x) = \frac{1}{1 + e^{-z}}, z = \beta_0 + \beta_1 X \tag{9}$$

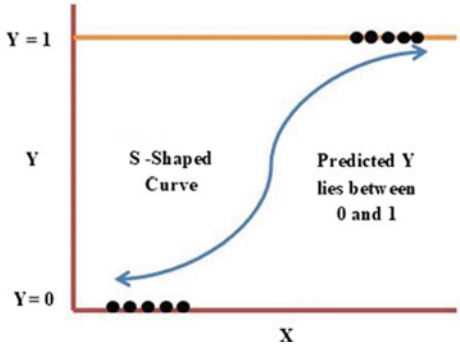

Fig. 2 Sigmoid function graph

The probability-based outcome or classes provided by the logistic regression classifier had probability score between 0 and 1.

$$\text{cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - (h_\theta(x))) & \text{if } y = 0 \end{cases} \tag{10}$$

The cost method serves as the goal of optimization. Optimizing the cost function in logistic regression to develop a precise model with minimal inaccuracies. The possibility of an event in the future is predicted using this model. The primary principle of logistic regression is to use a model based on the likelihood that an outcome will occur. Pseudocode 1 provides a description of the logistics regression model, which is used to train and test the data instance.

Pseudocode 1: Logistic Regression

1. **Input**: Featured Data
2. **Output**: Classified Data
3. For $i = 1$ to K
4. For Each data instance d_i
5. Set the Target Regression Value

$$Z = \frac{y_i - P(1 - d_j)}{[p - (1 - d_j).(1 - p(1 - d_j))]}$$

6. Initialize the weight of instance d_j to $P(1|d_j).\ (1 - P).\ (1|d_j)$
7. Finalize a $f(j)$ to the data with class value (z_j) and weights (w_j)
8. Assign (class label:1) if $P\ (1|d_j) > 0.5$, otherwise (class label:2).

4 Experimental Result

The student dataset, which has 400 instances and 30 attributes, is used in this work. The dataset statistics and description are given in Tables 2 and 3, respectively. The student dataset is used as the basis for performance analysis using the two different dimensionality reduction techniques, PCA and LDA, as well as logistic regression classifier. Dimensionality reduction during preprocessing was accomplished using the LDA and PCA methods. Then logistic regression is used to properly classify samples into defined groups. Prior to deploying a predictive model for implementation, it is crucial to ensure its effectiveness and accuracy. The results of the analysis and evaluation involve assessing various criteria, including Precision, Recall, and Accuracy. Table 5 illustrates the implemented model's performance metrics.

4.1 Employing Different Algorithms for Comparison

The student dataset is modeled with three distinct algorithms using the original dataset, PCA processed data, and LDA processed data in order to further assess how the model works. The outcome is shown in Table 4. LDA enhanced the performance accuracy of the other algorithms, but when Naive Bayes is employed an exception performance is found. As a result of PCA processing, the result in Table 4 shows decrease in Naive Bayes accuracy from 89 to 87%. Also, it was shown that LDA improved the algorithms' precision.

5 Discussion

The experimental findings shown that LDA improves classification accuracy than PCA. Jawad et al. [26] and Musso et al. [24] produced the similar finding, with a precision of 96% (Table 1). According to experimental findings, the proposed LDA approach increased logistic regression's classification accuracy for the student dataset. The accuracy of such model is determined by comparing it to the classification results published by other researcher's algorithms for academic prediction.

6 Conclusion and Future Work

The research work implemented an effective framework for predicting academic success. After carefully examining prior published works, this model combines the use of logistic regression for classification with LDA for dimensionality reduction.

Table2 Dataset statistics

	School	Sex	Age	Address	famsize	Pstatus	Medu	Fedu	traveltime	studytime
Count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
Mean	0.883544	0.526582	16.696203	0.777215	0.288608	0.896203	2.749367	2.521519	1.448101	2.035443
Std.	0.321177	0.499926	1.276043	0.416643	0.453690	0.305384	1.094735	1.088201	0.697505	0.839240
Min.	0.000000	0.000000	15.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000
25%	1.000000	0.000000	16.000000	1.000000	0.000000	1.000000	2.000000	2.000000	1.000000	1.000000
50%	1.000000	1.000000	17.000000	1.000000	0.000000	1.000000	3.000000	2.000000	1.000000	2.000000
75%	1.000000	1.000000	18.000000	1.000000	1.000000	1.000000	4.000000	3.000000	2.000000	2.000000
Max.	1.000000	1.000000	22.000000	1.000000	1.000000	1.000000	4.000000	4.000000	4.000000	4.000000
	freetime	goout	Dalc	Walc	Health	absences	*G*1	*G*2	*G*3	pass
Count	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000	395.000000
Mean	3.235443	3.108861	1.481013	2.291139	3.554430	5.708861	10.908861	10.713924	10.415190	0.670886
Std.	0.998862	1.113278	0.890741	1.287897	1.390303	8.003096	3.319195	3.761505	4.581443	0.470487
Min.	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000	3.000000	0.000000	0.000000	0.000000
25%	3.000000	2.000000	1.000000	1.000000	3.000000	0.000000	8.000000	9.000000	8.000000	0.000000
50%	3.000000	3.000000	1.000000	2.000000	4.000000	4.000000	11.000000	11.000000	11.000000	1.000000
75%	4.000000	4.000000	2.000000	3.000000	5.000000	8.000000	13.000000	13.000000	14.000000	1.000000
Max.	5.000000	5.000000	5.000000	5.000000	5.000000	75.000000	19.000000	19.000000	20.000000	1.000000

Table 3 Dataset description

Attribute	Explanation	Data type	Enum
medu	Mother's education	Numeric	{0, 1, 2, 3}
traveltime	Home to school travel time	Numeric	–
studytime	Weekly study time	Numeric	{1–10}
failures	Number of past class failures	Numeric	{n if $1 \leq n < 3$, else 4}
schoolsup	Extra educational support	Bool	{Yes, no}
internet	Internet access at home	Bool	{Yes, no}
Health	Current health status	Numeric	{1—very bad to 5—very good}
absences	Number of school absences	Numeric	{0–93}
$G1$	First period grade	Numeric	{0–20}
$G2$	Second period grade	Numeric	{0–20}
$G3$	Third period grade	Numeric	{0–20}
pass	Output target	Bool	{0, 1}

Table 4 Comparison of models using various methods

Method	Original dataset	PCA processed	LDA processed
Logistic regression	0.81	0.91	0.97
SVM	0.65	0.88	0.89
KNN	0.79	0.87	0.88
Naïve Bayes	0.89	0.87	0.94

Table 5 Performance metrics

Method	Precision	Recall	Accuracy
Logistic regression	0.89	0.72	0.81
PCA + logistic regression	0.90	0.92	0.91
LDA + logistic regression	0.98	0.96	**0.97**

Note: Bold represent better result

First, the LDA approach is used to our dataset with the goal of increasing classification accuracy. Although being a widely used approach, PCA's effectiveness in logistic regression has not garnered enough emphasis. In this research work, the integration of LDA and logistic regression can result better for predicting academic prediction. Also, the logistic regression model outperformed other algorithms employed in the work and findings from other studies in terms of prediction performance.

References

1. Antonio HB, Boris HF, David T, Borja NC (2019) A systematic review of deep learning approaches to educational data mining. Complexity 2019:1306039
2. Tsiakmaki M, Kostopoulos G, Kotsiantis S, Ragos O (2020) Implementing AutoML in educational data mining for prediction tasks. Appl Sci 10(1):90–117
3. Kausar S, Huahu X, Hussain I, Zhu W, Zahid M (2018) Integration of data mining clustering approach in the personalized E-learning system. IEEE Access 6:72724–72734
4. Buenaño-Fernandez D, Villegas W, Luján-Mora S (2019) The use of tools of data mining to decision making in engineering education—a systematic mapping study. Comput Appl Eng Educ 27(3):744–758
5. Feng G, Fan M, Chen Y (2022) Analysis and prediction of students' academic performance based on educational data mining. IEEE Access 10:19558–19571. https://doi.org/10.1109/ACCESS.2022.3151652
6. Turk M, Pentland A (2019) Face recognition using eigenfaces, computer vision and pattern recognition, proceedings CVPR'91. IEEE Comput Soc Conf Int J Emerg Technol Learn (iJET) 14(14):92
7. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell
8. Vikram M, Pavan R, Dineshbhai ND, Mohan B (2019) Performance evaluation of dimensionality reduction techniques on high dimensional data. In: 2019 3rd international conference on trends in electronics and ınformatics (ICOEI), Tirunelveli, India, pp 1169–1174. https://doi.org/10.1109/ICOEI.2019.8862526
9. Karthikeyan VG, Thangaraj P, Karthik S (2020) 'Towards developing hybrid educational data mining model (HEDM) for efficient and accurate student performance evaluation.' Soft Comput 24(24):18477–18487
10. Crivei LM, Czibula G, Ciubotariu G, Dindelegan M (2020) Unsupervised learning based mining of academic data sets for students' performance analysis. In: Proceedings of IEEE 14th internatonal symposium on application computer intelligence informatics (SACI), Timisoara, Romania, May 2020, pp 11–16
11. Javier BA, Claire FB, Isaac S (2020) Data mining in foreign language learning. WIREs Data Min Knowl Discov 10(1):e1287
12. Li S, Liu T (2021) Performance prediction for higher education students using deep learning. Complexity 2021:1–10
13. Imran M, Latif S, Mehmood D, Shah MS. Student academic performance prediction using supervised learning techniques
14. Pang Y, Yuan Y, Li X (2008) Effective feature extraction in high dimensional space. IEEE Trans Syst Man Cybern B Cybern
15. Zhu C, Idemudia CU, Feng W (2019) Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques. Inform Med Unlock 17:100179
16. Archana HT, Sachin D (2015) Dimensionality reduction and classification through PCA and LDA. Int J Comput Appl 122(17):4–8. Available at https://doi.org/10.5120/21790-5104
17. Karalar H, Kapucu C, Gürüler H (2021) Predicting students at risk of academic failure using ensemble model during pandemic in a distance learning system. Int J Educ Technol Higher Educ 18(1)
18. Ramaphosa KIM, Zuva T, Kwuimi R (2018) Educational data mining to ımprove learner performance in gauteng primary schools. In: 2018 ınternational conference on advances in big data, computing and data communication systems (icABCD), pp 1–6. https://doi.org/10.1109/ICABCD.2018.8465478
19. Han J, Kamber M, Pei J (2012) Data mining concepts and techniques, 3rd edn. Morgan Kaufmann Publishers, USA
20. Mishra P, Biancolillo A, Roger JM, Marini F, Rutledge DN (2020) New data preprocessing trends based on ensemble of multiple pre- processing techniques. TrAC Trends Anal Chem 132:116045

21. Fan C, Chen M, Wang X, Wang J, Huang B (2021) A review on data pre-processing techniques toward efficient and reliable knowledge discovery from building operational data. Front Energy Res 9:652801
22. Smith LI (2002) A tutorial on principal components analysis
23. Yağcı M (2022) Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn Environ 9(1)
24. Musso MF, Hernández CFR, Cascallar EC (2020) Predicting key educational outcomes in academic trajectories: a machine-learning approach. High Educ 80(5):875–894
25. Waheed H, Hassan SU, Aljohani NR, Hardman J, Alelyani S, Nawaz R (2020) Predicting academic performance of students from VLE big data using deep learning models. Comput Human Behav 104:106189
26. Jawad K, Shah MA, Tahir M (2022) Students' academic performance and engagement prediction in a virtual learning environment using random forest with data balancing. Sustainability 14(22):14795
27. Sassirekha MS, Vijayalakshmi S (2022) Predicting the academic progression in student's standpoint using machine learning. Automatika 63(4):605–617
28. Pujianto U, Agung Prasetyo W, Rakhmat Taufani A (2020) Students academic performance prediction with K-nearest neighbor and C4.5 on smote-balanced data. In: 2020 3rd international seminar on research of information technology and intelligent systems (ISRITI)
29. Tarbes BJ, Morales P, Levano M, Schwarzenberg P, Nicolis O, Peralta (2022) Explainable prediction of academic failure using Bayesian networks. In: 2022 IEEE ınternational conference on automation/XXV congress of the chilean association of automatic control (ICA-ACCA)
30. Echegaray-Calderon OA, Barrios-Aranibar D (2015) Optimal selection of factors using genetic algorithms and neural networks for the prediction of students' academic performance. In: 2015 Latin America congress on computational ıntelligence (LA-CCI)
31. Xu X, Wang J, Peng H, Wu R (2019) Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Comput Hum Behav 98:166–173