INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Chi-Square Feature Selection Technique for Student's performance prediction

Himanshi Bhoria[1]*, Amita Dhankhar[2], Kamna Solanki[2]

1 M.Tech. Student Department of CSE, UIET, MDU Rohtak, India
2 Associate Professor, Department of Computer Science & Engineering, University Institute of Engineering and Technology, MDU Rohtak, India

OPEN ACCESS

Received: 19-04-2023

Accepted: 29-08-2023

Published: 13-10-2023

Citation: Bhoria H, Dhankhar A, Solanki K (2023) Chi-Square Feature Selection Technique for Student's performance prediction. Indian Journal of Science and Technology 16(38): 3250-3257. https://doi.org/10.17485/IJST/v16i38.921

* **Corresponding author**.

himanshibhoria07@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The main goals of this study are: 1) To assess students' performance using several machine learning models. 2) To identify the attributes influencing the student's performance using feature selection. 3) To assess and compare machine learning model performance using accuracy, precision, recall, F-1 score, and AUC score (Area Under Curve) as performance indicators. 4) Compare the effectiveness of feature selection-based versus non-feature-based machine learning models. **Methods:** The student performance dataset from UCI has been taken for this study. It consists of 650 records with 32 features. The pertinent features are selected by applying the Chi-square method to facilitate the effective construction of the model. Further, the implementation has been performed by using the classification models. Lastly, how well the machine learning model has performed has been compared in terms of performance metrics namely accuracy, precision, recall, F-1 score, and AUC score. **Findings:** The findings related to the first objective showed that the outcome of the student performance is passed and failed. The experimental evaluation of the Decision tree (DT), random forest (RF), SVM, K-Nearest Neighbors Algorithm (KNN), and XGBoost are evaluated in terms of accuracy, precision, recall, F-1 score, and AUC score. The F-1 score achieved by the DT, RF, SVM, KNN, and XGBoost is 92.16, 95.06, 95.19,93.8 and 94.59 respectively. The finding to the second objective identifies the attributes: Failures, Schoolsup, First Period Grade(G1), Second Period Grade(G2), and Final Grade(G3) influence on students' performance. The finding of the third Objective shows that Support Vector Machine classification model outperforms the other models with F-1 score of 95.19%. The finding related to the fourth objective identifies that Models with use feature selection techniques give more performance than the model which does not use it. **Novelty:** Using machine learning to predict students' performance can revolutionize the education sector by providing a data-driven approach to evaluating academic performance. This research work proposed a new "Chi-Square Based Feature Selection" (CBFS) technique for the prediction of students' performance. Moreover, using chi-square for feature selection involves selecting only the most relevant features, which helps reduce the model's complexity and improves its performance.

Keywords: Machine Learning; Prediction; Dataset Problem; Early Warning System; Educational Data Mining

1 Introduction

Education is necessary for a country to advance and for an individual to succeed. Education institutions work hard to provide students with a high-quality education while making an effort to improve the learning process. The success of educational institutions is greatly impacted by student academic performance.

Studying educational datasets gathered from higher education institutions and e-learning environments uses machine learning (ML) and artificial intelligence (AI) techniques. When assessing many facets of education, EDM uses data mining techniques, including time series analysis, regression, classification, and association rule mining, to uncover insightful patterns and information. These EDM predictive models can offer insights supporting education and learning processes.

Educational institutions are increasingly incorporating Al technology to improve the learning experience for students. Providing high-quality education and improving student success rates pose a significant challenge for these institutions. Machine learning (ML) plays a vital role in education by predicting students' future academic performance and helping them attain higher grades. Anticipating students' academic success is crucial as it allows for the early identification of those at risk of failure during the semester. By identifying these students early on, educational institutions can provide them with suitable treatments to improve their academic results prior to the final assessment, hence enhancing the university's success rate [(1)].

Data Mining (DM) is the discovery of data [(2)]. Data Mining forecasts various educational outcomes, including achievement in performance, retention, dropout rate, and success [(3)]. Using data Mining methods in education is extremely beneficial, especially when analyzing and forecasting students' academic achievement.

The early prediction of students' academic performance during a semester is an invaluable tool for implementing timely interventions to improve their outcomes and decrease failure rates by the semester's end. However, accurately predicting academic performance poses a significant challenge due to various factors influencing student success, such as academic background, prior accomplishments, demographic factors, economic situation, behavioural traits, and other variables. In this context, Educational Data Mining (EDM) is vital to address this challenge [(4)]. One of the most common uses of EDM is forecasting students' future performance using previous academic data, which offers crucial insights to improve student accomplishments, lower failure rates, and obtain a thorough grasp of the learning process.

Massive amounts of educational data are being produced by academic institutions nowadays, This is used to enhance decision-making and improve student achievement through data analytics. This practice can improve educational settings holistically and foster a deeper comprehension of the learning process [(5)].

Several researchers have conducted studies in the field of prediction models, and here we provide an overview of some of their notable work. Dijana Oreški et al. [(6)] utilized an online dataset of 263 students from a Croatian university. They applied the CRISP-DM standards in decision trees and achieved an accuracy of 73.6%, which was the best-performing metric. Safira Begum et al. [(7)] used an online dataset from the UCI Repository and employed KNN, LDA, and SVM methodologies. They evaluated accuracy, normalization, and z-score as measuring metrics, and SVM yielded the best result with 67.69%. In [(8)], the author worked with a dataset of 6,807 records obtained from an online survey. Various machine learning methodologies such as Random Forest, Logistic Regression, Support Vector Classifier, Voting, Decision Tree, Bagging, MLP, and AdaBoost were utilised, with Random Forest achieving the highest F1-score

of 93.8%. M. Kumar et al.[9] collected an online dataset of 500 students and focused on accuracy scores. They compared different methodologies including Naïve Bayes, Decision Table, MLP, and J48 Ensemble Methods (Bagging, Random Sub Space, and AdaBoost), with AdaBoost achieving the best accuracy score of 80.33%. Y. K. Salal et al.[10] worked with an online dataset of 388 records and compared Random Forest, Logistic Regression, and K-Nearest Neighbour. Random Forest has the best accuracy of 93%. In J. Malini's[11], ANN, Bagging, and Boosting were employed, and accuracy, precision, recall, false positive rate, F1-measure, true positive rate, and confusion matrix were used as measuring metrics. Bagging achieved the highest accuracy score of 88% using an online dataset of size 649. J. Dhilipan et al[12] worked with student data using 10,12th and previous subject marks and it was found that binomial logical regression obtained the highest accuracy of 97.05%. J. Gajwani[13] utilized an online dataset of 500 records and employed Decision Tree, Logistic Regression, Naïve Bayes, and ensemble algorithms. The accuracy metric was used, and Boosting achieved the best result with an accuracy score of 75%. U. Pujianto et al.[14] worked with an online dataset of size 500 and compared Decision Tree and K-Nearest Neighbour. Based on accuracy score, Decision Tree performed the best with an accuracy of 71.09%. In K. A. Mayahi et al.'s research[15], an offline dataset of 550 records was used. Support Vector Classifier (SVC) and Naïve Bayes models were employed, and accuracy, precision, and recall were the measuring metrics. SVC was determined to be the best model, achieving an accuracy of 87%. These studies contribute valuable insights into prediction models, encompassing various methodologies, datasets,

Machine Learning models are primarily used to predict future outcomes based on available data. Educational institutions are increasingly interested in using these techniques to forecast the performance of their enrolled students. Previous studies have predominantly focused on finding the best prediction models, student performance in specific courses, grade inflation, and identifying struggling/failing students. However, there is a notable research gap in using irrelevant or redundant features in the dataset, and this impact the performance of the classifier.

To address this gap, our study aims to conduct a comparative analysis of various machine learning models with and without using Chi-Square feature selection techniques to show the difference how the selected attributes affect the data. It is applied to various classifiers in order to identify the importance of using the selected attributes by a selection technique. By analyzing the complete dataset, we can determine which attributes of the model have the most influence on students' outcomes. For this analysis, we employed Jupyter, an open-source data mining tool written in Python. Jupyter provides a comprehensive range of machine-learning algorithms specifically designed for real-world data mining challenges. In this study, we used a real dataset from the University of California, Irvine (UCI) to build our predictive models utilizing five methods: Decision Tree, Random Forest, SVM, KNN, and XGBoost. In our study, we have also compared the performance of different algorithms using Jupyter, analyze and evaluate student data in order to anticipate their success and failure in the Portuguese language, and thoroughly discussed the results acquired.

2 Methodology

Figure 1 shows the flow chart of the proposed Chi-Square feature selection technique.

2. 1 Data Collection

2.1.1 Data Selection Preparation

In the experiment we collected real data is from UCI i.e the University of California, Irvine, https://archive.ics.uci.edu/ml/datasets/student+performance of under-graduates students, where we have taken the Portuguese language. It contains 650 records of students and the records are based on 32 features which consist of their last 2 academic year, their final grade, their school, name, age, and many more.

2.2 Data Pre-Processing

2.2.1 Data Cleaning

Data cleaning plays a crucial role in preparing raw data collected from the real world for analysis, aiming to rectify errors and yield improved results. The initial step in data cleaning involves examining the presence of null values within the dataset. In the current dataset, no null values were identified. The chi-square test was used to examine the relationship between the category variables. The chi-square test assumes the absence of any correlation between categorical variables and was conducted with a significance level of 0.05. The analysis revealed correlations between 5 attributes that include Failures, Schoolsup, First Period Grade (G1), Second Period Grade (G2), and Final Grade (G3).

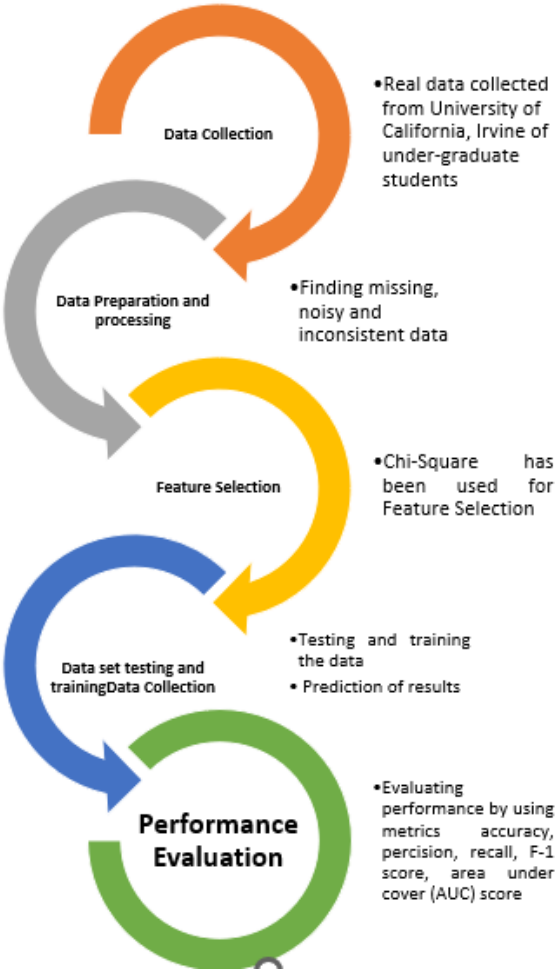

Fig 1. Flow chart of proposed approach

2.2.2 Data Processing

For the Processing of data, we convert the Final Grade(G3) into binary values to represent our data in the dataset. For that, the Grades in G3 greater than 9 are converted into 1 and less than 9 into 0 to make the algorithm learn more readily and get better outcomes as indicated in Table 1. In addition to that nominal values of Pass and Fail are also added in this study.

Table 1. Result of data processing

Input Feature	Student Marks	Class Label
1	>=9	Pass
0	< 9	Fail

So after processing the data, we evaluated that among 650 students, 489 students i.e 75.2 % Of the students pass and 161 students i.e. 24.8% of students will fail in the class.

2.3 Prediction

To examine the performance of the models in our study, we used five evaluation measures: accuracy, precision, recall, F1-score, and AUC (Area Under Curve) score. Accuracy, being a widely used metric, was utilized to evaluate the effectiveness of the machine learning models.

3 Results and Discussion

3.1 Environment Used

The experiment is carried out on a PC equipped with a core i5 processor and 16 GB of RAM. Anaconda Software (Jupyter) was utilized to analyze the predictive model.

The model used is Decision Tree, Random Forest, SVM, KNN, and XGBoost to experiment. The evaluated measures used are Accuracy, Precision, Recall, F1 Score, and AUC.

The performance metric that has been utilized the most often in prior research is accuracy. If the dataset has the same number of cases of each type, accuracy may be valuable. In the absence of this, accuracy is useless because it predicts the value of the majority class. Therefore, F-1 incorporates essential findings about the Classifier effectiveness in each category and is regarded as the average recall and accuracy value. When there are distinct class distributions, it is quite helpful.

The four categories were true positive (TP), true negative (TN), false positive (FP), and false negative (FN) into which the student's cases were placed. In our study, we employed the following criteria for quality assurance.

This example is given in the context of student's academics whether they will Pass or Fail.

True Positive (TP): A Student Actually Pass(Positive) and through experiment, classified as Pass(Positive). This is called True Positive.

True Negative(TN): A Student Actually Pass(Positive) and through experiment classified as Fail(Negative). This is called True Negative.

False Positive(FP): A Student Actually Fail(Negative) and through experiment classified as Pass(Positive). This is called False Positive.

False Negative(FN): A Student Actually Fail(Negative) and through experiment classified as Fail(Negative). This is called False Negative.

- **Confusion Matrix**

A confusion matrix is used to depict these classifier results based on TP, TN, FP, and FN, which may be displayed as

- **Accuracy**

It is described as several correct predictions and overall predictions. The formula is

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

The Accuracy score of Random Forest is the highest among all models which is 95.38% as shown in Figure 2 by using selection techniques with using feature selection and without using feature selection the highest accuracy is scored by SVM of 93.07%.

- **Precision**

It indicates how many positive forecasts are correct (true positives). The formula is

$$Precision = \frac{TP}{TP + FN}$$

The Precision Score of XGBOOST is the highest among all the models with a percentage of 95.75% as presented in Figure 3 with using feature selection and without using feature selection we get a precision score of 91.79% of XGBoost.

- **Recall**

It is a measure of how many positive cases the classifier anticipated correctly out of all the positive cases in the data. The formula is

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

The Recall Score of SVM and XGBOOST is the highest among all the models with a recall score of 99.09% and without using feature selection Recall score is highest at 99.08% of SVM.

- **Area Under Curve(AUC) Score**

It is an evaluation of the binary classifier's ability to distinguish among categories and acts as a summary of the ROC (Receiver Operator Characteristic) curve.

The AUC Score of Random Forest is the highest among all the models with 80.68% and without using feature selection AUC Score is 78.8% is the highest of the Random Forest Model.

- **F1-Score**

F1-Score is a measure combining both precision and recall. The term "harmonic mean" is often used to describe this middle ground between the two. The harmonic mean is an alternative to the arithmetic mean for establishing an "average" of numbers, and over the arithmetic mean, ratios (such as recall and accuracy) are frequently favored. Here's how to calculate an F1 score

$$F1 - Score = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right)$$

The F1 score of SVM is the highest among all the models with a percentage of 95.19% using feature selection and 93.57% without using the feature selection. Figure 2 shows the Experimental Results for Various Parameters using CBFS Technique.

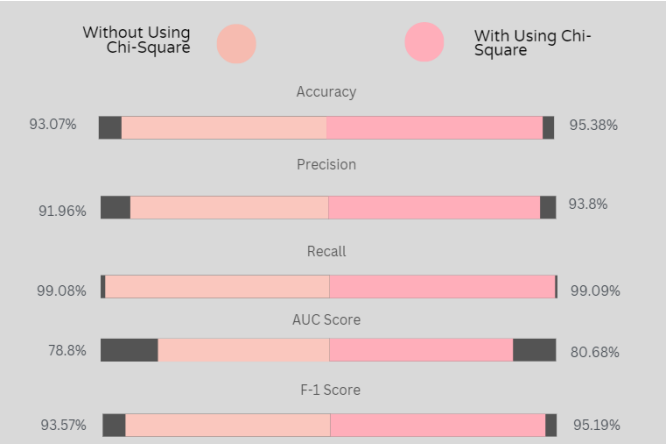

Fig 2. Experimental Results for Various ParametersUsing CBFS Technique

An experimental study was also conducted on the same dataset without using the Feature Selection Techniques and different Machine Learning Models. It is evident from Table 2 that applying the chi-square technique for feature selection in the dataset which consists of 650 records and 32 attributes resulted in selecting only the most relevant features which are 5 attributes namely Failures, Schoolsup, First Period Grade (G1), Second Period Grade (G2), and Final Grade (G3).

This process effectively reduced the model's complexity and led to improved performance. Consequently, the SVM model achieved an F-1 score of 95.19%, surpassing the F-1 score of 93.57% obtained when the selective technique was not utilized as shown in Figure 3.

So According to the evaluation, it was found that in the total set of 650 data, 498 Students will pass and 161 students will fail in academics, and along with that F1- Score of SVM(Support Vector Machine) is the highest with the percentage of 95.19% as shown in Table 3. In this research, we used 70% training and 30% testing data. So SVM is the best model to do prediction. And It is also found that by using the feature selection technique models gave more performance.

This work introduces a new machine learning-based model for predicting undergraduate students' final exam scores using intermediate grades as the data input, to calculate undergraduate students' final exam scores. The study examines the performance of numerous machine learning algorithms in predicting final exam scores, including decision trees (DT), random forest (RF), SVM (Support Vector Machine), K-Nearest Neighbours Algorithm (KNN), and XGBoost. Two main aspects are emphasized in this study. The first element is forecasting academic performance based on past achievement grades, while the second entails comparing the performance indicators of various machine learning algorithms utilising chi-square for feature selection.

Table 2. Performance Evaluation of Proposed CBFS technique against Contemporary techniques without feature Selection

	Evaluation Measures/ML Models	Accuracy	Precision	Recall	AUC	F-1 Score
Proposed Chi-Square Feature Selection Technique	Decision Tree (DT)	87.67	93.45	90.09	77.95	92.16
	Random Forest (RF)	95.38	93.8	96.36	80.68	95.06
	Support-Vector Machine (SVM)	93.07	91.59	99.09	75.54	95.19
	K-Nearest Neighbors (KNN)	90.76	92.98	94.64	75.09	93.8
	XGBOOST	91.53	93.75	99.09	80.22	94.59
Existing Technique without Feature Selection	Decision Tree (DT)	86.87	91.96	89.96	77.33	91.89
	Random Forest (RF)	92.38	90.59	96.33	78.8	92.45
	Support-Vector Machine (SVM)	93.07	89.2	99.08	75.34	93.57
	K-Nearest Neighbors (KNN)	87.67	91.3	94.59	70.98	92.92
	XGBOOST	90.48	91.76	98.4	76.76	91.3

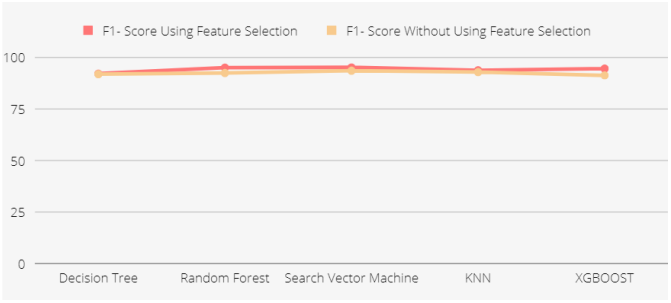

Fig 3. Comparison of the different models on measuring metrics F-1 Score with and without using Feature selection.

Table 3. Performance Comparison of CBFS technique

	Research Findings-Existing Techniques	Research Findings- Proposed CBFS Technique
(11)	Achieved the highest accuracy of 88% by using Bagging technique.	Using Proposed CBFS technique, SVM achieves the highest accuracy of 93.07%.
(7)	SVM achieved the highest accuracy of 67.69%.	Using Proposed CBFS technique, SVM achieve the highest accuracy of 93.07%.
(6)	This research employed on a smaller dataset of 263 records	A comparatively bigger dataset of 650 records.

The performance of proposed CBFS technique findings was compared to those of other studies that looked at how early-semester performance in one class correlates to later success in the same or similar classes. The performance comparison against contemporary research work has been shown in Table 3.

To address these limitations, using a real-world dataset, we not only find the model accuracy but we have also found that among 650 records of data, 498 Students will pass and 161 students will fail in academic study. As a result, we achieved an impressive accuracy of 95.38% and an F-1 Score of 95.19%. The proposed CBFS approach outperformed all other evaluated techniques, as evidenced by the experimental results.

4 Conclusion

The enormous quantity of educational data kept in educational settings must be analyzed, and this is where educational data mining comes in. It aids in decision-making processes, predicting students' academic performance early and uncovering valuable insights from educational data. However, one common challenge in predicting academic performance is dealing with imbalanced datasets, which can lead to suboptimal results.

In our study, we utilized a dataset obtained from the University of California, Irvine to develop predictive models using various machine learning algorithms, including Decision tree , random forest, SVM, K-Nearest Neighbors Algorithm (KNN), and XGBoost. Our goal was to forecast students' academic achievement in Portuguese based on grades from past courses taken during the academic year. To address this problem of feature selection, we used approaches such as the Chi-Square Test to improve the performance of the models.

The purpose of this study was to demonstrate the impact of feature selection on model performance and to investigate approaches to improve model performance. As a feature selection strategy, we specifically used the chi-square test. We employed two approaches for model validation: By using the feature selection technique and without using the feature selection technique.

Our findings highlighted the influence of feature selection on model performance. We observed that the classifiers' performance was unsatisfactory when dealing without Feature Selection. However, we achieved significant improvements and better outcomes when working with Feature Selective Technique. There are 32 attributes in the dataset resulting in selecting only the most relevant features which are 5 attributes. The Chi-Square test yielded superior results, we obtained reliable and accurate results. The attributes that affect the dataset are Failures, Schoolsup, First Period Grade (G1), Second Period Grade (G2), and Final Grade (G3). We used 70% training and 30% testing data. The SVM (Support Vector Machine) model showcased its superiority, achieving the F1 score of 95.19% Among the dataset of 650 records 495 students will pass and 161 students will fail in Studies. To enhance the model's accuracy for future work, we will expand our dataset and include data from additional semesters.

References

1) Embarak O. Apply Machine Learning Algorithms to Predict At-Risk Students to Admission Period. *2020 Seventh International Conference on Information Technology Trends (ITT)*. 2020;p. 190–195. Available from: https://doi.org/10.1109/ITT51279.2020.9320878.
2) Yağcı M. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*. 2022;9(1). Available from: https://doi.org/10.1186/s40561-022-00192-z.
3) Sekeroglu B, Dimililer K, Tuncal K. Student Performance Prediction and Classification Using Machine Learning Algorithms. In: Proceedings of the 2019 8th International Conference on Educational and Information Technology. ACM. 2019;p. 7–11. Available from: https://doi.org/10.1145/3318396.3318419.
4) Farissi A, Dahlan HM, Samsuryadi. Genetic Algorithm Based Feature Selection for Predicting Student's Academic Performance. *Advances in Intelligent Systems and Computing*. 2020;p. 110–117. Available from: https://www.researchgate.net/publication/337534319_Genetic_Algorithm_Based_Feature_Selection_for_Predicting_Student's_Academic_Performance.
5) Kouser F, Meghji AF, Mahoto NA. Early Detection of Failure Risks from Students' Data. *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*. 2020;p. 1–6. Available from: https://doi.org/10.1109/ICETST49965.2020.9080692.
6) Oreški D, Zamuda D. Machine Learning Based Model for Predicting Student Outcomes. *International Conference on Industrial Engineering and Operations Management Istanbul*. 2022. Available from: https://ieomsociety.org/proceedings/2022istanbul/967.pdf.
7) Begum S, Padmannavar SS. Prediction of Student Performance using Genetically Optimized Feature Selection with Multiclass Classification. *International Journal of Engineering Trends and Technology*. 2022;70(4):223–235. Available from: https://doi.org/10.14445/22315381/IJETT-V70I4P219.
8) Deepti A, Sonu M, Vikram B. Significance of NonAcademic Parameters for Predicting Student Performance Using Ensemble Learning Techniques. *International Journal of System Dynamics Applications (IJSDA)*. 2021;10:38–49. Available from: https://doi.org/10.4018/IJSDA.2021070103.
9) Kumar M, Mehta G, Nayar N, Sharma A. EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance in Academics. *IOP Conference Series: Materials Science and Engineering*. 2021;1022(1):012062. Available from: https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012062/pdf.
10) Salal YK, Hussain M, Theodorou P. Student Next Assignment Submission Prediction Using a Machine Learning Approach. In: Lecture Notes in Electrical Engineering;vol. 729. Springer International Publishing. 2021;p. 383–393. Available from: https://doi.org/10.1007/978-3-030-71119-1_38.
11) Malini J, Kalpana Y. Investigation of factors affecting student performance evaluation using education materials data mining technique. *Materials Today: Proceedings*. 2021;47:6105–6110. Available from: https://doi.org/10.1016/j.matpr.2021.05.026.
12) Dhilipan J, Vijayalakshmi N, Suriya S, Christopher A. Prediction of Students Performance using Machine learning. *IOP Conference Series: Materials Science and Engineering*. 2021;1055(1):012122. Available from: https://iopscience.iop.org/article/10.1088/1757-899X/1055/1/012122/pdf.
13) Gajwani J, Chakraborty P. Students' Performance Prediction Using Feature Selection and Supervised Machine Learning Algorithms. In: Advances in Intelligent Systems and Computing. Springer Singapore. 2021;p. 347–354. Available from: https://doi.org/10.1007/978-981-15-5113-0_25.
14) Pujianto U, Prasetyo WA, Taufani AR. Students Academic Performance Prediction with k-Nearest Neighbor and C4.5 on SMOTE-balanced data. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. 2020;p. 348–353. Available from: https://doi.org/10.1109/ISRITI51436.2020.9315439.
15) Mayahi KA, Al-Bahri M. Machine Learning Based Predicting Student Academic Success. *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. 2020;p. 264–268. Available from: https://doi.org/10.1109/ICUMT51630.2020.9222435.