Journal of Intelligent Systems and Internet of Things
Vol. 10, No. 02, PP. 24-37, 2023
DOI: https://doi.org/10.54216/JISIoT.100203

Predicting Student Performance Using Educational Data Mining and Learning Analytics Technique

Rahul Sharma [1], Shiv Shakti Shrivastava [2], Aditi Sharma [3,4*]

[1] Department of Computer Science and Engineering, Rabindranath Tagore University, Raisen, (M.P.), India
[2] Department of Computer Science and Engineering, Rabindranath Tagore University, Raisen, (M.P.), India
[3] Department of Computer Science and Engineering,
Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India
[4] IEEE Senior Member, Symbiosis Institute of Technology, Pune, India
Emails: sharma.rahul5656@gmail.com; shivshakti18@gmail.com; aditi.sharma@ieee.org
*Corresponding Author: aditi.sharma@ieee.org

Abstract

Data analysis is an essential component of decision support in various industries that includes industrial and educational institutions. This research proposes Data Mining (DM) techniques to improve the efficiency of higher education (HE) institutions. DM has a substantial impact on different higher education activities including student performances, management of student's life cycle, selection of courses, monitoring of retention rate, grants & funds management by using technique's such as clustering, decision trees (DT), and association. Educational Data Mining (EDM) is an interdisciplinary study topic that focuses on getting DM to the fields of education by leveraging methods from (ML) statistics, (DM), and (DA) to get important insights from educational sets of data. EDM is critical in transforming raw data into useful information, allowing for a greater knowledge of students and their academic settings, as well as promoting better teacher assistance and ESD (Educational System Decisions). The study's goal is to provide a complete overview of EDM (Educational Data Mining), highlighting its various applications and benefits in the context of higher education.

Received: April 07, 2023 Revised: July 02, 2023 Accepted: October 05, 2023

Keywords: EDM (Educational Data Mining); DM (Data mining) techniques; Data processing methods; Knowledge discovery in databases (KDD); Learning analytics (LA); EDM tools, and Visualizations tools were all examples of data mining strategies (DMS);

1. Introduction:

Educational Data Mining (EDM), and Learning Analytics have grown in popularity as alternative to standard frequentist, Bayesian methods (BM) to analyze educational data. DM, also known as 'Knowledge Discovery in Databases' (KDD), is the processes of examining enormous educational databases in order to identify novel and generalizable patterns and insights, rather than validating pre-existing views. With an ongoing expansion of data at universities due to advances in technology and decreasing IT expenses, the necessity for competent data assessment is essential. Using statistics, artificial intelligence (AI), neural networks (NN), machine learning(ML) and other techniques, data mining (DM) tools, approaches, and tactics can find hidden patterns and important information. It is an interdisciplinary science that focuses on making informed judgements by uncovering critical connections, patterns, and trends in massive datasets [1].
Data Mining has numerous uses in a variety of sectors. In finance, it analyses customer behavior data for greater client loyalty and reveals hidden linkages between financial indicators to detect fraudulent and non-fraudulent activity. It identifies links between diseases and their treatment outcomes in healthcare, which aids in the detection of healthcare insurance fraud. Mining data is utilized by law enforcement to identify illegal behaviors such as laundering funds and drug trafficking. Through

demographic research and behavior prediction, telecommunications enterprises use it to deliver personalized services, reduce customer, and increase profitability. Data mining in marketing and sales identifies trends in purchasing data, allowing market basket research and forecasting of future purchasing behaviors. Data mining is used in the banking industry for predicting client attrition and identify fraud and bankruptcy threats [2].

Although data mining may offer useful data, it additionally comes with several significant downsides, specifically with regard to user privacy and security. Information use and sharing procedures must be open and clear. Due to the processing of large datasets, selecting the most effective methods for analysis and competent IT skills for data preparation make data mining expensive to implement. Furthermore, since data mining (DM) is not perfect, errors could have adverse consequences and cost money. Figure 1 illustrates the users of educational data mining to examine the results of instructional methods. While students participate individually or in groups, lecturers plan and present lessons using a variety of techniques. Data mining offers helpful insights for educational development by classifying, spotting trends, and connecting concept with the data. [3]

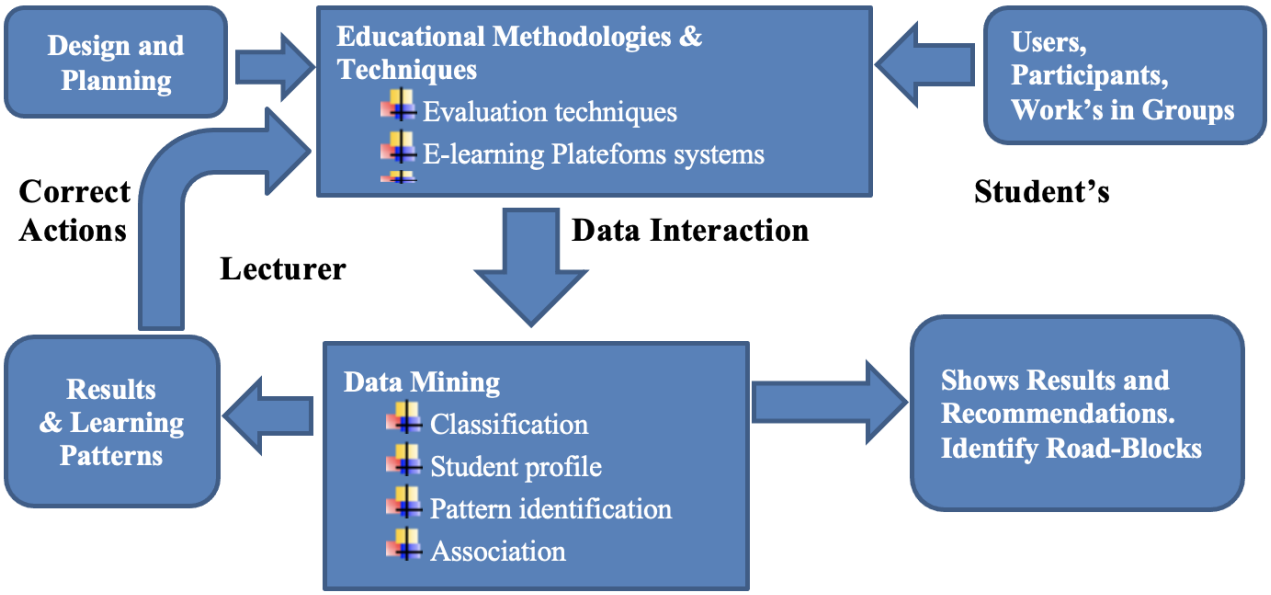

Figure 1: DM for identification of the SPL 'Student's Learning Pattern

2. Related Work

Previous evaluations on the topic have shown that academic research on data mining applications in education has achieved notable success. In these research, different methods of data mining have been used to support various educational tasks. These applications include anticipating student cognitive abilities in a classroom setting, locating at risk for students and slow learners, estimating courses &career choice, managing retention of student, predicting overall performance of student or learners. Predicting student success stands out among these element's as one of the most important and beneficial parts of educational data mining (EDM).

Estimating an unknown value, such as the students grades and marks, is required to be able to predicting a student's performance [4]. The use of this predictive skill by teachers and educational institutions can be very advantageous. Targeted interventions and individualized support can be given learners who are at risk of falling behind or becoming underperforming in order to enhance their learning outcomes. Predicting student performance can also help with curriculum design, method of instruction modification, and resource allocation optimization.

In the final analysis, the use of data mining in education facilitates evidence-based decision-making and provides educators with crucial insights into the development and needs of their learners. The field has a lot of promise for improving the total educational experience while promoting pupils to succeed as it progresses.

3. Educational Data Mining (EDM)

The necessity to manage significant educational data collections that are too large for manual analysis gave rise to the burgeoning discipline of EDM. It is an example of an interdisciplinary research area that uses data mining technologies to look at data gathered through education and learning activities. To be able to gain a deeper understanding of educational dynamics, EDM employs techniques from machine learning, statistics, and data analysis. The aim is to glean useful insights, previously undisclosed data, patterns from big/large data repositories [5].

3.1 EDM (Educational Data Mining Process)

There are four essential steps in the EDM process. First, problem definition entails defining research goals and themes along with converting a particular concern into an info mining issue. Second, most time-consuming stage, preparing data and gathering, focuses on making sure data quality by properly acquiring, cleaning, and formatting data. Thirdly, choosing the best parameters and using different methods of modelling are part of the modelling and evaluation step. Finally, during the deployment phase, data is arranged and shown in the form of graphs and reports. Because data mining is an iterative process, implementing one solution could prompt the use of new data sources and further data mining strategies [6].

3.2 Methods

The most often utilized applications of educational data mining consist of classification, clustering, prediction, association, which makes use of a variety of tools, algorithms, and methodologies. Among the techniques applied to data mining, decision trees, neural networks, regression analysis, & cluster analysis represent a few of the more widely utilized. Data inside dataset are categorized into distinct classes or categories by classification, a crucial step in data mining. It helps with data analysis and result forecasting by enabling the accurate prediction of target classes for each data case. Classification is commonly used in the educational industry to categories pupils according to a variety of criteria, age, gender, grades, educational achievements, knowledge, behavior, demography, or geographic characteristics [7].

A. Clustering Techniques

A technique for classifying a collection of abstract objects into groups of related components is to use clustering algorithms. These algorithms are useful for data analysis, pattern recognition, and image processing because they uncover patterns, similarities, or structures within data. Clustering makes it easier to find underlying structures or links in large datasets by gathering comparable data points together. Numerous industries use this method, including data analysis to identify client groups, machine learning to identify patterns, and image processing to classify pixels with comparable traits. Clustering is essential for gaining insightful understanding and streamlining the interpretation of intricate data patterns [7].

B. Decision Trees

Decision tree for understanding and quickly locating the most promising aspects, decision tree procedures are more reliable. To ascertain the relativity of two or more variables, it can also be used. The most popular algorithms used in decision trees are CHIAD, ID-3, CART, C4.5 and random tree.

C. Neural Networks

The capability of neural networks to find intricate patterns and connections in data based on given training data or prior knowledge is one of their main advantages. Neural networks are able to automatically learn and recognize complex connections between variables and predictors, in contrast to standard algorithms. They are excellent at detecting non-linear relationships and can modify their internal models to incorporate different aspects of the data. Because of this capabilities, neural networks are very effective at handling a variety of problems, including as speech and picture recognition, natural language processing, and predictive modelling. They are effective instruments for resolving difficult and complex problems in a variety of sectors because of their ability for detecting intricate patterns [7].

D. Bayesian Classifier

To calculate the parameters, it is an accurately straightforward process, minimal preparation of data is required. A Bayesian classifier makes obvious the class conditional connections between subsets of variables. It will be easier to learn if causal relationships are displayed graphically.

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Here's a breakdown of the terms in this equation:

- P(C|X) is the posterior probability of class C given the observed features X.
- P(X|C) is the likelihood, representing the probability of observing the features X given the class C.
- P(C) is the prior probability of class C, representing our initial belief about the distribution of classes.
- P(X) is the marginal likelihood or evidence, representing the overall probability of observing the features X across all possible classes.

E. Naive Bayes.

The conditional independence assumption is used in one simplification of the naive Bayes technique, supervised classifier, whereas the denominator is ignored in the other. The basis for it is the Bayes theorem while making strong naive assumption on the autonomy of the explanatory variables.

The Bayes rule of conditional probability is the foundation of the newly suggested Bayes categorization. The Bayes rule, often known as the Bayes theorem, is an approach for evaluating the probability of a property given a set of data as evidence or input.

$$P(h_i \mid x_i) = \frac{P(x_i \mid h_i)\, P(h_i)}{P(x_i \mid h_i) + P(x_i \mid h_2)\, P(h_2)}$$

F. K - nearest neighbor

K-nearest-neighbor classifiers are algorithms that depend on analogies gain knowledge by evaluating their performance with test training tuples that have comparable properties. The techniques can be applied to a tuple of unidentified numerical forecasts to generate a real-valued forecast. As a result, the method recovers the reasonable value related to the k-nearest neighbor (KNN) of the un-identified tuples. When used for predicting student's achievement, k-nearest neighbor worked well. Institutional investigators can better grasp how data mining can benefit them by looking at the tasks carried out and the technologies used. The various categories of data mining tasks include segmentation, estimation, classification, and description. The tasks and associated tools are listed in Table 1.

Mathematically, for a binary classification problem (two classes, 0 and 1), you can represent this as:

$$y\hat{}=\text{argmax}c\sum i=1kI(yi=c)$$

Where:

- $y\hat{}$ is the predicted class label for the new data point.
- k is the number of nearest neighbors to consider.
- y_i is the class label of the i -th nearest neighbor.
- I (y_i = c) is an indicator function that equals 1 if y_i is equal to class c and 0 otherwise.

Table 1: Classifications various data mining tasks and tools

Task	Unsupervised Data Mining	Supervised Data Mining
Classifications	Kohonen nets	**Memory based reasoning(MBR), genetic algorithm(GA), C&RT,link analysis, C5.0 and ANN**
Estimations	--	**C&RT and ANN**
Segmentations	Cluster detection, K-means,generalized rule induction and APRIORI	**Market basket analysis, memory based reasoning(MBR),link analysis and rule induction**
Descriptions	**Spatial visualization**	**Rule induction and market basket analysis**

Some different methods for predicting student's retention at various institutions or organizations throughout the world have been identified in the recent literature. These include SVM, NB, RF, LR, ANN, DT, EM, KNN, CLU, NN and C-45, along with CART.

3.3 Powerful EDM Tools

The following techniques have also been mentioned: - RF, NB, CLU, SVM (Support vector machine), LR, DT, EM, KNN (k-nearest neighbor), NN, ANN, C-45, and CART. Data mining has several uses, incorporating product marketing and promotion, services, products, study of artificial intelligence (AI), biological sciences, homicide investigation, high-level government intelligence (GI). Data mining technology have been developed over the due to its widespread use and the obstacles associated with establishing data mining applications. According to the most recent research for predicting student retention/dropout at various institutions/organizations universities throughout the world, each instrument has its own set of benefits and drawbacks [8].

A. Rapid Miner (YALE)

'Machine learning' (ML), 'Data mining' (DM), predictive analytics & business analytics are all compressed into one software platform named Rapid Miner. It supports every step of the data mining process and is used for commercial and industrial applications, research and education, training, fast prototyping or application development. A client/server programmed called Rapid Miner can be used using cloud infrastructures or as a SaaS.

Unlike most other data mining programmers, Rapid Miner's graphical programming language is more robust and has an extensive list of user-defined features. For instance, the Batch Cross Validation operator in Rapid Miner can perform cross-validation at several layers. This feature is an important leap above the graphical languages used by a majority of other software for data mining and is extremely useful for generalizability assessments. In order to aid in the evaluation of model fit, Rapid Miner additionally offers a set of metrics for model evaluations and visuals like Receiver-Operating Curves. Model can be exported as xml files, mathematical model which can be utilized in Rapid Miner code for applying the model to new data. A variety of task that the graphical programming languages could not perform can be handled using Rapid Miner Application Programmer Interface, which programmers using prefer Python or Java may utilise [9].

The subsequent list of Weka algorithms appears in Rapid Miner. Rapid Miner's latest versions include algorithm and parameter recommendations from the general public. You can use the multiple lessons included in Rapid Miner to help you learn the graphical programming language's use. For educational uses, Rapid Miner is free, whereas Rapid-I provides commercial licenses.

B. WEKA

Waikato Environment for Knowledge Analysis is referred to as Weka. It's collection of data mining (DM), machine learning(ML) techniques. Data mining (ML) experts utilize primarily for fact analysis and predictive modelling, this highly useful tool. It has a variety of advantages over quick miner and supports well-known (educational data mining) EDM tasks like: - selection, visualization, regression, and classification, clustering etc. These products can be performed directly upon data source and used in Java programmers. This Weka workbench is a collection of tool and techniques for visualizing, data analytics, predictive modelling, graphical user interfaces that make things simpler to use these features are offered as well. [10].

Various classification, clustering, and association mining methods are included with Weka and can be utilized individually, combination with strategies like boosting, stacking etc. GUI, Java API, or the command line may all be employed to access data mining algorithms. The GUI, which prevents users from making use of every one of extra features, is less powerful than the CLI and API's. Weka has the ability to generate Predictive Modelling Markup Language (PMML) file or mathematical models, which can be utilized with Weka scoring plug-in to run the model on fresh data [11].

C. Orange.

A Python-based open-source tool for EDM researchers is known as The Orange. The machine learning approach and the restricted bioinformatics and text mining functions offered by the orange tool are its primary benefits. It's a component-based software package for exploratory data analysis and visualization, and also for data mining and machine learning. It comes with Python binding's and modules. The programmer includes preparation of data, feature scoring & filtering, modelling, model evaluation, exploration approaches. [12]

D. KNIME

The Java-based 'KNIME' utility was built with Eclipse. This programmed did a good job of handling the key components of data beforehand strategy, transformation, extraction, and loading. Its tool will let creation of nodes for data processing through use of a GUI concept. BI (Business intelligence), financial analysis and reporting, an integration platform, and data analytics (DA) all were included in open-source plan. It also has numerous specialized algorithms, particularly many for

sentiment analysis and social network analysis.

E. Spark - MLlib

It's a distributed data processing architecture which enables for the processing of massive amounts of data over multiple processors. Java, Python, and SQL may all be used for distributed processing thanks to Spark's API, which connects to an extensive number of programming languages. The MLlib machine learning framework for Spark implements several of common ML and DL techniques. Although MLlib is still a rigorously coded tool with limited capabilities, it is a rapid and efficient alternative due to its distributed architecture.

3.3 Visualizations Tools

A. Tableau

'Tableau' is a set of interactive visualizations of data analysis tools. It's often utilized in education/educational circumstances to evaluate student's data, create useful insights, enhance teaching/tutoring methods, expedite education reporting, although primary focus on 'business intelligence' (BI). The fundamental benefit of Tableau is that it can evaluate enormous volumes of data from numerous sources with requiring for programming knowledge, opening providing a variety of visualizations to a wider audience. User may connect or import data from a variety of established format's (such as data warehouses, log data, etc.) using it. Users of Tableau may also create sophisticated, interactive dashboards that give users dynamic, real-time representations of data. On the other hand, Tableau has several restrictions, such as the inability to conduct relational data mining or analytics that are predictive. It's also not versatile, and since it is an industrial /commercial product, it doesn't support software platform interface.

B. D3js

The assistance of the JavaScript framework D3.js, researchers and professionals can make changes to create intricate, dynamic data visualization's and data-driven document's that are optimized for modern web browsers and require data processing. D3.js has many benefits namely the flexibility to reuse code, the ability to create an extensive range for visualizations of data without having to wait for installation's, the fact that it is open source (OS). On the other hand, greater implementation of educational research is fraught with challenges. The D3.js technology has a lots of compatibility issue's, limitation's on speed for huge collection of data, and requires substantial programming skills. Finally, pre-processing of data is necessary to protect 'privacy & security' because There is no way to keep data secret from users of visualization. [13].

3.4 Specialized Educational Data Mining and LA Applications

We covered some of the general-purpose EDM modelling and analysis tools in the article before. On the other hand, certain data and analysis objectives could call for the usage of additional specialized algorithm which are absent from this multipurpose equipment.

A. BKT ('Bayesian Knowledge Tracing') Tools.

A popular method to identify latent knowledge is called Bayesian Knowledge Tracing (BKT), which involves assessing the comprehension of students as they are taking an online course. The knowledge is evaluated while it has been learned online, compared to the testing-based educational measurement that's usually employed. This predicts, an intelligent tutoring systems or similar various applications, if a student has attained or not attained master in a certain talent. The two technique of brute force grid search (BFGS) and expectation maximization (EM) are frequently used to fit BKT models.

B. Text Mining(TM)

With several available programmers and API's for processing, tagging, recognizing textual data or text analysis is a quickly developing area of DM. Text analysis software can process word meaning, sentence structure, and word parts of speech.

4. Performance prediction from Classification Techniques

In order illustrate how the most commonly used educational performance prediction methods differ in their overall prediction precision from 2013 to 2023, the findings have been examined and displayed on a graph. In Figure 1, the schematic is shown.

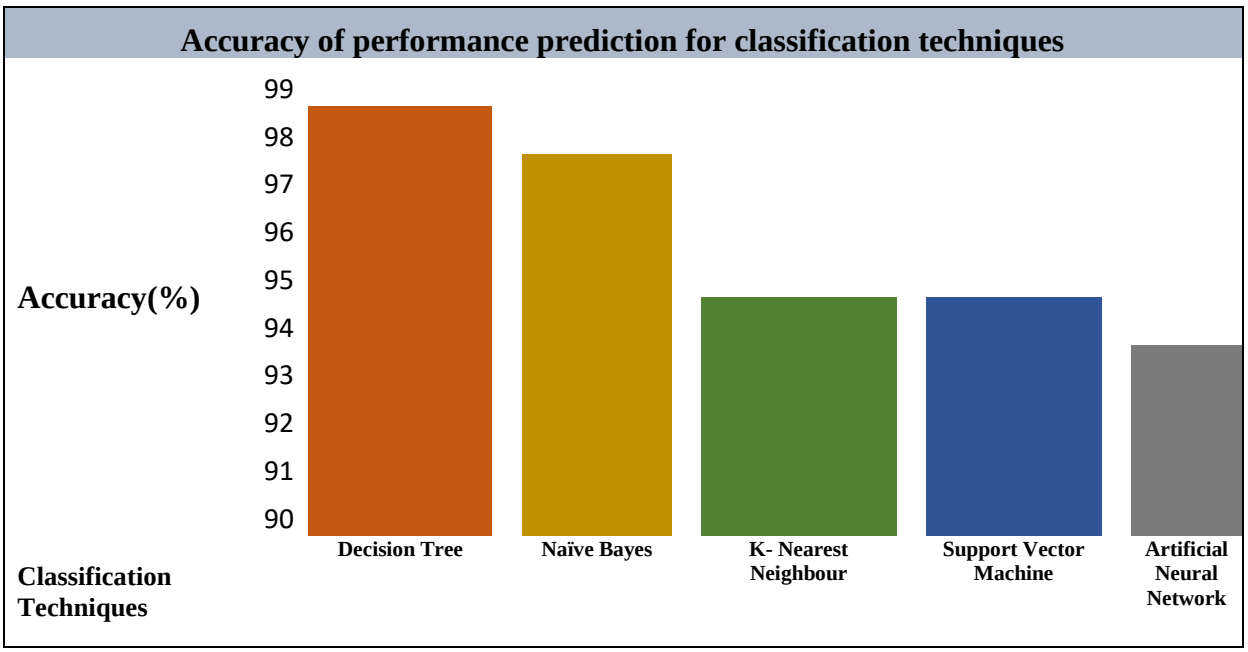

Figure 2: Accuracy of performance prediction (2013-2023)

The graph in Figure-2 shows the outcomes following the most frequently used strategies for forecasting student performance in earlier studies were examined and their results contrasted to see which approach had the highest accuracy. The general accuracy of the student method of estimation employed compared to conventional prediction methods.

The Figure illustrates the performance that the writers of this study looked into from 2013 to 2023, all organized by their algorithms. Figure-2 demonstrates the decision tree method's excellent prediction's and accuracy of 98.00 percent in compared to all other approaches. 'Naive Bayes' with a 97 percent accuracy rate, was the second-best approach. Then accompanied the SVM and KNN techniques, both of which had an accuracy rate of 94.00%. The accuracy of the 'artificial neural network' (ANN) in predicting student achievement is the lowest (93%) [14].

5. Parametric Attributes distribution in proposed architecture

This study yields the analysis as presented in Table 2 using the various student parameters. The parameters include the total of 22 attributes categorized in nominal and categorical data type with different values.

Table 2: Student's parametric attributes distribution

S.No.	Attributes	Data type	Values
1	Gender	Nominal	Male, Female
2	Percentage	Nominal	Poor, Good, Very good, Excellent
3	Stream	Nominal	Science, Commerce, Art, Design etc.
4	Qualification	Categorical	No formal education, HSC, HSC, Graduation, Masters, PhD
5	Flocculation	Categorical	Government worker, Private, Self-employed, NA
6	M_qulification	Categorical	No formal education, HSC, HSC, Graduation, Masters, PhD
7	M_occupation	Categorical	Government worker, Private, Self-employed, NA

8	No_of_sublings	Categorical	One, Two, Three, Four
9	Overall attendance	Nominal	Poor, Good, Very good, Excellent
10	Internal marks	Nominal	Poor, Good, Very good, Excellent
11	Assignment_ marks	Nominal	Poor, Good, Very good, Excellent
12	Practical knowledge	Nominal	Poor, Good, Very good, Excellent
13	Theory marks	Nominal	Poor, Good, Very good, Excellent
14	Internet_uses_learning	Nominal	Poor, Good, Very good, Excellent
15	Previous_sem_marks	Nominal	Poor, Good, Very good, Excellent
16	Subject Name	Nominal	Poor, Good, Very good, Excellent
17	Internal_Th_Marks	Nominal	Poor, Good, Very good, Excellent
18	Internal_Pr_Marks	Nominal	Poor, Good, Very good, Excellent
19	External_Th_Marks	Nominal	Poor, Good, Very good, Excellent
20	External_Pr_Marks	Nominal	Poor, Good, Very good, Excellent
21	Subject Result	Nominal	Poor, Good, Very good, Excellent
22	Semester_wise_result	Nominal	Poor, Good, Very good, Excellent

6. Experimental analysis using Weka

The semester-by-semester comparative result is described as per the following table. In this WEKA experimental analysis, we used a number of classification algorithms, including J48, Bayes Net, Decision Stump, Logistic Regression, Multi-layer Perception, the Naive Bayes algorithm, One R, Rep Tree, and Sequential Minimal Optimization.

Table 3: Time spent by each classifier in developing the model upon a semester-by-semester basis.

Semesters	**J-48**	**BN**	**DS**	**LS**	**MLP**	**NB**	**1R**	**RT**	**SMO**
I	0.02	0.8	0.9	2.41	56.43	0.09	0.18	0.11	2.25
II	0.0502	0.120	0.920	3.4302	66.4502	0.1002	0.1402	0.1202	1.2702
III	0.0625	0.1325	0.9325	3.4425	66.462	0.1125	0.152	0.1325	1.2825
IV	0.059	0.1299	0.9299	3.439	66.4599	0.109	0.1499	0.1299	1.2799
V	0.0573	0.1273	0.9273	3.437	66.457	0.1073	0.147	0.1273	1.2773
VI	0.03	0.1	0.9	3.41	66.43	0.08	0.12	0.1	1.25
Mean Value	0.046	0.234	0.9183	3.2616	64.781	0.0999	0.14832	0.1199	1.4349

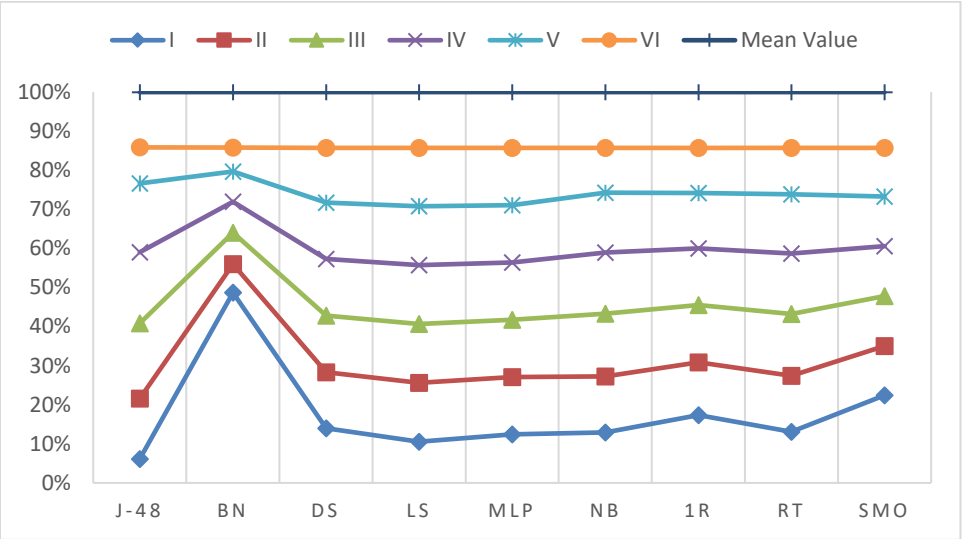

Figure 3: Time spent by each classifier in developing the model upon a semester-by-semester basis

Table 4: Through the use of several classifiers, semester-wise effectively categorized instances.

Semesters	J-48	BN	DS	LS	MLP	NB	1R	RT	SMO
I	99.434	97.422	60.876	97.528	87.157	97.482	78.29	97.391	87.1
II	99.12	96.4265	69.8764	97.3141	92.2976	97.3984	82.29	95.348	92.325
III	99.129	97.4365	71.886	97.8241	93.8976	98.8984	84.29	96.448	93.525
IV	99.269	97.536	71.986	98.0241	93.907	98.9184	84.59	97.438	94.025
V	99.325	97.6565	72.026	98.4241	94.7076	98.9284	84.75	97.458	94.125
VI	99.07	98.4827	59.8764	98.5389	89.157	98.482	79.29	98.391	89.1
Mean Value	99.224	97.4935	67.7547	97.942	91.854	98.351	82.25	97.079	91.700

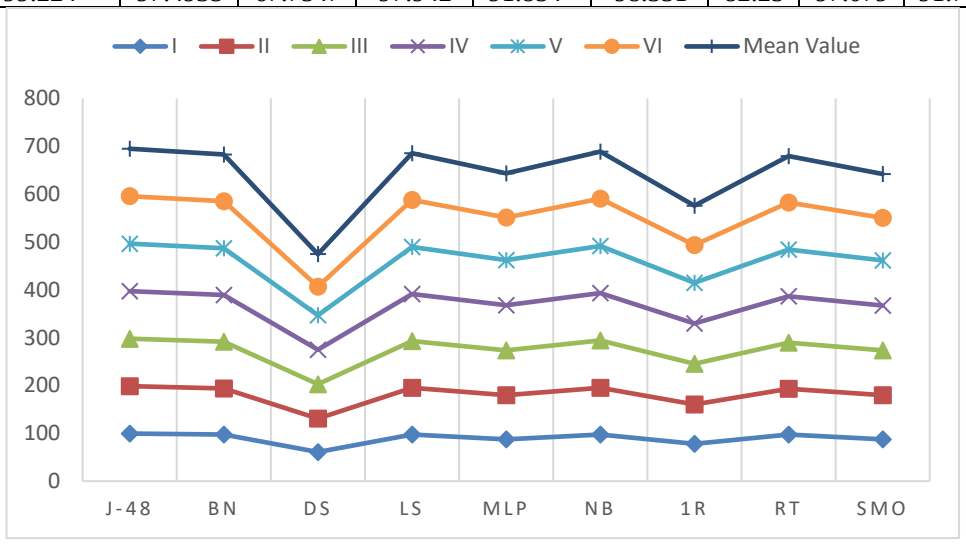

Figure 4: Through the use of several classifiers, semester-wise effectively categorized instances

Table 5: Through the use of several classifiers, semester-wise accurately categorized examples.

Semesters	J-48	BN	DS	LS	MLP	NB	1R	RT	SMO
I	0.3658	2.5173	39.1236	2.4611	12.8429	2.517	12.708	2.2082	12.899
II	0.88	3.5735	30.1236	2.6859	7.7024	2.601	17.7081	4.652	7.6743
III	0.871	2.5635	28.1136	2.1759	6.1024	1.101	15.7081	3.552	6.4743
IV	0.7305	2.4635	28.0136	1.9759	6.0924	1.081	15.4081	2.562	5.9743
V	0.6743	2.3435	27.9736	1.5759	5.2924	1.071	15.2481	2.542	5.8743
VI	0.92	1.5173	40.1236	1.4611	0.8429	1.517	20.708	1.2082	0.899
Mean Value	0.7402	2.4964	32.245	2.0559	6.47923	1.648	16.2481	2.7874	6.6325

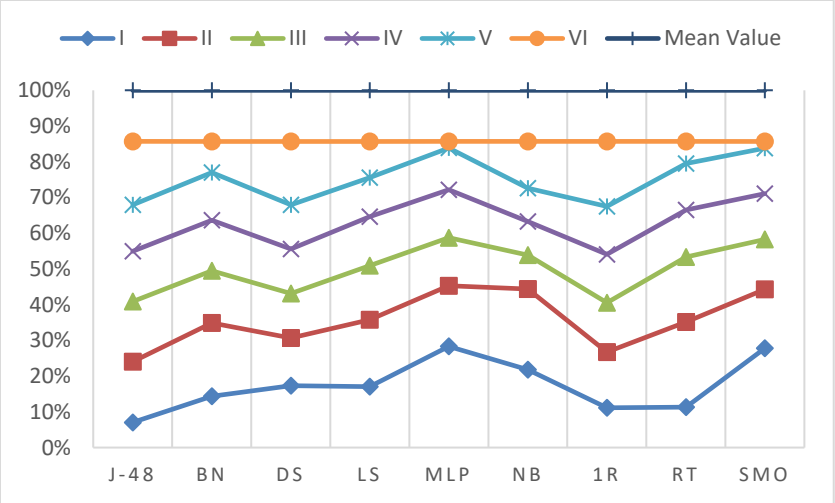

Figure 5: Through the use of several classifiers, semester-wise accurately categorized examples

Table 6: Kappa statistics are evaluated semester by semester using multiple classifiers.

Semesters	J-48	BN	DS	LS	MLP	NB	1R	RT	SMO
I	0.9752	0.9695	0.296	0.9680	0.9686	0.9695	0.6175	0.9737	0.97
II	0.9489	0.9219	0.65646	0.9308	0.8806	0.93168	0.7806	0.9111	0.88
III	0.9489	0.9320	0.67656	0.9359	0.8966	0.9466	0.8006	0.9221	0.89
IV	0.9503	0.9330	0.67756	0.9379	0.8967	0.94688	0.8036	0.9320	0.89
V	0.9509	0.9342	0.67796	0.9419	0.904	0.94698	0.805	0.9322	0.89
VI	0.98	0.9795	0.396	0.9702	0.9786	0.9795	0.7175	0.9737	0.977
Mean Value	0.9590	0.9450	0.5634	0.9474	0.9210	0.9535	0.7541	0.94085	0.92

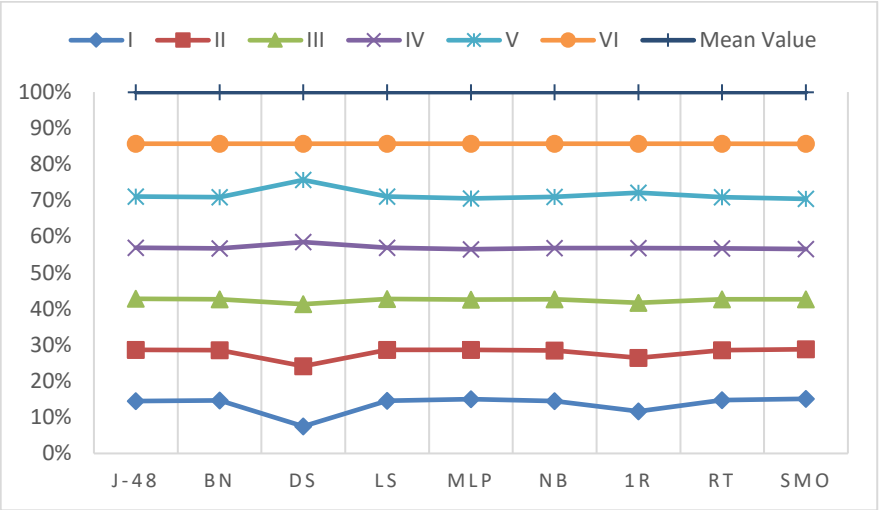

Figure 6: Kappa statistics are evaluated semester by semester using multiple classifiers

Table 7: Mean absolute Error (MAE) using different classifiers, semester-by-semester.

Semesters	J-48	BN	DS	LS	MLP	NB	1R	RT	SMO
I	0.0173	0.021	0.299	0.0256	0.1146	0.021	0.092	0.0256	0.14
II	0.0093	0.013	0.201	0.0076	0.0066	0.013	0.084	0.0076	0.2424
III	0.0103	0.014	0.202	0.0086	0.0076	0.014	0.085	0.0086	0.2434
IV	0.0113	0.015	0.203	0.0096	0.0086	0.015	0.086	0.0096	0.2444
V	0.0123	0.016	0.204	0.0106	0.0096	0.016	0.087	0.0106	0.2454
VI	0.005	0.011	0.199	0.0066	0.0069	0.011	0.082	0.0056	0.24
Mean Value	0.0109	0.0155	0.2189	0.01143	0.0256	0.0158	0.086	0.01126	0.2259

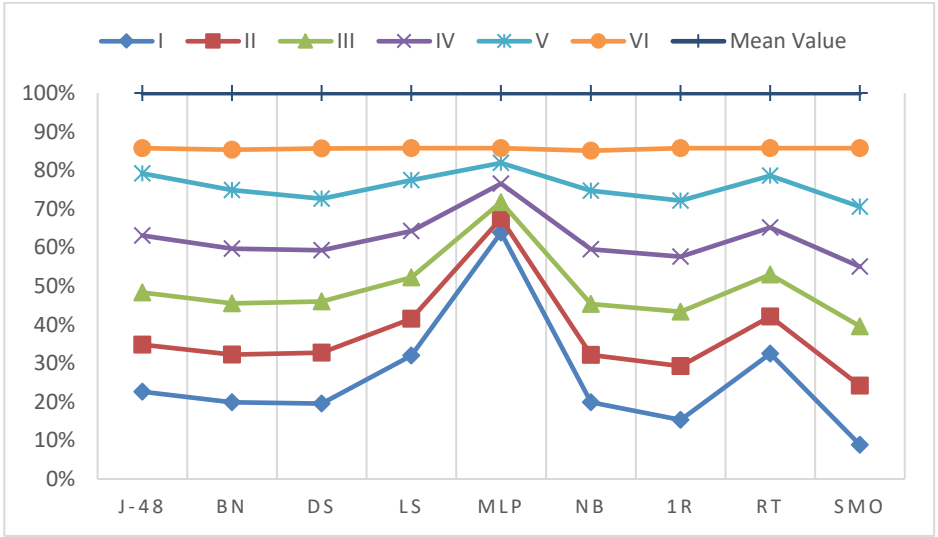

Figure 7: Mean absolute Error (MAE) using different classifiers, semester-by-semester

Table 8: Root Mean Squared Error Rate (RMSE) by multiple classifiers during the semester.

Semesters	J-48	BN	DS	LS	MLP	NB	1R	RT	SMO
I	0.0541	0.0667	0.2162	0.0634	0.0607	0.0672	0.1878	0.168	0.216
II	0.084	0.0967	0.3362	0.0934	0.0707	0.0972	0.3078	0.088	0.3364
III	0.0941	0.1067	0.346	0.1034	0.0807	0.1072	0.3178	0.098	0.3464
IV	0.104	0.1167	0.3562	0.1134	0.0907	0.117	0.3278	0.106	0.3564
V	0.1141	0.1267	0.3662	0.1234	0.1007	0.1272	0.3378	0.118	0.3664
VI	0.05	0.0767	0.3162	0.0734	0.0607	0.0772	0.2878	0.068	0.316
Mean Value	0.0834	0.0983	0.3228	0.0950	0.0773	0.0988	0.2944	0.108	0.3229

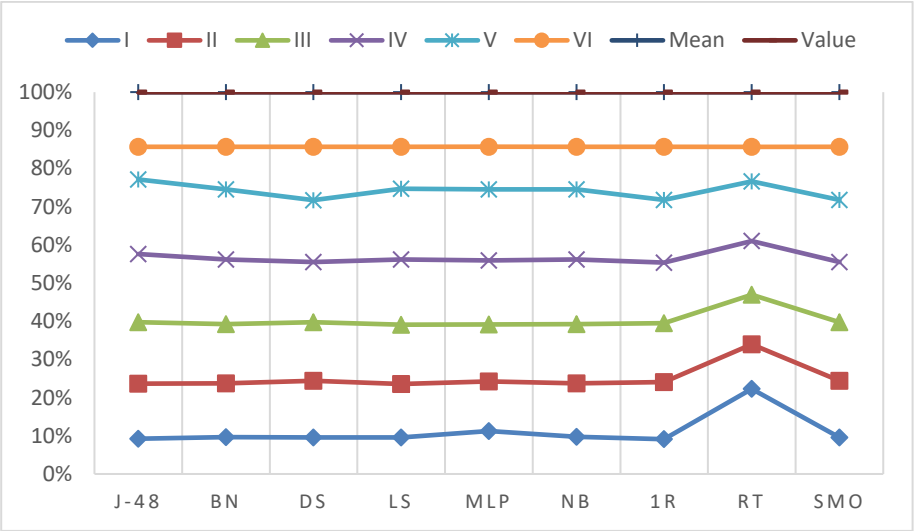

Figure 8: Root Mean Squared Error Rate (RMSE) by multiple classifiers during the semester

Used Abbreviation in above tables are:

BN - Bayes Net, DS - Decision stump, LS - Logistic Regression, MLP - Multi layer perception, NB - Naïve Bayes, IR - One R, RT - Rep Tree, SMO - sequential minimal optimization.

7. Result Analysis

This work integrates the EDM student parametric data and experimented with the tool weka that uses classification algorithms, including J48, Bayes Net, Decision Stump, Logistic Regression, Multi-Layer Perception, Naive Bayesian, One R, Rep Tree, and Sequential Minimal Optimization, were utilized and applied to the WEKA tool in this experiment analysis to obtain the semester-by-semester performance of the defined algorithm in regard to the used accuracy measured and error measured parameters. We used time to create the model, correctly identified instances, and incorrectly categorized instances as accuracy assessment metrics in the present investigation. In our research, we used kappa research, mean absolute error (MAE), and root mean square error (RMSE) as error measuring metrics. And this study concludes that J48 algorithm provides the most accurate result out of all the classification algorithms.

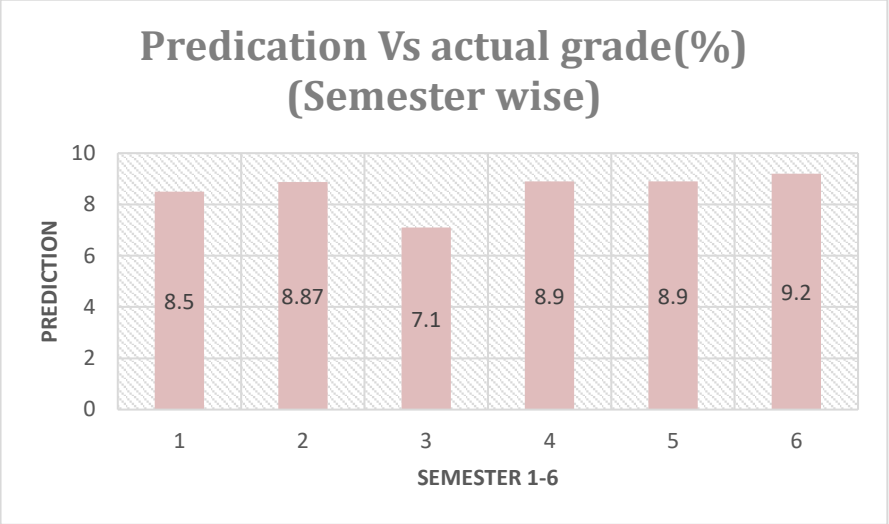

Figure 8: Percentage of predictions within a point of the actual grade.

8. Conclusion

All those involved in the process of learning can benefit greatly from the relatively recent discipline of 'educational data mining' (EDM). The methods of (DM) data mining have been developed automatically extracts hidden knowledge & pattern's in data. To be able to categories and forecast student performance, dropout rates, and instructor or educator performance, educational data mining can be employed. It may help both students in choosing courses and managing their education, as well as teachers in tracking student achievement to enhance the teaching process. EDM can be used to draw in, keep, and retain students, which is essential for a university to be profitable. Figuring out, recognizing, and understanding whether educational practices are effective requires analyzing student data. In this research study, we examined the advantages and uses of DM methods in various educational contexts. The main or primary objective of this research is to encourage others to use educational data mining techniques by demonstrating the advantages they offer. A significant problem in all educational institutions maintains the standard of guidance while evaluating students' performance. Data mining techniques are frequently used for analyzing the data that is already accessible and to extract knowledge and information to aid in decision-making. In this research study, various data mining classification techniques are employed to create a data mining model to predict student success based on their unique demographic and academic data. WEKA tool is used to conduct this analysis. The accuracy of the classifiers and the classifiers' error rate are the results. These findings are compared in order to figure out which algorithm is most appropriate for this kind of dataset. As an outcome of observation, it was determined that the J48 approach for both models provide better precision and less error. Only a few types of strategies were selected for this analysis; subsequent work will explore additional algorithms and the further research will explores the new dimensions of EDM via Augmented and Virtual Reality in Online learning environment.

References

[1] Dr. P. Nithya, B. Umamaheswari, A. Umadevi – "A Survey on Educational Data Mining in Field of Education" – International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016.

[2] Aykroyd, R.G.; Leiva, V.; Ruggeri, F. Recent developments of control charts, identification of big data sources and future trends of current research. Technol. Forecast. Soc. Chang., 144, 221–232, 2019.

[3] Hooshyar, D.; Pedaste, M.; Yang, Y. Mining educational data to predict students' performance through procrastination behavior, Entropy, Volume 22, Issue 12, 2020.

[4] Reem Atassi,Aditi Sharma. "Intelligent Traffic Management using IoT and Machine Learning." Journal of Intelligent Systems and Internet of Things, Vol. 8, No. 2, 2023 ,PP. 08-19.

[5] V. Gupta, N. Kumar, A. Sharma and A. Abraham, "Sensor Routing Protocol with Optimized Delay and Overheads in Mobile based WSN", Journal of Information Assurance & Security, vol. 16, no. 4, 2021.

[6] Bakhshinategh, B.; Zaiane, O.R.; Elatia, S.; Ipperciel, D. Educational data mining applications and tasks: A survey of the last 10 years. Educ. Inf. Technol., pp. 537–553, Volume 23, 2018.

[7] Del Bonifro, F.; Gabbrielli, M.; Lisanti, G.; Zingaro, S.P. Student Dropout Prediction. In Artificial Intelligence in Education; Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., Millán E., Eds.; Springer: Cham, Switzerland, 2020.
[8] V. Goar, A. Sharma, N. S. Yadav, S. Chowdhury and Y.-C. Hu, "IOT-based smart mask protection against the waves of covid-19", Journal of Ambient Intelligence and Humanized Computing, 2022.
[9] Lázaro, N.; Callejas, Z.; Griol, D. Predicting computer engineering student's dropout in cuban higher education with pre-enrollment and early performance data. J. Technol. Sci. Educ. 2020, 10, 241–258.
[10] J. R. Albert and A. Sharma, "Investigation on load harmonic reduction through solar-power utilization in intermittent SSFI using particle swarm genetic and modified firefly optimization algorithms", Journal of Intelligent & Fuzzy Systems, vol. 42, no. 4, pp. 4117-4133, 2022.
[11] Mduma, N.; Kalegele, K.; Machuve, D. Machine learning approach for reducing student's dropout rates. Int. J. Adv. Comput. Res., Volume 9, 156–169, 2019.
[12] P.Sinha, M. Arora, N. Mishra, "Framework for a Knowledge Management Platform in Higher Education Institutions", Volume 2, Issue 4, September 2012
[13] Goyal, R. Vohra, "Applications of Data Mining in Higher Education", International Journal of Computer Science Issues, Vol. 9, Issue 2, No1, March 2012.
[14] Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. (2014). Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses. Journal of Learning Analytics, 2(1), 156–184.
[15] A. U. Khasanah et al., "A comparative study to predict student's performance using educational data mining techniques," in IOP Conference Series: Materials Science and Engineering, vol. 215, p. 012036, IOP Publishing, 2017.
[16] M. Makhtar, H. Nawang, and S. N. Wan Shamsuddin, "Analysis on students' performance using naive bayes classifier." Journal of Theoretical & Applied Information Technology, vol. 95, no. 16, 2017.
[17] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of students' academic performance using weka," Indonesian Journal of Electrical Engineering and Computer Science, vol. 9, no. 2, pp. 447–459, 2018.
[18] N D Lynn and A W R Emanuel, Using Data Mining Techniques to Predict Students' Performance. A Review, IOP Conf. Series: Materials Science and Engineering 1096 (2021). doi:10.1088/1757-899X/1096/1/012083.
[19] Snježana Križanić, Educational data mining using cluster analysis and decision tree technique: A case study, International Journal of Engineering Business Management (2020). https://doi.org/10.1177/1847979020908675.
[20] Rahul Sharma and Dr. Shiv Shakti Shrivastava Predicting of Student Performance using Data Mining Classification Techniques, World Journal of Engineering Research and Technology (WJERT), Vol. 9, Issue 3, 2023.
[21] Yang, F., & Li, F. W. B. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining. Computers and Education, 123(April): 97–108. https://doi.org/10.1016/j.compedu.2018.04.006.
[22] B. Patel, C. Gondaliya, "Student Performance Analysis Using Data Mining Technique", International Journal of Computer Science and Mobile Computing IJCSMC, Vol.6 Issue.5, ISSN 2320–088X IMPACT FACTOR: 6.017, May-2017, pg. 64-71.
[23] Vilanova, R., Dominguez, M., Vicario, J., Prada, M. A., Barbu, M., Varanda, M. J., Alves, P., Podpora, M., Spagnolini, U., & Paganoni, A. (2019). Data-driven tool for monitoring of student's performance. IFAC-PapersOnLine, 52(9): 190–195. https://doi.org/10.1016/j.ifacol.2019.08.188
[24] M. C. R, S. Sharma, A. Sharma, M. Sunil Kumar, S. Kelkar and S. Vishal Deshmukh, "Cloud Top Management Role in Reducing Mobile Broadband Transmission Hazards and Offering Safety," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 1064-1068, doi: 10.1109/ICACITE57410.2023.10182893.
[25] Babandi Usman, Rabi'u Adamu and Sani Salisu, Prediction of Student Performance Using Classification Technique, International Journal of Information Processing and Communication (IJIPC) 8(1): (May, 2020).
[26] Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D. J., & Long, Q. (2018). Predicting academic performance by considering student heterogeneity. KnowledgeBased Systems, 161: 134–146. https://doi.org/10.1016/j.knosys.2018.0