Study of Educational Data Mining Approaches for Student Performance Analysis

R. Asad[1], S. Arooj[2], S. U. Rehman[3]

[1,2,3]*University Institute of Information Technology, PMAS - Arid Agriculture University, Rawalpindi, Pakistan*

[3]Saif@uaar.edu.pk

Abstract- Education is a vital component in the development of any country. In the education sector, research has been rapidly increasing using data mining techniques. The increase of e-learning resources, instrumental educational software, the use of the Internet in education, and the establishment of state databases of student information have created large repositories of educational data. Accurate prediction of students' progress and their potential at the beginning of the degree is crucial for recognizing weak students and preventing their dropout at early stages. Educational Data Mining (EDM) is the application of data mining techniques to this specific type of dataset that comes from educational environments to address important educational questions. EDM assists in selecting improved learning materials and learning activities, with the main focus line to discover the hidden facts and figures concerning the performance of the students. This research study aims to reinforce the students' academic performance prediction model, for higher studies using the Naïve Bayes classification method which has proved as the top classifier in making estimations accurately as compared to other classifiers of data mining. Different parameters like internal marks and sessional marks have been chosen to conduct this task. Internal marks are comprised of class performance, assignment marks, attendance marks, and presentation marks. Sessional marks are the results of exams conducted by the class instructors. In this way, early prediction can resolve the problem by indicating the factors that will cause their failure in academia.

Keywords- Educational Data Mining, Classification, performance prediction, student's dropout, learning analytics.

I. INTRODUCTION:

Education is a basic human right and people all over the world should get equal quality education [1]. It serves as the bedrock for the national integration of any country so it is a critical concern to consider education regarding the growth of country[2].We usually listen to "Education is Necessary" to live a better life. It makes the man a well-mannered human being by teaching many things like values, virtues, and ethics. Education makes the character of man bright moreover; it helps in building their moral character stronger. It gives a sense of self-confidence to the one, which aids them to pass over any hurdle that comes their way. Education helps in developing different skills in human making the society to generate the great personalities and good people. Value-based education develops the social skills, time management skills, decision making, and problem-solving skills in our youth, which helps to find the best direction for themselves[3].

Education is categorized into different levels alike: elementary, primary, secondary, and higher education. These educational levels and substages are the formal way of providing knowledge and developing skills in youth. For the past few years, there has been a lot of emphasis on education, especially in the higher education sector. Educational institutes are trying their best to improve their education quality so as to meet the requirement of time [4]. In the past years, the involvement of data mining in the educational sector has been seen to make improvements.

Data Mining (DM, hereafter) is the area of uncovering new facts and finding hidden useful information from the huge data by mining patterns[5]. The practical use of DM can be seen in a huge great number of fields like bioinformatics, retail sales, counter-terrorism, etc. In past years, the utilization of data mining to explore scientific questions within the research field of education has been raised. Life exists in an era where a huge aggregate of data is gathered daily. And there exists a great need to examine such data and DM can fulfill this requirement by supplying tools to investigate hidden knowledge and patterns from data. DM helps in many ways for examination of data like it helps in it aid companies gather well-founded details, it's an efficient, cost-effective solution compared to other data applications, aids in making beneficial production and operational adjustments, helps out in making informed decisions, make the data scientists enable to easily examine the bulk quantity of data quickly[5]. Figure 1 demonstrates the method for getting potentially useful and unknown patterns from datasets. Data scientists then can use this information to make automated predictions of behaviors and trends and investigate hidden facts. DM involves: tracking patterns, classification, association, outlier detection, clustering, regression, and predictions. Recently there has been a great exploration of educational data mining.

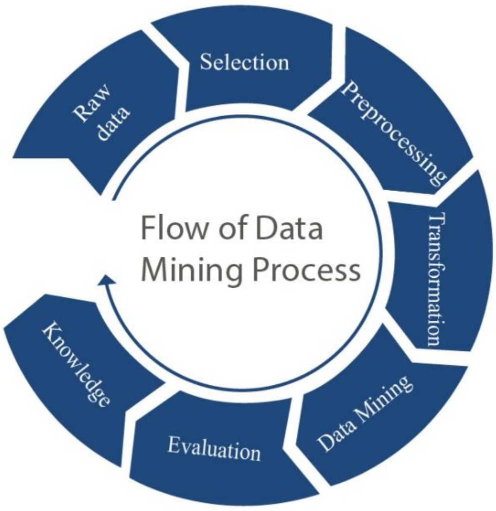
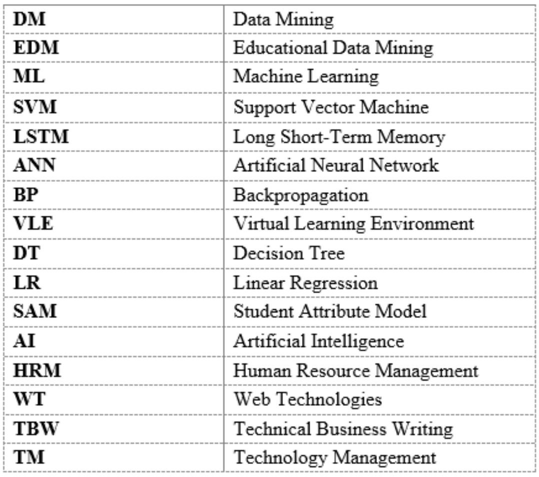


Figure 1: Flow of Data Mining Process [29]

Educational Data Mining (EDM, hereafter) is outlined as the area of the scientific investigation centered throughout the evolution of methods for making inventions within the special kinds of data that come from educational settings, and then making the use of these methods and techniques to find out how the students learning and the environment in which they live in [6]. EDM has attached some extraordinary benefits not only for the educational sectors but also for students. The benefits that are added up in educational sectors by EDM include: planning in a perfect manner, evaluation, estimation of different programs and strategies that are opting for learning in education, the content of the course, and educational result. It also helps to predict the productivity of students and then improve them by bringing variation in teaching strategies by analyzing the style that how a certain student learns. As well as obtaining awareness about different educational circumstances and resolution to the point complicate dissues. For students, the following benefits are provided like: examining the background of their performance in academia to make the predictions of their upcoming behavior at the earliest stage. Proposing the finest learning path for each student based on how any certain student performs in his academia, making the best use to find their attention and afterward recommending students those courses in which they are interested so that they can show the best progress.

Table 1: Acronyms used in paper



DM	Data Mining
EDM	Educational Data Mining
ML	Machine Learning
SVM	Support Vector Machine
LSTM	Long Short-Term Memory
ANN	Artificial Neural Network
BP	Backpropagation
VLE	Virtual Learning Environment
DT	Decision Tree
LR	Linear Regression
SAM	Student Attribute Model
AI	Artificial Intelligence
HRM	Human Resource Management
WT	Web Technologies
TBW	Technical Business Writing
TM	Technology Management

Figure. 2 depicts the flow that how data mining is applied in educational systems. It shows how in the educational sector educators are held responsible for upholding and scheming the plan of study. Afterward students need to interact with that plan decided by their teachers or mentors. This data of students is then collected, upon which certain classifier algorithms of data mining are applied for extracting potentially useful and unknown facts and findings and then giving recommendations based on of resulting outcome. EDM has used different techniques and datasets to make predictions some of its major applications are: making predictions for measuring the performance of students, discovering unpleasant behavior of students, a grouping of students, and student modeling[6].

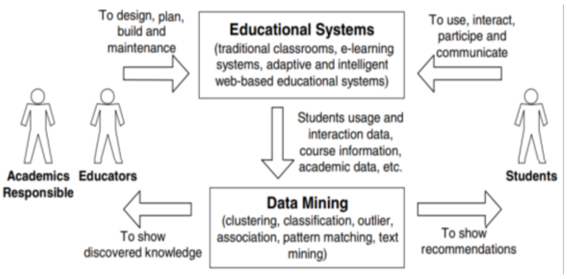

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics Responsible
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure 2: Cycle for applying data mining in the educational sector [46]

A student can go for choosing any of course and field in which he seems to be interested. Predicting a learner's performance is one of the important matters in education. After enrolling in the university, it is useful to see if a student will be able to perform well in the upcoming exams or not. This helps to make educational decisions more effective and efficient. Teachers can develop different learning strategies to improve student performance [7]. The use of learning analytics has been increased in the educational sector in the preceding decade[8]. There has been a rise in a huge volume of data [50].

Educational institutions are continuously gathering the data of students. Also, as it is the digital era means of e-learning have helped in producing data in big quantities. But along with that there paved many problems like there exist the gap between the mentors and learners due to which it's been difficult in collecting direct feedback from students, lack of understanding and interest of learners. Due to increased interest in higher education government has made efforts in sorting out this issue to get the outstanding outcome of student's performance in academia. This study aims to use educational DM to predict learners' performance before final exams in order to reduce students' dropout using educational DM[9].This may help the institution to improve learning services as well as lecturers to launch a learning plan that will help to reduce the student dropout rates in different courses.

EDM has been successfully employed for the measurement of students' performance. Figure.3 demonstrates the intersection of three main areas in the field of EDM: education, statistics, and informatics. This intersection among these three areas also generates other sub-fields, narrowly related to EDM, such as computer-based education, learning analysis [52], data mining (DM), and machine learning (ML) [10].

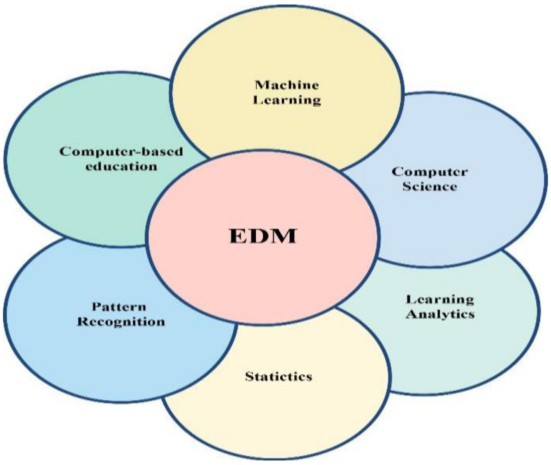


Figure 3: Educational Data Mining [10]

In this work, it has been presented a detailed comparative analysis of the different EDM approaches available in the literature. EDM is not a recently raised field, in past, a lot of the work has already been done using DM techniques.DM techniques in educational learning are used to get the accurate beneficial outcome for preventing student dropout by doing a deep analysis of their behavior. Attributes like internal marks, sessional marks, etc. in datasets are availed to perform correct predictions. Early detection of the reasons that influence their learning also helps in refining the grades of students. This work contributes to the educational sector by focusing on the evaluation and classification of different types of educational data using DM techniques in order to evaluate the efficiency of raw data and then getting useful patterns and knowledge out of it by mining this data.

The rest of the paper is organized as follows. Section 2 contains the literature review of educational data mining and a comprehensive analysis of other related work. Section 3 includes the proposed methodology, learner's data mining process as well as proposed framework. Section 4 will present experiments and results which will contain the experimental analysis and discussion followed by conclusion and future work in section 5.

II. Literature Review

Prediction of student performance with the help of machine learning and educational data mining is an ongoing research area. A lot of work has already been done and researchers are continuously doing their research to improve the progress. To prevent students from failing academic pathways a framework based on machine learning was suggested in [10]. The dataset used in their study includes 478 Physics students during the school years: 2015–2016, 2016–2017, and 2017–2018. Of the 196 students were male and 282 were female. The total number of students who were repeating concerns 52.Out of them, 25 were female and 27 were male. All students' age ranged from 16 to 19.Data is collected and prepared by the SMS-Massar. Machine learning Algorithms, multiple regression were used to generate the framework model. The model was composed of two sub-models: a model of the second semester (Model S2) and a model National Examination (Model NE). The model predicted that 10% of students will fail in the second semester and 52% will fail in the national exam. Results depicted that at the end of the 2nd semester 17% of students failed the exams while 50% of students were failed the national exam with a little difference of 2% [10].

Boran suggested that education is the necessary component for encouraging self-confidence in a person as well as vital for national growth in today's era. During the past few years, there are several problems come in the way of education. In order to tackle these problems, higher education institutions have made revolutionary changes in the traditional teaching methods by the use of technology. Their research work has contributed to the prediction performance of a student through the use of five machine learning algorithms while considering two datasets. They focused on the evaluation and classification of different types of educational data to analyze how efficient the machine learning algorithms work for raw data without using any preparation algorithm or data selection[11]. They have used BP, SVR, and LTSM algorithms for the research. They have considered two datasets that involve: Student Performance Dataset (SPD) and Students Academic Performance Dataset (SAPD) for their research.

The work done presented in [12] is mainly focused on e-learning. They defined E-learning as the transfer of skills as well as knowledge through electronic and web devices. For the past 3 years, e-learning is well developed in the educational field. Their research study was based on predicting the student's potential along with their performance and progress. Student Attribute Model (SAM) was proposed to measure the quantity of performance and non-performance-based learning attributes to analyze student performance and development. The model was trained on a back propagation neural network to predict the attribute causal relationship. Real academic data of 60 arts and science students were used by dividing them into subject groups and learning stage groups based on SAM.

In 2018, Raza et al., proposed a Virtual Learning Environment (VLE) by creating enormous volumes of data. Due to this huge amount of data, the use of data mining approaches has made organizations in the educational sector for exploring previously undiscovered knowledge and patterns from these data that will improve the learning methods and the performance of the students. They focused on exploring the performance of students in academia by using the decision tree algorithm of data mining by making use of features like information of students in academics and the activity of students[13]. They computed the results using WEKA, a well-known tool of data mining for evaluating decision tree algorithms and exploring the performance of students with Moodle access time for improving the grades of students. Data they have used for their proposed model was comprised of activity of student and performance data in the module of level 3 for computation of degree course at Middle East College. They employed data by making use of 22 students from Spring 2017, which get registered from undergraduate level, and have found that Random Forest, Naive Bayes, and SMO for training sets are the good agreement.

Romero et al., used prediction on a web-based course and predicted the likelihood of passing a course based on the student's work carried out in the system[14]. The study suggested a genetic algorithm. DRAL tool was used for the detection of e-learning activities based on the feature extraction of the attributes that a student needs in order to succeed in his course work. The proposed algorithm was compared with the traditional proposals and showed an improved result and accuracy. The datasets consist of academic information as well as the demographic information of the undergraduate engineering students. Three student performance datasets were used for the study. The dataset consists of complete (1000), complete (525), and outliers (960) instances respectively.

In another working student historical data from different semesters was collected and analyzed [15]. Student data from previous semesters were gathered and mapped into corresponding labels. Their work contributed to analyzing the performance value, constructing different classifiers, and their study as well as proposing a meta-based tree model for the prediction of student performance. A total of 400 records of students were fed to the model with 13 attributes or (features). Ensemble Meta-based tree model was trained on a dataset and its performance was increased by including preprocessing phase, construction phase, and evaluation phase. The proposed model combined two machine learning techniques to get more precise results compared to the traditional methods that were based on one algorithm learning and yielded an accuracy of 98.5%.

The downsides of existing EDM methods were acknowledged using an unsuitable training strategy and a parameter set to suggest a Multi Adaptive Neuro-Fuzzy Inference System with Representative Sets that is also called MAINFIS-S and put it to multi-input and the multi-output problem of student academic performance prediction[16]. For handling major two limitations of the currently used neuro-fuzzy MANFIS-S model concerning training strategy and parameter set. Confronting with the first problem, it finds out the betterment in accuracy and reasonable computational complexity. The time and accuracy of MANFIS-S were carried out per under with the numerous values of parameters. This work manifested the dominance in terms of the accuracy of MANFIS-S over the related algorithms. The proposed model was researched by taking datasets from KDD Cup [17]. Over three periods those datasets recounted the accurate answers of almost twenty thousand students. The suggested model was applied on MATLAB implemented in Matlab against MANFIS [18], OneR, ANFIS [19], and Random Tree [20]. Dataset 1 was taken from UCI [21] containing 649 objects with 33 features out of the last three variables were output. 2nd dataset was taken from VNU University of Science; Vietnam comprised of 260 objects was taken as student records. Attributes taken were total of 8 in number for input and the 9th one was taken as output. Each dataset has 3 input values (Hints, Step duration, and Incorrect), and 2 outputs. And these datasets were then set to normalize with the aid of exponential function.

In a recent study, 2022, Raza worked on video learning analytics with flipped teaching[22]. The transfer of learning techniques from traditional methods to digitalization had led to a model in which learning can take place outside the classroom by utilizing different attributes. They applied eight different kinds of the classification algorithms. The dataset contains 772 students of the sixth semester registered on e-commerce technologies modules at an HEI. The academic dataset was collected from Student Information System (SIS) and online activity from LMS. The results showed that the Random Forest algorithm predicted accuracy of up to 88.3% with an equal width and information gain ratio.

Lau et al., worked on students' CGPA prediction by using both conventional statistical analysis and Artificial Neural Network (ANN) approach[23]. Levenberg–Marquardt algorithm was deployed as a back propagation rule. Students' information about their social-economic background and entrance test results was used as data. It was collected from 1000 different students consisting of 810 boys and 175 girls. The model achieved a good accuracy of 84.8%. However, ANN did not perform well in classification with gender and generated high false-negative results.

In another recent work by Hajra et al., studied the VLE big data and students' interaction with the learning environment[24]. They used the VLE dataset to determine the effectiveness of deep learning models in predicting students' performance at early stages. A dataset of 32,593 students was collected from OULA, an openly accessible platform over 9 months. This study followed two-step approaches. The first approach contains mining the student demographic and VLE portal data. The second approach contains early intervention of categories. A deep ANN classification model was deployed to predict the students who were going to dropout of the courses. This yielded an accuracy of 84%. This model is not worked well for predicting students with distinction.

According to [25], predicting the failure of students is a critical issue to be addressed. Adopting a positive attitude towards this problem is beneficial bringing many positive changes in the educational sector. This study was performed to find the relation between the influences of ethnicity on the performance of students for predicting early dropout. This study was employed using Orange Software, making use of four algorithms (i.e, Neural Network [26], Classification Tree, Naïve Bayes, and Random Forest Algorithm). The dataset used for this study was taken from Badejo et al. [27] which contained 2413 records of students created by the department of Covenant University. The accuracy observed after experimenting was 53.2%. and was found that the factor of ethnicity doesn't have an impact on predicting the performance of students. Later on, this observed accuracy was increased to 79.8% by doing the oversampling.

It has been observed that predicting the performance of the student is important to improve their performance [28]. For that, he has conducted a study that was planned to scrutinize the use of transfer learning coming from the deep neural networks by conducting a study on students involved in higher education studies by evaluating their performance. For that reason, large numbers of examinations were carried out based on facts and features emerging from five obligatory course subjects of undergraduate programs. To perform this study simple deep neural network architecture was used through 3 process model as shown in Figure 4, which was mainly consisted of four layers: an input, an output, and two hidden layers. Dataset used to perform this study was comprised of 5-course subjects' data offered by Aristotle University of Thessaloniki in Greece. The dataset that was taken consisted mainly of six contrasting sorts of learning materials: assignments, resources, URLs, forums, folders, and pages. Conducting this experiment proves to be beneficial the resulted outcome was having satisfactory accuracy. In the future, further efficiency of the transfer learning model can be improved for getting better and more accurate results.

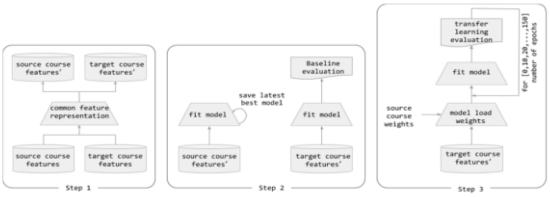


Figure 4: Three-process proposed method [28]

III. Prediction Techniques

Prediction techniques are used in educational data mining to predict the performance of students in academia. For performing such prediction-related tasks several algorithms are taken into account that including density estimation, decision tree, and artificial neural network. Prediction techniques can be classified as supervised and unsupervised learning methods as shown in fig.5.

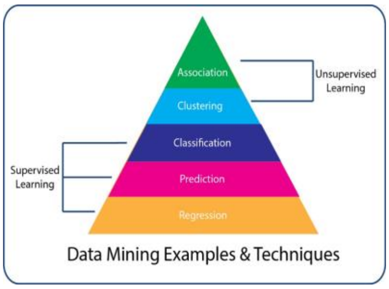


Figure 5: Data Mining Prediction Techniques [29]

Some of the techniques used for making predictions in EDM are as follows [51]:

A. Classification:
One of the most frequently applied techniques for prediction in educational data mining is classification, which has engaged a huge number of pre-classified instances to build a prototype or a model that can categorize the enormous records of the population. This process of classification technique involves two basic phases: classification and learning [30]. In the segment of learning the data that is used for training is examined by using a classification algorithm. On the other cross, in the classification, the data that is employed for testing is used for estimating the precision of classification rules. If in case the estimated precision is satisfactory the on tuples of new data, these rules can be applied. In the algorithm of classifier training, pre-defined instances are utilized in determining the set of data attributes for the discrepancy of objects in a proper way. Under the task of classification, several types of algorithms are used that include Artificial Neural Network, Support Vector Machine, and Bayesian. A generic classification algorithm is given below:

Algorithm Classification

```
1. Start with assigning a label to every object.
2. This data is utilized for building the model.
3. Then that model is used for classification.
4. Classified data into labels is compared with actual class labels.
5. Correct classification done by the model is then compared for accuracy calculation.
```

B. Decision Tree:
The best-known method for classification is a decision tree. It comes under the supervised algorithm of machine learning which utilizes the set of rules for making valid decisions that are potentially useful. It has a tree-like structure, starting with the tree node and expanding the remaining nodes to the branch nodes up to the leaf nodes. The work done for making valid decisions is done at the root node for performing certain actions. Internal nodes portray the attributes of the dataset, branches show the rules for making decisions, and leaf nodes depict the outcome. There is no further branching after the leaf node [31].

Algorithm Decision Tree

```
1. Starts from the root node of the image or pattern.
2. Find out the best attribute from the dataset.
3. Divide that into possible subsets for attribute selection
4. Take the best attribute & create a decision tree
5. Recursively repeat the whole process.
```

C. Artificial Neural Network:
Pattern recognition is greatly improved by artificial neural networks. These networks consist of various input and output ports that use data to perform learning to make valid predictions. Each connection in ANN contains certain weights which after combining with input make well-founded predicted outcomes. During the phase of learning the provided weights are adjusted with the input and then after making the correct predictions objects are classified into certain labels or classes. These are best known for determining design or drift in datasets and making a prediction on students' performance in academia [32].

Algorithm Artificial Neural Network

```
1. An input signal is given to ANN in form of an image or pattern.
2. Each input is then designed mathematically
3. Multiply each input to its weights.
4. Sum up all the weighted inputs.
5. If weighted sum==0:
6. Add bias value for a non-zero output
7. Set some threshold values.
8. An activation function is passed by the sum of weighted inputs.
```

D. Regression:
Another mostly used model that is used for making statistical analysis from data is known as regression. It is used to find the relationship between objects in which one variable is dependent and the remaining are independent. The relationship between these variables is conducted through a scatter plot. Its work is initiated with a set of data in which the target outcome value is already known. These models are tested by evaluating the contrast between the estimated and expected values [33].

Algorithm Regression

```
1. Draw up a model firstly in which one variable is dependent and the remaining are independent.
2. Fit the regression line to regression analysis
3. Testing model through:
    a. F-test for a testing model of regression
    b. Multiple F-tests for testing the importance of every intercept and coefficient.
```

E. Naïve Bayesian:
The classifier that is used for statistical analysis of dataset given by classifying them into classes by finding probabilities of class attributes. After finding probabilities of class attributes, classification of sample data is made. Naïve Bayes classifier made a supposition that the consequence of feature that has been made on the given class is

not dependent on the values of other allocated attributes in the dataset. Bayesian classifiers have also displayed huge precision and rapidity when it is applied on big databases. Also, the training of the dataset using this classifier is done fastly and it is also easy to evaluate it [34].The equation for Naïve Bayesian is stated below:

$$P\left(^{H}/_{E}\right) = \frac{P\left(^{E}/_{H}\right) X \;\; P(H)}{P(E)} \quad \ldots\ldots\ldots\ldots\ldots \qquad \ldots. (1)$$

Algorithm Naïve Bayesian

1. Firstly, the probability of attributes is calculated according to the classes in a dataset they fit in.
2. Sum up the dataset
3. Data condensation by class.
4. Then probability is calculated through a formula of the probability density function.
5. Then the calculated statistics from the dataset are used to calculate the further probability for new data.

Table 2 demonstrates the prediction techniques that have been most commonly used by researchers in educational data mining for envisaging the performance of students.

Table 2: Prediction algorithm as applied in EDM

No	Paper Reference	Algorithm	Tools Used	Results
1	[10]	Multivariate Regression machine learning algorithm	Python (scikit-learn v0.20.3)	Made the students' performance prediction more accurately
2	[4]	Random Forest Quadratic discriminant analysis K-Nearest Neighbor Linear Regression Support Vector Machine	Adaptive recommendation system	Recommended best ML algorithms for each department of faculty having higher accuracy
3	[35]	Bayesian Classification	Matlab	Predicted Student's Final Mark
4	[38]	Bayes Network Random Forest J48 and PART classifiers	Weka	Random forest showed the best accuracy as compared to other classifiers
5	[30]	ID3	Weka	Predicted the semester's end performance of students
6	[37]	C4.5 ID3	Weka	Performance of students is done based on their External and Internal marks
7	[39]	Bayesian Classification	Matlab	Evaluating the big prospective attributes that affect the performance of students
8	[34]	Decision Tree classifier Bayesian Classifier	Weka	Decision Tree has proved to be the best classifier in terms of accuracy for predicting the performance of students
9	[40]	Naïve Bayesian	Weka	Predicted performance of students
10	[17]	ANN Random Forest Tree Decision Tree	Rapid miner	Predicted performance of students in final exams at their graduation level

The above-given table 2 enlists the previous work that has been done in the sector of educational data mining for undertaking performance estimation of students in their academic course more precisely. This table also illustrates that for accompanying this task most of the researchers have used Random Forest Tree and Naïve Bayesian for performance evaluation than other remaining classifiers like: Decision Tree, ID3, ANN, KNN, J48, C4.5, and support vector machine. Some of these researchers have used a blend of these classifier algorithms for predicting the performance of students. The table also depicts the work of researchers in terms of tools that have been most frequently and rarely used for prediction. Most of the time WEKA tool has been seen to be used for making estimates by researchers. Rapid miner and MATLAB have been utilized hardly for doing performance evaluation.

IV. Proposed Methodology

This section introduces the proposed framework for predicting students' performance in the course. It uses students' educational information and performs classification model on it. Figure 6 explains the working of the proposed model. It starts with the gathering and preparation of students' data. Data pre-processing is done to clean the data. The model is generated by using a data mining classifier.

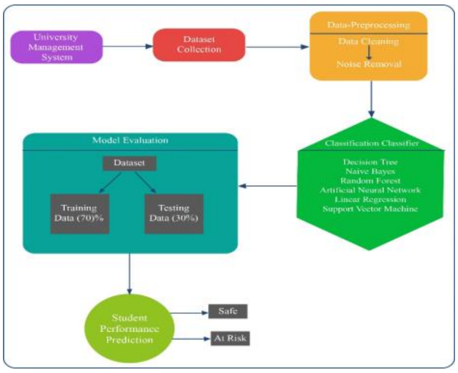


Figure 6: Proposed Data Mining Framework

Here we used six different classifiers i.e. Decision Tree (DT), Linear Regression (LR), naïve Bayes, Support Vector Machine (SVM), Artificial Neural Network (ANN), and random forest. The dataset is tested with each classifier to see the performance of each classifier with the dataset.

The performance of the model is assessed by a performance classification operator. The algorithm of the proposed framework is given below:

Proposed Algorithm Student Performance Prediction

```
INPUT: Student Educational Dataset
OUTPUT: Prediction (Safe/At Risk
1. Begin:
2. Collection of students' academic data.
3. Combine data
4. Apply pre-processing techniques
   a. Noise removal
   b. Filter data
5. Find the result of each course by naïve
   Bayes algorithm
   For each course do
   Apply model
6. Analyze result after prediction
7. End:
```

V. Experiments and Results:

This section converses the process of experiments alongside the results of the proposed methodology. Once the data is cleaned, it is ready to use for classification purposes. Sessional and internal marks are the attribute set used with the classifier. The naïve Bayes classifier is used to generate a prediction model.

5.1. Dataset

Dataset is the assembly of data used for data mining. Dataset used here for prediction is comprised of internal marks and sessional marks of students from the bachelor's program. Sessional marks are the scores students get after the lecturer conduct an examination. Internal marks include assignment, presentation, and attendance marks. The data of 179 bachelor degree students from the Information Technology department is used. Among the data of 176 bachelor's students it includes, 29 students from Artificial Intelligence (AI) course, 35 from Human Resource Management (HRM) course, 31 from Technical Business Writing (TBW), 39 from Technology Management (TM), and 42 from Web Technologies (WT) course. This dataset is divided into training and testing data. The dataset used for learning is termed training data and the dataset used for testing is known as testing data. Training and testing data is divided 70 and 30 percent respectively.

The given table below offers the data set of 25 students for predicting results of bachelor's degrees. Here's given a small glimpse of data used for experimentation purposes.

Table 3: Dataset for Bachelor Degree students

Roll No	Sessional Marks	Internal Marks
1	22	5
2	20	6
3	6	6
4	21	6
5	21	6
6	21	6
7	21	6
8	27	7
9	21	6
10	24	6
11	8	2
12	21	7
13	22	6
14	21	6
15	21	6
16	24	7
17	20	5
18	22	7
19	23	6
20	22	6
21	15	4
22	22	6
23	26	6
24	22	6
25	21	7

5.2. Preprocessing of Dataset

After the assemblage of data, it is made refined through the procedure known as "preprocessing". It is a complex and difficult task and occasionally this segment can take more than half of the time that is required to resolve the whole problem in the mining of data [47]. It is mandatory to clean the data before employing it for training the model because most of the data gathered are not in the useable format most of the time. So, it needs to refine the data through preprocessing it and filtering out the data intended to be useful according to the problem of data mining.

Data sometimes is not in a comprehensive form, to manage this issue data is made complete through the attributes of interest that are necessary to carry out certain tasks for resolving the problem. Old traditional techniques were previously used to filter data that were having some problems [48] which contain feature engineering. Before passing the dataset to any algorithm, it is made clean for eradicating the attributes and terms that are not related to the process of mining [49]. Through this stage, data that is missing is occupied and were made free of redundancy and data becomes free of having errors and mistakes in it.

5.3. Tools Used

There are diverse tools offered to facilitate the researchers to conduct experiments on different EDM approaches. Here, we discuss some of the well-known EDM tools.

- *Rapid Miner:*
  A platform utilized for analyzing datasets in data mining and for building models for evaluation is known as Rapid Miner. It contains many algorithms for regression and classification. It is used for performing tasks relevant to clustering and association rule mining. Their packages also support some functionalities of bootstrap. Although it is more influential than other tools of data mining in terms of having more functionalities, also contains some limitations for new features of engineering and feature selection. It's easy to import the dataset into it for training and testing models from them. It facilitates a graphical representation of outcomes by generating graphs [44].

- *Weka:*
  A platform that is open source and contains many functions for building models is known as WEKA. Provide users to call instances through Java environment and command-line interface for using built-in algorithms. It has a consistent interface that is easy to use. It provides an easy graphical representation of models and data. It contains many algorithms for association mining, clustering, and classification. It has more powerful techniques for implementing machine learning tasks but is not good enough in making statistical analyses [45].

5.4. Performance Metrics

Figure 7depictsthe performance metrics to evaluate the performance of the proposed methodology. It is the confusion matrix of 2x2 used for plotting the diversity between the proposed values of datasets and the predicted values estimated by the models for making different assessments.

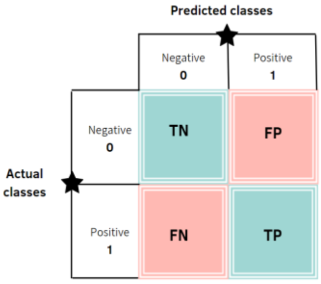


Figure 7: Confusion Matrix

i. *Accuracy:*
Accuracy [43] is the number of correct predictions in the dataset to the total number of given inputs and is collected by eq. 2.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \ldots. (2)$$

ii. *Precision:*
Precision can be defined as the number of correct predictions to the total number of inputs. The precision can be calculated by the following eq. 3.

$$Precision = \frac{Tp}{TP+FP} \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

iii. *Recall:*
It can be defined as the correct predictions of class to the total number of inputs of that class. Recall can be calculated with the help of eq. 4.

$$Recall = \frac{Tp}{TP+FN} \quad \ldots\ldots\ldots\ldots (4)$$

iv. *F1 Score:*
It is difficult to decide whether high precision or low recall is better or vice versa when comparing different models. F1 score combines both precision and recall to calculate results. Eq. 5 shows the f1 score:

$$F1 - Score = \frac{2*(Precsion*Recall)}{Precsion+ Recall} \quad \ldots. (5)$$

According to the proposed methodology mentioned in section 5 and the experiment conducted on the data, the performance of the system is evaluated in terms of accuracy, precision, recall, and f1 score.

5.5. Experiments

A total of 30 experiments were performed for the performance estimation of each of the six classifiers for five different courses. Students' data is fed to each classifier and the result is evaluated. For the justification of the proposed model, we have used the data of 50 students of five different courses which were taught during the session 2021-2022. Training and testing data were given to the model for prediction. Table 4 displays the total number of students' data used for experimenting with each course.

Table 4: Set of data used during experimentation

Courses	Students	Safe	At Risk	% At Risk
AI	29	17	12	41.3
HRM	35	31	4	11.42
TM	31	30	1	3.22
WT	39	33	6	15.38
TBW	42	36	6	14.28
TOTAL	**176**	**147**	**29**	**17.12**

Among 176 students the data of 147 students were those who were not at risk of failure and are safe. 29 students out of 176 were at risk. An average of 17.12% of students was at risk of dropout. To visualize the performance of an algorithm error matrix or confusion matrix is used.

Table 5 displays the confusion matrix that shows the difference between actual and predicted values of classifiers. Against each classifier of data mining predicted technique ratio of safe and at-risk students is predicted by the model accurately.

Table 5: Confusion matrix for different classifiers

Classifiers	Known Labels	Predicted Labels	
		Safe	At Risk
Decision Tree	Safe At Risk	36 0	4 15
Linear Regression	Safe At Risk	34 2	4 13
Naïve Bayes	Safe At Risk	35 1	2 15
SVM	Safe At Risk	34 2	3 23
ANN	Safe At Risk	35 2	3 14
Random Forest	Safe At Risk	34 2	2 15

The decision tree classifier predicted 4 out of 36 safe students as students who are at risk. The predicted value for all risk students was correct. 4 safe students were also considered at risk by the linear regression model. It also predicted 2 risk students as safe. Naive Bayes classifier gave two wrong values for safe students and predicted one at risk student as safe. 3 out of 34 safe students were reported as at risk by the SVM classifier. It also predicted 2 at risk students as safe. ANN classifier also got some wrong values with 3 out of 35 safe students as at risk and 2 out of 14 at risk students as safe ones. Random forest classifier depicts 2 out of 34 safe students as students who are at risk and 2 out of 15 at risk students as students who are safe. The accurate values are arranged in a diagonal line from top left to bottom-right of the matrix. More errors were made by predicted safe values as values at risk than vice versa. The true and false predicted values of all the classifiers are specified in table 5. A comprehensive study of different methods conducted on six different classifiers is shown in the confusion matrix.

From the confusion matrix of each of the models of five different courses, it has calculated the accuracy of models and it shaped the result that is displayed in below-given figure 8. It displays the results of the experiment conducted on five courses. It projected the number of students at risk and those who are safe correspondingly for each classifier.

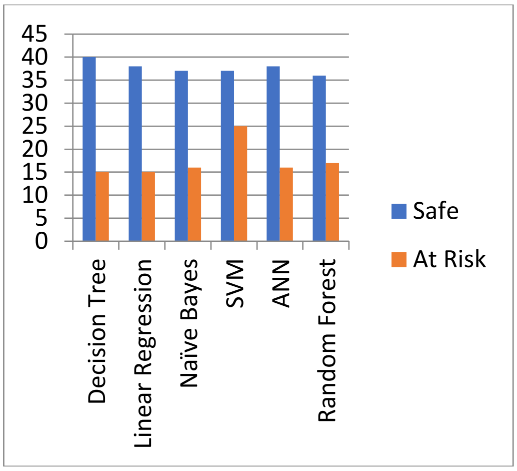


Figure 8: Predicted outcome of the experiment

Very strict decisions are followed in the case of the decision tree model. The students at the boundary of safe students will also be recognized as at risk. The normal students will also be under the supervision of experts. This will result in better performance of students at the end of the courses. Each classifier predicted how many students are at risk and how many are safe. According to the decision tree classifier, 40 students are those students who are safe and 15 students are at the edge of dropout. The linear regression classifier predicted 38 students to be safe and 15 students to be at risk. Calculation through Naïve Bayes shows a 37 to 16 ratio of safe and at risk students. According to SVM the number of safe students is also 37 and students who have a chance of dropout are 25. Artificial neural network classifier predicted a ratio of 38 to 16 for safe and at risk students. The ratio for a random forest classifiers was 36 to 17.

Table 6:Performance evaluation of the models

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision Tree	95.55	100	90	94.73
Linear Regression	88.03	94.44	89.47	91.88
Naïve Bayes	93.33	97.22	94.59	95.88
SVM	89.57	94.44	91.89	93.14
ANN	85.71	94.59	92.10	93.32
Random Forest	91.11	94.44	94.44	94.44

The above-given Table 6 depicts the performance evaluation degree of each of the classifiers. It comprises the accuracy, precision, recall, and f1 score values of all the classifiers against the provided data for 5 different courses. Calculating these performance metrics is a good way for the assessment of any model.From the above values, it can be seen that the decision tree is the classifier with the highest accuracy level of 95.55%. Naïve

Bayes has an accuracy value of 93.33, the second-best in this case. The F1 score of the decision tree is however less than Naïve Bayes. Naïve Bayes has an f1 score of 95.88% while the f1 score of the decision tree is that of 94.73%. Artificial neural network classifier proved to be least effective among six classifiers with an accuracy of 85.71%. Random forest classifier attained a constant result in terms of precision, recall, and f1 score. Although the accuracy of the decision tree is best greater does not mean that the model has a good performance on calculating the specific label. To tackle this other performance metrics are used. The results of the other classifiers are also not bad but Naïve Bayes has the highest f1 score value and hence it is proved to be the best classifier to predict student performance.

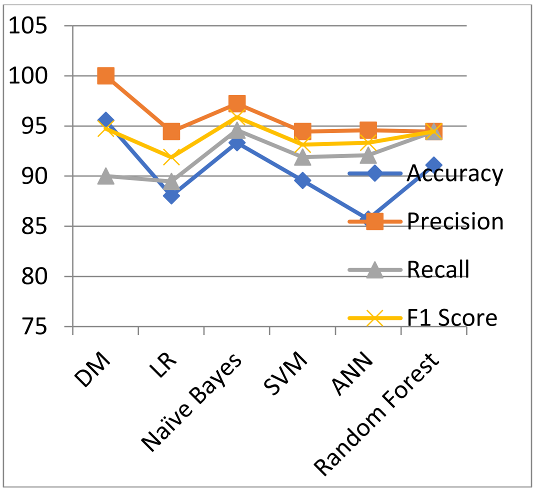


Figure 9: Performance evaluation of the classifiers

A graphical and pictorial view of the performance evaluation metricsof different classifiers has displayed in figure 9. It shows the accuracy, precision, recall, and f1 score values for all mentioned classifiers. Accuracy values range from 89.57 to 95.55 for artificial neural networks and decision trees. Both artificial neural networks and support vector machines have the same accuracy i.e. 89.57. The f1 score range from 91.88 to 95.88 for linear regression and naïve Bayes respectively. It shows the highest precision as well as accuracy value for decision tree and highest recall and f1 score value for naïve Bayes.Naïve Bayes with the highest f1 score value is the most effective classifier for predictive student performance in educational data mining.

VI. CONCLUSION:

Early prediction of students' performance is crucial for reducing students' dropout. To confront this subject, an examination of diverse classifiers is done in this paper. It assists in discovering weak areas of students where they are incapable to perform well. The lecturer can then take necessary actions that can facilitate them. Two attribute sets internal and sessional marks are fed to the model. The model then predicted the students who are at risk of dropout and those who are safe. The experiment has been conducted using the above defined proposed methodology for conducting the task of students' assessment. The work of former researchers was also been considered to get the knowledge of feeble areas in the field of EDM which required further work and improvements. Thus, working on some of those limitations has inclined to improvise the evaluation of students efficiently and more accurately.

For predicting at the bachelor level, the model was trained by the datasets of students against courses of AI, HRM, TM, WT, and TBW. This dataset was divided between training and testing data for producing possibly useful validations. A confusion matrix was used to investigate the actual and predicted outcome ratio from the model as compared to each classifier. Later, those fall outs were visualized through a bar graph for predicting outcome which showed the outcome for each classifier differently. The projected outcome was made to categorize students into two class labels, one for those who were safe and the second one for those who were at risk. Performance metrics were further deployed for the approximation of different classifiers to forecast the best one out of them. Then the resulted outcome showed that the Naïve Bayesian classifier is the greatest of all classifiers in terms of making valid predictions more accurately. F1 score for Naïve Bayesian was high among all the other classifiers due to its resulted best performance. Decision Tree was also proved good one having higher accuracy but have less F1 score than Naïve Bayesian. So, the decision tree is the second effective classifier for making useful evaluations. After them, comes the Random Forest, linear regression, and SVM respectively. The least and the last classifier which was proved to be less effective in making accurate predictions was ANN, showing less accurateness.

This study proved that the Naïve Bayesian is the best classifier and ANN is the least effective classifier for making estimates. Thus, this study proved efficacious in predicting the performance of students using Naïve Bayesian accurately. So, in the future, further work can be done taking the datasets of Master's students as well via the Naïve Bayes algorithm. Also, the dataset taken to perform this study was of 179 students at the Bachelor's level, further studies and research can be made in this area by considering large datasets of these Bachelor students to get more precise and effective predictions. In that way, weak students under the supervision of experts will be at less chance of dropout and will perform better during their academic sessions. This will also advantage these students to be employed by reputed firms that will aid them to make their future improved and safe.

REFERENCES

[1] E. Gomede, E. , Gaffo, F. H., Brigano, G. U., Barros, R. M.D. and Mendes, L. D. S., 2018. Application ofcomputational intelligence to improve education in smartcities, Sensors Journal in MDPI, 18, 267, 1-26.

[2] Gaviria, A. (2002). Los quessuben y los quebajab: education y movilidad social en Colombia. Fedesarrollo, Alfaomega..

[3] Dr. Ajay Bhardwaj, June 2016, 2[2], 23-38. Importance of education in human life: A holistic approach.

[4] Mohamed Ezz, Ayman Elshenawy, 2019, Adaptive recommendation system using machine learning algorithms for predicting student's best academic program

[5] Baker, E. (Eds.) International Encyclopedia of Education (3rd edition). Oxford, UK: Elsevier.

[6] Fatima Alshareef1, Hosam Alhakami2, Tahani Alsubait3, Abdullah Baz, International Journal of Advanced Computer Science and Applications, Vol. 11, No. 4, 2020. Educational Data Mining Applications and Techniques.

[7] Cohen L, Manion L, Morrison L (2007) Research methods in education, six. Routledge, Oxon, pp 211–216

[8] Hashmi Hamsa, Simi Indiradevi, Student academic performance prediction model using decision tree and fuzzy genetic algorithm, Procedia Technology 25 (2016) 326-332

[9] Concepción Burgos , MaríaL. Campanario , Daviddela Peña , JuanA. Lara, David Lizcano , María A. Martínez (2017), Data mining for modelling students' performance: A tutoring action plan to prevent academic dropout.

[10] Aimad Qazdar, Brahim Er-Raha, Chihab Cherkaoui, Driss Mammass (2019)A machine learning algorithm framework for predicting studentsperformance: A case study of baccalaureate students in Morocco

[11] Boran Sekeroglu, Kamil Dimililer, Kubra Tuncal, Student Performance Prediction and Classification Using Machine Learning Algorithms

[12] Fan Yang, Frederick W.B. Li (2018) Study on student performance estimation, student progress analysis, and student potential prediction based on data mining

[13] Raza Hasan, Sellappan Palaniappan, Abdul Rafi EZ Abdul Raziff, (2018) Student Academic Performance Prediction byusing Decision Tree Algorithm

[14] Zafra Amelia, Romero Cristóbal, and Ventura Sebastián, (2013) "DRAL: a tool for discovering relevant e-activities for learners," Knowledge and Information Systems, vol. 36, no. 1, pp. 211–250.

[15] Ammar Almasri, Erbug Celebi, and Rami S. Alkhawaldeh, (2018) EMT: Ensemble Meta-Based Tree Model for Predicting Student Performance

[16] Le Hoang Son, Hamido Fujita, Springer Science+Business Media, LLC, part of Springer Nature 2018, Neural-fuzzy with representative sets for prediction of student performance

[17] Stamper J, Niculescu-Mizilm A, Ritter S, Gordon GJ, Koedinger KR (2010) Data set from KDD Cup 2010 educational data mining challenge. Available at: http://pslcdatashop.web.cmu.edu/ KDDCup/downloads.jsp

[18] Kabakchieva D (2012) Student performance prediction by using data mining classification algorithms. Int J Comput Sci Manag Res 1(4):686–690

[19] Hidayah I, Permanasari AE, Ratwastuti N (2013) Student classification for academic performance prediction using neuro-fuzzy in a conventional classroom. In: Proceedings of the 2013 IEEE international conference on information technology and electrical engineering, pp 221–225

[20] Marquez-Vera C, Cano A, Romero C, Ventura S (2013) Predicting ´ student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Appl Intell 38(3):315–330

[21] Student Performance UCI machine learning repository. Available at: https://archive.ics.uci.edu/ml/datasets/Student+Performance

[22] Raza Hasan, Sellappan Palaniappan, Salman Mahmood, Ali Abbas,Kamal Uddin Sarker, and Mian Usman Sattar (2018) Predicting Student Performance in Higher Educational Institutions Using Video Learning

[23] Analytics and Data Mining TechniquesE. T. Lau,L Sun1,Q. Yang (2019) Modelling, prediction and classification of student academic performance using artificial neural networks

[24] Hajra Waheed, Saeed-Ul Hassan, Naif Radi Aljohani, Julie Hardman, Raheel Nawaz (2019), Predicting Academic Performance of Students from VLE Big Data using Deep Learning Models

[25] Aderibigbe Israel Adekitan, Odunayo Salau, 2019, Toward an improved learning process:

the relevance of ethnicity to data mining prediction of students' performance, Springer Nature Switzerland AG 2019

[26] Lau ET, Sun L, Yang Q (2019) Modelling, prediction and classification of student academic performance using artificial neural networks. SN Appl Sci 1(9):982. https://doi.org/10.1007/s4245 2-019-0884-7

[27] Badejo JA et al (2018) Data sets linking ethnic perceptions to undergraduate students learning outcomes in a Nigerian Tertiary Institution. Data Brief 18:760–764

[28] Maria Tsiakmaki, Georgios Kostopoulos, Sotiris Kotsiantis and Omiros Ragos, 2020, Transfer Learning from Deep Neural Networks for Predicting Student Performance, 26504 Rio Patras, Greece.

[29] Chady EI Moucary "Data mining for Engineering Schools", International Journal of Advanced Computer Science and Applications(IJACSA) Vol.2, 2011

[30] Birijesh Kumar, Saurabh Pal, "Mining Educational Data to Analyze Students' Performance", International Journal of Advanced Computer Science and Applications Vol 2, 2011

[31] A.Dinesh Kumar, R.Pandi Selvam2, K.Sathesh Kumar, Review on Prediction Algorithms in Educational Data Mining, International Journal of Pure and Applied Mathematics Volume 118 No. 8 2018, 531-537

[32] Elaf Abu, Amrieh, Thair Hamtini, Ibrahim Aljarah, "Mining Educational Data to Predict Student's Performance Using Ensemble Methods", International Journal of Database and Theory and Application,Vol 9, 2016.

[33] Febrianti Widyahastuti, Viany Utami Tjhin, "Predicting Students Performance in Final Examination using Linear Regression and Multilayer Perceptron", IEEE, 2017.

[34] Dorina Kabakchieva, Predicting Student Performance by Using Data Mining Methods for Classification, Cybernetics and Information Technologies.

[35] Umesh Kumar Pandey S.Pal, "Data Mining: A Prediction of performer or underperformer using classification", International Journal of Computer Science and Information Technologies (IJCSIT) Vol 2(2),2011.

[36] Mohamed Ezz & Ayman Elshenawy, Adaptive recommendation system using machine learning algorithms for predicting student's best academic program, Springer Science+Business Media, LLC, part of Springer Nature 2019.

[37] S.Anupama Kumar, Dr.Vijayalskhmi, "Efficiency of Decision Trees in Predicting Student's Academic Performance", CCSEA,2011.

[38] Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwi, Najoua Ribata, Educational Data Mining and Analysis of Students' Academic Performance Using WEKA, Indonesia Journal of Electrical Engineering and Computer Science.

[39] Birijesh Kumar Bharadwaj, Saurabh Pal," Data Mining: A Prediction for performance improvement using classification", International Journal of Computer Science and Information Security, Vol 9, 2011

[40] Azwa Abdul, Nur Hafieza, Fadhilah Ahmad, "Mining Students' Academic Performance", Journal of Theoretical and Applied Information Technology", Vol 53, 2013.

[41] Raheela Asif, Saman Hina, and Saba, "Predicting Student Academic Performance using Data Mining Methods", International Journal of Computer Science and Network Security (IJCSNS), VOL.17, 2017.

[42] Ida Bagus Adisimakrisna Peling, I Nyoman Arnawan, I Putu Arich Arthawan, and IGN Janardana, Implementation of Data Mining to Predict Period of Students Study Using Naive Bayes Algorithm, International Journal of Engineering and Emerging Technology, Vol. 2, No. 1.

[43] Vikash Kumar, Ditipriya Sinha:A robust intelligent zero-day cyber-attack detection technique,2021

[44] Stefan Slater, Srec´ko Joksimovic, Ryan S. Baker, Dragan Gasevic, Vitomir Jovanovic, Tools for Educational Data Mining: A Review, Journal of Educational and Behavioral Statistics Vol. XX, No. X, pp. 1–22 DOI: 10.3102/1076998616666808 # 2016 AERA. http://jebs.aera.net.

[45] Blaz Zupan, Ph.D. ,b, Janez Demsar, Open-Source Tools for Data Mining, Clin Lab Med 28 (2008) 37–54.

[46] C. Romero *, S. Ventura, Educational data mining: A survey from 1995 to 2005, Expert Systems with Applications 33 (2007) 135–146.

[47] Bienkowski, M., Feng, M., & Means, B. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief (pp. 1–57). Washington, DC: U.S. Department of Education, Office of Educational Technology.

[48] Romero, C., Romero, J. R., & Ventura, S. (2014). A survey on pre-processing educational data. In Educational data mining (pp. 29–64). Cham, Switzerland: Springer.

[49] Koutri, M., Avouris, N., & Daskalaki, S. (2004). Ch. A survey on web usage mining

techniques for web-based adaptive hypermedia systems.

[50] Mohammad Alsuwaiket, Anas H.Blasi, Ra'fat Al-Msie'deen, Formulating module assessment for Improved academic performance predictability in higher education, Engineering, Technology & applied science research, Vol 9, No.3, 2019, 4287-4291.

[51] Mohammad Alsuwaiket, Anas H.Blasi, Khawla Al arawneh, Refining student marks based on enrolled Modeules' Assessment methods using Data Mining Techniques, Engineering, Technology & applied science research, Vol 10, No.1, 2020, 5205-5010.

[52] Parisa Shayan, ErsunIscioglu, an assessment of students' satisfaction level from learning management systems: Case study of Payamnoor and Farhangian Universities, Engineering, Technology & applied science research, Vol 7, No.4, 2020, 5205-501