Analysis of Popular Techniques Used in Educational Data Mining

Satinder Bal Gupta*, Raj Kumar Yadav*, Shivani**
*Associate Professor, Indira Gandhi University, Meerpur, Rewari, India
**Faculty, Indira Gandhi University, Meerpur, Rewari, India
Email: satinderbal@gmail.com, rajyadav76@rediffmail.com,shivanigupta646@gmail.com

The importance of data mining is increasing in education field as it can help both in the improvement of education system and in the growth of students by making predictions. Educational Data Mining (EDM) is a young interdisciplinary work field that helps to deal with the data related to educational perspective. Today, educational institutions collect and archive massive quantities of data, such as students registration, attendance as well as the exam results. Mining of this data helps the institutions to understand students behaviour and interests by extracting all the useful information from the huge data available. Different data mining techniques are being used for mining the data in educational field. Now days, the Artificial Intelligence and Machine Learning techniques are more popular among the researchers to extract the information from the educational databases, as these provide more reliable results as compared to other techniques. In this paper, many popular data mining techniques have been reviewed that are being applied on the educational data to solve the different problems faced by the students so as to improve the learning outcomes of the students

Keywords: Data Mining, EDM, Neural Network, Clustering, Prediction, Classification.

1. INTRODUCTION

Data mining is a process of analysing large data sets and finding hidden knowledge that can be utilised. Data Mining is finding useful and hidden patterns of information from large datasets available by applying different classification techniques. The process of data mining is called as Knowledge Data Discovery (KDD) i.e. discovery of knowledge from the databases that can help in gathering the useful information from the large databases. Data mining can be considered as a subfield of computer science that can be used in the field of education. When the data mining technique is applied in the field of education, it is called as Educational Data Mining (EDM). Educational data mining acts as a bridge between education of the students and the field of computer science. It is used to uncover the hidden data from the raw data available. Every year, in the educational institutions a huge amount of data is generated Baker and Yacef [2009]. The quality of the analysis of large-scale data related to educational field can be improved with the help of EDM Siemens and Baker [2012]. This data requires to be analysed so that the behaviour of the students can be identified which can be helpful for the teachers to enhance the teaching process so that it can result in effective learning process. EDM helps to improve the educational system by improving the assessment process of students. EDM is a very useful technique to retrieve the hidden knowledge from the large database by applying DM techniques for classification. The different data mining techniques used are statistical techniques based on probability and mathematics such as Naive Bayes Algorithm, Linear and Logistic Regression, machine learning techniques such as supervised learning (Support Vector Machine(SVM) and Decision Tree (DT)) and unsupervised learning (Clustering in Fuzzy Logic) and Artificial Neural Network (ANN), Association Rule Mining etc. Mohamad and Tasir [2013] Manjarres et al. [2018]. In the field of education the most important thing is to identify the students performance and to improve their learning and in this Artificial intelligence and machine learning techniques plays a very important role as they provide more accurate results and can handle large amount of data

easily. An educational system has a large amount of educational data which includes data related to students, teachers, resources used etc. Silva and Fonseca [2017].
The data present in large educational databases is passed through the various steps involved in the process of data mining and then the useful data is filtered which is then used according to the future requirements. In the whole process, data collection and gathering is considered as a very important part. The data can be collected from online as well as offline resources Algarni [2016]. The time that is spent on collecting the data must also be taken care of so as to avoid delay. The data collected must be sufficient and complete so that it can help teachers and institutions to understand students Ekubo [2019]. This can be very helpful for the teachers to identify the students behaviour such as those who are slow learners and need special attention. Also, the learning outcomes can be identified and the students can be characterised into groups Peña et al. [2009] and according to that teaching methods can be changed as per need of the groups which can help the teacher to teach them. The use of these techniques can improve the success rate of the students which can directly affect the enrolment of students in different courses, retention rate, chances of dropouts etc. So, in education data mining plays a very important role.

2. RELATED WORK

In the educational institutions the data of the students keep on increasing every year. So, there is a need to analyze the data so as to use it to improve the teaching process in institutions. Educational Data Mining process can help to improve the teaching and learning process by extracting useful information from the educational data Abu Tair and El-Halees [2012]. With the help of EDM, higher educational institutions can be provided with effective ways so as to improve the effectiveness of institution and the learning process of the students Huebner [2013], Baker [2014]. A lot of research work has been done in this field by hundreds of researchers. These researchers use many mining techniques to mine the data from the educational databases. There are more than 15 such mining techniques that were used by the researchers in the education field. These are Decision Trees (DT), Regression Trees (RT), Markov Chains (MC), Association Rules (AR), Linear Regression (LR), Sequential Patterns (SP), Correlation Analysis (CA), Bayesian Networks (BN), Artificial Neural Networks (ANN), Classification, Clustering, Differential Sequence Mining (DSM), Fuzzy Logic (FL) and Genetic Algorithms (GA). All these techniques are not popular these days due to limitations of these techniques. From last few years, Artificial Intelligence (AI) and Machine Learning (ML) techniques are more popular among the researchers for doing work in educational data mining. The reason behind this popularity is that there is a huge scope to improve the results by using these AI techniques in educational data mining. We have reviewed only the EDM literature that uses these AI techniques that are popular among researchers. Some of the latest papers reviewed by the authors are given in the Table I.

Table I: Data Mining Techniques as applied in EDM

Sr. No.	Reference	Objective of the Paper	Technique Used	Source of Data	Observations

1	Alsuwaiket et al. [2020]	To solve the problem of gap between course work and exam based assessment	Random forest, Naive Bayes classifier	A record of 230,823 students was collected from six departments of a UK University.	The Module Assessment Index (MAI) was used and it was observed that this index helps to increase the accuracy of predicting the average of students obtained in the second year based on the average of the first year.
2	Rastrollo-Guerrero et al. [2020]	To Predict Students Performance	Supervised ML, Unsupervised ML, ANN	Dataset collected from Hellenic Open University included demographic characteristics and grades of students from some tasks.	Supervised Learning provided accurate and reliable results in case of predicting students behaviour.
3	THI and BA [2019]	To support students in selecting the courses	J48, K-Means, Supporting Courses Selection	Data was collected from Civil Engineering Department of Ho Chi Minh City University of Transport, Vietnam of period 2013-16.	The data mining technique were applied on the data collected as an experiment and was found that the experimental results might bring positive results.
4	Khasanah and Harwati [2019]	To Predict Students Performance	SVM, Linear Regression	Data included Senior high school grade, attendance in Ist Semester, GPA in Ist Semester	LR produced better result than SVM
5	Rogers [2019]	To classify students on the basis of performance	Fuzzy technique. (Linear Fuzzy Real Logistic Regression)	Data of 172 students was collected from four elementary schools in Blackbelt of Alabama and Mississippi, USA. Students Survey Answers, teachers evaluation were used as data.	The model used was able to do a successful classification upto 90%.

6	Aulck et al. [2019]	The disbursement of scholarship to students	GA	Data included information of about 72,589 students (DNR applicants from 2014-17)	The predictive classifier for enrolment of students was developed and GA was used with it which resulted in an increase of 23.8% in enrolment yield.
7	Toivonen et al. [2019]	To make knowledge discovery possible through AUI	Neural N-Tree model, Augmented Intelligence Method	Data was collected from 3rd or 4th year Computer Science students in robotics course at University of Eastern Finland, School of Computing.	It was observed that the AUI method produced more accurate results and with this method new knowledge could be discovered easily by the end-user.
8	Moscoso-Zea et al. [2019]	To predict graduation rate	Decision Tree, J48, Random Trees	Students data was collected in the computer science deptt. Of a private university	Random trees provided prcised results but the possibility of interpretation of results was better in case of J48 algorithm.
9	Adekitan and Salau [2019]	To determine the impact of performance of students on their result	NN, Random Forest, Decision Tree, Nave Bayes, Logistic Regression	The GPA of students for the first 3 academic years and final CGPA of 1841 students	Logistic Regression Algo. Achieved max. accuracy of 89.15% while NN achieved the least accuracy of 85.895%.
10	Kaunang and Rotikan [2018]	To Predict Students Academic Performance	Decision Tree, Random Forest	Data was collected using questionnaires which included demographics of students, previous GPA, family background information	DT was found to provide better results with accuracy of 66.9% as compared to Random Forest tech.
11	Wati et al. [2017]	To Predict Students Learning Result	Nave Bayes Classifier, Tree C4.5 algorithm	Academic data of students was collected from academic database	The accuracy and precision % of both algorithms was found. The average accuracy was above 60% but the precision average was only 58.82%.

12	Mousa and Maghari [2017]	To Predict Students Performance	Nave Bayes, Decision Tree, KNN	Data was collected from preparatory male school in Gaza strip. About 1100 records were collected from 7th, 8th and 9th grade students of year 2015-16.	DT classifier gave best results when applied on the collected data.
13	Burman and Som [2019]	To Predict Students Performance	Multi classifier SVM (Linear and Radial Basis kernel)	Psychological features of students was used as dataset	RBF produced better results in comparison with linear kernel.
14	Oloruntoba and Akinode [2017]	To Predict Students Academic Performance	DT, ANN, Bayesian Classifier, KNN, SVM	Students information was obtained from the CS deptt. Of Federal Polytechnic in Nigeria for 2015-16 session of graduated students	SVM gave 98% accuracy in prediction and the error rate was low.
15	Saa [2016]	To Predict Students Performance	Decision Tree (C4.5, ID3, CART, CHAID)	Data was collected using survey distributed to students in daily classes and using online survey with the help of Google Forms. About 270 records were collected which included personal as well as academic information.	It was observed that the performance of students could not be predicted only on the basis of their academic efforts. Other factors are also responsible for doing the prediction.
16	Atta-Ur-Rahman et al. [2018]	To help institutions in selecting better teaching and learning practices such as effective timetabling, medium of teaching etc.	K-means, Apriori algorithm	Data was collected by doing a survey which consisted of 38 questions related to teaching and learning in an educational institution.	It was observed that more aspects of EDM were covered using these algorithms when compared to others.

17	Costa et al. [2017]	To reduce the failure rate of students by identifying those students who are likely to fail at early stage.	SVM and other tech.	Data was collected from two independent and different courses on programming from Brazilian Public University.	SVM technique was found to perform best in order to identify the students who are likely to fail at early stage.
18	Al-Twijri and Noaman [2015]	To propose a model to filter students from the data who satisfy the eligibility criteria for admission.	Data Mining Admission Model (DMAM) using Rule Mining	Data was collected from Saudi University	The proposed method was found to be the best and was recommended to be used in Saudi University for admission process.

3. DATA MINING IN EDUCATION

The process of EDM involves following phases:

(1) The first phase includes identifying the relationship among the data collected and stored in the database. The consistent relationship is found out between different attributes of student by searching the data stored in the repository. To find this relationship different algorithms are used which includes classification, association, clustering, pattern evaluation.

(2) The validation is performed on the discovered relationships in order to avoid over fitting so that predictions can be made without any difficulty.

(3) The obtained relationships are used for making predictions so as to improve the learning of students and making the changes in the teaching process.

(4) Predictions are used and the changes are made in the process of improving education.

The different data mining techniques are used in this field that help both lecturers and students in finding new things and improving knowledge. It works in the form of a cycle and the detailed process can be shown with the help of Fig. 1.

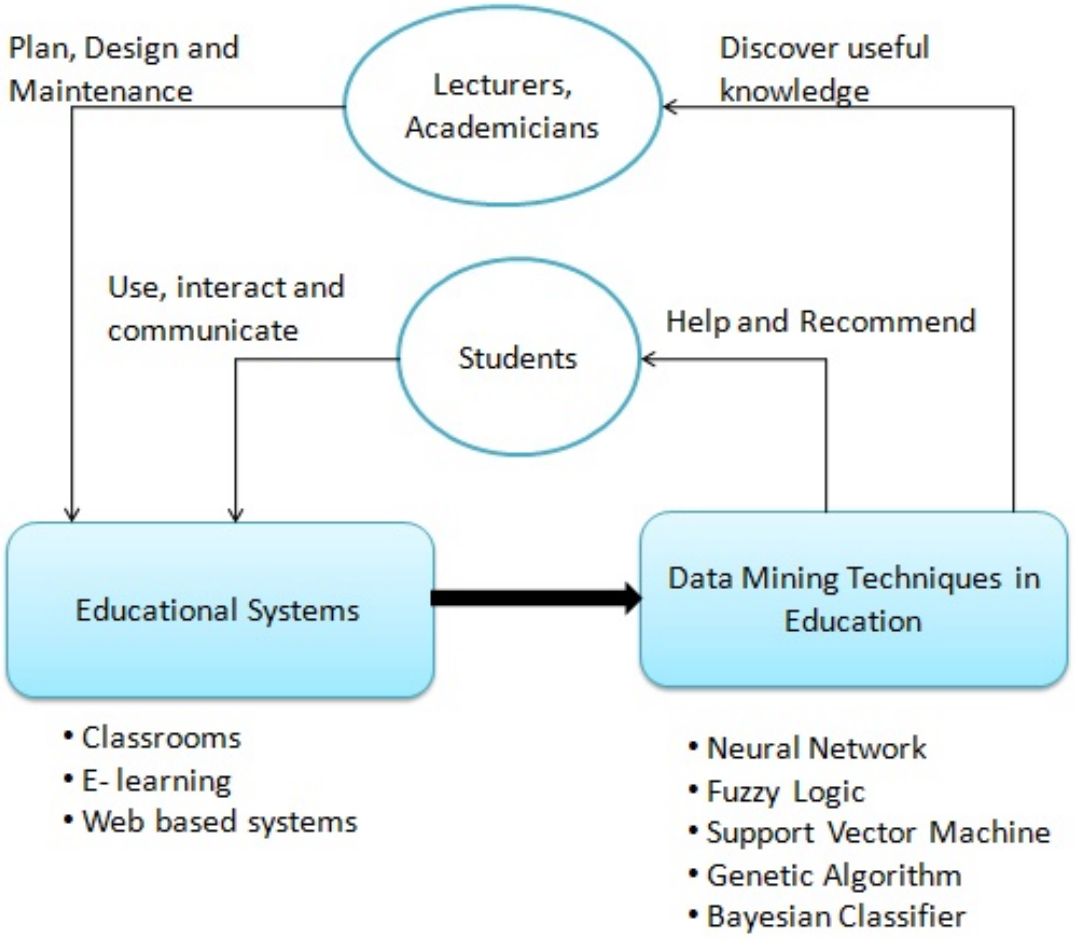

Fig.1. Cycle of DM in the Field of Education

The Fig. 1 shows the use of data mining techniques for both students and lecturers. The lecturers and academicians are mainly responsible for the planning, designing and maintaining the system of education. This is done in order to improve the teaching and learning process. The data mining process can help in discovering new knowledge by applying classification techniques on the students database in which data collected from the different institutions is stored. This knowledge is thus used by the lecturers to improve the course to be taught to the students and make some changes to it according to the requirements. The knowledge obtained from applying data mining techniques on data of students can be used for the benefit of students as it can help them to collect the suitable course and also students can interact with the educational system so as to improve their learning Zain et al. [2014].The EDM process can be explained with the help of Fig. 2.

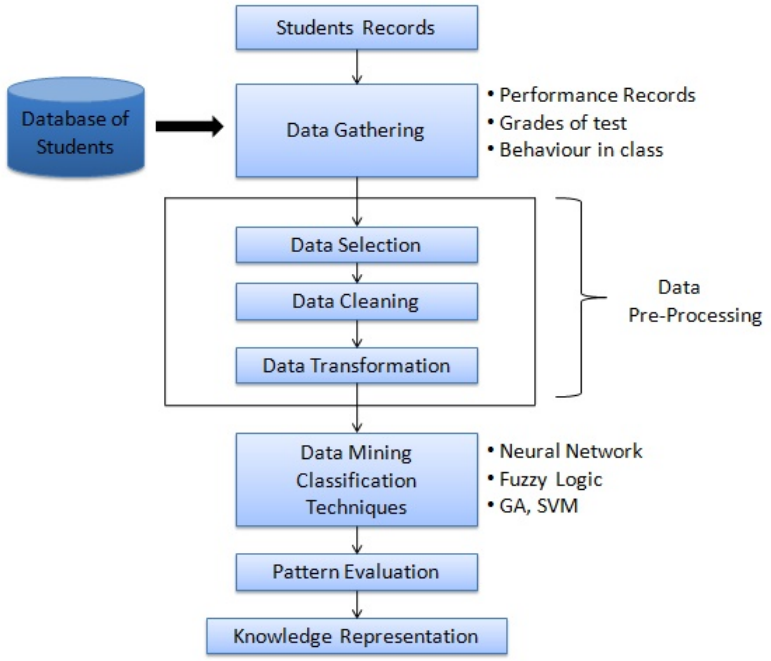

Fig. 2. EDM Process

(i) **Data in the Educational Environment (Raw Data):** The data in the educational domain can be classified into structured and unstructured data collected from multiple sources Zain et al. [2014] as shown in Fig. 3.
Structured Data: This type of data is obtained from any specific source and has very less possibility of being vague. This data is present in regulated form and has its own explanation. Some of the sources from where this data can be obtained are: Intelligent mentors, Learning Management Systems etc.
Unstructured Data: The data that is not obtained from any specific source and can contain errors. It can include the data obtained through e-mail messages or any audio files etc. and cannot be considered fully reliable.

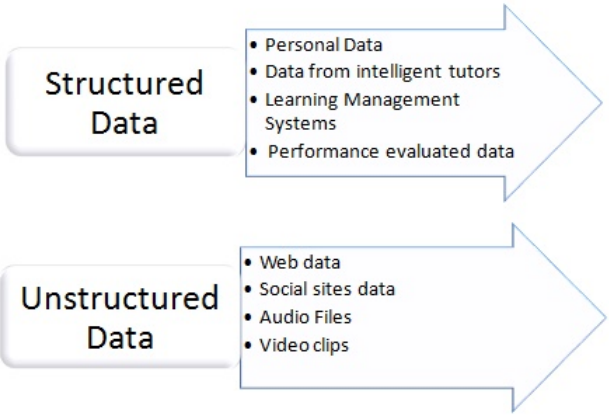

Fig. 3. Types of Data

(ii) **Data Processing:** The raw data collected from various sources in the previous step is now processed to obtain the useful data from the vast amount of data available.
   a) **Data Selection:** The data is selected from the database. In this step, the data which is needed in doing the analysis is selected. In education, the data of a student can be more but only useful data required for doing the analysis is selected.
   b) **Data cleaning:** The data obtained can contain missing values or errors and this type of data is considered irrelevant. So, data cleaning is done to handle this problem. It removes missing values, handles noisy data (data containing error). In case of students database, the data missing can be the marks of a student in any subject or there can be error in the values updated. So, these values must either be changed or removed.
   c) **Data Integration:** The same data obtained from multiple sources is put together and the conflicts are resolved. It involves reducing the redundancy present in the data collected. In education, the data related to students can be obtained from their teachers, tutors, scholarship performance etc. And all this needs to be combined to get information.
   d) **Data transformation:** This step performs transformation of data to obtain the data in a form that the data becomes suitable for applying the data mining process. In this case for example, if the data values are .5, .001,.9 , then these values can be represented as 0.50, 0.001, 0.90 which can be used anytime without causing error.

(iii) **Data Mining:** This can be considered as the most important part of the whole process. It includes tasks such as classification, prediction, analysis etc. The different techniques are applied to extract patterns from the data such as Neural Network, Fuzzy Logic, and Genetic Algorithms etc. In Neural network, the data of students obtained from previous step is given to the input units. Then, activation functions are applied to the input data and output is obtained. In Fuzzy Logic, the membership values are assigned to the input data and based on these values clusters are formed. One cluster contains data of similar type. Mamdani Fuzzy inference system is used which performs Fuzzification and Defuzzification on the data to obtain crisp output from the fuzzy output. Genetic algorithms are based on the phenomenon of natural selection and genetics.

(iv) **Pattern Evaluation:** The patterns extracted in the previous step are now evaluated to obtain the required information. This step provides all the useful information that is needed. The information extracted is according to the need of the analysis to be done on the students data.

(v) **Knowledge Representation:** This information thus obtained is represented in a form that is easy to understand such as tables, reports etc. which can be stored for future use also Baker [2014].

4. CHALLENGES OF EDUCATIONAL DATA MINING

The various challenges faced in the Educational Data Mining (EDM) are as follows:

(1) **Progressive nature of educational data:** As the data regarding students in the educational institutions is growing exponentially, it is becoming very difficult to store the data in the data warehouse. It is required to identify the interests and intentions of the students and the impact it puts on the educational institution. Also, the growing data must be aligned and translated properly. This increasing data becomes difficult to be utilised optimally.

(2) **Chances of uncertainty in the data:** Some uncertain errors can be present in the data collected regarding students and a model used cannot predict accurate results.

(3) **Relationship between teachers and students according to the expertise:** The students in the final year of any engineering institution have to complete a project which is a research done by the students in the area of their interest. In doing this research, supervisors are allotted to each student taken into consideration their area of expertise and availability.

But in reality it is not possible to assign the supervisor to all the students with same area of expertise. So, relation between interest areas is required to be found out and association rule mining can be used to solve this problem.

(4) **Lack of compatibility among the data:** The gap among the different types of data can be removed with the help of Neuro Fuzzy mining technique. This technique can create clusters of data according to the similarity between the data.

5. ANALYSIS OF TECHNIQUES USED IN EDM

The different AI techniques that can be applied on the data in Educational Data Mining (EDM) are discussed below briefly:

(i) **Neural Networks:** The neural network architecture involves layers such as input, hidden and output layers. The data is first taken by the first layer and the passed to the next layer after applying some activation function on the data. This layer passes the data to next layer and this process goes on till the desired result is obtained. The network is trained to obtain the desired output and this is done by updating the weights through which these layers are interconnected. The steps involved are:

1. *Data Gathering:* The data of students is collected from various institutions.
2. *Data Selection:* Now some common attributes of all the students are selected e.g. attendance, marks in a particular subject, labs work performance etc. The result of the group is predicted based on the values of these attributes by applying the classification technique i.e. by using neural network.
3. *Use of Neural Network:* The data values are now passed to the layers and output is obtained after applying the activation function.

There are different types of neural networks that can be used in EDM. The NN that can be used are as follows

Feed Forward Neural Network: It does not contain backward pass and data can flow in one direction only. The main problem of this network is that it can end in local minima i.e. it may provide a suboptimal solution at last instead of providing the optimal solution as these can perform calculations with limited patterns only. So, for this the next type is used which involve multi layers.

Convolutional Neural Network: It is a multi layer neural network which involves use of hidden layers in between the input and output layer. It does not contain cycle. The problem with this network is its high cost of computation, need of large data for training of network.

Recurrent Neural Network: It is a back propagation neural network which contains backward loop. It maintains a record of previous inputs in the memory. The main problem is that they require high performance hardware in order to train the network Hernández-Blanco et al. [2019].

Optimization of ANN architecture: There are many algorithms used for the optimization of ANN architecture. Some of them are PSO (Particle Swarm Optimization), NMPSO (New Model PSO), SGPSO (Second Generation PSO). These 3 algorithms considered the following three components for optimization: weights between the interconnected nodes, activation functions, connections between nodes. These algorithms made use of Mean Square Error and Classification Error in order to decrease the number of connections in the architecture of ANN Garro and Vázquez [2015]. One of the algorithms used is IPSONet which is designed to optimize the architecture of Feed forward Neural Network. This algorithm is designed by making improvement in the PSO algorithm.

(ii) **Fuzzy Logic:** This technique uses fuzzy values to obtain the result instead of crisp values. In EDM it can be used by following some steps which are as follows:

1. *Data Collection:* In this step, the data related to students is gathered from institutions.

2. *Database Design:* The data is stored in the database in order to find missed values and to update them Jahan [2015]. The data of students is stored in form of attributes and each attribute is assigned a value which is different for students.
3. *Data Automation:* The data stored in the database is obtained from different sources and thus needs to be sorted. This can be done with the help of SQL queries in the database.
4. *Applying Fuzzy Set Technique:* Fuzzy inference system is used in this step. It can be shown with the help of diagram as follows:

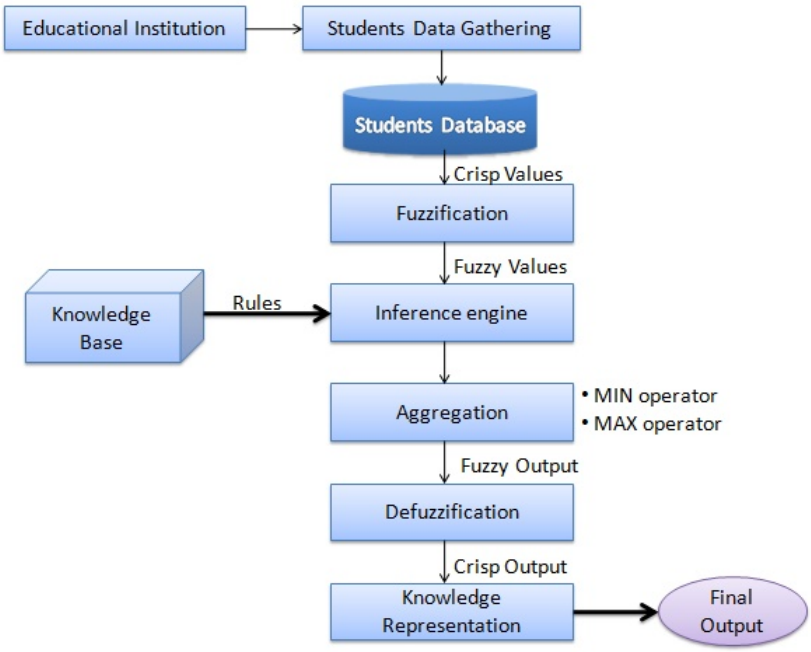

Fig. 4. Fuzzy Logic in Education System

The Fig. 4 shows that this system takes crisp values as input. The values assigned to the attributes in above step are used as input. Then fuzzification of the input is done and the crisp values are converted into fuzzy values e.g. Grades of different students can be converted into fuzzy form as Excellent, Good, Bad, and each variable can be assigned with a membership value. These can be represented in different forms such as triangular or trapezoidal form. These inputs are then passed to the inference engine where rules are applied to the membership functions and then aggregation of these functions is done using max, min operator to obtain the outputs. This fuzzy output is then passed through the defuzzification system to obtain a single crisp value as output Jahan [2015].
The different techniques that use fuzzy logic are as follows:
Clustering: In this technique, clusters or groups are formed based on the similarity. In educational data, the students can be grouped on the basis of their similarity in performance level in the class, attendance, class test marks etc. and then techniques can be applied on these groups to determine their future performance.
K-nearest neighbour: This technique is based on fuzzy clustering and it classifies the data using following steps:
a) The unclassified data of students is taken from any institution e.g. students marks in different subjects.
b) The Euclidean distance is measured of all the data sets from a single neighbour data already classified.

c) The k smaller distance is determined.
d) The list of classes are then determined and the class which has the shortest distance is selected and classified with the class Alfere and Maghari [2018].

(iii) **Genetic Algorithms:** Genetic Algorithms are heuristic search algorithms that are based on natural selection and genetics. GA was proposed by Holland. It is a heuristic method based on the theory given by Charles Darwin - "survival of the fittest." GA was discovered as a useful tool for search and optimization problems. The detailed genetic algorithm can be represented with the help of flowchart as shown in Fig. 5. Each of the steps of the flowchart is discussed below:
Initial state: Genetic Algorithms work on the population. Random solutions are selected to get the initial population. In EDM, the population consists of students and the data includes information related to students such as marks in case of performance prediction.
Fitness Evaluation: Each solution is assigned a fitness value depending on its chances of solving the desired problem. In EDM, the different attributes of student are represented in the form of values i.e. 0s and 1s. This process is called encoding. Each individual student attributes can be encoded in the form of a rule. Example: IF var1= val1, var2= val2...and so on (var= variables and val=values), THEN varX= valX, where Variables are the field names of database and values are the possible values of attributes used in the database. For assigning the fitness value, the precision of the rule formed is considered Altaher and Barukab [2018].
Selection: Based on the fitness value the attributes are selected for further evaluation. The attributes of student with high fitness value are selected. After this some basic operators

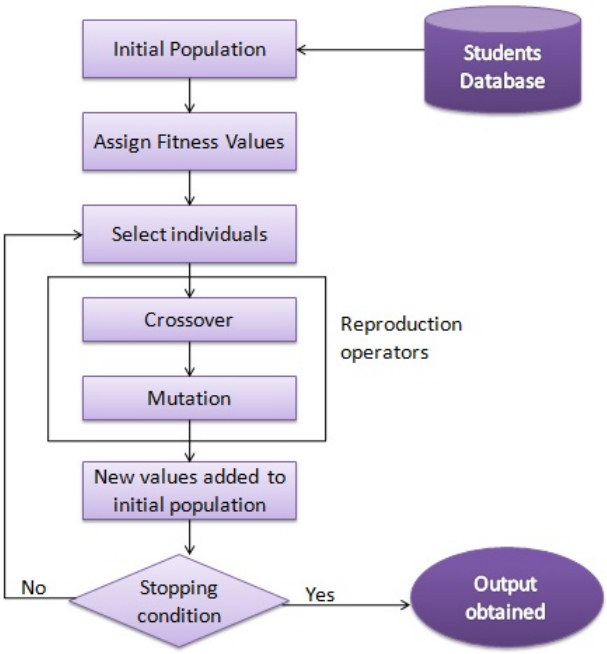

Fig. 5 Genetic Algorithm in EDM

are used which are as follows:

a) **Crossover:** After selection, the crossover operator is applied on the selected values. In this, the attributes of student represented in bits are interchanged according to the type of operator used so as to produce new attribute which is better than the previous one. Thus, the new population is produced and added to the initial population.

b) **Mutation:** It is used to maintain diversity in the population by making a slight change in the chromosome i.e. attribute of student represented in bits Michalewicz [1996]. Then, the fitness value is calculated and the features with the highest value is selected and added to the pool of values and the lowest value is removed from the pool.

Stopping condition: If the selected values are according to the result needed then the process is stopped otherwise the above steps are repeated.

(iv) **Decision Trees:** It is a data mining technique in which data is represented in the form of a tree starting from a root node and ending with leaf nodes. It uses divide and conquer approach to represent the data in tree form. The steps to obtain the final output are:
Data Collection: The data related to students is collected from the institutions. For e.g. Grades of student in different subjects.
Data Processing: The missing values are either replaced or removed.
Use of tools: The tools are used for e.g. WEKA tool to store the data so that data mining techniques can be applied on the data.
Decision Tree Formation: The classification technique is applied and the data stored in WEKA to predict the final GPA of students based on the grades obtained by them. Now, the decision tree algorithms are applied on the data such as ID3, C4.5 etc. These algorithms are used to build trees from the data using the information entropy concept. Then, the useful information can be gained from the tree Al-Barrak and Al-Razgan [2016].
The process of formation of tree can be defined with the help of following steps:

1. The dataset is obtained.
2. The continuous values present in the data set are converted into discrete values.
3. All the attributes present in the dataset are incorporated in the form of a single tree node.
4. If the dataset is found to be homogeneous, then the process is terminated.
5. Otherwise, the non-homogeneous data is represented in tree form by finding the different independent attributes from the students database. Then the nodes of the tree are split into child nodes on the basis of values of attribute chosen.
6. Then the data is checked again and if similar data is found then go to step 4, otherwise perform step 5 again.

These steps can be shown as: *Result Evaluation:* The result is evaluated from the decision tree.
Knowledge Representation: The result obtained can be represented and used later.
There are two phases of decision tree classifier:

1. Building phase: In this phase the decision tree is built by splitting the dataset recursively on the basis of different attributes and their values. But there is a risk of over-fitting of the data which is handled by the next phase.
2. Pruning phase: The tree obtained is generalised and the noisy, missing data is removed from the tree which results in increased accuracy. This step takes less time than building phase.

This technique uses algorithms to make decision trees such as ID3, C4.5, CART algorithms. In **ID3 algorithm**, the splitting attribute is chosen on the basis of the measurement of information gain. It makes use of only categorical attributes to build the tree. But in case of noisy data, accurate results are not obtained. For each attribute, the information gain is measured and the root node is measured by selecting the attribute with highest information gain and then the node is split on the basis of values of attributes. This algorithm uses discrete values for making tree and does not support pruning. **C4.5 algorithm** is an enhanced version of ID3 algorithm as it can handle both continuous and discrete values. It can also handle missing values. It makes use of Gain Ratio to build the decision tree. The attribute with the highest gain ratio is chosen as the root node. It supports pruning phase and uses pessimistic pruning to improve the accuracy. **CART (Classification And**

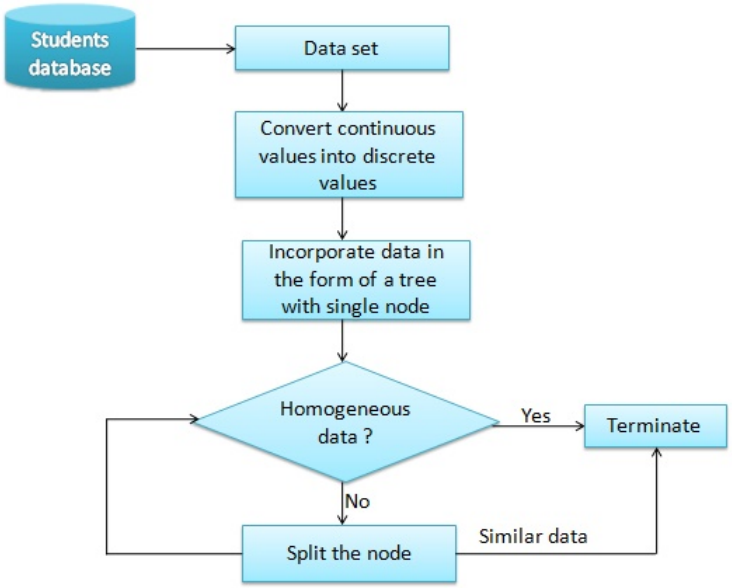

Fig. 6. Dataset conversion into decision tree

Regression Trees) algorithm uses Gini Index to find the root node. It can handle both continuous and discrete values. It uses cost complexity pruning in order to obtain accurate results Khasanah et al. [2017].

(v) **Support Vector Machines (SVM):** This data mining technique involves plotting of each data set as a point in the form of a graph. Then, the hyper-plane is find out that can differentiate the two classes. The steps involved are:
Fetching of Data: The data containing the information about students is collected from different institutions.
Cleaning of Data: The irrelevant data and the data containing errors is identified and removed.
Filtration and Transformation of Data: The large amount of data available is reduced and transformed in a better interpreted form so that data mining method can be applied to the data.
Support Vector Machine Classification Technique: The values assigned to the data sets are plotted on a plane. Then, a hyperplane is identified which can separate the data points into two different classes. The selected hyperplane must have the largest margin from the data point of the two classes Tian et al. [2012]. This technique involves following steps for doing classification of the data:

1) Firstly, the closest pair of data vectors is found out and these are known as the support vectors. These vectors are find from the training data set.
2) Then, a new point or vector is added in the support vector set and classification of this new point is done on the basis of previous closest pair of points.
3) The above two steps are repeated till all the points in the dataset are pruned.

SVM technique can be used to classify both linearly separable and linearly inseparable data. In case of linearly separable data, a hyperplane can be drawn easily that separates the data into different classes and classification can be done easily in case of linearly separable data. But in case of linearly non-separable data, a non linear line is required to separate the data. So, a kernel function is used in this case to handle the data. The kernel function maps the non-linear data into a high dimensional space so that classification of the data can be done.

There are different kernels used for this process such as linear kernel, polynomial kernel, and radial basis function Xia [2016]. General kernel function can be represented as:

$$K(x, y) = \Phi x^T \Phi y$$

where K is the kernel function x and y are the axis of the graph plotted
Linear Kernel: It is the simplest kernel function and is represented as:

$$K(x, y) = x^T.y + C$$

where c is a constant
Polynomial Kernel: It is not a stationery kernel and works for normalised data. It is represented as:

$$K(x, y) = [(x^T.y) + 1]^d$$

where d is a kernel parameter
Radial Basis Function: It is the important of all the kernels as it involves less difficulty in numerical computation and makes use of hyper parameters. It is represented as:

$$K(x, y) = exp(-||xy||^2/\sigma^2)$$

where σ is a kernel parameter

(vi) **Bayesian Classifier:** This data mining technique is based on probability and uses Bayes theorem which is:

$$P(a|b) = [P(b|a) * P(a)]/P(b)$$

where

$$\begin{aligned} P(a) &= \text{ probability of occurrence of } a \\ P(b) &= \text{ probability of occurrence of } b \\ P(a|b) &= \text{ Probability of a given } b \\ P(b|a) &= \text{ Probability of b given } a \end{aligned}$$

This technique involves following steps:

1. *Collection of Data:* The dataset is collected regarding performance of students in any institution.
2. *Feature Selection:* Cleaning of the data is done in order to extract the relevant data by removing missing data and errors. The dataset needs to be classified so as to obtain the result. The data set is stored in the database so that its classification can be done easily.
3. The Nave Bayes classification algorithm is used to classify the different categories of students performance for example, on the basis of grades of different subjects. Then the students are divided into different groups or classes on the basis of the data available. This classifier is applied to test the students performance in different categories given the sample Z is given, the classifier will predict that Z belongs to the class C having the highest probability, condition applied on Z that

$$Z \text{ belongs to class } C_i \text{ iff, } P(C_i|Z) > P(C_j|Z) \text{ for } 1 \leqslant j \leqslant m \text{ and } j \neq i$$

Thus, it is required to maximize $P(Z|C_i)P(C_i)$, for i = attributes assigned according to marks MAKHTAR et al. [2017].
This classifier is better when categorical attributes are used instead of numerical attributes.

The different techniques described above have different applications in the field of education and is shown in the Table II.

Table II: Data Mining Techniques in Education with Applications

Sr. No.	Techniques	Sub-techniques/ Algorithms	Applications	Merits/Demerits/Accuracy
1	**Neural Network** Adekitan and Salau [2019] Hernández-Blanco et al. [2019] Garro and Vázquez [2015] Rastrollo-Guerrero et al. [2020]	Feed Forward Neural Network	—Prediction of students performance —Recommendation of learning opportunities	**Merits:** —It is a technique that can be used to identify the hidden relationships between the data. —This technique can easily handle incomplete, noisy and the missing data. —It is a flexible technique as it can produce output even when the data available is incomplete. —It can work in parallel form as it has the capability to perform more than one job at a time. —The multi layer perceptron used is able to find out the complex relationships between the dependent variable and the independent variables.
		Convolutional Neural Networks	—Detection of undesirable behaviour of students —Students dropout prediction —Text classification	**Demerits:** —It works as Black box. —The transparency in this technique is poor. —It does not clearly define how the output is obtained using the ANN architecture. —The processing time required is high. It is not a robust technique and is evolving.
		Recurrent Neural Networks	—Prediction of proficiency of students —Prediction of students dropout —Performance Prediction —Improve accuracy of prediction	**Accuracy:** —The prediction accuracy can reach upto 99% depending upon the parameters and the type of ANN used. —The accuracy of multi layer ANN is high.

2	**Fuzzy Logic** Mousa and Maghari [2017] Jahan [2015] Alfere and Maghari [2018]	Clustering	Group of students can be created based on similarity	**Merits:** —It is a better technique to use in case of overlapping in the dataset. —It is an efficient technique. **Demerits:** —There is a difficulty in handling high dimensional datasets. —There is a problem of trapping in local minima. **Accuracy:** —The accuracy is near about 92.6% and may differ on the basis of parameters chosen.
		K-nearest neighbour	—To know the skills of students and their habits —Analysis of students behaviour —Performance analysis of students —Chances of dropout of students —To know the early grades of student and this helps to find slow learners	**Merits:** —In KNN, the linear separation of classes is allowed. —KNN can handle noisy data easily. —Accurate predictions can be made using this technique. **Demerits:** —When KNN is applied on large datase, the processing time required is high. **Accuracy:** —The accuracy is near about 97% and may differ on the basis of parameters chosen.

3	**Genetic Algorithms** Altaher and Barukab [2018] Michalewicz [1996]		—Selecting the best performer —Predicting Students future performance	**Merits:** —Easy to understand —Works well in case of noisy data in the database —Optimal solution can be obtained easily —Handles noisy data very well —Handles the individual population in parallel **Demerits:** —Expensive and consume more time. —Does not find optimal solution always. **Accuracy:** —The accuracy is approximately in between 74% and 83% which may differ on the basis of parameters chosen.

4	**Decision Trees** Adekitan and Salau [2019] Saa [2016] Khasanah et al. [2017] Kaunang and Rotikan [2018] Mousa and Maghari [2017] Al-Barrak and Al-Razgan [2016] Wati et al. [2017]	ID3 algorithm C4.5 algorithm CART algorithm	—Prediction of academic success rate of students —To identify students who need special attention —Helps educational institutions to improve the teaching —The expert area of students can be added by teachers	**Merits:** —Easy to understand. —Very simple and fast technique for doing classification. —Easily interpretable. —The data processing time in this technique is less. —Normalisation of the data is not required while using this technique. —Noisy data can be handled easily. —Domain knowledge is not required for construction of decision tree. **Demerits:** —Needs large amount of data in the database for performing better classification of the data. —It is a data sensitive technique as the data structure i.e. the tree structure can be changed if a small variation is made in the dataset. —Requires large memory. —The space and time complexity of applying this technique is high. **Accuracy:** —The prediction accuracy is in the range between 72.4% to 85.7% which may differ on the basis of the algorithm and the parameters used for doing classification. —ID3 algorithm gives better results.

5	**Support Vector Machine** Khasanah and Harwati [2019] Costa et al. [2017] Xia [2016] Tian et al. [2012]	Linear Kernel Polynomial Kernel Radial Basis Function	Classify students into different classes based on their performance	**Merits:** —Can work even when the data collected is not linearly separable. —Provides highly accurate results. —Can work with unstructured data easily. —Over-fitting risk is less. **Demerits:** —The training and testing of the dataset requires more speed. —Does not work well with large datasets as the training time required is very high. —Large memory is required for doing the classification. —The final model obtained is difficult to understand. —Interpretation is difficult. —The choice to be made between the kernel is very difficult. **Accuracy:** —Prediction accuracy is approximately 98% and the rate of error is low Al-Shehri et al. [2017].

6	**Bayesian Classifier** Adekitan and Salau [2019] Mousa and Maghari [2017] Wati et al. [2017] Alsuwaiket et al. [2020] MAKHTAR et al. [2017]	Naive Bayes Classification algorithm	—Prediction of course outcomes —Success rate of students in the next task —Classification of text documents —Prediction of future trends —Making intelligent decisions in distance education	**Merits:** —Implementation is simple. —The efficiency of doing the computation is good. —The classification rate is high. —Accurate results can be predicted. —Can handle discrete and continuous data. **Demerits:** —Accuracy of the result obtained through this technique decreases if small amount of data is used for training. —Large amount of data is required in the database to obtain good results. **Accuracy:** —The accuracy is between 95% to 98% which may differ on the basis of parameter chosen for classification.

6. COMMON APPLICATIONS OF EDM

There are some common applications that are indirectly involved with all the techniques used for educational data mining Hernández-Blanco et al. [2019]. They are as follows:

(i) Prediction of performance of students: The main objective here is to find out a value that could describe the performance of students.

(ii) Detection of undesirable behaviour of students: The main focus is to find out the undesirable behaviour of any student such as using mobile phones, bad habits such as cheating, talking in between or drop out before completion of the course.

(iii) Making clusters of students: The main point is dividing the students into groups or clusters on the basis of different parameters such as knowledge, performance, behaviour etc.

(iv) Analysis regarding social networking: The graph can be used to represent students and the relationships among different students.

(v) Maintenance of reports: The focus here is to maintain reports of students regarding activities performed by them while completing the course so that it can help teachers and administrators in order to provide them feedback.

(vi) Helping stakeholders: The main purpose is to forecast students attitude and detect unusual behaviour which could help the stakeholders to create alerts in time.

(vii) Plan and schedule teaching of course: The aim is to help teachers to prepare the plan and schedule of the course to be taught to the students in advance.

(viii) Creation of course work: The purpose is to help educators in deciding course materials using the information of students.

7. OBSERVATIONS AND RESULTS

The authors of this paper have done analysis on various techniques used for educational data mining. From the analysis, the authors have observed that Neural Networks can be used for prediction and classification. The different types of neural network are used for different purposes. Feed Forward Neural Network can be used for predicting the performance of students, taking into account their past activities. It also helps in recommending students about the learning opportunities after analysing their performance so that they can perform better in that particular field. Multilayer Neural Network helps to find out the students who show undesirable behaviour in the classrooms like doing side-talking, loss of focus in studying, bad behaviour with fellow students etc. This behaviour affects the learning of students and by identifying such students teachers can pay extra attention towards them and help them to focus in studies. It also helps in text classification and making dropout predictions. It finds the students who are likely to dropout without completing the course. The students take admission to improve their learning and make a growth in their career but due to some reasons they dropout in between which results in a loss for them as well as the educational institution. So, the data mining technique can help to make this prediction so that students can be stopped at the right time before they dropout. Recurrent Neural Network helps in predicting the proficiency level of the learners so that teachers can help the students reach up to that level and improve their learning. It also makes prediction about the dropout and the performance of students in the future.

Fuzzy Logic uses different Clustering and K-Nearest Neighbour techniques that help the educational institutions. Clustering helps to group students based on the similarity in their performance which indirectly helps to identify students who are slow in learning so that special attention can be given to such students which can result in their improvement. K-Nearest Neighbour helps in the prediction of performance and dropout, identify skills and habits of students.

Genetic Algorithms technique helps to predict the future performance of students and also helps in selecting the best performer on the basis of the data obtained related to students regarding their performance in the past.

Decision Trees technique helps to improve the teaching process by identifying students success rate in academics. It also helps to identify the students who need extra efforts in teaching so that they can improve their learning.

Support Vector Machine technique helps to classify students on the basis of their performance in academics and other activities.

Bayesian Classifier helps to make intelligent decisions in the distance education system by making prediction regarding the success rate of students. The students can be provided with special facilities to improve their learning. It also performs classification of text documents and helps in prediction of course outcomes. The course can be improved to enhance the learning of students.

8. CONCLUSION

The application of data mining techniques in the field of education plays a very important role as it helps in the overall improvement of the educational system. It helps not only students but also teachers and the other stakeholders involved in the education system. The predictions made using different data mining techniques discussed in the paper helps in the improvement of teaching-learning process. The teachers can get information about the needs of students so as to improve their future performance. The paper explains about the applications of all the techniques in the field of education and how the prediction made using them help in improving the teaching process.

References

Abu Tair, M. M. and El-Halees, A. M. 2012. Mining educational data to improve students' performance: a case study. *Mining educational data to improve students' performance: a case study 2,* 2.

Adekitan, A. I. and Salau, O. 2019. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon 5,* 2, e01250.

Al-Barrak, M. A. and Al-Razgan, M. 2016. Predicting students final gpa using decision trees: a case study. *International Journal of Information and Education Technology 6,* 7, 528.

Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., Alhiyafi, J., and Olatunji, S. O. 2017. Student performance prediction using support vector machine and k-nearest neighbor. In *2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 1–4.

Al-Twijri, M. I. and Noaman, A. Y. 2015. A new data mining model adopted for higher institutions. *Procedia Computer Science 65*, 836–844.

Alfere, S. S. and Maghari, A. Y. 2018. Prediction of student's performance using modified knn classifiers. *Prediction of Student's Performance Using Modified KNN Classifiers*.

Algarni, A. 2016. Data mining in education. *International Journal of Advanced Computer Science and Applications 7,* 6, 456–461.

Alsuwaiket, M. A., Blasi, A. H., and Altarawneh, K. 2020. Refining student marks based on enrolled modules assessment methods using data mining techniques. *Engineering, Technology & Applied Science Research 10,* 1, 5205–5010.

Altaher, A. and Barukab, O. M. 2018. An intelligent hybrid approach for predicting the academic performance of students using genetic algorithms and neuro fuzzy system. *INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY 18,* 10, 64–70.

Atta-Ur-Rahman, K. S., Aldhafferi, N., and Alqahtani, A. 2018. Educational data mining for enhanced teaching and learning. *Journal of Theoretical and Applied Information Technology 96,* 14, 4417–4427.

Aulck, L., Nambi, D., and West, J. 2019. Using machine learning and genetic algorithms to optimize scholarship allocation for student yield.

Baker, R. S. 2014. Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent systems 29,* 3, 78–82.

Baker, R. S. and Yacef, K. 2009. The state of educational data mining in 2009: A review and future visions. *JEDM— Journal of Educational Data Mining 1,* 1, 3–17.

Burman, I. and Som, S. 2019. Predicting students academic performance using support vector machine. In *2019 Amity International Conference on Artificial Intelligence (AICAI)*. IEEE, 756–759.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., and Rego, J. 2017. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior 73*, 247–256.

Ekubo, E. 2019. Data collection experience on educational data mining in nigeria. *Am J Compt Sci Inform Technol 7,* 2, 37.

Garro, B. A. and Vázquez, R. A. 2015. Designing artificial neural networks using particle swarm optimization algorithms. *Computational intelligence and neuroscience 2015*.

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., and Navarro-Colorado, B. 2019. A systematic review of deep learning approaches to educational data mining. *Complexity 2019*.

Huebner, R. A. 2013. A survey of educational data-mining research. *Research in higher education journal 19*.

Jahan, S. S. 2015. Educational data mining using fuzzy sets to facilitate usability and user experience-an approach to integrate artificial intelligence and human-computer interaction. Ph.D. thesis, Laurentian University of Sudbury.

Kaunang, F. J. and Rotikan, R. 2018. Students' academic performance prediction using data mining. In *2018 Third International Conference on Informatics and Computing (ICIC)*. IEEE, 1–5.

Khasanah, A. and Harwati, H. 2019. Educational data mining techniques approach to predict students performance. *International Journal of Information and Education Technology 9*, 115–118.

Khasanah, A. U. et al. 2017. A comparative study to predict students performance using educational data mining techniques. In *IOP Conference Series: Materials Science and Engineering*. Vol. 215. IOP Publishing, 012036.

MAKHTAR, M., NAWANG, H., and WAN SHAMSUDDIN, S. N. 2017. Analysis on students performance using naïve bayes classifier. *Journal of Theoretical & Applied Information Technology 95,* 16.

Manjarres, A. V., Sandoval, L. G. M., and Suárez, M. S. 2018. Data mining techniques applied in educational environments: Literature review. *Digital Education Review* 33, 235–266.

Michalewicz, Z. 1996. Genetic algorithms+ data structures= evolution programs, 3rd edn.© springer.

Mohamad, S. K. and Tasir, Z. 2013. Educational data mining: A review. *Procedia-Social and Behavioral Sciences 97,* 2013, 320–324.

Moscoso-Zea, O., Saa, P., and Luján-Mora, S. 2019. Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. *Australasian Journal of Engineering Education 24,* 1, 4–13.

Mousa, H. and Maghari, A. Y. 2017. School students' performance predication using data mining classification. *School Students' Performance Predication Using Data Mining Classification 6,* 8.

Oloruntoba, S. and Akinode, J. 2017. International journal of engineering sciences & research technology student academic performance prediction using support vector machine.

Peña, A., Domínguez, R., and Medel, J. d. J. 2009. Educational data mining: a sample of review and study case. *World Journal On Educational Technology 1,* 2, 118–139.

Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., and Durán-Domínguez, A. 2020. Analyzing and predicting students performance by means of machine learning: A review. *Applied Sciences 10,* 3, 1042.

Rogers, F. 2019. Educational fuzzy data-sets and data mining in a linear fuzzy real environment. *Journal of Honai Math 2,* 2, 77–84.

Saa, A. A. 2016. Educational data mining & students performance prediction. *International Journal of Advanced Computer Science and Applications 7,* 5, 212–220.

Siemens, G. and Baker, R. S. d. 2012. Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge*. 252–254.

Silva, C. and Fonseca, J. 2017. Educational data mining: a literature review. In *Europe and MENA Cooperation Advances in Information and Communication Technologies*. Springer, 87–94.

THI, Y. T. and BA, L. T. 2019. Educational data mining for supporting students courses selection. *International Journal of Computer Science and Network Security 19,* 7, 106–110.

Tian, Y., Shi, Y., and Liu, X. 2012. Recent advances on support vector machines research. *Technological and Economic Development of Economy 18,* 1, 5–33.

Toivonen, T., Jormanainen, I., and Tukiainen, M. 2019. Augmented intelligence in educational data mining. *Smart Learning Environments 6,* 1, 10.

Wati, M., Indrawan, W., Widians, J. A., and Puspitasari, N. 2017. Data mining for predicting students' learning result. In *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*. IEEE, 1–4.

Xia, T. 2016. Support vector machine based educational resources classification. *International Journal of Information and Education Technology 6,* 11, 880.

Zain, J. M., Herawan, T., et al. 2014. Data mining for education decision support: A review. *International Journal of Emerging Technologies in Learning 9,* 6.

Satinder Bal Gupta done his Doctorate in Computer Science from Kurukshetra University, Kurukshetra, in Year 2011. He is currently Associate Professor in Department of CSE of Indira Gandhi University, Meerpur, Rewari, Haryana. He has published more 30 papers in various International/National Journals/Seminars/Conferences. He has more than 15 books in his credit. He has research interest in Search engines, Data mining, Adhoc networks etc. He is a life member of ISTE.

Rajkumar Yadav done his Doctorate in Computer Science & Engineering from Maharshi Dayanand University, Rohtak in Year 2011. He is currently Associate Professor in Department of CSE of Indira Gandhi University, Meerpur, Rewari, Haryana. He has published more than 50 papers in various International/ National Journals/Seminars/Conferences. He has completed the Major Research project Granted by UGC, MHRD, Govt of India. He has research interest in Information hiding techniques, water marking; finger printing, Data mining etc. He is a life member of ISTE and Indian Science Congress.

Ms. Shivani done her M. Tech in Computer Science from Vaish College of Engineering, Rohtak, Haryana in the year 2015. She is currently working as a faculty member in Department of Computer Science, Indira Gandhi University, Meerpur, Rewari. She has two papers in her credit in International Journals. She has research interest in data mining, Information hiding techniques, search engines, discrete mathematics etc.