IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

AN VALUABLE APPROACH IN EDUCATIONAL DATA MINING TECHNIQUES TO IMPROVE THE PERFORMANCE IN DIPLOMA ENGINEERING STUDENTS

[1]*Mr.Sugin Lal.G*, [2] *Dr.K.Seetharaman*
[1] *Research Scholar*, [2]*Associate Professor and Head*,
[1]*Department of Computer Science*,
[1]*Bharathiyar University*, Coimbatore, India
[2]*Computer Science and Engineering Wing*,DDE
[2]*Annamalai University*, Annamalai Nagar, India

Abstract: Nowadays diploma engineering students performance occurs very poor mainly due to the unexpected reasons.based on this the performance of the students should be improved is very important to industrial growth of our country. EDM mainly prediction of weak engineering diploma student's performance. Various researches have been done so far for predicting the performance of weak students to make improvements in their performance. But in most research works, only few attributes like grades, results, assignments, gender, internal marks are considered in order to predict the students' performance. Though the teacher's maintain the performance of their students it is not correct in all cases. So a better mining algorithm has to be implemented to successfully identify the behavior of weak students. In most of the works, the attributes identified are irrelevant and are neglected as missing attributes leading to inconsistent results. The next drawback identified with the existing algorithm, is the similarity achieved the last results,also not help to improve the performance of the students. In order to reduce this drawback an efficient artificial neural network (ANN) Algorithm will be proposed in this work. Further this algorithm will be very useful to improve the performance of the diploma engineering students.

Index Terms **- Dataset, Data Reduction, ANN Algorithm.**

I.INTRODUCTION

Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.

Educational Data Mining focuses on developing new tools and algorithms for discovering data patterns. EDM develops methods and applies techniques from statistics, machine learning, and data mining to analyze data collected during teaching and learning. EDM tests learning theories and informs educational practice.

Educational data mining is emerging as a research area with a suite of computational and psychological methods and research approaches for understanding how students learn. New computer-supported interactive learning methods and tools—intelligent tutoring systems, simulations, games—have opened up opportunities to collect and analyze student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn. Data collected from online learning systems can be aggregated over large numbers of students and can contain many variables that data mining algorithms can explore for model building.

Goals of EDM:

1. Predicting students' future learning behavior by creating student models that incorporate such detailed information as students' knowledge, motivation, metacognition , and attitudes
2. Discovering or improving domain models that characterize the content to be learned and optimal instructional sequences;
3. Studying the effects of different kinds of pedagogical support that can be provided by learning software; and
4. Advancing scientific knowledge about learning and learners through building computational models that incorporate models of the student, the domain, and the software's pedagogy.

II. RESEARCH MOTIVATION

Current issues identified in EDM especially diploma engineering student performance occurs mainly due to the huge volume of records in learning databases. Additional issues identified in EDM mainly incorporate identification or prediction of weak diploma engineering student‟s performance. Several researches have been done so far for predicting the performance of weak students to make improvements in their performance. But in most research works, less than 10 attributes are considered in order to predict the students‟ performance. Though the teacher‟s preserve the performance of their students it is not correct in all cases. So a superior mining algorithm has to be implemented to successfully identify the behavior of weak students. So various attribute have to be considered other than the mostly utilized ten attributes, for identifying the student‟s frequent behavior. Thus a discovered pattern data mining algorithm has to be made by which performance prediction can be done. Also succeeding greater precision results for student enactment prediction is also a problem definition.

III.MAJOR CONTRIBUTION

There is huge of amount of data repositories available with education institutions that helps in data discovery and data mining. In the existing EDM techniques are already used for predicting the enactment of the students in educational institutions but performance of those methods in predicting is accurate only in case of small set of dataset.

A student‟s performance is mostly evaluated and predicted base on the marks which is an old method. Off late some other parameters were also used but were very limited in predicting in the performance. This work focused on using higher number of attributes to ensure the accuracy of prediction. In specific, we have explored and implemented innovate and efficient two algorithms, CURE and Neural Network that can be cluster and classify the various attributes to analyses the reasons for the reduced performance of the students. The experimental outcomes illustrate that the proposed calculations enhanced the expectation accuracy.

The central contribution of the thesis is to model a combination of clustering and classification algorithm in case of huge data sets. The objective of this research is data reduction, clustering and classification. The contribution to this research is organized as follows:

1. Data Preprocessing by dimensionality reduction method and data transformation.
2. Cure Algorithm for Clustering.
3. Mini-Batch gradient descent algorithm.

The first approach creates a dataset is created by involving 27 different attributes so that it is completely utilized and other nodes are kept open for any process requesting for the execution on the process. As most of the existing methods have considered only limited attributes that affects the accuracy level, we have introduced additional attributes to increase the accuracy level.

Secondly, a data preprocessing technique will be performed in the available dataset by different stages such as data reduction by dimensionality reduction method and data transformation by discretization. Since this paper aims on identification of reason for students (weak) performance by different attribute, here only the weak students list will be processed for further processing. Thus the best students‟ performance are neglected or reduced by preprocessing and thus making data transformation in the dataset and the algorithm flinches again with finding out the demanding match followed by finding out the closest match.

Thirdly the Cure Algorithm is explored for clustering of the data set. As the conventional Cure algorithm is mostly used in small data sets, we have used the random sampling methods to reduce the data set. Then the data set is partitioned and at a later stage it is labeled.

Finally classification is done with the help of Adaptive Artificial Neural Network (AANN). Rather than the usually utilized Multilayer Perceptron (MLP) in ANN in this work, Minibatch gradient descent optimization algorithm is utilized. This optimization algorithm will be responsible for achieving a best solution and will be best utilized during training of neural network. By this adaptive technique the error function occurred due to use of sigmoid activation function with existing algorithm.

IV. RESEARCH METHOD

DATA

The primary division of data is to focus on unstructured data. Generation of a raw data set incorporating co-related attributes, providing an insight into a student‟s personality and academic performance will be our major outline. Successively, the records in the data set will be grouped into dissimilar clusters. Post clustering, each cluster will be assigned a class label considering the overall student performance in that cluster. At this stage, the raw data set is separated into training and testing data sets. A data model can now be developed as a result of a knowledge algorithm which will be implemented on the training data set. Succeeding, the developed data model will be calculated based on accuracy using the testing data set. Lastly, the data model would be entreated from MATLAB for predicting a student‟s performance (given all the attributes).

Let Relation R = {Attribute1, ..., Attribute k} be a set of attributes, D as a data set, and r as a relation according to R. Feature extraction is a plotting f_e: D→r, which plots each data item d ∈D into a tuple t ∈ r:

t = f(d) = {(Attribute1 = ai), ..., (Attributek = ak)}, ai ∈ Dom(Ai).

Table 1: Summary of Numerical and categorical data types and relations.

Data type	sign	Space	Equivalence
1.Numerical	+	+	+
Discrete	+	+	+
continuous	+	+	+
2.Categorical	+/-	-	+
Nominal	-	-	+
ordinal	+	-	+

A dataset is created by involving 33 different attributes which are selected from 36 questionnaires not considered in other EDM such as:

1	School Medium	12	Purpose to choose	23	Family income
2	Board	13	Travel time	24	Parent's Health Condition
3	Living location	14	Student study hours	25	Daily Test Performance
4	Gender	15	Gaps in study	26	Weekly Test Performance
2	Family size	16	Back logs	27	Monthly Test Performance
6	Parent status	17	Assignment	28	First semester result
7	Mother's edification	18	Lab Practicals	29	Previous semester result
8	Father's edification	19	Attendance	30	Seminar
9	Parent/Mother's job	20	Social media	31	Mini Projects
10	Parent/ Father's job	21	Time spent with friends	32	In-Plant Training
11	Parent/ Guardian's job	22	Grade in High School	33	Sports Activity

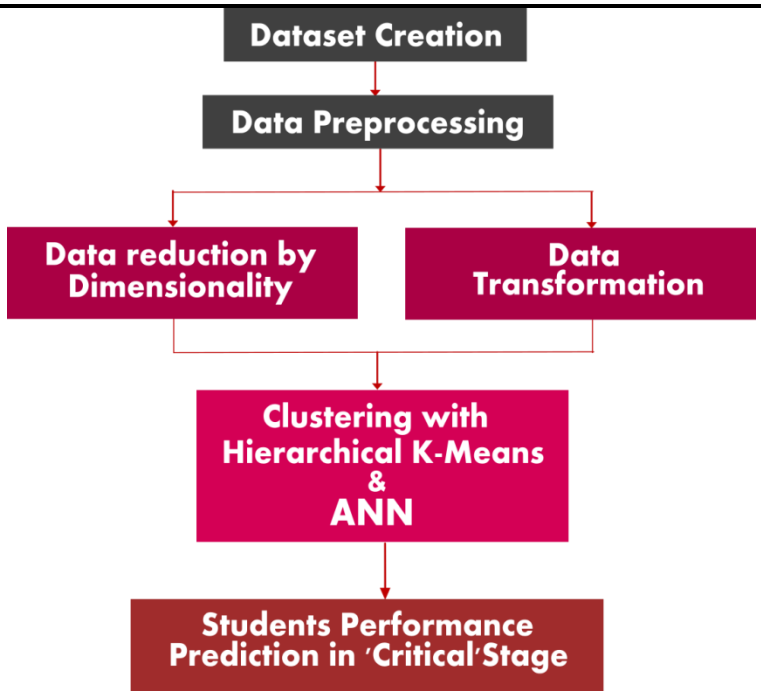

Figure 1: Proposed Research work Process Flow

V. DATA PREPROCESSING

Initially in EDM, a data preprocessing technique will be performed in the available dataset by different stages are data cleaning, data reduction and data transformation. Since this proposed research work aims on identification of reason for students (weak) performance by different attribute. Here, only the weak students list will be processed for further processing. Thus the best students‟ performance are neglected or reduced by preprocessing and thus making data transformation in the dataset.

Dimensionality Reduction

Things that describe the instance/object are known as the dimensions/features. Features vary from one dataset to another. If any datasets contain high dimensions then it increases the computational density of the data analysis algorithms. Hence it reduces quality of the result and some time it misleads to the algorithm. So there is a need of dimensionality reduction techniques. By using dimensionality reduction one can reduce the computation complexity, reduce the storage requirement, and better visualization of the data is possible. Noise and outliers are also being efficiently eliminated using these techniques.

A main challenge of the dimensionality reduction is to identify the important features from the given datasets; so that users are able do good analysis on their data. Dimensionality reduction is to abstract a minor group of structures that improves best of the predictability of the records.

Dimensionality reduction algorithms can be clustered in four different ways, they are:

- Linear or nonlinear
- Feature selection or feature extraction
- Supervised or unsupervised
- Local and global

A large portion of the dimensionality reduction procedures depend on highlight choice or highlight extraction. In the element determination, subset of the first highlights is chosen toward the end; in highlight extraction, new highlights are removed from the first arrangement of highlights utilizing either straight or nonlinear methods. Guideline part examination (PCA) is a case for straight strategy that utilizes direct mapping to remove new highlights from the first highlights. Sammons plotting and ISOMAP utilizes non-direct mapping strategies to digest the highlights from the datasets.

The most important algorithms based on feature extraction are implemented using PCA, Singular Value Decomposition (SVD), Latent Semantic Indexing (LSI), Linear Discriminative Analysis (LDA), and Sammon map. SVD determines most influential features having maximum eigen values for feature extraction, and LDA considers the class information for feature extraction. Feature selection can be implemented using wrappers, filters and embedded methods. Binding methods use a predictive model to score feature subsets.

Data Transformation by Discretization

Regularly data are assumed in the shape of unceasing values. If their range is enormous, model constructing for such information can be tough. Moreover, many facts insertion algorithms function best in discrete pursuit or variable space.

Discretization is commonly done single variable at once, referred to as static variable discretization. Strategies can likewise be call neighborhood or universal. In the prior, presently not all factors are discretized, and inside the finishing up all are discretized. In the succeeding, terms unsupervised and regulated are utilized.

The other experiential is to pick the quantity of periods, mXi, for every variable, Xi, i=1,…,p,

wherein p is the number of variables, as follows :

$$M=\{m_{x1},m_{x2},\ldots\ldots m_{xp}\}$$

There is a discretization sample D on variable X that discretizes the continuous variable X into m discrete intervals, bounded by using the pairs of numbers

$$D: [d_0, d_1],(d_1, d_2],\ldots,(d_{m-1}, d_m]$$

In which d_0 is the minimal and dm is the maximum of variable X, and the values are prepared in rising order.

Pre-processing strategies can have good sized drawbacks. Random sampling can throw out probable beneficial facts, while random sampling increases the dimensions of the dataset and consequently the schooling time. Random-Balance sustains the size of the schooling set and because it's far a method that's repeated several times, the problem of putting off important samples is reduced.

Pseudo code for the Random Balance ensemble method.

Need: Set A of Samples

Make sure: New set N of samples with Random Balance

1: *totalsize* ← |*A*|

2: $A_N \leftarrow \{(x_i, y_i) \in S \mid y_i = -1\}$

3: $A_P \leftarrow \{(x_i, y_i) \in S \mid y_i = +1\}$

4: *Majoritysize*← $|A_N|$

5: *Minoritysize*← $|A_P|$

6: new MajoritySize ←Random integer between new and totalSize-2

7: newMinoritySize ←totalSize-newMajoritysize

8:if newMajoritysize<Majority size then

9: $A \leftarrow S_p$

10: Get a random section of size newMajoritySize from A_N,add the section to N

VI. ARTIFICIAL NEURAL NETWORK ALGORITHM

BASIC CONCEPT OF ANN

An artificial neural network (ANN), normally called as “neural network" (NN), is a calculated model or scientifically model based on biological neural networks, or it is an artificial of biological neural system. It comprises of an interrelated group of artificial neurons and processes data expending a connectionist methodology to calculation.

The Biological Model

Artificial neural networks occurred later the arrival of simplified neurons. These neurons have been existing as models of biological neurons and as theoretical modules for paths that would communicate out computational tasks. The primary version of the neuron is originated upon the capability of an organic neuron. ."Neurons are the simple signaling devices of the worried system and each neuron is a distinct cellular whose numerous processes stand up from its movable structure".

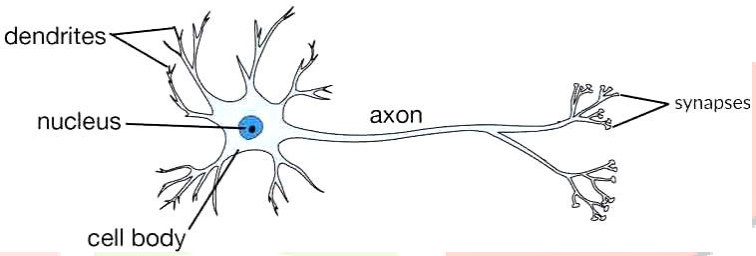

Figure No.2 Biological Neuron

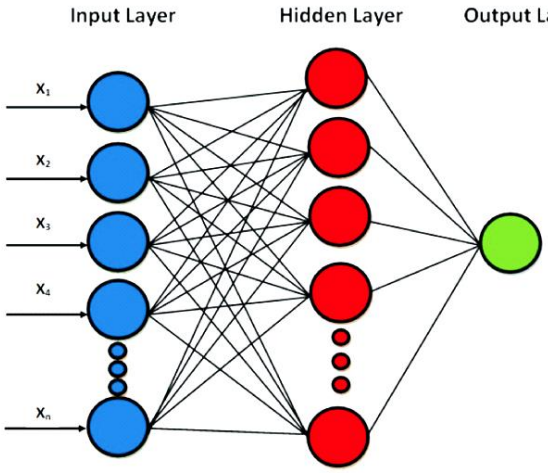

Figure No.3 ANN Network Architecture

Mathematical Model

While making a functional model of the natural neuron, there are three crucial added substances of significance. Fundamental, the neuron transmitters of the neuron are displayed as loads. The intensity of the connecting between information and a neuron is noted by utilizing the charge of the load. Negative weight esteems duplicate inhibitory associations, while positive qualities assign excitatory systems. The following two modules show the genuine action inside the neuron portable. At last, an activation work controls the liberality of the output of the neuron. A proper scope of output is usually among zero and 1, or - 1 and 1.

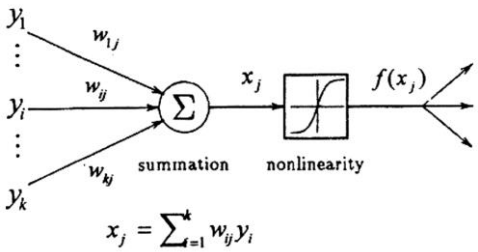

Figure No.4 Neural network architectures for Nonlinearity

Feed-forward neural networks

In a feed forward community, records streams in unique route along connecting pathways from the enter layer via the hidden layers to the very last output layer. There isn't any feedback (loops) i.e., the production of some layer does not have an effect on that same or earlier layer.

Recurrent neural networks

These systems shift from feed forward network structures in the vibe that there is somewhere around single criticism circle. Accordingly, in those systems, for instance, there might need to exist one layer with criticism systems. There can likewise be neurons through self-remarks joins, i.e., the yield of a neuron is feed returned into itself as arrive.

Multi-Layer Perception

A multilayer perception (MLP) is a feed forward neural network that maps sets of input records onto a set of proper output. An MLP includes of multiple layers of nodes in a directed graph, with each layer completely related to the next one. Every node is a neuron with a nonlinear activation function. MLP improves a supervised success to known method called lower back propagation for training the community.

Back Propagation

Back propagation is a method of training artificial neural networks in what way to perform a specified task. The propagation set of procedures is recycled in layered feed forward neural network. In this method that the artificial neurons are arranged in layers, and direct their pointers —forward and then the mistakes are propagated backwards. The back propagation algorithm practices supervised gaining knowledge of, which means that we supply the set of rules with samples of the input data and output data, need the unrestricted to calculate, after which the error is calculated.

Overall summary of the procedure

- Present a train data to the neural network.
- Relate the network result to the chosen output from that sample. Compute the fault in every result.
- For each data, compute what the output it produces, and a scaling highlights, how much lower or higher the output must be recycled to compare the preferred output. This is the local fault (error).
- Modify the loads (weights) of each node to lower the local error.
- Replication the steps above on the neurons at the prior level, using each one's "fault" as its error.

Mini batch Gradient Descent

Instead of implementation of gradient descent on the entire training set, we can separated our training set into smaller sets and implement gradient descent on each batch one after the other. In this way, we can get an perception of gradient descent before finishing entire training set. It makes the algorithm quicker and more efficient. It is called mini batch gradient descent.

ANN ALGORITHM IMPLEMENTATION

Classification will be done with aid of adaptive Artificial Neural Network (AANN) classification algorithm. Rather than the usually used Multilayer Perceptron (MLP) in ANN in this work a Mini-batch gradient descent optimization algorithm is utilized. This descent optimization algorithm will be responsible for achieving a best solution and will be best utilized during training of neural network. By this adaptive technique the error function occurred due to use of sigmoid activation function with existing algorithms, will be reduced since gradient descent computes a sum of squared errors for an entire training dataset.

Algorithm steps

- In the beginning we need to define our network model which will be three layer networks consists of input layer, middle layer and output layer.
- In the input layer given the mean of the each attribute of the students using the testing data.
- The middle layer contains the sum of the mean of the all attributes which obtained by the training data set.
- The output layer gives the desired output which will be compared with the target output obtained by the training data set which gives the error rate.
- In our work we used Mini-Batch gradient descent algorithm which is used to train the weights in an artificial neural network. This one of the method which is used to minimize the sigmoid error function.
- First we need to shuffle the training data then get portions from it. Then we will get the random subset with size of mini batch size of our complete training data. Each of individuals random subset will be fed to middle layer of the networks, and at that moment the gradients of that mini batch will be disseminate back to apprise the constraints/weights of the networks.
- We get the gradients of every single layer of the networks for the current mini batch. Then we use that gradient to modernize the weights of each layer of the networks. The update process is very easy. We just improve the gradient of particular weight matrix to our existing weight matrix.
- Recapitulate every data point in given mini batch, then direct it to the network and match the desired output with the accurate output from the training data. The error is precisely well-defined by the difference of the probability of accurate result with the probability of our calculation.

ADVANTAGES OF ANN

Capability to effort with imperfect knowledge: After ANN training, the data may produce output even with incomplete information. The damage of performance here depends on the importance of the missing information.

Devising error acceptance: Exploitation of one or more cells of ANN does not prevent it from producing output. This feature creates the networks error tolerant.

Devising a scattered memory: In order for ANN to be able to learn, it is necessary to determine the examples and to teach the network according to the preferred output by showing these examples to the network. The network's success is directly proportional to the selected instances, and if the event cannot be shown to the network in all its aspects, the network can produce false output.

DATA SET:

	A	B	C	D	E
1	Mention your Register Number	162586	248563	417585	257681
2	Mention your Branch of Study	DECE	DME	DCE	DCSE
3	Mention your Gender	Male	Female	Male	Female
4	What is your Living location	Hostel	Day scholar from own home	Day scholar from relatives	Day scholar from own home
5	What was your Grade in Tenth board	Outstanding (above	Very Good (above 80% to below	Good (above70% to	Pass (above60% to
6	What was your Grade in senior secondary	Outstanding (above	Very Good (above 80% to below	Good (above70% to	Pass (above60% to
7	In which board you were studied in Tenth Board	State board	CBSE	ICSE	State board
8	Mention your Student family size	Solitary	single parent	both parents	Big family
9	What is your parent status	Married	Widowed	Married	Married
10	What is your Father Qualification	Under Graduate	Post Graduate	Basic	Under Graduate
11	What is your Mother Qualification	Under Graduate	Basic	Basic	Basic
12	What is your Father Profession	Presently on job	Presently on job	Presently on job	Presently on job
13	What is your Mother Profession	Presently on job	N/A	N/A	N/A
14	How many numbers of Friends you have	One	Normal	Medium	High
15	How much Hours you spent with your friends per week	Very limited	Medium	Average	High
16	Do you have an interest to studying in the branch belonging to you	strongly agree	strongly agree	strongly agree	strongly agree
17	What about your Daily test performance	Very good	Pass	Excellent	Good
18	What about your Weekly test performance	Very good	Pass	Excellent	Good
19	What about your Monthly test performance	Very good	Pass	Excellent	Good
20	Are you having an interest to do your Lab practicals perfectly	strongly agree	strongly agree	strongly agree	strongly agree
21	What was your Result in first semester	good	pass	pass	very good
22	How do you could perform in the previous semester	good	Pass	pass	very good
23	Whether are you completed In-plant Training, mention the no	three	two	four	morethan four
24	Are you having knowledge to spent Hours in study after college timings	strongly agree	strongly agree	somewhat agree	strongly agree
25	Whether are you presented Miniprojects, mention the projects no	two	three	one	morethan four
26	Whether are you having any Gap in study	No	No	No	No
27	Whether are you interested in sports activity	Yes	Yes	No	yes
28	Whether your parent having any non-curable disease	No	no	No	No
29	Whether are you having Backlogs, mention the backlogs no	One	Two	One	No backlogs
30	Do you have an interest to access the Social media	Slightly Agree	strongly agree	strongly agree	strongly agree
31	What about your Average family income	high	Medium	Low	Above medium
32	Whether are you having the Sponsor/Guardian belonging to this studying	No	No	No	No
33	What is the motivation to choose this college	Status	Near to Home	Near to home	Status
34	Mention your travel time to college	1 to 2 hours	1 to 2 hours	1 to 2 hours	1 to 2 hours
35	How many hours you are using "Social media" in daily basis	2 to 3hours	2 to 2 hours	2 to 3hours	<1hour
36	Do you have interest to take seminars	Slightly Agree	strongly agree	strongly agree	strongly agree

Sheet1 Sheet2 Sheet3

VII. RESULTS AND CONCLUSION

The application of diverse data mining and machine learning knowledge of strategies inside the domain of training records mining offers which means facts that during flip allows inside the selection making method Though the significance of analysing and predicting the reasons for the bad overall performance of the scholars has been addressed in other disciplines like psychology and sociology, its significance has been much less addressed in the vicinity of tutorial records mining. The major objective of these studies portraits is to discover the motives that make a contribution to the poor performance of college students in educational institutions. As there are numerous techniques which are used for records type, the ANN set of rules is used here. Information‟s like Attendance, Seminar and Assignment and Test marks had been accrued from the scholar‟s preceding database, to are expecting the overall performance at the give up of the semester. The other attributes are amassed through college students like touring hours, own family profits, social media and their respective schools who realize the conduct of college students. This will benefit to the student‟s and the academics to progress the end result of the scholars who are at the threat of failure. This study may even work to pick out the ones students who wanted special interest to lessen failure degree and taking suitable motion for the following semester examination. The AANN proposed reduces the mistake prices and improves the overall performance. Experiments and consequences have proved our claims regarding prediction and accuracy.

REFERENCES

[1] Kaur, Parneet, Manpreet Singh, and Gurpreet Singh Josan. "Classification and prediction based data mining algorithms to predict slow learners in education sector." Procedia Computer Science Vol.57, pp: 500-508, 2015.

[2] Kiu, Ching Chieh. "Supervised Educational Data Mining to Discover Students' Learning Process to Improve Students' Performance." In Redesigning Learning for Greater Social Impact, pp. 249-258. Springer, Singapore, 2017.

[3] Natek, Srečko, and Moti Zwilling. "Student data mining solution–knowledge management system related to higher education institutions." Expert systems with applications Vol.41, no. 14, pp: 6400-6407, 2014.

[4] Lin, Chun Fu, Yu-chu Yeh, Yu Hsin Hung, and Ray I. Chang. "Data mining for providing a personalized learning path in creativity: An application of decision trees." Computers & Education Vol.68, pp: 199-210, 2013

[5] Chen, Xin, Mihaela Vorvoreanu, and Krishna Madhavan. "Mining social media data for understanding students' learning experiences." IEEE Transactions on Learning Technologies Vol.7, no. 3, pp: 246-259, 2014.

[6] Jishan, Syed Tanveer, Raisul Islam Rashu, Naheena Haque, and Rashedur M. Rahman. "Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique." Decision Analytics Vol.2, no. 1, pp: 1, 2015.

[7] Goga, Maria, Shade Kuyoro, and Nicolae Goga. "A recommender for improving the student academic performance." Procedia-Social and Behavioral Sciences Vol.180, pp: 1481-1488, 2015.

[8] Guarín, Camilo Ernesto López, Elizabeth León Guzmán, and Fabio A. González. "A model to predict low academic performance at a specific enrollment using data mining." IEEE Revista Iberoamericana de Tecnologias Del Aprendizaje Vol.10, no. 3, pp: 119-125, 2015.

[9] Asif, Raheela, Agathe Merceron, Syed Abbas Ali, and Najmi Ghani Haider. "Analyzing undergraduate students' performance using educational data mining." Computers & Education (2017).

[10] Márquez-Vera, Carlos, Alberto Cano, Cristóbal Romero, and Sebastián Ventura. "Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data." Applied intelligence Vol.38, no. 3, pp: 315-330, 2013