Students Performance Prediction Using Machine Learning Algorithms

E.PADMA[1], TAPALA YASIN[2], MADDIPATI SRIVATSAV[3]

[1]*Assistant Professor, Department of Computer Science and Engineering, SCSVMV, Kanchipuram*

[2]*B.E Graduate (IV year), Department of Computer Science and Engineering, SCSVMV, Kanchipuram*

[3]*B.E Graduate (IV year), Department of Computer Science and Engineering, SCSVMV, Kanchipuram*

Abstract— In numerous associations, data mining ways are used for assaying large quantum of available data's, information's for decision timber process. In educational sector, Data mining is used for wide variety of operation's similar as performance of the scholars like mark, attendance, staff opinion, social media, Extracurricular activities, Ragging and stress. The data mining ways used for relating the performance of the students using Naïve Bayes and KNN algorithms. These two algorithms identify and analyses the performance of the students.

Keywords— Data mining ways, educational sector, Data mining, Data Collection, Prediction.

1. INTRODUCTION

There are numerous tremendous enhancement exploration interests in using data mining in educational sector. This ultramodern arising sector, called educational data mining, concerned with bettered styles that prize knowledge from data come from the educational sector. Data mining is a fashion of sorting which is actually used to prize retired patterns from huge databases. This generalities and styles can be applied in colorful fields like marketing, drug, real estate, client relationship operation, engineering, web mining, etc. Educational data mining is a new arising or advanced fashion of data mining that can be applied on the data related to the field of education. The data can be collected from history used data and functional data live in the databases of educational institutes. The data of scholars can be particular information or academic performance. Likewise it can be achieve from learning database systems which have a huge quantum of data and information used by utmost institutes.

It uses numerous ways for proper perpetration of data mining generalities similar as Naïve Bayes and K-Nearest neighbour. Using these ways different kinds of knowledge can be discovered using association rules, bracket and clustering. By using this we prize knowledge that describes scholars performance in examination and all their detail information. From These huge quantities of data, the first task is to sort them out, cluster analysis is to classify the raw data in a reasonable way. Clustering is a bunch of physical or abstract objects, as per the degree of similarity between them, divided into several groups, and makes the same data objects within a group of high similarity and different groups of data objects which aren't analogous. As we know in moment's terrain, there's a lack of quality education, and also the competition is adding day by day. So there's a need for quality way to ameliorate the standard of the scholars and education also. For this several proponents give time to time suggestions and norms for performance enhancement. Still, the systems are lacking before. So experimenters had come to a conclusion that the technology can be an important factor for analysing the excrescencies that are present in the moment's system, and why we warrant before. And also the use of technology makes decision-making process easy, as it can induce reports and graphs for analysis purpose. Education could be an important issue for achieving fiscal progress. The Students scholastic performance focuses on different aspects, creating analysis little bit delicate. In forthcoming times, there has been a rise within the chance .in rate of interest and concern over individualists within the use of data mining for assaying academic rates. Data processing depicts growing and forthcoming areas of inquiries in education and it has separate separate requirements that some fields Warrant. During this design, the performance analyses of scholar's are mentioned. The thing at of this design is furnishing scholars' performance using given strategies through different algorithms. A lot of studies in this field are that probe the ways for applying ways associated with machine literacy in educational fields. It focuses on relating high- threat scholars and also pupil performance Ultramodern literacy institutions operate in a largely competitive and complex environment. Therefore, assaying performance, furnishing high-quality education, formulating strategies for assessing the scholars' performance, and relating unborn requirements are some challenges faced by utmost universities moment.

1.1 LITERATURE SURVEY

The study conducted by Kotsiantis et al (1) is one of the original studies which delved operation of machine literacy ways in distance literacy for powerhouse vaticination. The most significant donation by this study was that it was a colonist and sculpted the path for several similar studies. While machine literacy algorithms had been preliminarily enforced in several settings, this was maybe the first time that these ways were applied to an academic terrain.

Bhardwaj and Pal (2) conducted a study in India, Faizabad to determine factors that most heavily affected pupil performance. They used Bayesian Bracket for their study. The study by Erkan Er (3) was grounded upon Kotsiantis'as well as other analogous studies. It concluded that Naive Bayes indeed performed better than any other machine learning algorithm. Still, the pivotal donation of this study was that time-steady features may be mischievous to the machine literacy process, and hence are more left out of the study entirely. He also concluded that" Rather of demographic characteristics of scholars, using original attendance and schoolwork grades produces better vaticination rate at earlier stages."Bhardwaj and Pal (2) conducted a study on the pupil performance grounded by opting 300 scholars from 5 different degree council conducting BCA (Bachelorette of Computer Operation) course of Dr.R.M.L. Awadh University, Faizabad, India. By means of Bayesian bracket system on 17 attributes, it was plant that the factors like scholars' grade in elderly secondary test, living position, medium of tutoring, mama's qualification, scholars other habit, family periodic income and pupil's family status were largely identified with the pupil academic performance.

In the present study, those variables whose probability values were lesser than0.70 were given due considerations and the largely impacting variables with high probability values have been shown in Table 1. These features were used for vaticination model construction. For both variable selection and vaticination model construction, the publishers have used MATLAB. From the table, it's plant that the alternate high implicit variable for scholars' performance is their living position, and the third high implicit variable for scholars' performance is medium of tutoring. In Uttar Pradesh the mama lingo language of scholars is Hindi. Hence, scholars tend to be more comfortable in Hindi and other languages, than in the English language.

The study conducted by Erkan Er (3) proved precious in attesting the oneness of the proposed operation. His work concluded that all current operations of machine literacy in an academic setting were to prognosticate powerhouse rates in a distance literacy program. There's maybe no operation that attempts to prognosticate the absolute performance of the student. However, it has not been published yet, If one does live. Kotsiantis et al (1) compared five algorithms, viz. Decision Trees (C4.5), Naive Bayes algorithm (Bayesian networks), 3-NN (kNN), RIPPER (Rule Literacy) and WINNOW (Perceptron grounded neural networks). This study was composed of two experimental stages, training and testing. During these stages, number of attributes was increased step-by- step. For illustration, while only demographic data was included in the first step, performance attributes were added in the coming step. Five algorithms were tested for each these posterior way and also they were compared. This relative study helped in narrowing down campaigners for our own operation. Data Mining can be used in educational field to enhance our understanding of literacy process to concentrate on relating, rooting and assessing variables related to the literacy process of scholars as described by Alaa el-Halees. Mining in educational terrain is called Educational Data Mining. Han and Kamber describes data mining software that allow the druggies to dissect data from different confines, classify it and epitomize the connections which are linked during the mining process.

Pandey and Pal conducted study on the pupil performance grounded by opting 600 scholars from different sodalities of Dr.R.M.L. Awadh University, Faizabad, India. By means of Bayes Bracket on order, language and background qualification, it was plant that whether new adventurer scholars will performer or not.

Hijazi and Naqvi conducted as study on the pupil performance by opting a sample of 300 scholars from a group of sodalities combined to Punjab university of Pakistan. The thesis that was stated as "Student's station towards attendance in class, hours spent in study on diurnal base after council, scholars' family income, scholars' mama's age and mama's education are significantly related with pupil performance" was framed. By means of simple direct retrogression analysis, it was plant that the factors like mama’s education and pupil’s family income were largely identified with the pupil academic performance.

Khan conducted a performance study on 400 scholars comprising 200 boys and 200 girls named from the elderly secondary academy of Aligarh Muslim University, Aligarh, India with a main ideal to establish the prognostic value of different measures of cognition, personality and demographic variables for success at advanced secondary position in wisdom sluice. The selection was grounded on cluster slice fashion in which the entire population of interest was divided into groups, or clusters, and a arbitrary sample of these clusters was named for farther analyses. It was plant that girls with high socio-profitable status had fairly advanced academic achievement in wisdom sluice and boys with low socioeconomic status had fairly advanced academic achievement in general.

Galit gave a case study that use scholars data to dissect their literacy geste to prognosticate the results and to advise scholars at threat before their final examinations.

A Review on Data Mining ways and factors used in Educational Data Mining to prognosticate pupil amelioration.

Educational Data Mining (EDM) is an interdisciplinary ingenuous exploration area that handles the development of styles to explore data arising in a educational fields. Computational approaches used by EDM is to examine educational data in order to study educational questions. As a result, it provides natural knowledge of tutoring and literacy process for effective education planning. This paper conducts a comprehensive study on the recent and applicable studies put through in this field to date. The study focuses on styles of analysing educational data to develop models for perfecting academic performances and perfecting institutional effectiveness. This paper accumulates and relegates literature, identifies consequential work and mediates it to calculating preceptors and professional bodies. We identify exploration that gives well- fortified advise to amend edifying and amp the further impuissant member scholars in the institution. The results of these studies give sapience into ways for upgrading pedagogical process, prognosticating pupil performance, compare the perfection of data mining algorithms, and demonstrate the maturity of open source tools.

Data Mining Approach For Predicting Student Performance.

This work proposes a new approach- substantiated soothsaying-to take into account the successional effect in prognosticating pupil performance (PSP). Rather of using all literal data as other styles in PSP, the proposed styles only use the information of the individual scholars for vaticinating his/ her own performance. Also, these styles also render the" pupil effect" (e.g. how good/ clever a pupil is, in performing the tasks) and" task effect" (e.g. how delicate/ easy the task is) into the models. Experimental results show that the proposed styles perform nicely and much faster than the other state-of-the- art styles in PSP.

A new approach for upgrading Indian education by using data mining ways.

Education is the backbone of all developing countries. Elevation of the education system, upgrades the country to the world top ranking position. One of the major problems that the education system facing is prognosticating the geste of scholars from large database. This paper focus on upgrading Indian education system by using one of the ways in Data booby-trapping videlicet clustering. Cluster analysis solves the given data into some meaningful groups. Typically the performances of the scholars can be classified into different patterns as normal, average and below normal. In this paper we essay to dissect pupil's data in different angle beyond the below indicated patterns through recently proposed UCAM (Unique clustering with Affinity Measures) clustering algorithm.

2. Project Description

2.1 Problem Statement

The main end of this design is to extemporize the pupil performance in studies grounded on some important factors. Education is an essential element for the betterment and progress of a country. It enables the people of a country cultivated and well mannered. Now-a-days developing new styles to discover knowledge from educational database in order to assay pupil's trends and behaviours towards education. To assay the data from different confines classify it and to epitomize the connections. It motivated us to work on pupil dataset analysation. The data collection, categorization and bracket is being performed manually.

2.2 Existing System

As of now, Being system take only performance into consideration which isn't sufficient for having system, which can help us to estimate performance of a pupil. We aren't having a system which would help us to integrate the performance and undesirable into consideration. Disadvantages of Existing System ,Being system miss the undesirable data for the scholars And It may not check the social data for the pupil.

2.3 Proposed system

The work aims to develop a trust model using data mining ways, which mines needed information, so that the present education system may borrow this as a strategic operation tool. The proposed system use educational data mining ways to estimate performance and identify undesirable geste. In educational sector, Data mining is used for wide variety of operation’s similar as performance of the scholars like mark, attendance, staff opinion, adulterous conditioning, Ragging and stress. The data mining ways used for relating the performance of the pupil using K- means and KNN algorithms. Advantages of Proposed System, Educational database contain the useful information for Evaluating Students. The data mining techniques are more helpful in classifying educational database and help us in evaluating the performance and undesirable behavior of a student.

2.4 Module Description

The following modules are used in this project:

Data Collection:

In this module, Student data's will be collected from the council. Student's data like mark, attendance, staff opinion, Social media, extracurricular activities, Ragging and stress.

Preprocessing:

Data pre-processing is done to remove the deficient noisy and inconsistent data. Data must be pre-processed before using in point selection task.

Classification Module:

The data mining ways used for relating the performance of the pupil using Naïve Bayes and KNN algorithms. These two algorithm's identifies and analyses the performance of the pupil.

Prediction:

In this module, to prognosticate the pupil performance grounded upon pupil mark, attendance, staff opinion, Social Media, extracurricular activities, Ragging and stress.

3. Project Design and Analysis

3.1 Architecture

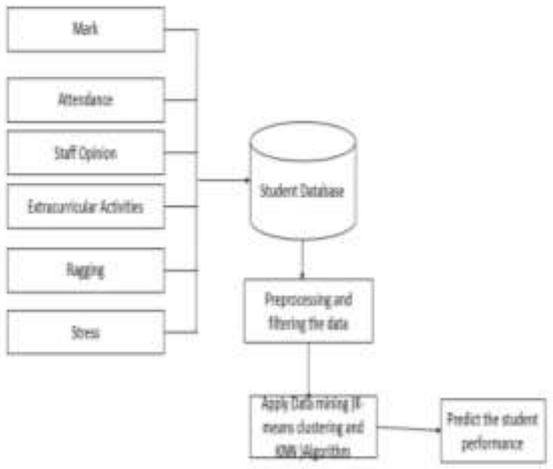
Mark
Attendance
Staff Opinion
Extracurricular Activities
Ragging
Stress
Student Database
Preprocessing and filtering the data
Apply Data mining (K-means clustering and KNN Algorithm
Predict the student performance

Fig.1 Architecture

3.1 Proposed Algorithms

K-NEAREST NEIGHBOR(KNN) CLASSIFICATION METHOD

K-NN is a type of case- grounded literacy, or lazy literacy, where the function is only approached locally and all calculation is remitted until bracket. The k-NN algorithm is among the simplest of all machine learning algorithms. The neighbors are taken from a set of objects for which the class (for k-NN bracket) or the object property value (for k-NN retrogression) is known.

```
STEP 1 BEGIN
STEP 2 Input D = (x1, c1),..., (xN, cN)
STEP 3 x = (x1... xn) new case to be classified
STEP 4 FOR each labelled case (xi, ci) calculate d (xi, x)
STEP 5 Order d (xi, x) from smallest to loftiest, (i = 1... N)
STEP 6 select the K nearest cases to x Dkx
STEP 7 Assign to x the most frequent class in Dkx
STEP 8 END
```

NAIVE BAYES ALGORITHM

It's a bracket fashion grounded on Bayes 'Theorem with an supposition of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular point in a class is unconnected to the presence of any other point. For illustration, a fruit may be considered to be an apple if it's red, round, and about 3 elevation in periphery. Indeed if these features depend on each other or upon the actuality of the other features, all of these parcels singly contribute to the probability that this fruit is an apple and that's why it's known as' Naive'.

Let's understand it using an illustration. Below I've a training data set of rainfall and corresponding target variable' Play' (suggesting possibilities of playing). Now, we need to classify whether players will play or not grounded on rainfall condition. Let's follow the below way to perform it.

Step 1 Convert the data set into a frequence table.

Step 2 Produce Liability table by chancing the chances like Heavy probability = 0.29 and probability of playing is0.64.

Step 3 Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the loftiest posterior probability is the out growth of vaticination.

3. UML Diagrams

UML is simply anther graphical representation of a common semantic model. UML provides a comprehensive memorandum for the full lifecycle of object- acquainted development. Advantages,To represent complete systems (rather of only the software portion) using object acquainted generalities .To establish an unequivocal coupling between generalities and executable law. To take into account the scaling factors that are essential to complex and critical systems. To creating a modeling language usable by both humans and machines.

UML defines several models for representing systems. The class model captures the stationary structure. The state model expresses the dynamic geste of objects. The use case model describes the conditions of the stoner. The commerce model represents the scripts and dispatches flows. The perpetration model shows the work units.The deployment model provides details that pertain to reuse allocation.

3.1 Usecase Diagram

Use case diagrams overview the operation demand for system. They're useful for donations to operation and/ or design stakeholders, but for factual development you'll find that use cases give significantly further value because they describe “ the meant” of the factual conditions. A use case describes a sequence of action that provides commodity of measurable value to an action and is drawn as a vertical cirque.

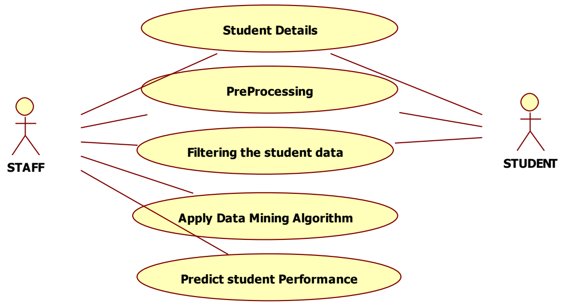
Student Details
PreProcessing
Filtering the student data
Apply Data Mining Algorithm
Predict student Performance
STAFF
STUDENT

Fig. 2 Usecase Diagram

3.2 Sequence Diagram

Sequence Illustration model the inflow of sense within your system in a visual manner, enabling you both to validate and validate your sense, and generally used for both analysis and design purpose. Sequence illustration are the most popular UML artifact for dynamic modeling, which focuses on relating the geste within your system.

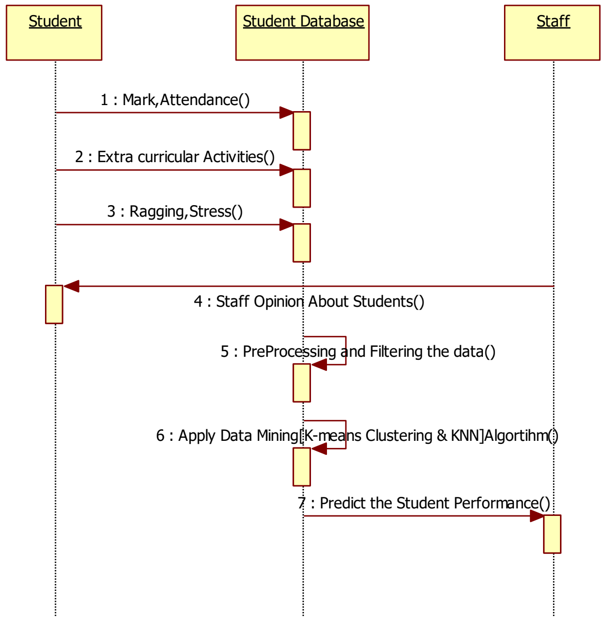
Student
Student Database
Staff
1 : Mark,Attendance()
2 : Extra curricular Activities()
3 : Ragging,Stress()
4 : Staff Opinion About Students()
5 : PreProcessing and Filtering the data()
6 : Apply Data Mining[K-means Clustering & KNN]Algortihm()
7 : Predict the Student Performance()

Fig. 3 Sequence Diagram

3.3 Collaboration Diagram

Another type of commerce illustration is the collaboration illustration. A collaboration illustration represents a collaboration, which is a set of objects related in a particular environment, and commerce, which is a set of dispatches exchange among the objects within the collaboration to achieve a asked outgrowth.

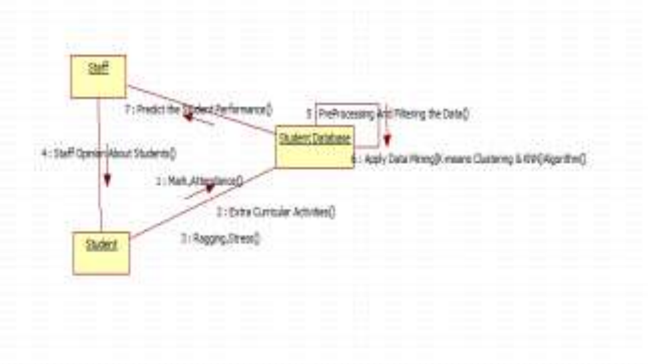
Staff
Student Database
Student
7 : Predict the Student Performance()
5 : PreProcessing And Filtering the Data()
4 : Staff Opinion About Students()
6 : Apply Data Mining(K-means Clustering & KNN)Algorithm()
1 : Mark,Attendance()
2 : Extra Curricular Activities()
3 : Ragging,Stress()

Fig. 4 Collaboration Diagram

3.4 Activity Diagram

Exertion illustration are graphical representations of workflows of accretive conditioning and conduct with support for choice, replication and concurrency. The exertion plates can be used to describe the business and functional step-by- step workflows of factors in a system. Exertion illustration correspond of Original knot, exertion final knot and conditioning in between.

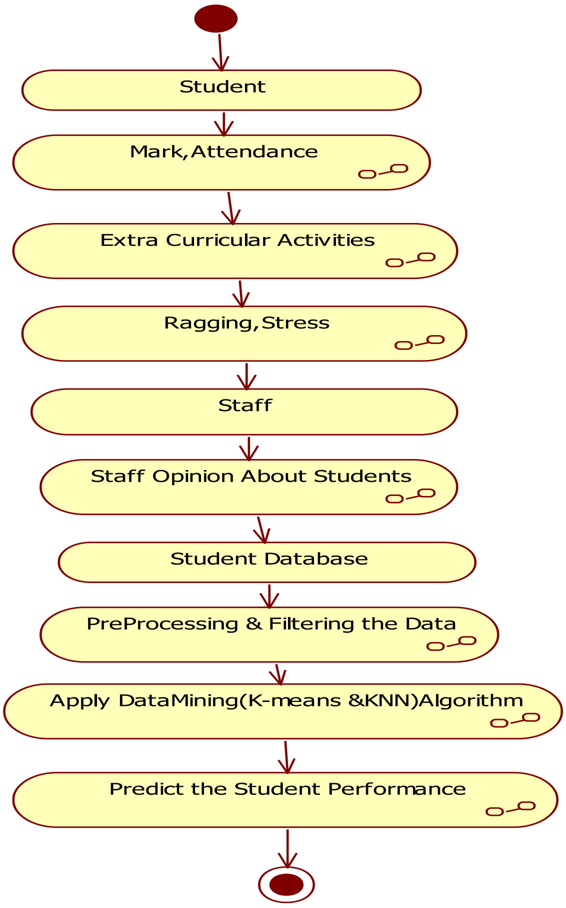
Student
Mark,Attendance
Extra Curricular Activities
Ragging,Stress
Staff
Staff Opinion About Students
Student Database
PreProcessing & Filtering the Data
Apply DataMining(K-means &KNN)Algorithm
Predict the Student Performance

Fig. 5 Activity Diagram

4. System Design

4.1 Input Design

The input design is the link between the information system and the stoner. It comprises the developing specification and procedures for data medication and those way are necessary to put sale data in to a usable form for processing can be achieved by examining the computer to redate from a written or published document or it can do by having people conciliating the data directly into the system. The design of input focuses on controlling the quantum of input needed, controlling the crimes, avoiding detention, avoiding redundant way and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the sequestration. Input Design considered the following effects ,What data should be given as input? How the data should be arranged or enciphered? The dialog to guide the operating help in furnishing input. Styles for preparing input attestations and way to follow when error do.

4.2 output Design

A quality affair is one, which meets the conditions of the end stoner and presents the information easily. In any system results of processing are communicated to the druggies and to other system through labors. In affair design it's determined how the information is to be displaced for immediate need and also the hard dupe affair. It's the most important and direct source information to the stoner. Effective and intelligent affair design improves the system's relationship to help stoner decision- timber. The affair form of an information system should negotiate one or further of the following objects. Convey information about once conditioning, current status or protrusions of the Future. Signal important events, openings, problems, or warnings. Detector an action. Confirm an action.

5. Conclusion

In this paper, the bracket task is used on students database to prognosticate the students division on the base of former database. As there are numerous approaches that are used for data bracket, the Naïve Bayesian Classifier and Weighted Naïve Bayesian Classifier are used then. Information's like Attendance, Class test, Forum and Assignment marks were collected from the pupil's former database, to prognosticate the performance at the end of the semester. This study will help to the scholars and the preceptors to ameliorate the division of the pupil. This study will also work to identify those scholars which demanded special attention to reduce fail portion and taking applicable action for the coming semester examination. This can help the scholars ameliorate in their academics, which ultimately leads to a good performance in their end examinations. By this the self-murder rates of scholars will also get reduced since the stress is reduced. This could help in our country development by furnishing good and effective masterminds to the country.

6. Future Enhancement

This study can be developed in numerous ways, and it's possible to perform unborn work in the following directions. New ensemble and mongrel classifiers can be introduced for having a better comparison and also achieving advanced performance. Also, point selection styles as a way of perfecting models results can be performed to get a better perspective on the significant features.

References

[1] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning systems using machine learning techniques," AI Techniques in Web-Based Educational Systems at Seventh International Conference on Knowledge-Based

Intelligent Information & Engineering Systems, pp. 3-5, September 2003.

[2] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[3] Erkan Er. "Identifying At-Risk Students Using Machine Learning Techniques", International Journal of Machine Learning and Computing, Vol. 2, No. 4, pp. August 2012.

[4] S. Kotsiantis, I.D. Zaharakis, and P. Pintelas, "Assessing Supervised Machine Learning Techniques for Predicting Student Learning Preferences"

[5] M.Durairaj, C.Vijitha .Educational Data mining for Prediction of Student performance Using Clustering Algorithms. The data mining techniques are more helpful in classifying educational database Which contain the useful information for predicting a student's performance.

[6] Kin Fun Li, David Rusk and Fred Song.Predicting Student Academic Performance,The performance predictors, if identified, can then be used effectively to formulate corrective action plans to improve the attrition rate.

[7] Achumba, I. E. and Azzi, D. and Dunn, V. L. and Chukwudebe, G. A. ”Intelligent Performance Assessment of Students’ Laboratory Work in a Virtual Electronic Laboratory Environment.” IEEE Transactions on Learning Technologies, vol. 6, pp. 103-116, Apr 2013.

[8] Chen, Hsuan-Hung and Chen, Yau-Jane and Chen, Kim-Joan. ”The Design and Effect of a Scaffolded Concept Mapping Strategy on Learning Performance in an Undergraduate Database Course”
IEEE Transactions on Education, vol. 56, pp. 300-307, Aug 2013.

[9] Doctor, Faiyaz and Iqbal, Rahat. ”An intelligent framework for moni- toring student performance using fuzzy rule-based Linguistic Summari sation.” 2012 IEEE International Conference on Fuzzy Systems, pp. 1-8, Jun 2012.

[10] Barney, Sebastian and Khurum. ”Improving Students With Rubric-Based Self-Assessment and Oral Feedback”, IEEE Transactions on Education, vol. 55, pp. 319-325, Aug 2012.

[11]Barney, Sebastian and Khurum. ”Improving Students With Rubric Based Self-Assessment and Oral Feedback”, IEEE Transactions on Education, vol. 55, pp. 319-325, Aug 2012.