International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2584
Volume 3, Issue 3, March-2017

Educational Data Mining on students' academic performance using Clustering technique

Selvakumar G [1], Sujasri V [2], Sweetha K B [3]

[1]*Associate Professor, CSE & Sri Shakthi Institute of Engineering and Technology*,
[2]*Student, CSE & Sri Shakthi Institute of Engineering and Technology*,
[3]*Student, CSE & Sri Shakthi Institute of Engineering and Technology*.

Abstract— The student database needs to be analysed and the useful information is to be extracted by the educational institutions to focus on improvement in the student records. The performance indicators of students are to be observed as they may count on varied factors including their learning methodology, environment they are associated with, their place of stay and so on. Data Mining comes up with its techniques and applications to satisfy the needs. Educational Data Mining, an emerging application of data mining used principally on student data is applied here to discover knowledge from the largely available data. Clustering of students is done using K-means based on their scores where students of the same cluster show homogenous characteristics with one another and the interesting hidden patterns among the students are mined. Here, R tool provides an extensive platform to accomplish the mining process. The reports are generated in the form of visualization through graphs

Keywords— Data Mining, Educational Data Mining, Knowledge Discovery Databases, Clustering, K-means

1. INTRODUCTION

Student's performance is one of the crucial parts in learning institutions because it becomes their main criteria to produce excellent record of academic achievement Academic performance of a student is concerned with various factors like personal, socio-economic, psychological and other environmental variables and there must be a performance indicator for the achievements throughout their study as students have different levels of knowledge, different understandings of teaching and learning, and different environmental conditions varying from each other. Currently, there are many techniques being proposed mainly to evaluate student's performance. Due to enormous availability of huge amount of data, data mining approach is carried out to meet the requirement. Data mining, also called Knowledge Discovery in Databases (KDD), which is the field of discovering potentially useful information from large databases. Educational Data Mining (EDM) is the application of Data Mining techniques on educational data. To carry out the mining process different algorithms have been proposed. Clustering (an unsupervised classification) is one such algorithm implemented to segment data into different clusters of similar data objects with some relation with one another. It identifies the factors associated with students that affect their academic performance so that the students, staffs and the management could provide extra care accordingly, to improve their records.

2. BACKGROUND OR RELATED WORK

2.1 EDUCATIONAL DATA MINING & STUDENTS' PERFORMANCE PREDICTION

Researches on the educational data mining field focuses on gathering knowledge, and the deliverables help students to manage their curriculum better and to help the educational institutions manage their students better. In this research, the students are classified based on the information analysed which is then used for creating the visuals of them with algorithms for enhance decisions. The need for classification is for extracting effective and similar data and categorizes them in groups. It also identifies the students performing low in the beginning of the course and so that some additional attention can be paid to those students for helping them to get better marks. The analysis is being beneficial because it not only focuses on the slow performers but also the well performers where they can concentrate more on their projects which will help them to gain more knowledge technically and getting help from their educators. Multiple classification methods are being used to represent the results in the form of multiple classifiers. The use of data mining methodologies is being used to represent the data collected from various sources where the exploration and visualisation of the data is represented. While exploring the data, it allows understanding the differences that can be seen before and after applying data mining algorithms for the huge data in interpreting results. Here the student's grade at the end of the semester is generated as a final outcome by using the method of multiple classifications [5]. This shows a comparative study for the overall performance of students. An algorithm called 10-fold cross validation which provides accuracy measures for verifying and validating the outcomes. The tools like Rapid Miner and WEKA was used to implement the data mining techniques. The data collected contained personal, social and academic related of the students. They were then made to undergo pre-processing steps that are suitable for data mining to be carried out. Predictive data mining models were created which were able to predict the grades of the students from the data collected. Interesting patterns were found using the Naive Bayes' model. Four decision trees algorithms have also been used. It was also found that the students' performance were dependent on other factors apart from their academic efforts on a greater level. This mining can be

used on a periodically to find the patterns to help students and their institutions in many ways. The mining can be done applying different algorithms as per the needs. Clustering can also be one among the interesting algorithms.

2.2 ANALYSING EDUCATIONAL DATA THROUGH EDM PROCESS: A SURVEY

Educational Data Mining involves the extraction of knowledge in the field of education involving activities of the institution whether it is academic, cultural, examination and training or placement. On filtering the data after several steps, it gives out the hidden patterns with the significant relationship between the variables of the filtered dataset. Classification methods like rule mining, decision tree, association rule with other techniques, Bayesian network etc are uses to predict the student performance. Among them, decision tree induction, association, logistic regression, Naive Bayes' are preferred [6]. EDM is usually proceeded through four phases. First, the relationships between the variables are found out and then in the second phase, their validations are checked. They are taken to the next phases of prediction to make guidelines for the institute. The data filtering is done based on the analyst.

2.3 A SURVEY ON EDUCATIONAL DATA MINING IN FIELD OF EDUCATION

The analysis on educational data not only aims to provide results for predicting students' performance or for the betterment of the institution but also for analysing huge data which in turn increases the computation power and use of computer based learning environment where the ability to process the data is shown efficiently. This is identified by translating them into a series of educational data mining. There are different educational data mining methods by which the concept of EDM is being applied whereas all these methodologies lie upon any one specific concepts relating to data mining, which can be:

Clustering: In the process of clustering, clusters are obtained by segmenting the data into groups based on their category. If the data is already classified then the method of clustering happens to be easier for processing. Here the each data is represented by data points where data point in one cluster should be more similar to other data points of the same cluster and if shows dissimilarity then that data point is represented in another cluster, in such a way various clusters are determined. There are algorithms for performing clustering techniques which can be implemented either by initiating clustering algorithms with a prior assumption or clustering algorithms with no prior assumption

Prediction: Predicting is a concept in which the results are predicted using the variable factors that are likely to influence the behaviours. It further categorized into classification, regression and Density estimation. Classification deals with binary or categorical values. Regression takes continuous input variables for predicting the results. Density estimation uses kernel functions in processing the data.

Relationship Mining: The most closely associated variables are being discover in this case. The various types of relationship mining are sequential pattern mining, casual data mining, association rule mining and correlation mining. In association rule mining, the if-then rule is applied to see if some set of variables appear to have specific values. The linear correlation mining of variables is found in correlation mining where in sequential pattern, temporal relationship is identified.

Distillation of Data for Human Judgment: Here classification and identification was the main purpose of distillation of data for easy recognize of well-defined patterns.

From the above factors, prediction technique is the most used method of data mining process[7] and a few application related to education sector are,

Analysis and Visualization of Data: There are many researches based on visualisation of educational details. For clear understanding and analysing of data, graphical methods are used for visualisation where the useful and meaningful information can be extracted for identifying each student's performance during the course and thenceforth for decision making.

Predicting Student Performance: The prediction involves in predicting the unknown values of a variable. To define the student's performance, the factors like in knowing the success and knowledge gained from that course and also the final grades obtained by them in the end of the course. Several techniques like decision trees, neural networks, and rule based systems and Bayesian networks are used for prediction of students' performance and linear regression for predicting the students' marks.

Grouping Students: The students are grouped base on the similar features like personal characteristics and behaviour, performance in the previous course, their marks and achievement. Effective learning groups are formed to improve the learning system. For this clustering algorithms are being used to group the students in different sector. Intelligent e-learning systems are used to see the learning style discriminating features and external profiling features.

2.4 ANALYSIS OF STUDENT RESULT USING CLUSTERING TECHNIQUES

The several issues faced in educational system are predicting the quality of students, their interaction, identifying the students need and personalised training for each individual in the organisation. To overcome these issues, Educational data mining technique is used for improving the learning experience of the students and also it contributes increase the standards of the organisation. The bottleneck of large data analysis is created by manual data mining but it does not generate transitions automatically so the concept of data mining is being used. For analysing the data from various dimensions and identifying relationship among them, data mining software is used which is further used for analysing the existing gaps and works. The students are being grouped into three ways by considering their final grades. Initially assigning labels with possible number of grades then secondly categorizing into three classes as High, Medium and Low and after that lastly into two classes Pass for marks greater than 40 and Fail for marks lesser than 40. The basic technique used for analysing data sets is clustering [8]. This analysis brings out a lot of merits such as decreasing students drop out, increasing improvement ratio in terms of education, increasing student's success, learning outcome, retention rate, transition rate and also maximizing the institution work on their progress efficiently. Data mining provides knowledge

and insights for decision makers in improving the educational system. This study makes use of cluster analysis to segment students into groups according to their characteristics and draw insights using data mining techniques.

3. PROPOSED METHODOLOGY AND DISCUSSION

3.1 PROBLEM STATEMENT

As each set of students graduate every year, the management of the institution and the staffs are responsible for analysing the student database on various aspects to adopt for better methodologies for the next batch of students. Their performance is to be tracked not only at the end of their final semester, but also during their period of study at intervals of each semester for betterment in the later parts of their curriculum and other participations.

3.2 OBJECTIVE

To meet the challenge mentioned above, an analytic platform is to be set up to mine the hugely available raw data. The resultant outcomes are obtained through different perceptions and a visual representation of the mined information is preferred for effective and efficient understanding. The mining process is to be carried out on a regular periodic basis. Based on the interesting patterns acquired, the performance indicator is identified and an appropriate environment to be provided for the betterment of the students and the institutional growth is made accordingly, benefiting all the actors involved in an educational system.

3.3 PROPOSED SYSTEM

The analysis is carried out through the process of KDD comprising of Data Integration, data selection, data cleaning, data mining, pattern Evaluation, knowledge representation and decision making. After the cleaning phase, the clustering technique is applied on the data to mine the patterns among the students of the same cluster. The outcomes of the data are visualized through graphs. The insights observed are then used to make remedial measures and define strategies for the improvement in the performance.

4. SYSTEM IMPLEMENTATION

4.1 STUDENT DATASET

About 250 students' data with their personal details, periodical test marks, university grades, SGPA, their learning methodology preferences (like usage of web-resources or library or class-notes, preferring a group study or self-study) are considered for analysis. The following is a real time sample data set:

	CS6003	Cs6701	CS6702	CS6703	CS6704	CS6711	CS6712	IT6006	SGPA	Attendance	UnivCS6003	UnivCS6701	UnivCS6702	UnivCS6703	UnivCS6704	UnivIT6006	Gender	Stay
1	17	18	18	15	17	18	19	15	7.136364	87.50000	C	C	C	B	E	E	F	H
2	16	18	19	16	18	19	19	15	7.409091	100.00000	C	C	C	B	E	C	F	H
3	15	16	16	11	14	18	19	11	6.125000	71.00000	E	E	U	E	U	E	M	D
4	16	16	16	13	17	18	19	12	7.136364	100.00000	C	C	D	C	E	C	F	D
5	15	17	15	14	17	18	20	14	6.318182	100.00000	D	E	E	E	C	E	F	D
6	16	18	16	13	16	19	18	15	6.842105	91.66667	U	C	E	C	E	D	F	D
7	16	16	15	12	14	18	18	11	6.090909	91.50000	C	E	E	E	E	E	M	D
8	18	18	18	16	17	19	18	15	6.909091	87.50000	C	C	E	B	E	D	F	H
9	16	17	14	13	14	19	19	10	6.110000	95.83333	D	U	E	E	E	E	M	D
10	16	16	16	13	18	18	18	13	6.318182	100.00000	E	D	E	D	D	E	F	D
11	17	18	18	16	19	19	19	15	7.681818	91.66667	C	C	C	B	C	C	F	D
12	16	17	15	13	17	19	19	15	7.136364	91.66667	C	C	E	B	E	C	F	D
13	15	16	15	12	14	18	19	12	6.370000	91.50000	E	U	E	D	D	E	M	D
14	16	17	16	15	15	18	19	13	6.625000	87.50000	E	E	U	C	E	U	F	H
15	17	16	18	18	15	18	18	13	6.590909	100.00000	C	NA	E	B	E	E	F	D
16	15	15	14	11	14	16	18	11	6.500000	83.50000	E	U	E	C	E	U	M	D
17	15	17	15	13	17	18	19	13	6.812500	95.83333	E	U	E	B	E	U	M	D
18	16	17	16	14	14	18	20	12	6.454545	100.00000	C	E	E	C	E	E	F	D
19	16	16	16	16	17	18	18	13	5.954545	91.66667	D	E	E	E	E	E	F	H
20	16	17	17	17	14	19	18	13	6.727273	100.00000	C	D	C	D	E	E	F	D
21	18	19	18	17	18	19	18	16	7.818182	100.00000	D	B	C	C	A	C	F	D
22	17	17	16	15	17	18	19	12	7.136364	100.00000	C	B	E	C	E	C	F	H
23	16	16	14	13	14	19	19	13	6.526316	96.00000	C	U	E	E	E	D	F	D
24	17	16	17	16	15	17	19	14	6.181818	100.00000	E	E	E	C	E	E	M	D
25	17	16	18	16	17	18	19	15	7.272727	100.00000	B	C	E	C	C	D	F	D
26	16	15	16	13	15	16	18	12	5.818182	87.50000	E	E	E	E	E	E	M	D

Fig I : A sample real time data set after the data cleaning process

4.2 CLUSTERING OF STUDENTS

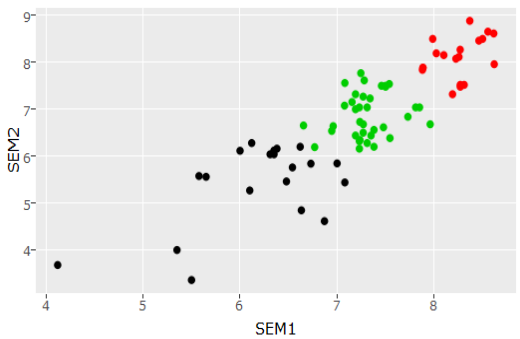

Fig II: Semester 1 SGPA vs. Semester 2 SGPA

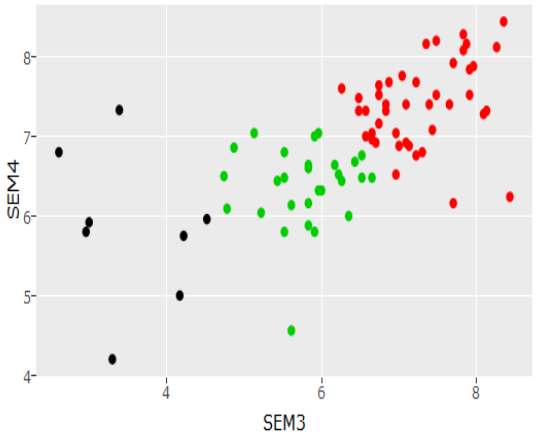

Fig III: Semester 3 SGPA vs Semester 4 SGPA

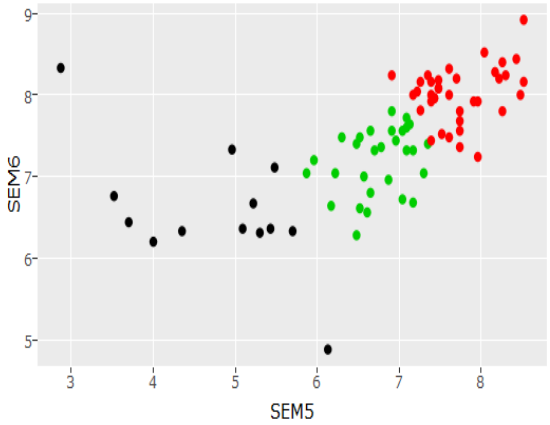

Fig IV: Semester 5 SGPA vs Semester 6 SGPA

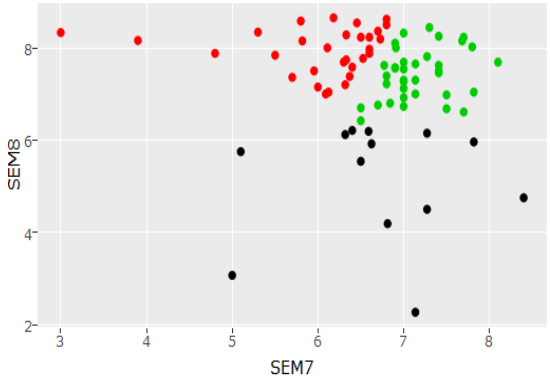

Fig V: Semester 7 SGPA vs. Semester 8 SGPA

Using the K-means clustering algorithm, defined with the squared Euclidean distance, we cluster the students into three, formed namely the high performers (cluster1), average performers (cluster2) and performers who need improvement (cluster3), based on their academic performance of each year, with the R IDE. The clusters are then represented in the form of graphs as displayed above. The students who show high performance that is, falling into cluster1 in all the four graphs are taken separately. The patterns of those students are analyzed.

5. EXPERIMENTAL RESULT

Cluster 1 students are 46.83% of the total number of students taken into consideration. On analysis, it is found that 75.94% of the total female students, 5.26% of the total male students, 45.76% of the total day-scholars and 50% of the total hostellers. The learning methodology of the students is analysed. Out of those, 70.27% of the students use the web

resources for their examination preparations and 77% of the students prefer studying in groups. On comparing the co-curricular and extra-curricular activities of the cluster 1 students, they show more participation in the co-curricular than the extra-curricular activities.

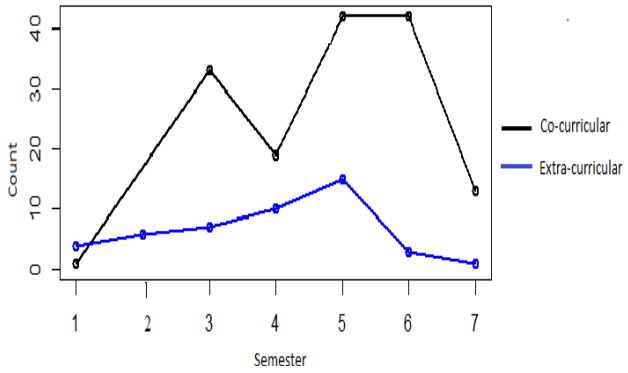

Fig VI: Co-curricular and Extra-curricular participations of cluster 1 students

Also the SGPA of female students of the whole dataset range between 4.17 and 8.92; between 2.61 and 8.04 for male students, whereas the high performers always fall within 7.02 and 8.92.

6. CONCLUSION AND FUTURE WORK

On mining the patterns among the students, we can conclude that female students perform better than the male students. Hostel students have shown better performance compared to day-scholars, as they work with a schedule. The performance indicator of the high performing students is observed to be their learning methodology principally. The more effective methodology of learning is the usage of web-resources and the students who prefer to study in groups have performed well. In future, prediction algorithm is to be applied to predict the students' performance. With the same mining strategies, the effective teaching methodology is to be observed.

7. ACKNOWLEDGMENT

We acknowledge all authors whose work gave knowledge to us. Thanks a lot.

8. REFERENCES

1. Mohammed I.Al-Twijri, Amin Y. Noaman, A New Data Mining Model Adopted for Higher Institutions
2. Parneet Kaur,Manpreet Singh,Gurpreet Singh Josan," Classification and prediction based data mining algorithms to predict slow learners in education sector "
3. Harwati, Ardita Permata Alfiani, Febriana Ayu Wulandari, "Mapping Student's Performance Based on Data Mining Approach (A Case Study)"
4. Pooja Thakar , Anil Mehta, Manisha, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue"
5. Amjad Abu Saa," Educational Data Mining & Students' Performance Prediction "
6. Tripti Dwivedi, Diwakar Singh ," Analyzing Educational Data through EDM Process: A Survey "
7. Dr. P. Nithya, B. Umamaheswari, A. Umadevi," A Survey on Educational Data Mining in Field of Education "
8. P.Veeramuthu, Dr.R.Periyasamy, V.Sugasini, "Analysis of Student Result Using Clustering Techniques"