Data Mining Techniques for Predicting Students' Academic Performance in Higher Educational Institutes

Alok Singh Chauhan, Dr. Yashpal Singh

Abstract — Higher education is a crucial zone for any successful nation. All exploration and developments for the most part originates from advanced education. Top most profitable experts like designers, supervisors, researchers are originates from the channel of higher education. The supervision of the scholarly execution of designing/understudies is fundamental amid a beginning time of their educational module. In reality, their evaluations in particular center/real courses and additionally their combined General Point Average (GPA) are conclusive when relating to their capacity/condition to seek after higher examinations. Information mining joins machine learning, measurements and perception methods to find and concentrate learning.
This paper proposes to apply data mining strategies to anticipate understudies' scholastic execution in higher instructive establishments. In this paper we additionally utilized information mining and factual strategies to enhance students' execution and defeat the issue of low evaluations of understudies.

Index Terms — Data Mining, K-means clustering algorithm, Decision Tree and Regression

I. Introduction

The approach of data innovation in different fields has lead the substantial volumes of information stockpiling in different organizations like records, records, archives, pictures, sound, recordings, logical information and numerous new information positions. The information gathered from various applications requires legitimate technique for removing learning from extensive vaults for better basic leadership. Knowledge discovery in databases, frequently called data mining, goes for the revelation of helpful data from vast accumulations of information [2]. The principle elements of information mining are applying different techniques and calculations keeping in mind the end goal to find and concentrate examples of put away information [3]. Information mining and learning revelation applications have a rich concentration because of its hugeness in basic leadership and it has turned into a basic part in different associations [13].

There are expanding research interests in utilizing information mining in instruction. This new rising field, called Educational Data Mining, worries with creating techniques that find information from information starting from instructive conditions [16]. Instructive Data Mining utilizes numerous procedures, for example, Decision Trees, Neural Networks, Naïve Bayes, K-Nearest neighbor, and numerous others. Utilizing these methods numerous sorts of information can be found, for example, affiliation principles, orders and grouping.

The found information can be utilized for expectation with respect to enrolment of understudies in a specific course, recognition of uncalled for implies utilized as a part of online examination, identification of strange esteems in the outcome sheets of the understudies, forecast about understudies' execution et cetera [7, 9]. Information grouping is often utilized as a part of numerous fields, for example, information mining, design acknowledgment, choice help, machine learning and picture division [4, 5 14]. As the most understood system for performing non-hierarchal grouping, the K-implies bunching [6] iteratively discover the k centroids and doles out each example to the closest centroid, where the facilitate of every centroid is the mean of the directions of the articles in the group.

The primary goal of this investigation is to apply Data Mining Techniques to anticipate understudy's scholarly execution in End Semester Examination which distinguishes understudies in danger early and enable the staff to give suitable exhortation in a convenient way.

II. Related Work

Information mining is a developing strategy utilized as a part of instructive field to upgrade our comprehension of learning procedure to center around recognizing, extricating and assessing factors identified with the learning procedure of understudies (Alaa el-Halees 2009). (Kifaya, 2009) K-implies grouping is a broadly utilized strategy that is simple and very easy to get it. (Galit, 2007) gave a contextual analysis that utilization understudies information to dissect their learning conduct to anticipate the outcomes and to caution understudies in danger before their last, most decisive tests. Kumar and Uma (2009) contemplated understudies' execution in the course utilizing information mining methods, especially characterization strategies, for example, Naïve Bayes and Decision tree in view of understudies ID and imprints scored in course.

Moreover, they propose that information mining procedure should be possible to the educators for characterizing execution which helps in enhancing advanced education framework. Information mining strategies causes understudies and educators to enhance understudies' execution. Quadri and Kalyankar (2010) directed an examination on information mining so as to discover the scholarly execution from understudies information in the instructive establishments utilizing the choice tree methods.

Manpreet Singh Bhullar (2012) led an examination on training segment with a specific end goal to recognize how the information mining methods encourages the instructive foundations to perform better. Information mining and its strategies encourages the instruction organization to associate more with the understudies' by its different highlights. Creator utilized characterization system alongside J48 calculation with a specific end goal to anticipate the understudies' outcome.

III. Data Mining Overview

Han and Kamber (2006) characterize information

Table 1: Selected Attributes

S. No.	Given Attributes	Description
1	Gender	Male, Female
2	10th Division	I, II, III
3	12th Division	I, II, III
4	Graduation Division	I, II, III
5	Previous Semester Marks	Less than 60%, 60 to 80%, Over 80%
6	Class test Grade	A, B, C
7	Seminar Performance	Good, Average, Below
8	Assignment	Good, Average, Below
9	Attendance	Less than 60%, 60 to 80%, Over 80%
10	Lab Work	Good, Average, Below
11	General Proficiency	Good, Average, Below
12	End Semester Marks	Less than 60%, 60 to 80%, Over 80%

mining as the way toward finding 'shrouded pictures', examples and learning inside vast measure of information and making expectations for results or practices [16]. Information mining is the self-loader disclosure of examples, affiliations, changes, peculiarities, rules, and measurably critical structures and occasions in information. That is, information mining endeavors to remove learning from information. Information mining utilizes a blend of an unequivocal learning base, complex expository aptitudes, and space information to reveal concealed patterns and examples. These patterns and examples shape the premise of prescient models that empower experts to create new perceptions from existing information.

Mining in instructive condition is called Educational Data Mining [10], worry with growing new strategies to find information from instructive databases (Galit, 2007) (Erdogan and Timor 2005), keeping in mind the end goal to examine understudies patterns and practices toward training (Alaa el-Halees , 2009) [12]. Absence of profound and enough information in higher instructive framework may forestall framework administration to accomplish quality destinations, information mining philosophy can help spanning this learning holes in advanced education framework [1, 9].

IV. Students Data Set And Preprocessing

In this research, dataset has been gathered from three foundations: Institute of Professional Studies (Center of Computer Education), University of Allahabad, Ewing Christian Institute of Management and Technology (built up by Ewing Christian College Society, Allahabad) and United Institute of Management, Naini, Allahabad. The dataset contains distinctive qualities. The entire depiction of dataset is appeared in table 1.

V. Methodology

The exploration technique adjusted depends on the inside and out investigation of the subject relating to the information mining and its application in advanced education. To satisfy the present examination goals, we have been gathered primary data by distributing questionnaires to gather comprehensive data linked to the domain being studied. This study used a questionnaire and survey technique as a data collection instrument. The participants for the investigation were chosen through arbitrary testing. The proposed methodology will be utilized to produce a database for the current study. For this examination, we have been gathered dataset of MCA students from three different institutes of Allahabad as mentioned above in source of data. Before processing of data we will be going to clean the information to expel commotion and inconsistency. To expel missing esteems in the dataset, we will utilize the cleaning methods. The scholastic exhibitions of the understudies are successfully estimated utilizing their Grade Point Average (GPA). With the end goal of this investigation, the contributing variables like class test marks, midterm test imprints, task and lab work are gathered and taken for measurable examination. The tests and perceptions will be done by utilizing information mining instrument WEKA and SPSS.

VI. Application Of Data Mining Techniques To Students Dataset: Results And Discussion

The current statistical analysis has used multiple tools to analyse the data collected. These statistical tools include frequency distributions, one way anova tests, independent sample tests, data clustering and decision trees. We might take a gander at each technique separately.

A. Analysis Technique 1: Frequency Distribution

As per frequency distributions of the current data, the following inferences can be drawn:

i. Nearly 51.3% of students previous semester score falls in first grade, 23.3% students previous semester score falls in second grade and 14.7% students previous semester score falls in third grade. That is, over 50% of the students have got good grades in their previous semester exams. Thus, one can derive that the execution of understudies has deteriorated as time has passed. There could be several reasons for the fall in performance. The chief amongst them might be either low attendance, poor performance in assignments or the fact that the studies could have become tougher.

ii. Nearly 74.7% of student's assignment scores are in good grade and 24% of students assignment scores are in average grade. Explanation behind this could be credited to both the efforts of the teachers and that of the students. Further, good assignment grades show that the students are studying properly.

iii. Nearly 79.3% of the students are having good attendance score and 10% of the students are having average attendance score. A good attendance score shows that students have been attending their classes regularly. Proper attendance ensures that the students are taught properly and that they finish their assignments on time. Further, regular attendance also signifies that the students are aware of what is important in terms of studies.

iv. Nearly 48.7% of students score falls in first grade, 29.3% students score falls in second grade and 12.7% students score falls in third grade. That is, over 50% of the students have secured good grades in their previous semester exams. Even though there is a slight slip in the first grade scores, we see that there is a significant increase in the second and third grade scores and a dip in fail grade. This shows that the performance of the students in the second and third semester falls compared to the first grade performance. A poor performance as stated previously could be credited to several reasons. The primary reason might be a dip in the attendance or the syllabus becoming tougher. This also indicates that teachers require giving more attention to the students especially in the second and third grades.

v. Nearly 39.3% students have GPA scores between 1 and 2, 38.0% students have GPA scores between 2 and 3, 12.7% students have GPA scores between 3 and 3.5 and 4.7% students have GPA scores above 3.5. This clearly indicates that majority of the students are not performing too well. This in turn signifies that the teachers require paying attention to the students and try to identify techniques by which the performance of the students would improve.

From the above, it can clearly be inferred that students have not been performing extremely well in the GPA. Also, students have been performing badly as compared to their first semester. However, this cannot be credited to poor attendance alone, as majority of the students have had a good attendance record. Hence, it is crucial for teachers to identify why the performance of the students has suffered as compared to their first semester.

B. Analysis Technique 2: One Way Anova

According to the current analysis conducted via One Way ANOVAs, the following inferences are drawn:

- The mean score in class test is very high for students who scored well in their assignments (mean score = 8.6875, std dev = 1.158). This indicates that students who are having good assignment score records tend to score higher marks in their class tests (F test statistic = 20.262, p – value = 0.000 < 0.05). Thus, it can be inferred that when students perform well in their assignments, they do equally well in their class tests. Hence, teachers should realise that students who desire to perform well in their class tests need to perform well in their assignments.
- The mean score in midterm test is very high for students who scored well in their assignments (mean score = 11.8036, std dev = 1.361). This indicates that students who are having good assignment score records tend to score higher marks in their midterm tests (F test statistic = 44.646, p – value = 0.000 < 0.05). This directly helps us to infer that students performing well at their tests usually score well in their assignments.
- The mean score in class test is very high for students who completes their lab work (mean score = 8.47, std dev = 1.08). This indicates that students who concentrates on their lab work tend to score higher marks in their class tests when compared to students who did not focus on their lab work (t test statistic = - 2.069, p – value = 0.040 < 0.05). This helps us to infer that students who complete their lab work satisfactorily perform better in their class tests.
- The mean score in midterm test is very high for students who completes their lab work (mean score = 11.43, std dev = 1.45). This indicates that students who concentrates on their lab work tend to score higher marks in their midterm tests when compared to students who did not focus on their lab work (t test statistic = - 3.853, p – value = 0.000 < 0.05). This also helps one to comprehend that students who score higher marks in their midterm tests are those who finish their lab work in time.

The One Way Anova test had made it evident that for students to perform well in both their class tests and their mid-term tests they need to perform well on their assignments. Further, they also need to ensure that they complete their lab work to score well on their tests. Thus, teachers need to ensure that for students to perform well, their students need to complete their assignments successfully an also complete their lab work.

C. Analysis Technique 3: Data Clustering

The current study has used GPA scores of the students in k- means clustering. The selected sample has been segregated into three groups. Usually the GPA scores have a range of 0 and 4. Students whose score above 95% marks have a GPA score of 4. However, students whose score below 60% have a GPA score of 0. The number of clusters in the current study was fixed as 3. The SPSS output according to the analysis was as under:

The first group was the 'High Score Group, the second group was the 'Medium Score Group' while the last group was the 'Low Score Group'. According to the analysis, it was discovered that majority - 55.3% of students GPA score belonged to the 'High Score Group' (GPA scores closer to 3.01); likewise, 39.33% students GPA score belonged to the 'Medium Score Group' (GPA scores closer to 1.77) while a mere 5.33% students GPA score belonged to the 'Low Score Group' (GPA scores closer to 0.75).

This made it clear that there were students who were performing extremely well in their GPA. However, there was a group of students who had failed to perform adequately. It was essential for the teachers to focus

Table: 2

Priority	Grade	GPA Scores	Attention Required
Less Priority	A+	> 3.5	No special attention is required. We can keep these students as a benchmark for the lower grade students
Medium Priority	A, A-	3 – 3.5	Not too bad. But bit more attention required on their Class test and other activities where he makes mistakes
High Priority	B+, B	2 – 3	Has to give high priority, as these group students have potential to get high GPA scores. Therefore, high preference is required.
Very high priority	Below B grade	< 2	Students are of lower standard falls in this category and they require lots of practice in class test, midterm test and lab work too.

on these students to ensure that their performance improved for the GPA. It was also essential for the teachers to identify the causes behind the poor performance of the students. Despite not being analysed, the study made it obvious, that the poor scores could be attributed to three aspects:

- Poor attendance
- Poor performance in the assignments
- Incomplete lab work

Thus, it was crucial that the students comprehended the significance of doing their assignments and completing the lab work. The teachers also needed to focus on this aspect so that the students whose performance was poor could improve their performance.

D. Analysis Technique 4: Decision Tree

The decision tree is a well-known technique employed in data mining. It is these decisions that act as rules to segregate the observations or gathered set of data. In the decision tree, every branch node stands for different choices while every leaf symbolises the decision made. This is a common technique employed to get details which would be required to make decisions. The decision tree is considered to be a useful technique as it is fast and can be used to identify the characteristic that is most beneficial.

As per the decision tree made in the current analysis, the inferences that were drawn included the following Table 2

From the given table makes it clear that the focus of the teachers needed to be on students who secured below B grade or even those securing a B+ or B grade. The different tests conducted had clearly outlined the reasons why students performed well in the class tests and the mid-term tests. It was crucial for teachers to comprehend and explain to the students the need for completing their assignments properly and also completing their lab work. It can be inferred that students who worked on a regular basis, attended class regularly and finished their assignments properly and also their lab work performed consistently well in their tests.

VII. Conclusion

In this paper, we gave a contextual investigation in the instructive data mining. It indicated how valuable data mining can be utilized as a part of higher education especially to enhance MCA understudies' execution. We have utilized measurable techniques Frequency distribution and One-Way ANOVA. We connected data mining methods to find learning. The changed strategies utilized as a part of the present examination in reference to data mining included k-means clustering and decision tree techniques. Every single one of these system can be utilized to enhance the execution of MCA understudy. The examination made it clear that the understudies who performed well in their mid-term tests and class tests were chiefly the individuals who had performed well on their assignments and the individuals who had finished their lab work. Along these lines, it can effectively be induced that instructors expected to recognize if the understudies who were getting low evaluations were additionally performing ineffectively in setting of their assignments and lab work.

References

[1] Shyamala K. and Rajagopalan S. P. (2006), 'Data Mining Model for a better Higher Educational System', Information Technology Journal, Vol. 5, No. 3, pp 560-564.

[2] Heikki, Mannila, Data mining: machine learning, statistics, and databases, IEEE, 1996.

[3] U. Fayadd, Piatesky, G. Shapiro, and P. Smyth, From data mining to knowledge discovery in databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–262 56097–6, 1996.

[4] Sun Shibao, Qin Keyun, "Research on Modified k-means Data Cluster Algorithm", I. S. Jacobs and C.P. Bean, " Fine Particles, Thin Films and Exchange Anisotropy", Computer Engineering, 33.

[5] Merz C. and Murphy P., UCI Repository of Machine Learning Databases.

[6] Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautics and Space Technologies July 2005 Volume 2 Number 2 (53-57).

[7] N. V. Anand Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique," European Journal of Scientific Research, vol. 34 (4), 2009.

[8] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M.Inayat Khan, "Data Mining Model for Higher Education System", European Journal of Scientific Research ISSN 1450-216X Vol.43 No.1 (2010), pp.24-29

[9] Brijesh Kumar Baradwaj and Saurabh Pal, "Mining Educational Data to Analyze Students. Performance",

International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
[10] M. N. Murty and A. K. Jain, "Data Clustering: A review", ACM computing surveys, 31, 1999.
[11] Alaa el-Halees (2009) Mining Students Data to analyze e-Learning Behavior: A Case Study.
[12] Fayyad, U.M., Data mining and knowledge discovery: making sense out of data. IEEE Expert 11, October 1996, pp. 20-25.
[13] Yuan F., Meng Z. H., Zhang H. X. Dong C. R., "A New Algorithm to Get the Initial Centroids", Proceeding of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26-29, August 2004.
[14] Berson, "Data Warehousing, Data-Mining & OLAP", TMH
[15] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.
[16] Chandra, E. and Nandhini, K. (2010) 'Knowledge Mining from Student Data', European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163.
[17] Mohammed M. Abu Tair, Alaa M. El-Halees (2012), "Mining Educational Data to Improve Students' Performance" International Journal of Information and Communication Technology Research, Volume 2 No. 2, February 2012, pp. 140-146.

AUTHORS' DETAILS

Alok Singh Chauhan
Assistant Professor,
IT Department, IMS Ghaziabad (University Courses Campus), India
Research Scholar, Bundelkhand University, Jhansi, India
Email: alokchauhan.1983@gmail.com

Dr. Yashpal Singh
Associate Professor & Head
Department of Computer Science & Engineering, BIET Jhansi, India
Email: yash_biet@yahoo.co.in

CITE THIS ARTICLE AS :

Alok Singh Chauhan, Dr. Yashpal Singh, "Data Mining Techniques for Predicting Students' Academic Performance in Higher Educational Institutes," International Journal of Technology and Science, vol. 5, Issue. 1, pp. 6-10, 2018