PREDICTION OF ACADEMIC PERFORMANCE FOR HIGHER EDUCATION STUDY USING DATA MINING CLASSIFIER AND MONGODB

[1]Harish Barapatre, [2]Dr. Yogesh Kumar Sharma, [3]Vaishali Shinde
JJTU University, Rajasthan Dept of Computer Scinece[1], JJTU University, Rajasthan Dept of Computer Scinece[2], JJTU University, Rajasthan Dept of Computer Scinece[3]
harishkbarapatre@gmail.com[1], dr.yogeshkumar@yahoo.in[2], vaishalishinde22@gmail.com[3]

ABSTRACT

Educational Data Mining is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyse educational data in order to study educational questions. This paper surveys the most relevant studies carried out in this field to date. Firstly, it introduces EDM and describes the different groups of users, types of educational environments and the data they provide. It then goes on to list the most typical/common tasks in the educational environment that have been resolved through data mining techniques and finally some of the most promising future lines of research are discussed. With the overwhelming successes gained in Big Data analysis in the Business Industry, it is little wonder that there is a strong belief in the academia that these successes can be replicated in the Education Sector. As new findings and outcomes of research crop up daily, it is my belief that amongst these successes potentially identifiable, prediction of students' academic performance can have strong positive influences in knowledge management and delivery in education thereby adding more quality to the learning experience.

Keywords— Data Mining, Student Database Mapping, Prediction analysis, Educational dynamics of behavior.

INTRODUCTION

Educational Data Mining (EDM) is the application of Data Mining (DM) techniques to educational data, and so, its objective is to analyse these types of data in order to resolve educational research issues. DM can be defined as the process involved in extracting interesting, interpretable, useful and novel information from data. It has been used for many years by businesses, scientists and governments to sift through volumes of data like airline passenger records, census data and the supermarket scanner data that produces market research reports. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn. On one hand, the increase in both instrumental educational software as well as state databases of student information has created large repositories of data reflecting how students learn. On the other hand, the use of Internet in education has created a new context known as eLearning or web-based education in which large amounts of information about teaching-learning interaction are endlessly generated and ubiquitously available. All this information provides a gold mine of educational data. EDM seeks to use these data repositories to better understand learners and learning, and to develop computational approaches that combine data and theory to transform practice to benefit learners. EDM has emerged as a research area in recent years for researchers all over the world from different and related research areas such as

- Offline education tries to transmit knowledge and skills based on face-to-face contact and also study psychologically on how humans learn. Psychometrics and statistical techniques have been applied to data like student behaviour/performance, curriculum, etc. that was gathered in classroom environments
- E-learning and Learning Management System (LMS). Elearning provides online instruction and LMS also provides communication, collaboration, administration and reporting tools. Web Mining (WM) techniques have been applied to student data stored by these systems in log files and databases.

- Intelligent Tutoring (ITS) and Adaptive Educational Hypermedia System (AEHS) are an alternative to the just-put-it-on-the-web approach by trying to adapt teaching to the needs of each particular student. Data Mining has been applied to data picked up by these systems, such as log files, user models, etc.

The EDM process converts raw data coming from educational systems into useful information that could potentially have a great impact on educational research and practice. This process does not differ much from other application areas of data mining like business, genetics, medicine, etc. because it follows the same steps as the general data mining process: pre-processing, data mining and post-processing. However, it is important to notice that in this paper the term data mining is used in a larger sense than the original/traditional DM definition. That is, we are going to describe not only EDM studies that use typical DM techniques such as classification, clustering, association rule mining, sequential mining, text mining, etc. but also other approaches such as regression, correlation, visualization, etc. that are not considered to be DM in a strict sense. Furthermore, some methodological innovations and trends in EDM such as discovery with models and the integration of psychometric modelling frameworks are unusual DM categories or not necessarily universally seen as being DM.

The full paper is written as follows: Section II review the related works of Educational data mining and prediction result using machine learning algorithm. The ML algorithm concerns and potential risks of huge data are debated in Section III. The integrated proposed system and methodology of different classification and challenges, objective, flow system, and different methods is discussed in Section IV and V. The experimental results and accuracy calculation and GD points with 2-layer classification are presented in Section VI. Finally, Conclusions and future work are presented in Section VII.

REVIEW OF LITERATURE

Despite the amount of research interest, the LA/EDM domain has received and is currently still receiving, some researchers are of the belief that there are a number of unexplored issues in this rapidly growing domain and have suggested the incorporation of some other emerging research technologies with LA/EDM. One of such suggested technologies is Game-Based Learning (GBL). Conolly et al., (2012), postulated that "playing computer games is linked to a variety of perceptual, cognitive, behavioral, affective and motivational impacts and outcomes." Papamitsiou and Economides (2014) acknowledged that GBL has positive impacts on learners and suggested that an interesting area of future research would be to find out if and how LA/EDM methods could be applied to report and visualize learning processes during GBL.

Another emerging technology rapidly developing is that of mobile and ubiquitous learning. Mobile Learning has been acknowledged for the unique opportunity it affords learners authentic learning experiences anytime and anywhere (Tatar et al., 2003). Papamitsiou and Economides (2014) suggest that LA/EDM research could investigate the appropriateness of the popular methods in mobile and ubiquitous learning contexts as illustrated by (Chen and Chen, 2009; Leong et al., 2012), in order to provide sophisticated, personalized learning services through mobile applications.

In the same vein, several regression techniques have also been used for predication. Kotsiantis and Pintelas (2005), used model trees, linear regression, neural networks, support vector machines and locally weighed linear regression to predict students' marks in an open university. Linear regression prediction models have been used for predicting end-of-year accountability assessment scores (Anozie and Junker, 2006) while a multivariable regression model was used by Yu et al., (1999), to predict student performance from log and test scores in web-based instruction.

Stepwise linear regression was used for predicting student academic performance (Golding and Donalson, 2006) while multiple linear regression was used for predicting time to be spent on a learning page (Arnold et al., 2005). Martinez (2001), identified variables that could predict success in college courses using multiple regression while Thomas and Galambos (2004), used regression and decision trees analysis for predicting university students' satisfaction. Linear regression was used for predicting exam results in distance education courses (Myller et al., 2002), for predicting end-of-year accountability assessment scores (Anozie and Junker, 2006) and also for predicting the probability that the student's next response is correct (Beck and Woolf, 2000). Logistic regression was used for predicting when a student will get a question correct and association rules to guide a search process to find transfer models to predict a student's success (Frey berger et al., 2004). Robust Ridge regression algorithm was used to predict the probability of a student giving the correct answer to a problem (Cetintas et al., 2009) while stepwise regression was used to predict a student's test score (Feng et al., 2005).

Correlation analyses have been applied together to predict web-student performance in on-line classes (Wang and Newlin, 2002), to predict a student's final exam score in online tutoring (Pritchard and Warnakulasooria, 2005) and for predicting high school students' probabilities of success in universities (McDonald, 2004

NEED OF THE STUDY

Data collection, analysis and management in the education sector in all education is problematic and non-standardized. Secondary and Tertiary institutions of learning manage their data and there is no regulatory form of sharing this data with other stakeholders in the industry. This leads to various agencies and bodies collecting and analysing their own data to meet their unique requirements. The resultant effect is a lot of data redundancy across the industry with minimal efficiency and non-optimization of its use.

Learning Analytics (LA) and Educational Data Mining (EDM) have provided significant results only in quantitative research and not in qualitative research. There is no doubting the availability of Big Data in the educational sector and with the Data Analytics and Data Mining successes in the business sector, there is no empirical qualitative evidence of its success or otherwise in the Education Sector. The tools and resultant information processing and management techniques and procedures, relevant for achieving overall business goals and objectives, are relevant also for educational providers, managers and other stakeholders to enhance efficacy and efficiency of learning management and delivery.

This research aims to provide a standardized framework/model for data collection and analysis on Educational Big Data in education organization by exploring LA and EDM methods and techniques and determining how they can be employed particularly in prediction of Students Academic Performance.

PRIDICTIVE ANALYTICS PROCESS

Predictive analytics Process involves the following model which shown in figure-1

1. Understanding The data: The first step is to identify the outcome of the project, the deliverables, business objectives, and based on which the data collection will be poised.

2. Data Collection: The next step is to collect data from multiple sources, to have a picture of the various customer interactions as a single view item.

3. Analysis: Further, the data is inspected, cleansed, transformed, and modelled to discover if it provides useful information and helps to conclude.

4. Statistics: This enables to validate if the findings, assumptions, and hypothesis are admirable to go ahead and test them using a statistical model.

5. Modelling: It helps in creating an error-free predictive model about the future that provides options to choose the best option that can be sorted through multi-model evaluation.

6. Deployment: Predictive model deployment provides an option to create and deploy the analytics results into productive decision-making. It also helps to generate results, reports, and other metrics.

7. Monitoring: Models that are prepared are further tracked to control and check for performance conformance to ensure that the desired results are acquired as expected.

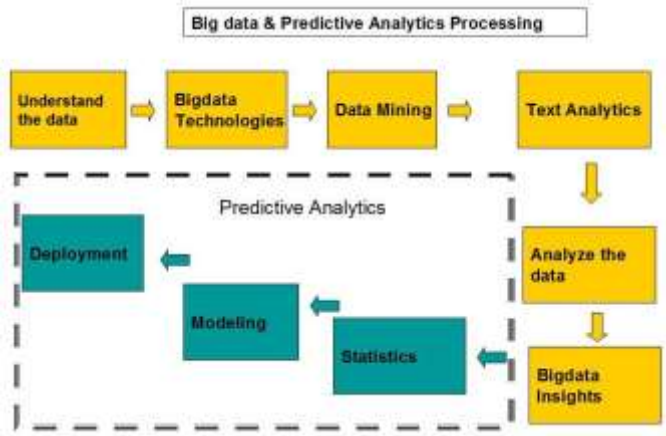


Fig-1: Predictive analytics Process

Predictive analytics is the branch of the advanced **analytics** which is used to make predictions about unknown future events. **Predictive analytics** uses many techniques from **data** mining, statistics, modelling, machine learning, and artificial intelligence to analyse current **data** to make predictions about future.

OBJECTIVE

The primary objective of this study is to provide a standardized framework/model for data collection and analysis on Educational Big Data in all schools to enable Learning Providers and Admin predict Students' Academic Performances.

The secondary objectives are as follows:

- To explore how LA and EDM in Educational Big Data can impact Knowledge Learning and Delivery in all education organization.
- To investigate whether we can get significant results in using qualitative research methods in LA/EDM
- To contribute to knowledge by investigating the potential possibilities of combining LA/EDM with GBL or mLearning in positively enhancing the overall learning experience.
- The main objective is to design and develop a Big Data Analytical Framework for use in the educational settings. This study is two folded, with the first phase focusing on constructing the conceptual framework for Big Data Analytics and the subsequent phase focusing on developing and testing the prototype.

PROPOSED SYSTEM

The underlying structure of the proposed framework mainly revolves around the stages of input, process and output (IPO) modelled after Garris et al, (2002). Hence, the identified components were grouped into four major stages for the new framework.

The four major stages identified are data acquisition, data processing, data mining and data display and is termed as **CoProMinD** (Collecting, Processing, Mining and Displaying results). These four stages form the major pillars for the proposed framework.

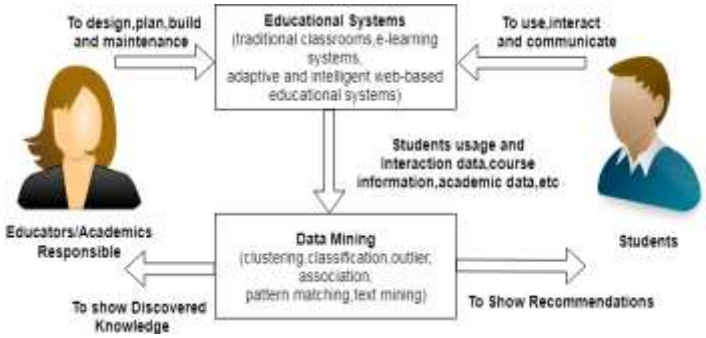


Fig-2: Proposed Framework

The Collection refers to the activities of acquiring data by identifying data, planning the acquisition, transporting and securing while in transit. Meanwhile, Processing responsible for whole lot of pre-processing task. In Mining (mine), the actual data analytics takes place by producing predictive models using appropriate algorithms. Finally, Display refers to visualization and the various reporting formats to show the end-result of the analytics. The placement of the components will follow the structure of CoProMind where the flow will start from Collecting and ends with Displaying as per Figure-2. Those components identified earlier will be placed in the CoProMind structure.

METHODOLOGY

A. Educational Data Mining

EDM is concerned with the application of data mining techniques on educational data with the goal of addressing educational challenges and discovering hidden insights in data. Data may include mining student demographic data and navigation behaviour within a learning environment, learning activities data such as quizzes, interactive class exercises/activities, as well as data from a group of students working together in an exercise, text chat forum, teacher data, administrative data, demographic data, and emotional data. EDM can be employed to access student-learning outcomes, enhance learning processes and supervise student learning to give feedback. Data can also be used to offer recommendations to suit the learning behaviour of students based on individuals, evaluation of learning design as well as the discovery of irregular learning behaviours. For instance, Ayesha, et al. reported on how the learning activities of students were predicted with the application of k-means clustering algorithm. Pal employed machine-learning algorithm to determine fresh engineering undergraduates that were anticipated to drop out in their initial year. Parack, et al. applied various data mining algorithms to carry out student profiling to categorise them given their academic records that include practical test scores, quiz scores, exam scores and assessment grades.

This study adapted the Design Science Research Methodology (DSRM) by (Peffers, Tuunanen, Rothenberger & Chatterjee, 2007) which has become a distinct line of research paradigm within the IS field. The DSRM method was employed to guide the research directions and to answer the research objectives and questions that were formulated. The core research activities of the design and development of the proposed framework and the prototype was based on a combination of "existing theory and prior research" and "exploratory research" (Maxwell, 2012). This combination makes up the basis or foundation of the new proposed framework. The existing theory or prior research provide insights into the current practices and research gaps while exploratory method provided avenue for possibilities of new findings. In depth analysis of related work was done through extensive systematic and narrative literature review. Comparative analysis was carried out to discover and comprehend existing findings and gather knowledge on the area of the study. The SP Theory of Intelligence and Bounded Rational Theory has become the theoretical basis for the development of the framework.

The Big Data community has focused largely on the use of analytics to discover unpredicted, novel, unforeseen, or unidentified circumstances and report back to the stakeholders as the analytic output (Szymczak, Zelik, & Elm, 2014). Similarly, educational agencies can use this data driven decision making method to improve institutional functions, spot trends and directions, create better learning opportunities, and predict student performance and the performance of the organization.

In embracing greater integration of ICT to enhance the effectiveness of education and training programs, many initiatives were deployed by the Ministry of Education. The Frog- VLE through 1-BestariNet, is perceived as a tool to revolutionize learning, to produce richer curricula, to enhance pedagogies, to lead to more effective organizational structures in schools and to produce stronger links between schools and society to empower learners (Sua, 2012). The Frog-VLE is the world's first nationwide deployment of school in the cloud, connecting all its schools on a single learning platform. Frog-VLE provides virtual access to classes, lesson content, homework, assessments, grades and other external resources. There is also a social component, which enables pupils and teachers to interact in threaded discussions or chat. The connectivity established through FROG-VLE between students, teachers, parents and school administration generates huge volume of data that will indeed be of great value to whole education system.

Moreover, sub divisions under the MOE have been tasked to develop and maintain various databases and manage remotely data related to schools, students and teachers. There are almost four hundred over databases that have been developed, managed and maintained separately by the respective divisions. These databases are often used by the relevant authorities to channel information to policy makers and stakeholders as and when the need arise. Substantially Couldry & Powell (2014) claimed that the actual process of data gathering, data processing and organizational adjustment associated with Big Data narratives constitute important facts which all stakeholders must deal with. It is critical for stakeholders to figure out how to capture, store, and analyse the data correctly in order to gain insights and to use the right information in the organization to accelerate institutional functions accurately. The proliferation of data across the databases makes it critical to integrate and manage these diverse data assets and leverage the power of insights from an integrated view. Big Data Analytics can be considered as a potential solution to orchestrate all these various databases to a single integrated platform.

B. Framework Development

The proposed framework is the culmination of various recommendations and research gap found during the extensive and in-depth reading. In developing the framework, this study drew on research in the big data, educational data mining, pedagogical related fields, Malaysian school's database systems, interviews with key stakeholders in Education Ministry of schools and practitioners. Systematic Literature Review (SLR) provided the related general big data educational frameworks. All together 6 Big Data Analytical frameworks were reviewed to identify and gather information on the key components of the framework. Comparing the components of the chosen 6 frameworks for big data analytics, it can be seen that the frameworks share some fundamental components or constructs. The components are retained or combined if they perform similar tasks. New components added to fill the gap to produce more polished framework. A total of 33 framework components from the 6 frameworks were identified.

C. Grouping of Components

The 33 components found from the review are grouped to newly consolidated groups forming part of the new framework proposed. The components that are logically related reduced and merged into nine distinguished

components to represent the essence of details in each group. The new groups' names either use the existing names from any of the framework or a new name is given. Table 1 shows the grouping of components and the new names coined from the consolidation.

USED TECHNOLOGY

Mongo DB is an open source document database and leading NoSQL database. MongoDB is written in C++. MongoDB is a cross-platform, document-oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document. Database is a physical container for collections. Each database gets its own set of files on the file system. A single MongoDB server typically has multiple databases. Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database. Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or related purpose.

Schema less MongoDB is a document database in which one collection holds different documents. Number of fields, content and size of the document can differ from one document to another. MongoDB supports dynamic queries on documents using a document-based query language that's nearly as powerful as SQL.

Data collection, analysis and management in the education sector in all education is problematic and non-standardized. Secondary and Tertiary institutions of learning manage their data and there is no regulatory form of sharing this data with other stakeholders in the industry. This leads to various agencies and bodies collecting and analysing their own data to meet their unique requirements. The resultant effect is a lot of data redundancy across the industry with minimal efficiency and non-optimization of its use.

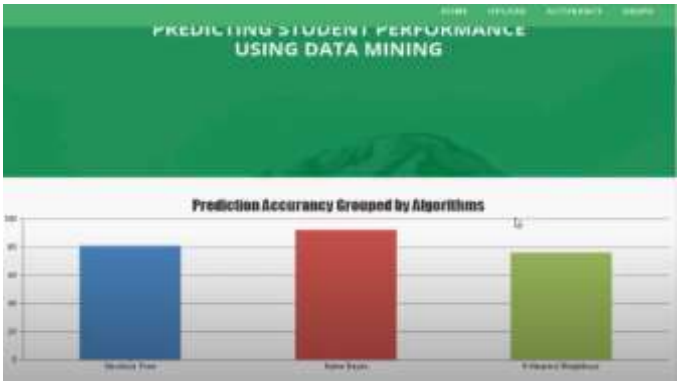
Learning Analytics (LA) and Educational Data Mining (EDM) have provided significant results only in quantitative research and not in qualitative research. There is no doubting the availability of Big Data in the educational sector and with the Data Analytics and Data Mining successes in the business sector, there is no empirical qualitative evidence of its success or otherwise in the Education Sector. The tools and resultant information processing and management techniques and procedures, relevant for achieving overall business goals and objectives, are relevant also for educational providers, managers and other stakeholders to enhance efficacy and efficiency of learning management and delivery.

This works aims to provide a standardized framework/model for data collection and analysis on Educational Big Data in education organization by exploring LA and EDM methods and techniques and determining how they can be employed particularly in prediction of Students Academic Performance.

ANALYSIS

Comparison of Algorithms for Student Performance Prediction:

The below screen shows a comparison of 3 different algorithms that are used to predict the student performance. It can be observed that Naïve Bayesian algorithm can predict the student performance with 85% accuracy.



We have modifed the NB and defining valid rules which improves the accurcy as compared to other algorithum

CONCLUSION

Educational data mining techniques are used to predict student performance at a very early stage due to the consideration of important factors such as student background and involvement of student in social activities. This helps the at-risk students at a very early stage and gives them scope to improve their performance. The teachers can also prepare early and come up with new techniques in teaching the topic.

In future, unsupervised machine learning algorithms and data mining techniques could be applied to discover the relationship and impact up the attributes in clusters. This attribute analysis and feature selection would help in building more accurate models to predict the student performance

REFERENCE

1. Koedinger, K., Cunningham, K., Skogsholm A., Leber, B. (2008), "An open repository and analysis tools for fine-grained, longitudinal learner data", In 1st International Conference on Educational Data Mining, Montreal, 157-166.
2. Baker, R. Yacef, K. (2009), "The State of Educational Data Mining in 2009: A Review and Future Visions" Journal of Educational Data Mining (JEDM), 1(1), 3–17.
3. Cetintas, A., Si, L., Xin, Y.P., Hord, C. (2009), "Predicting correctness of problem solving from low-level log data in intelligent tutoring systems" In International Conference on Educational Data Mining, Cordoba, Spain, 230- 238.
4. Baker, R. S. J. D. (2010), "Data mining for education. International encyclopedia of education" B. McGaw, P. Peterson, and E. Baker, Eds., 3rd ed. Oxford, U.K.: Elsevier, 7, 112-118.
5. Kleesuwan, S., Mitatha, S., Yupapin, P. P., Piyatamrong, B. (2010), " Business Intelligence in Thailand Higher Educational Resources Management" Procedia-Social and Behavioral Sciences, 2(1), 84-87.
6. Espejo, P., Ventura, S., Herrera, F. (2010) "A Survey on the Application of Genetic Programming to Classification" IEEE Transactions on Systems, Man, and Cybernetics-Part C. 40, 2, 121-144.
7. Arsad, P. M., Buniyamin, N., Ab Manan, J. L., & Hamzah, N. (2011), "Proposed academic students' performance prediction model: A Malaysian case study" , In Engi neering Education (ICEED), 2011 3rd International Congress on (pp. 90-94). IEEE.
8. Connolly, T. M., Boyle, E. A., MacArthur, E., Hainey, T.,Boyle, J. M. (2012), " A systematic literature review of empirical evidence on computer games and serious games" Computers Education, 59(2), 661-686.
9. Aziz, A. A., Ismail, N. H., Ahmad, F. (2013)," Mining Students & Academic Performance. Journal of Theoretical & amp", Applied Information Technology, 53(3).

10. Papamitsiou, Z., & Economides, A. (2014), "Learning Analytics and Educational Data Mining in Practice: A Systematic Literature Review of Empirical Evidence" Educational Technology Society, 17 (4), 49–64.
11. Sin, K., Muthu, L. (2015), "Application of Big Data in Education Data Mining and Learning Analytics–A Literature Review", ICTACT Journal on soft computing, 5(4).
12. Khan, S., Shakil, K. A., Alam, M. (2016), "Educational Intelligence: Applying Cloud-based Big Data Analytics to the Indian Education Sector" Proceedings of International Conference on Contemporary Computing and Informatics 2016 (IC3I 2016).
13. Singh, R. and Pal, S., 2021, "A Critical Review on Educational Data Mining Segment: A New Perspective" Data Intelligence and Cognitive Informatics, pp.341-347.
14. Shetu, S.F., Saifuzzaman, M., Moon, N.N., Sultana, S. and Yousuf, R., 2021 "Student's Performance Prediction Using Data Mining Technique Depending on Overall Academic Status and Environmental Attributes" In International Conference on Innovative Computing and Communications (pp. 757-769). Springer, Singapore.
15. Khan, A. and Ghosh, S.K., 2020, "Student performance analysis and prediction in classroom learning: A review of educational data mining studies", Education and Information Technologies, pp.1-36.
16. Salloum, S.A., Alshurideh, M., Elnagar, A. and Shaalan, K., 2020, April, "Mining in educational data: review and future directions", In Joint European-US Workshop on Applications of Invariance in Computer Vision (pp. 92-102). Springer, Cham.
17. Sunday, K., Ocheja, P., Hussain, S., Oyelere, S., Samson, B. and Agbo, F., 2020 "Analyzing student performance in programming education using classification techniques", International Journal of Emerging Technologies in Learning (iJET), 15(2), pp.127-144.
18. Kunjumon, L.T., Shaji, S., Saji, S.T., Naushad, T. and Joseph, N., 2019, "An Intelligent System to predict Students academic performance using Data Mining", International Journal of Information, 8(2).
19. Sharma, P. and Sharma, S., 2018, "Data mining techniques for educational data: A review", International Journal of Engineering Technologies and Management Research, 5(2), pp.166-177.