A Comprehensive Survey on Educational Data Mining and Use of Data Mining Techniques for Improving Teaching and Predicting Student Performance

M. Anoopkumar

Dr. A. M. J. Md. Zubair Rahman

Abstract

Educational Data Mining (EDM) is an interdisciplinary ingenuous research area that handles the development of methods to explore data arising in a scholastic fields. Computational approaches used by EDM is to examine scholastic data in order to study educational questions. As a result, it provides intrinsic knowledge of teaching and learning process for effective education planning. This paper conducts a comprehensive study on the recent and relevant studies put through in this field to date. The study focuses on methods of analysing educational data to develop models for improving academic performances and improving institutional effectiveness. This paper accumulates and relegates literature, identifies consequential work and mediates it to computing educators and professional bodies. We identify research that gives well-fortified advice to amend edifying and invigorate the more impuissant segment students in the institution. The results of these studies give insight into techniques for ameliorating pedagogical process, presaging student performance, compare the precision of data mining algorithms, and demonstrate the maturity of open source implements.

Index Terms: Data Mining (DM), Educational data mining (EDM), Academic Performance, Student Performance, learning styles, teaching, models, EDM Tools, Data Mining Techniques, Prediction.

I. Introduction

Applying data mining technique in Inculcative setting is called as Educational Data Mining (EDM); and is a field that exploits statistical, machine-learning, and data-mining (DM) algorithms over the variants of edifying data. Its main objective is to analyse these types of data in order to resolve scholastic research issues [1]. EDM is concerned with
developing methods to explore the unique types of data in inculcative settings and, utilizing these methods, to better understand students and the settings in which they learn [21].

EDM has emerged as a research area in recent years for researchers all over the world from different and cognate research areas, which are as follows. 1) Offline education 2) E-learning and learning management system and 3) Intelligent tutoring system (ITS) and adaptive educational hypermedia system (AEHS) are an alternative to precedent ones and hence endeavouring to habituate edifying to the desiderata of each particular student.

This process does not differ much from other application areas of DM, like business, genetics, medicine, etc. However, it is consequential to note that in this paper, we are going to describe not only EDM studies that use typical DM techniques, such as association-rule mining, sequential mining, classification, text mining, clustering, etc., but withal discuss other approaches, such as regression, correlation, visualization, etc., which are not considered to be DM in a rigorous sense. Furthermore, some methodological innovations and trends in EDM are additionally considered.

EDM sanctions, to potentially amend some aspects of the quality of edification, and to lay the substratum for a more efficacious cognition process [54]. Although this area is flourishing there are few hurdles which pull the application of EDM from the expected growth such as

2. Different Types of Data: Relationships, Intrinsic semantic information, and Multi-Level Hierarchy
3. Techniques: Although most of the traditional DM techniques can be applied directly, others cannot and have to be habituated to the categorical educational problem at hand.

Nowadays, there is a great variety of scholastic systems/ environments such as: the traditional classroom, e-learning, LMS, adaptive hypermedia (AH) educational systems, ITS, tests/quizzes, texts/contents, and others such as: learning object (LO) repositories, concept maps, social networks, forums, scholastic game environments, virtual environments, ubiquitous computing environments, etc.

The main objective of EDM viewed by different Researchers as [5] [6]: 1. Student Modelling, 2. Domain Modelling, 3. Learning System, 4. Building the computational models, 5. Study the effects of resources.
The data can be personal or academic which can be used to understand students' behaviour, to assist instructors, to improve teaching, to evaluate and improve e-learning systems, to improve curriculums and many other benefits. Utilizing these Educational Data Mining techniques many kinds of erudition can be discovered such as classification, association rules, and clustering. The discovered knowledge can be used for organization of syllabus, prediction regarding enrolment of students in a particular programme, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students and so on.

In the authentic world, presaging the performance of the faculties is a challenging task. But the major challenge of higher education is the decrease in the success rate of faculties. An early prediction of faculties' performance may help the management to provide timely actions as well training to increase success rate. The curricula changes and the different styles in pedagogical process is also a big concern in the success rates of faculties. We discuss different parameters used in evaluating faculty performance to be used with different classification algorithms that predicts faculty performance. The results says[14] that can predict the result of the faculty and then it becomes feasible for taking necessary action. It can be proved auxiliary for academic organization and performance and magnification of students.

The possibilities for data mining in education and the data to be reaped are illimitable. Erudition discovered by Data Mining Techniques can be used not only to avail teachers to manage their classes, understand their students' learning processes, and reflect on their own edifying methods, but withal to fortify a learner's reflections on the situation and provide feedback to learners [5]. All of these are helpful to ensure the progress of students in their academics and enforce few remedies if the progress is infeasible to the programme and institutional expectation. The main advantage is that these kind of analysis avails to establish solution for slow learners.

The purport and objective of this survey paper is to review, different Data Mining Methods especially the mostly used and trendy algorithms applied to EDM context. Legion studies have been conducted in this context, but most of them with disparate methods and tools. The knowledge about these researched are the driving force towards new and fresh initiatives in EDM to ameliorate teaching and soothsaying Academic Performance of Students in educational institutions. This survey paper is to bridge this gap and present a comprehensive review of most of the types of Data Mining methodologies applied to EDM till date.

ISBN: 978-0-9948937-1-0

This survey is organized as follows. Section II lists the most recent and cognate works in education that have been resolved by utilizing DM techniques. Section III is simply engenders what precisely the Data Mining Process and its distinction with Educational Data Mining Process. Section-IV describes some of the most prominent Data Mining
Techniques applied to Education. Section-V discusses the users and tasks in Education and the Data Mining Tools for most of the kinds. At Section VI methods to perform faculty performance analysis are mentioned. During the final stages at Section VII the paper concentrate more on to explore the discussions and researches on the students' academic performance analyses and cognate research lines. Determinately, this paper identifies and suggests few research opportunities and future scopes in EDM at Section VIII and conclusions are outlined in Section IX.

II. Related Works

Computational overtures of EDM uses to analyse educational data with an objective to study edifying questions. Even though there are many works to date, conducted survey like this work, is to provide a comprehensive resource of papers published on Scholastic Data Mining (EDM) during 2005 to 2015 that utilises a categorical DM Technology.

Contributions from C. Romero, S. Ventura, [1] have the most germane studies in this field to date. This paper introduces EDM and describes the different groups of utilizer, types of scholastic environments, and the data they provide. Then it endeavours to expound the most typical/mundane tasks in the scholastic environment that have been resolved through data-mining techniques, and at the final stage of the paper, some of the most promising future lines of research are conveyed.

V. Sarala, Dr.V.V.Jaya Rama Krishnaiah discuss [2] the applications of data mining in educational institution to extract useful information from the huge data sets and providing analytical tool to view and use this information for decision making processes by taking real life examples.

A. Dutt, S. Aghabozrgi, M. A. Binti Ismail, and H. Mahroeian [3] advise the researchers with the methodology of how the puissance of massive amounts of didactic data of the organisations can utilized for the strategic purposes utilizing Data Mining Techniques. Meanwhile it simplifies the design of system which learns cognizance from data, utilizing sundry data mining approaches like clustering, classification, and prediction algorithms, this paper focuses to consolidate the variants of clustering algorithms as applied in the context of educational Data Mining.

Jo E. Hannay, E. Arisholm, H. Engvik, and Dag I.K. Sjøberg [4] utilize a pristine statistical method to conclude that effort shall be spent on elaborating on the effects of personality on sundry measures of collaboration, which, in turn, may be habituated to prognosticate and influence performance.
A. MERCERON and K. YACEF [5] establishes how Data Mining algorithms can pick pedagogically important mines contained in the data stores obtained from the educational system. These knowledge help to manage classes understand students and practice it on their teaching and to support learner reflection and provide proactive feedback to learners.

S. Akinola, B.O. Akinkunmi and T.S. Alo use [6] Data Mining methods to soothsay the performance of students in programming. Result of the study shows that a priori cognizance of Physics and Mathematics are essential in order for a student to stand out in Computer Programming. This work will be of considerable usefulness in identifying students in jeopardy early, especially in profoundly and immensely colossal classes, and sanction the instructor to provide opportune advising in a timely manner.

J. Kumar [7] helps to learn and develop models for the growth of education environment. It provides decision makers a better understanding of student learning and the environment setting in as of EDM.

A. Pears, S. Seidman, L. Malmi, and other successive contributors of this paper [8] accumulate and relegates germane research literature subsists across several disciplines including education and cognitive science, identifies paramount work and arbitrates it to computing educators and professional bodies with good proposal to computing academics teaching initiatory programming.

E. Lahtinen, K. Ala-Mutka, H. Järvinen [9] study the difficulties in learning programming in order to initiate development of learning materials for elementary programming courses. The survey provides information of the hurdles felt, experienced and perceived when learning and teaching programming.


C. Kelleher and R. Pausch [11] presents a taxonomy of languages and environments designed to make programming more accessible to neophyte programmers of all ages. The paper explicates all categories in the taxonomy, puts up a abbreviated description of the systems in each class, and suggests some avenues for future work in tyro programming environments and languages.

R. Jamili Oskouei, M. Askari [12] identified several factors affect the performance of students in different countries and used several classification and prediction algorithms
for improving the accuracy of predicting students' academic results before examination. The experimental results show that, factors such as gender, family background, parent's level of education, style of living have important effects on students' academic performance in both countries of which data collected.

S. Hussain, J. Hazarika, P. Buragohain, G.C. Hazarika [13] were using the data of the students enrolled in various affiliated institutions of Dibrugarh University, to explores the effect of performance on the basis of gender and caste. Yet another analysis was accomplished to examine the trends of performance with respect to time using ARIMA Model. The authors considered impact of some of the socio-demographic factors on the performance of the students during their investigation.

Miss P. R Shah, Prof. D. B Vaghela and Dr. P. SHARMA [14] found different parameters utilized in evaluating faculty performance to be utilized with different relegation algorithms that prognosticates faculty performance utilizing the advantages of Distributed Data Mining.

P.Geethalakshmi and Dhivy [15] considered the intricacy of students' experiences reflected from social media content and the emphasise that the growing scale of data demands automatic data analysis techniques developed a specific method to integrate both qualitative analysis and sizably voluminous-scale data mining techniques. The paper here focused on engineering students social media posts to understand issues and problems in their educational experiences.

M. Esteves, B. Fonseca, L. Morgado and P. Martins [16] conducted an action research approach to the analysis of how do the teaching and learning of programming at the university level could be developed within the Secondary Life Virtual world. Results appreciate the belief that it is possible to utilize this environment for better efficacy in the cognition of programming.

E. Osmanbegović, M. Suljić, [17] compared Different methods and techniques of data mining, during the presage of students' success, applying the data accumulated from the surveys conducted during the summer at the University of Tuzla,during academic year 2010-2011, among first year students and the data taken during the enrolment. The impact of students' socio- statistical variables, achieved results from high school and from the entrance exam, and postures towards studying which can have an effect on prosperity, were all investigated.

B. K. Baradwaj and S. Pal [18] suggested, the relegation task to evaluate student's performance and since there are many approaches that are utilized for classification of data and the decision tree method is utilized. They extracted knowledge that describes
performance of students’ in end semester examination. The study brings out earlier in recognizing the dropouts and students who require most specific attention and sanction the edifier to provide congruous advising/counselling.

Md. H. Islam Shovon and M. Haque [19] recommended a hybrid procedure based on Decision Tree method and Data Clustering of Data Mining that enables academicians to predict student’s GPA and based on that instructor can take adequate and appropriate moves to improve student curriculum performance. Especially the K-Means and Decision Tree algorithms.

D. A. AlHammadi and M. S. Aksoy [20] in this paper reviewed several applications of data mining in education and their benefits, presented some classification techniques, tested some sample data, and then evaluated them against some selected criteria.

M. Pandey and V. Kumar Sharma [21] have considered significant factors that may affect the performance of students and it is compared with four different decision tree algorithms J48, NBtree, Reptree and Simple cart. The result says that, J48 decision tree algorithm is found to be the best suitable algorithm for model construction. Then the Cross validation method and percentage split method were used to evaluate the efficiency of the different algorithms.

A. F. El Gamal [22] proposed an educational data mining model for predicting student performance in programming courses. The proposed model includes three phases; data pre-processing, attribute selection and rule extraction algorithm.

V. Ramesh, P. Parkavi and K. Ramar [23] adopted a survey cum experimental methodology was adopted to engender a database and it was formulated from a primary and a secondary origin. The obtained results from hypothesis testing suggested that nature of school does not have a impact on student performance and parents’ vocation plays a major role in presaging grades.

S. Borkar and K. Rajeswari [24] evaluated students’ performance and some attributes are selected which generate rules by means of association rule mining. A Multi-Layer Perceptron Neural Network is employed for selection of interesting features using 10 – fold cross validation. It is observed that in association rule mining important rules are generated using these selected attributes.

Sweta Rai and A. K. Jain [25] studied about students’ instant dropout or after first semester, of the undergraduate courses of computer science using the simple and intuitive classifiers (decision trees) ID3 and J48. The main reason recorded for dropout of students at this residential university were personal factors.
A Comprehensive Survey on Educational Data Mining and Use of Data Mining Techniques for Improving Teaching and Predicting Student Performance

A students' performance prediction system using Multi Agent Data Mining is proposed by Dr. A. AL-Malaise, Dr. A. Malibari and M. Alkhozae, [26] to predict the performance of the students based on their data with high precision of prognostication and provide an aid to the weaker students by optimization rules. The proposed system has been implemented and evaluated by looking into the prediction exactitude of Adaboost.M1 and LogitBoost ensemble classifiers methods and with the single classifier method C4.5. The results emphasis that using SAMME Boosting technique improves the prediction exactitude and surmounted C4.5 single classifier and LogitBoost.

L. Dole, and J. Rajurkar [27] proposed a Decision Support System which uses Naive Bayes algorithm (NB) approach to predict graduating CGPA (Cumulative Grade Point Average) based on applicant data collected from the studies conducted during the summer sessions at the University of Tuzla, the Faculty of Economics, academic year 2010-2011, among first year students and the data taken during the enrolment.

M. Jayakameswaraiah and S. Ramakrishna [28] conducted a Study on Prediction Performance of Some classification and clustering algorithms using 10 fold cross validation methods. And the result compared and shown.

T. Ranbaduge [29] mainly focused on the use of different data mining techniques upon the educational data to identify or excavate the important knowledge on student learning which can be used to evaluate the students overall performances in the e-learning open systems and discover and how these are been used to make out different learning patterns of the students.

S. Borkar and K. Rajeswari [30] suggested a method of evaluation of student's performance using association rule mining. Research work has been exercised on estimating student's performance based on various attributes such as Assignment, Attendance, Unit Test Performance, Graduation Percentage and University Result. The main objective of this is, prediction of student's performance in university result on the basis of their performance in assignment, unit test, graduation percentage and attendance.

Anwar M. A. and N. Ahmed [31] adopted a data mining approach applied to discover students’ performance models in supervised and unsupervised assessment tools of a course in an engineering degree program.

N. Thai-Nghe, A. Busche, and L. Schmidt-Thieme [32] nominated a method to tackle the class imbalance for amending the prediction/classification results by over-sampling techniques as well as utilizing cost-sensitive learning (CSL). The paper shows that the results have been ameliorated when comparing with only utilizing baseline
classifiers such as Decision Tree (DT), Bayesian Networks (BN), and Support Vector Machines (SVM) to the pristine data sets.

Syeda Farha Shazmeen, Mirza Mustafa Ali Baig, M.Reena Pawar [33] conducted performance evaluation of different data mining classification algorithm and predictive analysis. The study shows how the algorithms have been applied on different dataset to find out the efficiency of the algorithm and ameliorate the performance by implementing pre-processing techniques of data and feature selection and also prediction of new class labels.

A. Abdul Aziz, N. Hafieza Ismail And F. Ahmad [34] survey the three elements needed to make prognostication on Academic Performances of Students which are parameters, methods and implements. This paper withal proposes a framework for soothing the performance of first year bachelor students in computer science course. Naïve Bayes Classifier is utilized to extract patterns utilizing the Data Mining Weka tool. The framework can be utilized as a substratum for the system implementation and prognostication of Students’ Academic Performance in Higher Learning Institutions.

N. J. Coull and Ishbel M M [35] derives ten requisites that a fortification tool should have used in order to ameliorate CS1 student prosperity rate with reverence to learning and understanding.

Matthew Butler and Michael Morgan [36] describes an investigation into the nature of the academic quandaries that face abecedarian programming students. This quandary were exacerbated by the vogue of learners to study individually, outside the classroom or in online modes, which further reduces the choices available for quality feedback on issues of high-level. This paper analyses the results of a survey given to students enrolled in an initiatory programming unit across three campuses at Monash University in 2007. This denotes that many students may achieve a caliber of understanding sanctioning near transfer of domain cognizance but fail to reach a caliber of understanding that enables far transfer.

Peter J Mellalieu [37] constructed a mining based prototype Decision Support System (ReXS) to provide his students the provisions to predict their personal academic success and final grade as they progressed through a first-year course 'Innovation and Entrepreneurship'.

R. P. Bringula [38] could able to determine which of the sources of errors would presage the errors committed by tyro Java programmers. Delineate statistics brought out that the respondents perceived that they committed the identified eighteen errors infrequently.
Factor analysis showed that there were five categories for the types of errors committed. Four of them were symbol- or keyword-cognate errors such as, Invalid symbols or keywords, Mismatched symbols, Missing symbols, and Extortionate symbols, and the fifth one was Denominating-cognate error similar as Infelicitous designating error.

Ms. Sunita N. Nikam [39] conducted a general survey of Data Mining Applications at sundry sectors to amend the performance.

N. Thai Nghe, P. Janecek, and P. Haddawy [40] proposed a comparative study on the precision of Decision Tree and Bayesian Network algorithms for presaging the academic performance. In this analysis, the Decision Tree was consistently 3-12% more precise than the Bayesian Network. The results of these case studies give insight into techniques for accurately prognosticating student performance, compare the precision of data mining algorithms, and demonstrate the maturity of open source implements.

S. Sembiring, M. Zarlis, D. Hartama, Ramliana S, and E. Wani [41] have suggested kernel method as data excavation techniques to analyze the affinity of association between behavior and their prosperity of student's and to develop the model of student performance auspicates. This is done by utilizing Smooth Support Vector Machine (SSVM) classification and clustering techniques like kernel k-means. The results of this study have revealed a model of student performance predictors by employing factors of psychometry as variables predictors.

A.A. Aziz and N.H. I. Fadhilah Ahmad [42] have proposed Students' Academic Performance prognostication models for the first semester Bachelor of Computer Science from University Sultan ZainalAbidin by utilizing three culled classification methods; Rule Based, Decision Tree, and Naïve Bayes. The result discovers the race is a most influential factor to the students' performance followed by gender, total income of family, ingress mode of university, and hometown location parameters. The prognostication simulation can be habituated to classify the students so the lecturer can instantiate an early action to ameliorate students' performance.

A.D. Kumar and Dr. V. Radhika [43] gave an overview about the recent research papers on EDM and Student's academic performance in educational environment predicated upon the psychological and environmental factor is presaged by different educational data mining techniques.

All of the above papers have further studied based on the Technique which each of the paper utilised and the observation is listed in the Table 1.
Table 1: List of EDM References Grouped According to Types of Technique Used

<table>
<thead>
<tr>
<th>Technique</th>
<th>References</th>
<th>Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. EDM Fundamental</td>
<td>1,10,12,29</td>
<td>4</td>
</tr>
<tr>
<td>2. Classification (Associative, Bayesian Network, Decision Tree, Rule Based, NN Back Propagation, SVM, GA etc.)</td>
<td>1,5,14,15,17,18,19,20,21,22,23,25,26,27,28,30,32,33,34,35,40,41,42,</td>
<td>23</td>
</tr>
<tr>
<td>3. Clustering (Partitioning: K-Means, Hierarchical: Agglomerative, Model Based etc.)</td>
<td>1,3,14,15,19,28</td>
<td>6</td>
</tr>
<tr>
<td>4. Association Rule Mining</td>
<td>1,5,23,28,30,31</td>
<td>6</td>
</tr>
<tr>
<td>5. Sequential Mining (pattern etc.)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6. Text Mining (Keyboard based, Tagging approach, Information Extraction etc.)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7. Interactive Mining</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8. Temporal Mining</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9. Neural Network</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>10. Distributed Data Mining</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>11. Web Mining</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>12. Regression Analysis</td>
<td>3,37,38</td>
<td>3</td>
</tr>
<tr>
<td>13. Correlation Analysis</td>
<td>6,9,30</td>
<td>3</td>
</tr>
<tr>
<td>14. Statistical Methods</td>
<td>1,4,12,13,17,18,22,23,25,23532</td>
<td>10</td>
</tr>
<tr>
<td>15. Visualisation Analysis Methods</td>
<td>1,5,11,12,13,17,25,27,32,33</td>
<td>10</td>
</tr>
</tbody>
</table>

The trend in the EDM is analysed as follows (see Figure-1) and the analysis is carried out in the later Sections most precisely at the section future works.
III. Data Mining Process

The data mining defined as “the non-nugatory process of identifying valid, novel, anteriorly unknown, potentially utilizable information, and ultimately understandable patterns from data in database”[25][50]. The field of Data Mining (DM) is concerned with finding incipient patterns in astronomically immense amounts of data. DM is a technology used in different disciplines[39] to probe for consequential relationships among variables in astronomically immense data sets. DM is mainly utilized in commercial applications. Barahate S R. and Shelake V M verbally express [46] that now a days, the researcher have shown great interest in utilizing data mining applications in the field of education to efficiently manage and extract undiscovered knowledge from the data. Data mining has different steps [26] which are mentioned in shortly in Figure 2.

Figure 1: Educational Data Mining Technique- A Method Analysis

![EDMT Comprehensive Trend between 2005 and 2015](image-url)
Advances in Innovative Engineering and Technologies

From figure 2, it is evident that, data mining has the following steps [12] [29] [48]:

A. Data Amassing and Processing: in this step data collection is executed. Predicated on our goals, data can be amass from different environments or offices, such as banks, schools, markets, educational environments and etc. After accumulating data, a sole data warehouse may be providing for keeping this data for subsequent processing steps. The amassed data further will be processed and all faults, redundant data and etc., will be abstracted.

B. Data Transformation: in this step predicated on utilization or implements which we will utilize for analysing data, we require to transform data. For example for utilizing with Weka we require to make .csv files and etc.

C. Pattern Revelation: in this step with applying data mining techniques such as clustering, classification, and etc., we will be endeavour to discover pattern from that data.

D. Knowledge revelation: in this step we will be endeavour to utilize the extracted patterns for more examination or extracting association rules or further analyses.

E. Evaluation: in this step, with testing our extracted cognizance, the percentage of efficiency of that knowledge will be declared.

F. Action: Determinately with discovering all impuissance or efficiency of these extracted cognizance we can utilize these knowledge for sundry usages or applications.

Various algorithms[5] and techniques like Classification[42], Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules[29], Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for Knowledge discovery from databases [18] and also Distributed Data Mining [14].

Currently the data mining techniques has been used in various and withal in educational environments. Application of data mining technique in Educational setting is called as Educational Data Mining (EDM). The Educational Data Mining community website [47]...
defines educational data mining as follows: —Educational Data Mining is a growing discipline, occupied with developing methods for exploring the unique types of data that come from educational framework, and utilizing those methods to better understand students, and the settings which they learn in[17].

The key constituents of EDM are Stakeholders of Education, DM Methods, Tools and Techniques, data and task of Education and Outcomes which meet the Educational objectives[10].

The process of Educational data mining [7] is an iterative, Knowledge Discovery process which consists of Hypothesis formulation, Testing and refinement [4] (see Figure-3).

All those who take part in the educational process could gain by applying data mining on the data from the higher education system (Figure 4). Since data mining represents the computational data process from divergent perspectives, with the goal of excavating in-explicit and fascinating samples (Witten and Frank, 2000), trends and information from the data, it can greatly avail every participant in the scholastic process in order to amend the understanding of the teaching process, and it canters on discovering, detecting and expounding educational phenomenon’s (ElHalees, 2008)
So with DM the cycle of data mining techniques, are built in educational system which consists of forming hypotheses, testing and training, and hence its utilization can be directed to the various acts of the educational process in accordance with specific needs (Romero and Ventura, 2007, pp. 136):[49]

A. of students,
B. professors and
C. Administration and supporting administration.

Thus, usage of data mining in educational systems can be directed to support the specific needs of each of the participants [27] in the educational process. The student is required to recommend additional actions, teaching handouts/materials and tasks that would favour and ameliorate his/her learning. Professors could collect the feedback, possibilities to classify students into various groups based on their need for guidance and monitoring, to find the most frequent mistakes, find the efficacious actions, etc. Administration and administrative staff will receive the parameters that will ameliorate system performance.

IV. Educational Tasks and Data Mining Techniques

There are several applications or tasks in teaching environments that are resolved through DM. Baker [52], [53] suggests four key areas of application for EDM: ameliorating student models, amending domain models, finding out the education support provided by learning software system, and research project into learning and learners; and withal five approaches/methods: prognostication, clustering, relationship mining, distillation of knowledge for human judgment, and discovery with models.

C. romero, S. ventura, [1] have given the most pertinent studies in this field and it is summarised as follows:

A. Analysis and Visualization of Data - Statistics and visualization information are the two main techniques that have been most widely utilized for this task.
B. Providing Feedback for Fortifying Instructors - Several DM techniques have been utilized in this job, albeit association-rule mining has been the most mundane and reveals intriguing relationships among variables.
C. Recommendations for Students - Several DM techniques have been utilized for this chore, but the most mundane are association-rule mining, clustering, and sequential pattern mining.
D. Presaging Student’s Performance - Presage of a student's performance is one of the oldest and most popular applications of DM in education, and different techniques and models have been applied (neural networks, Bayesian networks, rule-based systems, regression, and correlation analysis)
E. Student Modelling - Different DM techniques and algorithms have been utilized for this job (mainly, Bayesian networks). Various DM algorithms (naïve Bayes, Bayes net, support vector machines, logistic regression, and decision trees) have been correlated to detect student noetic/mental models.

F. Detecting Undesirable Student Comportments - Several DM techniques (mainly, relegation, and clustering) have been used to reveal these types of students in order to provide them with felicitous helpin plenty of time.

G. Grouping Students - The DM techniques utilized in this chore are classification (meant for supervised learning) and clustering (meant for unsupervised learning).

H. Social Network Analysis - Different DM techniques have been used to mine convivial networks in educational environments, but collaborative filtering is the most mundane.

I. Developing Concept Maps - Few DM techniques (mainly, association rules, and text mining) have been used to construct concept maps.

J. Constructing Courseware - Different DM techniques and models have been used to build up courseware. The clustering of students and naïve algorithms have been proposed to construct personalized courseware by building a personalized Web tutor tree [51].

K. Orchestrating and Scheduling - Different DM techniques have been utilized for this chore (mainly, association rules).

There are many techniques in EDM to meet and achieve its objectives. Very few of them are categorised into three and are follows:

**Prediction:** This technique is utilized to derive predicted variable (single variable) from predictor variables (cumulation of variables). Prediction is utilized analyse student performance and drop out. P.V. Praveen Sundar (2013) and Dekker, G., Pechenizkiy, M., and Vleeshouwers J. (2009), for detecting student comportment. It is relegated into three types.

A. Classification: used to presage class label from (discrete or perpetuate). Some popular classification methods possess logistic regression, SVM and decision trees.

B. Regression: used to presage from continuous variable. Some of the famous regression methods of educational data mining include linear regression, neural networks.

C. Density Estimation: probability density function is utilized to predicted variable. Density estimator can be predicated on variety of kernel functions, considering Gaussian function also.

**Clustering:** Clustering is an unsupervised relegation process. It is utilized for grouping objects into classes of kindred objects. Data items are partitioned into groups or subsets (clusters) predicated on their neighbourhood and connectivity.
within N-dimensional space. In educational data mining, it uses clustering to group students according to their cognition.

**Relationship mining:** Relationship mining is utilized to determine relationship between variables in a data set and form rules for categorical purport. Relationship mining is relegated into four types:
A. **Association rule mining:** This method is utilized to identify relationship between attributes in data set, extracting intriguing correlations, frequent patterns among data items for finding students' mistakes most often recrudesce together while solving exercises.
B. **Correlation mining:** This method is utilized to find Linear correlations between variables (positive or Negative).
C. **Correlation analysis:** It is utilized to find the most vigorously correlation attributes.
D. **Sequential pattern mining:** This method is specifically utilized to find inter-session patterns such as the presence of a set of items followed by another items in a time-coherent set of sessions or episodes predicated on temporal relationship between variables to soothsay which group a learner belongs to. Wang et. al. proposes a four phase learning portfolio mining approach.
E. **Causal data mining:** This method is utilized to find causal relationship between variables by analysing the covariance of two events or by utilizing information about how one of the events was trigger.

Other Methods are
A. **Distillation of data for human judgment:** The objective of this method is to present data in summarize and visualized way for e.g. (3D graph etc.), to focus on congruous information and support decision making. In EDM it is utilized for identification and relegation.
B. **Discovery with models:** This type of model is utilized as component in other analysis such as relationship mining or prognostication.
C. **Knowledge Tracing:** This method is used to monitor student cognizance and skills over time. It is ineffective method in cognitive tutor system.

V. **Educational Data Mining Users and Tools**
Various people are involved with educational data mining of which there are four main users and stakeholders. These include:
A. Learners - Learners are interested in understanding student needs and methods to improve the learner's experience and performance.
B. Educators - Educators attempt to perceive the learning process and the methods they can use to improve their teaching methods. Educators can utilise extensibility of EDM to determine how to organize and structure the curriculum, the best methods to
present course information and the tools to use to engage their learners for optimal learning outcomes. In particular, the purification of data used for human judgment technique give an opportunity for educators to get benefit from EDM because it helps educators to quickly identify demeanour patterns, which can fortify their teaching methods during the tenure of the course or to improve future courses. Educators can decide on indicators that show student gratification and engagement of course material, and withal monitor learning progress.

C. Researchers - Researchers fixate on the development and the evaluation of data mining techniques for efficacy.

D. Administrators - Administrators are responsible for allocating the resources for implementation in institutions.

DM tools are normally designed more for power and flexibility than for simplicity [1]. Due to the rapid magnification of educational data, there is a desideratum to summarize the tools according to their function/features, integrated techniques and working platforms. EPRules, Gismo, TADAEs, O3R, Synergo/CoLAT, LISTEN Mining tool, MINEL, LOCO, CIECoF, PDinamet, Meerkat, MMT tool are examples of EDM tools, Agathe MERCERON and Kalina YACEF[5] used a range of tools. At the inception, worked with Excel and Access to perform simple SQL queries and visualisation. Then used Clementine for clustering and usual data mining platform for teachers, Tada-Ed, for clustering, classification and association rule (Clementine is more versatile and powerful but Tada-Ed has got pre-processing facilities and withal visualisation of results more tailored to our needs). Finally used SODAS to perform symbolic data analysis. Jo E. Hannay, Erik Arisholm, Harald Engvik, and Dag I.K. Sjøberg used Statistical tools to do all the performance analyses. Jasvinder Kumar [7] have given a vestige of WEKA, Moodle Tool, Rapid Miner, KEEL, TADA-ED and Decision Tools. R.Jindal and M. D. Borah [10] tried to explore a couple of few tools such as, Intelligent Miner (IBM), MSSQL Server 2005 (Microsoft), MineSet (SGI), Oracle DataMining (Oracle Corporation), SPSS Clementine (IBM), Enterprise Miner (SAS Institute ), Insightful Miner (Insightful Incorporation), CART (Salford Systems), TreeNet(R)(Salford Systems), RandomForests (Salford Systems), GeneSight (Inc. of El Segundo, CA), PolyAnalyst (Megaputer Intelligence), iData Analyzer (Microsoft), See5 and C5.0 (RuleQuest), TANAGRA (SPAD), SIPINA (Ricco Rakotomalala Lyon, France), ORANGE (University of Ljubljana, Slovenia a.), ALPHA MINER (E-Business Technology Institute), WEKA (University of Waikato, New Zealand), Carrot etc.

Arnold Pears, Stephen Seidman, Lauri Malmi [8] have proposed about

1. Visualization tools – ITICSE, Jeliot, jGRASP, JHAV’E and MatrixPro, DDD, Tango, Polka and ALVIS Live

2. Automated judgement/assessment tools - CourseMarker, BOSS, WebCAT, TRAKLA2
4. Other tools – Jplag, MOSS, and YAP


Apart from these, many more tools can be searched out by conducting a more in depth survey on Data Mining Tools and their context.

VI. Faculty Performance Analysis

Considerably less number of works have found in this area and the ones in which studies have incorporated are includes here in favour of the methods they have used.

- A. MERCERON and K. YACEF [5] establishes how Data Mining algorithms can pick pedagogically paramount mines contained in the data stores obtained from the educational system. These erudition help to manage classes understand students and contemplate it on their teaching and to support learner reflection and provide proactive feedback to learners.
- A. Pears, S. Seidman, L. Malmi, L. Mannila, and other successors [8] accumulate and classifies pertinent research literature subsists across several disciplines including education and cognitive science, identifies consequential work and arbitrates it to computing educators and professional bodies with good advice to computing academics teaching prelusive programming.
- Miss P. R Shah, Prof. D. B Vaghela and Dr. P. SHARMA [14] found different parameters utilized in evaluating faculty performance to be utilized with different classification algorithms that prognosticates faculty performance utilizing the advantages of Distributed Data Mining.
- M. Esteves, Be. Fonseca, L. Morgado and P. Martins [16] carried out an action research advance to the analysis of how teaching and learning of computer programming at the university level could be created within the Second Life virtual world. Results support the notion that it is possible to utilize this environment for better efficacy in the cognition of programming.

Besides these papers less contribution have found in this comprehensive survey. It is to be handled in the next researches.
VII. Student Performance Evaluation Methods

Various evaluation methods and factors used for the same are given below. Many of the papers cover few of common practices and hence it is classified into five category of papers in the section VIII.

- C. romero, S. ventura, [1] have given the most germane studies in this field to date. This paper introduces EDM and describes types of educational environments, and the data they provide. Then it endeavors to explicate the utilization of regression, classification, various neural networks such as back-propagation and feed-forward, Bayesian networks, rule-based systems, Bayesian networks, correlation analysis, and regression, so as to predict performance evaluation in educational data mining, and withal Naïve Bayes, Bayes net, SVM, logistic regression and decision trees are additionally discovered to simulate and analyse student noetic models. At the later stage, besides some of these, various types of clustering, association-rule mining, human reliability analysis and Markov chain analysis are withal mentioned as implements to discover student comportment. Conclusively, few of the prominent future lines of research are conveyed.

- A. Dutt, S. Aghabozrgi, M. A. Binti Ismail, and H. Mahroeian [3] advise the researchers with the methodology of how the puissance of massive amounts of didactic data of the organisations can utilized for the strategic purposes utilizing Data Mining Techniques. Meanwhile it simplifies the design of system which learns cognizance from data, utilizing sundry data mining approaches like clustering, classification, and prediction algorithms, this paper focuses to consolidate the variants of clustering algorithms: ANN, K-means, Hierarchical Clustering, Simple k-Means and X-Means, C-Means clustering, Markov Clustering, UCAM (Unique Clustering with Affinity Measure), Two Phase Clustering (TPC), Ward’s Hierarchical clustering along with corresponding data-set used, as applied in the context of educational Data Mining.

- Jo E. Hannay, E. Arisholm, H. Engvik, and Dag I.K. Sjøberg [4] utilize a pristine statistical method like univariate and multivariate analysis to conclude that effort shall be spent on elaborating on the effects of personality on sundry measures of collaboration, which, in turn, may be habituated to prognosticate and influence performance.

- S. Akinola, B.O. Akinkunmi and T.S. Alo use [6] Data Mining methods to soothsay the performance of students in programming. A specific Multi-Layer Feed-Forward Back Propagation Neural Network was used and result of the study shows that a priori cognizance of Physics and Mathematics are essential in order for a student to stand out in Computer Programming. The work provided usefulness in identifying students in jeopardy early, especially in profoundly and immensely colossal classes, and sanction the instructor to provide chance to advising in a timely manner.

- E. Lahtinen, K. Ala-Mutka, H. Järvinen [9] study the difficulties in learning programming in order to initiate development of learning materials for elementary programming courses. The survey using questionnaire and correlation analysis provides information of the hurdles felt, experienced and perceived when learning and teaching
programming. And a careful attention is required to design materials and approaches used in teaching programming and skill development in students.

C. Kelleher and R. Pausch [11] presents a taxonomy of languages and environments designed to make programming more accessible to neophyte programmers of all ages. The paper explicates all categories in the taxonomy, puts up an abbreviated description of the systems in each class, and suggests some avenues (to reduce unnecessary syntax, native language closure, alternate programming, etc...) for future work in tyro programming environments and languages.

S. Hussain, J. Hazarika, P. Buragohain, G.C. Hazarika [13] were using the data of the students enrolled in various affiliated institutions of Dibrugarh University, to explore the effect of performance on the basis of gender and caste. Yet another analysis was accomplished to examine the trends of performance with respect to time using ARIMA Model. The authors considered impact of some of the socio-demographic factors on the performance of the students during their investigation.

P. Geethalakshmi and Dhivy [15] considered the complexity of students' experiences reflected from social media content and the emphasise that the growing scale of data demands automatic data analysis techniques developed a workflow to incorporate both qualitative analysis and data mining techniques on large-scale. The paper here focused on engineering students social media posts to understand issues and problems in their educational experiences by implementing a multi-label classification algorithm.

M. Esteves, B. Fonseca, L. Morgado and P. Martins [16] conducted an action research approach to the analysis of how do the teaching and learning of programming at the university level could be developed within the Secondary Life Virtual world. Results appreciate the belief that it is possible to utilize virtual environment with interactive learning for better efficacy in the cognitive programming.

E. Osmanbegovic, M. Suljic, [17] compared different methods and techniques of data mining, during the presage of students' success, applying the data accumulated from the surveys conducted during the summer at the University of Tuzla. Classifiers such as NB, MLP and J48 upon Chi-square, One R, Info Gain and Gain Ratio test are conducted. The impact of students' socio-statistical variables, achieved results from high school and from the entrance exam, and postures towards studying which can have an effect on prosperity, were all investigated.

D. A. AlHammadi and M. S. Aksoy [20] in this paper reviewed several applications of data mining algorithms in education and their benefits, presented some classification techniques, tested some sample data, and then evaluated them against some selected criteria by excluding K-Means as it has shown lesser accuracy.

A. F. El Gamal [22] proposed an educational data mining model for predicting student performance in programming courses. The proposed model includes three phases; data pre-processing, attribute selection and rule extraction algorithm (decision tree)
and conveys the impact of programming aptitude and mathematical skills on programming performance.

- V. Ramesh, P. Parkavi and K. Ramar [23] adopted a survey cum experimental methodology was adopted to engender a database and it was formulated from a primary and a secondary origin. Utilised the implementation of Naïve Bayes, SMO, and Multi-Layer Perception and J48 algorithms and found MLP as a best performer of these Algorithms. The obtained results from hypothesis testing suggested that nature of school does not have an impact on student performance and parents' vocation plays a major role in presaging grades.

- S. Borkar and K. Rajeswari [24] evaluated students' performance by means of association rule mining. A Multi-Layer Perceptron Neural Network is utilised for selection of interesting features using 10 – fold cross validation. It is observed that in association rule mining important rules generated using these selected attributes and correctly classifies when apriori is applied to it.

- Sweta Rai and A. K. Jain [25] studied about students' instant dropout or after first semester, of the undergraduate courses of computer science using the simple and intuitive classifiers (decision trees) ID3 and J48. And j48 is rated as successful over ID3 when classification efficiency is considered.

- A students' performance prediction system using Multi Agent Data Mining is proposed by Dr. A. AL-Malaise, Dr. A. Malibari and M. Alhozae, [26] to predict the performance of the students based on their data with precision of Multi Agent Data Mining prognostication and provide an aid to the weaker students by optimization rules. The proposed system has been implemented and evaluated by looking into the prediction exactitude of Adaboost.M1 and LogitBoost ensemble classifiers methods and with the single classifier method C4.5. The results emphasis that using SAMME Boosting technique improves the prediction exactitude and surmounted C4.5 single classifier and LogitBoost.

- L. Dole, and J. Rajurkar [27] proposed a Decision Support System which uses Naïve Bayes (NB) approach to predict graduating CGPA (Cumulative Grade Point Average) based on applicant data collected from the studies conducted during the summer sessions at the University of Tuzla, the, academic year 2010-2011, among first year students and the data taken during the enrolment.

- T. Ranbaduge [29] mainly focused on the use of different data mining techniques such as KNN, and classifiers like Rule-based, Decision Trees, Bayesian and instance-based learner classifiers upon the educational data to identify or excavate the important knowledge on student learning which can be used to evaluate the students overall performances in the e-learning open systems. And also discover and how these are been used to make out different learning patterns of the students.

- S. Borkar and K. Rajeswari [30] suggested a method of evaluation of student's performance using association rule mining. Research work has been exercised on
estimating student’s performance based on various attributes such as Assignment, Attendance, Unit Test Performance, Graduation Percentage and University Result.

- N. J. Coull and Ishbel M M [35] derives ten requisites that a fortification tool genetic algorithms should have used in order to ameliorate CS student prosperity rate with reverence to learning and understanding.

- Matthew Butler and Michael Morgan [36] describes an investigation into the nature of the academic quandaries that face abecedarian programming students. This quandary were exacerbated by the vogue of learners to study individually, from outside the normal classes or in online open access modes, which further reduces the culls available for quality feedback on issues of high-level. This denotes that many students may achieve a caliber of understanding sanctioning near transfer of domain cognizance but fail to reach a caliber of understanding that enables far transfer.

- Peter J Mellalieu [37] constructed a mining based feature rich prototype Decision Support System (ReXS) with spreadsheet mining to provide his students the provisions to predict their personal academic success and final grade as they progressed through a first-year course 'Innovation and Entrepreneurship'.

- R. P. Bringula [38] could able to determine which of the sources of errors would presage the errors committed by tyro Java programmers. Factor analysis showed that there were five categories for the types of errors committed. Four of them were symbol- or keyword-cognate errors such as, Invalid symbols or keywords, Mismatched symbols, Missing symbols, and Extortionate symbols, and the fifth one was Denominating-cognate error similar as Infelicitous designating error.

- N. Thai Nghe, P. Janecek, and P. Haddawy [40] proposed a comparative study on the precision of Decision Tree and Bayesian Network algorithms for presaging the academic performance. In this analysis, the Decision Tree was consistently 3-12% more precise than the Bayesian Network. The results of these case studies give insight into techniques for accurately prognosticating student performance, compare the precision of data mining algorithms, and demonstrate the maturity of open source implements.

- A.A.Aziz and N.H. I.Fadhilah Ahmad [42] have Proposed Students' Academic Performance prognostication models for the first semester Bachelor of Computer Science from University Sultan ZainalAbidin by utilizing three culled classification methods; Rule Based, Decision Tree, and Naive Bayes. The result discovers the race is a most influential factor to the students’ performance followed by gender, total income of family, ingress mode of university, and hometown location parameters. The prognostication simulation can be habituated to classify the students so the lecturer can instantiate an early action to ameliorate students’ performance.

- A.D. Kumar and Dr.V.Radhika [43] gave an over view about the recent research papers on EDM and Student's academic performance in educational environment predicted
upon the psychological and environmental factor is presaged by different educational data mining techniques and the impacts were analysed.

VIII. Future Work And Research Pools
The information given in the Table 2 is all about the contribution of the reference papers into the contextual design of this comprehensive study.

| Table 2: List of EDM selected references based on the design of the paper. |
|-----------------------------------------------|-----------------|---|
| I. DM PROCESS | 3, 5, 7,10,12,14,17,18,25, 26,27,29,31,39,42 | 15 |
| II. EDUCATIONAL DATA MINING TECHNIQUES | 1,2,4,6,7,10,12,14, 18,20,26,28,29, 32,33,34,38,40,41, 42 | 20 |
| III. EDUCATIONAL MINING USERS, TASKS AND DATA MINING TOOLS | 1,2, 4, 5, 7, 8,10,11,13,14,23,34, 35,37,41 | 15 |
| IV. FACULTY PERFORMANCE ANALYSIS | 5,8,14,16 | 4 |
| V. STUDENT PERFORMANCE EVALUATION METHODS | 1,3,4,6,9,11,13,15,16,17,20,22,23, 24,25,26,27,29,30,35,36,37,38,40, 42 | 25 |

As per the perception from the Figure 1 and Table 2, the trend of EDM is running around few Methodologies such as classification (very few algorithms), statistical and visualisation methodology. Even though it is limited to these area, ample number of papers, reviews and surveys are available to address these. Even though many of the papers are concentrated on distinct issues, only few papers have been evolved in some specific areas. Though it is smaller in amount or the other in larger in numbers, problems still exist, which are yet to be addressed. It is found that the relation of Pedagogical factors on the Students’ Performance Analysis or identification of the co-existence of concurrent factors and their analysis have scopes in the future researches.

Besides these, the studies related to open source e-learning platforms [55] with regards to performance, support, security, flexibility, easy of using, interoperability, administration tools, management, communication tool, content development and course delivery of tools says that Moodle, Caroline, mambo, and A-tutor systems deliver the best results. And it is as followed in Table 3:
Table 3: List of Open Source E-learning tools and their best of features.

<table>
<thead>
<tr>
<th>No</th>
<th>Open E-Learning Systems</th>
<th>Best of feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Moodle</td>
<td>security, performance, Support, interoperability, flexibility, communication tool and course delivery tools</td>
</tr>
<tr>
<td>2</td>
<td>Caroline</td>
<td>Easy of Using</td>
</tr>
<tr>
<td>3</td>
<td>Mambo</td>
<td>Management</td>
</tr>
<tr>
<td>4</td>
<td>A-tutor</td>
<td>Administration Tools and Content Development</td>
</tr>
</tbody>
</table>

The paper withal says that Moodle is considerable the best and if integrate this four platform it produces a total weight of 97.72 effectiveness while the best Open Source Moodle 1.9 is 89.4. Hence the studies towards difficulties and effectiveness of these integration is laying a future scope in educational data mining using e-learnings systems.

IX. Conclusion

The goal of this paper has been to give a comprehensive survey towards the research papers which would have discussed different Data Mining Methods especially the mostly used and trendy algorithms applied to EDM context. These survey are very helpful for achieving good overview of educational data mining methods and tools which is used presently to bring about improvements in teaching and predicting the performance of Students so as to predict Academic Performance in Learning Programming.

References

1. Cristobel Romer, Sebastian Ventura "educational data mining: a review of the state of the art" ieee transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 40, no. 6, november 2010
2. v. Sarala, Dr. v.v. Jaya Rama krishnaiah, "empirical study of data mining techniques in education system", international journal of advances in computer science and technology (ijacst), vol. 4 no.1, pages: 15 – 21 2015
A Comprehensive Survey on Educational Data Mining and Use of Data Mining Techniques for Improving Teaching and Predicting Student Performance


11. CAITLIN KELLEHER, and RANDY PAUSCH, "Lowering the Barriers to Programming: a survey of programming environments and languages for novice programmers", ACM 2003

12. Rozita Jamili Oskouei, Mohsen Askari, "Predicting Academic Performance with Applying Data Mining Techniques (Generalizing the results of two Different Case Studies)", Computer Engineering and Applications Journal 2014.


15. P. Geethalakshmi and S. Dhivy, "Mining Social Media Data for Understanding Students Learning Experiences", International Journal of Advances in Engineering, 2015, 1(3), 373 - 377, ISSN: 2394-9260 (printed version); ISSN: 2394-9279 (online version)


35. NATALIE J. COULL, Ishbel M M Duncan, "Emergent Requirements For Supporting Introductory Programming", ITALICS 2011

36. Matthew Butler and Michael Morgan, "Learning challenges faced by novice programming students studying high level and low feedback concepts", Proceedings asclite Singapore 2007


42. Azwa Abdul Aziz, Nor Hafieza Ismail and Fadhilah Ahmad, "First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms", International Conference on Artificial Intelligence and Computer Science (AICS 2014), September 2014
47. www.educationaldatamining.org.