Data Mining in Education: State of the Art and Proposal

Suleiman Khalifa Arafa
University of Al Neelain
Khartoum, Sudan
Email: suliemankhalifa [AT] yahoo.com

Dr. Mahmoud Ali Ahmed
Under Secretary of the University of Khartoum
Khartoum, Sudan
Email: mali [AT] uofk.edu

Abstract— **Data Mining (DM) is a new promising technology that has been successfully applied to education to address many educational issues. In recent years, there are an increasing numbers of researches that interest in using data mining in education system to improve education and facilitate education research. Institutions of higher education (IHE) are at the core of educational systems in which extensive research and development is performed in a competitive environment. Applying DM in education is an emerging interdisciplinary research field known as educational data mining (EDM). It concerns on the development of methods for exploring the unique types of data that come from educational environments, so as to develop better ways to handle data and support future decision making. Furthermore, those methods are used to better understand students and settings what the students learn in. In this paper, we aim to present a review of the works in which DM techniques were used to predict students' performance based on significant criteria, in order to find suitable DM methods to fulfill the gaps in existing works. Then, considering the result of this review we suggest a work methodology to obtain an optimal predictive model for predicting students' performance. The rest of this paper is organized as follows: Section I offer an introduction. Section II reviews previous related studies that focused on EDM applied for predicting student's performance, and Section III points out the proposed research methodology, data collection and preparation procedures. Finally, conclusions are drawn in Section V.**

KeyWords: Data Mining, Educational Data Mining, Institutions of Higher Education, Predicting Students Performance

I. INTRODUCTION

Data mining (DM) has a wide range of applications in different areas, including marketing, banking, telecommunications, scientific discovery, and educational research [1]. The education system is one of these domains in which the primary concern is the evaluation and, in turn, enhancement of educational organizations. The use of DM application has recently expanded for a variety of educational purposes. The assessment of students' needs, students' dropout rate, and performance are some important emerging data mining applications in education.

Today, Institutions of higher education (IHE) are actively collecting and storing big educational data. The explosive growing of data is currently stored in educational databases that contain various hidden information that can help to improve the academic performance of students. Predicting students' performance becomes more challenging due to the lack of existing system to deeply analyze and predict student's performance. Currently, there is an increasing desire from the IHE towards predicting student's performance to successfully preparing their future [2]. This would allow the IHE to take corrective measures to assist students at-risk and to enhance their performance. One way how to predict student's performance is by discovering knowledge from educational data. DM is one step at the core of the knowledge discovery process. Data mining is thus used to study available data and extract the hidden information. Data mining is the process of extracting the hidden information from a given database and is therefore a valuable tool for converting data into usable information. This information can be used in several educational processes such as predicting student's performance and estimating student dropout rate [3].

Recently, DM techniques have been extensively applied to find interesting patterns, build descriptive and predictive models from large volumes of data accumulated through the use of different educational systems [4] [5]. The results of DM can be used for getting a better understanding of the underlying educational processes, for generating recommendations and advices to students, for improving resource management.

A. Educational Data Mining

DM in the field of education known as Educational Data Mining (EDM) is a new growing research area. Furthermore, EDM can be defined as the application of data mining techniques on raw data that come from educational systems to respond to the educational questions and problems, and also to discover the information hidden after this data [6] [8]. Over the last few years, the popularity of this field enhanced a large number of research studies that is difficult to surround and to identify the contribution of data mining techniques in educational systems. In fact, understanding the raw data collected from educational systems can be ''a gold mine'' to help the designers and the users of these systems improving their performance and extracting useful information on the behaviors of students in the learning process. EDM concerns with developing, researching, and applying computerized methods to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the enormous volume of data within which they exist [7]. Its goal is to better understand how students learn and identify the settings in which they learn to improve educational outcomes. EDM can be applied [9] to assess students' learning performance, to improve the learning process and guide students' learning, to provide feedback and adapt learning recommendations based on students' learning behaviors, to evaluate learning materials and courseware, to detect abnormal learning behaviors and problems, and to achieve a deeper understanding of educational phenomena. Also EDM is useful in identifying at-risk students, identifying priority learning needs for different groups of students, increasing graduation rates, effectively assessing institutional performance and optimizing subject curriculum renewal.

B. EDM Objectives

In the last several years, EDM has been applied to address a wide number of goals that are all parts of the general objective of improving learning [12]. Several studies [10] [11] [12] [13] [14] address a list of these objectives. Romero and Ventura [12] proposed to classify EDM objectives depending on the viewpoint of the final user (Students, Educator, Administrator, and Researcher) as shown in Fig 1:

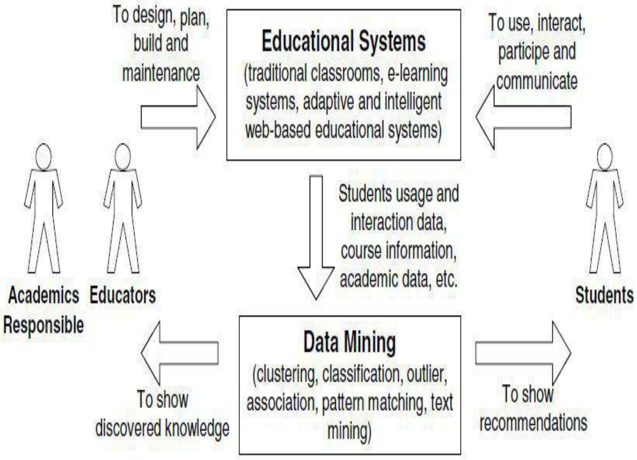

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Students usage and interaction data, course information, academic data, etc.
Academics Responsible
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure 1. The Cycle of Applying Data Mining in Educational Systems [5]

- Learners. The goal is to support a learner's reflections on the situation, to provide adaptive feedback or recommendations to learners, to respond to student's needs, to improve learning performance, etc.
- Educators. They need to understand their students' learning processes and reflect on their own teaching methods, to improve teaching performance, to understand social, cognitive and behavioral aspects, etc.
- Administrators. The objective is to evaluate the best way to organize institutional resources including human and material.
- The above EDM goals clearly shows the benefit of EDM applications to the end user, but it is difficult to classify all EDM application goals according to these four actors, especially when an objective is related to more than one actor. That is why, based on the work of [10] [14] [15] [16] that focused on the related research goal of EDM applications, we distinguish between the following EDM general goals:
- Student Modeling. The common objective here is to create or improve a student model from usage information. User modeling in the educational domain incorporates such detailed information as students' characteristics or states such as knowledge, skills, motivation, satisfaction, meta-cognition, attitudes, experiences and learning progress, or certain types of problems that negatively impact their learning outcomes.
- Predicting Students' Performance and Learning Outcomes. The objective is to predict a student's final grades outcomes based on data from course activities.
- Generating recommendation. The objective of EDM in this context is to recommend to students, which task is the most appropriate for them at the current time.
- Analyzing learner's behavior. The objective is to group student according to their profile, and for adaptation and personalization purposes.
- Communicating to stakeholders. The objective is to help course administrators and educators in analyzing students' activities and usage information in courses.
- Domain structure analysis. The objective is to determine domain structure and improving domain models that characterize the content to be learned and optimal instructional sequences, using the ability to predict the student's performance as a quality measure of a domain structure model.
- Advancing scientific knowledge about learning and learners through building, discovering or improving models of the student, the domain, and the pedagogical support.

C. EDM Methods

EDM methods are drawn from a variety of literatures, including data mining and machine learning, psychometrics and other areas of statistics, information visualization, and computational modeling. In EDM method, predictive modeling is usually used in predicting student performance. In order to build the predictive modeling, there are several tasks used. The most popular classification of these methods is the one proposed in Baker [9]: prediction, clustering, relationship mining, distillation for human judgment and discovery with models. Bienkowski et al. [13] then Romero and Ventura [15] extended this these techniques into the following methods:

- Prediction. The goal is to develop a model which can suppose a single predicted variable from some combination of other predictor variables. Types of predictions methods are classification when the predicted variable is a categorical value, regression when the predicted variable is a continuous value or density estimation when the predicted value is a probability density function.

- Clustering. The process of finding instances that naturally group together and can be used to split a full dataset into categories. Once a set of clusters has been determined, new instances can be classified by determining the closest cluster.
- Relationship mining. Used for discovering relationships between variables in a dataset. There are different types of relationship in mining techniques such as association rule mining (any relationships between variables), sequential pattern mining (temporal associations between variables), correlation mining (linear correlations between variables), and causal data mining (causal relationships between variables).
- Distillation of data for human judgment. It is a technique that involves depicting data in a way that enables a human to quickly identify or classify features of the data. This approach uses summarization, visualization and interactive interfaces to highlight useful information and support decision-making.
- Discovery with models. Its goal is to use a validated model of a phenomenon as a component in further analysis such as prediction or relationship mining.
- Outlier Detection. The goal of outlier detection is to discover data points that are significantly different than the rest of data.

D. *Process of Applying EDM*

Romero,Ventura , Sachin and Vijay proposed a process of applying EDM as in Fig 2.

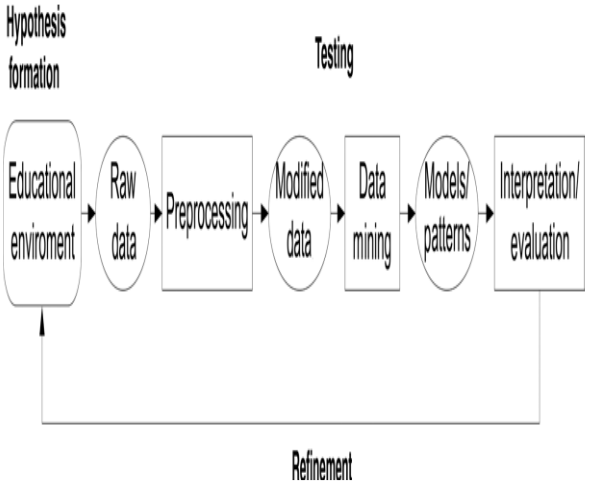


Figure 2. Overview of how EDM methods are applied [16] [17]

The process starts with collecting the data to study from the educational environment. Next, the obtained raw data require cleaning and preprocessing for the purpose of treatment of missing and incorrect values, converting the data to an appropriate form, feature selection, etc. in this phase, the process requires the use of some data mining techniques. Once the data preprocessed, the appropriate EDM method/technique is applied to identify any latent valuable nuggets of information. Many techniques have been used to perform the common data mining activities of associations, clustering, classification, modeling, sequential patterns, and time series forecasting. Finally, the last step is the interpretation and the assessment of the obtained results to see if additional domain knowledge was discovered and to determine the relative importance of the facts generated by the mining algorithms.

II. Review of the related Works

Many works have been proposed, as well, different approaches or/and methods have been applied for predicting students performance. However, in all these researches, accuracy of prediction is key and what researchers look forward to try and improve. In this section, we present the reviewed studies to which data mining methods are applied in predicting student performance. Most of the studies were gathered from the conferences, and journals publications as stated in Table 1.

TABLE I. Studies that focused on Educational Data Mining

Author (year)(Ref#)	**Target**	**Predictor**	**EDM Methods & Techniques**
Lakshmipriya &Arunesh (2017) (20)	Loss Of Academic Status	Academic &other data	Classification: Decision tree, SVM, Apriori Naïve Beyes
Singh,Bhalla (2016) (21)	Pass/Fail Examination	Academic data	Classification: Decision Tree
Mueen Zafar & Manzoor (2016) (22)	Course Grade	Academic data	Classification: Naïve Bayes, Neural Network, Decision Tree
Abu Saa (2016) (23)	GPA	Academic &other data	Classification: decision tree
Muluken (2015) (24)	Academic Failure/Success .	Academic & other data	Classification: Decision tree, NaiveBayes
Fadhilah, Hafieza , Azwa(2015) (25)	First Year's GPA	Academic & other data	Classification: Decision Tree, Naïve Bayes, Rule Based
Asif, Merceron, & Pathan (2015) (26)	Graduation GPA	Academic & other data	Classification: Decision Trees, K-NN, Naive Bayes, Neural Networks.
Ogwoka, Cheruiyot,& Okeyo(2015) (27)	Final Grade	Academic data	Classification: Decision tree Clustering: K-means

Abu-Oda M..ElHalees (2015) (28)	Pass/Fail Semester	Academic data	Classification: Decision Tree, Naive Bayes
Al-Barrak & Al Razgan (2015) (29)	Course Grade	Academic data	Classification: Naïve Bayes, JRip rule-based, decision tree
AhmetTekin (2014) (30)	Graduation GPA	Academic data	Classification: Neuralnetworks, SVM, ELM
Badr,Elaraby (2014) (31)	Final Grade	Academic data	Classification: Decision Tree (ID3)
Azwa,Hafieza, Fadhilah (2014) (32)	First Year's GPA	Academic &otherdata	Classification: Decision Tree, Naïve Bayes, Rule Based
Osmanbegović, Agić, & Suljić, (2014.) (33)	Final Grade	Academic &otherdata	Classification: Decision Tree (J48), Random Forest, Naive Bayes, Multilayer Perceptron
Tiwari,Singh & Vimal(2013) (34)	Final Grade	Academic &otherdata	association rules, classification rules, Clustering: K-means
Bhise, Thorat, Supekar (2013) (35)	Final Grade	Academic data	Clustering: K-means
Kabakchiea (2 013) (36)	Final Grade	Academic Data & Others	Classification: decision tree, Naïve Bayes, k-Nearest Neighbour
Sajadin, Zarlis, Dedy, Ramliana ElviWani (2011) (37)	Final Grade	Academic &otherdata	Classification: SVM Clustering: k-means

Brief Discussion on each Study

In India, Lakshmipriya. K, Dr. Arunesh P.K [18] presented a paper where they applied data classification methods for data mining techniques to predict the performance of college students. For the case studies, real data set for the college students was collected. Several techniques were applied and analyzed using R Studio programming Open Source Tool. This study aims to identify those students which needed special attention to reduce fail ratio and taking necessary action for the future career.

Sohajbir Singh Ubha, Gaganpreet Kaur Bhalla in [19] Applied decision tree algorithms to predict the results of students in Senior Secondary Schools in the state of Punjab. For the subject of the study, they used real data of 838 students. The results of the experiments, helps in identifying the weak students and those students, whose chances of failure in the examination are high. The study helps schools to concentrate more on weak students in order to minimize the failure rate and drop out ratio so that the overall performance of the school can be improved.

Another paper presented by Ahmed Mueen, Bassam Zafar, and Umar Manzoor [20] used data mining techniques to predict and analyze students' academic performance based on their academic record and forum participation. For that purpose, datasets of undergraduate students who had taken the Programming Fundamental and Advanced Operating System courses from August 2014 to May 2015 have been used. For the study purpose, three classification algorithms (Naïve Bayes, Neural Network, and Decision Tree) have been applied. The prediction performance of three classifiers are measured and compared. It was observed that Naïve Bayes classifier outperforms other two classifiers by achieving overall prediction accuracy of 86%.This study helps teachers to early detect students who are expected to fail to provide special attention and help them to improve their performance.

In another study done by Amjad Abu Saa [21] multiple data mining tasks used to create qualitative predictive models to predict the students' grades. For that purpose, a survey was constructed that has targeted university students and collected multiple personal, social, and academic data. Meanwhile, he applied four decision trees as well as, the Naïve Bayes algorithms to identify the factors assumed to affect students'.The study found that the student's performance is not totally dependent on their academic efforts, in spite; there are many other factors that influences as well. In conclusion, this study can motivate and help universities to perform data mining tasks on their students' data regularly to find out interesting results and patterns which can help both the university as well as the students.

Muluken Alemu Yehuala [22] gave a case study to investigate the possible application of data mining in the Ethiopian higher education at Debre Markos University. For the subject of the study, a regular undergraduate student data set consists of 11,873 records was used. Then, classification technique based on the decision tree and Bayes was applied to predict the students' performance. The research findings indicated pre-college result, Sex, Number of students in a class, number of courses given in a semester, and field of study are the major factors affecting the student performances. The study offered a helpful and constructive recommendations to the academic planners in universities of learning to enhance their decision making process. Also helps Students to decide about their field of study before they are enrolled in specific field of study.

Among other studies, Fadhilah Ahmad, Nur Hafieza Ismail and Azwa Abdul Aziz [23] conducts a comparative analysis of three classification techniques; DT, NB, and RB for predicting students' academic performance of first year bachelor students in Computer Science course. For the study purpose, data were collected from 8 year period range between July 2006/2007 until July 2013/2014 that contains the students' demographics, previous academic records, and family background information. The experimental result shows that the RB has the best classification accuracy compared to NB and DT. However, the limitation of this study is the small size of data due to incomplete and missing value in the collected data. In conclusion, this study helps the lecturers to take early actions to help and assist the poor and average category students to improve their performance.

Raheela Asif, Agathe Merceron, Mahmood K. Pathan [24] presents a case study on predicting performance of students at the end of a university degree at an early stage of the degree program. For the subject of the study, Datasets comprising 347 undergraduate students' pre-admission data and the examination scores of the courses of first and second academic years have been used to predict the students' overall performance at the end of the degree. To predict the graduation performance, several data mining classification algorithms have been used like decision trees, rule induction, 1-nearest neighbor, naive Bayes and neural networks. The overall results show that it is possible to predict the graduation performance in 4th year at university using only pre-university marks and marks of 1st and 2nd year courses, no socio-economic or demographic features, with a reasonable accuracy. In conclusion, this study helps universities to identify students with low academic performance and find ways to support them.

In Malaysia, Amirah Mohamed Shahiria, Wahidah Husaina, Nur'aini Abdul Rashida [25] provided a systematical reviews on the data mining techniques that have been used to predict students' performance. This study has reviewed previous studies on predicting students' performance with various analytical methods. It results that Most of the researchers have used cumulative grade point average (CGPA) and internal assessment as data sets. While for prediction techniques, the classification method is frequently used in educational data mining area. Under the classification techniques, Neural Network and Decision Tree are the two methods highly used by the researchers for predicting student's performance. In conclusion, this study will help the educational system to monitor the students' performance in a systematic way.

In Kenya, a paper presented by Thaddeus Matundura Ogwoka, Wilson Cheruiyot and George Okeyo[26] applied Decision tree and K-means data mining algorithms to create a model for predicting students performance. For the subject of the study, they have considered dataset of undergraduate students pursuing Bachelor of Science in Technology (BTIT), Bachelor science in Information Technology (BSIT) both government sponsored and self-sponsored, part-time and full-time students at the department of computer science and information technology (CSIT) in the Technical university of Mombasa. Next, they evaluated the experiments conducted using Decision tree using J48 and K-means. In conclusion, this study helps the University management in predicting student performance in advance in order to devise ways of assisting weak students and even make more decisions on how to select students for particular courses.

A paper presented by Ghadeer S. Abu-Oda and Alaa M. El-Halees[27] focused on examining and predicting students' dropouts through their university programs. For the case study, they selected a total of 1290 records of computer science students Graduated from ALAQSA University between 2005 and 2011. The collected data included student study history and transcript for courses taught in the first two years of computer science major in addition to student GPA , high school average , and class label of (yes ,No) to indicate whether the student graduated from the chosen major or not. To classify and predict dropout students, different classifiers have been trained on data sets. The overall performance shows excellent performance in predicting students' dropouts. In conclusion, The study found that digital design and algorithm analysis courses has a great affect on predicting student persistence in the major and decrease student likelihood of dropout.

In Saudi Arabia, A paper presented by Mashael A. Al-Barrak, Mona S. Al-Razgan [28] provides a case study to analyze students' grades in different evaluative assignments for a course on data structures. They compare three different classifiers using real female students' grades data in a one year period from King Saud University to predict students' performances. They apply classification techniques to both numerical and categorized attributes. The overall experimental results showed that the model based on the Naïve Bayes algorithm provides the most accurate predictions. The main objective of this study was to find the best classifier to predict students' performance obtained on the final exam in the course Data Structures.

Ahmet Tekin [29] applied several prediction techniques in data mining to assist educational institutions to predict their students' GPAs at graduation. For the case studies, he used a sample consisted of 127 student. For the study purpose, three prediction methods—NN, SVM, and ELM were used to estimate student GPAs at graduation. This study has two scenarios. The first scenario was designed to estimate the GPAs at graduation of students according to their GPAs of coursework completed during their first 2 years of study. In the second scenario, the GPAs after the first 3 years of coursework were used as data. Moreover, for all methods, results indicated that GPAs after the first 3 years of coursework demonstrated improved predictions for all methods. This study helps educational institutions earlier to predict students GPAs If students are predicted to have low GPAs at graduation, then extra efforts can be made to improve their academic performance and, in turn, GPAs.

Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby [30] applied classification task to predict the final grade of students. They collect the data set from a student's database session 2005 to 2010 with initially size of 1548 records. This study will help the student's to improve the student's performance, as well as, helps the administration to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time.

Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby [30] applied classification task to predict the final grade of students. They collect the data set from a student's database session 2005 to 2010 with initially size of 1548 records. This study will help the student's to improve the student's performance, as well as, helps the administration to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time.

Another similar study from Azwa Abdul Aziz, Nor Hafieza Ismail and Fadhilah Ahmad [31] Attempts to develop Students' Academic Performance prediction models for the first semester Bachelor of Computer Science from University Sultan ZainalAbidin. They applied three selected classification methods; Naïve Bayes, Rule Based, and Decision Tree. To conduct this study, five independent parameters (gender, race, hometown, family income, university entry mode) have been selected. The result reveals the models of Rule Based and Decision Tree algorithm gives the highest prediction accuracy value of 68.8%. However, the model failed to predict the poor students. This study helps the university to classify the students so the lecturer can take an early action to improve students' performance.

Edin Osmanbegović, Mirza Suljić and Hariz Agić [32] conducted a research on data mining in education with the aim of emphasizing the data mining possibilities that can be of significant help during monitoring, decision-making and management in education. For the subject purpose of the study, a sample of 907 students from secondary schools located in Tuzla Canton (Bosnia and Herzegovina) school year 2011/12 and 2012/13 used to predict students final grade by applying and comparing four data mining algorithms, Decision Tree (J48), Random Forest, Naive Bayes and Multilayer Perceptron. Results gained by the selected algorithms for classification of performances of students in secondary school, indicate that prediction rate varies between 65-75%. This study helps school management decision makers at all levels to make decisions to improve the quality of education and student performance.

A paper by Mahendra Tiwari , Randhir Singh, and Neeraj Vimal [33] provided an experiential Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education. For that purpose, they collected the student data of B.Tech second year (CS & IT branch) from database management system course held at the United College of Engineering and Research Naini Allahabad (Affiliated to GBTU) in fourth semester of 2011/2012 and they used questionnaire to collect the real data that describing the relationships between learning behavior of students and their academic performance. In conclusion, this study helps higher education to improve their students' academic performance.

Bhise R.B., Thorat S.S., Supekar A.K. in [34] applied clustering task using K-Means to predict students' performance. The data set used in this study was obtained from department of Bachelor of computer Application (B.C.A.), B.J. College, Ale, in Nov-2012. This study will help the teacher to reduced drop-out ratio to a significant level and improve the performance of students.

A paper presented by Dorina Kabakchieva [35] applied different data mining algorithms for classification on the university sample dataset contains 10330 instances. For the study purpose, , well known data mining algorithms, including two rule learners, a decision tree classifier, two popular Bayes classifiers and a Nearest Neighbor classifier were used. The prediction performance of three classifiers are measured and compared. The results achieved indicate that the prediction rates are not remarkable (vary between 52-67 %). The main objective of this study is to implement data mining project at the University of National and World Economy. This study helps the University management to focus more on the profile of admitted students at the entry point of the university to decide about their field of study before they are enrolled in specific field of study.

III. Research Proposal and Methodology

In this section, we describe the proposed research methodology. The methodology starts from the data collection, then preprocessing which are discussed in the introduction and the data set and preprocessing sections, then the data mining methods which are classification, clustering, association, followed by the evaluation of results and patterns, finally the knowledge representation process as shown in Fig. 3.

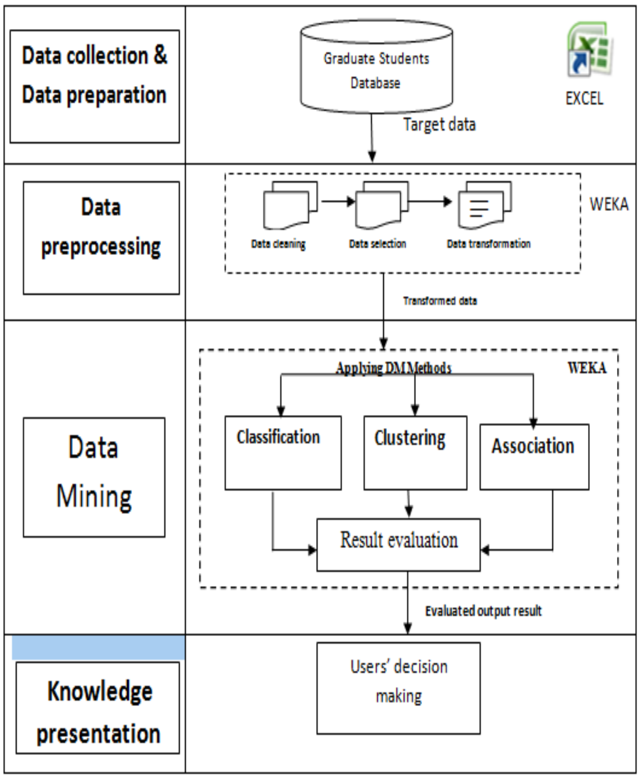


Figure 3. Proposed Research Methodology

The detail explanation about the data collection and data preparation will be described next.

A. Data Source for Proposed Work

The dataset used for the purpose of the research is obtained from a set of graduate students in the program of Computer Science and Information Technology from Comboni College of Science and Technology in Sudan from the period from 2005 till 2017. The college grants their graduates a bachelor degree and diploma in three specialties, including one specialty for bachelor degree in Computer Science, and two specialties for diploma which are Information technology (Business Administration, Accounting). Table 2 presents the attributes and their description that exists in the dataset .

- Grade Assessment policy

  Exams were evaluated out of 100 points possible. Course scores were calculated according to the contribution of the mid-term exam (40%) and the Final exam (60%). The mark out of 100 is divided by 20 to get the points as follows depending on the academic affairs in the college: A=4.0 and more; B= 3.5 up to 3.9; C= 3.0 up to 3.4; D= 2.5 up to 2.9; < 2.5 means Failure: F.

  Different subjects have different weights according to their relevance. That is translated into the credit hours of the subject. To get the result of a semester, the credit hours of each subject are calculated according to the points scored, and an average of all is made (GPA or Grade point Average).

- Graduation criteria’s

  A student is considered graduated after the decision of the academic Council at the end of the 4 year’s course for the Bachelor degree or the 3 year’s course for the Diploma program. A student must have passed in all subjects of the program and got a cumulative Grade point Average (cGPA) of at least 2.50 at the end of studies depending on the academic affairs in the college as follows:

- Grade Point Average (GPA)

  It is the average of the student’s score that is rounded up to only two decimal numbers as follows:

$$\text{GPA} = \text{Sum of}\ \frac{\text{(scores of every subject} \times \text{credit hours)}}{\text{Sum of the class credit hours}}$$

- Cumulative Grade Point Average (CGPA)

  It is the average of the student’s scores during the semesters and rounded up to two decimal numbers:

$$\text{CGPA} = \frac{\text{sum of (every studied subject} \times \text{its credit hours}}{\text{Sum of the credit hours of the studied subjects}}$$

B. Data Collection

For the study purpose, we collected a sample containing approximately 1800 graduate students’ records found in a spreadsheet format. During this phase, data collection process for Graduate students at CCST is studied, in order to select the appropriate dataset to work with. The rules and procedures for collecting data about examination results of the college students are also reviewed. We used students’ demographic profile, namely the gender, race, and previous school, and the CGPA obtained on all semesters of graduate students’ as the variables for the student data. Data related to the students’ college admission has been collected through the enrolment form filled by the students while admitting in the college from the Registrar office. Data related to the students’ grades achieved at the exams on the different subjects including (course code, course name, credits hours, grade obtained (0 to 5), grade point average (GPA) of a semester, and the cumulative grade point average (cGPA) has been collected through the academic affair office.

C. Data Preparation

Data pre-processing allows the original data to be transformed into a suitable shape to be used by a particular data mining algorithm or framework. So, before applying a data mining algorithm, a number of general data pre-processing tasks can be addressed (data cleaning, data transformation, data integration, and data reduction) The data being collected contains different information starting from pre-university education (high school grades, type of the certificate), demographics data, the CGPA obtained in all semesters throughout the different study phases in the college. The data is also being studied for missing values, and for obvious mistakes, are corrected. The provided data is subjected to many transformations.

Next, we perform a discretization of numerical values in order to increase the interpretation and comprehensibility. Discretization divides the data in categorical classes that are easier to understand (categorical values are more familiar than precise magnitudes and ranges). In this case, we have used the manual method for the mark attribute. The following steps are performed as part of the preparation and preprocessing of the data set depending on the academic affairs in the college as follows:

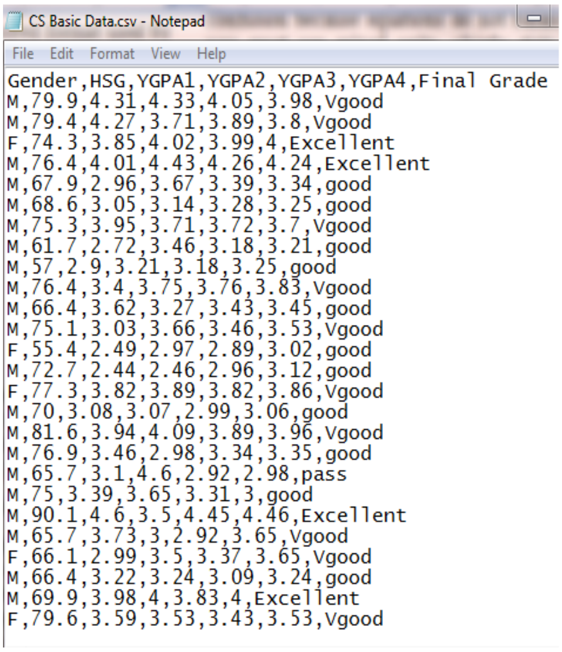
The CGPAs are classified into four intervals, namely, “Excellent” if the CGPA obtained is >=4 and <=5, “V.Good” if the CGPA obtained is >=3.5 and <4, “Good” if the CGPA obtained of the student is >=3 and <3.5, “Pass” if the CGPA obtained is>=2.5 and <3.

Also, all numerical values for the mark attribute in higher secondary school (HSS) labeled as “Excellent” if the mark obtained is >=90 and <=100, “V.Good” if the mark obtained is >=80 and <90, and “Good” if the student mark obtained is >=70 and <80, and “Average” if the student mark obtained is >=60 and <70, and “Poor” if the mark obtained is>=50. Furthermore, Table II describes the attributes of the data and their possible values.

TABLE II. ATTRIBUTES DESCRIPTION AND POSSIBLE VALUES

Attribute	Description	Possible value
Student id	The student ID	The student ID
Student name	The name of the student	Student Full name
Gender	The Sex of Student's	Male or female
Country	The original country for the student	Sudan, South Sudan, Eritrea, Ethiopia
HSS mark	Higher Secondary School mark obtained	Excellent, V.good, Good, Average, poor
HSS certificate	Type Of HSS Certificate for college entrance	Science, art, others
Specialty	The specialty of study of the Student	CS,IT(Business Administration, Accounting)
Qualification		Bachelor degree, Diploma
YGPA1	The Cumulative Grade Point Average obtained for the first year	Excellent, V.good , Good, pass
YGPA2	The Cumulative Grade Point Average obtained for the second year	Excellent, V.good , Good, pass
YGPA3	The Cumulative Grade Point Average obtained for the third year	Excellent, V.good , Good, pass
YGPA4	The Cumulative Grade Point Average obtained for the fourth year	Excellent, V.good , Good, pass
Graduation final grade	The graduation grade for the student obtained	Excellent, V.good, Good, pass

Finally, we transform the data to the required format used by the data mining algorithm. We show here the way how it can be transformed into the CSV format which is appropriate for data mining as stated in Fig. 5.



```
Gender,HSG,YGPA1,YGPA2,YGPA3,YGPA4,Final Grade
M,79.9,4.31,4.33,4.05,3.98,Vgood
M,79.4,4.27,3.71,3.89,3.8,Vgood
F,74.3,3.85,4.02,3.99,4,Excellent
M,76.4,4.01,4.43,4.26,4.24,Excellent
M,67.9,2.96,3.67,3.39,3.34,good
M,68.6,3.05,3.14,3.28,3.25,good
M,75.3,3.95,3.71,3.72,3.7,Vgood
M,61.7,2.72,3.46,3.18,3.21,good
M,57,2.9,3.21,3.18,3.25,good
M,76.4,3.4,3.75,3.76,3.83,Vgood
M,66.4,3.62,3.27,3.43,3.45,good
M,75.1,3.03,3.66,3.46,3.53,Vgood
F,55.4,2.49,2.97,2.89,3.02,good
M,72.7,2.44,2.46,2.96,3.12,good
F,77.3,3.82,3.89,3.82,3.86,Vgood
M,70,3.08,3.07,2.99,3.06,good
M,81.6,3.94,4.09,3.89,3.96,Vgood
M,76.9,3.46,2.98,3.34,3.35,good
M,65.7,3.1,4.6,2.92,2.98,pass
M,75,3.39,3.65,3.31,3,good
M,90.1,4.6,3.5,4.45,4.46,Excellent
M,65.7,3.73,3,2.92,3.65,Vgood
F,66.1,2.99,3.5,3.37,3.65,Vgood
M,66.4,3.22,3.24,3.09,3.24,good
M,69.9,3.98,4,3.83,4,Excellent
F,79.6,3.59,3.53,3.43,3.53,Vgood
```

Figure 4. Sample of CSV Graduate Student's data

IV. CONCLUSION

Educational data mining (EDM) is an area full of exciting opportunities for many researchers. A key area of EDM is to predict students' performance in order to recommend the improvement ways on the current educational practice. Therefore, predicting student's performance continues to be one of the most popular goals for EDM. Nowadays, Institutions of higher education (IHE) are very focused on their students' academic performance. Predicting students' performance is useful to help instructors and students improving their learning and teaching process. In this paper, we present comprehensive survey among the previous studies involving several works tackling the application of data mining in education for predicting students' performance. The EDM methods and techniques that we have explored in this paper, when properly applied, can achieve high rates of success. In particular, from the analysis of the papers we have reviewed, we conclude that the researchers have been able to predict academic performance using various factors such as graduation GPA, course grade, pass/fail determinations as Target variables. Predictor variables: most of the researchers used academic and non-academic data as Predictor variables. While the EDM methods: Classification is by far the most popular EDM method used to predict students' performance, other methods such as clustering, association rule mining have also been successfully tried.

Currently, we are finishing the data collection and data preparation that support the proposed research work. Data collection and data preparation is very complex and difficult process which takes more time and effort to prepare the needed data. Very soon, our future work will be focused on how we could develop the proposed methodology by using DM classification technique for predicting the graduation grade for the graduate students.

ACKNOWLEDGEMENTS

This work is supported by Comboni College of Science and Technology (CCST). We would like to thank CCST for providing the data

REFERENCES

[1] Han, J. Kamber, M. (2008). Data Mining: concepts and techniques. 2nd Edition, Morgan Kaufmann publishers.

[2] Baker, R.S.J.d.: Data Mining for Education. In: McGaw, B., Peterson, P., Baker, E. (eds.) To appear in International Encyclopedia of Education, 3rd edn. Elsevier, Oxford (2010)

[3] Yukselturk, E., Ozekes, S., Turel, Y. K. (2014). Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program, European Journal of Open, Distance and e-Learning, 17(1), 118-133.

[4] Romero, C., Ventura, S., and Garcia, E., Data mining in course management systems: MOODLE case study and tutorial. Comput. Educ. 51, 368–384, 2008.

[5] Romero, C. and S. Ventura, S., Educational data mining: A survey from 1995 to 2005. Expert Syst. Appl. 33(1), 135–146, 2007.

[6] Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. In 3rd Information Systems International Conference, 2015 (Vol. 72, pp. 414–422). Shenzhen: Elsevier. http://doi.org/10.1016/j.procs.2015.12.157

[7] Romero C, Ventura S, Pechenizky M, Baker R. Handbook of Educational Data Mining. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Chapman and Hall/CRC Press; 2010.

[8] Romero, C., & Ventura, S. (2013). Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(1), 12–27. http://doi.org/10.1002/widm.1075

[9] He, W.: Examining students' online interaction in a live video streaming environment using data mining and text mining. Comput. Hum. Behav. 29(1), 90–102 (2013)

[10] Baker, R.S.J.d.: Data mining for education. In: McGaw, B., Peterson, P., Baker, E. (eds.) International Encyclopedia of Education, vol. 7, 3rd edn., pp. 112–118. Elsevier, Amsterdam (2010)

[11] Calders, T., Pechenizkiy, M.: Introduction to the special section on educational data mining. ACM SIGKDD Explor. 13(2), 3–6 (2011)

[12] Romero, C., Ventura, S.: Data mining in education. Wiley Interdisc. Rev.: Data Min. Knowl. Discovery 3(1), 12–27 (2013)

[13] Bienkowski, M., Feng, M., Means, B.: Enhancing teaching and learning through educational data mining and learning analytics: an issue brief. US Department of Education, Office of Educational Technology, pp. 1–57 (2012)

[14] Scheuer, O., McLaren, B.M.: Educational data mining. In: Seel, N.M. (eds.) Encyclopedia of the Sciences of Learning, pp. 1075–1079. Springer, US (2012)

[15] Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d.: Introduction. In: Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (eds.) Handbook of Educational Data Mining, Chapman and Hall/CRC Data Mining and Knowledge Discovery Series, pp. 1–5. CRC Press, Boca Raton (2011)

[16] Romero, C., Ventura, S.: Educational data mining: a review of the state of the art. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 40(6), 601–618 (2010)

[17] Sachin, B.R., Vijay, S.M.: A survey and future vision of data mining in educational field. In: Proceedings of IEEE 2nd International Conference on Advanced Computing and Communication Technologies, pp. 96–100. ACM, New York (2012)

[18] Lakshmipriya. K, Dr. Arunesh P.K. "Predicting Student Performance Using Data Mining Classification Techniques". International Journal of innovation Research in Science and Engineering Vol. No.3, Issue 02, February 2017

[19] Sohajbir Singh Ubha, Gaganpreet Kaur Bhalla. "Data Mining for Prediction of Students' Performance in the Secondary Schools of the State of Punjab". International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 8, August 2016

[20] Ahmed Mueen, Bassam Zafar, and Umar Manzoor. "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques". I.J. Modern Education and Computer Science, 2016, 11, 36-42

[21] Amjad Abu Saa. "Educational Data Mining & Students' Performance Prediction". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016

[22] Muluken Alemu Yehuala. "Application of Data Mining Techniques for Student Success and Failure Prediction: The Case of Debre_Markos University". International Journal Of Scientific & Technology Research Volume 4, Issue 04, April 2015.

[23] Fadhilah Ahmad, Nur Hafieza Ismail and Azwa Abdul Aziz. "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques". Applied Mathematical Sciences, Vol. 9, 2015, no. 129, 6415 – 6426

[24] Raheela Asif, Agathe Merceron, Mahmood K. Pathan. " Predicting Student Academic Performance at Degree Level: A Case Study" I.J. Intelligent Systems and Applications, 2015, 01, 49-61

[25] Amirah Mohamed Shahiria, Wahidah Husaina, Nur'aini Abdul Rashida. "A Review on Predicting Student's Performance using Data Mining Techniques". Procedia Computer Science 72 (2015) 414 – 422 Available online at www.sciencedirect.com

[26] Thaddeus Matundura Ogwoka, Wilson Cheruiyot and George Okeyo. "A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms". International Journal of Computer Applications Technology and Research Volume 4– Issue 9, 693 - 697, 2015, ISSN: 2319–8656

[27] Ghadeer S. Abu-Oda and Alaa M. El-Halees. "Data Mining In Higher Education: University Student Dropout Case Study". International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[28] Mashael A. Al-Barrak, Mona S. Al-Razgan. "Predicting Students' Performance through Classification: A Case Study". Journal of Theoretical and Applied Information Technology 20th May 2015. Vol.75. No.2

[29] Ahmet TEKIN. "Early Prediction of Students' Grade Point Averages at Graduation: A Data Mining Approach" Eurasian Journal of Educational Research, Issue 54, 2014, 207-226

[30] Abeer Badr El Din Ahmed, Ibrahim Sayed Elaraby. "Data Mining: A prediction for Student's Performance Using Classification Method". World Journal of Computer Application and Technology 2(2): 43-47, 2014

[31] Azwa Abdul Aziz, Nor Hafieza IsmailandFadhilah Ahmad. "First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms". Proceeding of the International Conference on Artificial Intelligence and Computer Science (AICS 2014), 15 - 16 September 2014, Bandung, INDONESIA. (E-ISBN978-967-11768-8-7).

[32] Edin Osmanbegović, Mirza Suljić, and Hariz Agić2. "Determining Dominant Factor for Students Performance Prediction by using Data Mining Classification Algorithms". Vitez-Tuzla-Zagreb-Beograd-Bucharest, Vol. XVII, July - December 2014

[33] Mahendra Tiwari, Randhir Singh, and Neeraj Vimal. "An Empirical Study of Applications of Data Mining Techniques for Predicting Student Performance in Higher Education". IJCSMC, Vol. 2, Issue. 2, February 2013, pg.53 – 57

[34] Bhise R.B., Thorat S.S., Supekar A.K. "Importance of Data Mining in Higher Education System". IOSR Journal of Humanities and Social Science (IOSR-JHSS) ISSN: 2279-0837, ISBN: 2279-0845. Volume 6, Issue 6 (Jan. - Feb. 2013), PP 18-21

[35] Dorina Kabakchieva. "Predicting Student Performance by Using Data Mining Methods for Classification". CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume 13, No. 1, 2013

[36] Sajadin Sembiring, M. Zarlis, Dedy Hartama, Ramliana S, and Elvi Wani. "Prediction of Student Academic Performance by an Application of Data Mining Techniques". 2011 International Conference on Management and Artificial Intelligence IPEDR vol.6 (2011) © (2011) IACSIT Press, Bali, Indonesia Graduation: A Data Mining Approach". Eurasian Journal of Educational Research, Issue 54, 2014, 207-226