B.Sc. in Computer Science and Engineering Thesis

Parametric Study of Student Learning in IT Using Data Mining

Submitted by

Reshad Reza
201214015

Rubyeat Islam
201214027

Shakil Zaman
201114031

Supervised by

Dr. Syed Akhter Hossain

Professor and Head of the Department

Department of Computer Science and Engineering

Daffodil International University (DIU)

Dhaka, Bangladesh

Department of Computer Science and Engineering
Military Institute of Science and Technology

December 2015

CERTIFICATION

This thesis paper titled **"Parametric Study of Student Learning in IT Using Data Mining"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in December 2015.

Group Members:

Reshad Reza

Rubyeat Islam

Shakil Zaman

Supervisor:

Dr. Syed Akhter Hossain
Professor and Head of the Department
Department of Computer Science and Engineering
Daffodil International University (DIU)
Dhaka, Bangladesh

CANDIDATES' DECLARATION

This is to certify that the work presented in this thesis paper, titled, "Parametric Study of Student Learning in IT Using Data Mining", is the outcome of the investigation and research carried out by the following students under the supervision of Dr. Syed Akhter Hossain, Professor and Head of the Department, Department of Computer Science and Engineering, Daffodil International University (DIU), Dhaka, Bangladesh.

It is also declared that neither this thesis paper nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Reshad Reza
201214015

Rubyeat Islam
201214027

Shakil Zaman
201114031

ACKNOWLEDGEMENT

We have been blessed by the Almighty for the successful completion of our thesis with ease. We are highly obliged and profoundly indebted to our supervisor, Dr. Syed Akhter Hossain, Professor and Head of the Department, Department of Computer Science and Engineering, Daffodil International University (DIU), Dhaka, Bangladesh, for his constant dedication, guidance and encouragement throughout the tenure. His deep interest on the topic and valuable opinions throughout the study was incredibly helpful for the completion of our thesis.

We are obliged to the Department of Computer Science and Engineering (CSE) of Military Institute of Science and Technology (MIST) for providing their enduring support during our thesis work. We are also grateful to rest of the departments of MIST who have helped us during the survey conduction.

Finally, we would like to thank our friends and admirers for their appreciable support, patience, instructions and suggestions during the course of our thesis.

Dhaka
December 2015

Reshad Reza

Rubyeat Islam

Shakil Zaman

ABSTRACT

In the present context of digitalization, the role of information technology is dramatic. The growth of connected society today is leading a new dimension of living. This brings data mining as the agent for predictable computation. The higher educational institutions are constantly starving for excellence in education and research where student plays the central role. It is very important to determine the factors that affects quality of learning of students at the university. In predicting Students' Academic Performance (SAP) model, Data Mining plays a great role in educational field. Data Mining used in educational field is called Educational Data Mining (EDM). The objective of this research is to examine students' learning using Information Technology (IT) by implementing Data Mining. At first a survey was designed and conducted for the students of Military Institute of Science and Technology (MIST) with the help of the authority of MIST. The surveys' questionnaires have been divided into five learning fields: 'About English Competency', 'About Learning Process', 'About Class Items Interaction', 'About Communication and Interaction with Teachers and with Classmates'. In order to complete the thesis, three classification algorithms: Decision Tree (J48), Naive Bayes, Rule Based (OneR) and one clustering algorithm: K-Means clustering are applied on the data set. The dependent parameter of this data set is Cumulative Grade Point Average (CGPA) of the student. As a tool WEKA 3.7 is used to apply the algorithms on the finalized data set. For the highest accuracy, the comparison of classification and clustering is measured. Besides the comparison of three classifications models has been completed to find out the most efficient classification model for predicting SAP. The outcome of the experiment declares that classification gives better accuracy than clustering and error proficiency is so much less comparing to clustering. The result also reveals, the 'Rule Based (OneR)' model is the best model to give the highest accuracy value comparing to Decision Tree (J48) and Naive Bayes. Rule Based gives the highest accuracy for the excellent and good students but fails to predict the accuracy for the average students. While Decision Tree model gives the accuracy value for excellent, good and average students but its' efficiency is less than the Rule Based model. This prediction model is really useful to determine the efficiency of Information Technology in students' academic skill and can be implemented to upgrade their performance. In the future, this research can have extended mining covering higher spectrum of students' learning.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

List of Algorithms

LIST OF ABBREVIATION

DM : Data Mining
SAP : Student Academic Performance
EDM : Educational Data Mining
KDD : Knowledge Discovery Database
IHL : Institutions of Higher Learning
WEKA : Waikato Environment for Knowledge Analysis
IT : Information Technology
ARFF : Attribute Relation File Format
CSV : Comma Separated Value
CGPA : Cumulative Grade Point Average

LIST OF SYMBOLS

D	: Training Data Set
T	: Testing Data Set

CHAPTER 1
INTRODUCTION

Data Mining (DM) is the convenient way used for data analyzing process. It is the combination of machine learning, statistical and visualization techniques to unveil and concentrate knowledge in such a manner such that humans can understand readily.

DM is being used widely in various fields namely medical, economic, engineering and even education these days. The main purpose of Educational Data Mining (EDM) is to analyze the hidden and unveiled pattern in the students' data.

1.1 Information Technology

The application of computers and telecommunications equipment for the storage, transmission and retrieval of data for their proper processing and operations is called Information Technology (IT). IT is usually comprised of computers, their corresponding networks, telephones, televisions or other electronic devices and the data transmission within them. Various industries are associated with information technology, including computer hardware, software, electronics, semiconductors, internet, telecommunication equipment, engineering, health-care, e-commerce and computer services.

Since the development of writing from 3000 BC, humans were trying to store, retrieve and manipulate information. The term Information Technology first appeared in 1958 in an article published in 'Harvard Business Review' by Harold Leavitt and Thomas Whistler. The term was defined basing on three categories: techniques for processing, the application of statistical and mathematical methods to decision-making, and the simulation of higher-order thinking through computer programs.

Basing on the storage and processing technologies employed, it was possible to distinguish four distinct phases of IT development: pre-mechanical (3000 BC 1450 AD), mechanical (1450 1840), electromechanical (1840 1940) and electronic (1940 present).

1.1.1 Significance of Information Technology (IT)

- accessible to variety of learning resources
- collaborative learning
- multimedia approach to education
- authentic and up to date information
- access to online libraries
- teaching of different subjects made interesting
- distance education
- multiple communication channels e-mail, chat, forum, blogs, etc.
- better accesses to children with disabilities
- reduces time on many routine tasks

1.1.2 Information Technology (IT) in Education

Education is a continuous learning process and it is limitless. Education supplies us with information which can be used further for the development of new theories and discoveries. Information System has the ability of speeding up the transmission of data and thus it can be used in educational fields. Most of the points are illustrated below.

- Plenty of Educational Resources:

  Information technology makes it easy to access academic information at any time. Both students and teachers use Information technology to acquire and exchange educational material.

- Instant Access to Educational Information:

  Information technology speeds the transfer and distribution of information. Students can easily access academic data using computers and new technologies like mobile phone application. These mobile phones are replacing past methods of borrowing books from the store, library instead they are downloaded from stores such as e-books from the internet.

- Full Time learning:

  Unlike in the past when learning was limited to a physical classroom, students and teachers could only access academic information while at school. Today, all that has

changed, a student will access information at any given time of the day. A person sitting in Africa can access lectures made in another corner of the earth instantly by means of IT.

- Use of Audio Visual:

  Information technology has changed the way we learn and interpret information. The use of audio-visual education helps students learn faster and easily. As opposed to text and blackboard notes, students get bored in this form of education. Visual learning is always more interesting and convenient to adapt than phonetically learning as students can always indulge themselves practically more with this technology.

1.2 Objectives

To improve the students' academic performance is the main concern of Institutions of Higher Learning (IHL) . IT can play a great role in this sector as nowadays each and every student is using IT medium for communication as well as for learning. So the main objective of this research is to find out the efficiency of IT in education field. We have completed our research by following the below objectives. The objectives are

- To understand how a student learn using IT.
- To determine critical success factors for a student.

1.3 Motivation

Numerous methods are associated to conduct data experiments such as classification, clustering and association rules. This analysis uses classification and clustering techniques to develop a sample for Students Academic Performance (SAP). This research aims for comparative analysis of classification and clustering. It also performs the comparison of three classification algorithms: Decision Tree (J48), Naive Bayes and Rule Based (OneR) to undermine the best predictive model. It also undermines the parameter to demonstrate the SAP in engineering course at MIST. The results generated from the thesis will help the students to improve their performance using IT and also guide the teachers to conquer the issues of low grades acquired by the students.

In our research our data collection has been branched into five classes. Students' English competency, their learning process, class items for interactions, communication and interaction for learning with teachers and also with classmates, and others are the main classes. In every class, we used CGPA as the goal parameter in our thesis and we classified it into three

groups: average, good and excellent. All other features are being used as an autonomous parameter. In our investigation, we used Waikato Environment for Knowledge Analysis (WEKA) which is an open source tool. It is mainly used for classification model development. WEKA is well known tool among researcher that's widely used for research purpose in the DM field [1–3].

1.4 Our Contribution

In this research our contribution is to build SAP model based on IT learning. Bangladesh is a developing country. Here also students use IT for educational purpose but in a little amount. Our main aim is to develop SAP model based on IT related learning in Bangladesh. For this analysis we have made a survey and conducted it to the students of MIST. The SAP model shows the highest accuracy value for the excellent, good and average students. This model evaluates that Rule Based (OneR) gives the highest accuracy for excellent students comparing to Decision Tree (J48) and Naive Bayes. This model also determines that classification predicts better efficiency than clustering. Error of classification is very much less than clustering. So to analyze students' performance using IT, classification is better technique than clustering.

In this paper we have visualized how a student learns using IT. From the analysis we have find out that this SAP model is very much efficient for students of excellent category. And we also determine that the most popular communication medium is 'Facebook' and mostly used mediums for group discussion are 'Skype and email'.

This model also shows that our SAP model is very much fruitful for average students using Decision Tree (J48). The root mean squared error of the accuracy value in classification model is very much low for all sections (approximately 0.543).

1.5 Report Layout

In rest of the research paper chapter 2 covers the literature review where all the previous experiments related to SAP and DM has been enclosed. Chapter 3 contains research methodology. In chapter 4 we have discussed EDM using WEKA, classification and clustering methods are also covered. Chapter 5 is covered by results of all of the sections and discussion of all of the outputs and also the comparison of classification and clustering is measured. And at last we conclude our research in chapter 6 and future expansion has also been explained.

CHAPTER 2
LITERATURE REVIEW

Data mining (DM) is a current field of research in higher education and this region of research is acquiring fame because of its latency to educational principle. Classification is one of the most eligible fields in DM. The main challenge in our country is to improve our educational performance. To achieve this aim DM is the best field. SAP is the main concern of previous researches because of the demand of qualified students not only in govt. but also in private institutes/sectors. So, SAP prediction is very much necessary to allocate SAP through which proper initiatives can be taken to improve students' performances.

2.1 Data mining

Data mining is a versatile field drawing works from statistics, database technology, artificial intelligence, pattern recognition, machine learning, information theory, knowledge acquisition, information retrieval, high performance computing, and data visualization [47] . Knowledge Discovery Database (KDD) is another form of DM. The main objective of KDD is to extract or mine knowledge from huge amounts of data [42].

Data mining is an approximately new and encouraging technology. It can be specified as the process of determining meaningful new interrelationship, design, and trends by drilling into (mining) large amounts of data stored in warehouse, using statistical, machine learning, artificial intelligence (AI), and data visualization methods [47].

Data mining is the nontrivial withdrawal of hidden, previously unidentified and potentially functional information from the data stack. It is the regular finding of innovative facts and interaction in data that are like precious nuggets of commerce records. It is not a difficult question where the user previously has a suspicion about a correlation in the data and wants to drag all such information. Data mining is the orientation of logical models and patterns from a database. It is the process of extracting previously unidentified, applicable, and actionable data from large databases and then using the data to make vital business decisions. Data mining is integrating the conversion of lots of information into significant facts. It is a method that assists to identify new opportunities by finding fundamental truths in apparently random data. The patterns revealed can shed light on application problems and assist in more useful, proactive decision making.

Han and Kamber claimed that data mining software which allows the users to examine data from different scope, classify it and summarize the relations which are recognized during the mining process [48]. Alaa el-Halees described that DM can be used in educational area to increase our perceptive of learning procedure to focus on distinguishing, extracting and judging variables associated to the education system of students [49].

2.2 Educational Data Mining

Educational Data Mining (EDM) is the mining of data in educational environment [42]. DM is the technique through which we can come across different relationships and models invented among different areas of databases. EDM is said to be a distinct research field where the application of data mining, machine learning and statistics to information which are generated from educational systems are specified. Predicting students' future learning behavior, advancing scientific knowledge, effects of educational support are the main objectives of EDM [28].

EDM is an emergence regulation that focuses on applying data mining mechanisms and methods to data which are related to education [38]. EDM is the technique of determining fruitful facts from raw data which are produced and accumulated from educational organizations so that different associations can use this information [31].

Data collection from historical and operational data can be used which exist in the databases of educational organizations. Students' personal or academic data can also be used. Most institutes use e-learning systems which have a huge amount of data so that these can also be collected for data mining to improve students' performance [45].

2.3 Application of DM in Monitoring Performance

Institutes of Higher Learning (IHL) mainly concern about SAP. Administrators, teachers and also students can change their techniques to improve the academic performance by acquiring the right pattern of SAP. It will also be helpful to classify the students according to their results. SAP can be obtained by using versatile data set, algorithms, tools and techniques.

The research for developing SAP models for the first semester Bachelor of Computer Science from University Sultan Zainal Abidin (UniSZA) used three selected classification models: Naive Bayes, Rule Based and Decision Tree. The total data set selected for the research was 399 from 497 during the data pre-processing stage. Five independent parameters: gender, race, hometown, family income, university entry mode have been selected to conduct this research. The result from the analysis technique discovered that the race is the most effective parameter to the students' performance [1].

The research to prevent undergraduate student dependency was accomplished by 5793 records consist of six special subjects that were economic sciences, law, civil engineering, languages, medicine, and pedagogy course. During the refining procedure 13 parameters were chosen from 21 parameters. The analysis will be helpful for the IHL to verify the weak and good students [50].

The analysis described the model to improve the SAP and identify the dropouts and students who need special attention. This analysis was conducted according to the data set obtained from VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method of computer Applications department of course MCA (Master of Computer Applications) from session 2007 to 2010. This analysis consists of eight parameters: PSM (Previous Semester Marks), CTG (Class Test Grade), SEM (Seminar Performance), ASS (Assignment), GP (General Proficiency), ATT (Attendance), LW (Lab Work), and ESM (End Semester Marks). This research result produced a model through which special attention can be taken to decrease the number of dropouts and it also allows the lecturers to offer suitable suggestion/recommendation [42].

The experiment of educational data was analyzed for progressing SAP and to conquer the problem of low grades of graduate students at College of Science and Technology, Khanyonis, Gaza. This data set consists of 3314 records and 18 parameters by categorizing the grade into four classes: excellent, very good, good and average as goal parameter. Three DM techniques were implemented to the data this are association rules, classification, and clustering. A set of rules were generated from the analysis which producing the relation between the parameter and the parameters that provided to the SAP [45].

2.4 Existing Frameworks

Lee, J. Stolfo, and W. Mok proposed a framework for building intrusion detection models. The vital proposal was to employ analyzing programs to dig out an expanded set of features that illustrate each network correlation or host session, and implement data mining programs to learn rules that exactly depict the behavior of intrusions and regular actions. These policies can then be applied for abuse recognition and inconsistency detection [12].

A framework for machine learning and data mining in the cloud which was based on graphlab was proposed by Low, Gonzalez, Kyrola, Bickson, Guestrin, and Hellerstein. Sometimes it has been noticed that map reduce, simplify the design and implementation is being difficult because these don't efficiently support various data mining process and machine learning. To minimize this problem the graphlab framework has been established [8].

Chen, Chung, Jennifer, Gang, Qin, and Chau presented a general framework for crime data mining. By incrementing competence and diminishing errors, crime data mining methods

can ease police job and facilitate investigators to give their time to other precious tasks [9].

An electric energy consumer characterization framework based on data mining techniques was established by Vera, Ftima, Zita, and Joaquim. Two main modules: the load profiling module and the classification module constitute this framework [10].

2.5 Challenges in EDM

Challenges of electronic document management were claimed by Sprague. The challenge of this research was to build up some construction for understanding the fast emerging field. This paper examines the range and consequence of EDM in more detail and describes how it broadens our view of information managing [51].

Ritika Saxena applied two algorithms of classification: Decision Tree (J48) and clustering: k-means clustering so as to predict the performance of both the algorithms. They were implemented on the marks of the student collected from the database of the university so as to grade the students based on their up to date performances [1].

CHAPTER 3
RESEARCH METHODOLOGY

In this chapter a framework has been suggested for this research to give a prediction of the SAP by applying DM classification rules. This proposed framework prefers a prediction model of parametric study of students learning using IT. This framework has been established through several steps:

- Main Goal
- Survey Design & Conduction
- Data Acquisition & Normalization
- Preparing for EDM
- Data Transformation
- Data Mining using WEKA
- Results & Discussion
- Implementation

The framework has been shown in Fig. 3.1.

3.1 Proposed Methodology

This section covers first four steps of our research methodology. The first step describes the goal of our analysis. Through survey design and conduction second step is performed. Then data acquisition and normalization has been accomplished in third step. After all these steps data are prepared for EDM at the next step. Then we implement WEKA for classification, analysis, comparison, and for decision making.

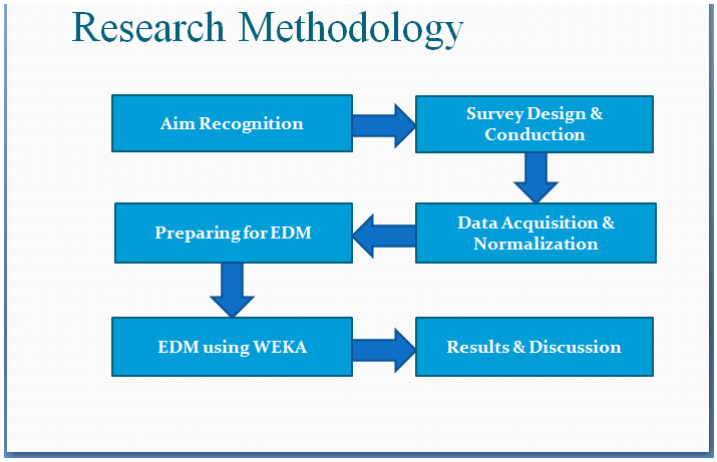


Figure 3.1: The Framework of the Research

3.1.1 Main Goal

The main goal of this research is to develop SAP of students' learning using IT by applying some classification rules implemented in WEKA that have already been mentioned. The techniques are:

- Decision Tree (J48)
- Naive Bayes
- Rule Based (OneR)

The WEKA is an open source tool which is composed in java language and also consists of a combination of state-of-the-art machine learning and DM classification algorithms for the test action in this experiment [2,26]. After analysis process, comparison of classification efficiency among the selected algorithms is observed. By verifying efficiency, the highest efficiency is checked and from the confusion matrix, percentage of prediction efficiency is calculated.

3.2 Survey Design and Conduction

To achieve the aim a survey has been designed for students who are the users of our research. Some questions related to IT have been set for them. These questions were divided into five categories. The categories are:

- About English Competency
- About Learning Process

- About Class Items Interaction
- About Communication and Interaction With Lecturers For Learning
- About Communication and Interaction With Classmates For Learning

Conduction process has been started after finalizing the survey. The survey was conducted manually to the B.Sc. students of level two, three, and four in Military Institute of Science and Technology (MIST), Dhaka, Bangladesh. The main aim is to find out their CGPA according to their answer of given questions so that we can find out whether IT related learning is really effective or not?

3.3 Data Acquisition and Normalization

At the preliminary step 300 data were collected from the B.Sc. students in MIST. Some data were eliminated from the collected data during preparation for EDM. The data contains information about students' interest in IT. Why do they use IT? How do they get access to it? And in which section IT is being used most? These are the main targets of the analysis. For this experiment four excels files were created:

- About English Competency. xls:

  In this excel sheet some questions about competency of English were set as parameters which is shown in table 3.1. Through these questions we can get idea about the effects of IT on the capability of English of students.

Table 3.1: Parameters About English Competency

Type	Parameter		Category	Priority
	Variable	Question		
Independent	B1	Clearly understanding topics while reading without any help?	{Very much,Somewhat, Undecided, Disagree, Strongly Disagree}	{5, 4, 3, 2, 1}
	B2	Clearly understanding topics while reading with the help of Dictionary?		
	B3	Clearly understanding topics while watching any video tuto rial without subtitle?		
	B4	Clearly understanding topics while watching any video tuto rial with subtitle?		
	B5	While reading do you find new wor ds?		
	B6	While reading do you find very few new words?		
	B7	Most of the new words are 'Scienti fic Terms'?		
	B8	Most of the new words are in 'Eng lish Dictionary'?		
	B9	I am competent in English for my studies		
	B10	I feel comfortable speaking in Eng lish		
Dependent	CGPA	CGPA?	Average:- 2.50-2.99 Good:- 3.00-3.4 Excellent:- 3.50-4.00	

- About Learning Process. xls:

  Few learning items of IT were set as parameter in this excel sheet. So, the most effective and helpful IT medium can be detected. All these parameters are illustrated in Table 3.2.

Table 3.2: Parameters About Learning Process

Type	Parameter		Category	Priority
	Variable	Learning Item		
Independent	C1	Video Tutorial	{I have no idea, Totally Ineffective, Not Satisfied, Satisfied, Completely Effective}	{1, 2, 3, 4, 5}
	C2	Audio Tutorial		
	C3	Pdf/Document		
	C4	Blog		
	C5	Simulator/Software		
	C6	Virtual Classroom		
	C7	Group Chat		
	C8	Online Shared ppt		
	C9	Webinar(Web Based Seminar)		
	C10	Face to face instruction with instructor		
	C11	Participation Online Test		
	C12	FB Group/Page		
	C13	Google Classroom		
	C14	Google Docs		
	C15	Blog & Wiki		
	C16	Google Calendar		
	C17	Google Drive		
Dependent	CGPA	CGPA?	Average:- 2.50-2.99 Good:- 3.00-3.49 Excellent:- 3.50-4.00	

- Class Items for Interaction. xls:

  Some teaching criteria which are applied during class time belong to this excel sheet. Few are helpful to students and few are not, this suggestion is given by students. Table 3.3 is drawn for showing these criteria.

Table 3.3: Parameters of Class Items for Interaction

Type	Parameter		Category	Priority
	Variable	Items		
Independent	C18	Example Based Teaching	{Very much, Somewhat, Undecided, Disagree, Strongly Disagree}	{5, 4, 3, 2, 1}
	C19	Hands on Learning		
	C20	Exercise in the Classroom		
	C21	Lecture Only		
	C22	Online Testing on Knowledge		
	C23	Assignment at Home		
	C24	Virtual Interaction on Google		
	C25	Problem Based Learning		
Dependent	CGPA	CGPA?	Average:- 2.50-2.99 Good:- 3.00-3.49 Excellent:- 3.50-4.00	

- About Communication and Interaction for Learning. xls:

  This excel sheet has been divided into two sections: interaction with lectures and interaction with classmates. Some IT related learning item has been included in this sheet as parameters. Table 3.4 is representing this data.

Table 3.4: Parameters of Communication and Interaction

Type	Parameter				Category	Priority
	Interaction with Teachers		Interaction with Classmates			
	Variable	Learning Item	Variable	Learning Item		
Independent	D1	Skype	D9	Skype	{Very much, Somewhat, Undecided,Disagree, Strongly Disagree}	{5, 4, 3, 2, 1}
	D2	Facebook	D10	Facebook		
	D3	Google Classroom	D11	Google Classroom		
	D4	Email	D12	Email		
	D5	Text via Phone	D13	Text via Phone		
	D6	Face to Face Interaction After Class	D14	Face to Face Interaction After Class		
	D7	Visiting Teachers Website	D15	Blog and Wiki		
Dependent	CGPA	CGPA?			Average:- 2.50-2.99 Good:- 3.00-3.49 Excellent:- 3.50-4.00	

3.4 Preparing For EDM

During this process some data have been eliminated due to the absent and partial value. Data pre-processing is performed to give the qualified data [2]. This lacking of data and partially completed value has been removed from the collection. At first around 56 data were removed from 300 data due to the missing value of CGPA. In the next step 66 data were eliminated due to the incomplete value. And at last 178 data have been selected for data transformation and mining.

Thus all the processes have been performed in this section to give the finalized data. The ultimate records are reduced to 178 from 300. This selected data have been filtered in WEKA.

CHAPTER 4
EDM USING WEKA

In this research WEKA 3.7 is used as an assessment tool. It was first invented by the University of Waikato, Hamilton, New Zealand. WEKA is an open source tool which is used for DM and also very familiar to the EDM. This tool is also well known to implement SAP. Many recognizable algorithms are built in it. A data mining tool is described by Bhullar. He applied WEKA classification algorithms because this tool is really fruitful for offering a steadiness among accuracy, swiftness and comprehensible of outcomes [27]. WEKA consists of machine learning algorithms [28]. To obtain perfect outputs the clustering and classification methods are tested in WEKA [35].

4.1 Fitting Data in WEKA

To fit data in WEKA at first all the excel sheets are converted into Comma Separated Value (CSV) file formats. And CSV file was converted to Attribute Relation File Format (ARFF) file inside WEKA. The ARFF file was fitted into the WEKA 3.7 explorer. Then string values are converted into nominal value for analysis. Classification, clustering, and regression techniques are applied to the finalized data set. To find out the efficiency, errors in data and also for establishing SAP for students and lecturers, these rules are really useful. Classifications are done by using three rules: Decision Tree (J48), Naive Bayes, Rule Based (OneR) and K-means clustering is applied as the clustering technique.

4.2 Classification

Classification is the one of the most frequently tested procedure in EDM that allocates class lies in the data set. Variables/objects are grouped into predefined class. Classification algorithms are already built in WEKA through which SAP can be established. In our thesis we are using three algorithms.

4.2.1 Decision Tree (J48)

Decision tree is very effective and well-known algorithm in classification. This algorithm is used to form a top-down tree from the input data set. This tree is drawn from top to bottom by various branches and leaf nodes. The leaf node is the ultimate result of the data set. It represents a class.

In WEKA J48 is a java implementation of the C4.5 algorithm. This algorithm shows the relation of different parameters of the data set [6].

4.2.2 Naive Bayes

Naive Bayes is actually a Bayesian classifier which can handle any number of parameters in spite of considering whether these variables have quantity value or quality value. Naive Bayes algorithm claims that the variables used in this classifier are independent of other variables [35].

4.2.3 Rule Based (OneR)

For classification of data Rule Based approach is a great technique in WEKA. It implements if-then rules in data set and gives generally highest accuracy of data. OneR is the simplest and easiest algorithm among three. It generates a decision tree which has only one level [1].

Azwa, Nor, and Fadhilah did an experiment on the first semester Computer Science students from UniSZA to present SAP which was used to discover a parameter that contributes to students' success most. Three classification algorithms: Naive Bayes, Decision Tree (J48), Rule Based (OneR) were used. The output declared that the models of Rule Based and Decision Tree algorithm provide the highest prediction efficiency value of 68.8%. This model also proved the outstanding performance of predicting average students with the efficiency value of 100%. But this model failed to give poor students efficiency [1].

4.3 Clustering

Clustering is the most efficient process in DM. For statistical data analysis it is also a very popular method. It is applied to cluster objects or variables of similar characteristics into one group or class. The characteristic of one cluster is different from other. In data mining clustering is used to find out data point that can be grouped into one class and this cluster defines the data set of similar data point. While determining of one group has been finished, new examples can be classified by determining the nearest group [5]. In EDM clustering, regression, association rules are being used to advance SAP and also to develop the teaching

process. Clustering is also being used in solving homework assignment and organization of course material related problem [26]. This technique is so much helpful in data mining. Clustering is being used in different areas: pattern recognition, image analysis, bioinformatics, artificial intelligence, machine learning and so on. There are several clustering algorithms such as k-means, hierarchical clustering, fuzzy clustering, Expectation maximization algorithm (EM), Density-Based Spatial Clustering Of Applications with Noise (DBSCAN) etc. Among these k-means clustering is used in this research.

4.3.1 K-Means Clustering

K-means clustering is very efficient, easiest and simple clustering algorithm. It gives approximately the highest accuracy value. Most of the time in case of clustering k-means algorithm is applied to acquire accurateness in results. It is actually Kernel k-means clustering [28]. James MacQueen in 1967 first invented the factor 'k-means' and it is also known as Forzy's method [13]. Anderberg in 1973 also claimed that k-means is very efficient for large data set so it is very fruitful for DM. Jain and Dubes in 1988 said that k-means is suitable only to numeric data that means a ratio scale is maintained in measuring variables since a cost task is minimized by changing the means of clusters [16].

The k-means algorithm is mainly applied on different type of variables and then it classifies the variables into k groups so that's the reason it's called k-clustering. At first the sum of square distances between the given data and cluster centroid is performed. And the minimum distance is calculated and the variables of this minimum distance are kept in a cluster. So the steps of k-means algorithm are:

- Find the centroid.
- Calculate the sum of square distances.
- Cluster the variables of minimum distance.
- This group is the nearest group of k-means.

K-means is basically a entitle to the classic K-Means algorithm taking: an primary initial dot, data set and the number of clusters K, returning a set of K to d-dimensional vectors, the estimates of the centroids of the K clusters [13].

So the main objective of k-means is to minimize the mean square distance. It is the most admirable algorithm in clustering to find out the data points which are close to the centroid of the cluster. K-means clustering is used by Ritika for mining large amount of data in order to evaluate this model for future expansion [28].

4.4 Comparative Analysis

Classification is supervised learning while clustering is unsupervised learning. In classification objects are classified into predefined classes. And in clustering objects of similar characteristics are clustered into newly defined class.

In classification, we have a training set containing data that have been previously categorized. Based on this training set, the algorithm finds the category to which the new data points belong. In clustering, we do not know the characteristics of similarity of data in advance. Using statistical concepts, we split the data sets into sub-data sets such that the sub-data sets have similar data.

For example:

In classification, we use training data set which categorized customers that have churned. Now based on this training set, we can classify whether a customer will churn or not.

In clustering, we use a data set of customers and split them into sub-data sets of customers with ‘similar’ characteristics. Now this information can be used to market a product to a specific segment of customers that has been identified by clustering algorithm.

CHAPTER 5
RESULTS AND DISCUSSION

After performing all the calculation, classification, clustering, association, and visualization we have at last reached our destination and got expected results. From the output, we can analyze the accuracy value both of classification and clustering. This value determines how efficient the SAP model is. For classification we split data set in fold cross validation. Data set have been separated into 3, 5, 7 and 10 data subsets in fold cross validation using WEKA. For clustering the most efficient method k-means clustering is applied in WEKA 3. 7.

5.1 Effect on Performance Matrix

Classification and clustering both is done for all of the five data matrices. In clustering among all of data sets, error for 'Class Item Interaction' is the lowest. And in classification the data set for 'About English Competency' gives the highest accuracy for excellent students in five fold cross validation method. That is 94.5%.

The efficiency of the classification model for all of the data sets is shown in different sections and then we compare the values to get the highest accuracy. Besides the clustered instances for all of the five data sets have been evaluated in clustering using k-means algorithm. For clustering we have used three parameters to cluster the objects into three expected classes although the default value of parameter is two in WEKA 3.7.

5.1.1 Performance Matrix of About English Competency

The table 5.1 shows the efficiency value of the three classification algorithms in WEKA 3. 7. The highest efficiency using Decision Tree is 50% in seven fold cross validation; Naive Bayes gives 47.2% highest accuracy also in seven fold cross validation, Rule Based determines the highest accuracy in five fold cross validation which is 55.7%. Among three the highest efficiency value is 55.7% which is obtained from Rule Based (OneR) in five fold cross validation.

Table 5.1: Efficiency Values of Classification Model for English Competency

Fold Cross Validation.	Classifier Efficiency		
	Decision Tree (J48)	Naive Bayes	Rule Based(One R)
3	46.6%	44.9%	52.3%
5	46.5%	41.5%	55.7%
7	50%	44.2%	55.6%
10	43.2%	42.6%	55.6%
Highest Efficiency	50%	47.2%	55.7%

From figure 5.1 it can be observed that the data set are characterized into three clusters. Cluster 0, 1, 2 contain 46, 40 and 90 instances respectively.

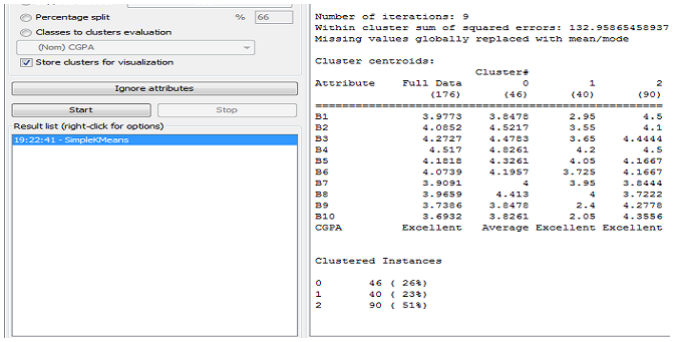


Figure 5.1: Cluster Instances of 0, 1, 2 clusters About English Competency

5.1.2 Performance Matrix of About Learning Process

The table 5.2 shows the efficiency value of the three classification algorithms in WEKA 3. 7. The highest efficiency using Decision Tree is 54% in ten fold cross validation; Naive Bayes gives 45.5% highest accuracy in five fold cross validation, Rule Based determines the highest accuracy in ten fold cross validation which is 53.4%. Among three the highest efficiency value is 54% which is obtained from Decision Tree (J48) in ten fold cross validation. Figure 5.2 shows the tree which is obtained from J48 algorithm where the leaf nodes represents the class.

Table 5.2: Efficiency Values of Classification Model for Learning process

Fold Cross Validation.	Classifier Efficiency		
	Decision Tree (J48)	Naive Bayes	Rule Based(One R)
3	43.2%	44.3%	44.9%
5	51.7%	45.5%	52.3%
7	42.0%	42%	53.4%
10	54%	44.3%	53.4%
Highest Efficiency	54%	45.5%	53.4%

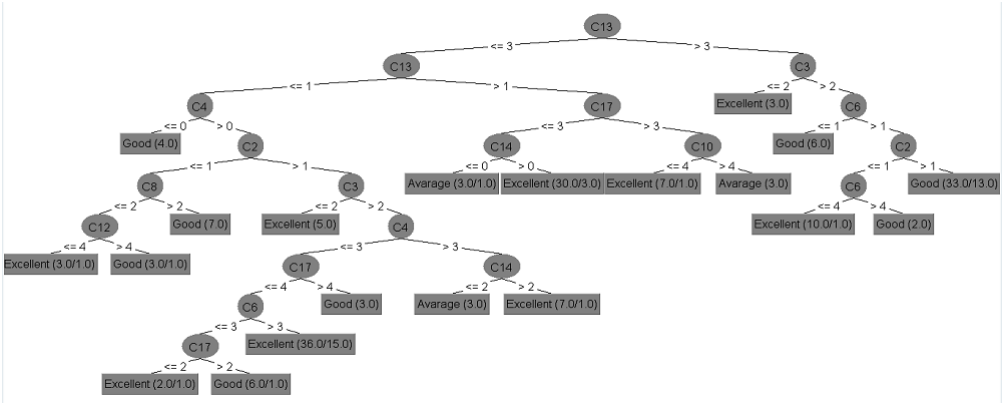


Figure 5.2: Decision Tree Visualization of J48 Algorithm

From figure 5.3 it can be shown that the data set are parameterized into three clusters. Cluster 0, 1, 2 contain 38, 48 and 90 instances respectively.

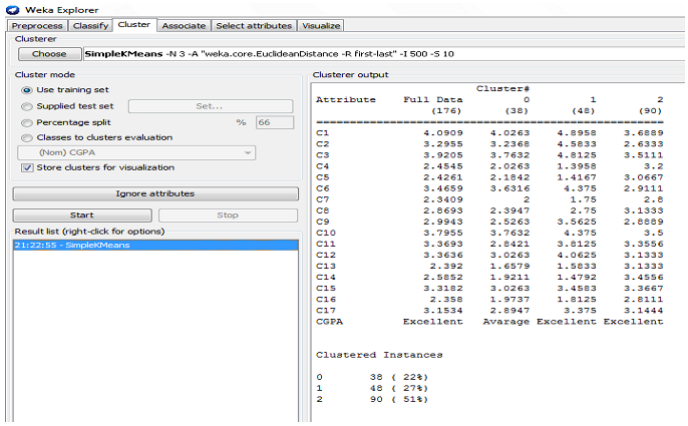


Figure 5.3: Cluster Instances of 0, 1, 2 clusters About Learning Process

5.1.3 Performance Matrix of About Class Items Interaction

The table 5.3 shows the efficiency value of the three classification algorithms in WEKA 3. 7. The highest efficiency using Decision Tree is 41.5% in three fold cross validation; Naive Bayes gives 43.2% highest accuracy in seven fold cross validation, Rule Based determines the highest accuracy in ten fold cross validation which is 55.1%. Among three the highest efficiency value is 54% which is obtained from Rule Based (OneR) in ten fold cross validation.

Table 5.3: Efficiency Values of Classification Model for Class Item Interaction

Fold Cross Validation.	Classifier Efficiency		
	Decision Tree (J48)	Naive Bayes	Rule Based(One R)
3	41.5%	39.8%	50.6%
5	38.6%	42%	52.3%
7	39.2%	43.2%	49.4%
10	39.2%	42%	55.1%
Highest Efficiency	41.5%	43.2%	55.1%

From figure 5.4 it can be shown that the data set are parameterized into three clusters. Cluster 0, 1, 2 contain 44, 87 and 45 instances respectively.

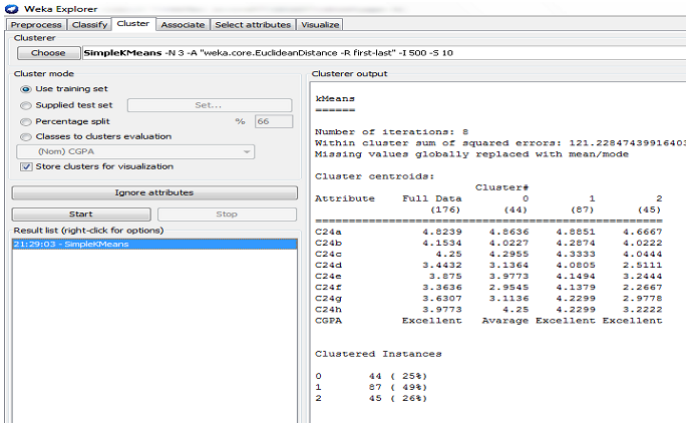


Figure 5.4: Cluster Instances of 0, 1, 2 clusters About Class Items Interaction

5.1.4 Performance Matrix of About Communication and Interaction

The table 5.4 shows the efficiency value of the three classification algorithms using WEKA tool. This section is divided into two categories:

- Interaction with teachers
- Interaction with classmates

For interaction with teachers, the highest efficiency using Decision Tree is 44.3% in three fold cross validation; Naive Bayes gives 47.7% highest accuracy in five fold cross validation, Rule Based determines the highest accuracy in ten fold cross validation which is 53.4%. Among three the highest efficiency value is 53.4% which is obtained from Rule Based (OneR) in ten fold cross validation.

Table 5.4: Efficiency Values of Classification Model for Interaction

Fold Cross Validation.	Classifier Efficiency					
	Interaction with Teachers			Interaction with Classmate		
	Decision Tree (J48)	Naive Bayes	Rule Based (One R)	Decision Tree (J48)	Naive Bayes	Rule Based (One R)
3	44.3%	43.2%	47.7%	44.3%	46%	51.7%
5	42.6%	47.7%	47.2%	48.9%	51.1%	53.4%
7	39.2%	44.9%	52.8%	51.1%	48.9%	54%
10	40.9%	46.6%	53.4 %	47.2%	50.6%	54%
Highest Efficiency	44.3%	47.7%	53.4%	51.1%	51.1%	54%

For interaction with classmates, the highest efficiency using Decision Tree is 51.1% in seven fold cross validation; Naive Bayes gives 51.1% highest accuracy in five fold cross validation, Rule Based determines the highest accuracy in ten fold cross validation which is 54%. Among three the highest efficiency value is 54% which is obtained from Rule Based (OneR) in ten fold cross validation.

From figure 5.5 it can be determined that the data set are characterized into three clusters. Cluster 0, 1, 2 contain 44, 84 and 48 instances respectively.

From figure 5.6 it can be determined that the data set are catheterized into three clusters. Cluster 0, 1, 2 contain 41, 59 and 76 instances respectively.

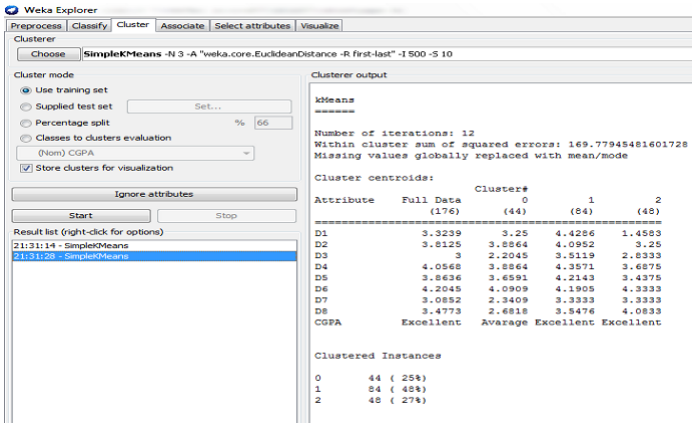


Figure 5.5: Cluster Instances of 0, 1, 2 clusters About Interaction with Teachers

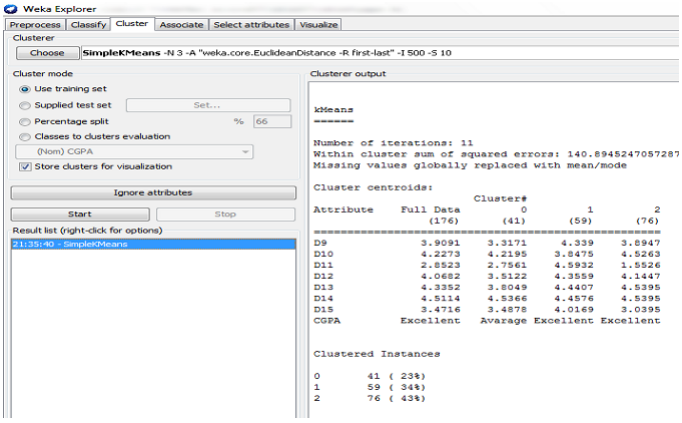


Figure 5.6: Cluster Instances of 0, 1, 2 clusters About Interaction with Classmates

5.2 Impact of Classification and Clustering

After observing all of the performance matrices both in classification and clustering, impact of them can be determined in this section. Errors are compared between the two techniques to find out which one is better than other in this research.

5.2.1 About English Competency

The confusion matrix table for Rule Based 5 fold cross validation is represented in table 5.5. From the table it is notified that prediction gives the highest achievement of 94.5% for excellent class, followed by 20% of good students and this prediction fails to predict average students for 'About English Competency'.

Table 5.5: Confusion Matrix for Rule Based of 5 Fold Cross Validation

Classified as				Sum of Accurately Classified Data	Prediction Achievement
	Excellent	Good	Average		
Excellent	86	5	0	86	94.5%
Good	48	12	0	12	20.0%
Average	24	1	0	0	0%

Figure 5.7 shows the cluster visualization of data for this section.

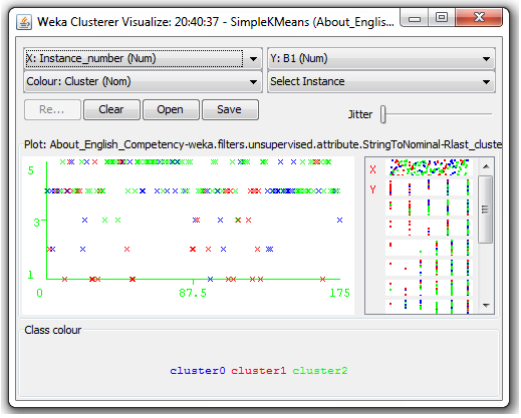


Figure 5.7: Clustering Visualization About English Competency

From two techniques it can be observed in figure 5.1 that the sum of squared errors is 133% in clustering while in Rule Based root mean squared error is 0.5436 which is very much less than clustering.

5.2.2 About Learning Process

The confusion matrix table for Decision Tree 10 fold cross validation is represented in table 10. From the table it is notified that prediction gives the highest achievement of 64.9% which is less than Rule Based for excellent class, followed by 55% which is better than rule

Based of good students and this also predicts average students' achievement 44% for 'About Learning Process'.

Table 5.6: Confusion Matrix for Decision Tree(J48) of 10 Fold Cross Validation

Classified as				Sum of Accurately Classified Data	Prediction Achievement
	Excellent	Good	Average		
Excellent	59	21	11	59	64.9%
Good	21	33	6	33	55%
Average	17	5	3	11	44%

Figure 5.8 shows the cluster visualization of data for this section.

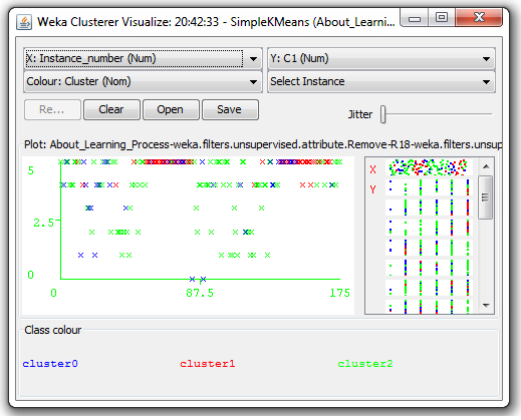


Figure 5.8: Clustering Visualization About Learning Process

From two techniques it can be observed that in figure 5.3 the sum of squared errors is 264.3% in clustering while in Decision Tree root mean squared error is 0.5017 which is very much less than clustering.

5.2.3 About Class Items Interaction

The confusion matrix table for Rule Based 10 fold cross validation is represented in table 11. From the table it is notified that prediction gives the highest achievement of 93.4% for excellent class, followed by 20% of good students and this prediction fails to predict average students for 'About Class Items Interaction'.

Figure 5.9 shows the cluster visualization of data for this section.

Table 5.7: Confusion Matrix for Rule Based of 10 Fold Cross Validation

Classified as				Sum of Accurately Classified Data	Prediction Achievement
	Excellent	Good	Average		
Excellent	85	6	0	85	93.4%
Good	48	12	0	12	20%
Average	24	1	0	0	0%

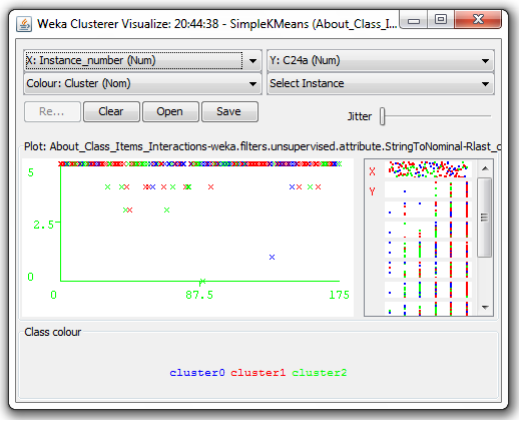


Figure 5.9: Clustering Visualization About Class Items Interaction

It can be observed from figure 5.4 that the sum of squared errors is 121.2% in clustering while in Rule Based root mean squared error is 0.547 which is very much less than clustering.

5.2.4 About Communication and Interaction

The confusion matrix table for Rule Based 10 fold cross validation is represented in table 12 for both teachers' and classmates' interaction. From the table it is notified that prediction gives the highest achievement of 91.2% and 90.1% for excellent class, followed by 18.3% and 21.8% of good students respectively and this prediction fails to predict average students for 'About Communication and Interaction'.

Figure 5.10 and 5.11 shows the cluster visualization of data for this section.

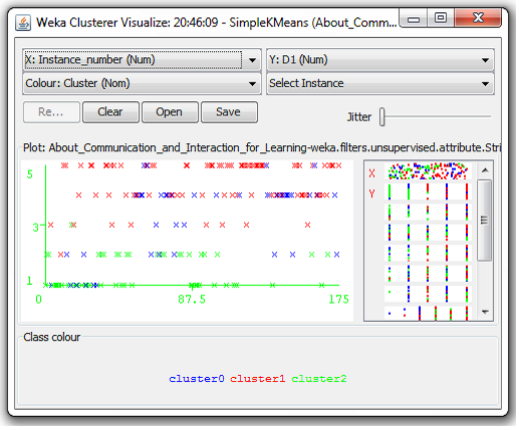


Figure 5.10: Clustering Visualization About Interaction with Teachers

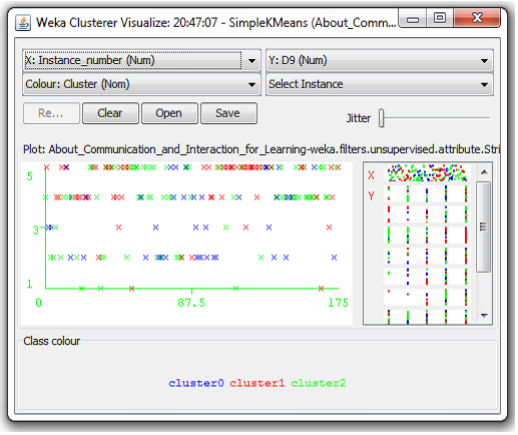


Figure 5.11: Clustering Visualization About Interaction with Classmates

Table 5.8: Confusion Matrix for Rule Based of 10 Fold Cross Validation for Interaction

Interaction with Teachers				Interaction with Classmates			Sum of Efficiency Classified Data		Prediction Achievement	
	Excellent	Good	Average	Excellent	Good	Average	Interact with Teachers	Interact with Classmates	Interact with Teachers	Interact with Classmates
Excellent	83	8	0	82	9	0	83	82	91.2%	90.1%
Good	49	11	0	47	13	0	11	13	18.3%	21.8%
Average	23	2	0	21	4	0	0	0	0%	0%

From two techniques it can be observed that in figure 5.5 and 5.6 the sums of squared errors are 169.8% and 140.9% in clustering for teachers' and classmates' interaction respectively while in Rule Based root mean squared errors are 0.5573 and 0.5539 respectively which are very much less than clustering. So for this SAP it can be concluded that classification is better than the clustering method and Rule Based (OneR) method is very much efficient to give the highest accuracy value among the three classification algorithms.

CHAPTER 6
CONCLUSION AND FUTURE EXPANSION

6.1 Conclusion

There are various kinds of data each and everywhere. Information related to education is huge and the educational sector is vast. So it is difficult to acquire all the data at the same time and use them for data mining. That is why it is a great challenge to mine all the educational data for predicting SAP model and give better suggestion to improve the education system for IHL. Moreover IT plays a great role in educational field. Students as well as teachers are using IT in their working sector for improving skill in their own sector. Even in schools and colleges there are versatile uses of IT which is really helpful both for the students and teachers and even also for the administrators. In this research our main aim was to find out that 'Is really IT effective for students' learning?'. In this research, at first a survey was designed. By conducting it to the B.Sc. students of MIST the data acquisition was complete. And then data was normalized from the acquired data, due to some missing and incomplete value some of the data was removed from the normalized data. After doing all the normalization and pre-processing, data set were transferred to WEKA 3.7. Performing classification and clustering in WEKA the accuracy value was calculated and then the errors between the two techniques were compared. The output reveals that the model of Rule Based (OneR) is very efficient to give the highest accuracy for the excellent and good students for all sections. It fails to predict the average students but the model J48 predicts all the excellent, good and also average students though its' accuracy is less than the Rule Based. K-means clustering is the best method for clustering data. It divided the data set into three clusters as there are three classes of CGPA. Error is greater in clustering comparing to classification. So it can be concluded that classification gives better accuracy than clustering.

6.2 Limitations

During data pre-processing due to absence and partial value some of the data were eliminated. So whatever we acquired we could not utilize them. This is a limitation of this research. Rule Based (OneR) fails to predict the average students while Decision Tree (J48) predicts all of the students' accuracy. But accuracy of the Decision Tree (J48) is less than the Rule Based (OneR). So these are the limitations of this research.

6.3 Future Scope

This research has been performed only on the parametric study of students' learning in IT. In future we will accomplish the study of teaching process using IT. It is a great medium to spread knowledge. So this can be utilized in teaching process also. Many teachers use ppt. to deliver their lecture, they use iphone, laptop and so many digital equipment and techniques to make the lectures easy to students. Even most of the teachers use email, facebook and so many groups to upload the lectures, mark sheets and also essential educative elements so that the students can get them easily without any rush and confusion. That's why our next target will be to find out effectiveness of IT in teaching systems. So that the teachers can improve their teaching technique and can propose a very efficient SAP model to guide the students. It will really be helpful for upgrading students' performance.

REFERENCES

[1] A. A. Aziz and N. H. I. Ahmad, "First semester computer science students academic performances analysis by using data mining classification algorithms,"

[2] S. Pal, "Mining educational data using classification to decrease dropout rate of students," *arXiv preprint arXiv:1206.3078*, 2012.

[3] D. Kabakchieva, "Predicting student performance by using data mining methods for classification," *Cybernetics and Information Technologies*, vol. 13, no. 1, pp. 61–72, 2013.

[4] R. Ferguson, "Learning analytics: drivers, developments and challenges," *International Journal of Technology Enhanced Learning*, vol. 4, no. 5-6, pp. 304–317, 2012.

[5] U. K. Pandey and S. Pal, "Data mining: A prediction of performer or underperformer using classification," *arXiv preprint arXiv:1104.4163*, 2011.

[6] J. Gholap, "Performance tuning of j48 algorithm for prediction of soil fertility," *arXiv preprint arXiv:1208.3943*, 2012.

[7] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: a survey," *IEEE transactions on neural networks*, vol. 13, no. 1, pp. 3–14, 2002.

[8] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein, "Distributed graphlab: a framework for machine learning and data mining in the cloud," *Proceedings of the VLDB Endowment*, vol. 5, no. 8, pp. 716–727, 2012.

[9] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *Computer*, vol. 37, no. 4, pp. 50–56, 2004.

[10] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," *Power Systems, IEEE Transactions on*, vol. 20, no. 2, pp. 596–602, 2005.

[11] R. T. Ng and J. Han, "E cient and e ective clustering methods for spatial data mining," in *Proc. of*, pp. 144–155, 1994.

[12] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in *Security and Privacy, 1999. Proceedings of the 1999 IEEE Symposium on*, pp. 120–132, IEEE, 1999.

[13] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering.," in *ICML*, vol. 98, pp. 91–99, Citeseer, 1998.

[14] G. Jagannathan and R. N. Wright, "Privacy-preserving distributed k-means clustering over arbitrarily partitioned data," in *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 593–599, ACM, 2005.

[15] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, *et al.*, "Constrained k-means clustering with background knowledge," in *ICML*, vol. 1, pp. 577–584, 2001.

[16] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data mining and knowledge discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[17] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[18] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li, "Automated variable weighting in k-means type clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 5, pp. 657–668, 2005.

[19] M. Chau, R. Cheng, B. Kao, and J. Ng, "Uncertain data mining: An example in clustering location data," in *Advances in Knowledge Discovery and Data Mining*, pp. 199–204, Springer, 2006.

[20] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, p. 37, 1996.

[21] S. Tahir and S. R. Naqvi, "Factors affecting studentsperformance," *Bangladesh e-journal of sociology*, vol. 3, no. 1, p. 2, 2006.

[22] G. Ben-Zadok, A. Hershkovitz, R. Mintz, and R. Nachmias, "Examining online learning processes based on log files analysis: A case study," in *5th International Conference on Multimedia and ICT in Education (m-ICTE09)*, 2009.

[23] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees," in *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*, 2006.

[24] A. Osofisan, O. Adeyemo, and S. Oluwasusi, "Empirical study of decision tree and artificial neural network algorithm for mining educational database," 2014.

[25] J. Doshi, "Result mining: Analysis of data mining techniques in education,"

[26] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, and E. M. Al-Shawakfa, "A comparison study between data mining tools over some classification methods," *IJACSA) International Journal of Advanced Computer Science and Applications*, pp. 18–26, 2011.

[27] M. S. Bhullar and A. Kaur, "Use of data mining in education sector," in *Proceedings of the World Congress on Engineering and Computer Science*, vol. 1, pp. 24–26, 2012.

[28] R. Saxena, "Educational data mining: Performance evaluation of decision tree and clustering techniques using weka platform," *International Journal of Computer Science and Business Informatics, IJCSBI. ORG*, vol. 15, no. 2, 2015.

[29] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "Data mining algorithms to classify students," in *Educational Data Mining 2008*, 2008.

[30] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," *Expert systems with applications*, vol. 33, no. 1, pp. 135–146, 2007.

[31] M. S. G. Kulkarni, M. G. C. Rampure, and M. B. Yadav, "Understanding educational data mining (edm)," *International Journal of Electronics and Computer Science Engineering*, vol. 2, no. 2, pp. 773–777, 2013.

[32] K. Sharma, D. Ashok, and H. Rohil, "A study of sequential pattern mining techniques," *space*, vol. 20, p. 21.

[33] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *Knowledge and data Engineering, IEEE Transactions on*, vol. 8, no. 6, pp. 866–883, 1996.

[34] N. Manouselis, H. Drachsler, K. Verbert, and E. Duval, *Recommender systems for learning*. Springer Science & Business Media, 2012.

[35] M. P. R. SHAH, D. B. VAGHELA, and D. P. SHARMA, "Predicting and analysing faculty performance using distributed data mining,"

[36] P. Blikstein, "Using learning analytics to assess students' behavior in open-ended programming tasks," in *Proceedings of the 1st international conference on learning analytics and knowledge*, pp. 110–116, ACM, 2011.

[37] M. Parikh, B. Chaudhari, and C. Chand, "A comparative study of sequential pattern mining algorithms," *International Journal of Application or Innovation in Engineering and Management*, vol. 2, no. 2, pp. 103–109, 2013.

[38] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, no. 6, pp. 601–618, 2010.

[39] R. Baker *et al.*, "Data mining for education," *International encyclopedia of education*, vol. 7, pp. 112–118, 2010.

[40] S. G. Brainard and L. Carlin, "A longitudinal study of undergraduate women in engineering and science," in *Frontiers in Education Conference, 1997. 27th Annual Conference. Teaching and Learning in an Era of Change. Proceedings.*, vol. 1, pp. 134–143, IEEE, 1997.

[41] E. Dahlstrom, J. Walker, and C. Dziuban, "Ecar study of undergraduate students and information technology," tech. rep., 2012, 2012.

[42] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv preprint arXiv:1201.3417*, 2012.

[43] M. Pandey and V. K. Sharma, "A decision tree algorithm pertaining to the student performance analysis and prediction," 2013.

[44] M. Waran *et al.*, "Employee performance evaluation using machine learning algorithm," *International Journal of Computer Communications and Networks (IJCCN)*, vol. 4, no. 2, 2014.

[45] M. M. A. Tair and A. M. El-Halees, "Mining educational data to improve students performance: a case study," *International Journal of Information*, vol. 2, no. 2, 2012.

[46] S. Singh and V. Kumar, "Performance analysis of engineering students for recruitment using classification data mining techniques,"

[47] S. Sumathi and S. Sivanandam, *Introduction to data mining and its applications*, vol. 29. Springer, 2006.

[48] J. Han and M. Kamber, "Data mining: concepts and techniques," 2001.

[49] A. El-Halees, "Mining students data to analyze e-learning behavior: A case study," *Department of Computer Science, Islamic University of Gaza PO Box*, vol. 108, 2009.

[50] H. R. B. d. Silva and P. J. L. Adeodato, "A data mining approach for preventing undergraduate students retention," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pp. 1–8, IEEE, 2012.

[51] R. H. Sprague Jr, "Electronic document management: Challenges and opportunities for information systems managers," *MIS Quarterly*, pp. 29–49, 1995.

APPENDIX A
ALGORITHMS

A.1 Sample Algorithm

Algorithm

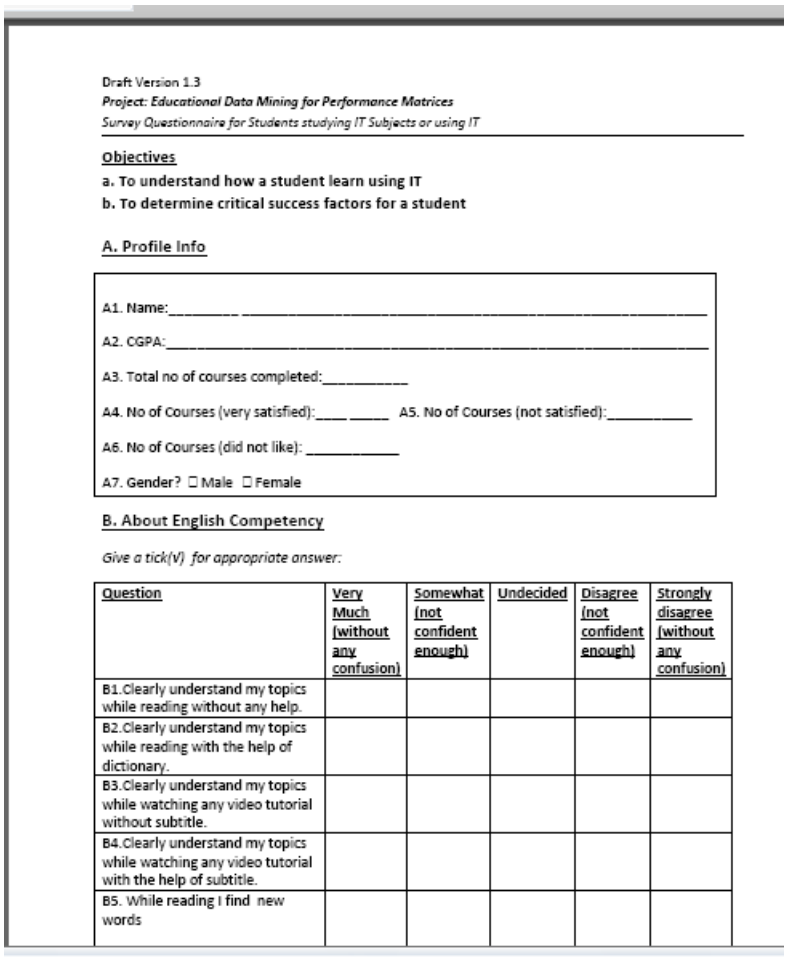
Algorithm 1 Sample Algorithm

Input: : D = $x_1, x_2, ..., x_n//$ Training data set, D, which contains a set of training instances and their associated class table. Output: T, Decision Tree. Method:
1: T = 0
2: Determine based splitting attribute
3: T = Create the root node and label it with the splitting attribute
4: T = Add to the root node and label it with the splitting attribute
5: For each arc do
6: D = Data set created by applying splitting predicate to D
7: If stopping point reached for this path then
8: T' = Create a leaf node and label with an appropriate class
9: Else
10: T = DT Build (D)
11: Else if
12: T = Add T' to arc
13: End for

APPENDIX B
SURVEY PAPER

B.1 Survey Paper

This is the survey paper



Draft Version 1.3
Project: Educational Data Mining for Performance Matrices
Survey Questionnaire for Students studying IT Subjects or using IT

Objectives

a. To understand how a student learn using IT
b. To determine critical success factors for a student

A. Profile Info

A1. Name:_____ _____

A2. CGPA:_____

A3. Total no of courses completed:_____

A4. No of Courses (very satisfied):____ _____ A5. No of Courses (not satisfied):_____

A6. No of Courses (did not like): _____

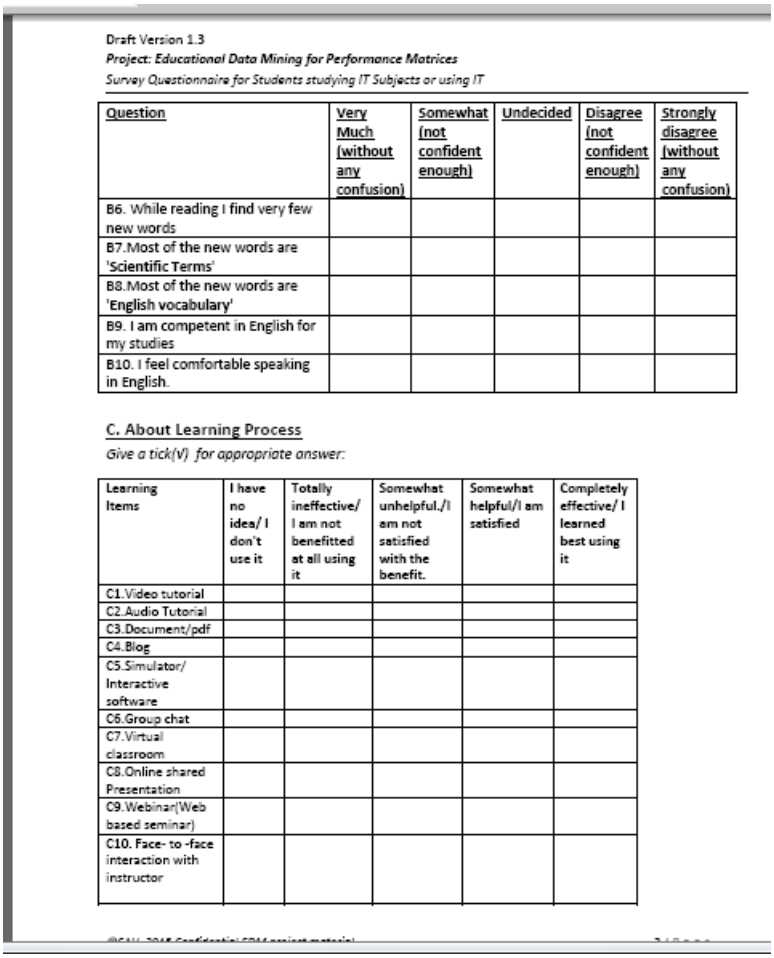
A7. Gender? ☐ Male ☐ Female

B. About English Competency

Give a tick(V) for appropriate answer:

Question	Very Much (without any confusion)	Somewhat (not confident enough)	Undecided	Disagree (not confident enough)	Strongly disagree (without any confusion)
B1.Clearly understand my topics while reading without any help.					
B2.Clearly understand my topics while reading with the help of dictionary.					
B3.Clearly understand my topics while watching any video tutorial without subtitle.					
B4.Clearly understand my topics while watching any video tutorial with the help of subtitle.					
B5. While reading I find new words					

Figure B.1: 1st Page of Questions



Draft Version 1.3

Project: Educational Data Mining for Performance Matrices

Survey Questionnaire for Students studying IT Subjects or using IT

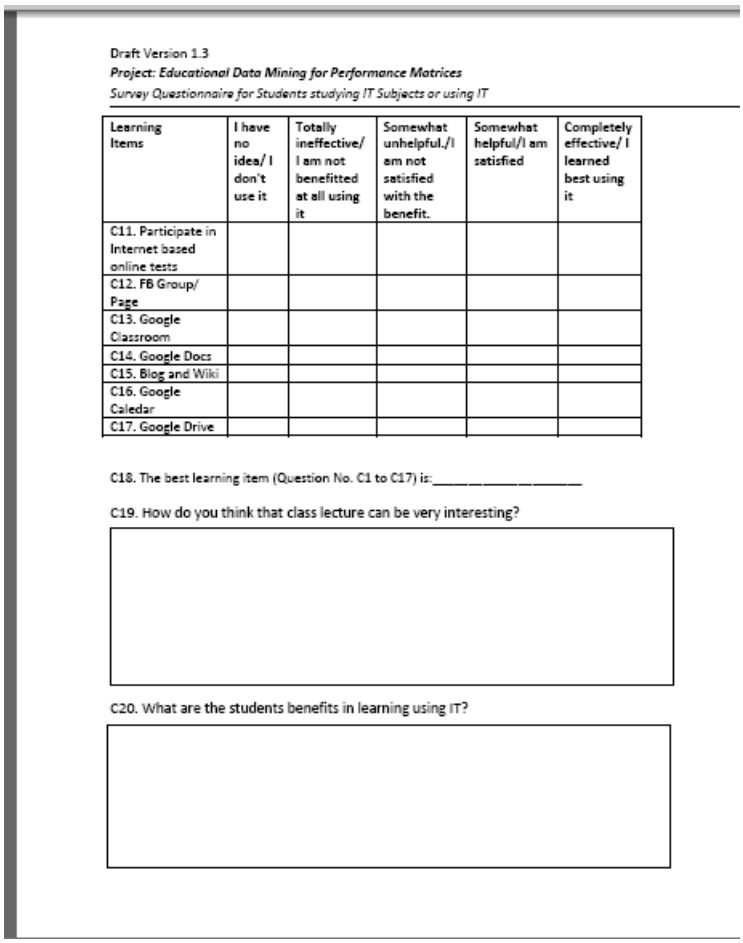
Question	Very Much (without any confusion)	Somewhat (not confident enough)	Undecided	Disagree (not confident enough)	Strongly disagree (without any confusion)
B6. While reading I find very few new words					
B7.Most of the new words are 'Scientific Terms'					
B8.Most of the new words are 'English vocabulary'					
B9. I am competent in English for my studies					
B10. I feel comfortable speaking in English.					

C. About Learning Process

Give a tick(V) for appropriate answer:

Learning Items	I have no idea/ I don't use it	Totally ineffective/ I am not benefitted at all using it	Somewhat unhelpful./I am not satisfied with the benefit.	Somewhat helpful/I am satisfied	Completely effective/ I learned best using it
C1.Video tutorial					
C2.Audio Tutorial					
C3.Document/pdf					
C4.Blog					
C5.Simulator/ Interactive software					
C6.Group chat					
C7.Virtual classroom					
C8.Online shared Presentation					
C9.Webinar(Web based seminar)					
C10. Face- to -face interaction with instructor					

Figure B.2: 2nd Page of Questions



Draft Version 1.3

Project: Educational Data Mining for Performance Matrices

Survey Questionnaire for Students studying IT Subjects or using IT

Learning Items	I have no idea/ I don't use it	Totally ineffective/ I am not benefitted at all using it	Somewhat unhelpful./I am not satisfied with the benefit.	Somewhat helpful/I am satisfied	Completely effective/ I learned best using it
C11. Participate in Internet based online tests					
C12. FB Group/ Page					
C13. Google Classroom					
C14. Google Docs					
C15. Blog and Wiki					
C16. Google Caledar					
C17. Google Drive					

C18. The best learning item (Question No. C1 to C17) is:_____

C19. How do you think that class lecture can be very interesting?

C20. What are the students benefits in learning using IT?

Figure B.3: 3rd Page of Questions

Draft Version 1.3

Project: Educational Data Mining for Performance Matrices

Survey Questionnaire for Students studying IT Subjects or using IT

C21. What are your obstacles in learning using IT?

C22. Why do you think about virtual platform for learning? (e.g. Google Classroom etc)

C23. You find learning easier when teacher uses:

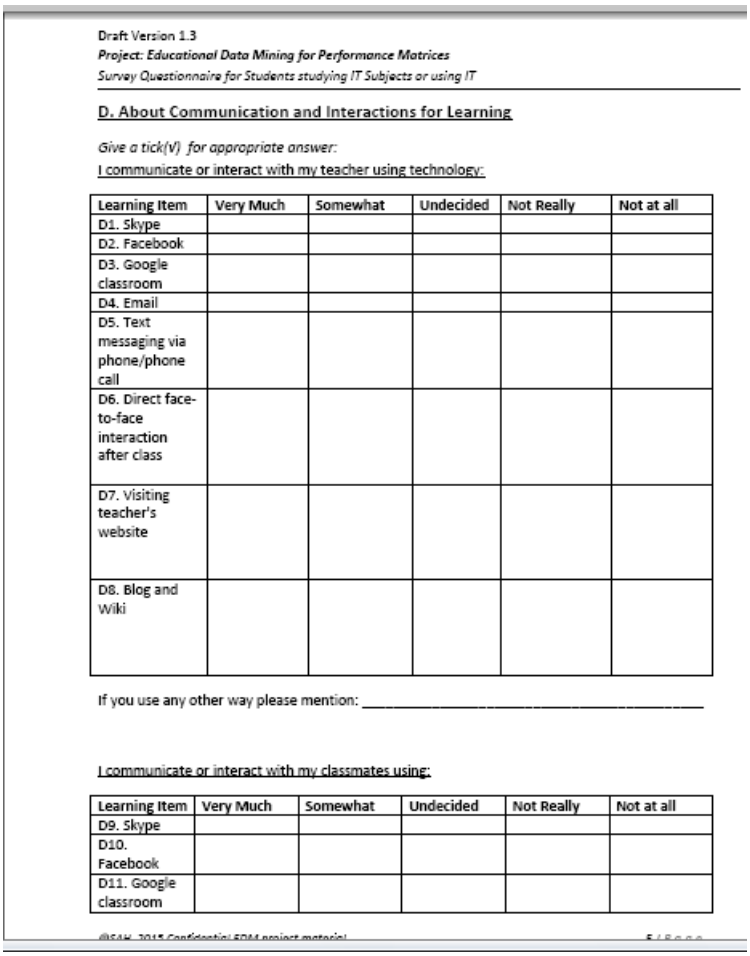
a. Multimedia b. White-board c. Any other _____

C24. You enjoy class when teacher uses: (please tick whichever appropriate)

Class Items for Interactions	Very Much (without any confusion)	Somewhat (not confident enough)	Undecided	Disagree (not confident enough)	Strongly disagree (without any confusion)
a. Example based teaching					
b. Hands-on-learning					
c. Exercise in the classroom					
d. Lecture only					
e. Online testing on knowledge					
f. Assignment at home					
g. Virtual interaction on Google					
h. Problem based learning					
i. Others:					

C25. How many course projects you have completed either in a team or individual? _____

Figure B.4: 4th Page of Questions



Draft Version 1.3
Project: Educational Data Mining for Performance Matrices
Survey Questionnaire for Students studying IT Subjects or using IT

D. About Communication and Interactions for Learning

Give a tick(√) for appropriate answer:
I communicate or interact with my teacher using technology:

Learning Item	Very Much	Somewhat	Undecided	Not Really	Not at all
D1. Skype					
D2. Facebook					
D3. Google classroom					
D4. Email					
D5. Text messaging via phone/phone call					
D6. Direct face-to-face interaction after class					
D7. Visiting teacher's website					
D8. Blog and Wiki					

If you use any other way please mention: _____

I communicate or interact with my classmates using:

Learning Item	Very Much	Somewhat	Undecided	Not Really	Not at all
D9. Skype					
D10. Facebook					
D11. Google classroom					

Figure B.5: 5th Page of Questions

Draft Version 1.3

Project: Educational Data Mining for Performance Matrices

Survey Questionnaire for Students studying IT Subjects or using IT

Learning Item	Very Much	Somewhat	Undecided	Not Really	Not at all
D12. Email					
D13. Text messaging via phone call					
D14. Direct face-to-face interaction after class					
D15. Blog and Wiki					

If you use any other way please mention: _____

D16. Which online storage do you use?

a. Google drive b. Dropbox c. I don't use online storage

d. If something else please mention:_____

E. Miscellaneous

E1. On an average I download number of e-books for education purpose per month

a. I don't prefer e-book. b. 0-3 c. 3-5 d. more than 5

E2. I learn Using Video Tutorial per week

a. 0-2 hrs b. 2-5 hrs c. 5-10 hrs d. more than 10 hrs.

E3. I use interactive conversation via internet to clear any confusion with classmate

a. 0-1hr /week b. 1-2.5 hr/week c. 2.5-5 hr /week d. more than 5 hr/week

E4.To clear any kind of confusion I browse internet for

a. I don't browse internet b. 0-1 hr per week c. 1-5 hr per week d. 5-10 hr per week

e. more than 10 hr per week

E5. If using mobile apps for learning, please mention out of 10 you would like to score:

a. 0-2.5 b. 2.5-5 c. 5-7.5 d. 7.5-10

_____ *Thank you for your valuable contributions* _____

Figure B.6: 6th Page of Questions