Analysis and Prediction of Student's Academic Performance in University Courses

Garima Sharma
Gyan Ganga Institute
of Technology and Sciences, India.

Santosh K. Vishwakarma
Gyan Ganga Institute
of Technology and Sciences, India.

ABSTRACT
Management of huge amount of data has always been a matter of concern. With the increase in awareness towards education, the amount of data in educational institutes is also increasing. The increasing growth of educational databases, have given rise to a new field of data mining, known as Educational Data Mining (EDM). With the help of this one can predict the academic performance of a student that can help the students, their instructors and also their guardians to take necessary actions beforehand to improve the future performance of a student. This paper deals with the implementation of ID3 decision tree algorithm to build a predictive model based on the previous performances of a student. The dataset used in this paper is the semester data of the students of a private institute of India. Rapidminer, an open source software platform is used to obtain the results.

General Terms
Data mining, Knowledge Discovery in Databases, Classification, ID3 induction algorithm, Performance prediction

Keywords
Data Mining, EDM, KDD, Classification, Decision tree, ID3, Student's academic performance prediction

1. INTRODUCTION
Data mining is the process of discovering interesting patterns from massive amounts of data. As a knowledge discovery process, it typically involves data cleaning, data integration, data selection, data transformation, pattern discovery, pattern evaluation, and knowledge presentation [1]. Several data mining techniques include associations, classifications, sequential patterns and clustering. Classification involves finding rules that partition the data into disjoint groups. Classification takes training data set as input whose class labels are already known. Prediction is one of the two goals of data mining, which makes use of existing variables in the database in order to predict unknown or future values of interest [2]. Educational Data Mining (EDM) is concerned with developing, researching, and applying computerized methods to detect patterns in large collections of educational data patterns that would otherwise be hard or impossible to analyze due to the enormous volume of data they exist within [3]. This paper aims to predict the performance of the students in the final examination studying in Computer Science (CS) department, Gyan Ganga Institute of Technology and Sciences (G.G.I.T.S.), Jabalpur, India. Prediction is done by considering various tests conducted during that semester and the respective performances of the students.

2. LITERATURE REVIEW
A lot of study is already been done in this area and are proven to be beneficial. The first study, entitled Predicting Critical Courses Affecting Students Performance: A Case Study [4] builds a predictive model based on records of female students using the ID3 decision tree algorithm to reveal the courses affecting low academic performance at the IT department, King Saud University, Riyadh, Saudi Arabia. They have built several models based on ID3 decision tree algorithm. They divided the dataset into three groups to build separate model for each group. Results suggested that the classification model based on performance in the second year is the most accurate. The student performance in IT 221 and the two programming courses, CSC111 and CSC113, is a great indicator of student level of achievement.

The second study entitled Predicting Students Performance in University Courses: A Case Study and Tool in KSU Mathematics Department [5] builds an application to predict student's performance in a programming course based on their previous performances in specific mathematics and English courses. In addition, the model aims to reduce dropout rates by helping students predict their performance in programming courses before enrolling for them. Two experiments were conducted using the CBA rule-generation algorithm. The first used students grades in two English courses and two mathematics courses, which generated four rules with accuracy of 62.75%. The second used students grades only in two English courses, generating four rules with accuracy of 67.33%. These results showed that students performance in English courses has a significant predictive effect on their performance in the programming course.

Third study entitled Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students [6] designed a system to justify that various data mining techniques such as classification that can be used in educational databases to suggest career options for the high school students and also to predict the potentially violent behaviour among the students by including extra parameters other than academic details. Three decision trees ID3, C4.5 and CHAID were compared for their various performance measures.

Another study entitled A Review on Predicting Students Performance using Data Mining Techniques [7] provides an overview on the data mining techniques that have been used to predict student's performance. This paper also focuses on how the prediction algorithm can be used to identify the most important attributes in a student's data. The meta-analysis is based on the highest accuracy of prediction methods and also the main important factors that may influence the student's performance. Prediction accuracy that uses classification method grouped by algorithms for predicting student's performance since 2002 to 2015 are presented. Neural Network has the highest prediction accuracy by (98%) followed by Decision Tree by (91%). Next, Support Vector Machine and K Nearest Neighbor gave the same accuracy,

which is (83%). Lastly, the method that has lower prediction accuracy is Naive Bayes by (76%).

3. METHODOLOGY

Knowledge Discovery in Database (KDD) is a process, which is used to extract useful information by performing various actions on the dataset. Data mining and KDD are often used interchangeably to serve the purpose of mining. The former method is one of the steps involved in the KDD process while the latter is the overall process of mining. KDD involves following steps:

(1) Data Selection.

(2) Data Preprocessing.

(3) Data Transformation.

(4) Data Mining.

(5) Data Interpretation and Evaluation.

3.1 Data selection and description

Data selection allows us to select a subset of data from huge database. This subset can now be used for our study. A university conducts several exams which tests the theoretical and practical knowledge of the students in that field. These exams help the students to get the better understanding of each course. A certain score is defined by the university/ institute which must be attained by every student in order to continue the degree program. It is very necessary for the student to attain that score based on which he/she can be considered to be passed. The dataset used in this paper is the semester data of the students at Computer Science (CS) department, G.G.I.T.S, Jabalpur, India. This data provides the information about students' performance during the entire semester and before the final exams. The aim of this study is to predict the performance of the students in final university exams based on their previous performance in that semester. This study will also provide the information about how much a student is learning from these tests and how well they have performed. Each semester consists of five core subjects of computer science department. Two midsem and two assignment tests are being conducted of each subject during the entire semester. A sample of original dataset is shown in Figure 1.

The dataset contains seven attributes (columns) Roll No., Name, Assignm1, Assignm2, Midsem1, Midsem2 and final. Initially the attributes Roll No. and Name are categorical attributes and the remaining are numeric attributes. The attribute final is a special attribute which defines class label. Description of each attribute is shown below:

- Roll No. and Name - These field tells the name of the student along with the unique identity number (i.e., Roll number) which is provided by the institute to that student for the degree program.
- Midsem 1 and 2 - These field tells the marks obtained by each student in Midsem exam 1 and 2.
- Assignm 1 and 2 - These field tells the marks obtained by each student in assignment exam 1 and 2.
- Final - This field shows the information about final performance of the students in that semester.

3.2 Preprocessing of data

Data preprocessing is the process of converting data into more readable and understandable format. Raw data is of no use until useful information is extracted from it. Quality of the data is the major concern while dealing with data mining methods. Data obtained from various sources may be inconsistent, redundant or of low quality that may leads us to improper and low quality data mining results. To improve the quality, and efficiency of the dataset and, consequently, of the data mining results, proper preprocessing steps for the dataset needs to be performed. Without this step proper and useful results from data mining cannot be guaranteed. These steps when applied can provide us with more accurate, useful and efficient mining results. A sample of preprocessed data is shown in Figure 2. Preprocessing is being done as follows:

		Assignm 1					midsem I					Assignm 2					midsem 2					final		
Roll No.	Name	CS501	CS502	CS503	CS504	CS505	CS501	CS502	CS503	CS504	CS505	CS501	CS502	CS503	CS504	CS505	CS501	CS502	CS503	CS504	CS505	CS501	CS502	CS503
		10	10	10	10	10	20	20	20	20	20	10	10	10	10	10	20	20	20	20	20	40	40	40
0206CS141001	AALAP SWAMI	2	4	7	2	4	7	5	2	8	3	4	4	7	4	2	7	12	16	10	8	14	17	22
0206CS141002	ABHIJEET GUPTA	4	4	8	5	4	2	3	6	5	0	6	5	7	2	7	5	10	5	3	8	12	15	18
0206CS141003	ABHISHEK CHAKRABORTY	4	4	1	5	4	7	5	4	7	4	8	8	8	3	6	3	12	0	4	8	15	20	9
0206CS141004	ABHISHEK SHRIVASTAVA	3	4	4	5	3	6	7	7	9	4	6	7	7	3	5	8	11	7	5	9	16	20	17
0206CS141005	ABHISHEK UPADHYAY	1	3	5	0	0	8	5	7	7	5	8	5	0	3	8	8	5	3	8	5	17	12	10
0206CS141006	ABHYUDAY SHUKLA	3	2.5	6	1	5	3	7	2	10	2	8	0	8	0	5	5	12	6	4	2	13	15	15
0206CS141007	ADITI VERMA	7	6	6	4	2	14	11	12	14	14	10	7	9	5	6	8	15	16	10	13	26	26	29
0206CS141008	AGNI CHATURVEDI	2	0	3	2	0	5	7	10	10	16	8	8	0	0	0	5	9	12	15	11	14	16	17
0206CS141009	AKASH ALUNG	6	5	3	0	4	2	3	5	5	3	3	0	0	1	0	5	3	6	5	2	11	8	10
0206CS141010	Akrati Sahu	5	5	3	5	4	10	5	11	5	3	10	8	10	5	6	12	8	10	13	5	25	18	23
0206CS141011	AKSHADA HUMNE	3	1	10	0	0	11	9	12	5	14	10	8	10	3	5	8	9	13	13	5	22	18	30
0206CS141012	AKSHAT TIWARI	2	4	7	1	2	5	6	1	2	1	8	3	0	0	0	2	6	8	3	9	12	13	11
0206CS141013	AKSHAY PATHAK	1	2	10	2	5	6	6	7	2	4	8	4	5	1	0	3	7	9	1	4	12	13	21
0206CS141014	AKSHAY SWARNKAR	2	4	8	1	0	5	7	4	12	15	7	5	8	6	7	7	5	6	8	6	14	14	18
0206CS141015	AMAN JAIN	6	8	8	0	8	11	7	10	7	14	8	7	7	3	8	11	10	10	7	13	24	22	24
0206CS141016	AMBER JAIN	3	2.5	4	6	5	13	11	7	5	8	9	7	5	5	5	15	5	0	14	8	27	17	11
0206CS141017	AMEYA VERMA	0	5	6	0	1	4	4	2	7	5	9	0	6	0	5	8	13	5	8	8	14	15	13
0206CS141018	AMIT KUMAR MEHRA	0	0	5	1	2	0	3	0	0	1	2	4	5	7	7	2	5	10	11	15	3	8	14
0206CS141019	ANAM KHAN	3	0	8	0	0	20	16	20	14	18	10	10	10	3	10	20	20	11	15	17	36	31	33
0206CS141020	ANAMIKA VISHWAKARM	0	3	6	4	5	9	3	5	1	2	9	8	10	0	0	5	2	8	2	5	16	11	20
0206CS141021	ANCHITA MISHRA	4	2	5	0	2	16	14	11	14	11	9	8	10	5	8	17	15	10	11	10	31	26	24

Fig. 1. Sample of Original Data Set

- Data cleaning is done to remove noise, resolve inconsistency and replace missing values from the existing dataset. The attributes which have missing values are handled by replacing them with some value of similar consistency. In our case, the original dataset contains some tuples /attributes in which students are absent in exams are left blank or having values like Abs, absent or A, are replaced by giving them a score of 0 in that exam.
- Not all attributes in the entire dataset is useful for mining. Thus data reduction is done to reduce the size of the dataset to make it suitable and easier for mining without any loss of information. The attributes which are relevant for our study are taken into consideration while others are ignored. The column which provides the name of the student is discarded because it is irrelevant for our work. We have Roll number column to differentiate among

  each student.
- Similarly, the original dataset contains all four columns (midsem1, midsem2, assignm1, assignm2) as multi-valued columns (i.e., contains marks obtained in five subjects). So every column is taken separately and an average of marks obtained in five subjects is calculated. Each test now contains the average marks of five subjects of the students in that test.

Roll No.	Midsem 1 (20)	midsem 2 (20)	Assign 1 (10)	Assign 2 (10)	Final (40)
CS 01	5	10.6	3.8	4.2	16.2
CS 02	3.2	6.2	5	5.4	13.6
CS 03	5.4	5.4	3.6	6.6	14.4
CS 04	6.6	8	3.8	5.6	16.4
CS 05	6.4	5.8	1.8	4.8	12.6
CS 06	4.8	5.8	3.5	4.2	12.6
CS 07	13	12.4	5	7.4	25.4
CS 08	9.6	10.4	1.4	3.2	16.6
CS 09	3.6	4.2	3.6	0.8	8.6
CS 10	6.8	9.6	4.4	7.8	19.4
CS 11	10.2	9.6	2.8	7.2	20
CS 12	3	5.6	3.2	2.2	9.6
CS 13	5	4.8	4	3.6	11.8
CS 14	8.6	6.4	3	6.6	16.6
CS 15	9.8	10.2	6	6.6	22.2
CS 16	8.8	8.4	4.1	6.2	18.6
CS 17	4.4	8.4	2.4	4	13
CS 18	0.8	8.6	1.6	5	11
CS 19	17.6	16.6	2.2	8.6	30.4
CS 20	4	4.4	3.6	5.4	12
CS 21	13.2	12.6	2.6	8	24.4
CS 22	11.8	8.4	5.2	5.8	21.2
CS 23	18.2	16.4	3.6	8.4	31.4

Fig. 2. Sample of Pre-processed Dataset

3.3 Transformation of data

Data discretization, a form of data transformation, is the process of reducing the number of values for a given continuous attribute, by dividing the attribute into a range of intervals. A sample of transformed data is shown in Figure 3. Certain rules are made to identify how well the students have performed in each test.

- Students who scored marks between 14 to 20 corresponds to good, marks between 7 to 13 corresponds to average and marks less than 7 corresponds to bad performance of a student in midterm exams.
- Similarly, in assignments exams, students who scored marks between 7 to 10 corresponds to good, marks between 4 to 6 corresponds to average and marks less than 4 corresponds to bad performance of a student.
- Final marks are considered as class label. Students who scored marks between 27 to 40 corresponds to good, marks between 13 to 26 corresponds to average and marks less than 13 corresponds to bad performance of a student in final exams.

Roll No.	Midsem 1 (20)	midsem 2 (20)	Assign 1 (10)	Assign 2 (10)	Final (4
CS 01	bad	avg	bad	avg	Averag
CS 02	bad	bad	avg	avg	Averag
CS 03	bad	bad	bad	avg	Averag
CS 04	bad	avg	bad	avg	Averag
CS 05	bad	bad	bad	avg	Bad
CS 06	bad	bad	bad	avg	Bad
CS 07	avg	avg	avg	good	Averag
CS 08	avg	avg	bad	bad	Averag
CS 09	bad	bad	bad	bad	Bad
CS 10	bad	avg	avg	good	Averag
CS 11	avg	avg	bad	good	Averag
CS 12	bad	bad	bad	bad	Bad
CS 13	bad	bad	avg	bad	Bad
CS 14	avg	bad	bad	avg	Averag
CS 15	avg	avg	avg	avg	Averag
CS 16	avg	avg	avg	avg	Averag
CS 17	bad	avg	bad	avg	Averag
CS 18	bad	avg	bad	avg	Bad
CS 19	good	good	bad	good	Good
CS 20	bad	bad	bad	avg	Bad
CS 21	avg	avg	bad	good	Averag
CS 22	avg	avg	avg	avg	Averag

Fig. 3. Sample of Transformed Data Set

3.4 Data Mining

Data mining is the most important step which extracts interesting patterns existing in the transformed data. Depending on the goal of the study or the type of knowledge that should be discovered, suitable data mining tasks are chosen (Here classification task is chosen). Classification technique aims to predict certain outcome based on the given set of inputs. For the prediction, ID3 (Iterative Dichotomizer 3) algorithm of decision trees, proposed by Quinlan is used. This algorithm follows a greedy approach and constructs the tree in top down recursive divide-and-conquer manner. Several models were built using this algorithm. The cross validation operator is used which depicts the number of values predicted wrong in the model. This operator consists of two sub-processes, i.e., the training sub-process and the testing sub-process. In the training section, ID3 algorithm is applied along with filter examples operator to build the model.

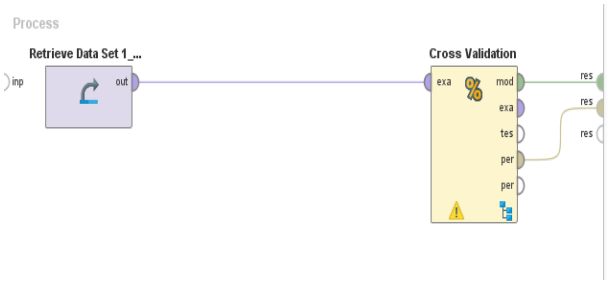

Fig. 4. Cross Validation operator.

The filter examples operator is used to reduce the number of observations. Figure 4 depicts how the Cross validation operator is applied. In the testing section, model is applied to check its performance by using apply model and performance operator. Figure 5 depicts various operators used inside cross validation. The dataset consists of 70 records in total. Initially we put the entire data into use to build a model and test its

performance on the previously known data. In order to check the performance of the model on unknown data we divide the entire dataset into training and testing datasets. Hence 50 records were used for training and the remaining 20 records were used for testing.

3.5 Interpretation and Evaluation of data

By applying the data mining technique (here classification), a pattern is obtained from the transformed data. This step assigns meanings to the result obtained from the previous step. Based on the given training set, the ID3 algorithm will builds a decision tree which is depicted in Figure 6.

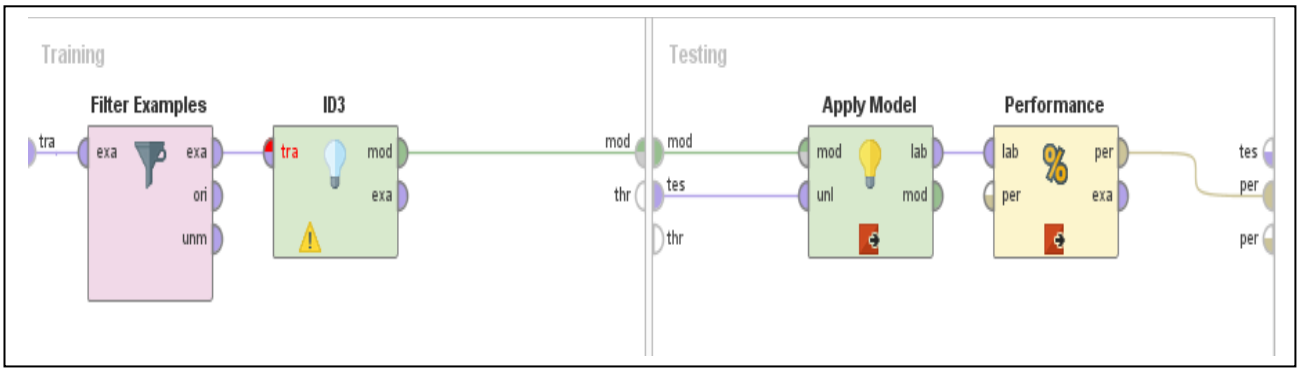

Fig. 5. Configuration inside cross validation

Following rules can be inferred from the decision tree:

1. Rule 1 - if Midsem1 performance = average, then the final performance = average.
2. Rule 2 - if Midsem1 performance = bad, Assign2 performance = average and Assign1 performance = average, then the final performance = average.
3. Rule 3 - if Midsem1 performance = bad, Assign2 performance = average, Assign1 performance = bad and Midsem2 performance = average, then the final performance = average.
4. Rule 4 - if the performance of the students in Midsem1 performance = bad, Assign 2 performance = average, Assign 1 performance = bad and Midsem2 performance = bad, then the final performance = bad.
5. Rule 5 - if Midsem1 performance = bad and Assign 2 performance = bad, then the final performance = bad.
6. Rule 6 - if Midsem1 performance = bad and Assign 2 performance = good, then the final performance = average.

Rule 7 - if Midsem1 performance = good and Midsem2 performance = average, then the final performance = average.

7. Rule 8 - if Midsem1 performance = good and Midsem2 performance = good, then the final performance = good.

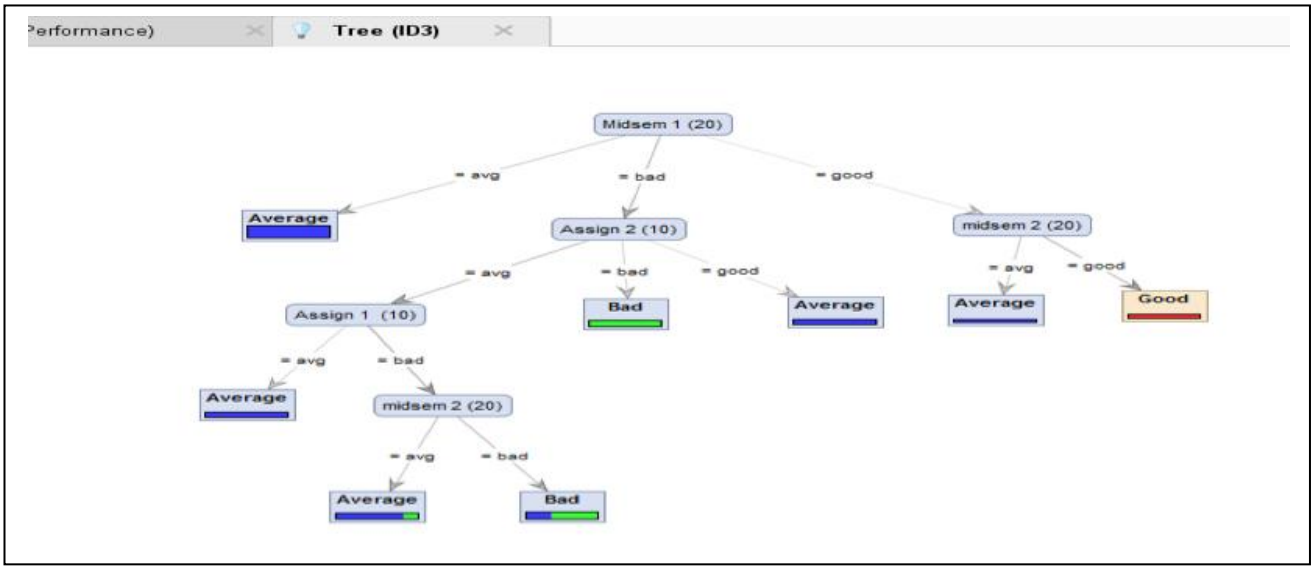

Fig. 6. Tree generated after applying ID3 algorithm.

The apply model operator applies trained data model on the test data for prediction purpose. In our study, we want to predict the final performance of the students in that semester based on previous performance. Prediction is shown in Figure 7.

Evaluation is done to check the performance of the model over unseen data. Evaluation is based on cross validation. To evaluate the model, three performance measures are used

Roll No.	Final (40)	prediction(Final (40))
CS 51	?	Good
CS 52	?	Average
CS 53	?	Average
CS 54	?	Average
CS 55	?	Average
CS 56	?	Average
CS 57	?	Average
CS 58	?	Average
CS 59	?	Bad
CS 60	?	Bad
CS 61	?	Average
CS 62	?	Average
CS 63	?	Average
CS 64	?	Bad
CS 65	?	Bad
CS 66	?	Average

Fig. 7. Predicted values of Final performance

(1) Accuracy - Accuracy is the total number of values that are classified correctly.

Accuracy = (TP + TN)=(P + N) (1)

(2) Precision - Precision is the number of the predicted positive values that were correct.

Precision = TP=(TP + FP) (2)

(3) Recall - Recall is the number of positive values that were correctly identified.

Recall = TP=P (3)

Where TP, TN, FP, P, N refer to the number of true positive, true negative, false positive, positive, and negative samples, respectively. Accuracy in our model is comes out to be 90%. Performance of model is presented in the form of confusion matrix.

4. CONCLUSION AND FUTURE WORK

This study makes use of the ID3 decision tree algorithm to predict the final performance of the students based on their previous performances. The algorithm uses the criteria of information gain to make splitting. The more the information gain of an attribute, the more is the split possibility of that attribute. Through this model, we have achieved the accuracy of 90%. We hope that this study will help the students, instructors and the guardians to take necessary action to improve the performance of the students in future. For future work, we would like to refine our work by taking more number of example set and come up with more accuracy and other techniques to help students in their educational careers.

accuracy: 90.00% +/- 8.94% (mikro: 90.00%)

	true Average	true Bad	true Good	class precision
pred. Average	32	1	0	96.97%
pred. Bad	3	11	0	78.57%
pred. Good	1	0	2	66.67%
class recall	88.89%	91.67%	100.00%	

Fig. 8. Confusion Matrix

5. REFERENCES

[1] Han, J., Kamber, M., Pei, J.. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann; 3rd ed.; 2012. ISBN 978-0-12- 381479-1.

[2] Pujari, K., A.. Data Mining Techniques. ISBN-10:8173713804.

[3] Scheuer, O. & McLaren, B.M. (2011). Educational Data Mining. In the Encyclopedia of the Sciences of Learning, Springer.

[4] Altujjar, Y., Altamimi,W., Al-Turaiki, I., Al-Razgan, M., Predicting Critical Courses Affecting Students Performance: A Case Study. Procedia Computer Science 82 (2016) 65 71.

[5] Badr, G., Algobail, A., Almutairi, H., Almutery, M., Predicting Students Performance in University Courses: A Case Study and Tool in KSU Mathematics Department. Procedia Computer Science 82 (2016) 80 89.

[6] Elakia, Gayathri, Aarthi, Naren J., Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4649-4652.

[7] Shahiri,A., M., Husain,W., Rashid,N.,A., A Review on Predicting Students Performance using Data Mining Techniques. Procedia Computer Science 72 (2015) 414 422.

[8] Rapidminer Studio Documentation http://docs.rapidminer.com/studio/