Using Data Mining Classifier for Predicting Student's Performance in UG Level

Surbhi Agrawal
Gyan Ganga Institute
of Technology and Sciences,
Jabalpur, India

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

Akhilesh K. Sharma
Manipal University Jaipur
India

ABSTRACT
In the realm of digitalization, the competition in the field of education has expanded drastically. To analyze this increment the education data mining has played a vital role. In this paper, student's historical record and the relevant features like their living habits, backgrounds and so forth are utilized as data set (corpus). The performance of students is evaluated using four distinct classifiers named as decision tree, random forest, naive bayes and rule induction. Different classifiers show different accuracy depending on different algorithms used in it. These analyzed results are explicitly used to predict the upcoming grades of the students and the relevant features (like access to the Internet, study time, etc.) which affect the academic performance of the students.

General Terms
Education data mining, Classifiers, Algorithms, prediction, Students performance index

Keywords
Corpus, decision tree, random forest, naïve bayes, rule induction, Data mining.

1. INTRODUCTION
Most importantly, data mining, it is a procedure of analyzing the given data collection from different perspective and discovering the information and knowledge out from it, by summering it. The outcomes provide the co-relations or patterns among the data inside the data set.

When the data of data set is originated from education environment, the mining done on such data set is termed as education data mining. Now days, education data mining is done on the high scale as education has become remarkable for the personal and economic growth of an individual. Figure 1 shows the steps involved throughout the process of education data mining.

The main objective of the education institute is to achieve success by minimizing the failure rate of students. The success of the education institute is hidden inside the quality of the education system, and the academic rules and regulations which they follow. This objective can be achieved by performing analysis of the academic performance of the students and the factors affecting the student's performance. Further this analysis can be effectively utilized to recommend the decision makers of the education institute.

As in this work, the analysis on the previous academic records of the students is performed. Using the outcome of the analyses, prediction for the coming final semester grades is made so that the failure rate can be minimized by giving extra attention towards the weak students. This further enhances the overall result of the education institute.

2. REVIEW OF LITERATURE
Nagy et al. [2], gave an intelligent student advisory frame work for education dome. This frame work advises first-year university students to follow a certain academic path depending on their academic records. They aim by proposing this framework to increase success rate of students by minimizing the failure rate. They have classified [3] and also clustered students into suitable departments, then at last combined the results for accurate outcome.

Tair et al. [4], gave a case study, to show the importance of data mining to improve the performance of graduate students by applying several data mining techniques and their methods such as association rule, classification, clustering [5] and outlier detections [8].

Shovon et al. [6], performed clustering and produced a decision tree on student data sets in order to predict student's capability using the data mining tools so that teachers can take necessary steps before the final exam to reduce the dropout rate.

Cortez et al. [11], predicted the student's grades by analyzing the past obtained grades and found that school environment, family environment and social circle where a student resides also affect the student's academic performance. For this, they have tested binary and five-level classifications and regression. They have used different data modules, which are decision tree, random forest, neural network, support vector machine.

Jovanovic et al. [12], have applied classification for analysis and predicting student's academic performance and further also have applied clustering to cluster students on the basis of cognitive style in e-learning. At last, they have given a module which helped teachers to distinguish students depending on their academic strength. So that extra attention could be given to weak students.

3. METHODOLOGY
3.1 Materials (Data Set)
Table 1 shown below is the metadata of the data set which is taken online. This metadata includes family, social, and academic information of the students from two schools, of Portugal during the year 2005 – 2006. The metadata includes the attribute with their data types.

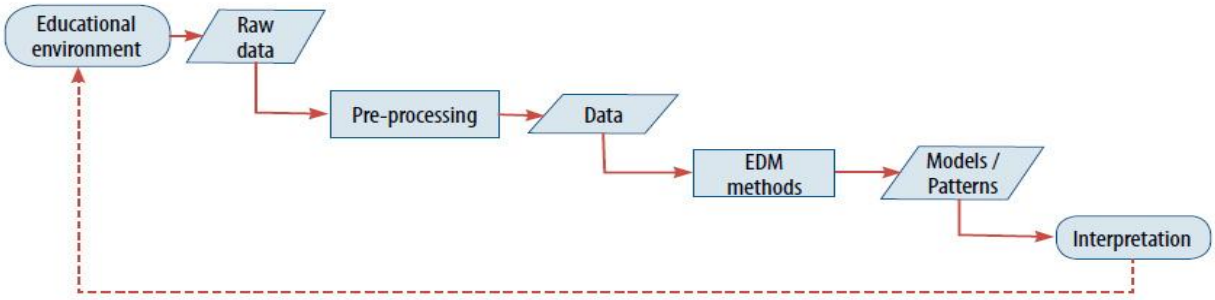

Fig 1: Education data mining process model

Table 1. The preprocessed student related variables

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: Gabriel Pereira or Mousinho da Silveira)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4a)
Mjob	mother's job (nominalb)
Fedu	father's education (numeric: from 0 to 4a)
Fjob	father's job (nominalb)
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
failures	number of past class failures (numeric: n if 1 ≤ n < 3, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Further, for this work the table 1 has been pruned and limited to eight attributes with 100 examples, including the information of 100 students from two schools. The data set is in excel format. Figure 2 shows the screenshot of the training data set, including attribute and values. The description of the eight attribute is given below:

1 School: This field defines the name of the school where they study.

2 Absence: This is a numeric field which states the number of absence of the student in school.

3 Activity: This is a binomial field which defines 'yes' if the student takes part in extra activity organized in school and 'no' for vice versa.

4 Study Time: this is a numeric field which defines the number of hour's student study at home.

5 The internet : This is a binomial field, which states 'YES' if the student has the Internet facility at home and 'NO' for vice versa.

6 G1: This is a polynomial attribute of the data set which defines the grades 'A, B, C, D' of the first pre-semester of the students.

7 G2: This is a polynomial attribute of the dataset which defines the grades 'A, B, C, D' of the second pre-semester of the students.

8 G3: This is a polynomial attribute of the dataset which defines the grades 'A, B, C, D' of the Final semester of the students. This is a labeled attribute.

In the Test dataset G3 is not present as the final semester grades of the students is to be predicted. So that extra attention can be given to the weak students can be given.

3.2 Computational Environment (Rapid Miner)

All the experiment conducted for this work is carried out using an open source integrated software platform named rapid miner developed by the company with the same name rapid miner. This environment is used for applying various data mining techniques for extracting several hidden information and knowledge from the available data set. It is used in various applications like business, commercial, research, training, development, education, prototyping, etc.

3.3 Methods And Models Used

3.3.1 Method: classification

Classification is a method of data mining, which is used to recognize, differentiate, categorize, and understood items to the predefined class. Classification is a supervised machine learning method where the machine is made to learn through examples. So the machine is trained by providing dataset as training data in training phase. This can also be viewed as analyzing the data. Figure 3 shows the screenshot of the training phase, where the dataset is given to the cross validation operator which uses different classifiers, which stores the understanding in form of models in the store.

	A	B	C	D	E	F	G	H
1	school	Absence	Activities	Studytime	Internet	G1	G2	G3
2	CS	4	Y	2	Y	C	B	B
3	CS	2	Y	3	Y	C	C	C
4	XS	6	N	4	Y	A	D	B
5	XS	8	N	5	Y	A	D	B
6	CS	4	Y	2	Y	C	C	C
7	CS	1	Y	3	N	D	B	C
8	XS	9	N	4	Y	A	C	B
9	CS	0	Y	2	Y	C	D	C
10	XS	8	N	5	Y	A	C	B
11	CS	1	Y	3	N	D	B	C
12	CS	3	Y	2	Y	C	A	B
13	CS	0	Y	3	Y	C	C	C
14	XS	6	N	4	Y	A	B	A
15	CS	3	Y	2	N	D	A	C
16	CS	1	Y	3	N	D	D	D
17	CS	2	Y	2	Y	C	B	B
18	XS	6	N	5	Y	A	A	A
19	CS	1	Y	2	Y	C	B	B
20	CS	2	Y	3	Y	C	B	B
21	XS	9	N	4	N	B	D	C
22	CS	3	Y	2	N	D	B	C
23	CS	2	Y	3	Y	C	B	B
24	CS	1	Y	2	Y	C	D	C
25	CS	4	Y	3	Y	C	C	C

Fig 2 : Database Screenshot

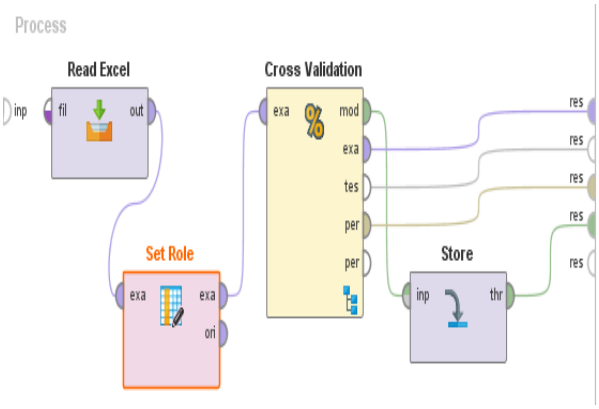

Fig 3 : Cross validation, Training Phase

Figure 4 shows four different classifiers, which are used during this phase. The accuracy given by each classifier at different folds is calculated here.

Then the understanding of the system is tested. In one sense, the system is asked to categorize the items or examples of the test data to different predefined classes. Or can say, the system is asked to predict the class of each item of the data set. Here the final semester grades are predicted (i.e. G3 attribute) of the students on the basis of previous grades of two pre-semesters. Figure 5 shows the screenshot of the operators used in predicting (testing) phase in rapid miner. There are several models under classification for this work. In this work four types of models are used namely decision tree, random forests, naïve Bayes [19], rule induction. They are discussed further.

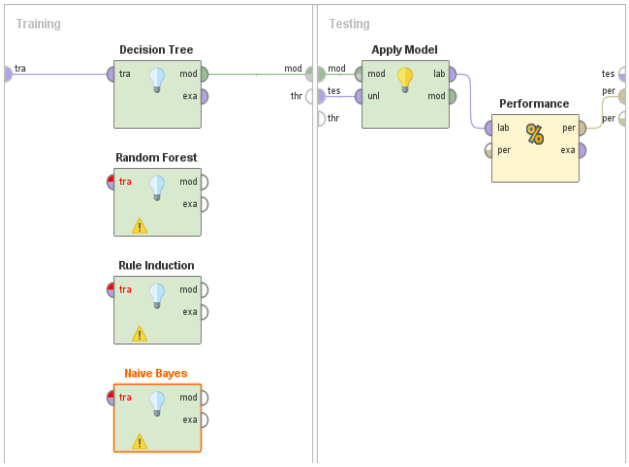

Fig 4 : Process inside cross validation, Training Phase

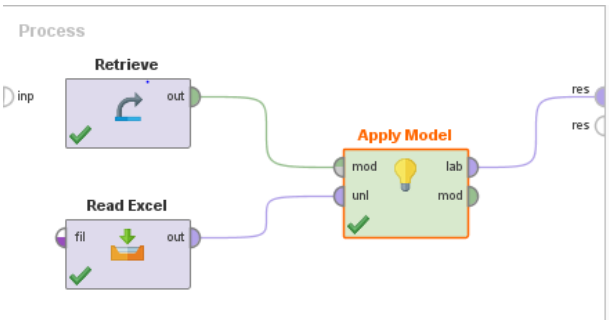

Fig 5 : Testing Phase screenshot

3.3.2 Models

3.3.2.1 Decision Tree

Decision tree is the way of modeling data in tree-like structure. It is a predictive model in tree form, where the top node is called root node, and last node is called leaf node, which is the outcome of the data. The decision Tree can be modeled using c4.5 algorithm of by using CART algorithm. The decision tree is used to classify the items into predefined classes. When the decision tree is used for classification purpose, it is also known as classification tree. "If, then" rule can be further induced from the decision tree to understand the data well and categorize them into respective classes correctly [7].

3.3.2.2 Random Forest

Random forest is an ensemble learning method for classification, in which it constructs multiple of unpruned classification trees in the training phase, by bootstrap sampling method on the training data. The final predicted output for a random selected feature is given by finding the mean from all unpruned classification trees in the testing phase [18].

3.3.2.3 Naive Bayes

Naive Bayes classification is the extended form of Bayesian classifiers which include naïve assumption too. Bayesian classifiers are statistical classifiers, which is based on Bayes' theorem. Bayesian classifiers can predict probability that a given sample belongs to a particular class, i.e. can predict class membership probabilities. According to the naïve assumption, the changes in an attribute value on a given class are independent of the changes in the values of the other attributes. This assumption is also known as 'class conditional independence' [19].

Suppose there are k different classes denoted as C_1, C_2 …C_k. Let X = {x_1, x_2 … x_n}, depicting n measured values of the n attributes, A_1, A_2. . .A_n respectively. Then X is predicted to belong to the class C_i if and only if

$P(C_i|X) > P(C_j|X)$ for $1 \le j \le m$, $j \ne i$.

Where P(C|X) is the posterior probability of class C conditioned on predictor X, P(C) is the prior probability of class C,P(X|C) is the probability of predictor X conditioned on class C and P(X) is the prior probability of predictor X. By Bayes' theorem:

$P(C_i|X) = (P(X|C_i)\, P(C_i))/ P(X)$

Mathematically, the naive assumption can be written as:

$$p(X \mid C_i) \approx \prod_{k=1}^{n} p(x_k \mid C_i)$$

This can be written as:

$$p(C_i \mid X) \propto p(C_i) \prod_{k=1}^{n} p(x_k \mid C_i)$$

3.3.2.4 Rule Induction

Rule induction uses sequential covering algorithm to extract the "if, then" rule, and they are directly extracted from the training data set. These rules are learned sequentially. One rule covers multiple examples present in the database hence, termed as the sequential covering algorithm. The collection of rules extracted represents full model. [4]

4. RESULT

In this study, two datasets are used, one for the training phase and another for the predicting (testing) phase. In the training phase, the training data got analyzed by different classifiers and the accuracy percentages given by each classifier at different folds are noted. Table 2 shows the accuracy percentage given by four different classifiers at five different numbers of folds. The graph shown in the Figure 6 clearly shows that decision tree at 50 numbers of folds gives the highest accuracy percentage of 90.00. Figure 7 shows the screenshot of the highest accuracy percentage given by decision tree. In the testing phase, test data was applied to predict the final semester grades (G3) of the students. Different classifiers gave different prediction of the grades. Figure 8 shows the prediction given by the highest accuracy classifier i.e. decision tree.

Table 2. Accuracy table given by different classifiers

Folds	Decision Tree	Naive bayes	Random Forest	Rule induction
10	88.00%	82.00%	84.00%	86.00%
20	88.00%	85.00%	82.00%	82.00%
30	89.72%	84.72%	86.67%	84.72%
40	87.50%	85.42%	87.08%	83.33%
50	90.00%	84.00%	85.00%	82.00%

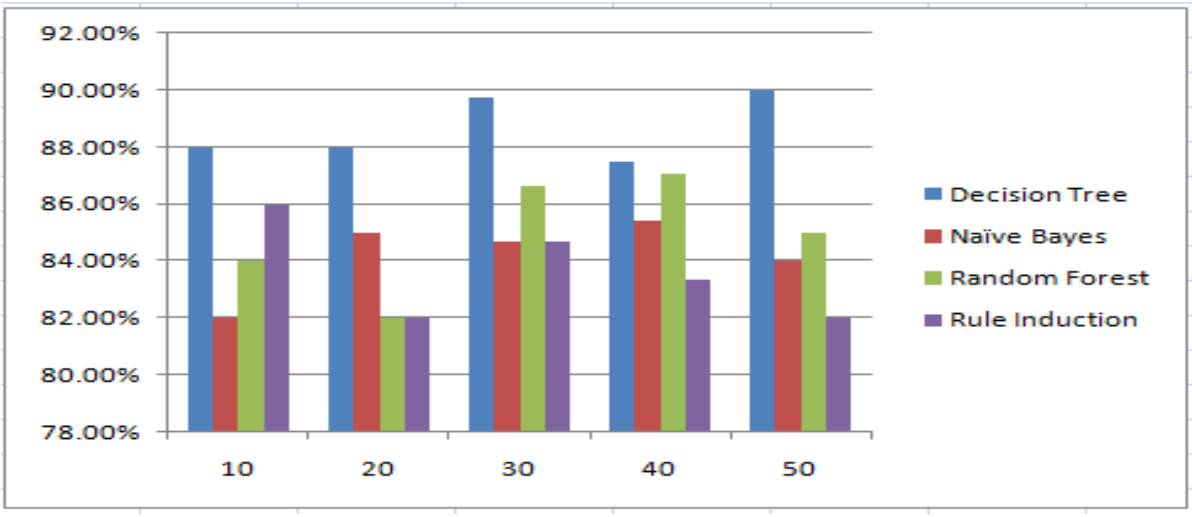

Fig 6 : Accuracy performance given by different classifiers at several folds

accuracy: 90.00% +/- 20.00% (mikro: 90.00%)

	true B	true C	true A	true D	class precision
pred. B	28	1	1	0	93.33%
pred. C	4	48	0	2	88.89%
pred. A	2	0	10	0	83.33%
pred. D	0	0	0	4	100.00%
class recall	82.35%	97.96%	90.91%	66.67%	

Fig 7 : Training Phase, Highest Accuracy Screenshot (Decision Tree)

ExampleSet (49 examples, 5 special attributes, 7 regular attributes) Filter (49 / 49 examples): all

Row No.	prediction(G3)	confidence(B)	confidence(C)	confidence(A)	confidence(D)	school	Activities	Studytime	Internet	Absence	G1
1	B	0.500	0.500	0	0	CS	Y	2	N	7	C
2	C	0	1	0	0	CS	N	3	Y	5	A
3	B	1	0	0	0	XS	Y	4	N	0	B
4	B	0.609	0.391	0	0	XS	N	5	Y	4	A
5	A	0.091	0	0.909	0	CS	Y	4	Y	6	D
6	B	0.500	0.500	0	0	XS	Y	5	N	7	D
7	B	0.500	0.500	0	0	CS	Y	2	N	7	A
8	B	1	0	0	0	XS	Y	3	N	4	B
9	B	0.500	0.500	0	0	CS	Y	4	N	6	B
10	C	0	0.833	0	0.167	CS	N	2	Y	5	C
11	C	0	0.833	0	0.167	CS	Y	4	Y	5	C
12	C	0	1	0	0	XS	Y	2	N	4	A
13	B	1	0	0	0	XS	N	5	Y	5	B
14	B	0.609	0.391	0	0	CS	Y	3	N	3	A
15	B	1	0	0	0	CS	N	2	Y	4	B
16	B	0.500	0.500	0	0	XS	Y	3	N	6	A

Fig 8 : Testing Phase, the screenshot shows the predicted values of G3 by decision tree

5. CONCLUSION

In the current educational system, there have been different issues related to student's performance evaluation and assessment. In this paper, few data mining techniques have been experimented. The prediction of the final semester grades of the students is generated before the actual final semester exam held. This can help the education institutes to minimize the dropping of the overall result by giving extra attention to the weak students. The results claimed in the paper are useful for the further evaluation and predication and can be vividly applied in many other fields. In future, this paper can be further extended to increase the accuracy percentage by using several other classifiers at different numbers of folds whereas, there can be several other relevant features other than used in this study, that can be taken into consideration for better and actual results.

6. REFERENCES

[1] Haque and Shovon 2012. Prediction of student's academic performance by an application of k-means clustering algorithm. International Journal of Advanced Research in Computer Science and Software Engineering.

[2] Nagy, Aly, and Hegazy 2013. Introduced an Educational Data Mining System for Advising Higher Education Students. World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering.

[3] Bhardwaj and Pal 2011. Data mining: A prediction for performance improvement using classification. (IJCSIS) International Journal of Computer Science and Information Security.

[4] Tair and El-Halees 2012. Mining Educational Data to Improve Students' Performance: A Case Study, International Journal of Information and Communication Technology Research.

[5] Bakar, Mohemad, Ahmad and Deris. A Comparative Study for Outlier Detection Techniques in Data Mining. 2006 IEEE.

[6] Shovon and Haque 2012. An approach of improving student's academic performance by using k-means clustering algorithm and decision tree. (IJACSA) International Journal of Advanced Computer Science and Applications.

[7] Q. Al-Radaideh, E. Al-Shawakfa and M. Al-Najjar 2006. Mining Student Data Using Decision Trees. International Arab Conference on Information Technology.

[8] S. Ramaswamy, R. Rastogi and S. Kyuseok 2000. Efficient algorithms for mining outliers from large datasets. In Proc.of the ACM SIGMOD International Conference on Management of Data.

[9] E. Chandra and K. Nandhini 2010. Knowledge Mining from Student Data. European Journal of Scientific Research.

[10] Kifaya 2009. Mining student evaluation using associative classification and clustering.

[11] Dr. V. Kumar and A. Chadha 2011. An Empirical Study of the Applications of Data Mining Techniques in Higher Education. (IJACSA) International Journal of Advanced Computer Science and Applications.

[12] P. Veeramuthu, Dr. R. Periyasamy and V. Sugasini 2014. Analysis of Student Result Using Clustering Techniques. (IJCSIT) International Journal of Computer Science and Information Technologies.

[13] Mrs. P. Patil and Dr. R. S. Kamath 2016. Assessment of Graduate Students' Performance using Data Mining: Proposed Research. IJSRD - International Journal for Scientific Research & Development.

[14] Breiman and Leo. Random forests. Machine learning 2001. Communications of the IBIMA 2009 .

[15] P. Cortez and A. Silva. Using data mining to predictsecondary school student performance.

[16] M. Jovanovic, M. Vukicevic, M. Milovanovic and M. Minovic 2012. Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. International Journal of Computational Intelligence Systems.

[17] K. Umamaheswari and S. Niraimathi 2013. A Study on Student Data Analysis Using Data Mining Techniques. International Journal of Advanced Research in Computer Science and Software Engineering.

[18] R. Jindal and M. D. Borah 2013. A survey on educational data mining and research trends. International Journal of Database Management Systems.

[19] McCallum, Andrew and K. Nigam 1998. A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization*.