International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Predicting Student's Academic Performance using Data Mining Techniques

Surbhi Agrawal, Santosh K. Vishwakarma

Abstract: ***To meet the change in world in terms of digitalization and progress, the need and importance of education is known to everyone. The increasing awareness towards and digitization has given rise to increase in size of education field's database. Such database contains information about students. The information includes students behavior, their family background, the facility they have, the society environment which surrounds them, their academic records etc. The increasing technology in data sciences can help utilize this huge education field database in a productive way by applying data mining on it. When the techniques of Data mining are applied on the database relating education records, then this process is called as education data mining. This process helps us understand the area and the students on whom the attention and the amendments are required. This increases the level of education system and also affects the success rate and understanding of the students in academics in positive direction. In this paper four different classification algorithms are used to predict grades of the students, by referring student's previous academic records. Out of the four algorithms, the one which gave the most accurate prediction is considered as the final prediction. The performance accuracy of different algorithm is compared through accuracy performance percentage.***

Keywords: ***Accuracy performance percentage, Data mining, Algorithm, Classification, Education data mining, Data sciences.***

I. INTRODUCTION

Data mining is a process in which the data is being studied deeply, understood deeply and analyzed hard to extract, dig out or can be said to mine useful information and knowledge out of it about the data and field to which data relates. In world of digitization, the size and type of data is increasing rapidly, this increase in size and complexity of data has given rise to data mining. So that effective utilization of such data could be made. Data mining evolves out with great facts, figures, patterns and ideas about the data which helps data related areas, organizations and institute to grow really well with good profits. Using this method we can even predict the future market about any product, field or even about the organization. Our work relates to education data mining, so most importantly, education data mining is exactly extracting useful, hidden, potential data, important facts, figures and correlation among data which has emerged out from any education field that may be any education institute coaching organizations etc. The main goal of education data mining is high up the level of education and understandings. This could be done if somehow we could predict the week students, so that extra facility and attention could be given to them or area where education institute really need to change their plans, rules or way of running their organizations in whole the area which is leading to failure. The data mining can even help suggest students about their future study line depending upon their interest or past performance and achievements etc.

In this work four different classification algorithms are used namely rule induction decision tree, random forest and Naive Bayes for mining education database which is taken online. The online taken database is pruned first, it contains other related information of students also like social circles, family environment, facilities student using, past academic records, etc. Using data mining algorithms and its techniques, we can predict the performance of the students in final semester, so that the needy students can be given extra attention, to get better result in future as compared to the one predicted.

II. REVIEW OF LITERATURE

Haque and Sovon [3] gave a model, using which final exam results of the student can be predicted in advanced on the basis of student's performance in mid exams, assignments of mid and final exam submitted by students, quizzes etc. The final exam predictions are then used to give extra attention to week students. In this paper, students are clustered on the basis of their academic records. For clustering, K-means clustering algorithm is used.

Bharadwaj and Pal [4] gave a predictive model which can predict future performance of the students on the basis of students living location, family environment, social life, parent's qualification and occupation, academic records of students etc. This paper have used Bayesian classification for predicting results, have successfully classified slow and higher learners students and showed that student's other relative information apart from academic records also effects the student's performance.

Bakar et al [5] apart from predicting student's future performance, they have detected outliers in the education data and states that outliers should not be ignored as it can lead to loss of valuable information. They have used association rule, classification, clustering and there different algorithm for caring out the task.

Revised Manuscript Received on February 05, 2020.
* Correspondence Author
Surbhi Agrawal, computer science, Gyan Ganga Institute of Technology and Sciences, Jabalpur, India. Email: Surbhiagrawal177@gmail.com
Santosh K. Vishwakarma, computer science, Manipal University, Jaipur, India. Email: Santoshscholar@gmail.com

They have also compared statistical based linear regression and control chart technique and found control chart technique working better. Further they perform outlier detection using distance based Manhattan technique and found distance based Manhattan technique to be the best out of all when compared.

Al-Radaideh et al [7] gave a method using which performance of the students could be predicted. The study here is restricted to only those students who took C++ course. Decision tree classification is used to predict result. To build the model CRISP - DM methodology is used. Further they compared accuracy performance percentage of three classification algorithms which are ID3, Naive Bayes and C4.5 using two methods namely K-cross validation method (K-CV) and holdout method.

Chandra and Nandhini [10] have performed data mining on records containing failure as results they have used association rules to find facts and figures among students who got failed and the courses in which they got failed. At last they have given facts in form of rules which can leads to failure. This actually helps students to get success, increase success rates, planning and plotting.

III. METHODOLOGY

A. Corpus

Database on which education data mining is done in this work is taken online. This database contains information of students from the school situated in the Portugal. The information of students contains academic performance of the students, social circle of the students, family environment of the students, life style, daily routine, parents education and occupation, the facilities students has etc. As these information also affects the student's studies.

- *Training Database:*

The online taken database is pruned first before performing data mining on it, and limited to eight attributes and hundred examples giving eight behaviors of hundred students. The attributes with their data types are described in table 1.

Table- I: Training Database

Attributes	Data type
School Name	Character
Absence of students	Numeric
Extracurricular Activity	Binomial (yes and no)
Study Time	Numeric
Internet availability	Binomial (yes and no)
Grades of semester 1(G1)	Polynomial (A, B, C, D)
Grades of semester 2(G2)	Polynomial (A, B, C, D)
Grades of final semester(G3)	Polynomial (A, B, C, D)

Figure 1 shows the screenshot of the training database. This training database is given to train the machine to build the model. The model build can predict the performance of the students in final exam. The value to be predicted is present in the training database as this database is used to first train machine. This makes machine understand that on what factors does value to be predicted depends on.

	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. 1.Screenshot of training database

- Testing Database:

Test database is the database used to test the understanding of machine on which machine. Machine used to summarize this understanding inform of rules inside the model. This model is further used to find the proposed result or labeled attribute. Here predicting final semester grades G3 of students is the labeled attribute. Test database is same as real problem which model will be encountered too. The figure 2 below shows the screenshot of the test database.

school	Activities	Studytime	Internet	Absence	G1	G2
CS	Y	2	N	7	C	A
CS	N	3	Y	5	A	C
XS	Y	4	N	0	B	A
XS	N	5	Y	4	A	B
CS	Y	4	Y	6	D	D
XS	Y	5	N	7	D	C
CS	Y	2	N	7	A	D
XS	Y	3	N	4	B	B
CS	Y	4	N	6	B	A
CS	N	2	Y	5	C	D
CS	Y	4	Y	5	C	D
XS	Y	2	N	4	A	C
XS	N	5	Y	5	B	B
CS	Y	3	N	3	A	B
CS	N	2	Y	4	B	A
XS	Y	3	N	6	A	A
CS	Y	4	Y	5	D	B
CS	Y	2	Y	3	D	C
CS	Y	3	N	6	A	C
XS	Y	4	N	6	C	D
CS	N	2	N	3	B	B

Fig. 2.Screenshot of test database

B. Computational Environment

In our work we have used RapidMiner as our tool to perform data mining on the education database. RapidMiner is a software platform which is available in internet at free of cost and is developed by RapidMiner. This software platform provides the environment for machine learning, data preparation, text learning, deep learning, predictive learning and descriptive learning integrated at one place. RapidMiner is widely in use in the business, education, training, commercial, research, prototyping, application development and in machine learning process, etc. Earlier RapidMiner was known by the name YALE (yet another learning environment). RapidMiner is written in Java programming language. RapidMiner have used very intelligent GUI (graphics user interface) and well define work flow. RapidMiner's functionality can be further extended using R and Python scripts through plugins which are made available in RapidMiner market place. This software has millions of downloads. RapidMiner stands at the top in the field of data science.

C. Methods and Algorithms Used

- Methods:

In this work, classification is used to classify the students in different predefined classes. Here predefined classes are 'A', 'B', 'C', 'D' which are the grades of final semester of the students. The grades of the final semester (G3) are the labeled attribute. In one sense we can say that predicting the grade G3 of the students. We have used classification for carrying out this task because the classes are predefined well. Classification uses supervised learning method to train the machine and summarize the understanding inside the model. This model can be inform of set of rules, tree structured etc. Supervised learning method make machine to learn through example and create understanding about the result in different situation. These different situations are depicted through several different examples given as training database during training phase of the machine. During training phase, machine goes through each example in the training database using different algorithms but one at a time. As every algorithm has different technique to train the machine. This difference in technique of making machine to learn creates difference in it performance accuracy percentage. Performance accuracy percentage is the accuracy percentage of the algorithm with which they performed the task. In this paper four different classification algorithms are used for performing classification which are Rule induction, decision tree, Naive Bayes and Random Forest. After training phase, the understanding of the machine is tested using test database in testing phase. Here the set of examples are given to the machine, which are termed as testing database, in which labeled attribute is not present as is asked machine to predict the classes or values of labeled attribute i.e. here final semester grades G3 of the students. The prediction given by the algorithm having highest performance accuracy percentage is considered as the final prediction for the student's final semester grades G3.

The figure 3 shows training phase, the classification is done using cross validation operator in the RapidMiner tool. The figure 4 shows the process inside the cross validation where four different algorithm are used for classification. The figure 5 shows the testing phase.

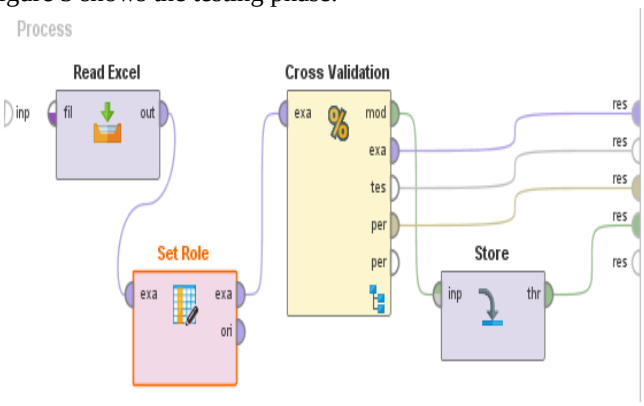

Fig. 3.Cross Validation,Training Phase

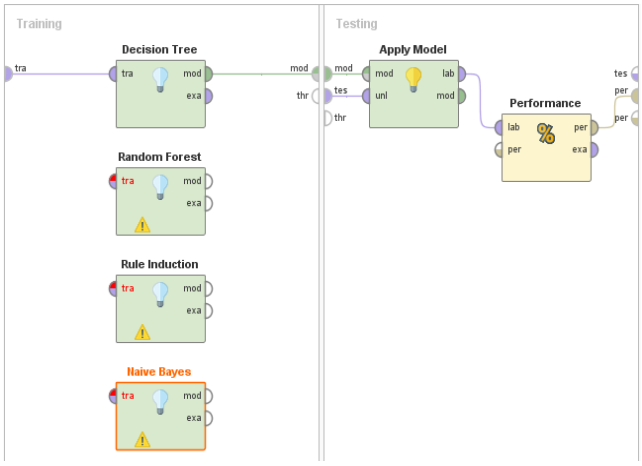

Fig. 4.Process inside Cross Validation,Training Phase

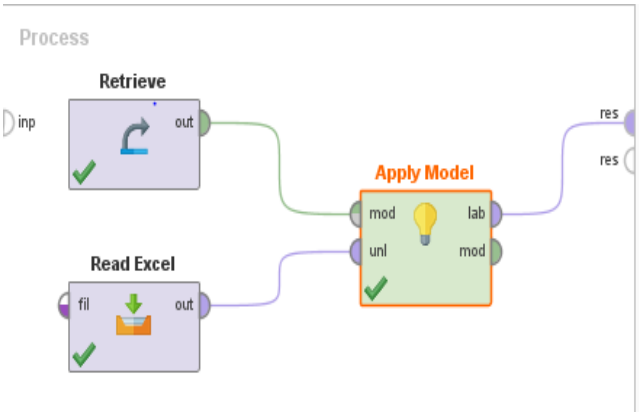

Fig. 5.Training Phase

- Algorithms:

a. Decision Tree:

This is a technique used for performing classification on the database. Various algorithms like CART, C4.5, CHAID, ID3 uses this techniques for performing classifications of the data set.

In this technique, the understanding of the machine is created through examples which are given in the training data set. The understanding is summarized inside the model inform of trees. This tree is termed as decision tree. Further using this tree, rules are formulated. These rules are inform of "if - then". The learning is given to the machine so that it can classify any upcoming example into predefined classes or can be said that the machine could predict the classes or value for the labeled attribute. As classification and prediction is closely related to each other.

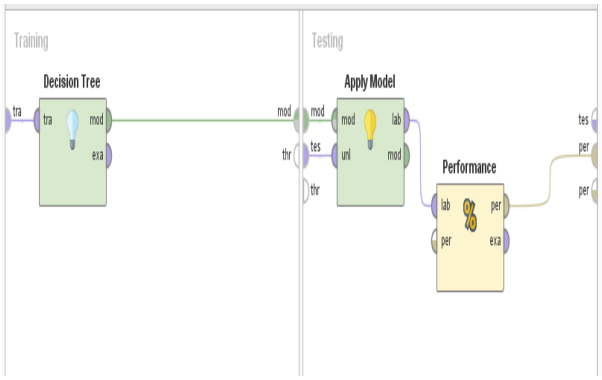

Fig. 6.Training through decision tree

b. Random Forest:

This is a technique used for classification. In this technique numbers of decision trees are generated randomly, each giving its prediction for the class of labeled attribute. This group of randomly generated evolved out as a forest, hence called random forest. The prediction of the decision tree getting highest vote, is considered as the final prediction at the end. The trees generated here are uncorrelated to each other, as each tree uses bagging and feature randomness, while its generation.

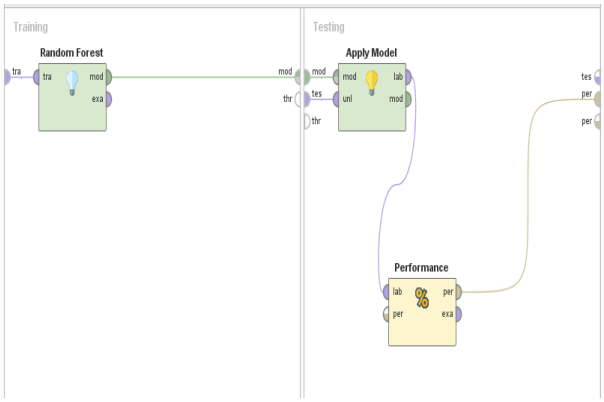

Fig. 7.Training through Random Forest

c. Naïve Bayes:

Naive Bayes classification is Bayesian classifiers which include Naive assumption too. Bayesian classifier is statistical classifiers, which is based on Bayes' theorem. Bayesian classifiers can predict probability. Like a given sample belongs to a particular class, it means, it can predict class membership probabilities. Further in Naive assumption, the changes in value of attribute of a given class are independent of the changes in the values of the other attributes. This assumption is also known as 'class conditional independence'.

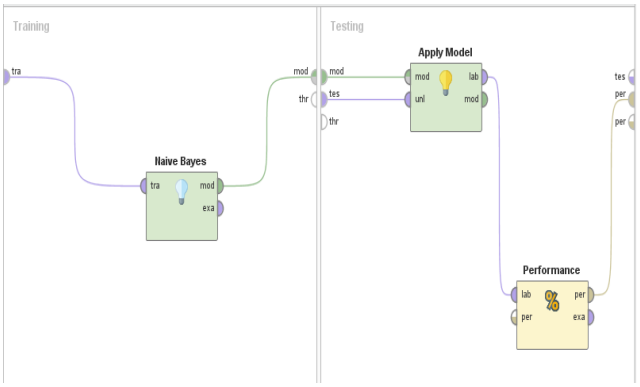

Fig. 8.Training through Naïve Bayes

d. Rule Induction:

Rule induction is a technique used for classification. According to this technique, after training phase, the understanding of the machine should be summarized in the model in form of rules. These rules could be formulated from any observation like decision tree, confusion matrix, hypothesis conditions, etc. Further this model comprising rules which represents the dataset is used for classification and prediction for any new upcoming example. Any learning method which formulates rules at the end to construct model comes under rule induction technique. For example association rule learning, rough set rules, decision tree, hypothesis rules, confusion matrix, etc.

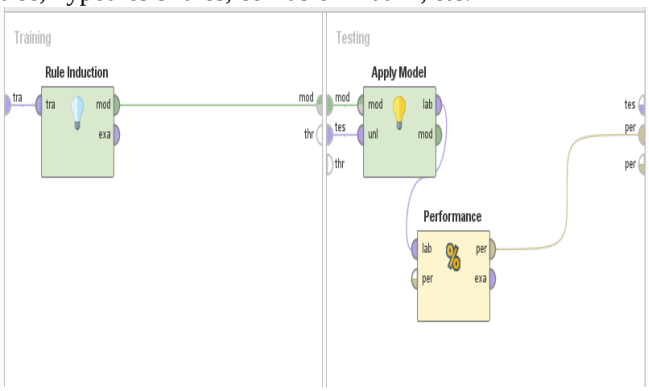

Fig. 9.Training through Rule Induction

IV. RESULT AND DISCUSSION

In our work we have used classification for predicting the performance of the students in final semester. Classification comprises of training and test phase. In training phase, training database was given to train machine using four different classification algorithms which are random forest, decision tree, Naive Bayes and rule induction each at five different number of folds, they are '10', '20', '30', '40' and '50'. The percentage of performance accuracy of each algorithm at each fold is noted and tabulated in the table 2 shown below. The figure 10 shows the bar graph showing performance accuracy percentage given by all the algorithms at each fold.

The Table- II and the bar graphs in figure 10 show that the decision tree at fifty number of fold gave the maximum performance accuracy percentage. The screenshot shown in the figure 11 depicts the accuracy percentage of decision tree at fold fifty which is highest among all. The figure 12 shows the prediction given by decision tree at fold fifty for the final semester grade G3 of the student which is considered as the final prediction.

Table- II: Accuracy Percentages of all the algorithms at different folds

No of Folds	Decision Tree	Naive Bayes	Random Forest	Rule induction
10	88.00	82.00	87.00	82.00
20	88.00	85.00	84.00	88.00
30	89.72	84.72	83.61	88.89
40	87.50	85.42	84.58	83.75
50	90.00	84.00	86.00	89.00

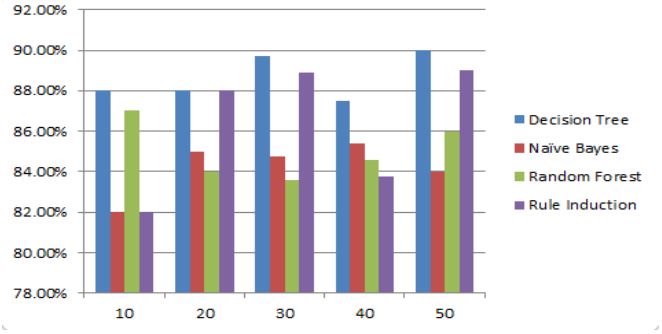

Fig. 10. Graph representation of accuracy percentage.

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. 11. Accuracy percentage by decision tree at fold 50

Row No.	prediction(G3)	confidence(B)	confidence(C)	confidence(A)	confidence(D)
1	B	0.500	0.500	0	0
2	C	0	1	0	0
3	B	1	0	0	0
4	B	0.609	0.391	0	0
5	A	0.091	0	0.909	0
6	B	0.500	0.500	0	0
7	B	0.500	0.500	0	0
8	B	1	0	0	0
9	B	0.500	0.500	0	0
10	C	0	0.833	0	0.167
11	C	0	0.833	0	0.167
12	C	0	1	0	0
13	B	1	0	0	0
14	B	0.609	0.391	0	0
15	B	1	0	0	0
16	B	0.500	0.500	0	0
17	C	0	1	0	0

Fig. 12. Grade(G3) predicted by decision tree at fold 50

V. CONCLUSION

This work deals with education data mining, where data mining is done on the education database containing academic performance records and all the information concerning students. Further a model is given which predicts the academic performance of the students. This prediction mainly helps the students for whom low grades are given. This process could increase the success rate and quality of education. The accuracy of the prediction given by this model is 90 percent. In future this work can be extended, by use of more algorithms which can give better performance accuracy percentage than 90. The size of the database could be increased in terms of attributes and examples so that more accurate result could be obtained.

REFERENCES

1. G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in Plastics, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
2. W.-K. Chen, Linear Networks and Systems (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.
3. H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch. 4.
4. B. Smith, "An approach to graphs of linear forms (Unpublished work style)," unpublished.
5. E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," IEEE Trans. Antennas Propagat., to be published.
6. J. Wang, "Fundamentals of erbium-doped fiber amplifiers arrays (Periodical style—Submitted for publication)," IEEE J. Quantum Electron., submitted for publication.
7. C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.
8. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interfaces(Translation Journals style)," IEEE Transl. J. Magn.Jpn., vol. 2, Aug. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].
9. M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.
10. (Basic Book/Monograph Online Sources) J. K. Author. (year, month, day). Title (edition) [Type of medium]. Volume(issue). Available: http://www.(URL)
11. J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com
12. (Journal Online Sources style) K. Author. (year, month). Title. Journal [Type of medium]. Volume(issue), paging if given. Available: http://www.(URL)

AUTHORS PROFILE

Surbhi Agrawal has obtained her Bachelor of Engineering degree in Computer Science & Engineering from Oriental Engineering College, Jabalpur in the year 2012. She worked in a multinational company named Cognizant for 2 years as a Programmer Analyst. Currently she is pursuing her MTech in Computer Science & Engineering from GGITS, Jabalpur. Her specialization includes Machine learning and data mining algorithms.

Dr. Santosh K. Vishwakarma is working as Associate Professor in the department of CSE, School of Computing & IT, Manipal University Jaipur. He completed his bachelor's and master's degree in Computer Science & Engineering. He is a doctorate in the field of Information Retrieval. His teaching specialization includes database management system, operating system, compiler design and theory of computation. He holds 15 years of Teaching Experience in reputed Institute. His research interest includes data mining, text mining, predictive analysis. He has been invited in various national and international forums for delivering sessions on databases, big data, predictive analysis and machine learning algorithms.