Research Journal of Computer and Information Technology Sciences _____ISSN 2320 – 6527
Vol. **6(6),** 1-10, October **(2018)** *Res. J. Computer and IT Sci.*

Predicting the academic performance of college students through machine learning techniques

R. Kaviyarasi* and T. Balasubramanian
Department of Computer Science, Periyar University, Salem-636011, Tamilnadu, India
arasikavi@gmail.com

Available online at: www.isca.in
Received 18th July 2018, revised 26th September 2018, accepted 18th October 2018

Abstract

Data Mining is one of the interdisciplinary subfield of Computer Science and by means of data analysis; it explains the past and predicts the future. Educational Data Mining (EDM) is one of the applications of Data Mining, Machine Learning and Statistics to generate the information from various educational settings such as universities and intelligent tutoring systems that has a vital impact on predicting students' academic performance. To predict and explore the factors affecting the performance of college students, many empirical researches are carried out. The main focus of this research is to identify the slow learners from the taken dataset which contains the students' profile details associated with their internal examination details. The student dataset is tested and applied on several classification models such as J48, Naïve Bayes and REPTree using an open source tool WEKA. The statistics are generated to predict the best accuracy based on classification algorithms and comparison of these classifiers is done to find the best performing classifier among others. This study explores the classifier models to predict the academic performance of students in the field of Educational Data Mining.

Keywords: Data mining, EDM, learners, J48, Naïve Bayes, REPTree.

Introduction

In earlier education systems, the responsibilities of educators were limited only with teaching the lessons in the classroom to expand the knowledge of students. But today, the teachers' contribution should be in overall improvement of the students such as to achieve optimum development of their abilities and harmonious personality development. Hence it is the responsibilities of the academic institutions to provide proper guidance to the students' for choosing the right carrier according to their abilities and aptitudes, so that they can achieve success and obtain personal satisfaction in their life. Many factors determine the level of academic performance of the students. Few are given below: i. Student abilities and their personal characteristics, ii. Faculties abilities and their personal characteristics, iii. Level of interaction between students and faculties, iv. Infrastructural facilities available in the college, v. External environmental influences on the students'.

Related studies have been carried out in this area. It identifies the poor performers and analyses the factors that affects the students' academic performance at schools, colleges and even at universities[1]. This proposed research aims at to analyses what could be the reason behind the non-academic performance of the students'.

Educational data mining: EDM develops methods for exploring data of distinctive types that comes from educational settings. Also through those methods students' are provided with better understating and learning process. The Educational Data Mining researchers set many goals for their research. Few are listed below[2]: i. Predicting students' future academic performance, ii. Analyzing the factors that characterize the performance of students' learning, iii. Studying the effects of different kinds of educational support that can be provided by learning software, iv. Advancing the scientific knowledge about learning and learners.

The users and stakeholders of EDM are: i. Learners, ii. Educators, iii. Researchers, iv. Administrators.

Literature Review: Raheela Asif, Agathe Merceron, Syed Abbas Ali and Najmi Ghani Haider[3] used Data Mining methods to study the performance of undergraduate students. Here the authors focused on two aspects of student performance. First, predicted students' academic achievements and next, study through typical progression throughout the academic years has taken out. Later combinations of the progression and predictions results are formulated.

Ankita A. Nichat and Anjali B Raut[4] done a research on the improvement of prediction, classification techniques that are used in analyzing and predicting the students' academic performance. This has been carried out using Data Mining technique, Decision Tree.

R. Sumitha and E.S. Vinothkumar[5] designed student's data model using J48 algorithm which proved to be an efficient algorithm in terms of accuracy identified by a comparative study of data mining classification algorithms.

Syed Tahir Hijazi and S.M.M. Raza Naqvi[6] focused the students performance in intermediate examinations associated with their' profile and the research used Data Mining Techniques which is based on student profile developed on the bases of information and data collected through survey from students of a group of private colleges.

Amirah Mohamed Shahiria, Wahidah Husaina, Nur'aini Abdul Rashida[7] focused on the prediction algorithms to identify the most important attributes in the student dataset. Also the research paper provided an overview of data mining techniques to improve the students' achievement.

Methodology

Before implementing Data Mining methods on our dataset, we should frame methodology for our work. The Figure-2 gives the work flow of our work used in this paper. Here the methodology starts from the problem statement, then proceeded to data collection, later preprocessing the dataset has been carried out, then we move to Data Mining Classification followed by the evaluation of results and finally we entered into knowledge representation process.

Data Collection: The one part of dataset for this research is obtained from the student database of various private colleges under Periyar University – Salem. These data were consulted with the experts in the educational field, experienced senior faculties and psychiatrist. The other part is their internal academic performance collected from the institutions.

Selection of Algorithms: Using Experimenter window, we have designed our own experiments of running algorithms on our datasets and analyzed the results as shown in the Figure-3. From this test output, the top three classifier algorithms such as Naive Bayes, REPTree and J48 have been selected for the implementation process.

Experimental Analysis: For this work, the dataset is processed by implementing certain classification algorithms.

Data Mining with WEKA**:** WEKA is an open source Tool. It is issued under the GNU, General Public License. It is fully implemented in Java programming language, so it is portable. It provides a collection of Data Mining, Machine Learning and Preprocessing tools. Also it includes algorithms for Classification, Clustering, Association Rule Mining, Regression and Attribute Selection.

Attribute Selection: The Data miner tool supports many in-built Machine Learning algorithms. We have applied one of the filter methods under supervised option, because Classification comes under Supervised Learning Method.

For this work the dataset with 100 records has been created in Excel 2007 and later it has been saved in the format of CSV. Later the CSV formatted dataset has been opened and saved in the format of ARFF which is accepted file format for our mining process. The attributes taken for the dataset are listed in Table-1.

Among 48 attributes, 12 attributes were selected using select attribute option. For this attribute selection process, we have chosen Info Gain Attribute Eval as Attribute Evaluator and Ranker as Search Method. The attributes we selected for this work are show in Figure-4.

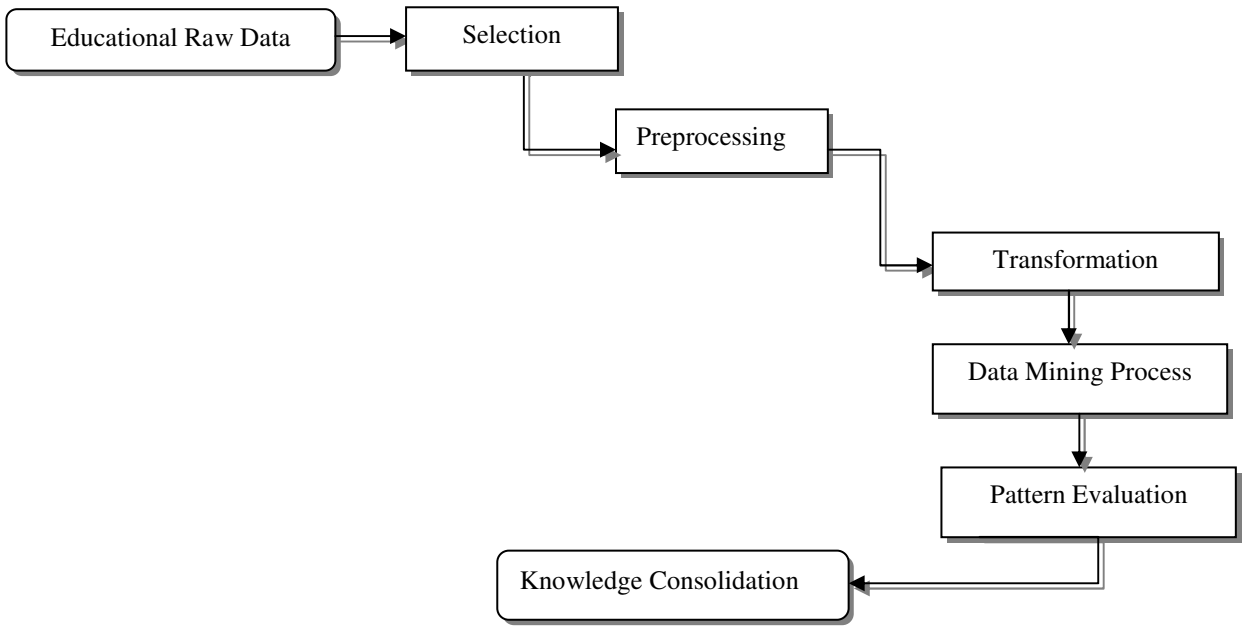


Figure-1: Phases in EDM.

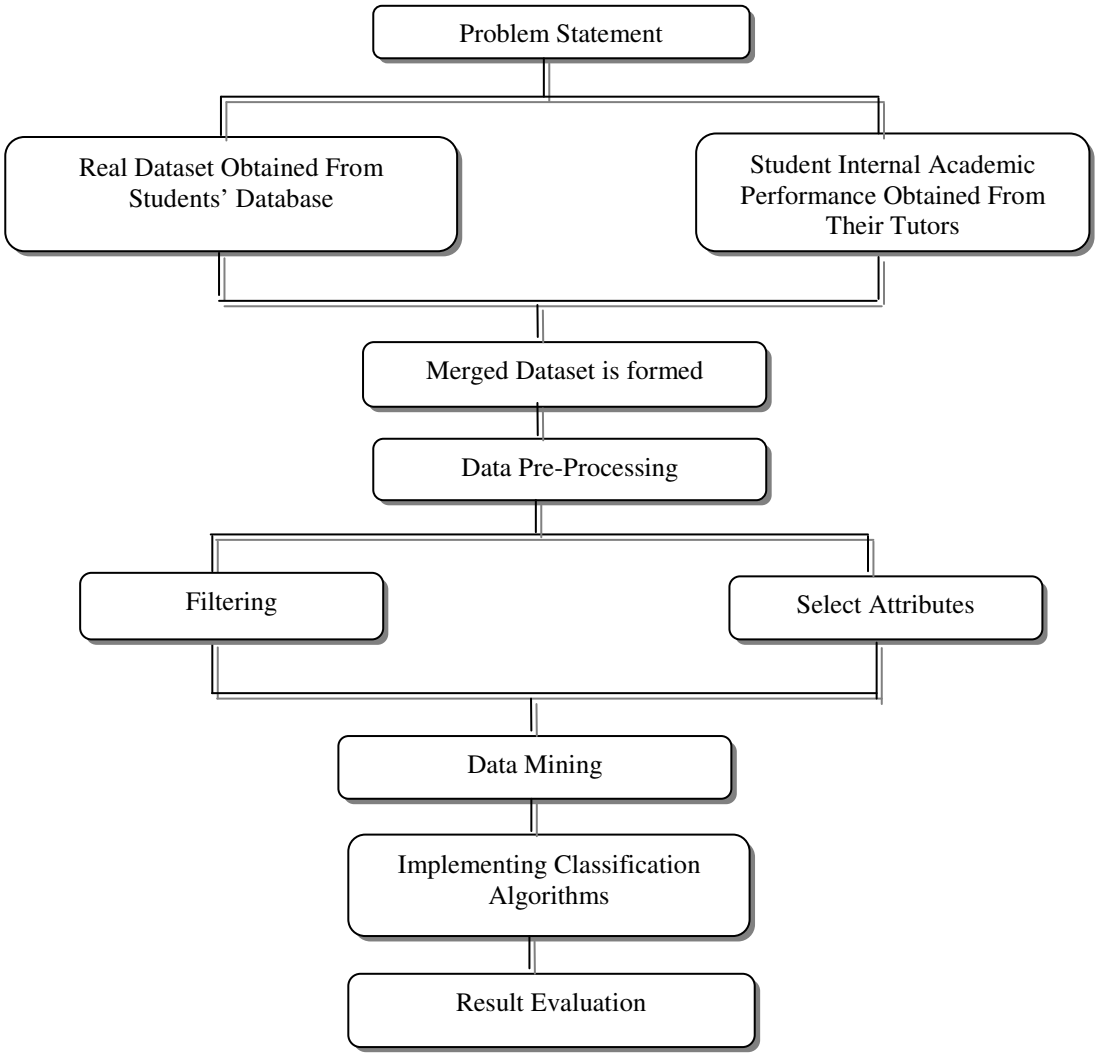


Figure-2: Workflow of Proposed Methodology.

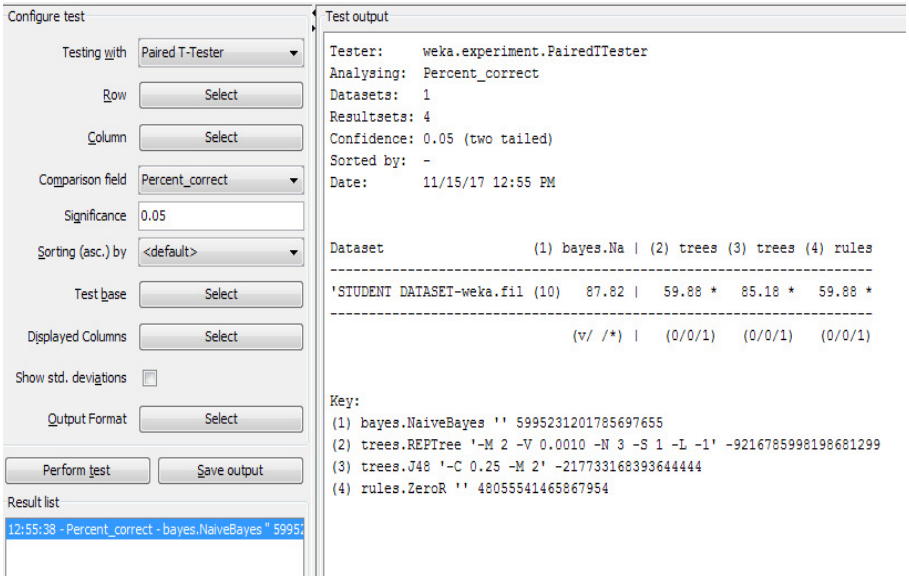


Figure-3: Comparison of Classifiers Test Output.

Table-1: Attributes taken for our work.

S. No.	Attributes	S. No.	Attributes
1	Name	25	Have mobile
2	Age	26	Student Using Mobile
3	Gender	27	Computer/laptop at home
4	Accommodation	28	Net access
5	Taken Care By	29	Social network id
6	Living Location	30	Study hours
7	Parental Status	31	Past arrears
8	Cohabitation Status	32	Extra college support
9	Fathers Education	33	Extracurricular activities
10	Fathers Job	34	Extra paid classes
11	Mothers Education	35	Going outings
12	Mothers Job	36	Alcohol consumption
13	Family size	37	Health status
14	10th grade	38	Any learning disabilities
15	12th grade	39	Place to study
16	Medium	40	Guidance
17	School	41	Care at home
18	Secondary syllabus	42	Interest in course
19	Group at Secondary	43	Attention in class
20	Any Part Time	44	Quality of study materials
21	Study Interest	45	Attendance percentage
22	Reason to choose this college	46	Semester percentage now
23	Travelling way	47	Internal test 1
24	Travel time	48	Internal test 2

Attribute selection output

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 13 CLASS):
        Information Gain Ranking Filter

Ranked attributes:
 1.371    1 NAME
 0.651   11 INTERNAL TEST 1
 0.578   12 INTERNAL TEST 2
 0        4 FAMILY SIZE
 0        5 HAVE MOBILE
 0        2 GENDER
 0        3 TC BY
 0        9 GUIDANCE
 0       10 QUALITY OF STUDY MATERIALS
 0        6 STUDY HRS
 0        8 ANY LD
 0        7 GOING OUTINGS

Selected attributes: 1,11,12,4,5,2,3,9,10,6,8,7 : 12
```

Figure-4: Attribute Selection Output.

The main attributes with expected relation are given here. The Table-2 explains the expected results for the attributes with high possible values. The attributes listed here are considered to be main factors that results for the students' attitude towards the academic performance.

As per the analyses and consideration with the teaching experts, the degree of performance of the students has been made into three categories as target values of class variables. i. Fast learners, ii. Average learners, iii. Slow learners.

The students who are found with the expected relation to be positive in 6 and above attributes positive internal test results are considered to be Fast Learners.

The students who are found with the expected relation to be positive in 4 and 5 attributes with 50% positive or negative internal test results are considered to be Average Learners.

The students who are found with the expected relation to be positive only below 4 attributes with negative internal test results are considered to be Slow Learners.

Table-2: Main attributes with expected relation.

Attributes	Expected Relation	Description
Taken Care By	Positive (Parents)	Students taken care by parents should be good
Family Size	Positive (<=5)	Family with limited members can take care the children
Computer/ Laptop	Positive (Yes)	Computer/ Laptop helps student in self learning
Mobile	Negative (Yes)	Usage of mobile reduces student involvement in studies
Study Hrs	Positive (>=2 hrs)	More study hours results in good performance
Going Outings	Negative (Yes)	More roaming hours reduces students study time hour as well as study interest
Guidance	Positive (Yes)	Guidance results in good performance
Quality of Study Materials	Positive (Easy)	Results in good academic output

Classification: WEKA supports number of Classification algorithms. One of the main benefits of its platform is supporting Machine learning algorithms for our machine learning problems. The classification algorithms used in this work will be discussed here.

J48: J48 is an open source Java implementation of the C4.5 algorithm in the Data Mining tool, where J for Java and 48 for C4.8, hence J48 name. It is a minor extension of the C4.5 algorithm. C4.5 generates the decision tree used for classification. It builds a decision trees from a set of training data. At each node of decision tree, C4.5 chooses the attributes of the data that effectively splits its sets of sample into subset enriched in one class or the other.

This splitting is the normalized information gain and the attribute with the highest normalized information gain is chosen to make the decisions[8].

In this work, the J48 algorithm is used to predict the slow learners among the given dataset and through this algorithm the decision tree is constructed.

Naïve Bayes: In Machine Learning, Naïve Bayes is not a single algorithm but a family of Classification algorithms based on Bayes rule of conditional probability[10]. It analyses the data individually for their dependency as well as the independency among each other by making use of all the attributes in the dataset. In this proposed work, Naïve Classifiers performs best compared to other classification algorithms such as J48 and REPTree.

REPTree: REPTree algorithm is the fast decision tree learner which uses the regression tree logic and builds multiple trees. Later it selects best one among the generated trees. It is based on C4.5 algorithm which can produce classification or regression trees. It generates decision tree using information gain/variance. It prunes it using reduced error pruning.

Classifier output

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          85               85      %
Incorrectly Classified Instances        15               15      %
Kappa statistic                          0.7283
Mean absolute error                      0.1262
Root mean squared error                  0.2909
Relative absolute error                 33.585  %
Root relative squared error             67.3125 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.95      0          1         0.95      0.974      0.974    SLOW LEARNER
                 0.9       0.2        0.871     0.9       0.885      0.868    FAST LEARNER
                 0.6       0.088      0.632     0.6       0.615      0.767    AVERAGE LEARNER
Weighted Avg.    0.85      0.138      0.849     0.85      0.849      0.869

=== Confusion Matrix ===

  a  b  c   <-- classified as
 19  0  1 |  a = SLOW LEARNER
  0 54  6 |  b = FAST LEARNER
  0  8 12 |  c = AVERAGE LEARNER
```

Figure-5: J48 Classifier Output.

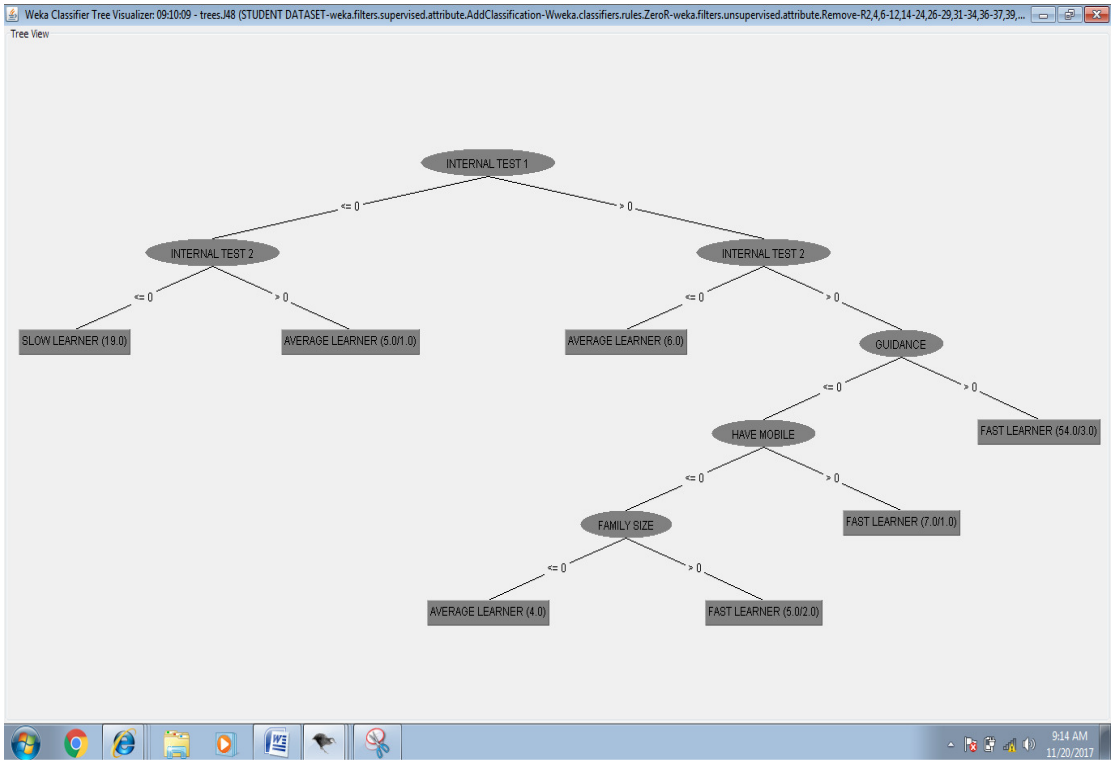


Figure-6: Decision Tree generated through J48 Classifier.

Classifier output

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          93               93      %
Incorrectly Classified Instances         7                7      %
Kappa statistic                          0.8732
Mean absolute error                      0.1117
Root mean squared error                  0.2121
Relative absolute error                 29.7468 %
Root relative squared error             49.0892 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.95      0          1         0.95      0.974      0.996    SLOW LEARNER
                 0.967     0.1        0.935     0.967     0.951      0.983    FAST LEARNER
                 0.8       0.038      0.842     0.8       0.821      0.959    AVERAGE LEARNER
Weighted Avg.    0.93      0.068      0.93      0.93      0.929      0.98

=== Confusion Matrix ===

  a  b  c   <-- classified as
 19  0  1 |  a = SLOW LEARNER
  0 58  2 |  b = FAST LEARNER
  0  4 16 |  c = AVERAGE LEARNER
```

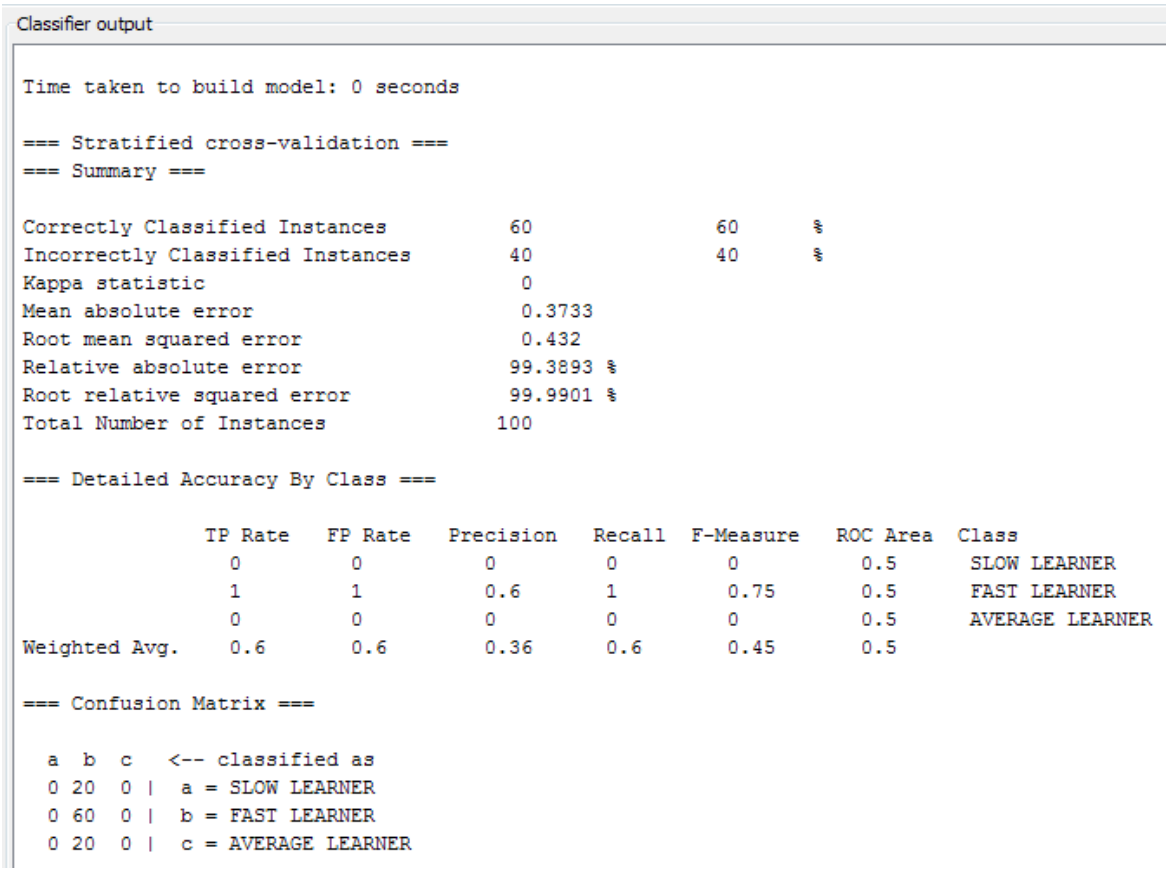
Figure-7: Naïve Bayes Classifier Output.

Results and discussion

Using three classification algorithms J48, Naïve Bayes and REPTree, the dataset taken for this research work has been tested and analyzed. Also comparisons among these algorithms have been done and concluded that Naïve Bayes is best among the implemented algorithms.

The confusion matrix is a possibility table. In our work, there are three classes, and therefore a 3X3 confusion matrix is formed. In confusion matrix, the sum of the diagonals represents the number of correctly classified instances and all others are incorrectly classified instances. The Stratified cross-validation summary is given in the Table-3 where Kappa statistic is a measurement that compares an Observed Accuracy with an Expected Accuracy. Also increase in Kappa statistic value indicates increase in classifier accuracy. From this comparative analysis, Naïve Bayes has highest Kappa statistic rate 0.8732 then J48 and REPTree algorithms. The classified accuracy percentage of Naïve Bayes is 93% which is higher than other classifiers. This two parameters show that the Naïve Bayes algorithm of classification performance is best for Educational Data Mining. From our dataset, 20% is classified as slow learners as per the confusion matrix.

The Table-4 shows the statistical result of these three algorithms comparison.



```
Classifier output

Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          60               60      %
Incorrectly Classified Instances        40               40      %
Kappa statistic                          0
Mean absolute error                      0.3733
Root mean squared error                  0.432
Relative absolute error                 99.3893 %
Root relative squared error             99.9901 %
Total Number of Instances              100

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0          0         0         0          0.5      SLOW LEARNER
                 1         1          0.6       1         0.75       0.5      FAST LEARNER
                 0         0          0         0         0          0.5      AVERAGE LEARNER
Weighted Avg.    0.6       0.6        0.36      0.6       0.45       0.5

=== Confusion Matrix ===

  a  b  c   <-- classified as
  0 20  0 |  a = SLOW LEARNER
  0 60  0 |  b = FAST LEARNER
  0 20  0 |  c = AVERAGE LEARNER
```

Figure-8: REPTree Classifier Output.

Table-3: Comparative Analysis: Stratified cross-validation.

Classification Algorithms	Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
J48	85	15	0.7283	0.1262	0.2909	33.59%	67.31%
Naïve Bayes	93	7	0.8732	0.1117	0.2121	29.75%	49.09%
REPtree	60	40	0	0.3733	0.432	99.39%	99.99%

Table-4: Comparative Analysis of Classifiers using Cross Validation Test Option.

Classification Algorithms	Class	TPRate	FPRate	Precision	Recall	F-Measure	ROC Area
J48	Slow Learners	0.95	0	1	0.95	0.974	0.999
	Fast Learners	0.9	0.2	0.871	0.9	0.885	0.868
	Average Learners	0.6	0.088	0.632	0.6	0.615	0.767
Naive Bayes	Slow Learners	0.95	0	1	0.95	0.974	1
	Fast Learners	0.967	0.1	0.935	0.967	0.951	0.983
	Average Learners	0.8	0.038	0.842	0.8	0.821	0.959
REPTree	Slow Learners	0	0	0	0	0	0.5
	Fast Learners	1	1	0.6	1	0.75	0.5
	Average Learners	0	0	0	0	0	0.5

Table-5: Accuracy Percentage of all Classifiers on the basis of Correctly Classified Instances Using the Test Option Cross Validation.

Classification Algorithms	Accuracy Percentage
J48	85%
Naïve Bayes	93%
REPtree	60%

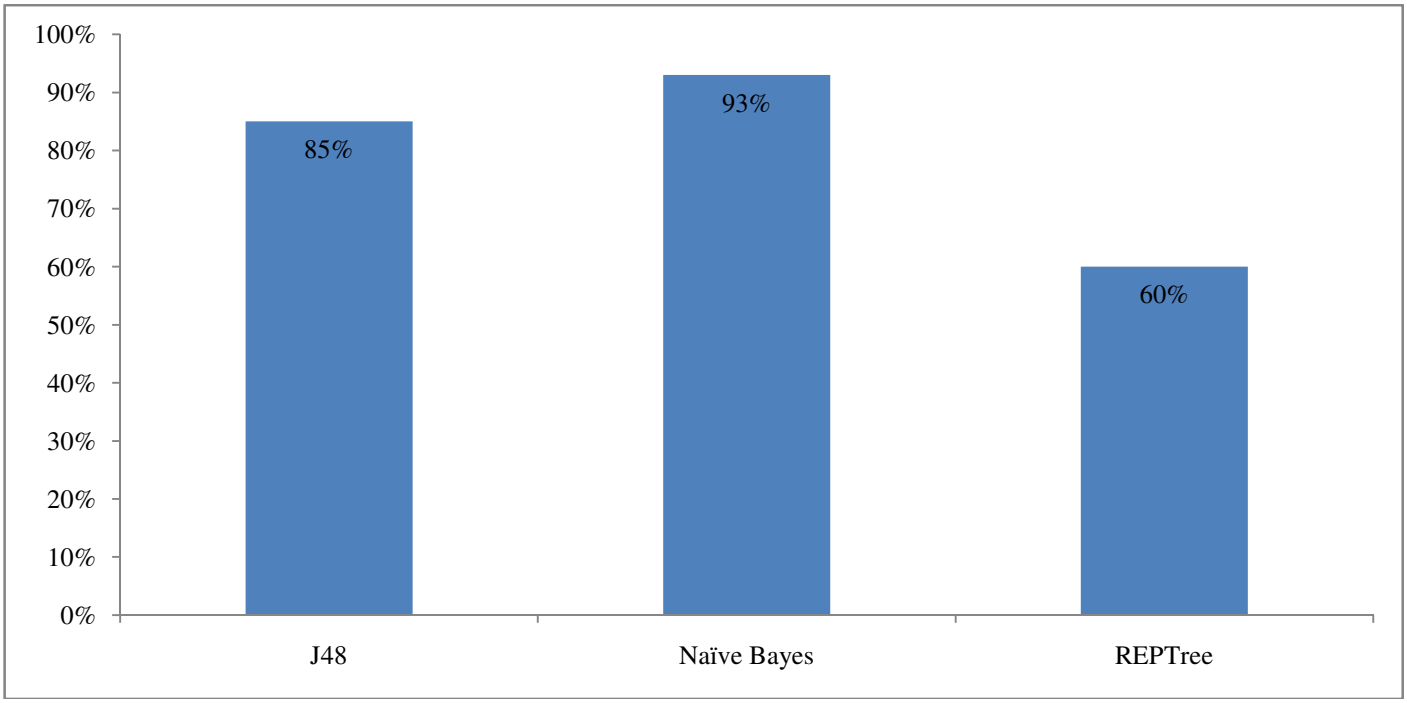


Figure-9: Comparison of Classifiers Output.

The accuracy percentage of these classifiers based on their performance has been given in the Figure-9.

Conclusion

In this work, we have identified the factors affecting the students' academic performance for the dataset of 100 records using Data Mining Tool. The main attributes were selected and the performance of the students' was predicted using certain classifiers J48, Naïve Bayes and REPTree. Also the comparison among these classifiers has been done and Naïve Bayes is proved to be the best among these algorithms. From the collected sample dataset, as per the confusion matrix, we observed that 20% of candidates are classified as slow learners. The future work of this research is to implement this in other Machine Learning Techniques such as Clustering, Neural Network, SVM and Fuzzy Logic to predict impacted attributes of slow learners which would be useful to counsel them and change their attitudes.

Acknowledgement

We would like to thank our Management and Principal of Sri Vidya Mandir Arts and Science College, Uthangarai, Tamilnadu for supporting this research by providing necessary student data.

References

1. Ali Shoukat, Haider Zubair, Munir Fahad, Khan Hamid and Ahmed Awais (2013). Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus. *Am J Educ Res.,* 1, 283-289. 10.12691/education-1-8-3.

2. Report (2018). What is Educational Data Mining (EDM)?. https://www.edtechreview.in/dictionary/394-what-is-educational-data-mining. April 2018.

3. Asif R., Merceron A., Ali S.A. and Haider N.G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers and Education*, 113, 177-194. 10.1016/j.compedu.2017.05.007.

4. Nichat Ankita A. and Anjali B.Raut Dr. (2017). Predicting and Analysis of Student Performance Using Decision Tree Technique. *International Journal of Innovative Research in Computer and Communication Engineering*, 13(7), 1735-1741.

5. Sumitha R. and Vinothkumar E.S. (2016). Prediction of Students Outcome Using Data Mining Techniques. *International Journal of Scientific Engineering and Applied Science (IJSEAS)*, 2(6), 132-139.

6. Hijazi S.T. and Naqvi S.M.M. (2006). Factors affecting students' performance. *Bangladesh e-Journal of Sociology*, 3(1), 1-10.

7. Shahiri A.M. and Husain W. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414-422.

8. Geeks for Geeks (2018). Naïve Bayes Classifiers. Retrived from http://www.geeksforgeeks.org/naive-bayes-classifiers/. Apr 2018.