Journal of Computer Science

Original Research Paper

Predicting Students' Academic Performance in the University Using Meta Decision Tree Classifiers

[1]Shanthini, A., [2]G. Vinodhini and [2]R.M. Chandrasekaran

[1]Department of Information Technology, SRM University, Chennai, India
[2]Department of Computer Science and Engineering, Annamalai University, Chidambaram, India

Article history
Received: 05-05-2017
Revised: 05-10-2017
Accepted: 19-02-2018

Corresponding Author:
Shanthini, A.
Department of Information Technology, SRM University, Chennai, India
Email: shanthini.aus@gmail.com

Abstract: Student performance prediction is an area of concern for educational institutions. At the University level learning system, the method or rule adopted to identify the candidates who pass or fail differs depending on various factors such as the course, the department of study and so on. Predicting the result of a student in a course is an issue that has recently been addressed using machine learning techniques. The focus of this work is to find a way to predict a student's academic performance in the University using the machine learning approach. This is done by using the previous records of the student rather than applying course dependent formulae to predict the student's final grade. In this work, meta decision tree classifier techniques based on four representative learning algorithms, namely Adaboost, Bagging, Dagging and Grading are used to construct different decision trees. REPTree is used as the decision tree method for meta learning. These four meta learning methods have been compared separately with respect to the training and test sets. Adaboost is found to be the best meta decision classifier for predicting the student's result based on the marks obtained in the semester.

Keywords: Decision Trees, Meta Classifiers, REPTree, Machine Learning

Introduction

Universities operate in very energetic and effective viable environments. A massive volume of data with reference to the students is available in digital form. However, the effective conversion of this voluminous data into knowledge for decision making is a huge problem. Predicting a student's performance is one such challenging issue faced by the educational sector (Asogbon *et al*., 2016). In recent years, many research works have focused on data mining techniques in higher education institutions to enhance the method of learning. The existing work in predicting students' performance includes analysing students' enrolment data to prevent dropouts, to predict student detention at an early stage and to analyse the quality and usage of learning materials. Developing an automated system for this will help educators to monitor their students' achievements (Buldu and Ucgun, 2010; Delen, 2010; Marquez-Vera *et al*., 2016) and the students to enhance their learning skills. The automated system will also help the administrative staff to upgrade the institutions' performance. Thus, the application of data mining techniques can be focused on particular applications of an automated system (Amrieh *et al*., 2016; Chen and Bai, 2010; Hien and Haddawy, 2007; Nespereira *et al*., 2015).

An ensemble of different machine learning algorithms is an effective method used for acquiring a high level of predictive accuracy However, such improvements are predicated based upon the existing ensembles. If every instance in the ensemble behaves in nearly the same way, a little more is achieved by combining their predictions. Decision trees are best adopted for ensemble methods as they are fast and stable (Buldu and Ucgun, 2010). The main purpose of this work is to compare the performance of various meta decision tree algorithms in predicting the performance of students in both training sets and test sets. The remainder of this paper is structured as follows. Section 2 presents the background. Section 3 describes the data models used. Section 4 presents the methods adopted and section 5 discusses the results followed by section 6 with the conclusions and future research possibilities.

Background

Predicting a student's performance has been studied previously in educational data mining research in the context of student attrition. Wolff *et al*. (2014) explored the effectiveness of predictive modelling methods for identifying students who will benefit most from tutor interventions in distance learning. The students and tutor will not meet face to face in the case of distance learning. The methods analysed for distance learning included decision-tree classification, support vector machine, general unary hypotheses automaton, Bayesian networks and linear and logistic regression. Romero *et al*. (2013) investigated how the accuracy of prediction was affected by factors like selection of instances and attributes, the usage of classification algorithms and the date when the data was gathered. A new Moodle's module was developed for gathering forum indicators. Using this Moodle, different experiments were carried out using real data from 114 university students in a first-year course in computer science. The results achieved proved its effectiveness both in terms of final prediction at the end of the course and early prediction before the end of the course.

Marquez-Vera *et al*. (2016) proposed a technique and a classification algorithm to construct a prediction model for student dropout as early as possible. The data set used for their research was obtained from 419 high school students in Mexico. Several experiments were carried out to predict dropout at different levels of the course. Some classical and imbalanced well-known classification algorithms were compared with their proposed algorithm to give the best indicator of dropouts. The results obtained in their research work showed that the algorithm devised by them was effective enough to predict student dropout within the first 4-6 weeks of the course. The algorithm can be used as an early warning system. Ramesh *et al*. (2013) adopted an experimental methodology to generate a database constructed from primary and secondary sources. The results obtained by this work reveal that parents' occupations play a major role in predicting the students' grade, whereas the type of school did not influence the students' results. Such findings can help institutions to identify the weak students at risk and concentrate on providing additional training to them.

Zhang *et al*. (2015), in order to improve the academic level at the undergraduate stage and achieve a better graduation thesis grade, researched the hidden relations between courses and graduation thesis grades and employed the support vector machine to construct a prediction model for predicting the graduation thesis grades of undergraduates. Some other prediction models (Neural Network, Decision Tree and Naïve Bayes) were also built to predict the graduation thesis grades of undergraduates, but the result showed that the Support Vector Machine (SVM) performed better in this case study. Strecht *et al*. (2015) addressed the problem of predicting the success or failure of a student in a course or a program using data mining techniques. They evaluated some of the most popular classification and regression algorithms in this problem. They addressed two problems in particular: Prediction of approval/failure and prediction of grade. The algorithms with best results overall in classification were decision trees and SVM while in regression they were SVM, Random Forest and AdaBoost (Illanas Vila *et al*., 2013). Arsad *et al*. (2013) used an Artificial Neural Network (ANN) model to predict the academic performance of engineering students pursuing a bachelor's degree. The study takes the Grade Point (GP) scored by the students in fundamental subjects as inputs without considering their demographic background, while it takes the Cumulative Grade Point Average (CGPA) as output. Schalk *et al*. (2011) built a machine-learning-based predictive system to determine which students were at risk of failing introductory courses in mathematics and physics. The system used the Random Forest technique to model data coming from previous years of SAT. While their results were good, the method designed was neither meant to be maintained over time nor to make progressive predictions based on incremental information.

In summary, various researches have been investigated to solve educational problems using data mining techniques. However, very few researches have shed light on students' behaviour during the learning process and its impact on the students' academic success. This proposed research will focus on the impact of the academic system on the students' performance. The performance of the student's predictive model was evaluated by a set of classifiers namely, ANN, Naïve Bayesian and Decision tree algorithms. In addition, we applied ensemble methods to improve the performance of these classifiers. The extracted knowledge will help schools to enhance students' academic performance and help administrators to improve learning systems. This work also concentrates, in particular, on the performance of students to predict whether the student will pass or fail at the end of the degree to differentiate between the strong and weak students. The current research study varies from other works by limiting the variables in predicting performance to marks only; no demographic or socioeconomic data were used. It takes one training set to build a model and another test set to evaluate it, thus allowing for some measurement of how well findings can be generalized.

Methodology

In this study, we introduce a performance model for students using ensemble methods. An ensemble method is a learning approach that combines multiple models to solve a problem. In contrast to traditional learning approaches which train data by one learning model, ensemble methods try to train data with a set of models

and then combine them to take a vote on their results. The predictions made by ensembles are usually more accurate than predictions made by a single model. Figure 1 shows the methodology design of this research work.

The following are the steps involved in methodology design:

- Collect data and identify the features for datasets
- For a dataset, develop the following prediction models using the respective training dataset
    - Bagging.
    - Boosting.
    - Dagging and
    - Grading.
- Predict the marks for all the models mentioned above
- Compare the prediction results with actual results

REPTree method is used as base prediction method for all ensemble methods employed. REPTree algorithm is based on the principle of calculating the information gain with entropy and reducing the error arising from variance. The complexity of the decision tree model is decreased by "reduced error pruning method" and the error arising from variance is reduced.

Boosting

In Boosting, a random subset of training samples *d*1 is selected without replacement from the training set D to train a weak learner C1. To select a second random training subset *d*2 without replacement from the training set, add 50 percent of the samples that were previously misclassified to train a weak learner C2. Find the training samples *d3* in the training set D on which C1 and C2 disagree to train a third weak learner C3. Combine all the weak learners via majority voting (Petkovic *et al*., 2012).

Bagging

Bagging represents Bootstrap AGGreg at ING. The steps involved in bagging are as follows:

- Generate *n* different bootstrap training samples
- Train the algorithm on each bootstrapped sample separately
- Average the predictions at the end

Dagging

Dagging Meta classifier creates a number of disjoint, stratified folds out of the data and feeds each chunk of data to a copy of the supplied base classifier. With majority voting, predictions are made on base classifiers that are accumulated into the Vote Meta classifier. A number of training instances useful for base classifiers that are quadratic or worse in time behaviour, are identified (Sorour *et al*., 2015).

Grading

The underlying idea of grading is to predict whether the prediction for a particular example is correct or not for each of the original learning algorithms. Therefore, one classifier is trained for each original learning algorithm with the training set of original examples. These original examples have class labels that encode whether the prediction of the learners was correct in this particular example (Yoo and Kim, 2014).

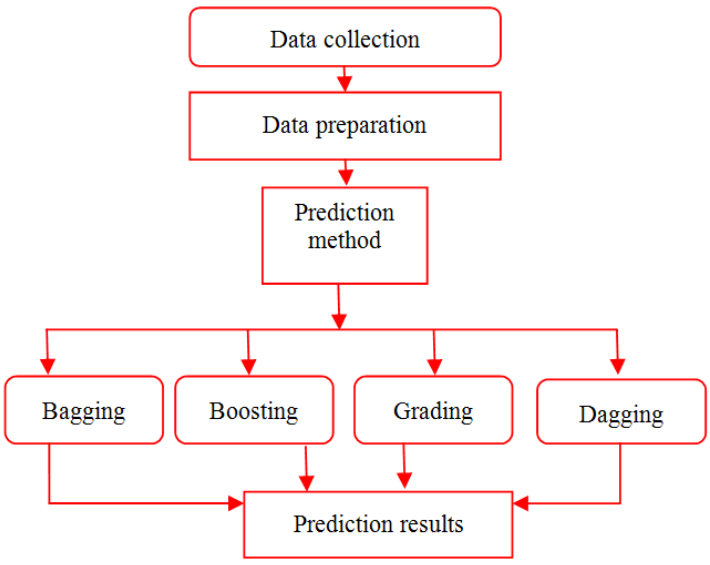

Fig. 1: Methodology design

Results

In this study, we used the marks obtained by computer science and engineering students in two semesters. Each semester has five subjects. We collected the marks of 401 undergraduate students who were enrolled in the academic year of 2014-15 and used it as the training set. The data contains variables related to the students' university examination marks in various subjects that were taught in the first and second semester. The description of the dataset used is shown in Table 1. For the test set, the marks obtained by students in the third semester were used. Table 2 shows the attribute description of the dataset used.

Figure 2 shows the tree model generated by the training dataset using REPTree.

The performance of the following Meta classifiers, Bagging, Boosting, Dagging and Grading were examined using Weka knowledge flow environment. Figure 3 shows the sample tree generated during an iteration for Bagging REPTree classifier. The attribute S3 is identified as the root node.

Table 1: Dataset description

Properties	Training set	Test set
Number of instances	401	80
Number of attributes	12	12
Class Attributes	1	1
Number of students passed	292	55
Number of students failed	109	25
Dimension	401×13	80×13

Table 2: Attributes description

Attribute Result	Description Final result	Type binominal	Range (Training set) Pass (292), Fail (109)	Range (Test set) Pass (55), Fail (25)
Sex	Sex	binominal	M (185), F (216)	F (50), M (30)
Age	Age	integer	[15.000; 22.000]	[15.000; 20.000]
S1	Subject 1 marks	integer	[15.000; 99.000]	[15.000; 98.000]
S2	Subject 2 marks	integer	[18.000; 99.000]	[28.000; 97.000]
S3	Subject 3 marks	integer	[12.000; 99.000]	[27.000; 95.000]
S4	Subject 4 marks	integer	[14.000; 99.000]	[38.000; 99.000]
S5	Subject 5 marks	integer	[18.000; 99.000]	[18.000; 99.000]
S6	Subject 6 marks	integer	[11.000; 99.000]	[41.000; 99.000]
s7	Subject 7 marks	integer	[11.000; 99.000]	[40.000; 97.000]
S8	Subject 8 marks	integer	[15.000; 99.000]	[28.000; 96.000]
S9	Subject 9 marks	integer	[12.000; 99.000]	[26.000; 94.000]
S10	Subject 10 marks	integer	[15.000; 99.000]	[28.000; 96.000]

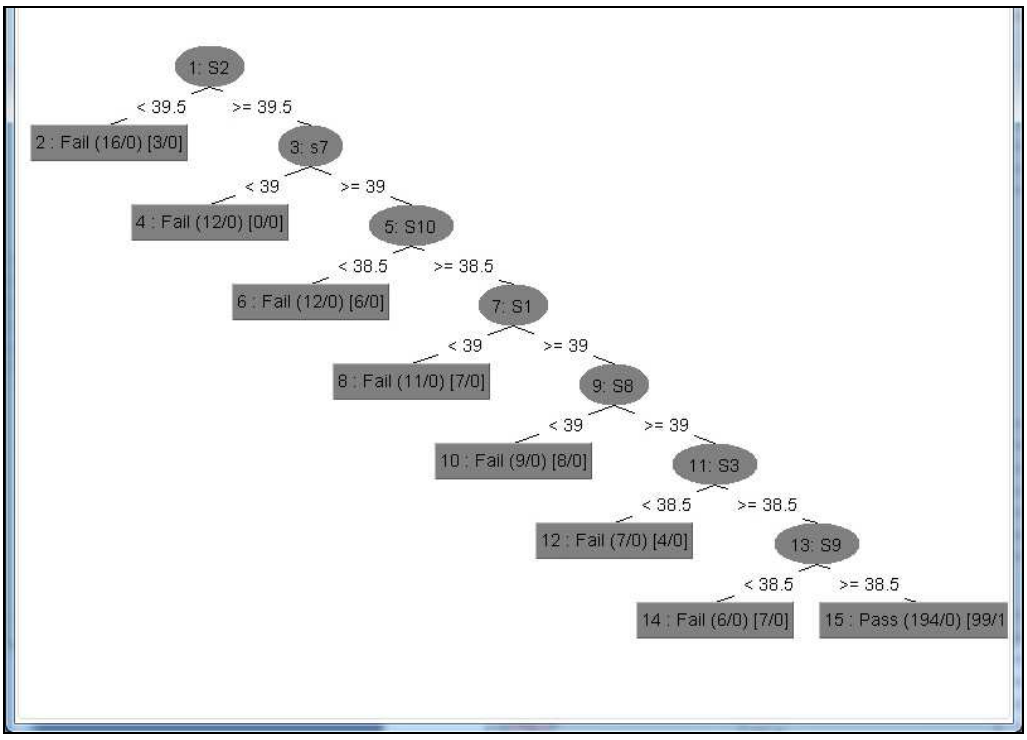

Fig. 2: Model generated by REPTree

```
S3 < 38: Fail (15/0) [8/0]
S3 >= 38
|  s7 < 37.5: Fail (14/0) [4/0]
|  s7 >= 37.5
|  |  S2 < 39: Fail (12/0) [2/0]
|  |  S2 >= 39
|  |  |  S10 < 38: Fail (12/0) [5/0]
|  |  |  S10 >= 38
|  |  |  |  S1 < 38.5: Fail (8/0) [6/0]
|  |  |  |  S1 >= 38.5
|  |  |  |  |  S8 < 38: Fail (5/0) [4/0]
|  |  |  |  |  S8 >= 38
|  |  |  |  |  |  S9 < 32.5: Fail (4/0) [3/0]
|  |  |  |  |  |  S9 >= 32.5: Pass (197/0) [102/3]

Size of the tree : 15
```

Fig. 3: Model generated by Bagging REPTree

```
s7 < 81.5
|  S5 < 46.5
|  |  S5 < 40.5: Fail (7.64/0) [0.26/0]
|  |  S5 >= 40.5: Pass (27.32/0.52) [13.95/0.39]
|  S5 >= 46.5
|  |  age < 16.5: Fail (165.76/6.59) [30.75/4.26]
|  |  age >= 16.5
|  |  |  s7 < 39: Fail (12.79/0) [0.13/0]
|  |  |  s7 >= 39: Pass (12.14/2.07) [17.83/1.03]
s7 >= 81.5
|  S2 < 40.5: Fail (6.86/0) [0.13/0]
|  S2 >= 40.5: Pass (95.34/1.16) [10.09/1.03]
Size of the tree: 13
Weight: 3.11
```

Fig. 4: Model generated by Boosting REPTree

```
S10 < 38.5: Fail (3/0) [1/0]
S10 >= 38.5: Pass (23/4) [13/3]

Size of the tree : 3
```

Fig. 5: Model generated by Dagging REPTree

```
S3 < 38.5: Fail (14/0) [2/0]
S3 >= 38.5
|  S10 < 38.5: Fail (13/0) [5/0]
|  S10 >= 38.5
|  |  S8 < 38.5: Fail (11/0) [7/0]
|  |  S8 >= 38.5
|  |  |  S2 < 40.5
|  |  |  |  S3 < 70.5: Fail (11/0) [8/2]
|  |  |  |  S3 >= 70.5
|  |  |  |  |  S2 < 38.5: Fail (2/0) [2/0]
|  |  |  |  |  S2 >= 38.5: Pass (2/0) [1/0]
|  |  |  S2 >= 40.5
|  |  |  |  S9 < 38.5: Fail (8/0) [4/0]
|  |  |  |  S9 >= 38.5
|  |  |  |  |  S1 < 38: Fail (7/0) [6/0]
|  |  |  |  |  S1 >= 38
|  |  |  |  |  |  s7 < 37.5: Fail (6/0) [3/0]
|  |  |  |  |  |  s7 >= 37.5: Pass (193/1) [96/1]

Size of the tree: 19
```

Fig. 6: Model generated by Grading REPTree

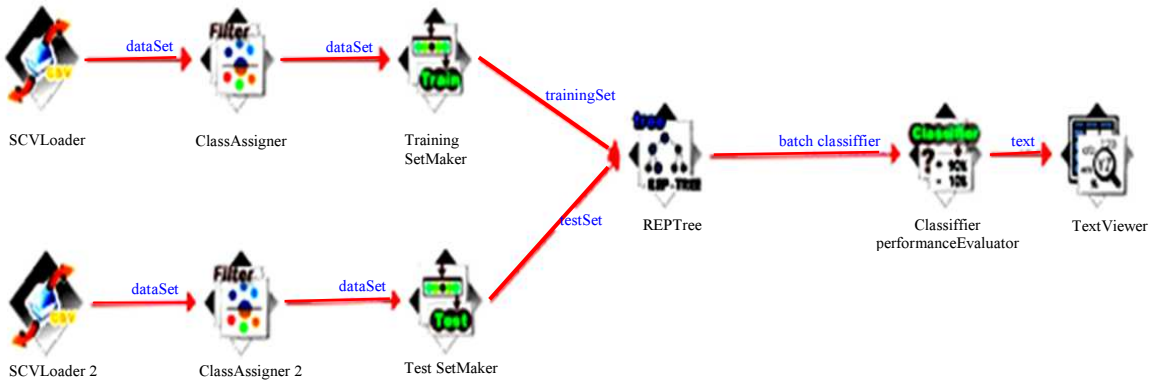

Fig. 7: Knowledge flow layout of REPTree

Figure 4 shows a model tree generated for Boosting REPTree classifier.

Figure 5 depicts the model tree obtained for Dagging REPTree.

Figure 6 shows the model generated by grading meta classifier.

As mentioned earlier, the Meta decision tree classifier using REPTree method was developed using Weka knowledge flow environment. The parameter setting available as default in the tool is used for all the algorithms. A sample knowledge flow layout for REPTree classifier is shown in Fig. 7.

The performances of the classifiers depend on the characteristics of the data to be classified. The performance results of the chosen algorithms are used to measure the accuracy, precision, recall and F-score. The performance measures obtained for individual REPTree classifier used for the training set and the test set are shown in Table 3.

The performance of meta decision tree classifiers are measured in terms of precision, recall and F-score. To evaluate the classification model, the training set and test set have been used. Four different meta decision tree algorithms Bagging, Boosting, Dagging and Grading

have been used for comparison. The results of the comparison are given in Table 4 in terms of precision. The highest precision of 99.0% has been achieved by Bagging and Boosting of REPTree algorithm for the training dataset. For the test set used, the highest precision is obtained for Boosting rather than Bagging.

The recall values of meta decision tree classifiers employed are shown in Table 5. The recall values obtained prove that Bagging and Boosting of REPTree classifiers performed better than Dagging and Grading of REPTree classifiers for both the training set and the test set. The highest recall of 99.0% is obtained for Bagging and Boosting REPTree classifiers. For the test set, the highest recall is obtained for Boosting REPTree classifier (97.5%).

Table 3: Performance of REPTree classifier

Measures	Training set	Test set
Precision	97.3	89.4
Recall	97.3	87.5
F-score	97.2	86.3

Table 4: Precision of meta decision tree classifiers

Meta classifiers (REPTree)	Training set	Test set
Bagging	99.0	96.4
Boosting	99.0	97.6
Dagging	80.3	50.8
Grading	98.0	92.5

Table 5: Recall of Meta decision tree classifiers

Meta Classifiers (REPTree)	Training set	Test set
Bagging	99.0	96.3
Boosting	99.0	97.5
Dagging	73.1	71.3
Grading	98.0	92.5

Table 6: F-score of Meta decision tree classifiers

Meta classifiers (REPTree)	Training set	Test set
Bagging	99.0	96.2
Boosting	99.0	97.5
Dagging	61.9	51.4
Grading	98.0	92.5

Table 7: Comparison with previous works

Author	Prediction Method	Accuracy obtained
Wolff *et al*. (2014)	Bayesian Network	89.6%
Romero *et al*. (2013)	Support Vector machine	92.3%
Ramesh *et al*. (2013)	Decision Tree	91.9%
Zhang *et al*. (2015)	Neural Networks	93.6%
Our approach	Ensemble method	97.5% (bagging)

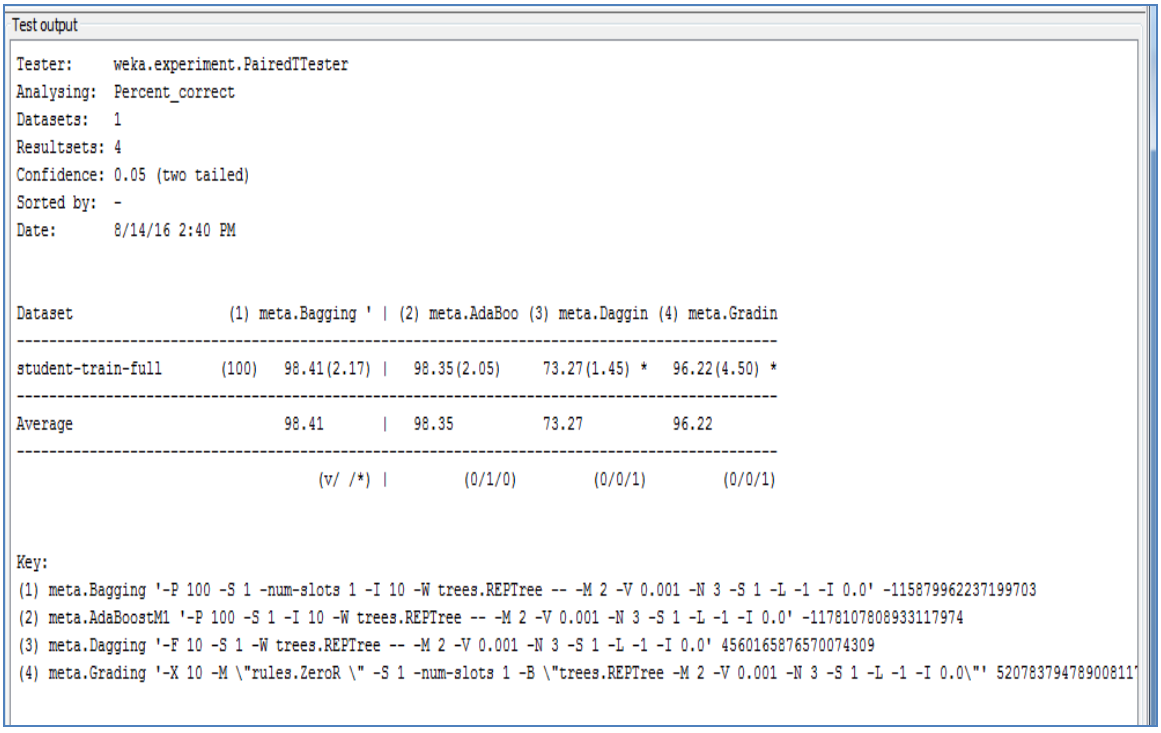

Fig. 8: Comparison of accuracies using Weka Experimenter

The results of F-Score for the meta decision tree classifiers employed are shown in Table 6. The F-score is a weighted combination of Bagging and Boosting. Thus almost similar observations of precision and recall are noted for the F-score of the meta decision tree classifiers. The F-score values also show that Bagging and Boosting of REPTree classifiers perform better than Dagging and Grading of REPTree classifiers for both the training set and the test set.

The accuracy of the Meta decision tree classifiers are calculated and compared using Weka experimenter. The accuracy result obtained show that Bagging shows better accuracy than the other meta decision tree classifiers (Fig. 8). The boosting REPTree classifier ranks next in the hierarchy. Similar results are noted for the test set also.

The results obtained in this research work are compared with existing work and the results are tabulated in Table 7. The results show that the ensemble method proves to be better than other individual methods employed earlier in predicting students' results in educational institutions.

Conclusion

Predicting the student's performance is the most effective way to a dedicators and learners in upgrading their teaching and learning processes. Better inferences could be drawn with the classification approach resulting in better prediction of whether a student will pass or fail in a course. Further analysis is necessary to better understand and improve these results. In addition to the problems studied in this work, it would be interesting to predict an interval for a grade. This method will aid the educational institutions to monitor the performance of students in an effective and systematic way. Lastly, this model can help educators understand learners, identify weak learners, to improve learning processes and bring down academic failure rates. It also can help the administrators to improve the learning system outcomes.

In our future work, we will focus more on analysing behavioural features on the students' performance model. This result may prove to be a more realistic predictive model. Some optimization could be made using a parameter selection method such as feature selection. In conclusion, the meta-analysis on predicting a student's performance has inspired us to conduct further research to be applied in various educational institutions.

Author's Contributions

Shanthini, A.: Participated in all experiment data and part of writing.

G. Vinodhini: Data analysis and research.

R.M. Chandrasekaran: Research plan and writing.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and there are no ethical issues involved.

References

Amrieh, E.A., T. Hamtini and I. Aljarah, 2016. Mining educational data to predict student's academic performance using ensemble methods. Int. J. Database Theory Applic., 9: 119-136. DOI: 10.14257/ijdta.2016.9.8.13

Arsad, P.M., N. Buniyamin and J.L.A. Manan, 2013. A Neural Network Students' Performance Prediction Model (NNSPPM). Proceedings of the IEEE International Conference on Smart Instrumentation, Measurement and Applications, Nov. 25-27, IEEE Xplore Press, Kuala Lumpur, Malaysia, pp: 1-5. DOI: 10.1109/ICSIMA.2013.6717966

Asogbon, M.G., O.W. Samuel, M.O. Omisore and B.A. Ojokoh, 2016. A multi-class support vector machine approach for students academic performance prediction. Int. J. Multidisciplinary Curr. Res., 4: 210-215.

Buldu, A. and K. Ucgun, 2010. Data mining application on students' data. Proc. Soc. Behav. Sci., 2: 5251-5259. DOI: 10.1016/j.sbspro.2010.03.855

Chen, S.M. and S.M. Bai, 2010. Using data mining techniques to automatically construct concept maps for adaptive learning systems. Expert Syst. Applic., 37: 4496-4503. DOI: 10.1016/j.eswa.2009.12.060

Delen, D., 2010. A comparative analysis of machine learning techniques for student retention management. Decis. Support Syst., 49: 498-506. DOI: 10.1016/j.dss.2010.06.003

Hien, N.T.N. and P. Haddawy, 2007. A decision support system for evaluating international student applications. Proceedings of the 37th Annual Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, Oct. 10-13, IEEE Xplore Press, Milwaukee, WI, USA, pp: F2A-1-F2A-6. DOI: 10.1109/FIE.2007.4417958

Illanas Vila, A.I., J.R. Calvo Ferrer, F.J. Gallego-Durán and L.F. Largo, 2013. Predicting student performance in translating foreing languages with a serious game. Proceedings of the INTED Conference, Mar. 4-6, Valencia, Spain, pp: 52-59.

Marquez-Vera, C., A. Cano, C. Romero, A.Y.M. Noaman and H. Mousa Fardoun *et al*., 2016. Early dropout prediction using data mining: A case study with high school students. Expert Syst., 33: 107-124. DOI: 10.1111/exsy.12135

Nespereira, C.G., E. Elhariri, N. El-Bendary, A.F. Vilas and R.P.D. Redondo, 2015. Machine learning based classification approach for predicting students performance in blended learning. Proceedings of the 1st International Conference on Advanced Intelligent System and Informatics, Nov. 28-30, Springer, Beni Suef, Egypt, pp: 47-56. DOI: 10.1007/978-3-319-26690-9_5

Petkovic, D., K. Okada, M. Sosnick, A. Iyer and S. Zhu *et al*., 2012. Work in progress: A machine learning approach for assessment and prediction of teamwork effectiveness in software engineering education. Proceedings of the Frontiers in Education Conference, Oct. 3-6, IEEE Xplore Press, Seattle, WA, USA, pp: 1-3. DOI: 10.1109/FIE.2012.6462205

Ramesh, V., P. Parkavi and K. Ramar, 2013. Predicting student performance: A statistical and data mining approach. Int. J. Comput. Applic., 63: 35-39. DOI: 10.5120/10489-5242

Romero, C., M.I. López, J.M. Luna and S. Ventura, 2013. Predicting students' final performance from participation in on-line discussion forums. Comput. Educ., 68: 458-472. DOI: 10.1016/j.compedu.2013.06.009

Schalk, P.D., D.P. Wick, P.R. Turner and M.W. Ramsdell, 2011. Predictive assessment of student performance for early strategic guidance. Proceedings of the Frontiers in Education Conference, Oct. 12-15, IEEE Xplore Press, Rapid City, SD, USA, pp: S2H-1-S2H-5. DOI: 10.1109/FIE.2011.6143086

Sorour, S.E., K. Goda and T. Mine, 2015. Student performance estimation based on topic models considering a range of lessons. Proceedings of the International Conference on Artificial Intelligence in Education, Jun. 22-26, Springer, pp: 790-793. DOI: 10.1007/978-3-319-19773-9_117

Strecht, P., L. Cruz, C. Soares, J. Mendes-Moreira and R. Abreu, 2015. A comparative study of classification and regression algorithms for modelling students' academic performance. Proceedings of the 8th International Conference on Educational Data Mining, Jun. 26-29, At Madrid, Spain.

Wolff, A., Z. Zdrahal, D. Herrmannova and P. Knoth, 2014. Predicting Student Performance from Combined Data Sources. In: Educational Data Mining, Peña-Ayala, A. (Ed.), Springer International Publishing, pp: 175-202.

Yoo, J. and J. Kim, 2014. Can online discussion participation predict group project performance? Investigating the roles of linguistic features and participation patterns. Int. J. Artificial Intell. Educ., 24: 8-32. DOI: 10.1007/s40593-013-0010-8

Zhang, W., S. Zhang and S. Zhang, 2015. Predicting the graduation thesis grade using SVM. Int. J. Intell. Inform. Process., 5: 60-68.