Comparative Analysis of Models for Student Performance with Data Mining Tools

A. K. Shrivas[1], Pragya Tiwari[2]

[1]Assistant Professor, Department of IT, Dr. C. V. Raman University, Chhattisgarh, India
[2]Research Scholar, Department of IT, Dr. C. V. Raman University, Chhattisgarh, India

Abstract: *In modern time the analysis of student performance is very challenging task for every educational institutions. The main reason behind that rapid growth of population and increasing number of schools and colleges claiming that they can give their students quality education and provide the best environment for quality learning and many other aspects through which they can increase the performance capabilities in each and every student. There are different researcher have worked in the field of analysis of student performance, but they have not achieved satisfactory result. In this research work, we have used various data mining techniques for analyzing of student performance using WEKA , Rapid Miner, Tanagra and Orange data mining tools in case of both Portuguese and Mathematics Dataset . Random forest gives best accuracy as 93.52% in Weka data mining tool while 73.65% of accuracy in Tanagra data mining tool in binary and multiclass problem respectively with Portuguese data set. Similarly, in case of Mathematics dataset, Radom forest achieved 92.40% of accuracy in Weka data mining tool while 74.43% of accuracy in Orange data mining tool with binary and multiclass problem respectively. Finally, Random forest is robust model for classification of student performance.*

Keywords: *Data Mining, Classification, Student Performance.*

I. INTRODUCTION

In past few years there were so many changes occurred in field of higher education. These changes are done specially after the invention of many new trends and technologies in field of computer science such as “Data Mining”. Data Mining is nothing but mining the data for fruitful information i.e. “knowledge”. For a healthy growth of any educational institution, it is a must that they have substantial amount of knowledge. This knowledge need can be fulfilled with the help of Knowledge Discovery process that extracts the knowledge from available datasets. All this knowledge can be stored in a knowledgebase or a repository and can be used by the institution when needed for prediction of the student performance. These institutions classify there students by their academic performance only. But there are also some other factors that influence on the performance of any student and hence they should be come out in the lime light [14].

As more as the population increases so as the educational institutions are also increasing. And now all they are competing to provide quality education, for this reason they need to extract information related with their students. There are several Data mining techniques that are useful for deriving the hidden knowledge. This derived knowledge can be student specific such as his/her academic performances, courses, failures etc. All these facts that define student performance can be used for predicting the overall performance of the student by using a number of available data mining classification algorithms [1]. In this paper we will try to analyse the performance of the student in different categories with the help of these well known classification techniques.

II. LITERATURE REVIEW

J. Ruby et al. [2], have used various classification techniques in which MLP classification got an accuracy of 64.5% in an average of 5 runs . Kumar S. Anupama et al. [3], applied C4.5 decision tree algorithm to the internal marks of the MCA students and predict their performance in terms of pass or fail in final exam. They have compared the predicted results with actual results which indicates, that there was a significant improvement in results as the prediction helped a lot to identify weak and good students and help them to score better marks. They also compared the model with ID 3 decision tree algorithm and prove that the developed model is better in terms of efficiency and time taken to build the decision tree. Mohd Maqsood Ali et al. [4], have presented the roll of data mining in education sector. O. F. Naoh et al. [5], applied the K-Means clustering algorithm of data mining for discovering knowledge from data that come from educational environment for improving students’ performance. Brijesh Kumar Baradwaj et al. [8], have used decision tree method for analyzing students performance. Z. J. Kovacic [9], presented a study on educational data mining to identify to predict student's success. The algorithms CHAID and CART were used and CART gives the highest accuracy as 89.85%. M. Ramaswami et al. [10], have suggested CHAID based predictive model and used data set with 772 records to test the dataset with minimum number of features through feature selection. The accuracy of the presented model was compared with other models and it has been found to be satisfactory. M. Pandey et al. [11], have collected 524 records with 18 attributes each from a college of Faridabad and applied C4.5 (J48) classification algorithm having gain ratio as feature selection under cross-validation method and got an accuracy of 80.15%. P. Cortez et al. [14], have used a dataset with 649 records of the Portuguese students and 395 records of the maths students and applied different techniques and achieved an accuracy of 93.0 % with decision tree and 91.9 % with naive bayes respectively. J. Ruby et al. [15], have used two datasets in which first is of 165 records and the second one is 396 records. They have used MLP classification algorithm and achieved 64.5% and 91.42% accuracy respectively as an average of 5 runs .S. K. Gupta et al. [16], have used data set with 1282 records of the students willing to be admitted for any technical under graduation course. They have used 9 different classifiers and found that MLP achieved better accuracy for classification. A. TEKIN [17], has used data set with 127 undergraduate students and applied 3 different classifiers NN, SVM, ELM and has got an accuracy of 93.06% with SVM. Q. Al-Radaideh et al. [18], have worked on decision tree method and naive bayes method and got highest accuracy with ID3 method which is 38.46%.M.S. Mythili et al. [21], have used data set with 260 records and applied different classifiers and has got an accuracy of 89.23% with Random Forest algorithm.

III. PROPOSED OBJECTIVE

Figure1 shows that proposed architecture of student performance analysis. The proposed architecture consists data set, partition of data set with 10-fold cross validation, various data mining tools, data mining techniques and performance of models. In this proposed architecture, Portuguese and Mathematics data set with 10- applied into different data mining techniques using different data mining tools.

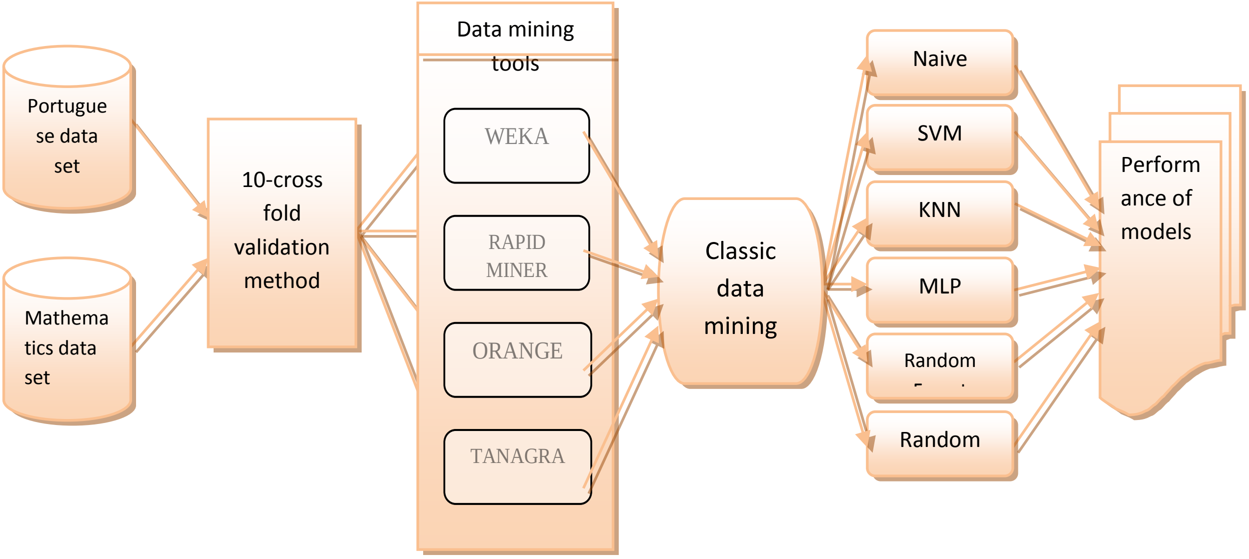

Fig 1: Proposed Architecture

Finally, we calculated the performance measure as accuracy of various model. We have selected the best model based on the accuracy for analysis of performance of students.

IV. METHOD AND MATERIALS

Techniques and tools are very important role in every field of research area. In this research work, we have used various data mining based classification techniques, different data mining software and data set for analysing student performance.

i. Data Mining Techniques

Data classification is a process of building a learning model i.e. classifier and training it to classify the next given test set as accurately as possible. In this analysis many classification techniques have been used such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbour, and K-NN and Support Vector Machine.

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or *terminal node*) holds a class label. The topmost node in a tree is the root node. A decision tree depicts rules for dividing data into groups. In this research work, we have used Random forest and Random Tree for student performance analysis.

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naïve" [12].

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. An SVM model uses a nonlinear mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for the linear optimal separating hyperplane (that is, a "decision boundary" separating the tuples of one class from another). With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane. The SVM finds this hyperplane using *support vectors* ("essential" training tuples) and *margins* (defined by the support vectors). In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [12].

Multi Layer Perceptron is a feed forward artificial neural network model trained with the standard back propagation algorithm that maps sets of input data onto a collection of acceptable output[12]. The perceptron is a simple neural network, proposed in 1958 by Rosenblatt [Ros58], which became a landmark in early machine learning history. Its input units are randomly connected to a single layer of output linear threshold units. Nearest-neighbour classifiers are based on learning by analogy, that is, by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by *n* attributes. Each tuple represents a point in an *n*-dimensional space. In this way, all of the training tuples are stored in an *n*-dimensional pattern space. When given an unknown tuple, a ***k***-nearest-neighbour classifier searches the pattern space for the *k* training tuples that are closest to the unknown tuple. These *k* training tuples are the *k* "nearest neighbours" of the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance [12].

ii. Dataset

In this paper we have used two datasets which is taken from UCI Machine Learning repository. Both datasets are related to two distinct subjects Portuguese and Mathematics of two Portuguese schools. First dataset consist of 649 records and second dataset is consist of 395 record each of which is a multivariate dataset having 32 normal attributes and 1 special attribute that has a capability to become a label. Both datasets are modelled under five-level classification and binary classification as well.

iii. Software Tools

In this experiment, we have used 4 different data mining software are WEKA, Rapid Miner, Orange and Tanagra. All these software are top data mining tools that are available as open source tools for data mining. In these tools WEKA is the tool with is based on java is very sophisticated tool for data analysis. It is free under general public license by which users can customize it however they please [24]. Whereas RAPID MINER which is also written in java programming language offers advanced analytics through template-based frameworks. In template-based data mining tools next comes ORANGE data mining tool which is written in python and is very powerful and easy to learn tool. On the other hand TANAGRA is a suite of machine learning software developed by Ricco Rakotomalala at the Lumiere University Lyon 2 France as an academic project. It supports several data mining methods[25].

V. RESULT ANALYSIS

In this paper, the experiment is carried out using various data mining tools in window 7 environment with i5 system. The main focus of this research work is to analysis of student performance using WEKA, RAPID MINER, TANAGRA and ORANGE data mining tools. We have used two data set ie. Mathematics and Portuguese collected from UCI repository. These data sets with binary and multiclass applied on the different classification techniques like Random Trees (RT), Random Forests (RF), Multi Layer Perceptron Network (MLP), Support Vector Machines (SVM), K-Nearest Neighbour (KNN) and Naive Bayes (NB). The obtained results reveal that it is possible to achieve a high predictive accuracy,

when all the dataset is modelled under different data mining models having relevant features, such as number of absences, reason to choose school, extra educational school support, student's age, parent's job and education going out with friends and alcohol consumption. Table 1 shows that accuracy of models with different data mining tools in case of Mathematics and Portuguese data set with binary classification. Results shows, that Random forest gives best accuracy as **93.52% with** Portuguese dataset while 92.40% of accuracy with Mathematics dataset in case of WEKA data mining tools. Similarly, Table 2 that accuracy of models with different data mining tools in case of Mathematics and Portuguese data set with multi class classification. Results shows, that SVM gives best accuracy instead of Random forest with ORANGE data mining tool. From the table 1 and table 2 shows that accuracy of models different form tool to tool. From the results, we conclude that Weka data mining tools is better in case of binary classification problem while ORANGE data mining tool is better in case of multiclass classification problem.

Table 1: Accuracy of models in 10-fold cross validation with binary class

	Classifiers	Weka	Rapid Miner	Tanagra	Orange
Portuguese Dataset	Naive Bayes	88.44	87.83	89.22	88.32
	SVM	89.67	84.59	91.72	88.77
	MLP	89.67	77.83	**92.03**	NA
	K-NN	82.12	**91.06**	85.78	91.05
	Random Forest	**93.52**	86.59	NA	**92.26**
	Random Tree	88.13	85.05	90.00	NA
Mathematics Dataset	Naive Bayes	70.37	85.80	86.92	86.33
	SVM	89.11	57.25	**87.95**	84.07
	MLP	87.34	82.34	**87.95**	NA
	K-NN	64.05	**86.33**	73.59	87.31
	Random Forest	**92.40**	68.36	NA	**89.88**
	Random Tree	85.82	70.33	84.36	NA

Table 2: Accuracy of models in 10-fold cross validation with multi class

	Classifiers	Weka	Rapid Miner	Tanagra	Orange
Portuguese Dataset	Naive Bayes	68.25	**66.72**	65.63	69.37
	SVM	**56.70**	24.64	52.38	53.30
	MLP	57.16	37.30	57.47	63.28
	K-NN	31.74	61.17	**62.25**	35.16
	Random Forest	**72.57**	30.97	**73.65**	NA
	Random Tree	48.38	30.97	NA	**62.50**
Mathematics Dataset	Naive Bayes	63.02	70.37	**71.41**	64.05
	SVM	53.16	55.18	30.42	61.77
	MLP	NA	53.16	50.86	60.25

	K-NN	**61.03**	31.89	56.96	59.49
	Random Forest	**66.71**	**71.13**	32.92	**74.43**
	Random Tree	NA	54.43	34.46	NA

VI. CONCLUSIONS

Analysis of Student performance is very import role in education sector. Higher education play very important role for grading student based on their performance. Data mining based classification techniques play very important role for categorization of students. In this research work, we have used for data mining tools and used similar classification techniques for categorization of students. Results show that performances of classification techniques are different from one tool to another tool.

In this research work, we have analysed the both data set the data set with individuals models. In future we can develop the robust ensemble and hybrid model to achieve the better classification accuracy. We can also apply the feature selection techniques to computationally improve the performance of models.

REFERENCES

[1]. J. Ruby & K. David, “A study model on the impact of various indicators in the performance of students in higher education”, International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, pp. 750-755, 2014.

[2]. J. Ruby and K. David, “Analysis of Influencing Factors in Predicting Students Performance Using MLP – A Comparative Study “,International Journal of Research in Engineering and Technology, Vol. 3, Issue 2 , pp.1085-1092, 2015.

[3]. Kumar S. Anupama and M. N Vijayalakshmi, “Efficiency of Decision Trees in Predicting Students Academic Performance”, Computer Science & Information Technology , Vol. 02, .pp. 335–343, 2011.

[4]. M. Ali, “Role of Data Mining in Education Sector”, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, pg. 374-383, 2013.

[5]. O. F. Noah and B. Barida, “Evaluation of Student Performance Using Data Mining Over a Given Data Space”, International Journal of Recent Technology and Engineering (IJRTE), Volume-2, Issue-4, pp.101-104, 2013.

[6]. J. Rowley, “Is higher education ready for knowledge management”, International Journal of Educational Management, vol. 14(7), pp. 325–333, 2000.

[7]. E. A. Hanushek and M. E. Raymond,” Does School Accountability Lead to Improved Student Performance”, pp.1-52, 2004.

[8]. B.K. Baradwaj and S. Pal , “ Mining Educational Data to Analyze Students Performance” IJACSA, Vol.2, No.6, pp.63-69, 2011.

[9]. Z. J. Kovacic , “Early prediction of student success: Mining student enrolment data”, Proceedings of Informing Science & IT Education Conference 2010.

[10]. M. Ramaswami, and R. Bhaskaran, “CHAID Based Performance Prediction Model in Educational Data Mining“, International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, pp.10-18, 2010.

[11]. M. Pandey and V. K. Sharma,”A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction”, *International Journal of Computer Applications , Volume 61– No.13, pp.1-5, 2013*.

[12]. J. Han and M Kamber, “Data mining concepts and techniques”, San Francisco, USA, Morgan Kaufmann, 2001.

[13]. M. O. Mansur, M. Sap and M. Noor, ”*Outlier Detection Technique in Data Mining: A Research Perspective*“, In Postgraduate Annual Research Seminar, pp.23-31, 2005.

[14]. P. Cortez and A. Silva, “Using Data Mining to Predict Secondary School Student Performance”. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference 2008, pp. 5-12, Porto, Portugal, 2008, EUROSIS, ISBN 978-9077381-39-7.

[15]. J. Ruby & K. David, “Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study “, International Journal for Research in Applied Science & Engineering Technology, Volume 2 Issue XI, November pp.80-84, 2014.

[16]. S. K. Gupta, S. Gupta & R. Vijay,” Prediction Of Student Success That Are Going To Enroll In The Higher Technical Education”. International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), Vol. 3, Issue 1, pp.95-108, 2013

[17]. A. TEKIN,” Early Prediction of Students’ Grade Point Averages at Graduation: A Data Mining Approach”. Eurasian Journal of Educational Research, Issue 54, pp.207-226, 2014.

[18]. Q. Al.Radaideh, E. Al-Shawakfa and M. Al-Najjar,”*Mining Student Data Using Decision Trees*“, The 2006 International Arab Conference on Information Technology – Conference Proceedings. Pp.1-5, 2006

[19]. E. Chandra, and K. Nandhini, “*Knowledge Mining from Student Data*“, European Journal of Scientific Research, vol. 47, no. 1, pp. 156-163, 2010.

[20]. V. Kumar, and A. Chadha, “*An Empirical Study of the Applications of Data Mining Techniques in Higher Education*“, International Journal of Advanced Computer Science and Applications, vol. 2, no. 3, pp. 80-84. 2011.

[21]. M.S. Mythili, Dr. A.R.Mohamed Shanavas. “*An Analysis of students' performance using classification algorithms*” IOSR Journal of Computer Engineering (IOSR-JCE), Issue 1, Ver. III , pp. 63-69,2014

[22]. Komal S. Sahedani, B Supriya Reddy ,” A Review: Mining Educational Data to Forecast Failure of Engineering Students “. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, pp.628-635, 2013

[23]. M. Bray, *The shadow education system: private tutoring and its implications for planners*, 2nd ed. UNESCO, PARIS, France, 2007.

[24]. Svetlana, S. Aksenova, Machine Learning with WEKA WEKA Explorer Tutorial for WEKA Version 3.4.3, 2004.

[25]. http://eric.univ-lyon2.fr/~ricco/tanagra/index.html. (Browing date: dec:2016)