Prediction Of Students Academic Performance Using Ensemble Methods Through Educational Data Mining

BOYA BHARGAVI
PG Scholar, Dept. Of CSE., G.Pullareddy Engineering College, Kurnool, A.P, India.

ABSTRACT: In the last decade Data mining (DM) has been applied in the field of education, and is an emerging interdisciplinary research field also known as Educational Data Mining (EDM). One of the goals of EDM is to better understand how to predict student academic performance analysis of students' characteristics. Another goal is to identify factors and rules that influence educational academic outcomes. In this paper, we use multiple classifiers (Decision Trees-J48, Naïve Bayes and Random Forest) to improve the quality of student data by eliminating noisy instances, and hence improving predictive accuracy. The prediction performance of three classifiers are measured and compared. It was observed that Naïve Bayes classifier outperforms other two classifiers by achieving overall prediction accuracy of 80%. This study will help teachers to improve student academic performance.

Keywords: Data Mining (DM), Educational data mining (EDM), Education System

I. INTRODUCTION

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique and increasingly large-scale data that come from educational settings and using those methods to better understand students, and the settings which they learn in. weather educational data is taken from students' use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities, it often has multiple levels of meaningful hierarchy, which often need to be determined by properties of the data itself, rather than in advance. Issues of time, sequence, and context also play important roles in the study of educational data.

EDM can use different DM techniques, each technique can be used for certain educational problem. As Example, to predict an educational model the most popular technique is classification. There are several algorithms under classification such as Decision tree, Neural Networks and Bayesian networks.

2.1 Need for the study

Various researches have been investigated to solve the educational problems using data mining techniques. However, very few researches shed light on student's behavior during learning process and its impact on the student's academic success. This research will focus on the impact of student interaction with the e-learning system. Furthermore, the extracted knowledge will help schools to enhance student's academic success and help administrators in improve learning systems.

2.2 Methodology

Design: Student’s Performance Prediction Model Research design

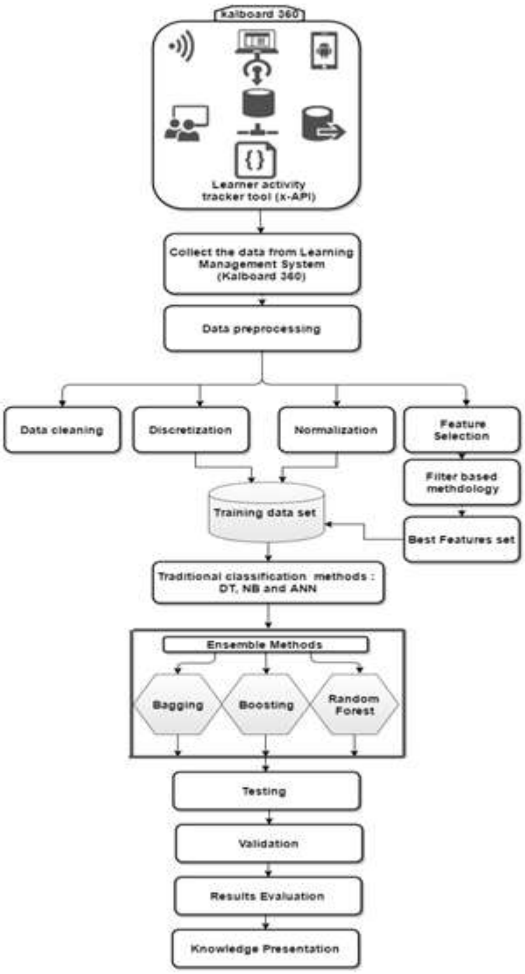


Figure 1

2.3 Data collection:

In this paper, data was collected from Kalboard 360 (LMS) system using experience API (xAPI)

In the current paper that data set extends into 500 students with 16 features. The features are classified into three main categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational Stage, grade Level and section. (3) Behavioral features, such as raised hand on class, Parent School Satisfaction.

2.4 Procedure:

We use discretization mechanism to transform the students’ performance from numerical values into nominal values, which represents the class labels of the classification problem. To accomplish this step, we divide the data set into three nominal intervals (High Level, Medium Level and Low Level) based on student’s total grade/mark such as: Low Level interval includes values from 0 to 69, Middle Level interval includes values from 70 to 89 and High Level interval includes values from 90-100. The data set after discretization consists of 127 students with Low Level, 211 students with Middle Level and 142 students with High Level. Then, we use normalization to scale the attributes values into a small range [0.0 to 1.0]. This process can speed up the learning process by preventing attributes with large ranges from outweighing attributes with smaller ranges. After that, feature selection process is applied to choose the best feature set with higher ranks. As shown in Figure7, we applied filter- based technique for feature selection.

In this paper, ensemble methods are applied to provide an accurate evaluation for the features that may have an impact on the performance/grade level of the students, and to improve the performance of student’s prediction model. Ensemble methods are categorized into dependent and independent methods. In a dependent method, the output of a learner is used in the creation of the next learner. Boosting is an example of dependent methods. In an independent method, each learner performs independently and their outputs are combined through a voting process. Bagging and random forest are example of independent methods. These methods resample the original data into samples of data, then each sample will be trained by a different classifier. The classifiers used in student's prediction model are Decision Trees (DT), Neural Networks (NN) and Naïve Bayesian (NB). Individual classifiers results are then combined through a voting process, the class chosen by most number of classifiers is the ensemble decision.

2.5 Data preprocessing:

This section will intensively talk about the data preprocessing. Data preprocessing is the step before applying data mining algorithm, it transforms the original data into a suitable shape to be used by a particular mining algorithm. Data preprocessing includes different tasks as data cleaning, feature selection and data transformation [23].

II. EXPERIMENTS AND RESULTS

3.1 Experiment:

We ran the experiments on the PC containing 6GB of RAM, 4 Intel cores (2.67GHz each). For our experiments, we used WEKA [25] to evaluate the proposed classification models and comparisons. Furthermore, we used 10-fold cross validation to divide the dataset into training and testing partitions.

3.2 Evaluation Measures

In our experiments, we use four common different measures for the evaluation of the classification quality: Accuracy, Precision, Recall and F-Measure [26, 27]. Measures calculated using Table 1, which shows classification confusion matrix based on the Equations respectively.

Table 1. Confusion Matrix

		Detected	
		Positive	Negative
Actual	Negative	False positive (FP)	True Negative(TN)
	Positive	True Positive (TP)	False Negative(FN)

Accuracy is the proportion of the total number of predictions where correctly calculated. Precision is the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases. Recall is the ratio of correctly classified cases to the total number of unclassified cases and correctly classified cases. In addition, we used the F-measure to combine the recall and precision which is considered a good indicator of the relationship between them [27].

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

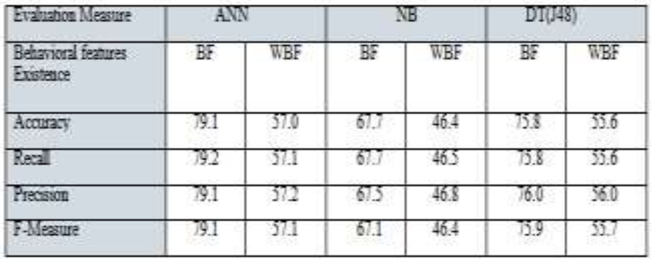
$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Fc} = 2\,\frac{Precisionc * Recall\ c}{Precisionc + Recall\ c}$$

3.3 Evaluation Results Using Traditional DM Techniques

There are many features directly or indirectly affecting the effectiveness of student performance model. In this section, we will evaluate the impact of behavioral features on student's academic performance using different classification techniques such as (DT, ANN and NB).

Table 2. Classification Method Results With Behavioral Features (BF) and Results without Behavioral Features (WBF)



Evaluation Measure	ANN		NB		DT(J48)	
Behavioral features Existence	BF	WBF	BF	WBF	BF	WBF
Accuracy	79.1	57.0	67.7	46.4	75.8	55.6
Recall	79.2	57.1	67.7	46.5	75.8	55.6
Precision	79.1	57.2	67.5	46.8	76.0	56.0
F-Measure	79.1	57.1	67.1	46.4	75.9	55.7

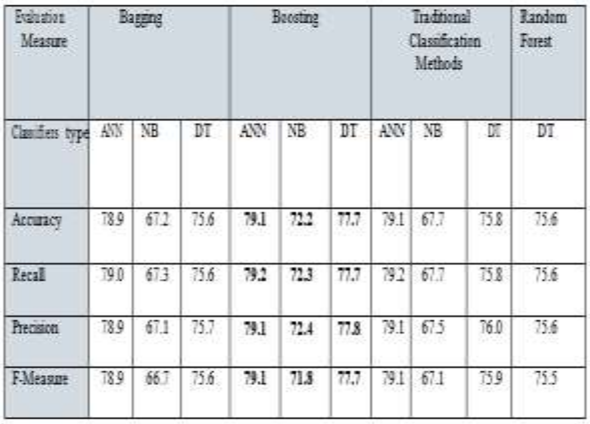
As shown in Table 2, we can notice that the ANN model outperforms other data mining techniques. ANN model achieved 79.1 accuracy with BF and 57.0 without behavioral features. The 79.1 accuracy means that 380 of 480 students are correctly classified to the right class labels (High, Medium and Low) and 100 students are incorrectly classified.

For the recall measure, the results are 79.2 with BF and 57.1 without behavioral features. The 79.2 recall means that 380 students are correctly classified to the total number of unclassified and correctly classified cases.

For the precision measure, the results are 79.1 with BF and 57.2 without behavioral features. The 79.1 precision means 380 of 480 students are correctly classified and 100 students are misclassified.

For the F-Measure, the results are 79.1 with BF and 57.1 without behavioral features. The experimental results prove the strong effect of learner behavior on student's academic achievement. We can get more accurate results by training the data set with ensemble methods.

Table 3. Classification Method Results Using Ensemble Methods



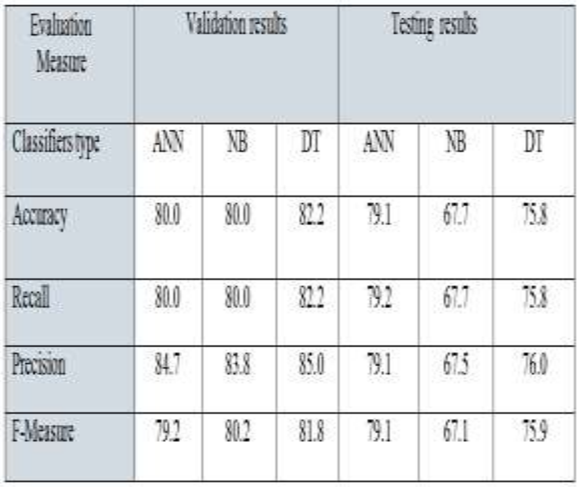
Evaluation Measure	Bagging			Boosting			Traditional Classification Methods			Random Forest
Classifiers type	ANN	NB	DT	ANN	NB	DT	ANN	NB	DT	DT
Accuracy	78.9	67.2	75.6	**79.1**	**72.2**	**77.7**	79.1	67.7	75.8	75.6
Recall	79.0	67.3	75.6	**79.2**	**72.3**	**77.7**	79.2	67.7	75.8	75.6
Precision	78.9	67.1	75.7	**79.1**	**72.4**	**77.8**	79.1	67.5	76.0	75.6
F-Measure	78.9	66.7	75.6	**79.1**	**71.8**	**77.7**	79.1	67.1	75.9	75.5

Boosting also achieved a noticeable improvement with NB model, in which the accuracy of NB using boosting increased from 67.7 to 72.2, which means the number of correctly classified students increased from 324 to 346 of 480 students. Recall results increased from 67.7 to 72.3, which means that 347 students are correctly classified to the total number of unclassified and correctly classified cases. Precision results are also increased from 67.5 to 72.4, which means 347 of 480 students are correctly classified. ANN model performance using boosting method is not differed much from ANN model results without boosting. Once the classification model has been trained using 10-folds cross validation, the validation process starts. Validation is an important phase in building predictive models, it determines how realistic the predictive models are. In this research, the model is trained using 500 students and the model is validated using 25 newcomer students. In validation, the data set contains unknown labels to evaluate the reliability of the trained model. Table

4, shows the evaluation results using several classification methods (ANN,

NB and DT) through testing process and validation process.

Table 4. Classification methods results through validation and testing



Evaluation Measure	Validation results			Testing results		
Classifiers type	ANN	NB	DT	ANN	NB	DT
Accuracy	80.0	80.0	82.2	79.1	67.7	75.8
Recall	80.0	80.0	82.2	79.2	67.7	75.8
Precision	84.7	83.8	85.0	79.1	67.5	76.0
F-Measure	79.2	80.2	81.8	79.1	67.1	75.9

As shown in table 4 we can notice that the evaluation measure results increased for the three prediction models through validation process. The three prediction models achieved accuracy more than 80%, which means that 20 of 25 new students are correctly classified to the right class labels (high, medium and low) and 5 students are incorrectly classified. The results of the validation process prove the reliability of the proposed model.

III. RESULT

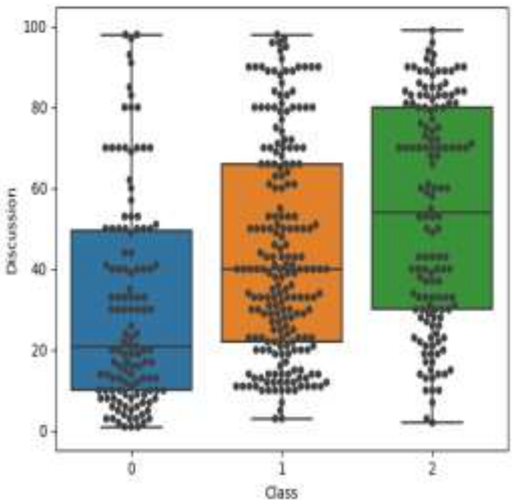


Figure 2

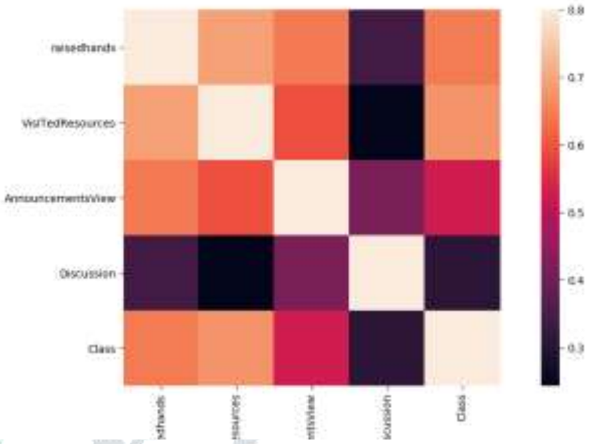


Figure 3

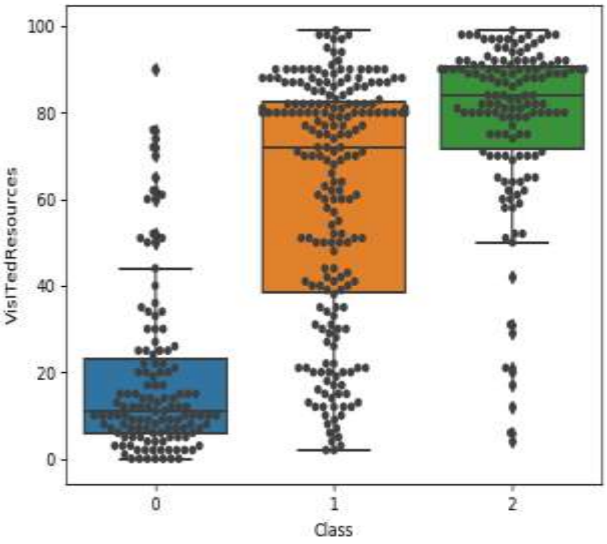


Figure 4

IV. CONCLUSION

The main objective of Educational Data Mining (EDM) is to improve teaching-learning process. Predicting students‘ performance is one of the major applications of EDM. So using decision treestudents‘ performance can be predicted. The students, whose performancse is poor, can be warned. The management can take necessary action to improve their performance by giving more attention, taking extra lectures etc. Due to such measures student performance can be improved. The number of failures can be reduced. Ultimately college results also get improved.

V. REFERENCES

[1] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert systems with applications, vol. 33, no. 1, **(2007)**, pp. 135-146.

[2] M. Hanna, "Data mining in the e-learning domain", Campus-wide information systems, vol. 21, no. 1,
(2004), pp. 29-34.

[3] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics", Part C: Applications and Reviews, IEEE Transactions on, vol. 40, no. 6, **(2010)**, pp. 601- 618.

[4] M. E. Zorrilla, E. Menasalvas, D. Marin, E. Mora and J. Segovia, "Web usage mining project for improving web-based learning sites", In Computer Aided Systems Theory–EUROCAST 2005, Springer Berlin Heidelberg, **(2005)**, pp. 205-210.

[5] A. M. Shahiri and W. Husain, "A Review on Predicting Student's Performance Using Data Mining Techniques", Proceeding Computer Science, vol. 72, **(2015)**, pp. 414-422.

[6] "Kalboard360-E-learning system", http://kalboard360.com/ (accessed February 28, 2016).

[7] G. Kakasevski, M. Mihajlov, S. Arsenovski and S. Chungurski, "Evaluating usability in learning management system Moodle", Information Technology Interfaces, 2008. ITI 2008. 30th International Conference on IEEE, **(2008)**, pp. 613-618.

[8] S. Rothman, "School absence and student background factors: A multilevel analysis", International Education Journal, vol. 2, no. 1, **(2001)**, pp. 59-68.

[9] J. DeKalb, "Student truancy. (Report No. EDO-EA-99-1). Washington, DC: Office of Educational Research and Improvement", (ERIC Document Reproduction Service No. ED429334), **(1999)**.

[10] S. Gunuc and A. Kuzu, "Student engagement scale: development, reliability and validity", Assessment & Evaluation in Higher Education, vol. 40, no. 4, **(2015)**, pp. 587-610.

[11] G. D. Kuk, "Assessing what really matters to student learning", Change, vol. 33, no. 3, **(2001)**, pp. 10- 17.

[12] I. Stovall, "Engagement and Online Learning. UIS Community of Practice for ELearning. http://otel.uis.edu/copel/EngagementandOnlineLearning.ppt, **(2003)**.

[13] C. S. Ong, and J. Y. Lai, "Gender differences in perceptions and relationships among dominants of e- learning acceptance", Computers in human behavior, vol. 22, no. 5, **(2006)**, pp. 816-829.

[14] C. Romero, S. Ventura, P. G. Espejo and C. Herv´as, "Data mining algorithms to classify students", in: Educational Data Mining, vol. 2008, **(2008)**.

[15] J. Ermisch and M. Francesconi, "Family matter: Impacts of family background on educational attainment", Economical, vol. 68, **(2001)**, pp. 137-156.

About Authors:

BOYA BHARGAVI is current pursuing M.Tech in CSE. dept., G.Pullareddy Engineering College, Kurnool , AP**.**