IOSR Journal of Engineering (IOSRJEN)
e-ISSN: 2250-3021, **p**-ISSN: 2278-8719
Vol. 3, Issue 2 (Feb. 2013), ||V1|| PP 37-42

A COMPARATIVE STUDY FOR PREDICTING STUDENT'S ACADEMIC PERFORMANCE USING BAYESIAN NETWORK CLASSIFIERS

P.V.Praveen Sundar
Research Scholar Hindusthan College of Arts & Science Coimbatore

Abstract: The main objective of educational institutions is to provide high quality of education. Providing a high quality of education depends on predicting the unmotivated students before they entering in to final examination. In this paper, we compare the Bayesian network classifiers for predicting the student's academic performance and generates a Model. This model helps earlier in identifying the drop outs and students who need special attention and allow the teacher to provide appropriate counselling / Advising. In Addition to this, Accurately predicting student performance is useful in many different contexts. For example, identifying exceptional students for scholarships is an essential part of the admissions process and identifying weak students who are likely to fail is also important for allocating limited tutoring resources.

Keywords: *Bayesian Networks Classifiers,Classification, Educational Data Mining(EDM), Prediction*

I. INTRODUCTION

Knowledge Discovery and Data Mining (KDD) is an interdisciplinary area focusing upon methodologies for extracting useful knowledge from data. The ongoing rapid growth of online data due to the Internet and the widespread use of databases have created an immense need for KDD methodologies. The challenge of extracting knowledge from data draws upon research in statistics, databases, pattern recognition, machine learning, data visualization, optimization, and high-performance computing, to deliver advanced business intelligence and web discovery solutions. In addition to this, recently there are increasing research interests in Educational Data Mining (EDM). EDM is a field that exploits statistical, machine-learning, and data-mining algorithms over the different types of educational data. Its main objective is to analyze these types of data in order to resolve educational research issues. EDM is concerned with developing methods to explore the unique types of data in educational settings and, using these methods, to better understand students and the settings in which they learn[1]. Whether 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 in the data itself, rather than in advance. Issues of time, sequence, and context also play important roles in the study of educational data.

The main objective of educational institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the students performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students' academic performance and trim down failure rate, to better understand students' behavior, to assist instructors, to improve teaching and many other benefits[2],[4].

Educational Data mining can be implemented in many techniques such as decision trees, neural networks, k-nearest Neighbor, Naive Bayes, support vector machines and many others. using these methods many kind of knowledge can be discovered such as association rules, clasification, clustering, pruning the data.

The main objective of this paper is to predict the student academic performace and make a comparative study on bayesian network classifers, through that we compute which classifier predicts more students when compared to other classifiers. In this paper, student's information like Previous Semester Performance, Attendance, Seminar , Assignment marks,Internal marks, and whether the student has attend any Co-curricular Activities are collected from students to predict the performance at the end of the semester examination.

II. RELATED WORKS

Although data mining in education is a recent research field, there are many works are already done in this area. that is because of its potiential to educational institutes.[4]gave a case study that used educational data mining to analyze students learning behaviour.[5][6] gave a case study that used educational data mining to identify behaviour of failing students to warn students at risk before final exam. [7] used educational data mining to identify and then enhance educational process in higher educational system which can improve their decision making process. [8] applied the classification of data mining technique to evaluate student performance, they used decision tree method for classification. the goal of their study is to predict the final grade of the students.the outcome of their results indicated that Decision tree model had better prediction than other models.

[9] applied the classification as data mining technique to evaluate student' performance, they used decision tree method for classification. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising.[10]applied the classification as data mining technique to evaluate student'performance, they used decision tree method for classification. This study allows the University management to prepare necessary resources for the new enrolled students and indicates at an early stage which type of students will potentially be enrolled and what areas to concentrate upon in higher education systems for support.[11] applied the association rule mining analysis based on students' failed courses to identifies students' failure patterns. The goal of their study is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students' performances.

[12]used k-means clustering algorithm to predict student's learning activities. the information generated after the implementation of data mining technique may be helpful for instructor as well as students.[13]used Bayesian Classification Method as a data mining technique and concluded that students grade in senior secondary exam, living location, medium of teaching, mother's qualification,students other habits, family annual income and students family status were highly correlated with the student academic performance.[14] used simple linear regression analysis and it was found that the factors like mother's education and student's family income were highly correlated with the student academic performance. [15]conducted study on the student performance using association rule technique and they find the interestingness of student in opting class teaching language.

III. DATA MINING PROCESS

Current Education System in India, a student's Performance is determined by their performance based on Internal marks and semester marks. The internal marks is carried out by the teacher based upon students performance in educational activities such as seminars taken, assignments, co-curicullar activities and performance in Internal exams. The end semester examination is one that is scored by the student in semester examination. each student has to get minimum marks to pass a semester in internal as well as end semester examination.

3.1. Data Preparations

Table-I Students dataset description.

Attribute	Description	Possible Values	Selected
Student_id	The Student _id		
Student_name	The Name of the Student		
Quota	The Quota in which student Joins.	{Management, Counselling}	
PSP	Average of Previous Semester Performance	{First ≥ 60%Second ≥ 45 & <60% Third ≥36 & <45% Fail <36%}	□
IM	Performance in Internal Exam	{Poor,Average, Good}	□
SEM	Performance in Seminars	{Poor, Average, Good}	□
ASS	Assignment	{Yes.No}	□
ATT	Attendance	{Poor,Average, Good}	□
CUR	Whether the student participate any co-curicullum activities	{Yes, No}	□
ESM	End Semester Marks , which is declared as response variable.	{First ≥ 60% Second ≥ 45 & <60% Third ≥36 & <45% Fail <36%}	□

In our comparision we contains details of First Year students of MCA Hindusthan college of Arts & Science- Coimbatore in the period of 2012-2013. Intially student dataset contains 48 record and 10 Attribute. Table –I presents the attributes and their description that exist in the data set as taken from source database. As part of the data preparation and preprocessing of the dataset and to get better input data for datamining techniques, we did some preprocessing for the collected data before loading the data set to the data mining software, irrevelant attributes should be removed. The attributes selected as seen in Table-I are processed via the Weka software to apply the data mining methods on them. The attributes such as the Student_Name or Student_ID,Quota are preprocessed using unsupervising filter "remove" and they are not selected to be part of the mining process; this is because they do not provide any knowledge for the data set processing and they present personal information of the students, also they have very large variances or duplicates information which make them irrelevant for data mining.

3.2.Model Construction

For our comparative Process, we use WEKA tool[17]. The WEKA Tool is a Open Source software which is fully implemented in the Java programming language and runs on any modern computing platform. it contains a comprehensive collection of data pre-processing and modelling techniques. Weka supports several standard data mining tasks like data clustering, classification, regression, pre-processing, visualization and feature selection. These techniques are predicated on the assumption that the data is available as a single flat file or relation.

After Prepropssing the data using Weka, Student_data.arff is created. This file was loaded into WEKA explorer. The classify panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. There are 13 algorithms under bayes classifiers like AODE, AODEsr, BayesNet,HNB, etc., which is implemented in WEKA. The algorithms used for our proposed work is AODEsr, WAODE, HNB, Naivebayes Updateable. Under the "Test options", the 10-fold cross-validation is selected as our evaluation approach. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model.

3.3. Performance Metrics

Once Predictive model is created, it is necessary to check how accurate it is, The Accuracy of the predictive model is calculated based on the precision, recall values of classification matrix.

PRECISION is the fraction of retrieved instances that are relevant. It is calculated as total number of true positivies divided by total number of true positivies + total number of false positivies.

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

RECALL is fraction of relevant instances that are retrieved.It is usually expressed as a percentage. It is calculated as total number of true positivies divided by total number of true positivies + total number of false negativies.

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negativies}}$$

ACCURACY is the overall correctness of the model and is calculated as the sum of correct classifications divided by the total number of classifications.

IV. EXPERIMENTAL RESULTS

The Table-II shows the accuracy of Naive bayes Updateable, HNB, WAODE,AODEsr algorithms for classification applied on the above data sets using 10-fold cross validation as follows:

Table-2 Classifiers Accuracy

Algorithm	**Correctly Classified Instances**	**Incorrected Classified Instances**
Naive Bayes Updateable	56.25	43.75
HNB	60.42	39.58
WAODE	62.5	37.5
AODEsr	64.58	35.42

Table –II shows that AODEsr algorithm has highest accuracy of 64.58 compared to other methods. Naive bayes Updateable, HNB, WAODE also showed an acceptable level of accuracy.
The Classifiers Accuracy can be represented in the form of graph.

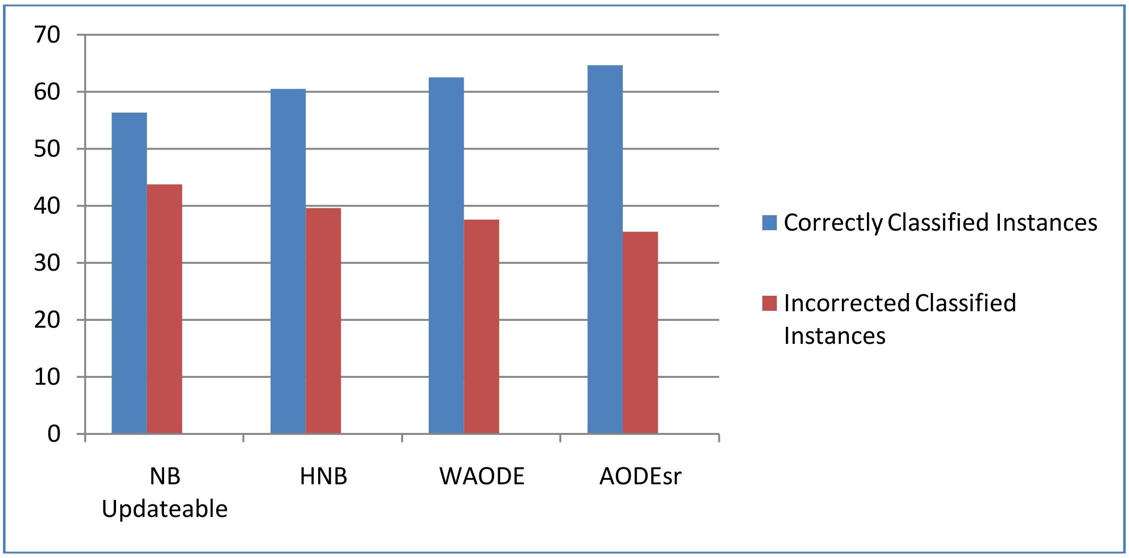
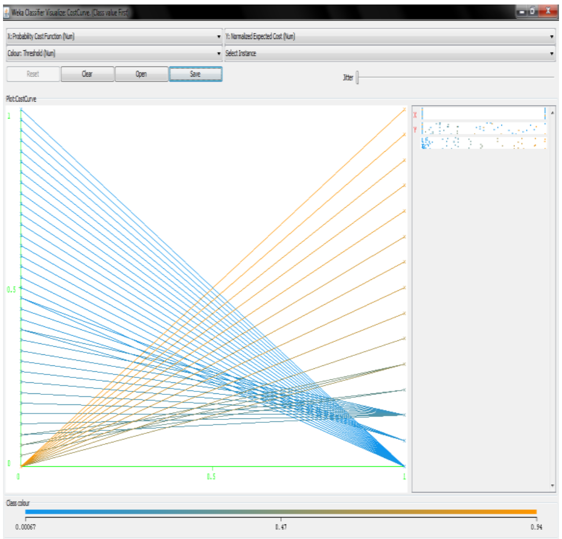


Fig 1. Comparision of Classifiers

The Classification matrix has been presented in Table-III,IV,V,VI, which compared the actual and predicted classifications.


Weka Classifier Visualize: CostCurve. (Class value First)

X: Probability Cost Function (Num)

Y: Normalized Expected Cost (Num)

Colour: Threshold (Num)

Select Instance

Fig 2. Cost Curve for AODEsr

Table-III
Classification matrix – Naive Bayes Updateable

ESM		**Predicted**				**Precision (%)**
		First	**Second**	**Third**	**Fail**	
Actual	**First**	8	4	2	0	100
	Second	0	10	4	0	55.56
	Third	0	4	5	4	35.71
	Fail	0	0	3	4	50
Recall(%)		57.14	71.43	38.46	57.14	

Table-IV
Classification matrix – Hidden Naive Bayes

ESM		**Predicted**				**Precision (%)**
		First	**Second**	**Third**	**Fail**	
Actual	**First**	10	3	0	1	90.91
	Second	1	10	3	0	55.56
	Third	0	5	6	2	46.15
	Fail	0	0	4	3	50
Recall(%)		71.43	71.43	46.15	42.86	

Table-V
Classification matrix – WAODE

ESM		**Predicted**				**Precision (%)**
		First	**Second**	**Third**	**Fail**	
Actual	**First**	10	2	1	1	90.91
	Second	1	10	3	0	62.5
	Third	0	4	7	2	46.67
	Fail	0	0	4	3	50
Recall(%)		71.43	71.43	53.85	42.86	

Table-VI
Classification matrix – AODEsr

ESM		**Predicted**				**Precision (%)**
		First	Second	Third	Fail	
Actual	First	11	2	1	0	91.67
	Second	1	10	3	0	62.5
	Third	0	4	7	2	46.67
	Fail	0	0	4	3	60
Recall(%)		78.57	71.43	53.85	42.86	

Table-VII
Overall Accuracy of Classifiers

Algorithm	**Overall Accuracy**	**Kappa Measure**
Naive Bayes Updateable	56.3%	0.408
Hidden Naive Bayes	60.4%	0.459
WAODE	62.50%	0.49
AODEsr	64.6%	0.51

The Table-VII shows that AODEsr algorithm has provides high overall accuracy rate than other algorithms.

.

V. CONCLUSION

In this paper, the classification task is used on student database to predict the students academic performance. As there are many approaches that are used for data classification, we use Bayesian Network Classifiers. Informations like Previous semester marks,Internal Marks,Performance on Seminars,Assignment,Attendance, Co-Curricular Activities were collected from the student‟ s database, to predict the performance of the end semester marks.

This study will help the students improve their performance and also it helps teacher to identify those students which needs a special attention to reduce failing ration and taking appropriate action at right time. Based on the Experimental Results we got AODEsr Algorithm predict more accuracy than any other Algorithms [3][5][6][7][8][11].

REFERENCES

[1] www.educationaldatamining.org.

[2] Mohammed M.Abu Tair,Alaa M.El-Hales," Mining Educational Data to Improve Student's Performance: A Case study", International Journal of Information and Communication Technology Research(ICT Jounal), 2012.

[3] *Heikki, Mannila, "Data mining: machine learning, statistics, and databases", IEEE, 1996.*

[4] Surjeet Kumar Yadav, Brijesh Bharadwaj,Saurabh Pal, *"Data Mining Applications:A Comparative study for Predicting Students Performance," International Journal of Innoviative Technology & Creative Engineering, 2011.*

[5] *Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009.*

[6] Merceron, A. and Yacef, K.,"Educational Data Mining: a Case Study" In Proceedings of the 12th International Conference on Artificial Intelligence in Education AIED 2005, Amsterdam, The Netherlands, IOS Press. 2005.

[7] *Galit.et.al, "Examining online learning processes based on log files analysis: a case study". Research, Reflection and Innovations in Integrating ICT in Education 2007.*

[8] Beikzadeh,M. and Delavari, N., "A New Analysis Model for Data Mining Processes in Higher Educational Systems". On the proceedings of the 6th Information Technology Based Higher Education and Training 7-9 July 2005.

[9] Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006) 'Mining Student Data Using Decision Trees', The 2006 International Arab Conference on Information Technology (ACIT'2006) – Conference Proceedings.

[10] Baradwaj, B. and Pal, S. (2011) 'Mining Educational Data to Analyze Student s' Performance', International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.

[11] Shannaq, B. , Rafael, Y. and Alexandro, V. (2010) 'Student Relationship in Higher Education Using Data Mining Techniques', Global Journal of Computer Science and Technology, vol. 10, no. 11, pp. 54-59.

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

[13] *Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data mining model for higher education system", Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.*

[14] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[15] *S. T. Hijazi, and R. S. M. M. Naqvi, "Factors affecting student's performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.*

[16] *U. K. Pandey, and S. Pal, "A Data mining view on class room teaching language", (IJCSI) International Journal of Computer Science Issue, Vol. 8, Issue 2, pp. 277-282, ISSN: 1694-0814, 2011.*

[17] http://www.dicom.uninsubria.it/~marco.vanetti/cfmatrix/

[18] www.cs.waikato.ac.nz/ml/weka/