Evaluation of Regional Benchmark Impact in EDM

P.V. Praveen Sundar[1], Dr.A.V. Senthil Kumar[2]

[1]Research Scholar, Hindusthan College of Arts & Science, Coimbatore, India.

[2] Director of MCA, Hindusthan College of Arts & Science, Coimbatore, India.

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 and motivates them before they enter into the final examination. There are so many factors which leads to unmotivated the students, such as college infrastructure, their living area, family annual income, Parents qualification, Past academic performance, students own interest on the course, other habits of the student, etc., There are many researches undertaken on the above factors and predict the unmotivated students, In this paper we mainly focus on how the geographical region plays a role on student's academic performance.

KEYWORDS:

Educational Data mining, Classification, Hidden naive Bayes, Place based learning.

1 INTRODUCTION

Educational data mining is a new research area that utilizes statistical, machine learning and data mining algorithms over the different types of educational data. The main objective of the Educational data mining is to adapt and solve the research issues on educational data and understand the students settings in which they learn. EDM allows discovering new knowledge based on students' usage data; it helps to validate/evaluate educational systems and to potentially improve some aspects of the quality of education, and to lay the groundwork for more effective learning process.

To provide a high quality of education, it is important to identify the unmotivated students and motivate them before they enter into the final exam. Generally Students' academic performance is based upon diverse factors like personal, social, psychological and other environmental variables[8][9][10]. A promising tool to attain this objective is the use of Data Mining. Data mining techniques are used to operate on large amount of data to discover hidden patterns and relationships which are helpful in decision making. Classification is a predictive data mining technique which makes prediction about values of data using known results found from different data [1].

In this study, we investigate how geographical region will make an impact on student's academic performance. Generally there is a myth called rural students are good in studies but have a less presentation skills, similarly urban students may or may not be good in studies but they have a good presentation skills, if so what is the status of the students who perform their studies in mixed area. Whether such students make any impact in educational settings.

To find out the solution for this problem, we use hidden naive bayes algorithm to classify the student dataset and then perform the independent t-test analysis to perform the significant difference among the study.

The rest of the paper is organized as follows Section-2 provides background information of this study. Section-3 describes the data mining process of the student data. Section-4 describes the experimental results and finally, conclusions and future works are outlined in Section-5.

2 RELATED WORKS

Although Educational data mining is a recent research field, there are many related research works done in this area, that is because of its vast potential to educational institutes. [2] which clearly indicates that rural and urban students are different in their learning styles and it suggest that students in rural schools appear to be more concerned and engaged in the educational process than urban students. Barcinas [3] indicates that students from the two areas are quite different in ethnicity. The rural students appears to be quite homogeneous, however the urban students seemed to have a greater mix of race and cultures. The lack of opportunity for rural students to interact with persons of varying backgrounds may be a limiting factor in their educational and sociological development and the educational level of the parents was higher in urban areas than in rural areas. Urban parents were more likely to expect their children to advance their education beyond high school. All these factors shows the difference in their social context between rural and urban areas. These differences help to explain the aspirations of students and finally the work suggests that students from rural areas should learn to live and work in an urban area. Bhardwaj and Pal [4] concluded that students living location are highly correlated with the student's academic performance. Yang[5] study reported that the decisions of youth to enter college was strongly influenced by the expectations of their parents

3 DATA MINING PROCESS

In this study, we have collect data from Hindusthan College of Arts & Science, Coimbatore, Tamilnadu, India. These data are analyzed using classification method to predict the student's performance. In order to apply this technique following steps are performed in sequence:

3.1 DATA PREPARATIONS

In our comparison we collect details from First Year and Second year students of MCA, Hindusthan college of Arts & Science- Coimbatore in the period of 2012-2013. Initially student dataset contains 200 record and 18 Attribute.

Table-I Students dataset description.

Attribute	**Description**	**Possible Values**
SSLC_Place	SSLC Area of Study	{Rural, Urban}
SSLC_Grade	Grade Obtained in SSLC	{Distinction ≥ 80 & <100% First ≥ 60 & <80% Second ≥ 45 & <60% Third ≥36 & <45%}
HSC_Study	HSC Area of Study	{Rural, Urban}
HSC_Grade	Grade Obtained in HSC	{Distinction ≥ 80 & <100% First ≥ 60 & <80% Second ≥ 45 & <60% Third ≥36 & <45%}
UG_Study	UG Area of Study	{Rural, Urban}
UG_Grade	Grade Obtained in UG	{Distinction ≥ 80 & <100% First ≥ 60 & <80% Second ≥ 50 & <60% Third ≥40 & <49%}
Status	Performance Status of the Student	{Improved, Decreased, No Change, Not Stable}

Table –I represents the attributes and their description of the final database for classifying the student's dataset. As a part of the data preparation and pre-processing of the dataset and to get better input data for data mining techniques, we have done some pre-processing for the collected data before loading the data set to the data mining software, irrelevant attributes should be removed. The attributes used in Table-I are processed via the Weka software to apply the data mining methods on them. The attributes such as the Student_Name , Student_rollno, Semester, SSLC_School_name, UG_School_name, Degree, HSC_School_name are not selected to be a part of the mining process; this is because they do not provide any knowledge on 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. In the above dataset Status is declared as a Response variable.

3.2 MODEL CONSTRUCTION

To classify Student's dataset, we use WEKA tool[6]. The WEKA Tool is an 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 Pre-processing 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. Among those algorithms we used Hidden Naive Bayes. 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.

Table –II shows the classifiers accuracy of the given dataset.

Table-II - Classifiers Accuracy

Algorithm	**Correctly Classified Instances**	**In Corrected Classified Instances**
HNB	**195**	**5**

Table –III shows the classification matrix of the given dataset using Hidden naive bayes algorithm.

Table-III- Classification matrix

Status		Predicted			
		Decreased	No Change	Improved	Not Stable
Actual	Decreased	43	0	0	0
	No Change	0	82	0	0
	Improved	0	2	43	0
	Not Stable	0	3	0	27

3.MEASURING KAPPA STATISTIC

Cohen's kappa measures the agreement between two raters, who each classify N items into C mutually exclusive categories.

The equation for κ is:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category.

$$\Pr(a) = \frac{(43+82+43+27)}{200} = 0.975$$

$$\Pr(e) = \frac{43}{200} * \frac{43}{200} + \frac{82}{200} * \frac{87}{200} + \frac{45}{200} * \frac{43}{200} + \frac{30}{200} * \frac{27}{200}$$

$$\Pr(e) = 0.293$$

$$\kappa = \frac{0.975-0.293}{1-0.293} = 0.9646$$

Kappa Coefficient value is 0.9646 and the calculation shows that there is 95% confidence interval: From 0.934 to 0.995 and the strength of agreement is considered to be 'very good'.

3.4 MEASURING PRECISION AND RECALL VALUES.

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 confusion matrix.

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

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

$$\text{Precision}= \frac{82}{82+2+3} =0.943$$

From the above we calculate the precision value for “No change” value.

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

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

$$\text{Recall}= \frac{43}{43+2} = 0.956$$

From the above we calculate the recall value of “Improved” .

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

$$\text{Overall Accuracy} = \frac{195}{200} =0.975$$

Table-IV

Status	Precision	Recall
Decreased	1	1
No Change	0.943	1
Improved	0.956	0.956
Not Stable	0.9	0.9
Overall Accuracy		0.975

Table-IV shows that the Precision and Recall value is greater than 0.5, it shows that the accuracy is very good.

4 EXPERIMENTAL RESULTS

In Addition to the above data mining process, we have conducted the two different experiments to find out the student academic score based on their area of study.

The following table describes the independent t-test analysis between uniform study area (Rural & Urban) and academic score of the students.

Table V: t-test between Uniform study area and Academic Score

A.Score / Study Area	N	Mean	SD	t-value	p-value
Rural	33	66.59	6.35	-0.149	0.881
Urban	74	66.80	6.72		

P-value is tested at 5% level

Among the total number of students considered for this study, 107 students were belonging to uniform group, which means (SSLC, HSC and College in uniform area). There are 33 students belonging to Rural area, 74 of them belonging to Urban area and their mean, standard deviation values respectively 66.59±6.35, 66.80±6.72. The t-value is -0.149 and p-value is 0.881, which is greater than the level of significance 0.05. Hence it is concluded that uniformity of study area have not found significant difference between Rural and Urban cases; therefore it can be considered as-a-whole. The subsequent table describes another t-test carried out between student's study area known as mixed area and uniform area.

Table VI: t-test between Study Area and Academic Score

A.Score **Study Area**	**N**	**Mean**	**SD**	**t-value**	**p-value**
Mixed	93	61.66	6.26	-5.562	0.000*
Uniform	107	66.74	6.58		

*P-value is tested at 5% level and it is significant for this case

The study area mixed represents the pattern of study area combination of Urban and Rural, whereas the uniform study area represents constant study area. The mean, standard deviation respectively 61.66±6.26, 66.74±6.58; t-value is -5.562, p-value is less than the level of significance 0.05, hence it is concluded that there is a significant difference of academic score present between mixed and uniform study area. It is observed from the study result that mixed area of study influences less academic score and in uniform case it is appreciated.

5.CONCLUSION

Through this study, the work found that the students who have completed their past studies in Uniform area (Either Rural/Urban) has no significant difference but the students who completed in Mixed Area has found a significant difference. While learning in a new mode, students struggled to adapt the new mode, on that occasion, teachers care for their learning, and assist them in the best way possible. Teachers may help students to recognize the need and enhance their learning capabilities as well, by emphasizing less frequently used ways of learning. Once the struggle is been overcome, the student will develop a complete, mature and integrated approach for learning. Thus area of study has created an impact on academic performance of students and it can be considered for futuristic educational frameworks.

6.REFERENCES

[1] AI-Radaideh,Q. A., AI-Shawakfa, E.M., and AI-Najjar, M. I.,"Mining Student Data using Decision Trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[2] David E. Cox,Elizabeth Kendall Sproles And George B. Sproles, "Learning Style Variations Between Rural and Urban Students", "Research in Rural Education", Volume5, Number 1, 1988.

[3] J. David McCracken and Jeff David T. Barcinas,"Differences Between Rural and Urban Schools, Student Characteristics, and Student Aspirations in Ohio",Journal of Research in Rural Education, Winter, 1991, Vol. 7, No.2, (pp. 29-40).

[4] Brijesh Kumar Bhardwaj and Saurabh Pal, "Data Mining: A prediction for performance improvement using classification",(IJCSIS) International Journal of Computer Science and Information Security,Vol. 9, No. 4, April 2011.

[5] Yang, S. W. (1981). Rural youths 'Decision to attend college: Aspirations and realizations. Paper presented at the annual meeting of the Rural Sociological Society, Guelph, Ontario, Canada (ERIC Document Reproduction Service No. ED207765).

[6] Weka (2007). www.cs.waikato.ac.nz/ml/weka/

[7] Cristobal Romero and Sebastian Ventura, "Educational Data Mining: A Review of the State of the Art", IEEE Transactions on Systems, Man, And Cybernetics, Vol. 40, No. 6, November 2010.

[8] F. Weng, F. Cheong and C. Cheong, "IT education in Taiwan relationship between Self-efficacy and Academic Integration among Students" Pacific Asia Conference on Information Systems (PACIS), Association for Information Systems, 2009, 22

[9] L. P. Womble, "Impact of Stress Factors on College Students Academic Performance"Undergraduate Journal of Psychology. 2003;16

[10] Olufunke O. Oladipupo1 , Olanrewaju. J. Oyelade and Dada. O. Aborisade, "Application of Fuzzy Association Rule Mining for Analysing Students Academic Performance", International Journal of Computer Science Issues, Vol. 9, Issue 6, No 3, November 2012.