Using Data Mining to Predict Primary School Student Performance

Manmohan Singh, Harish Nagar, Anjali Sant

Reserch Scholer Department of Computer Science and Engg. Mewar University, Rajasthan, India

ABSTRACT

Predicting the performance of a student is a great concern to the Primary education. In this paper is to identify the factors influencing the performance of students in previous examinations result and find out a suitable data mining algorithm to predict the grade of students so as to a give timely and an improve student performance at next year exam result. This paper intends to analysis the Betul dist. In M.P India, Rural and Urban primary school students' performance in different categories of measurements in the present investigation, a survey cum experimental methodology was adopted to generate a database and it was constructed from a primary and a secondary source. The obtained results from hypothesis testing reveals that type of school is not influence student performance and area of school and student pervious result' occupation a major role in predicting of student s grades/result.

Keywords*: Data Mining, Decision Tree, Naïve Bayes Algorithm. Zero R Classification Algorithms.*

1. INTRODUCTION

Measuring of academic performance of students is challenging since students academic performance hinges on diverse factors like personal, socio-economic, psychological and other environmental variables. The scope of this paper is to predict the student marks and what are the factors that influence the performance of the students [3].

Data mining provides many tasks that could be used to study the student's performance. In this paper, the classification task is used to evaluate performance of a student and as there are many approaches that are used for data classification, the decision tree and Naïve Bayes, other methods was used here [4]. For this study, recent real world data were collected from different Rural and Urban primary schools in Betul district M.P. Two different sources, prepare one questionnaires were used. Information's like many filled were collected from students through questionnaire.

The main objectives of this study are

- Identification of highly influencing predictive variables on the academic performance of Primary school student .
- Find the best Decision Tree, Naïve Bayes Algorithm. Zero R Classification Algorithms on student data.
- Predict the grade/result or batter performance in examination.
- Apply data mining suitable techniques using Rural and Urban primary school student's data.

2. REVIEW OF LITERATURE

A number of reviews pertaining to not only the diverse factors like personal, socio-economic, psychological and other environmental variables that influence the performance of students but also the models that have been used for the performance prediction are available in the literature and a few specific studies are listed below for reference.

M.Ramaswami and R.Bhaskaran [1] have used CHAID prediction model to analyze the interrelation between variables that are used to predict the outcome of the performance at higher secondary school education. The features like medium of instruction, marks obtained in secondary education, location of school, living area and type

of secondary education were the strongest indicators for the student performance in higher secondary education. The CHAID prediction model of student performance was constructed with seven class predictor variable.

Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme [2] have applied machine learning techniques to improve the prediction results of academic performances for two the real case studies. Three methods have been used to deal with the class imbalance problem and all of them show satisfactory results. They first re balanced the datasets and then used both cost-insensitive and sensitive learning with SVM for the small datasets and with Decision Tree for the larger datasets. The models are initially deployed on the local web.

Arockiam et al. [3] used FP Tree and K-means clustering technique for finding the similarity between urban and rural students programming skills. FP Tree mining is applied to sieve the patterns from the dataset. K-means clustering is used to determine the programming skills of the students. The study clearly indicates that the rural and the urban students.

Banerjee, Banerji et.al [4] the authors write that impact of a program implemented by Parham in partnership with the state governments of Uttarakhand and Bihar. It attempted to scale up remedial instruction in public schools, and find that summer camps conducted by regular teachers using remedial material were effective in raising test scores. They find that there was no impact of other models that attempted to incorporate this pedagogy in the regular school day. The authors interpret their findings to suggesting that the remedial pedagogy was successful, but that it was difficult to get teachers to implement new curriculums during school hours.

Jayaraman et.al [5] the authors Write intake of school going children in rural Madhya Pradesh, by 50% to 100%. Using a difference estimation strategy that relies on a staggered roll out across schools, attendance rates for girls are estimated to increase by 12 percentage points in rural Madhya Pradesh and 5 percentage points overall in Delhi However, these papers do not study the impact of mid-day meals on test scores. They data from thirteen states to construct triple-difference estimates using private schools as a control group and find that the mid-day meal program is associated with a 6.8% increase in enrolment, but has no impact on test scores.

3. METHODOLOGY

Now Through extensive search of the literature and discussion with experts on student performance, a number of factors that are considered to have influence on the performance of a student were identified. These influencing factors were categorized as input variables. The output variables on the other hand represent some possible grades. The primary data were collected from the primary school students using to questionnaire.

For this study, recent real world data were collected from primary school student data in dist Rural and urban area private and government school students. Fifty schools were randomly selected from Betul district. A sample of 1000 students was taken from a group of schools and class only III to V Students were grouped in a classroom where they were briefed clearly about the questionnaire and it took on average half an hour to fill the questionnaire. Selection of students was at random [4].

The primary data was collected using a questionnaire which includes questions related to several personal, socio-economic, psychological and school related variables that were expected to affect student performance. The questionnaire was reviewed by the professionals and tested on a small set of 100 students in order to get a feedback. The final version contained 24 questions and it was answered by more than 1000 students. Latter a sample of 600 were selected from the whole. All 500 questionnaires were filled with the response rate of 100% out of which 198 were females and 302 were males.

4 .TOOLS AND TECHNIQUES

Classification trees are widely used in different fields such as Rural area government and Private school and Urban area private Government School .at Class only III to V student, helping to make them easier to interpret than they would be if only a strict numerical interpretation were possible. For this study WEKA"s implementation of Decision Tree Algorithm.Naïve Bayes Algorithm. Zero R Classification Algorithms [9].

4.1. Data pre-processing

As it is common in data mining, before running tests on data instances, it is necessary to clean and prepare the data for use into the WEKA workbench. An important piece here was the need to convert string data into nominal data

From the ARFF file. This was done based upon the requirements constraints of the algorithms used, as they do not accept string data for processing. In addition, it was important to look at relevance of the attributes to remove redundant, noisy, or irrelevant features. In the data, two attributes students register number and their name were removed. In this study, replace missing values file in WEKA was used to replace all missing. The distribution towards the mean value of the most frequent value for the related attribute then select attributes" were used to rank the attributes. Attribute selection involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction. To do this, two objects must be set up: an attribute evaluator and a search method. The evaluator determines what method is used to assign a worth to each subset of attributes. The search method determines what style of search is performed [8].

5. RESULTS

The following are the attributes and the corresponding hypothesis to verify the relationship between the attributes.

Chi-square test (χ^2) is one of the simplest and most widely used parametric as well as non parametric tests in statistical work. The Chi-square value is used to judge the significance of population variance. We used Chi-square test to find the significance between the different attributes and grade obtained by student. The results of hypothesis testing are given in Table 1.

Table 1Result Analysis

	Naïve Bayes	Decision Tree	Zero R
A	76	76	71
B	58	58	54
C	70	70	53
D	94	94.45	76

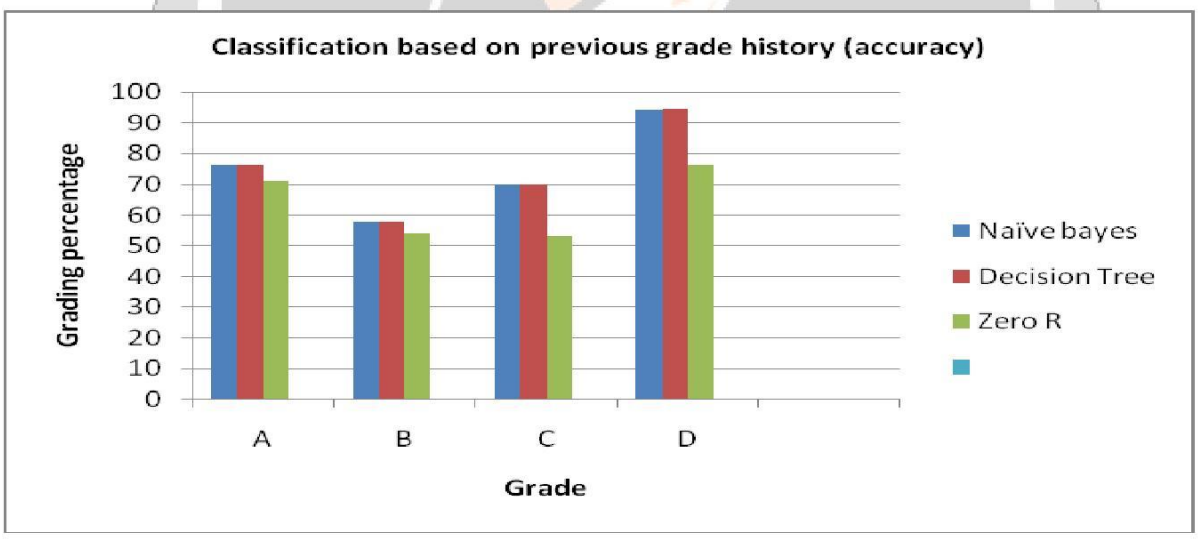

Figure 1 Result Graph

In this Paper study show that student"s previous grade can be used to generate a predictive model using decision tree algorithm and naïve Bayes and zero R, which can be used for predicting the final grade of primary School Student. The accuracy of the model is 85.53 percentages which means that the model is successfully predicting the final grade of student"s higher grade out of 600 students has been successfully classified. The teacher, students and their parents can improve the result of student who is likely to pass in low grade through proper counseling.

6. CONCLUSION AND FUTURE WORK

Data mining techniques allow a high level extraction of knowledge from raw data, offering interesting possibilities for the primary education domain. In this study a data mining model was developed based on some selected input variables collected through questionnaire method. After testing some hypothesis, some of most influencing factors were identified and taken to predict the obtained results from hypothesis testing reveals that type of school is not influence student performance. As a previous result, having the information generated through our experiment, Furthermore, we intent to enlarge the experiments to collect additional features like psychological factors which disturb the students ,motivational efforts taken by the teachers and applying suitable data mining clustering and neural network approach primary school students data .

REFERENCES

[1] M.Ramaswami and R.Bhaskaran, "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science Issues Vol. 7, Issue 1, No. 1, January 2010.

[2] Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt- Thieme, "Improving Academic Performance Prediction by Dealing with Class Imbalance", 2009 Ninth International Conference on Intelligent Systems Design and Applications.

[3] L.Arockiam, S.Charles, I.Carol, P.Bastin Thiyagaraj, S. Yosuva, V. Arulkumar, "Deriving Association between Urban and Rural Students Programming Skills", International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 687-690

[4] Banerjee, Abhijit, Rukmini Banerji, Esther Duflo, and MichaelWalton. 2012. Effective Pedagogies and a Resistant Education System:Experimental Evidence on Interventions to Improve Basic Skills in Rural India. MIT.

[5] Jayaraman, Rajshri, Dora Simroth, and Francis De Vericourt. 2010. The Impact of School Lunches on Primary School Enrollment: Evidence from India's Mid-Day Meal Scheme. Indian Statistical Institute l.

[6] Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto. 2013, 'Predicting School Failure and Dropout by Using Data Mining Techniques' , IEEE Journal Of Latin-American Learning Technologies, Vol. 8, No. 1, Feb 2013 .

[7] Dréze, Jean and Kingdon, Geeta, 2000. "School Participation in Rural India," Review of Development Economics.

[8] Delavari, N., Beikzadeh, M.R. (2004). A new model for using data mining in higher education system, 5th international Conference on Information Technology based Higher education and training: ITEHT (04), Istanbul, Turkey, 31st m\May-2nd June 2004.

[9] Das J,S, dercon , J Habyarimana ,P Krishna ,K Muralidharana and V.sundararaman 2013 " School Input ,hose hold substitution and test scores" American Economic Journal Applied Economics 5(2), p-29-29.

[10] Ghaida Abu, D, and Stephan Klasen (2004), the Costs of Missing the Millennium Development Goal on Gender Equality."0World Development32 (7): 1075–107.