IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Prediction Accuracy of Academic Performance of Students using Different Datasets with High Influencing Factors

Jai Ruby[1], Dr. K. David[2]

Research Scholar, Research & Development Centre, Bharathiar University, Tamilnadu, India[1]

Assistant Professor, H.H. Rajahs College, Pudukkottai, Tamilnadu, India[2]

Abstract: For the past few years, a lot of higher educational institutions tussle for providing quality education. To provide a quality education the institutions need information regarding their students. In higher education institutions a substantial amount of knowledge is hidden and need to be extracted. Various Data mining techniques are useful for deriving such hidden knowledge. The knowledge can be any student specific information like academic performance, dropouts, course preference, subject specialization etc. The quality of the students in a higher education institution is categorized by their academic performance. Many socio economic, non academic and academic factors influence the performance of the students. The factors that describe student performance can be used for predicting students performance by using a number of well - known data mining classification algorithms such as ID3, Simple CART, J48, NB Tree, MLP, Bayesnet etc. The model is mainly focused on finding the prediction accuracy of academic performance of students using two different datasets. The experimental model also proves that the student attributes considered are highly influential in predicting the results using MLP classification algorithm.

Key Words: Educational Data Mining, Academic Performance, Prediction, Classification, Influencing Factors.

I. INTRODUCTION

In this modern era, population growth leads to the establishment of new Educational Institutions. Educational institutions are becoming more competitive because of the number of institutions growing rapidly. To stay afloat, these institutions are focusing more on improving various aspects and one important factor among them is quality learning. For providing quality education, the institutions need to know about their strengths which are explicitly seen and which are hidden. To be competitive, the institutions should identify their own strengths hidden and implement a technique to bring it out. In recent years, Educational Data Mining has put on a massive recognition within the research realm as it has become a vital need for the academic institutions to improve the quality of education. For the higher education institutions to enhance their quality it is a must for them to extract a substantial amount of hidden knowledge. The technique behind the extraction of the hidden knowledge is Knowledge Discovery process that extracts the knowledge from available dataset and should create a knowledge base for the benefit of the institution. Higher education does categorize the students by their academic performance. Many factors influence the academic performance of the student. The study model [1] is mainly focused on exploring various indicators that have an effect on the academic performance of the students. The extracted information that describes student performance can be stored as intelligent knowledge and can be used by the institutions principally for predicting the student's performance in advance.

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering and extracting hidden and potentially useful information from large amounts of data. Recently, Data mining is widely used on educational dataset and it is termed as Educational Data mining (EDM). EDM has become a very useful research area [2]. Educational Data Mining refers to techniques, tools, and research designed for automatically extracting the pattern from large repositories of data generated by or related to people's learning activities in educational settings. Key uses of EDM include learning and predicting student performance in order to recommend improvements to current educational practice. EDM can be considered as one of the learning sciences, as well as an area of data mining [3]. Some of the benefits of data mining in an education sector are identifying students' preferences towards course choices, their selection of specialization and predicting students' knowledge, grades, and final results [4]. Students' data was collected, pre-processed and a combination of data mining techniques were applied and a substantial amount of hidden knowledge was extracted that described students' behaviour [5]. Institutions of Higher Learning (IHL) are similar to knowledge businesses, in that both are involved in knowledge creation, dissemination, and learning [6].

However, people in business world are concerned with the profit they could gain by exploiting knowledge through the implementation of KMS whereas IHL consider that KMS could improve the quality of service deliveries and sustained competitive advantages in the academic world [7]. In this paper, the researcher applied MLP, a data mining classification algorithm over two different

DOI 10.17148/IJARCCE.2016.5217

educational domain dataset for predicting students' academic performance. The researcher also made an effort to compare the prediction accuracy of the two different datasets. The experiment also proved that various factors identified in the study model [1] are highly influencing factors in predicting the results.

This study is the third in a series of studies on analysis of student's academic performance using classification data mining algorithms. This paper makes a new attempt to look into the higher educational domain of data mining to analyze the students' performance with different datasets. Section 2 gives the methodology and techniques. Section 3 provides the general account of the model and the dataset under study. Section 4 predicts the academic performance of students using MLP classification algorithm over the datasets and the comparative analysis. Conclusion and a discussion on future work are in the final section.

1.1 Related Work

J. Shana and T. Venkatachalam [8], used various feature selection methods and have found out the influence of features affecting the student performance. The authors have used a selected number of attributes and have not taken attributes like attendance, theory, laboratory etc. Brijesh Kumar Baradwaj, Saurabh Pal, in [9] conducted a study on a data set of size 50 Post Graduate students for mining educational data to analyze students' performance. Decision tree method was used for classification and to predict the performance of the students. Different measures that are not taken into consideration were economic background, technology exposure etc. El-Halees.A [10] has done a work on mining students data to analyze learning behavior. The data size considered was 151. The details include personal and academic records of students. Classification based on Decision tree is done followed by clustering and outlier analysis. The knowledge extracted describes the student behaviour. Jai Ruby and David [1], presented a study on the student data and identified that 7 factors are high influencing factors for predicting the students' academic performance. From the 16 initial factors Medium of Study, UG Percentage, Theory marks obtained, Stay, Extra Curricular Activities and Family Income and whether the student was good in Previous Course studied were identified as the influencing factors by using various feature selection techniques like Chi square, Information Gain, Correlation, Linear Regression and Gain Ratio.

Z. J. Kovacic, in [11] presented a study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success. The algorithms CHAID and CART were used. K.Shanmuga Priya and A.V.Senthil Kumar [12]., applied a Classification Technique in Data Mining to enhance the student's performance by extracting the knowledge from the end semester mark. Ramaswami M., and Bhaskaran in [13] have constructed a predictive model using 772 students' records with 7-class response variables by using highly influencing predictive variables obtained through feature selection.

Bengio Y, et.al [14], discussed that neural networks are suitable in data-rich environments and are typically used for extracting embedded knowledge in the form of rules, quantitative evaluation of these rules, clustering, self-organization, classification and regression. Neural networks have an advantage, over other types of machine learning algorithms, for scaling. The use of classification models such as neural network, decision tree and Naïve Bayes has also been popular in the education field to predict students' behavior as proved in this research work [15].

M. Wook, Y. Hani Yamaya, N. Wahab, M. Rizal Mohd Isa, N. Fatimah Awang and H. Yann Seong compared two data mining techniques which are: Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting and classifying student's academic performance. As a result, the technique that provides accurate prediction and classification was chosen as the best model. Using the proposed model, the pattern that influences the student's academic performance was identified [16]. Jai Ruby & David in [17] compared various data mining algorithms using the student dataset considering only the influencing factors and proved that MLP, a neural network based classification show best result of 74.8% accurate prediction which is followed by ID3 showing an accuracy of 73%.

Romero and Ventura [18], have a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it has a specific requirements not presented in other domains. Kuyoro' et.al [19] done a work on identifying the optimal algorithm suitable for predicting first-year tertiary students academic performance based on their family background factors and previous academic achievement. Five decision tree algorithms -Random forest, Random tree, J48, Decision stump and REPTree and five rule induction algorithms – JRip, OneR, ZeroR, PART, and Decision table and a multilayer perceptron, an artificial neural network function were taken for the study. It is discovered that random tree performance is better than that of other algorithms used in this study. Jai Ruby & David [20], presented a comparative study model to predict the accuracy of the academic performance of the students using Multi Layer Perceptron algorithm. The experiment proved that the attributes identified in the study [1] are practically high influencing factors in predicting student performance.

II. METHODOLOGY AND TECHNIQUES

Data mining also termed as Knowledge Discovery in Databases (KDD) refers to extracting or "mining" knowledge from large amount of data Han & M. Kamber , [30]. Fig.1 shows the process of extraction of well defined pattern as a result of mining the data.

Fig. 1 – Conversion of data into a pattern

DOI 10.17148/IJARCCE.2016.5217

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Knowledge Discovery process involve various steps like Data cleaning, transformation, data mining, pattern evaluation in extracting knowledge from data. Knowledge Discovery is involved in a multitude of tasks such as association, clustering, classification, prediction, etc. Classification and prediction are functions which are used to create models that are constructed by analyzing data and then used for assessing other data. Classification techniques can be applied on the educational data for predicting the students' behaviour, performance in examination etc. Basic techniques for classification are decision tree induction, Bayesian classification and neural networks. A number of well - known data mining classification algorithms such as ID3, REPTree, Simplecart, J48, NB Tree, BFTree, Decision Table, MLP, Bayesnet, etc., exist.

2.1MLP

Multilayer Perceptron algorithm is one of the most widely used and common neural networks. Multilayer 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. An MLP consists of multiple layers of nodes in a directed graph, with every layer totally connected to the consequent one. These are supervised networks so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification and prediction.

2.2 Feature Selection Techniques:

An educational dataset may contain a number of student oriented attributes. Various attribute selection methods do exists to identify the attributes that make great impact. Some of the notable methods are chi-square, information gain, correlation, gain ratio, and regression. Chi-square test is a statistical method used to identify degree of association between variables [21]. Information Gain and Gain Ratio are used to determine the best attribute. Linear Regression involves finding the best line to fit two variables so that one variable can be used to predict other and to find a mathematical relationship between them. Correlation is used to assess the degree of dependency between any two attributes.

III. DATASET

Two different datasets are considered for the study. One of the dataset used for this study for performance was taken from PG Computer Application course offered by an Arts and Science College between 2007 and 2012. The data of 165 students were collected. Student personal and academic details along with their attendance were collected from the student information system. The collected information was integrated into a distinct table. Student dataset contains various attributes like Theory Scores, Laboratory scores, Medium of study, UG course, Family Income, Parental Education, First Generation Learner, Stay, Extracurricular activities etc. Among the different attributes initially present using feature selection techniques like chi square, info gain, gain ratio, correlation and regression it is found that the high impact attributes that contribute for the performance of the students are Theory, Medium of Study, Previous Course studied, UG Percentage, Stay, Extra Curricular Activities and Family Income [1]. The influencing attributes are selected and are used to classify and predict the student performance using MLP one of the best classification algorithm identified [17].

The second data set considered was taken from UCI Machine Learning Repository. The dataset consists of 396 records. The dataset is a multivariate dataset for student performance with 33 number of attributes. The attributes include student grades, demographic, social and school features. The dataset is collected using reports and questionnaires. Among the 33 attributes, most influencing attributes identified in the study [1] are extracted into a distinct table. The attributes considered were fathers job, mothers job, travel time, first period grade, extracurricular activities, second period grade and previous class performance. These attributes were selected for predicting the students' academic performance using MLP classification algorithm.

Weka tool was used for the study. The classify panel in the weka tool facilitates to apply classification algorithms and to estimate the accuracy of the predictive model. Among the different classifiers of ID3, J48, NBTree, RepTree, Multi Layer Perceptron (MLP), SimpleCart and Decision table, the study model [17] show that MLP learning algorithm proved to be the best. The aim of this study is to justify that the found out high impact attributes that contribute for the performance of the students using feature selection are true and to justify that it is true for different sets of data.

IV. RESULTS AND DISCUSSION

The experiment was carried out using the data set with 165 records using only 7 high impact attributes (family income, previous course studied, UG percentage, stay, medium of study, theory marks, extracurricular activity). The 8th attribute in Fig.2 represents the unknown 'Result' attribute that is to be predicted by the algorithm.

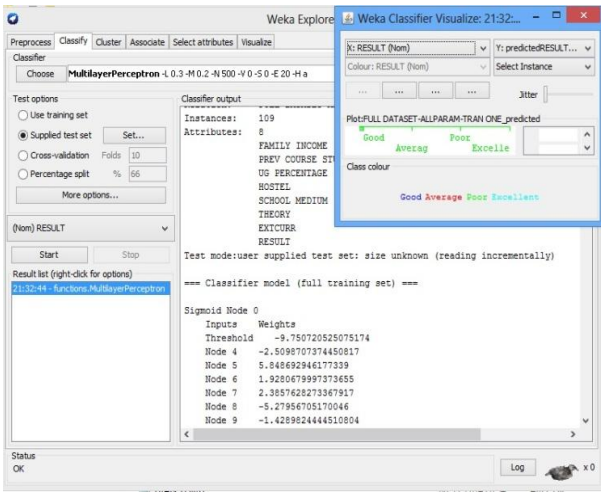

Fig.2 A sample run on the first dataset uses a train set and a test data with an unknown class label - Result

DOI 10.17148/IJARCCE.2016.5217

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Was split into two sets consisting of two-third as training set and one-third as testing set. The training set is used to build a model and the test set is used to estimate the accuracy of the classifier and if it is acceptable then it is used for the prediction of data for which the class label is unknown. MLP was the data mining classification algorithm chosen for the study via weka. The dataset set was divided into 5 sets (train) of distinct two-third records and 5 sets (test) of distinct one-third records. These sets were used in Run1 through Run5 respectively. In each run, 3 new set of data whose class labels are unknown was given for prediction. Since we use three different sets of new data for prediction, the average of the 3 results was considered for each run. Since we have 5 training data set and 5 test data set we get 5 results.

The same experiment was repeated for the second data set obtained from UCI Machine Learning Repository considering only the high influencing attributes (fathers' job, mothers' job, travel time, first period grade, extracurricular activities, second period grade and previous class performance). The 8th attribute is the "Result" attribute that is unknown and to be predicted. Fig.3 shows the sample run of the UCI dataset with a train and test test with a unknown class label – Result.

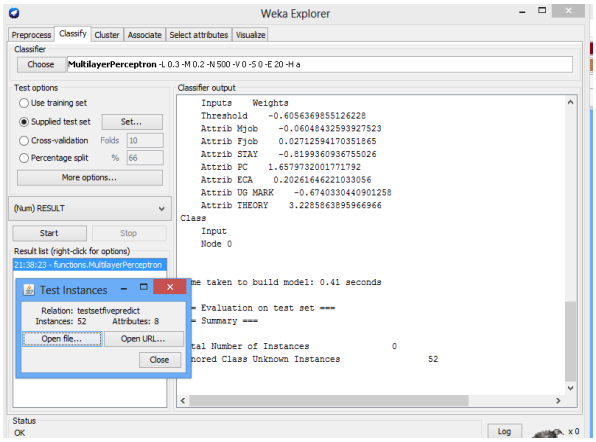

Fig.3 A sample run on the UCI dataset uses a train set and a test data with an unknown class label - Result

Table-1 shows prediction accuracy of the MLP classification algorithm for the two different datasets, ie from Arts and Science College and from UCI machine learning repository.

Table -1 prediction accuracy of MLP algorithm for the two different dataset for different runs

Data Set	Run 1	Run 2	Run 3	Run 4	Run 5
Arts and Science College Data Set	73.7	60	57.3	49	82.6
UCI Machine Learning Repository	89.9	93.3	88	94.6	91.3

In both cases, only the high influence attributes were considered. In each run, 3 new set of data whose class labels are unknown was given for prediction and the average of the 3 results was considered in terms of percentage.

Table 2 shows the prediction percentage by computing the average of the 5 runs using for the two different datasets.

Table -2 Average prediction accuracy of the two different dataset using MLP

Data Set	Prediction Percentage
Arts College	64.5
UCI Machine Learning Repository	91.42

The results show that prediction percentage of dataset with high influence attributes behave alike whatever may be the datasets. The prediction accuracy of the two different dataset using MLP in different runs is found to vary between 50% and 95%. The first dataset shows a average prediction percentage of 65%. The UCI dataset shows a better prediction percentage of 91 % as the dataset has a very good and ideal train and test data.

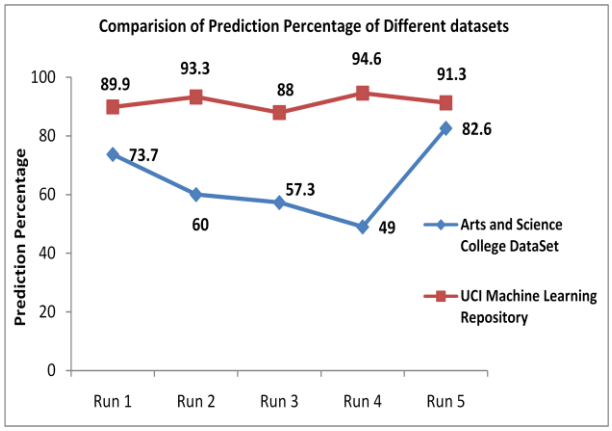

Fig.4 Comparison of Prediction Percentage of different datasets with high influence attributes using MLP

Fig.4 shows that the UCI dataset gives a prediction accuracy above 88% for different test data. For the Arts and Science dataset it is found that the variation ranges from 49% to 82.6% for different test data. Thus it is observed that if the dataset is an ideal dataset the prediction accuracy is found to be high irrespective of the train and test dataset. The result shows that the attributes identified in the study are practically high influencing factors in predicting student performance for any type of dataset.

V. CONCLUSION

This model is mainly focused on analyzing the prediction accuracy of the academic performance of the students using only influencing factors by Multi Layer Perceptron algorithm using two datasets of varying size. Both the datasets comprises of all academic, personal and economic factors of the students. The study proves the attributes

DOI 10.17148/IJARCCE.2016.5217

chosen from the original dataset are really high influence using MLP irrespective of the datasets. This study paper helps the institution to know the academic status of the students in advance and can concentrate on weak students to improve their academic results. A new hybrid algorithm for better classification and prediction using the high influence attributes would be the future work.

REFERENCES

[1] Jai Ruby & K. David, "A study model on the impact of various indicators in the performance of students in higher education", IJRET International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, May-2014, pp.750-755.
[2] Baker R.S.J.D., & Yacef K, 2009,' "The state of educational data mining in 2009: A review and future vision", Journal of Educational Data Mining, I, 3-17.
[3] Monika Goyal & Rajan Vohra, "Applications of Data Mining in Higher Education" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012, pp.130-120.
[4] Mohd Maqsood Ali, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, April- 2013, pg. 374-383
[5] Alaa M El-Hales, "Mining Educational Data to Analyze Learning Behaviour A case study", 2009
[6] Rowley, J., "Is higher education ready for knowledge management?", International Journal of Educational Management, 2000, vol. 14(7), pp. 325–333.
[7] Lubega, J. T., Omona, W., & Weide, T. V. D., "Knowledge management technologies and higher education processes_: approach to integration for performance improvement", International Journal of Computing and ICT Research, 2011, vol. 5(Special Issue), pp. 55–68.
[8] J. Shana, T. Venkatachalam, "Identifying Key Performance Indicators and Predicting the Result from Student Data." International Journal of Computer Applications (0975 – 8887) Vol.25-No.9 July 2011.
[9] Brijesh Kumar Baradwaj, Saurabh Pal," Mining Educational Data to Analyze Students'Performance" IJACSA, Vol.2, No.6, 2011
[10] El-Hales-A.(2008),"Mining Students Data to Analyze Learning Behavior: A Case Study", The 2008 International Arab Conference of Information Technology(ACIT2008)- Conference Proceedings, University of Sfax, Tunisia,Dec 15-18.
[11] Kovacic Z. J., "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science & IT Education Conference 2010.
[12] Shanmuga Priya. K, & Senthil Kumar A.V., "Improving the Student's Performance Using Educational Data Mining, 2013.
[13] Ramaswami M., & Bhaskaran R., "CHAID Based Performance Prediction Model in Educational Data Mining", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, 2010.
[14] Bengio Y., Buhmann J. M., Embrechts M., & Zurada J. M., "Introduction to the special issue on neural networks for data mining and knowledge discovery," IEEE Trans. Neural Networks, vol. 11, pp. 545-549, 2000.
[15] Vasile P. B., "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment". Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, IEEE, (2007).
[16] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, and H. Y. Seong, "Prediction NDUM student's academic performance using data mining techniques," presented at the International Conference on Computer and Electrical Engineering, 2009.
[17] Jai Ruby & K. David, "Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study ", IJRASET International Journal for Research in Applied Science & Engineering Technology, Volume 2 Issue XI, November 2014
[18] Romero, C. and Ventura, S. (2007) 'Educational data mining: A Survey from 1995 to 2005', Expert Systems with Applications (33), pp. 135-146.
[19] Kuyoro 'shade o, Nnicolae Goga, Oludele Awodele and Samuel Okolie "Optimal algorithm for predicting students' academic performance" International journal of computers & technology volume 4 no. 1, jan-feb, 2013 Issn 2277-3061
[20] Jai Ruby & K. David, "Analysis of Influencing Factors in Predicting Students Performance Using MLP - A Comparative Study ",International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 2, February 2015
[21] Anne F. Maben, 2005, "Chi-square test adapted from Statistics for the Social Sciences".

DOI 10.17148/IJARCCE.2016.5217