INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X

Impact factor: 4.295

(Volume 4, Issue 3)

Available online at: www.ijariit.com

A review on tracking of student performance using decision tree

Anjali Sharma
anjalisharma.aso2@gmail.com
Centre for Computers and Communication Technology, South Sikkim, Sikkim

Nigita Pradhan
nigitapradhan123@gmail.com
Centre for Computers and Communication Technology, South Sikkim, Sikkim

Sneha Gupta
nehag1079@gmail.com
Centre for Computers and Communication Technology, South Sikkim, Sikkim

Ong Tshering Lepcha
Lepcha191@gmail.com
Centre for Computers and Communication Technology, South Sikkim, Sikkim

Arvind Lal
arvindlal121@gmail.com
Centre for Computers and Communication Technology, South Sikkim, Sikkim

ABSTRACT

The main objective of this paper is an attempt to use data mining methodologies to study and track the student's academic performance in the subject, is to help in enhancing the educational institutions by evaluating and classifying student data to study the main attributes that may affect the student performance in the subject. This paper focused on improving student academic performance based on their semester marks, class assignments, and extra curriculum activity. Tracking students' performance will help the learner to know about their performance and it gives a chance to improve their performance in future. The dataset used for the tracking students 'academic performance include semester marks, class assignments, extra curriculum activity. This paper is mostly focused on a C4.5 algorithm to track the student performance.

Keywords: *Data mining, C4.5 algorithms, decision tree, WEKA (Waikato Environment for Knowledge Analysis) tool.*

1. INTRODUCTION

Data mining is the process of extracting useful information from a large amount of data. To mine the unknown data, different techniques were used such as the Supervised and unsupervised learning technique, pattern mining, clustering, classification technique, prediction, Association rule etc. Data mining is also called Knowledge Discovery in databases, in the field of determining useful information from huge amounts of data.[10] In this paper, decision tree techniques are used. The decision tree is the model that consists of the root node, branch and leaf node. The root node is the top most nodes in the tree structure, each internal node specifies the test on attributes, the class label is held by the leaf node, and the branch node is used to hold the test results. A decision tree is easy and fast method since it does not require any domain knowledge. [7] Under Decision tree there are some algorithm i.e. Iterative Dichotomiser 3 (ID3), C4.5 (Successor of ID3), Classification & Regression Tree (CART), and CHI-squared Automatic Interaction Detector (CHAID) in which C4.5 algorithm is used to track the student performance by generating a decision tree which is to be used for classification and it is an extension of ID3 algorithm. [10]

Currently, the data mining techniques have been used in educational environments. Applying data mining technique in Educational filed is called as educational data mining. Educational Data Mining is reconnoitering discipline, fretful with developing methods for determining the unique types of data that come from an educational setting and using those methods to better understand students, and setting which the student learn in. [8] Data mining is an important data analysis methodology that has been successfully employed in many domains, with numerous applications in educational problems. The Educational Data Mining was defined as "The process of converting raw data from educational systems to useful information that can be used to inform design decisions and answer research questions. Educational data mining is concerned with developing, researching, and applying computerized methods

to detect patterns in large collections of educational data that would otherwise be hard or impossible to analyze due to the huge volume of data within which they exist. There are many popular methods are there in Educational Data Mining. Some of them are widely used as a prediction, classification, clustering, and regression. Prediction attempts to form patterns that permit it to predict performance or learning outcomes based on the data from the available data. The most often used techniques for this type of goal are classification, clustering and association. In Educational Data mining, the prediction has been used for tracking student performance and for identifying student behaviors. [9]

2. REVIEW ON RELATED WORK DONE ON DIFFERENT TYPES OF DATA MIMING METHODOLOGIES USED IN TRACKING STUDENT PERFORMANCE

[1] Elia Georgiana Petre, "A Decision Tree for Weather Prediction", Buletinul UniversităŃii Petrol-Gaze din Ploieşti Vol. LXI No. 1, 2009.

In this paper analyze meteorological data registered during the last years in the capital of China and have tried to forecast the future temperature values in Hong Kong. In order to have a detailed outline of the weather parameters, the paper used the data between 2002 and 2005. The data used to create database include year, month, average pressure, relative humidity, clouds quantity, precipitations and average temperature. Author predict, with a certain accurate, the average temperature for a future month. For that decision tree built with CART algorithm implemented software in data mining – Weka is used.

[2] Brijesh Kumar Baradwaj, Saurabh Pal,"Mining Educational Data to Analyze Students Performance, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

In this paper, the classification task is used on student database to predict the student's division on the basis of the previous database. As there are many methods that are used for data classification, the decision tree method is used here. Information's like Attendance, Class test, Seminar and Assignment marks were collected from the student's previous database, to predict the performance at the end of the semester.

[3] Edin Osmanbegović , Mirza Suljić ,"Data mining approach for predicting student performance", Journal of Economics and Business, Vol. X, Issue 1, May 2012.

In this paper, three supervised data mining algorithms were applied i.e. Naïve Bayes algorithm, C4.5 algorithm and J48 algorithm on the preoperative assessment data to predict success in a course (either passed or failed) and the performance of the learning methods were evaluated based on their predictive accuracy. The results indicate that a good classifier model has to be both accurate and 0": left comprehensible for professors. This used to help students and teacher to improve student's performance and reduce failing ratio by taking the appropriate action.

[4] Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal3, Ashish Kumar Sen, "A Heart Disease Prediction Model using Decision Tree", IOSR Journal of Computer Engineering (IOSR-JCE) 2278-8727Volume 12, Issue 6 2013.

In this paper, author develop a heart disease prediction model that can help medical professionals in predicting heart disease status based on the clinical data of patients. Firstly, author select 14 important clinical features, i.e., age, sex, chest pain type, treetops, cholesterol, fasting blood sugar, resting ECG, max heart rate, exercise-induced angina, old peak, slope, number of vessels colored, that and diagnosis of heart disease. Secondly, author develops a prediction model using J48 decision tree for classifying heart disease based on these clinical features against unpruned, pruned and pruned with reduced error pruning approach. Finally, the accuracy of Pruned J48 Decision Tree with Reduced Error Pruning Approach is better than the simple Pruned and Unpruned approach. The result obtained that which shows that fasting blood sugar is the most important attribute which gives better classification against the other attributes.

[5] Mukesh Kumari, Dr. Rajan Vohra, Anshul arora,"Prediction of Diabetes Using Bayesian Network", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4), 5174-5178, 2014.

In this paper author aim at the discovery of a decision tree model for the diagnosis of diabetes. Pre-processing is used to improve the quality of data. The techniques of pre-processing applied are attributes identification and selection, data normalization, and numerical discretization. Next, the classifier is applied to the modified dataset to construct the Bayesian model. Finally, weka will be used to do simulation, and the accuracy of the model is calculated and compared with other algorithms efficiency. Classification with Bayesian network shows the best accuracy, 99.51 percent and error in the classification is .48 percent when the results were compared to clinical diagnosis. The mean absolute error (MEA) =.0053 and root mean squared error (MRES =.0596). The total time required to build the model is also a crucial parameter in comparing the classification algorithm.

[6] Thaddeus Matundura Ogwoka, Wilson Cheruiyot, George Okeyo, "A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms", International Journal of Computer Applications Technology and Research Volume 4– Issue 9, 693 - 697, ISSN: 2319–8656, 2015.

In this paper author predicting students' academic performance using Decision tree and k-means algorithms. First evaluated decision tree and k-means algorithms in terms of their operations using WEKA free software tool and other written literature. Then applied decision tree and k-means algorithms and created a model of predicting students' academic performance and analyzed 173 undergraduate students of Technical University of Mombasa's Computing and information technology department using first semester results to predict second Semester results. And finally tested the model of predicting students' academic performance and realized an accuracy of 98.8439% at an execution time of 20 milliseconds.

[7] K. Rajalakshmi, Dr. S. S. Dhenakaran ,"Analysis of Data mining Prediction Techniques in Healthcare Management System", International Journal of Advanced Research in Computer Science and Software Engineering Volume 5, Issue 4, 2015.

In this paper, authors analyze the various data mining application in the healthcare domain to discover a new range of pattern information. Firstly, predicted Diabetics using Chi-square algorithm of the Decision tree (SPSS), the C4.5 algorithm of the Decision tree and SMO algorithm of Support Vector Machine (SVM). SMO algorithm has the highest accuracy is 94.3%. Secondly, predict Eye disease using Decision tree/ neutral network technique, Back propagation algorithm has accuracy is 92%. Thirdly predict urinary system disease using learning algorithm of neutral network data mining technique has accuracy is 99%. Fourthly, predict Lung cancer using Naïve Bayes algorithm of classification technique has accuracy is 84.14%. Fifthly, predict Breast cancer using C4.5 algorithm using decision tree technique and process the dataset in WEKA tool has accuracy is 86.7%. Sixthly, predict Parkinson's disease using Regression tree technique has accuracy is 93.75%. Lastly, predict Heart disease using Naïve Bayes, Laplace smoothly, K-nearest neighbor's algorithm of classification. K-nearest neighbor's algorithm has the highest accuracy is 98.24%.

[8] Shaleena K. P, Shaiju Paul, "Data Mining Techniques for Predicting Student Performance", IEEE International Conference on Engineering and Technology (ICETECH), 20th March 2015, Coimbatore, TN, India, 2015.

In this paper, data mining technique for predicting the performance of a student is been discussed. Several white box classification methods like decision trees and rule induction algorithms are been discussed. The problem of imbalanced data is solved by data rebalancing followed by cost-sensitive classification. First prepared a sample dataset of student and then they apply the data mining techniques. Preprocessing methods such as data cleaning, the transformation of variables, and data partitioning have been applied. Data mining algorithms are applied to predict student failure like a classification problem. Then rebalanced data using SMOTE (Synthetic Minority Over-sampling Technique) that is available in Weka as a supervised data filter. Then applied Cost-sensitive classification to rebalanced data set. Any classification algorithm can be made cost sensitive by using the Meta classification algorithm Cost-Sensitive Classifier and setting its base classifier as the desired algorithm using Weka tool. The performance of classifiers can be evaluated using a confusion matrix. This matrix gives information about the actual and predicted classifications. In this study, the students' marks along with social, economic, and cultural attributes are used.

[9] R. Sumitha, E.S. Vinothkumar, "Prediction of Students Outcome Using Data Mining Techniques", International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-6, 2016.

In this paper, the model focuses on analyzing the prediction accuracy of the student's performance. The dataset that comprises of all academic and personal factors of the students. The results of the data mining algorithms for the classification of the students based on the attributes selected reveals that the prediction rates are not uniform among the algorithms. The range of prediction varies from (80-98%). Thereby by comparative analysis of classification algorithms (such as Naïve Bayes, Multilayer Perception (MLP), SMO, Decision Table, REP tree, J48) using WEKA tool, it is proven that the attributes chosen from the original dataset have high influence using J48 with an accuracy of 97% under analysis and used for predicting test data set for future outcome as best, good, average or poor.

[10] Ankita A Nichat, Dr.Anjali B Raut, "Predicting and Analysis of Student Performance Using Decision Tree Technique", International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 5, Issue 4, 2017.

In this paper, authors focus on analysis student academic performance by using the advantage of data mining techniques model. Data mining is utilized to analyses course evaluation questionnaires.

Here, the most important variables that separate "satisfactory" and "not satisfactory" student performances and their weakness' in particular subject or field. Firstly, collect data from student database and stored in different tables was joined in a single table after joining process errors were removed. Then data were selected which were required for data mining. Then used the decision trees generated by C4.5 can be used for classification. And compare ID3 and C4.5 algorithms accuracy C4.5 has the highest accuracy than ID3 and also compare execution time for ID3 and C4.5 algorithms ID3 has highest execution time.

Table 1: Inferences Drawn From Literature Survey

Author Name and publication	Title of the paper	Techniques used	Limitation
Petre, Elia Georgiana UniversităŃii Petrol–Gaze din Ploieşti (2009).	A Decision Tree for Weather Prediction	Decision tree, CART algorithm.	The prediction can not be used in winter due to wind speed, wind direction or radiation. Can be enlarged database with records from other years not only from 2002 to 2005 in future.
Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining educational data to analyze students' performance." arXiv preprint arXiv: 1201.3417 (2012).	Mining Educational Data to Analyze Students‟ Performance	Decision tree method used, the ID3 algorithm used	This paper can not identify those students which needed special attention to reduce fail ratio and to take appropriate action for the next semester examination.
Osmanbegović, Edin, and Mirza Suljić. "Data mining approach for predicting student performance." Economic Review 10, no. 1 (2012): 3-12.	Data mining approach for predicting student performance	Two supervised data mining algorithms: Naive Bayes algorithm and Multilayer perceptron algorithm	Research has not used other data mining algorithms such as Neutral algorithm, K-mean, J48 algorithm and different software such as SQL server data tools (SSDT) to get a broader approach, and more valuable and accurate outputs.
Pandey, Atul Kumar, Prabhat Pandey, and K. L. Jaiswal. "A heart disease prediction model using a decision tree." IUP Journal of Computer Sciences 7, no. 3 (2013): 43.	A Heart Disease Prediction Model using Decision Tree	J48 decision tree	The J48 algorithm is short-off giving an accurate value for that some other classification algorithm such as CART and C4.5 can be used to get more accurate value.
Kumari, Mukesh, Rajan Vohra, and Anshul Arora. "Prediction of diabetes using the Bayesian network." (2014).	Prediction of Diabetes Using Bayesian Network	Classification algorithm i.e. Bayesian network, Weka (Waikato Environment for Knowledge Analysis).	The fuzzy set method can be introduced to improve Bayes Network to do prediction. Also, in order to find the best prediction model, other machine learning methods such as Neural Network can be tested to compare the predicting results.
Shaleena, K. P., and Shaiju Paul. "Data mining techniques for predicting student performance." In Engineering and Technology (ICETECH), 2015 IEEE International Conference on, pp. 1-3. IEEE, 2015.	Data Mining Techniques for Predicting Student Performance	Decision trees and rule induction algorithms.	Has not explained the predictions in a higher level for that it can be explained in the form of IF-THEN rules.
Rajalakshmi, K., and Dr. SS Dhenakaran. "Analysis of Datamining Prediction Techniques in Healthcare Management System." International Journal of Advanced Research in Computer Science and Software Engineering 5, no. 4 (2015): 1343-1347.	Analysis of Datamining Prediction Techniques in Healthcare Management System	Decision tree algorithm, Support Vector Machine (SVM) classification technique, Naïve Bayes, Regression tree.	Rather using so many algorithms, researchers can use only one data mining technique to analysis the health care diagnosis system.

George Okeyo, Wilson Cheruiyot, and Thaddeus Matundura Ogwoka. "A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms". International Journal of Computer Applications Technology and Research Volume 4–Issue 9, 693 - 697, 2015, ISSN: 2319–8656.	A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms	Decision tree and K-means data mining algorithms	WEKA does not update automatically on test dataset predicted as is the case on training dataset, hence to view the results will save in future.
Sumitha, R., and E. S. Vinothkumar. "Prediction of Students Outcome Using Data Mining Techniques." International Journal of Scientific Engineering and Applied Science (IJSEAS) 2, no. 6 (2016): 8.	Prediction of Students Outcome Using Data Mining Techniques	J48 algorithm and K-mean clustering algorithm. And WEKA(Waikato Environment for Knowledge Analysis)	The paper can not analysis student extra-curricular skills and provide suggestions on communication and technical skill development by which students can be built in professional aspect of talents.
Ankita A Nichat and Dr.Anjali B Raut. "Predicting and Analysis of Student Performance Using Decision Tree Technique". International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2017.	Predicting and Analysis of Student Performance Using Decision Tree Technique	C4.5 and ID3 algorithms	Researchers have not given any attention to analysis of other performance such as extra curriculum, technical skill and also has not used other data mining techniques and software.

3. COMPARATIVE STUDY

After studying different papers on data mining, the comparative study is displayed on the basis of information collected form these papers.

The Accuracy and Execution time of J48 is highest than other algorithms. J48 and C4.5 are reliable than other algorithms.

Table 2: Comparison Table of Different Data Mining Algorithm

Metrics	CART algorithm	ID3 Algorithm	C4.5 Algorithm	Naïve Bayes	k-means Algorithm	J48 Algorithm
Accuracy	Less	Less accurate than C4.5	High accurate than ID3 and Naïve Bayes	Medium	High	High
Reliability	Partially reliable	Partially reliable	reliable	Partially reliable	Reliable	Reliable
Execution time	medium	Highest execution time than C4.5	Less than ID3	medium	Highest	Highest

4. CONCLUSION

This research is an initial attempt to use data mining methods to track and evaluate student academic performance and to enhance the quality of the institute. The review of this study indicates that Data Mining Techniques (DMT) capabilities provided effective improving tools for student performance. On reviewing different researcher's paper we found that data mining can be used in higher education in particularly to track the final performance of the student. After reviewing the different author's paper we came to know that there is a number of methods and techniques like C4.5, J48, CART and Naïve Bayes are available to track the student performance accurately by increasing the more number of parameter.

5. REFERENCES

[1] Petre, Elia Georgiana. "A decision tree for weather prediction." Buletinul UniversităŃii Petrol-Gaze din Ploieşti (2009).
[2] Baradwaj, Brijesh Kumar, and Saurabh Pal. "Mining educational data to analyze students' performance." arXiv preprint arXiv: 1201.3417 (2012).
[3] Osmanbegović, Edin, and Mirza Suljić. "Data mining approach for predicting student performance." Economic Review 10, no. 1 (2012): 3-12.
[4] Pandey, Atul Kumar, Prabhat Pandey, and K. L. Jaiswal. "A heart disease prediction model using a decision tree." IUP Journal of Computer Sciences 7, no. 3 (2013): 43.
[5] Kumari, Mukesh, Rajan Vohra, and Anshul Arora. "Prediction of diabetes using the Bayesian network." (2014).
[6] Shaleena, K. P., and Shaiju Paul. "Data mining techniques for predicting student performance." In Engineering and Technology (ICETECH), 2015 IEEE International Conference on, pp. 1-3. IEEE, 2015.
[7] Rajalakshmi, K., and Dr. SS Dhenakaran. "Analysis of Datamining Prediction Techniques in Healthcare Management System." International Journal of Advanced Research in Computer Science and Software Engineering 5, no. 4 (2015): 1343-1347.
[8] George Okeyo, Wilson Cheruiyot, and Thaddeus Matundura Ogwoka. "A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms". International Journal of Computer Applications Technology and Research Volume 4– Issue 9, 693 - 697, 2015, ISSN: 2319–8656.
[9] Sumitha, R., and E. S. Vinothkumar. "Prediction of Students Outcome Using Data Mining Techniques." International Journal of Scientific Engineering and Applied Science (IJSEAS) 2, no. 6 (2016): 8.
[10] Ankita A Nichat and Dr.Anjali B Raut. "Predicting and Analysis of Student Performance Using Decision Tree Technique". International Journal of Innovative Research in Computer and Communication Engineering Vol. 5, Issue 4, April 2017.

