A LITERATURE REVIEW FROM 2011 TO 2014 ON STUDENT’S ACADEMIC PERFORMANCE PREDICTION AND ANALYSIS USING DECISION TREE ALGORITHM

Hardeep Kaur
Assistant Professor
UniverseGroup of Colleges, Adampur, Patiala
hardeep28bawa@gmail.com

Abstract—Success of any educational institute depends upon the success of the students of institute. Student’s performance prediction and its analysis are essential for improvement in various attributes of students like final grades, attendance etc. This prediction helps teachers in identification of weak students and to improve their scores. Various data mining techniques like classification, clustering, are used to perform analysis. In this paper implementation of various decision tree algorithms ID3, J48/C4.5, random tree, Multilayer Perception, Rule Based and random forest have been studied for student’s performance prediction and analysis. The WEKA tool is used to perform evaluation. To evaluate the performance percentage split method or cross validation method is used. Main objective behind this analysis is to improve student’s performance. This review paper explores the use of various decision tree algorithms for student’s academic performance prediction and its analysis.

Keywords—EDM, Decision tree, J48, random tree, ID3, Multilayer Perception, CART, IBI.

I. INTRODUCTION

A. Data Mining and Educational Data Mining (EDM)

Data mining is the process of extracting useful patterns and information from large amount of data. Data mining is the logical process that is used to find useful data from huge amount of data. Main aim of data mining is to find out the patterns that will be helpful for decision making in future. Data mining is the pain part of Knowledge Discovery in Database (KDD).[1]

Educational Data Mining (EDM) is a process which is concerned with developing various techniques or methods for extracting the different types of data that come from educational settings, and use of those methods for better understanding of students. The main area of EDM is analyzing student’s performance. [2]

B. Student Performance Prediction and Analysis

Success of any educational institute depends upon the success of the students of institute. Student’s performance prediction and its analysis are essential for improvement in various attributes of students like final marks, seminar performance, knowledge and attendance etc. This prediction helps teachers in identification of weak students and to improve their scores. During student performance prediction, we predict some unknown values like student end semester marks, behavior, internal marks etc. Student’s performance prediction and analysis is a very popular application of EDM. Different data mining techniques like clustering, classification, regression are used to perform analysis. Various models decision trees, rule based systems, neural networks, Bayesian networks etc. are used for predicting student’s performance and its analysis. This is helpful in predicting student’s performance i.e. prediction about their final grade, prediction about their learning using student’s data. [2][3]

This paper is organized as follows: In section II we present work related to student performance prediction and analysis. In section III we present comparative study of survey. Conclusion is presented in section IV. In section V we discuss future scope.

II. RELATED WORK

Considering the improvements required in students grades or scores, literature survey has been surveyed based on student performance prediction and analysis using decision tree algorithms.

R. R. Kabra, R. S. Bichkar [11] (Dec. 2011) collected data from S.G.R. college of engineering and management, Maharashtra. They collected data from 346 students of engineering first year. Evaluation was performed using J48 algorithm by 10 fold cross validation. The accuracy of J48 algorithm was 60.46%. This model is successful in identifying the students who are likely to fail. So it will be helpful for increasing performance of students.

Surjeet Kumar Yadav, BrijeshBharadwaj, Saurabh Pal [15] (Dec. 2011) collected data of 48 MCA students from VBS Purvanchal University. They have used ID3, C4.5 and CART algorithms for their implementation. They used 10 fold cross...
validation for implementation. After implementation it has been observed that CART algorithm has the highest accuracy of 56.25% than other algorithms. CART algorithm has the highest TPR for class first. From the implementation it is clear that CART algorithm is best in performance than other algorithm. It will help in improvement of student’s performance.

Diego Garcia Saiz, Marta Zorrilla [16] (2011) collected students data from University of Cantabria. They created two different datasets each with 194 records. They have used J48, Rtree, JRip, OneR and NB classification models and proposed Meta algorithm. When they performed experiment using first data set then JRip algorithm has the highest 81.95% accuracy and J48 has the TPR of 95.62. When they performed experiment using second data set J48 has the highest accuracy of 87.11% and TPR was high (98.83) for JRip algorithm. They compared all these results with proposed Meta algorithm. It has been observed after comparison that proposed algorithm gives better results than other algorithms.

Brijesh Kumar Baradwaj, Saurabh Pal [5] (2011) have discussed that students performance is examined by internal marks and final results. Data set of 50 students was used in this study which was taken from MCA department of VBS Purvanchal University, Uttar Pradesh. Information like previous semester marks, attendance, and assignment and class test marks from previous database of students. They have used decision tree algorithms for student performance prediction and analysis. This overall study will help faculty members in improving student’s scores for future examinations.

Surjeet Kumar Yadav, Saurabh Pal [6] (2012) conducted analysis on 90 students of engineering department (session 2010) from VBS Purvanchal University, Uttar Pradesh. ID3, C4.5 and CART decision tree algorithms were used for evaluation. Evaluation was performed using 10 fold cross validation method. It has been found that C4.5 has higher accuracy 67.7778% than ID3 and CART algorithm. Model’s True Positive rate for class Fail is high 0.786 for ID3 and C4.5 which means it will successfully identify the fail students. This study will be helpful for those students that need special attention from teachers.

Manpreet Singh BhuJar, Amritpal Kaur [10] (2012) have taken data set of 1892 students from various colleges for student performance prediction and evaluation. J48 algorithm was chosen for evaluation using 10 fold cross validation. Success rate of J48 algorithm was 77.74%. In this way it will be helpful in identifying weak students so that teachers can help them before failure.

Mrinal Pandey, Vivek Kumar Sharma [4] (Jan. 2013) compared J48, Simple Cart, Reptree and NB tree algorithms for predicting performance of engineering students. They have taken data of 524 students for 10 fold cross validation and 178 students for percentage split method. It has been found that J48 decision tree algorithm achieved higher accuracy 80.15% using 10 fold cross validation method. By using percentage split method higher accuracy 82.58% is achieved by J48 algorithm. From this comparison it has been found that J48 performs best than other algorithms in both the cases. J48 decision tree algorithm will be useful for teachers in improving performance of weak students.

Kuyoro Shade O, NicoleaGoga, Dr. OludeleAwodele, Dr. Samuel Okolie [17] (Jan-Feb 2013) performed analysis on 1500 records of computer science students taken from Babcock University Nigeria. They compared random tree, random forest, JRip, J48, OneR, decision stump, ZeroR, REPTree, PART and MLP classification models. To perform evaluation 10 folds cross validation and holdout test option was used. After evaluation it has been found that random tree algorithm has the highest accuracy 96.07% using 10 fold cross validation. Using hold out method again random tree has the highest accuracy of 85.69%. Random tree has the highest TPR of 0.961 and lowest FPR of 0.013 than all other models while using 10 folds cross validation. Random tree has again the highest TPR of 0.857 and lowest FPR of 0.043 than other classification models while using holdout method. After this comparison it has been observed that random tree is the best amongst other algorithms in predicting performance of Babcock University students.

Anju Rathee, Robin prakashmathur [18] (March-April 2013) conducted analysis on student’s data for predicting their performance in final exams using ID3, C4.5, CART decision tree algorithms. They have used 10 fold cross validation method to perform analysis. They compared these three algorithms using student data and it has been observed that C4.5 has the highest accuracy of 67.7778% as compared to other algorithms. C4.5 and ID3 has the highest 0.786 TPR for class Fail which means this model will successfully find out those students who are likely to fail. From experiment results it is clear that C4.5 is the best algorithm in predicting those students who are likely to fail in exams.

T.MirandaLakshmi, A.Martin, R.Mumtaj Begum, Dr.V.PrasannaVenakatesan [19] (May 2013) compared C4.5, ID3 and CART decision tree algorithms using qualitative data of 120 students. They have used 10 folds cross validation test option for their implementation. After implementation it was observed that CART algorithm has the highest accuracy of 55.83% than other algorithms means it has more correct classification of instances than other methods. TPR of CART algorithm is highest 0.914 for class A which means it is successful in identification of students who has highest marks. It is clear from the results that CART is the best in terms of accuracy and TPR. This implementation assumes that the student’s academic performance and other activities are influenced by qualitative factors.

Anuja Priyam, Abhijeet, Rahul Gupta, Anju Rathee, and SaurabhSrivastava [12] (June 2013) compared ID3, C4.5 and CART decision tree algorithms on the basis of students data. Evaluation was performed using 10 fold cross validation method. It shows that the CART algorithm has higher accuracy 56.2500%. Model’s True Positive rate for class Fail is high 0.786 for ID3 and C4.5 which means it will successfully identify the fail students. So this model will help teachers in reducing failure rates.

Ramanathan L, SakshamDhanda, Suresh Kumar D [14] (June-July 2013) performed analysis on 50 students data. They
were used naïve bayes, J48 and proposed algorithm (Weighted ID3) for evaluation. It shows that WID3 has higher accuracy 93% than J48 and naïve bayes. In future you can make user friendly software using WID3 which will be very helpful for teachers.

Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao [7] (September 2013) performed analysis on data set of 182 students using ID3 and C4.5 decision tree algorithms. When they performed bulk evaluation on data set of 173 students both algorithms have same accuracy of 75.145% and when they performed singular evaluation on data set of 9 students then both algorithms have accuracy 77.778%. For 182 students accuracy was approximately 75.257.

Jaimin N. Undavia, Dr. Atul Patel, Dr. P. M. Dolia [20] (September 2013) conducted analysis on data of 128 students of MBA and MCA using J48, simple CART and Random tree decision tree algorithms. Under test option 10 folds cross validation was used for implementation. They have compared performance of these algorithms using student’s data. After implementation it has been observed that J48 and simple CART performs best in terms of accuracy and both have higher accuracy of 68.75% then random tree algorithm. J48 takes less time in building a model than CART. This model will work well in predicting the students who has highest grade. So J48 is the best in performance than other algorithms.

Dorina Kabakchieva [21] (2013) performed analysis on 10330 students data with 20 attributes. J48, KNN, naïve bayes, OneR and JRip classification models were used. Under test option 10 folds cross validation and percentage split methods were used. J48 algorithm has the highest classification accuracy of 67% amongst all other algorithms. From all algorithms J48 algorithm has the highest TPR. Using 10 fold cross validation in J48 TPR is highest 0.84 for class Bad and when using percentage split then also class bad has highest TPR of 0.84. It means this model is successful in predicting those students whose performance is bad.

Mrs. M. S. Mythili, Dr. A. R. Mohamed Shanavas [9] ( Jan. 2014) compared J48, Random Forest, Multilayer Perception, IBI and decision tree algorithms using data set of 260 students from various schools. 10 fold cross validation was chosen for evaluation. It has been found that Random Forest has higher accuracy 89.23% and less execution time amongst all other algorithms. This study will be helpful for educational institutions.

Jyoti Namdeo, Naveenkumar Jayakumar [13] (Feb. 2014) collected 51 students data from MBA 2007 batch. Decision tree algorithms used in evaluation were Naïve Bayes, Multilayer Perception, J48 and Random Forest. These algorithms were trained on 2007 batch data and tested on 2008 batch data. Evaluation was performed using training, cross validation, percentage split and test on 2008 data. After testing on 2008 data it has been found that naïve bayes has higher accuracy 31.57% amongst other algorithms but this accuracy is not according to requirement.

Azwa Abdul Aziz, Nor Hafieza Ismail and Fadhilah Ahmad [8] (September 2014) conducted analysis on 399 records of students using naïve bayes, rule based and J48 decision tree algorithm. They have used cross validation and percentage split method for evaluation. In cross validation 3, 5, 10 fold cross validation was performed and in percentage split method training: testing 10:90, 20:80, 30:70, 40:60, 50:50, 40:60, 30:70, 20:80, 10:90 percentage split were used. After comparison of 3 classification algorithms it has been found that rule based and J48 decision tree algorithm has higher accuracy 68.8%.

Mrinal Pandey, S. Taruna [22] (September 2014) performed analysis on 1000 students data with 11 attributes and under test options they have selected 10 folds cross validation option for their experiment. They have used J48 (MLCM) decision tree, MLP, IBK and NBTree algorithms for prediction and analysis of student’s final grades. After implementation they found that DT (MLCM) has the highest accuracy of 99.79% and each class accuracy class A- 100%, class B- 99.82%, class C- 98.67% and class F- 100% accuracy. So it has been observed that this model accurately predict the students who has the highest score and who are likely to fail. From all the four classification models decision tree J48 (MLCM) performs best in terms of efficiency, TPR and FPR. It is assumed more suitable for predicting grades of students.

### III. COMPARATIVE STUDY OF SURVEY

<table>
<thead>
<tr>
<th>Paper Name</th>
<th>Year of Publication</th>
<th>Size of Data Set (No. of students)</th>
<th>Algorithms Used</th>
<th>Test Options Used</th>
<th>Algorithm with Higher Accuracy</th>
<th>Accuracy (in %) of Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance Prediction of Engineering Students using Decision Trees</td>
<td>Dec. 2011</td>
<td>346</td>
<td>J48</td>
<td>Cross Validation</td>
<td>J48</td>
<td>60.46%</td>
</tr>
<tr>
<td>Data Mining Applications: A comparative Study for Predicting Student’s performance</td>
<td>Dec. 2011</td>
<td>48</td>
<td>ID3 C4.5 CART</td>
<td>Cross Validation</td>
<td>CART</td>
<td>56.25%</td>
</tr>
<tr>
<td>A promising classification method for predicting distance students’ performance</td>
<td>2011</td>
<td>194</td>
<td>J48 Rtree JRip OneR NB Proposed algorithm</td>
<td>---</td>
<td>Proposed algorithm</td>
<td>85.05%</td>
</tr>
<tr>
<td>Title</td>
<td>Year</td>
<td>Total</td>
<td>Algorithm(s)</td>
<td>Technique</td>
<td>Proposed Algorithm</td>
<td>Accuracy</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>------</td>
<td>-------</td>
<td>--------------------------------------------------</td>
<td>-----------</td>
<td>--------------------</td>
<td>----------</td>
</tr>
<tr>
<td>Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification</td>
<td>2012</td>
<td>90</td>
<td>ID3, C4.5, CART</td>
<td>Cross Validation</td>
<td>C4.5</td>
<td>67.7778%</td>
</tr>
<tr>
<td>Use of Data Mining in Education Sector</td>
<td>2012</td>
<td>1892</td>
<td>J48</td>
<td>Cross Validation</td>
<td>J48</td>
<td>77.74%</td>
</tr>
<tr>
<td>A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction</td>
<td>Jan. 2013</td>
<td>524</td>
<td>J48 Simple cart Reptree NB tree</td>
<td>Cross Validation</td>
<td>J48</td>
<td>80.15%</td>
</tr>
<tr>
<td></td>
<td>178</td>
<td></td>
<td>J48 Simple cart Reptree NB tree</td>
<td>Percentage Split</td>
<td>J48</td>
<td>82.58%</td>
</tr>
<tr>
<td>Optimal algorithm for predicting students academic performance</td>
<td>Jan.-Feb. 2013</td>
<td>1500</td>
<td>JRip J48 OneR decision stump ZeroR REPTree PART MLP</td>
<td>Cross Validation</td>
<td>Random tree, Holdout Random tree</td>
<td>96.07% 85.69%</td>
</tr>
<tr>
<td>Survey on Decision Tree Classification algorithms for the Evaluation of Student Performance</td>
<td>March-April 2013</td>
<td></td>
<td>ID3, C4.5, CART</td>
<td>Cross Validation</td>
<td>C4.5</td>
<td>67.7778%</td>
</tr>
<tr>
<td>An Analysis on Performance of Decision Tree Algorithms using Student’s Qualitative Data</td>
<td>May 2013</td>
<td>120</td>
<td>C4.5 ID3 CART</td>
<td>Cross Validation</td>
<td>CART</td>
<td>55.83%</td>
</tr>
<tr>
<td>Comparative Analysis of Decision Tree Classification Algorithms</td>
<td>June 2013</td>
<td></td>
<td>ID3 C4.5 CART</td>
<td>Cross Validation</td>
<td>CART</td>
<td>56.2500%</td>
</tr>
<tr>
<td>Comparison of Classification Algorithms to Predict Students’ Post Graduation Course in WEKA Environment</td>
<td>September 2013</td>
<td>128</td>
<td>J48 Simple CART Random tree</td>
<td>Cross Validation</td>
<td>J48 CART</td>
<td>68.75%</td>
</tr>
<tr>
<td>Predicting Students Performance using ID3 and C4.5 Classification Algorithms</td>
<td>September 2013</td>
<td>173</td>
<td>ID3 C4.5 for bulk evaluation</td>
<td>Cross Validation</td>
<td>ID3 C4.5</td>
<td>75.145%</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td></td>
<td>ID3 C4.5 for singular evaluation</td>
<td>Cross Validation</td>
<td>ID3 C4.5</td>
<td>77.778%</td>
</tr>
<tr>
<td>Predicting Student Performance by Using Data Mining Methods for Classification</td>
<td>2013</td>
<td>10330</td>
<td>J48 KNNe Naive bayes, OneR JRip</td>
<td>Cross Validation</td>
<td>Percentage Split Test</td>
<td>J48</td>
</tr>
<tr>
<td>An Analysis of students’ performance using classification algorithms</td>
<td>Jan. 2014</td>
<td>260</td>
<td>J48 Random Forest Multilayer Perception IBI</td>
<td>Cross Validation</td>
<td>Random Forest</td>
<td>89.23%</td>
</tr>
<tr>
<td>Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts</td>
<td>Feb. 2014</td>
<td>51</td>
<td>J48 Random Forest Multilayer Perception Naive Bayes</td>
<td>Cross Validation</td>
<td>Percentage Split Test</td>
<td>Naive Bayes</td>
</tr>
<tr>
<td>First Semester Computer Science Students’ Academic Performances Analysis by Using Data Mining</td>
<td>September 2014</td>
<td>399</td>
<td>Naive Bayes J48 Rule Based</td>
<td>Cross Validation</td>
<td>Percentage Split</td>
<td>J48 68.8%</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Classification Algorithms</th>
<th>September 2014</th>
<th>99.79%</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Multi Level Classification Model Pertaining to the Student’s Academic Performance Prediction</td>
<td>1000</td>
<td>J48 (MLCM)</td>
</tr>
<tr>
<td>J48 (MLCM) decision tree</td>
<td>MLP</td>
<td>Cross Validation</td>
</tr>
</tbody>
</table>

IV. CONCLUSION

Educational data mining’s (EDM) importance is increasing day by day as the student’s performance prediction and analysis requirements are increasing for improvement of student’s academic performance. As given above various authors have implemented different decision tree algorithms: J48, random forest, random tree, ID3, multi layer perception, naive bayes, rule based, IBI, REPTree, NB tree and CART using different data sets. Some authors performed comparison of algorithms to find out the best algorithm from them on the basis of accuracy. The survey done in this paper shows that most probably J48/C4.5 decision tree algorithm is considered best algorithm in terms of accuracy for different data sets. So it is clear from survey that J48 performs well for any size of data set. This is the reason behind wide use of J48 algorithm amongst all decision tree algorithms.

Survey done in the section II will be helpful to various researchers that are working in the field of student’s performance prediction and analysis using decision tree algorithms.

V. FUTURE WORK

For growth of any educational institute, student’s academic performance is main contributor. If students perform well academically then institution growth rate goes high. It is necessary in these days to focus on the student’s results so there is a wide scope in this field. To increase student’s performance, student performance prediction and analysis is used. For this purpose decision tree algorithms are used mainly. Various researchers have done lot of research in this field by performing evaluation using single algorithm or by comparing three or four algorithms.

In future researchers can enhance the research by comparing large number of algorithms using large size data sets. So there is a wide scope for researchers in this field.

ACKNOWLEDGMENT

First of all I express my sincerest debt of gratitude to the Almighty God who always supports me in my endeavors.

I would like to thank Prof. Neena Madan for their encouragement and support. Then, I would like to thank my family and my friends. I am thankful to all those who helped me in one way or the other at every stage of my work.

REFERENCES


