© 2015 IJEDR | Volume 3, Issue 4 | ISSN: 2321-9939

EDM: An analysis of learner's academics performance based on Frequent Pattern Tree Algorithm

Karan Sukhija
Research Scholar,
Panjab University, Chandigarh, India.

Abstract **- The education system performance of school education in India is a turning point in the academic lives of all learners. As this academic performance is influenced by many factors, it is essential to develop predictive data mining model for learners to determine factors that influence the learner's performance. Educational data mining is used to analyse the data available in the educational field and elicit the hidden knowledge from it. In this study, a survey cum experimental methodology is implemented to generate a database and it was constructed from school education department. The raw data is pre-processed in terms of filling up missing values, transforming values in one form into another and relevant attribute/ variable selection. As a result, we had 10,000 student examination records, which is use in implementation stage. This paper implement the generalized sequential pattern mining algorithm for finding frequent patterns from learner's database and frequent pattern tree algorithm to build the tree based on frequent patterns. This tree can be used for predicting the learner's performance as pass or fail.**

Index Terms **- Data Mining, Education Data Mining, Knowledge Discovery in Databases (KDD), Frequent Pattern Mining Algorithm, Decision Tree Classifier, C4.5 and C5.0 algorithm.**

I. INTRODUCTION

Data mining has increasing research interests in education field which is termed as Educational Data Mining. This field is concerned with developing methods that discover knowledge from data instigated from various educational environments [1]. EDM applies a variety of techniques viz. Decision Trees, Neural Networks, Naive Bayes, K-Nearest neighbour, and many others to discover hidden patterns.

With the rapid growth of educational data, the main goal of any educational institution is to improve education quality. Prediction of learner's performance in education institution is one way to attain the good quality in education system. Learners' academic performance hinges on diverse factors like personal, socio-economic, psychological and other environmental variables. All these variables are necessitated by prediction models for the effective prediction of the performance of the learners. The prediction of student performance assists in identifies the learners with low academic achievements and that learners can be individually assisted by the educators to improve their performance in future.

A number of data mining models and statistical models have been designed with various factors as inputs that influence the performance of the learners [3] [4]. Some of the models applied in educational environments are discussed as follows: Prediction can be classified into: classification, regression, and density estimation. In classification, the predicted variable is a binary or categorical variable. Some popular classification methods are: decision trees, logistic regression and support vector machines. In regression, the predicted variable is a continuous variable. Some popular regression methods within educational data mining include linear regression, neural networks, and support vector machine regression. Classification techniques like decision trees, Bayesian networks etc. can be used to predict the learner's behaviour in an educational environment.

EDM is the process of transforming raw data compiled by education systems in useful information that could be used to take conversant decisions and answer research questions [2]. The various techniques of data mining like classification, clustering and rule mining can be applied to bring out various hidden knowledge from the educational data. The core objective of this paper is to study the learners' performance in the school examination system by applying the frequent pattern mining algorithm [2]. Classification technique is applied to evaluate learners' performance along with the decision tree approach.

II. RELATED WORK

Z.N. Khan, et al [7] conducted a performance study on 400 learners comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream.

Al. Radaideh, et al [8] applied a decision tree model to predict the final grade of learners who studied the C++ course in Yarmouk University, Jordan in the year 2005. Three different classification methods namely ID3, C4.5, and the Naïve Bayes were used. The outcome of their results indicated that Decision Tree model had better prediction than other models.

Pandey, et al [9] conducted study on the student performance based by selecting 600 learners from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes Classification on category, language and background qualification, it was found that whether new comer learners will performer or not.

S.Anupama, et al [10]M.N concluded that Decision rule and One R rule algorithms can be used to predict the result of the fifth semester of student in higher education based on the marks obtained by the learners in the previous four semesters. Rule based algorithm can provide efficiency in predicting the learner's performance in higher education using the previous historical data.

Kin Fun Li, et al [11]concluded that in Canada there was a failure rate of more than 30% after the first two years in the Faculty of Engineering. Different data mining techniques such as clustering and classification approaches e.g. K-means and hierarchical clustering, and K-nearest neighbour and naïve Bayes classifiers can be used to predict the failure rate of learners.

Bhardwaj, et al [12] conducted study on the student performance based by selecting 300 learners from 5 different degree college conducting BCA course of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayesian classification method on 17 attributes, it was found that the factors like learners' grade in senior secondary exam, living location, medium of teaching, mother's qualification, learners other habit, family annual income and learner's family status were highly correlated with the student academic performance.

III. DATA COLLECTION AND PRE-PROCESSING

The data set consists of 10,000 records of metric students for the session 2013-14 which has been obtained from school education department of different states (* name is hidden subject to confidentiality constraint) based on the sampling method. According to the stages of data mining, the pre-processing of dataset has performed after data collection. Initially only required fields are taken into consideration for mining process. Further, some derived variables are also considered and some of the information for the variables has been extracted from the database. Furthermore, so many techniques are applied to remove the anomalies in the dataset viz. handling of missing values, data type conversion etc. Table 1 depicts various variables of data set based on parameters variable name, description and possible values.

Variable Name	Description	Possible Values
Sex	Gender information	Male, Female
Area	Location	Rural, Urban
Eng_Res	English Subject Result	Pass, Fail
Eng_Grade	English Subject Grade	A+,A,B,C,D,E
Math_Res	Math Subject Result	Pass, Fail
Math_Grade	Math Subject Grade	A+,A,B,C,D,E
Punjabi_Res	Punjabi Subject Result	Pass, Fail
Punjabi_Grade	Punjabi Subject Grade	A+,A,B,C,D,E
Hindi_Res	Hindi Subject Result	Pass, Fail
Hindi_Grade	Hindi Subject Grade	A+,A,B,C,D,E
Sci_Res	Science Subject Result	Pass, Fail
Sci_Grade	Science Subject Grade	A+,A,B,C,D,E
SS_Res	SS Subject Result	Pass, Fail
SS_Grade	SS Subject Grade	A+,A,B,C,D,E
Final Result	Overall Result	Pass, Fail, Reappear
Final_Grade	Overall Grade	A+,A,B,C,D,E

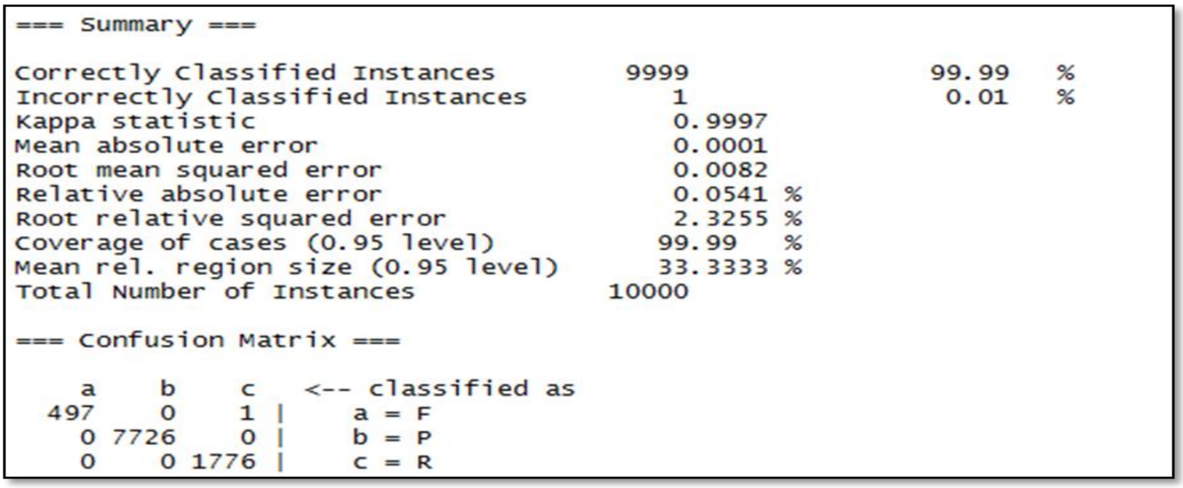
Table -1 Dataset Variable Description

IV. IMPLEMENTATION OF EDM MODEL

R-language the powerful statistical open source software is used for implementation of different types of algorithm and statistical analysis used in this stage. Here, it is used for mining algorithm specifically on educational data. The above mentioned dataset is prepared in .csv format which is actually required [13] for implementation stage. There are so many decision tree algorithms like ID3, J48, C4.5, C5.0 etc. are available for classification of dataset. This paper implement the generalized sequential pattern mining algorithm for finding frequent patterns [14] from learner's database and further frequent pattern tree algorithm is applied to build the tree based on frequent patterns. A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a final decision [15] of learner's performance as pass or fail. During this research, C4.5 and C5.0 decision tree classification algorithm is implemented on frequent dataset and decision tree is build according to classification segment which is used for analysis in future.

Implementation of C4.5

The popular decision tree classifier algorithm C4.5 is implemented in R language on dataset which contain records of student examination. The above mentioned Table 1 described the detail of each fields. The following summary shows the statistics of implementation stage. It described all types of error rate and confusion matrix which is based upon implementation [16].



```
=== Summary ===

Correctly Classified Instances        9999               99.99   %
Incorrectly Classified Instances         1                0.01   %
Kappa statistic                          0.9997
Mean absolute error                      0.0001
Root mean squared error                  0.0082
Relative absolute error                  0.0541 %
Root relative squared error              2.3255 %
Coverage of cases (0.95 level)          99.99   %
Mean rel. region size (0.95 level)      33.3333 %
Total Number of Instances            10000

=== Confusion Matrix ===

    a    b    c   <-- classified as
  497    0    1 |    a = F
    0 7726    0 |    b = P
    0    0 1776 |    c = R
```

Fig.1 C4.5 algorithm implementation

Decision tree based on C4.5

The above mentioned summary described the statistics which is based on implementation of C4.5 algorithm. The C4.5 is a decision tree classifier algorithm that provides the output in the form of tree which represents each attribute as nodes. The root node depicts the initial level and leaf nodes depict the decision nodes. The following tree depicts the nodes based on above said education dataset.

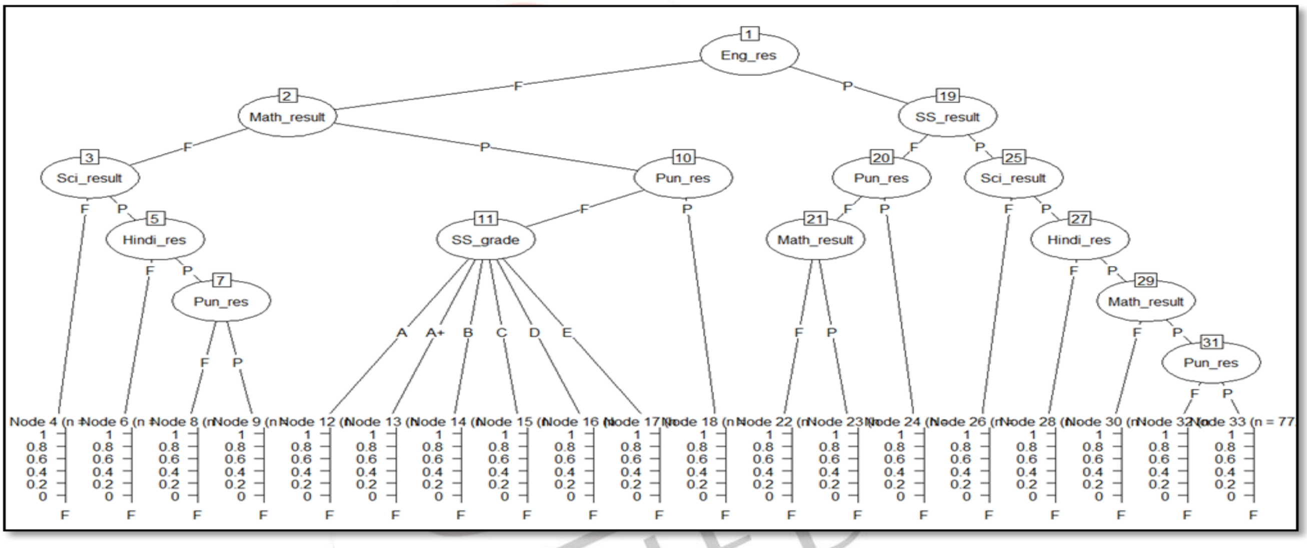
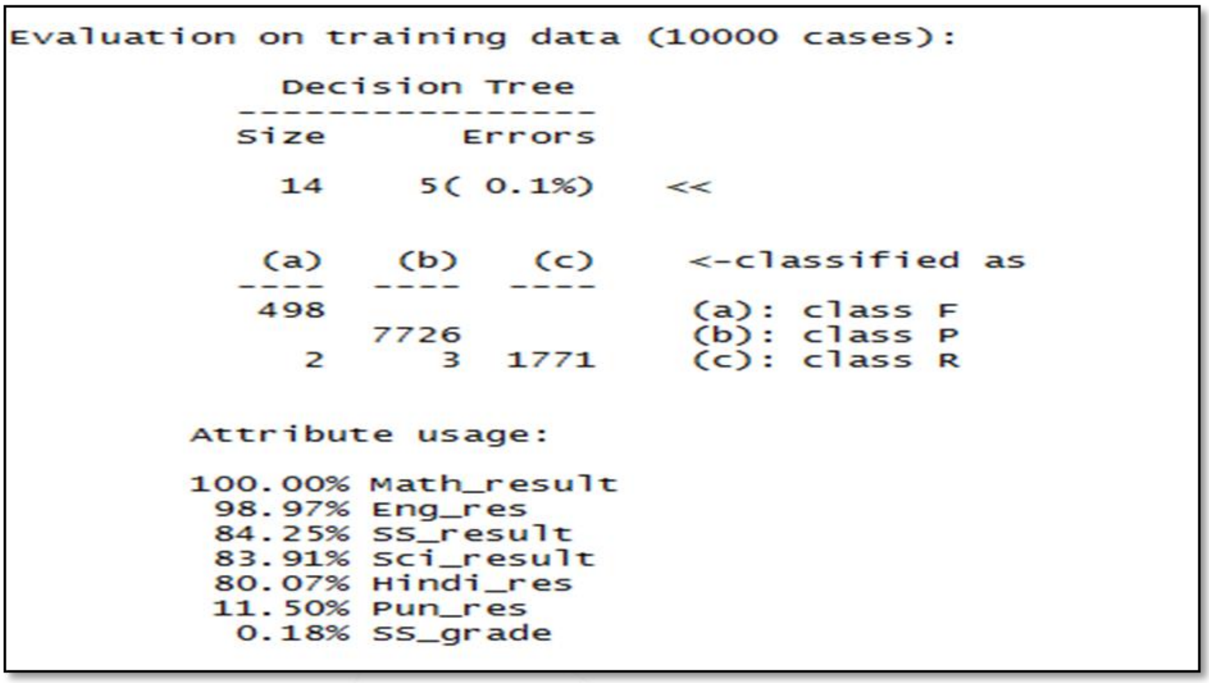


Fig.2 C4.5 algorithm decision tree

Implementation of C5.0

The extension of C4.5 decision tree classifier i.e. C5.0 is also implemented in R language on same dataset to evaluate the performance and also make comparison between both. The above mentioned Table-1 described the detail of each fields. The following summary shows the evaluation on training datasetin implementation stage. It described error rate in form of percentage and confusion matrix according to decision variable. Finally it also described the attribute usage in terms of percentage.



```
Evaluation on training data (10000 cases):

            Decision Tree
          ----------------
          Size      Errors

            14     5( 0.1%)   <<


           (a)   (b)   (c)    <-classified as
          ----  ----  ----
           498                (a): class F
                7726          (b): class P
             2     3  1771    (c): class R


        Attribute usage:

        100.00% Math_result
         98.97% Eng_res
         84.25% SS_result
         83.91% Sci_result
         80.07% Hindi_res
         11.50% Pun_res
          0.18% SS_grade
```

Fig.3 C5.0 algorithm implementation

Decision tree based on C5.0

The above mentioned summary described the statistics which are based on implementation of C5.0 algorithm. The C5.0 is prune the tree in better way as compare to C4.5. The root node shows the Initial level and leaf nodes described the decision nodes. The following tree depicts the nodes based on above said education dataset.

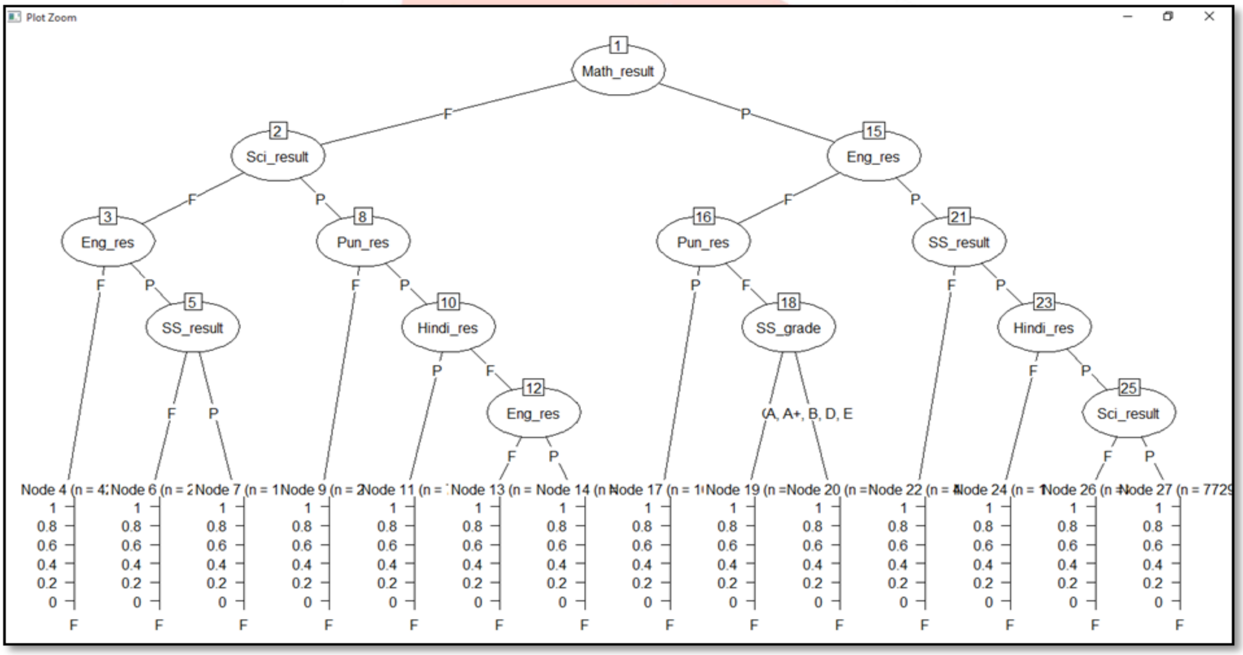


Fig.4 C5.0 algorithm decision tree

V. RESULTS AND DISCUSSION

Based on aforementioned implementation, it is concluded that C5.0 gives better result as compared to C4.5 both in terms of tree representation and performance. The algorithm C4.5 provides the detailed representation of nodes that depicts the outcome in elaborated manner. Moreover, the output represented by both decision tree algorithm C4.5 and C5.0 can be extracted and represented in the form of association rules. It is very easy to understand the representation of decision tree in the form of association rules in terms of predicate and consequent where one classification rule can be generated for each path from each terminal node to root node.

VI. CONCLUSION

The various data mining techniques can be effectively implemented on educational data. From the above results it is clear that classification techniques can be applied on educational data for predicting the learner's outcome and improve their results. This prediction will help the tutors to identify the weak learners and help them to score better marks. The C4.5 and C5.0 decision tree algorithm has applied to learner's examination data to predict their performance in the final exam. The outcome of the decision tree predicts the number of learners who are likely to fail or pass. The result given to the tutor aids in step forward to improve the performance of the learners who has been predicted to fail. Since the application of data mining contributes advantageous for school education system, these techniques can be applied in the other areas of education to optimize the resources, to predict the retainment of faculties in the institution and find out the solution of how to reduce student dropout rate in school education system.

VII. REFERENCES

[1] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[2] Cecily Heiner, Ryan Baker y KalinaYacef, -Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems Jhongli, Taiwan.,2006.

[3] Z. N. Khan, "Scholastic Achievement of Higher Secondary Learners in Science Stream", Journal of Social Sciences, Vol. 1, No. 2, 2005, pp. 84-87.

[4] S. T. Hijazi, and R. S. M. M. Naqvi, "Factors Affecting Learner's Performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1, 2006.

[5] Y. Ma, B. Liu, C.K. Wong, P.S. Yu, and S.M. Lee, "Targeting the Right Learners Using Data Mining", Proceedings of KDD, International Conference on Knowledge discovery and Data Mining, Boston, USA, 2000, pp. 457-464.

[6] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Prediction of Learner's Performance in Distance Learning Using Machine Learning Techniques", Applied Artificial Intelligence, Vol. 18, No. 5, 2004, pp. 411-426.

[7] Z. N. Khan, "Scholastic achievement of higher secondary learners in science stream", Journal of Social Sciences, Vol. 1, No. 2, pp. 84-87, 2005.

[8] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar, "Mining student data using decision trees", International Arab Conference on Information Technology(ACIT'2006), Yarmouk University, Jordan, 2006.

[9] U. K. Pandey, and S. Pal, ―Data Mining: A prediction of performeror underperformer using classification, (IJCSIT) InternationalJournal of Computer Science and Information Technology, Vol. 2(2),pp.686-690, 2011.

[10] S.Anupama Kumar, Vijayalakshmi M.N., ―Mining of Student Academic Evaluation Records in Higher Education‖, International Conference on Recent Advances in Computingand Software Systems (RACSS), IEEE conference publication, pp 67–70, Apr,2012.

[11] Kin Fun Li, D. Rusk, F. Song,―Predicting Student Academic Performance, Seventh International Conference on Complex, Intelligent, and Software Intensive Systems, IEEE conference publication, pp 27-33, July 2013.

[12] B.K. Bharadwaj and S. Pal. ―Data Mining: A prediction forperformance improvement using classification‖, International Journalof Computer Science and Information Security (IJCSIS), Vol- 9, No.4, pp. 136-140, 2011.

[13] Ramaswami, M., and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining." arXiv preprint arXiv:1002.1144 (2010).

[14] Kumar, S. Anupama, and M. N. Vijayalakshmi, "Efficiency of decision trees in predicting learner's academic performance." First International Conference on Computer Science, Engineering and Applications, CS and IT. Vol. 2. 2011.

[15] Baradwaj,Brijesh Kumar, and Saurabh Pal, "Mining educational data to analyze learners' performance." arXiv preprint arXiv:1201.3417 (2012).

[16] M.Z. Ashrafi, D. Taniar, and K.A. Smith: Redundant association rules reduction techniques, *International Journal of Business Intelligence and Data Mining*, 2(1): 29- 63 (2007)

[17] P. Williams, C. Soares, and J.E. Gilbert: A Clustering Rule Based Approach for Classification Problems. *International Journal of Data Warehousing and Mining* 8(1): 1-23 (2012)

[18] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, pp. 27-34, 1996.