ANALYSIS OF DECISION TREE ALGORITHMS BASED ON STUDENT RECORD FACTS

Dr.V.Shanmukha Rao, G.Chakradhar, K.Mohith and K.Phani
*Professor and B.Tech students, Department of Information Technology, Andhra Loyola Institute of Engineering and Technology, Vijayawada - 520008, Andhra Pradesh

***Abstract*—** Data Mining means analysis of data from large relational databases and forming related knowledge from it. The main goal of the data mining analysis is to extracting the valuable information from the relational databases. To analyze the different decision tree algorithms used in data mining, a record set of student academic facts are considered. The classifications of decision tree algorithms like ID3, C4.5, and C5.0. Etc., are taken and the applied to the collected facts of the student from the educational institution. This analysis will helps to the institution for their improvement of the student performances. The performance classification analysis of the decision tree algorithms are generated using Rapid Miner software tool. Based on the data set selected and applied for the algorithms, the effectiveness of the results will be depends.

Keywords **—Relational database, data mining, ID3, C4.5, C5.0, Rapid Minner.**

I. INTRODUCTION

Classification is a data mining technique widely used for the prediction purposes. Mainly, the classification algorithm involves two steps. One step is learning and another one is Classification. In Learning use the classification algorithms to analyze the training dataset and construct the decision tree. In Classification phase, the accuracy of the present classification rules is approximated by taking the test data. Decision tree algorithms are generate rules, which used in the classification of data. By considering these generated rules, apply the test data set. By using these rules get the accuracy of that algorithm. Decision tree algorithms like ID3, C4.5 and C5.0 are used to generate the decision tree rules.

ID3 algorithm invented by Ross Quinlan used to generate the decision tree.ID3 is based on the Entropy (H(S)). Information Gain chooses as selection criteria, the attribute with highest Information Gain IG(S, A) is chosen as the splitting attribute.ID3 can over fit the training data.ID3 is difficult to use on continuous data.

$$H(S) = \sum_{x \in X} -p(x)\log_2 p(x)$$

$$IG(S, A) = H(S) - \sum_{t \epsilon T} p(t)H(t) = H(S) - H(S|A).$$

C4.5 algorithm used to generate a decision tree developed by Ross Quinlan.C4.5 is an extended version of ID3 algorithm C4.5 algorithm is also a statistical classifier. C4.5 algorithm works on same concept of information entropy of ID3 algorithm. C4.5 takes the information gain as the splitting criteria it produces the most effective splits in the decision tree.C4.5 can handles continuous and discrete attributes.

C5.0 algorithm is an updated version of C4.5 algorithm. C5.0 is the classification decision tree algorithm. C5.0 is more efficient than C4.5. C5.0 algorithm works on the principle of C4.5 algorithm. The C5.0 algorithm split the sample subset data based on the information gain. In C5.0 algorithm, best split can be obtained based on the biggest information gained field of sample data.C5.0 algorithm can handles the continuous and discrete attributes of the sample data.

Hence, the student marks record facts are considered in the database and applied the classification decision tree algorithms ID3, C4.5 and C5.0 for predicting the student performances. The performance accuracy of those algorithms are compared and finds the best algorithmic results.

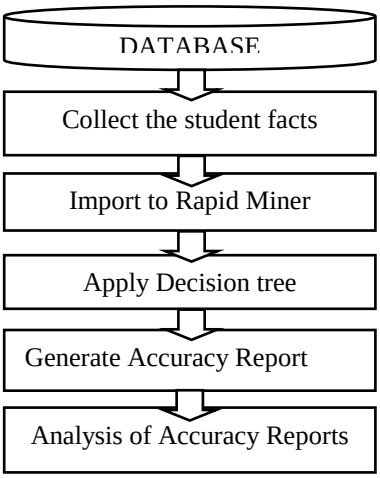
II. LITERATURE REVIEW

Analyzing the student details and factors like Roll No., HSC, UG, Board, Communication, placed (y/n) are considered and rules are generated by using the decision tree algorithms C4.5, ID3 and CHAID. Among these algorithms, the accuracy of the ID3 algorithm is considered as the best algorithm [1].The performances are analyzed by taking the student data consists of student marks in the entrance exam and results in the first year of the previous batch of students and applied on ID3 and C4.5 classification algorithms, the results are same for the algorithms [2]. Decision tree algorithms C4.5 and C5.0 are used to predict the areas which are lagging in placement records of the college and accuracy of C5.0 is higher than the C4.5 [3]. Educational Institutions has lot of student record facts in the relational databases. By considering these student data, analyze the data and get the valuable information from the data. Educational Data mining (EDM) is the procedure of applying the data mining tools and techniques to this educational institution data [4]. EDM is the help a crucial role in the education institution student marks data. By using different techniques we create a set of patterns of data by considering the any one of the attribute in the table of database.

III. METHODOLOGY

Collect the student marks data from the university results pdf. Converting and extracting the pdf data into .csv formats to import into the developed system database. Student data table contains fields are htno, subject name, internal marks, external marks, result. These datasets are considered to the Decision

tree algorithms in the Rapid minner.



3.1 Fig: 1 Data analysis procedure

Steps to Import the extracted data into Rapid minner

1.Open new blank sheet in the rapid minner.
2.Click on add data.
3.Choose the data path which can be local or database.
4.Click confirm.

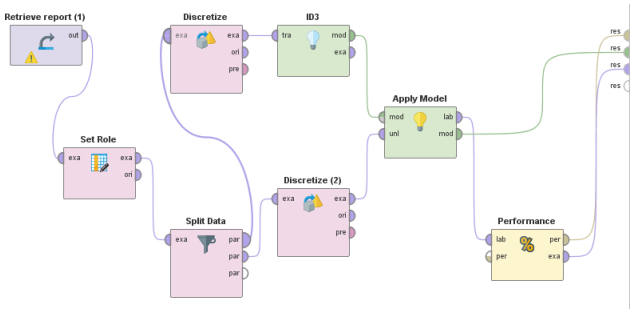
Apply Decision tree algorithms

In the Rapid minner Split the input data into two parts: one is training data set and testing data set. Apply decision tree algorithms to the training data set. Apply the decision tree output to the model along with the testing dataset.

IV. IMPLEMENTATION

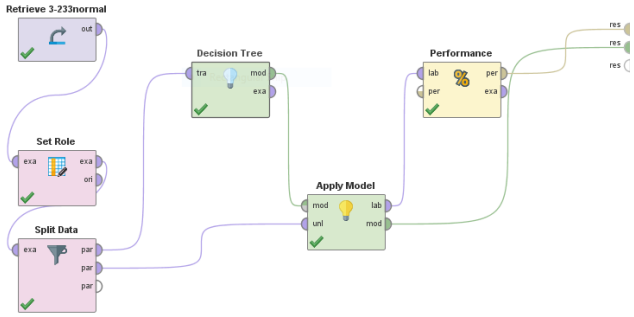
ID3 Implementation in Rapid Minner

In ID3 algorithm is implemented in the Rapid Minner by using the ID3 decision tree module. For input dataset the field result is set as the role. From the Set Role module the dataset is moved to split data module. In Split Data module, split the dataset into two datasets. They are Training Dataset and Testing Dataset .The Training Dataset is passed to the ID3 algorithm module. From the ID3 module the trained data moved to apply model module where the test dataset is applied to it.



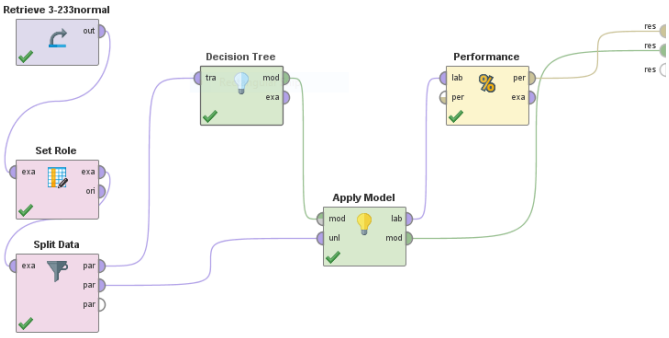
C4.5 Implementation in Rapid Minner

In C4.5 algorithm is implemented in the Rapid Minner by using the Decision tree module. For input dataset the field result is set as the role. From the Set Role module the dataset is moved to split data module. In Split Data module, split the dataset into two datasets. They are. Training Dataset and Testing Dataset .The Training Dataset is passed to the Decision tree algorithm module. From the Decision tree module the trained data moved to apply model module where the test dataset is applied to it.



C5.0 Implementation in Rapid Minner

In C5.0 algorithm is implemented in the Rapid Minner by using the Decision tree module. For input dataset the field result is set as the role. From the Set Role module the dataset is moved to split data module. In Split Data module, split the dataset into two datasets. They are. Training Dataset and Testing Dataset .The Training Dataset is passed to the Decision tree algorithm module. From the Decision tree module the trained data moved to apply model module where the test dataset is applied to it.

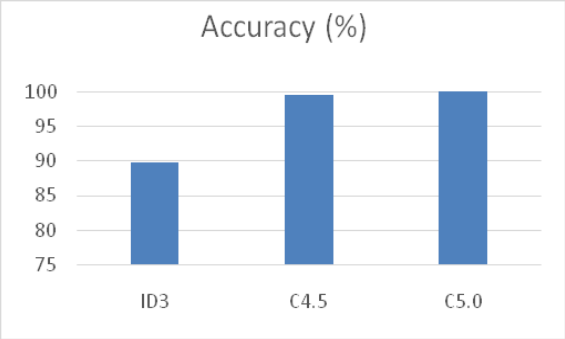


Generate Accuracy report

In rapid minner, the output of the model connected to the performance method. From the performance module, retrieve the accuracy of the decision tree algorithms applied to the given input data.

Analysis of Accuracy report:

After getting the accuracy reports of each Decision tree algorithms ID3,C4.5 and C5.0. Based on the research on analysis of these accuracy reports we can say that the accuracy of the C5.0 algorithm is more accurate than the other decision tree algorithms.



Results

Decision Tree Algorithms	Number of Inputs	Number of correctly predicted	Number of wrongly predicted	Accuracy (%)
ID3	383	343	40	89.75%
C4.5	383	380	3	99.45%
C5.0	383	383	0	100%

Table 4.1 performance comparisons of decision tree algorithms

V. CONCLUSION

Hence, according to the research study has done on the decision tree algorithms for the student marks data of the college. By the accuracy reports of the algorithms ID3, C4.5 and C5.0.Among them C5.0 algorithm has more accuracy when compared to the other algorithms for the given data.

REFERENCES

[1] T. Jeevalatha,N.Ananthi,,D .Sravana Kumar "Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms" , International Journal of Computer Applications (0975 – 8887) Volume 108 – No 15, December 2014

[2] PREDICTING STUDENTS' PERFORMANCE USING ID3 AND C4.5 CLASSIFICATION ALGORITHMS,KalpeshAdhatrao, Aditya Gaykar, AmirajDhawan, RohitJha and VipulHonrao ,Department of Computer Engineering, Fr. C.R.I.T., Navi Mumbai, Maharashtra, India

[3] 3.K.Nasaramma1,M.Bangaru Lakshmi2, D.kiranmayi3,"Prediction and Comparative Analysis of Students Placements Using C4.5 &C5.0", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 8 (3) , 2017,367-368, 2,3 CSE Department, Vignan's Institute of Information Technology.

Dr.V.Shanmukha Rao presently working as a Associate Professor in the Department of Information Technology branch in Andhra Loyola Institute of Engineering and Technology, Vijayawada, affiliated to Jawaharlal Nehru Technological University, India. He has a total of 18 years of rich experience comprising teaching and research. He has published the papers in International and national journals. His current research interests are in the areas of Computer Networks, cloud computing, datamining, Web Mining and Semantic web technologies.