International Organization of Research & Development (IORD)
ISSN: 2348-0831
Vol 11 Issue 01 | 2023

Student Grade Prediction Using C4.5 Decision Tree

V.S.Talole
Assistant Professor, Department of Computer Science & Engineering, Sanmati College of Engineering, Washim, Maharashtra

ABSTRACT

Analyzes data mining methods and techniques students' data to construct a predictive model for students' academic performance. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is also use for sorting the educational problem by using analysis techniques for measuring the student performance, instructor performance. In this paper, measuring student performance using classification technique such as decision tree. The task can be processed based on the several attributes to predict the performance of the student activity respectively. In this research, the paper have been focused the improvement of Prediction/ classification techniques which are used to analyze the skill expertise based on their academic performance by the scope of knowledge. Giving the details about the results, and the specific needs of studies to improvement, such as the accompaniment of students along their learning process, and the taking of timely decisions in order to prevent academic risk and desertion. Lastly, some recommendations and thoughts are laid out for the future development of performance. Helps to analyze the slow leaner that are likely study in poor which are used to improve their skill as early to achieve the goal.

KEYWORDS: Data mining, Classification algorithms, decision trees.

1. INTRODUCTION

Data mining is the analysis step of the "knowledge discovery in databases" process. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyses data from many different dimensions or angles, categorize it, and summarize the relationships identified. In recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational research, an area of inquiry termed educational data mining [3]. An ability of student performance is essential in education environment, which is influenced by many qualitative attributes like Student Identity, gender, age, Specialty, Lower class Grade, higher Class Grade, Extra knowledge or skill, Resource, Attendance, Time spend to study, Class Test Grade (Internal), Seminar Performance, Lab Work, Quiz, E-Exercise, E-Homework, Over all Semester exam Percentage are included for forming the data set. Educational data mining applied many techniques are K- nearestneighbor, decision tree, Naïve Bayes, Neural network, Fuzzy, Genetic and other techniques are applied in the environment [4].

2. RELATED WORK

Mustafa Agaoglu [1] research in educational mining focuses on modeling student's performance instead of instructors' performance. One of the common tools to evaluate instructors' performance is the course evaluation questionnaire to evaluate based on students' perception. In this study, four different classification techniques, –decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis– are used to build classifier models. Their performances are compared over a dataset composed of responses of students to a real course evaluation questionnaire using accuracy, precision, recall, and specificity performance metrics.

Tripti Mishra,Dr. Dharminder Kumar,Dr. Sangeeta Gupta [2] use different classification techniques to build performance prediction model based on students' social integration, academic integration, and various

emotional skills which have not been considered so far. Two algorithms J48 (Implementation of C4.5) and Random Tree have been applied to the records of MCA students of colleges affiliated to Guru Gobind Singh Indraprastha University to predict third semester performance.

Crist´obal Romero [3] Educational data mining (EDM) is an emerging interdisciplinary research area that deals with the development of methods to explore data originating in an educational context. EDM uses computational approaches to analyze educational data in order to study educational questions. This paper surveys the most relevant studies carried out in this field to date.

3. PROPOSED SYSTEM

We are going to propose the system by using which the user can give a test on specific educational or subject categories. When student complete the test, system will calculate the performance of the user by using the algorithm decision tree. The system will suggest to the teacher that on which topics the user is weak or need to study again. To solve the problems faced with manual examination writing, there is need for a computerized system to handle all the works. We propose an application that will provide a working environment that will be flexible and will provide ease of work and will reduce the time for report generation and other paper works. Today many organizations are conducting online examinations worldwide successfully and issue results online but they are not measuring the performance of the student and teacher not know about the weak points of the students and we are focusing on this issue. The main advantage is that it can be evaluation of answers can be fully automated for all questions and other essay type questions can be evaluated manually or through automated system, depending on the nature of the question's and the requirements. To bring, efficiency, transparency and reliability, universities should also adopt this new technology for managing the examination system.

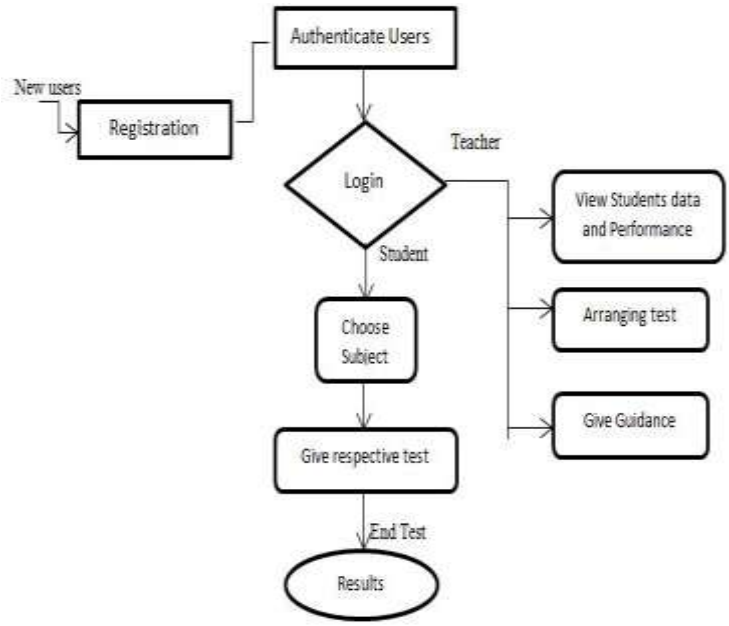

Fig. 1. Data Flow Diagram

4. PROPOSED METHODOLOGY

The major objective of the proposed methodology is to build the classification model that classifies a students' performance. The classifiers, has been built by combining the Standard for Data Mining that includes student data and finally application of data mining techniques which is classification in present study. In other words, using this Decision tree algorithm, we wanted to be able to guide student towards achievement of good score that we felt they would enjoy doing. Tree-based methods classify instances by sorting the instances down the tree from the root to some leaf node, which provides the classification of a particular instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute [5].The benefits of having a decision tree are as follows –

- It does not require any domain knowledge. It is easy to comprehend. The learning and classification steps of a decision tree are simple and fast.
- Decision Tree Induction Algorithm

A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). Later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algorithm, there is no backtracking; the trees are constructed in a top-down recursive divide-and-conquer manner.

4.1 Decision Trees

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labeled with distinct outcomes of the test. Each leaf node has a class label associated with it.

5. DATA MINING TECHNIQUE AND CLASSIFICATION

Data mining is very promising as a new effective technique for decision making processes. Through Educational data mining is an analysis of discipline to developing the methods for exploring the unique types of data from educational settings and it is used for improvement of students in better way [3] .Data mining techniques are applied in higher education more and more to give insights to educational and administrative problems in order to increase the managerial effectiveness. However, most of the educational mining research focuses on modeling student's performance. Data mining technique can give the input for the teachers and students about the student academic results. This technique can analysis the database patterns to forecast student performance, so this allows the teachers to prepare like a remedial program (needing extra help for learning) or more additional assignments for the students.

Classification is one of the most studied problems in machine learning and data mining .It consists in predicting the value of a categorical attribute (the class) based on the values of other attributes (predicting attributes).

5.1 Data Preparations

The data set used in this study was obtained from a student's database which we are created for our application. In this step data stored in different tables was joined in a single table after joining process errors were removed.

5.2 Data selection and transformation

In this step only those fields were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class. None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class. Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value. To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting.

Algorithm: Generate_decision_tree

Step 1- Start
Step 2-Take input which is given by User
I_n= {I_1,……..,I_n}
Step 3- Dataset preparation
D_n = { {I_1,…I_n}D}
Step 4- Dataset elaboration
D_I ={S_1,….S_n, C_1,…..C_n, I_1,…..I_n, a_1,…..a_n}
Step 5- Processing
While(D_n!=0)
{
If (a_n==I_n)
Check C_n, S_n;
}
Step 6- Result Generation

R={ S_c, S_n , C_n};
Where,
I_n = Input given by users
D_n = Dataset
D = Database
D_I = Dataset contents
S_c=Score
a_1,….a_n= Answer
S_1,….S_n= Subject
C_1,….C_n= Category

We also using Generalized Sequential Pattern mining algorithm for predicting the student's performance as pass or fail. Once the student is found at the risk of failure he/she can be provided guidance for performance improvement.

Generation Sequential pattern Mining Algorithm
Step 1-Start
Step 2- Take input from Dataset
Step 3- Processing
While (Cn!=0)
{
Qn={{Q1,…..Qn} Cn}
PData= {count(Qn), So(Qn), Cr(Qn)}
Rc= PData{(count(Qn)-Cr(Qn)) Cn}
}
Step 4- Result Generation
R=Rc;
R will show the weak category of student.
Where,
Qn=Questions(Total)
{(Qn)Cn}= Questions regarding to category.
So= Solved Questions
Cr= Correct questions
Rc= Result Category

It provides weak categories or subjects of students from his/her performance.

6. SIMULATION AND RESULTS:

The simulation studies involve comparison of ID3 and C4.5 accuracy with different data set size, this comparison is presented graphically in Fig.2.

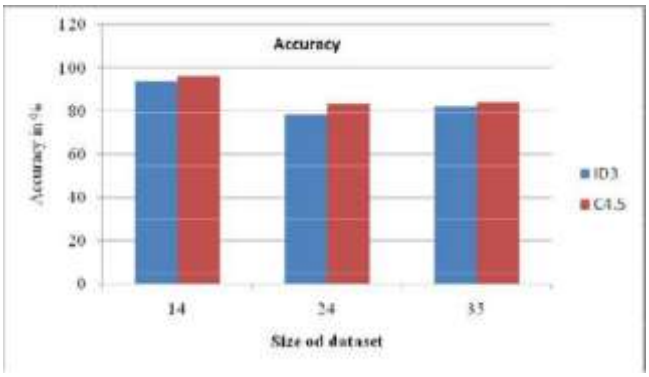

Fig.2. Comparison of Accuracy for ID3 & C4.5 Algorithm

The 2nd parameter compared between ID3 and C4.5 is the execution time which is show in Fig.3.

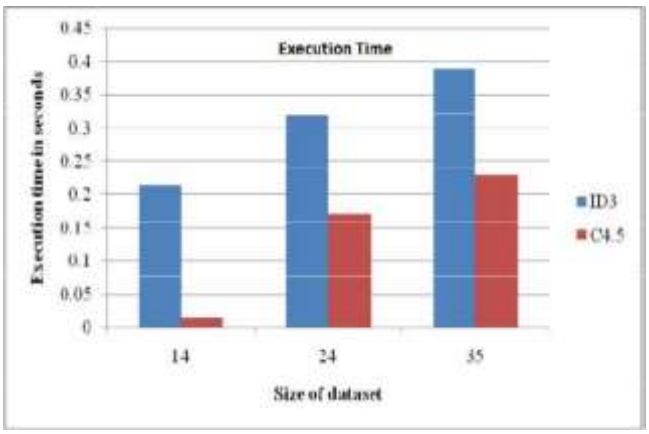

Fig.3. Comparison of Execution Time for ID3 & C4.5 Algorithm

In result student getting the all records of given test is show in Fig. 4. Figure shows the test given by student marks obtains, total marks, date oaf test and provide suggestion to study for improving performance.

Fig.4. Student Record

In Fig.5 shows the student performance record of test given by that particular student which involve subject, total marks, marks obtain, date of conducting test, weak concept shoes the category in which student is weak.

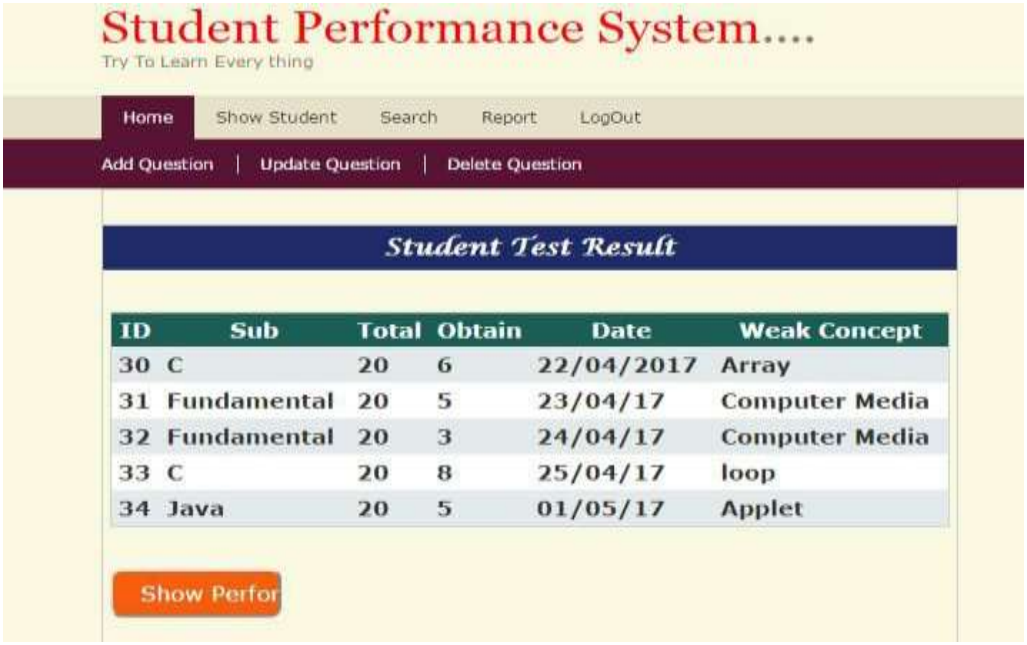

Fig.5. Student performancereport

In Fig.6 indicate performance of student in graphical to teacher.

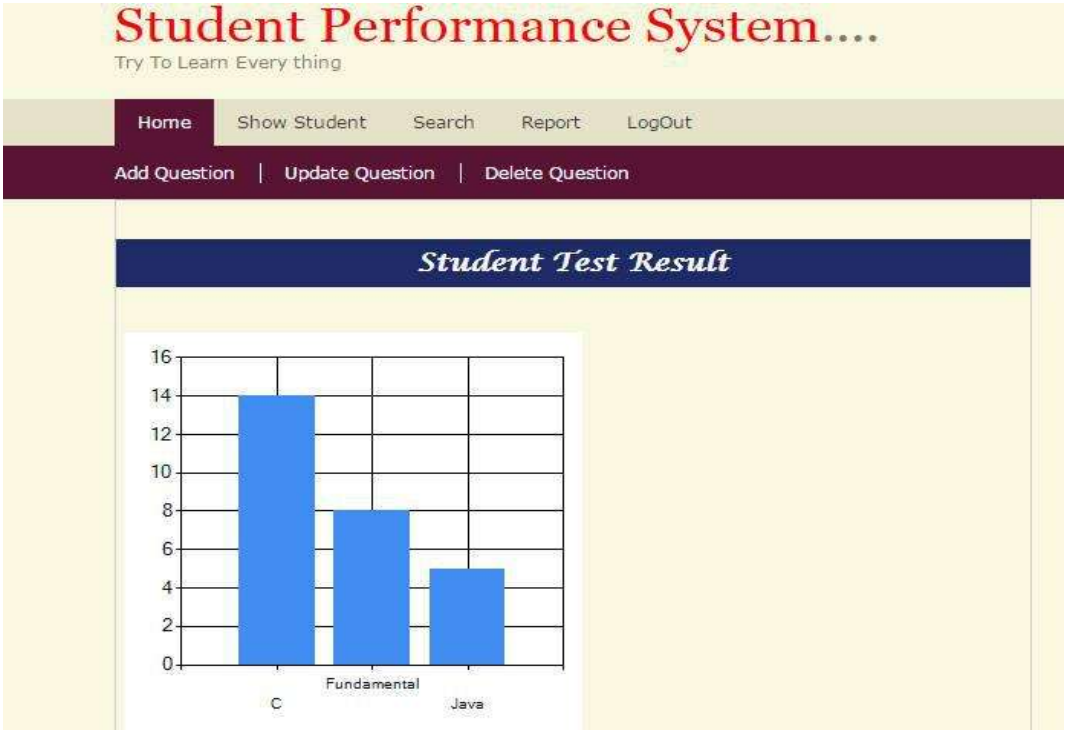

Fig.6. Performance result

7. CONCLUSION

Academic success of students of any professional Institution has become the major issue for the management. An early analysis of students at risk of poor performance helps the management take timely action to improve their performance through extra coaching and counseling. The result of this study indicates that data mining techniques capabilities provided effective improving tools for analysis student performance. In this paper, data mining is utilized to analyses course evaluation questionnaires. Here, the most important variables that separate “satisfactory” and “not satisfactory” student performances and there weakness’ in particular subject or field. Hopefully, these can help instructors to improve their performances. Tree-based methods classify instances by sorting the instances down the tree from the root to some leaf node, which provides the classification of a particular instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute. This paper focuses on analysis student academic performance by using advantage of data mining techniques model.

8. REFERENCES

[1] Higher Education," IEEE Access , Volume: 4 ,2016Tripti Mishra,Dr. Dharminder Kumar,Dr. Sangeeta Gupta,"Mining Students’ Data for Performance Prediction," in fourth International Conference on Advanced Computing & Communication Technologies,2014.
[2] Mustafa Agaoglu, “Predicting Instructor Performance Using Data Mining Techniques in Higher Education," IEEE Access , Volume: 4 ,2016
[3] Tripti Mishra,Dr. Dharminder Kumar,Dr. Sangeeta Gupta,"Mining Students’ Data for Performance Prediction," in fourth International Conference on Advanced Computing & Communication Technologies,2014.
[4] Crist´obal Romero," Educational Data Mining: A Review of the State of the Art," IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 40, No. 6, November 2010.
[5] Carlos Márquez Vera, Cristóbal Romero Morales and Sebastián Ventura Soto,“Predicting of school failure and dropout by using data mining techniques”, The IEEE Journal of Latin-American Learning Technologies (IEEE-RITA) , Vol. 8, No. 1, pp 7-14, Feb 2013.
[6] R.S.J.D Baker and K.Yacef, “The State of Educational Data Mining in 2009: A Review and Future Visions”, Journal of Educational Data Mining, 1, Vol 1, No 1, 2009.