PREDICTION OF STUDENT OUTCOME IN EDUCATIONAL SECTOR BY USING DECISION TREE

M.Pavani¹, A.Ravi Teja², A.Neelima³, G.Bhavishya⁴, D.Sai Sukrutha⁵

Abstract: Data Mining is defined as extracting information from massive sets of data. We can say that data mining is the procedure of mining knowledge from data. Classification is a data mining function that assigns items in a collection to target categories or classes, coming to prediction attempts to form patterns that permit it to predict the next event's given the available input data. Now-a-day's Educational data mining is an emerging discipline. Educational Data Mining field concentrates on Prediction more often as compared to generate exact results for future purpose. Observing the changes occurring in the schedule of day-to-day inquiry. Though it is the biggest educational challenge to improve the condition of student's performance. This title mainly focuses on identifying the student's outcome by using decision tree based classification algorithms. The data set of student academic records is tested and applied various decision tree based classification algorithms such as C4.5. The outcome of statistics is generated based on decision tree based classification algorithms, and comparison of two classifiers is also done to predict the accuracy. This project promotes the importance of prediction and decision tree classification based data mining algorithms in the field of education and also presents some promising future lines.

Key Words: Educational Data Mining, Classification, Prediction, C4.5.

I. INTRODUCTION

 Nowadays Educational Data Mining is an emerging discipline. Educational Data Mining describes a research field concentrated with the application of data mining, machine learning, and statistics Educational Data Mining field mainly focuses on prediction more often as compared to produce exact results for future purpose to information generated from educational settings. Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people's learning activities in educational settings. Quite often, this data is extensive, fine-grained, and precise. For example, several learning management systems (LMSs) track information such as when each student accessed each learning object, how many times they accessed it, and how many minutes the learning object was displayed on the user's computer screen. Educational Data Mining (EDM) describes a research field concerned with the application of data mining, machine learning and statistics to information generated from educational settings.

II. LITERATURE REVIEW

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes. Data Mining can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of student. Mining in educational environment is called Educational Data Mining. Educational data mining is emerging as a research area with a suite of computational and psychological methods and research approaches for understanding how students learn.

[1] Data Mining or Knowledge Discovery is needed to make sense and use of data. Knowledge Discovery in Data is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. [2] [3] The paper depicts the users, components as well as the various approaches in EDM. Educational data mining (EDM) is concerned with developing methods for exploring data from educational settings with the purpose of providing quality education to students. Another study of predicting the performance of students on educational web based system was done in which data mining techniques were applied on the vast amount of data on user patterns collected during the interaction of students with the web based education system. The approach followed is to analyze the logged data of students and then classify students based on the data and predict the final grade of the students, the authors have used multiple classifiers and applied genetic algorithm to improve the accuracy of the prediction. In this paper a strategy to improve the student’s performance is mentioned by mapping the student’s record using classification algorithm and grouping datasets into cluster but there is no future performance prediction. Applied the classification as data mining technique to evaluate student’s performance, they used decision tree method for classification. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising. The reference provides a methodology for Improving Quality of Educational Process by using classification, c4.5 algorithm on the other part decision making process for enhancing the quality of educational activities is mentioned.

[7] The paper provide a prediction of Applying data mining technique to identify whether students’ online learning experiences can be assessed based on their log files but it is limited to the available data in online database while factors such as students’ position in the collaborative group and Structure of the collaborative tasks is not considered. [8] Applied the classification as data mining technique to evaluate student’s performance, they used decision tree method for classification. This study allows the University management to prepare necessary resources for the new enrolled students and indicates at an early stage which type of students will
potentially be enrolled and what areas to concentrate upon in higher education systems for support.

III. PROPOSED WORK

Educational data mining has vast amount of data that has to be organized in a consistent manner. To organize, analyze and classify students details in classification algorithm we are using C4.5 algorithm is been used based on academic records.

Our Architecture of proposed work is as follows.

![Flowchart of proposed work](image)

But it simply specifies the current scenarios whereas no future prediction is available and variables used for analysis are only based on demographic and academic records. Therefore to enhance the existing system the proposed model is designed by collecting Students Personal and Academic data from the senior students of the institution and Thereby Grouping the student’s performance based on certain conditions as

- Poor
- Good
- average

Classification consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. There is a work methodology which governs a series of stages. The methodology starts from the problem definition, then data collection from questionnaire and Students Database. The analysis of C4.5 algorithm is done to predict student’s performance by creation of student model.

C4.5 is a computer program for inducing classification rules in the form of decision trees from a set of given instances. C4.5 is a software extension of the basic ID3 algorithm designed by Quinlan.

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data set is given S = s1, s2, ..., sn of already classified samples. Each sample xi consists of a p-dimensional vector (x1, x2, ..., xp), where the values (pi) are the feature values for attribute ai.

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.

Entropy

In general, if we are given a probability distribution P = (p1, p2, ..., pn) and a sample S then the Information carried by this distribution, also called the entropy of P is given by:

\[
\text{Entropy} (p) = \sum_{i=1}^{n} p_i \times \log p_i
\]

The gain information G (p, T)

We have functions that allow us to measure the degree of mixing of classes for all sample and therefore any position of the tree in construction. It remains to define a function to select the test that must label the current node.

It defines the gain for a test T and a position p

\[
\text{Gain} (p, T) = \text{entropy}(p) - \sum_{j=1}^{n} (p_j \times \text{entropy}(p_j))
\]

Where values (pj) is the set of all possible values for attribute T. We can use this measure to rank attributes and build the decision tree where at each node is located the where values (pj) is the set of all possible values for attribute T. We can use this measure to rank attributes and build the decision tree where at each node is located the attribute with the highest information gain among the attributes not yet considered in the path from the root. A survey cum experimental methodology is used. Through extensive search of the literature and discussion with experts on student performance, a number of factors that are considered to have influence on the performance of a student are identified. These influencing factors are categorized as input variables.

For this work, recent real world data is collected from college. This data is then filtered out using manual techniques. Then student data is finding accurately by using C4.5 algorithm.

This algorithm has few base cases to be checked

1. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for decision tree that saying to choose to that class.
2. 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.
3. Instance of previously-unseen encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

The pseudo code for C4.5 algorithm is as follows:

1. Check for the above base cases.
2. For each attribute a, find the normalized information gain ratio from splitting on a.
3. Let a best be the attribute with the highest normalized information gain.

www.ijtre.com

Copyright 2017. All rights reserved.
4. Create a decision node that splits on a_best.
5. Recur on the sublists obtained by splitting on a_best, and add those nodes as children of node.

IV. CONCLUSION
In this paper, c4.5 algorithm is used to predict the student outcome on the dataset of 250 students. These data sets are used to predict the student outcome (good or average or poor). In this study, a model was developed based on selected students related attributes collected from university data set. By using our model faculty can analyze the student performance as well as student also can view his/her outcome. Educators who don’t have technical knowledge can also understand student performance. New factors like No of projects, No of Papers Published, No of certifications, No of conferences also plays a key role during calculation of student performance.

REFERENCES