EDUCATIONAL DATA MINING AND STUDENT’S PERFORMANCE PREDICTION

V.MADHUBALA ¹, T.JEYA²

¹ Research scholar, department of computer science
Sri adi chunchanagiri women’s college, cumbum.

² Assistant professor, department of computer science sri adi chunchanagiri women’s college, cumbum.

ABSTRACT-

Educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. The performance in higher secondary school education in India is a turning point in the academic lives of all students. It is essential to develop predictive data mining model for student’s performance so as to identify the slow learners and make necessary steps for the improvement of the students. In this paper, a new system that will predict students’ higher secondary grades based on academic and personal details of the students. ID3 decision tree algorithm was used to train the data of the school students sets. The knowledge represented by decision trees were extracted and presented in the form of IF-THEN rules. A set if prediction rules were extracted from id3 decision tree algorithm and the efficiency of the generated model was found.

Keywords- Data mining, decision trees, id3 algorithm, prediction rules, if-then rules.
I. INTRODUCTION

Educational Data Mining (EDM) is a new trend in the data mining and Knowledge Discovery in Databases (KDD) field which focuses in mining useful patterns and discovering useful knowledge from the educational information systems, such as, admissions systems, registration systems, course management systems (moodle, blackboard, etc…), and any other systems dealing with students at different levels of education, from schools, to colleges and universities. Researchers in this field focus on discovering useful knowledge either to help the educational institutes manage their students better, or to help students to manage their education and deliverables better and enhance their performance.

Analysing students’ data and information to classify students, or to create decision trees or association rules, to make better decisions or to enhance student’s performance is an interesting field of research, which mainly focuses on analysing and understanding students’ educational data that indicates their educational performance, and generates specific rules, classifications, and predictions to help students in their future educational performance. Classification is the most familiar and most effective data mining technique used to classify and predict values.

II. DATA MINING PROCESS

In present day educational system, a student’s performance is influenced by psychological and environmental factors. Students should be properly motivated to learn. Motivation leads to interest, interest leads to success. Proper assessment of abilities helps the students to perform better. Students requires proper study atmosphere both at school and home. Poor economic condition also affects the performance of the students as most of them are unable to get proper education. Uneducated family background also affects the students’ performance. In this study consider environmental factors and educational institute factors. This helps the tutor to identify the factors that are related with the three types of learners and take appropriate action to improve their performance.

A. Data Preparations

The data set used in this study was obtained from different colleges on the questionnaire method of Computer Science department of
course B.Sc (IT), B.Sc, (CS) and B.E of session of 2013 to 2015. Initially size of the data is 300. In this step data stored in different tables was joined in a single table after joining process errors were removed. **B. Data Selection and Transformation**

In this step only those field were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables extracted from the database. All the predictor and response variables which were derived from the database. The parameter values for some of the variables have detailed below to give brief explanation about each attributes for the current investigation as follows:

**FI** - to predict student level, Family Income (FI) plays vital role among all the students’, by the help of given property values (i.e., Low, Medium and High).

**ME** - If mothers are educated they can contribute to improve the performance of the students. In this study, ME considered to predict student’s results with the help of selected property values by the students (i.e., Low, Medium and High).

**MW** - how mother education is doing vital role to educate their children, likewise their working status has considered with the name of MW attribute. Because, in a situation a particular student mother doesn’t work, then their mother can spend more time with them. Those data have been organized by the help of specified property values (i.e., Yes or No).

**SH** - Study hours, it represents how many hours a student spends on study after attending the class in school. Again it shows how much serious the student takes studies. The possible values are High, Less, Never.

**RE** - to predict student performance, relation or behaviors of the teacher with the student, which have collected by the name of handling basis (RE: Relation), and given to students to select according to their need. (i.e., casual, strictly and friendly).

**LS** - Learning style, students are following different learning styles. It’s commonly believed that most of the students follow some particular method of interacting with, taking in and processing information. This collected by the help of specified property values (i.e., AL, VL, and TL)

**RESULT** - it’s our main constant which collects and keeps the entire students final
results in separate place to predict student’s performance with the help of allocated property values (i.e., Below Average, Average, Excellent).

C. Decision Trees

Decision tree induction is the learning of decision trees from class- labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node [11].

D. The ID3 Decision Tree

ID3 is a simple decision tree algorithm introduced by Ross Quinlan in 1986 [11]. It is based on Hunts algorithm. The basic idea of ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. The tree is constructed in two phases. The two phases are tree building and pruning. ID3 uses information gain measure to choose the splitting attribute. It accepts only categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise pre-processing technique has to be used.

E. C4.5

C4.5 algorithm is developed by Quinlan Ross that generates the decision trees which can be used for classification problems [11]. It is the successor of ID3 algorithm by dealing with both categorical and continuous attributes to build a decision tree. It is also based on Hunt’s algorithm. To handle the continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. It uses Gain Ratio as an attribute selection measure to build a decision tree. C4.5 removes the biasness of information gain when there are many outcome values of an attribute.

III. LITERATURE SURVEY

Baradwaj and Pal [1] conducted a research on a group of 50 students enrolled in a specific course program across a period of 4 years (2007-2010), with multiple performance indicators, including “Previous Semester Marks”, “Class Test Grades”, “Seminar Performance”, “Assignments”,

V. MADHUBALA And T. JEYA
“General Proficiency”, “Attendance”, “Lab Work”, and “End Semester Marks”. They used ID3 decision tree algorithm to finally construct a decision tree, and if-then rules which will eventually help the instructors as well as the students to better understand and predict students’ performance at the end of the semester. Furthermore, they defined their objective of this study as: “This study will also work to identify those students which needed special attention to reduce failing ration and taking appropriate action for the next semester examination” [1].

Baradwaj and Pal [1] selected ID3 decision tree as their data mining technique to analyze the students’ performance in the selected course program; because it is a “simple” decision tree learning algorithm.

Abeer and Elaraby [2] conducted a similar research that mainly focuses on generating classification rules and predicting students’ performance in a selected course program based on previously recorded students’ behavior and activities.

Abeer and Elaraby [2] processed and analysed previously enrolled students’ data in a specific course program across 6 years (2005–10), with multiple attributes collected from the university database. As a result, this study was able to predict, to a certain extent, the students’ final grades in the selected course program, as well as, “help the student's to improve the student's performance, to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time” [2].

Pandey and Pal [3] conducted a data mining research using Naïve Bayes classification to analyse, classify, and predict students as performers or underperformers. Naïve Bayes classification is a simple probability classification technique, which assumes that all given attributes in a dataset is independent from each other, hence the name “Naïve”.

IV. CONCLUSION

The need of prediction over student performance is to help teachers and parents to concentrating their students and children to improvise their performance as well as researcher to select among the decision tree classifier algorithm to find the best classifier for predicting the student performance. The results show that ME (Mothers Education), SH (Students Study Hour), FI (Family income), FE (Fathers
Education), FI (Family Income), MW (Mother Working Status) and RE (Teachers relationship) more affect the student performance. This survey will also help to identify those students are low performers they needed special attention. Finally C4.5 is discovered as the best algorithm for predicting student performance.

REFERENCES


