JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

PREDICTION OF STUDENTS ACADEMIC PERFORMANCE USING J48 ALGORITHM IN EDUCATIONAL DATA MINING

[1]O. YAMINI,[2] S R AJITHA,[3]Dr. G V RAMESH BABU

[1]Research Scholar,[2] Research Scholar,[3] Research Supervisor

[1] Department of Computer Science,

[1]S V University, Tirupati, A P

Abstract: Data mining is a major advancement in the type of the analytical tools and a multi-disciplinary field of a combination of machine learning, statistics, database technology and AI. This technique includes a number of phases: Business understandings, Data understanding and preparation, Modeling, Pattern Evaluation, and Deployment [1]. Data mining has proven to be very beneficial in the field of Education as it increases accuracy of student's performance prediction, to reduce failure rate of student. Educational data mining is the process of applying the data mining tools and techniques to analyze data at the educational institutions. In this paper, in educational data mining we used data mining classification algorithms like J48 algorithm from decision tree to predict the student's performance basing on previous result and present attitude.

IndexTerms **- Educational Data Mining, Classification, Decision tree, J48.**

I. EDUCATIONAL DATA MINING INTRODUCTION

The availability of the educational data has been growing rapidly day by day in the real world. There is a need to analyze the huge amounts of data generated from educational ecosystem, Educational Data Mining (EDM) field that has emerged. EDM is the process of applying data mining tools and techniques to analyze the data at the educational institutions. Now-a-days EDM is evolving and helping the educational sector to adapt the new teaching techniques for learners. Especially this area of research is gaining popularity because of having potential benefits to the educational field [2]. Educational institutions used educational data mining (EDM) to gain deep and through the knowledge to enhance its assessment, evaluation, planning, and decision-making in its educational programs. EDM will help academic programs to extract the patterns that can be used to predict the student performance and behaviors very easily. [3].

Universities have been using many and different types of data mining techniques to analyze the educational report stored in the educational institutions. Here Decision tree can be easily or simply converted into the classification tree.

II. DECISION TREE

A decision tree is a flow chart like structure, where each node signify test on attribute value, each branch represents the out-turn of the test, and tree leaves represent classes. The drive model can be represented in different forms such as If-Then rules, decision tree, mathematical formula or neural networks [6]. Decision trees are simple to understand and provide good results even with small data.

Decision tree induction algorithms may be used for classification in several application areas, such as Education, Medicine, Manufacturing, Production, Financial analysis, Fraud Detection and Astronomy etc. There are many data processing algorithms like C4.5, ID3, CART, J48, NB Tree, REP Tree etc.

From Comparative Analysis of Data Mining Techniques on Educational Dataset, [4], comparative analysis of data mining techniques and algorithms it is clear that the comparison of classification algorithms Decision tree can handle any type of data (discrete and continuous) and also work well with numeric data in the classification. And also, in the comparison of Decision tree algorithms like C4.5, ID3, CART, J48, it is very clear that J48 can handle both the nominal and numeric values. J48 can also handle the missing values.

III. J48 ALGORITHM

J48 is associate open supply Java implementation of the C4.5 algorithm in the Weka data mining tool.C4.5 may be a program that makes a choice tree supported a group of labeled computer file. This algorithm was developed by Ross Quinlan.C4.5 is associate extension of Quinlan's earlier ID3 formula. The decision trees generated by J48 can be used for classification and builds decision trees from a set of labeled training data using the concept of information entropy. It uses the very fact that every attribute of the info may be wont to build a choice by cacophonous the info into smaller subsets. J48 examines the normalized the information gain (difference in entropy) that results from choosing an attribute for splitting the data. To make the choice, the attribute with the highest normalized information gain is used.

Then the algorithm recurs on the smaller subsets. The cacophonous procedure stops if all instances in a very set belong to identical category. Then a leaf node is made within the call tree telling to decide on that category. But it also can happen that none of the options offer any data gain. In this case J48 creates a choice node to a higher place within the tree victimization the expectation of the category.

J48 will handle each continuous and separate attributes, coaching information with missing attribute values and attributes with differing prices.

Further it provides an option for pruning trees after creation.

Algorithm J48:

INPUT:
D //Training data
OUTPUT
T //Decision tree
DTBUILD (*D)
T= {create|produce">produce root node and label with rending attribute;
T= Add arc to root node for every split predicate and
label;
For each arc do
D= Database created by applying splitting
predicate to D;
If stopping point reached for this path, then
T'= create leaf node and label with
appropriate class;
Else
T'= DTBUILD(D);
T= add T' to arc;
}

While building a tree, J48 ignores the missing values i.e. the value for that item will be expected supported what's renowned concerning the attribute values for the opposite records. The basic plan is to divide the information into vary supported the attribute values for that item that square measure found within the coaching sample. J48 permits classification via either call trees or rules generated from them [4][5].

.

IV. PROPOSED METHODOLOGY

4.1 Data Collection or Data Set Sample

We collected students' data from [10] that describing the academic performance and learning behavior of students. The collected data was organized in Microsoft Excel sheet. Each student record had the following attributes such as Previous Semester Marks (PSM), Class Test Grade (CTG), Seminar Performance (SEM), Assignment (ASS), Attendence (ATT), Lab Work (LW), End Semester Marks (ESM).

4.2 Tools Used-WEKA

WEKA stands for Waikato Environment for Knowledge Analysis. WEKA is a data mining tool that is developed at the University of Waikato, Newzeland. It uses GNU general public licenses and is freely available on following link: http://www.cs.waikato.ac.nz/~ml/weka. It is implemented in the programming language like java and has GUI for loading data, running analysis and producing visualization of result.

WEKA supports several data processing techniques and algorithms like classification, clustering, feature selection, data preprocessing, regression, visualization and clustering [4].

The GUI Interface of WEKA is shown in figure below

Fig1 : GUI Interface of WEKA

This is WEKA GUI Chooser. It provides four interfaces to work on:

Explorer

It is used for exploring the data with WEKA by providing access to all the facilities by the use of menus and forms.

Experimenter

WEKA Experimenter permits you to make, analyze, modify and run large scale experiments.

Knowledge Flow

It functions as explorer. It supports incremental learning. It handles data on incremental basis. It uses incremental algorithms to process data.

Simple CLI

CLI stands for command line interface. It simply provides all the practicality through command interface.

4.3 IMPLEMENTATION

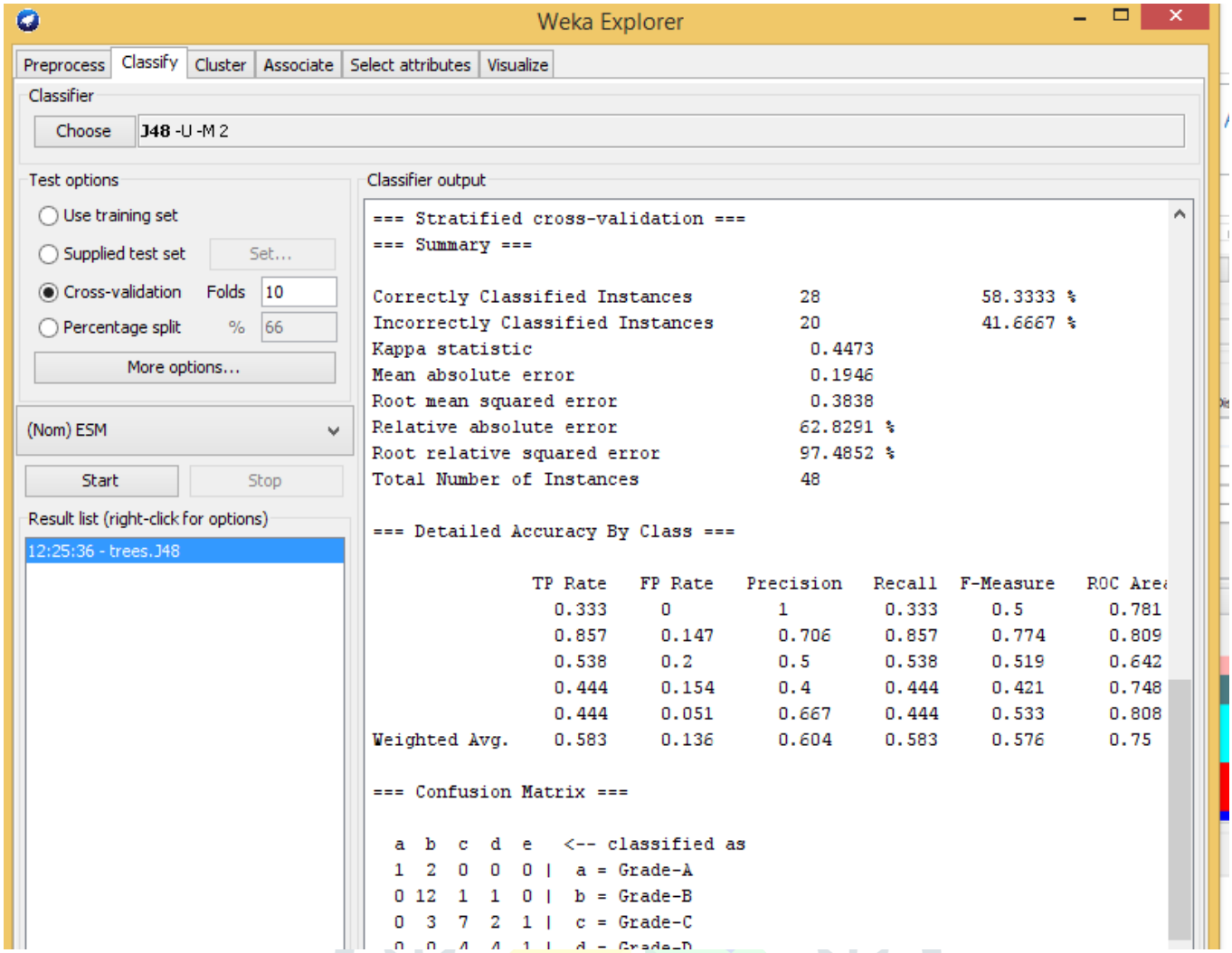
The WEKA is recognized as a most widely used tool for research in Data Mining and has achieved widespread acceptance in academia and as well as in business [17]. WEKA provides an excellent implementation of classification and clustering algorithms in the data mining with a good graphical user interface. Since this research also use data mining classification, therefore, WEKA is the best available option to use. In order to utilize WEKA software, the data needs to be transformed into a format compatible with WEKA which is ARFF format.

The following steps were being taken to bring out the information in desired format:

a. Data was collected from [10].
b. This data was then saved in CSV format in Excel (i-e filename.csv.)
c. WEKA environment provides the facility to convert data from CSV format into ARFF format. This facility was used and Student.csv was converted into Student.arff format.

Student.arff file was loaded into WEKA explorer. The Classify choice in wood hen permits USA to use classification formula on the dataset, to get calculable accuracy of the model, show the confusion matrix. We choose "J48" decision tree algorithm for classification. From the "Test Option" we select 10-fold cross-validation because we have not separated training data set and was therefore required to get a logical idea of accuracy for Evaluation. The attribute "Grade" was selected for prediction. The resultant model is generated in the Form of decision tree [9].

V. RESULTS AND FINDINGS



.

Fig2: J48 Algorithm Results from WEKA

Number of Leaves : 17
Size of the tree : 25
Correctly Classified Instances 28 58.3333 %
Incorrectly Classified Instances 20 41.6667 %
Kappa statistic: 0.4473
Mean absolute error: 0.1946
Root mean squared error: 0.3838
Relative absolute error: 62.8291 %
Root relative squared error: 97.4852 %
Total Number of Instances: 48

Table 1 Confusion Matrix

a	b	c	d	e	Classified as
1	2	0	0	0	A = GRADE-A
0	12	1	1	0	B = GRADE-B
0	3	7	2	1	C = GRADE-C
0	0	4	4	1	D = Grade-D
0	0	2	3	4	E = GRADE-E

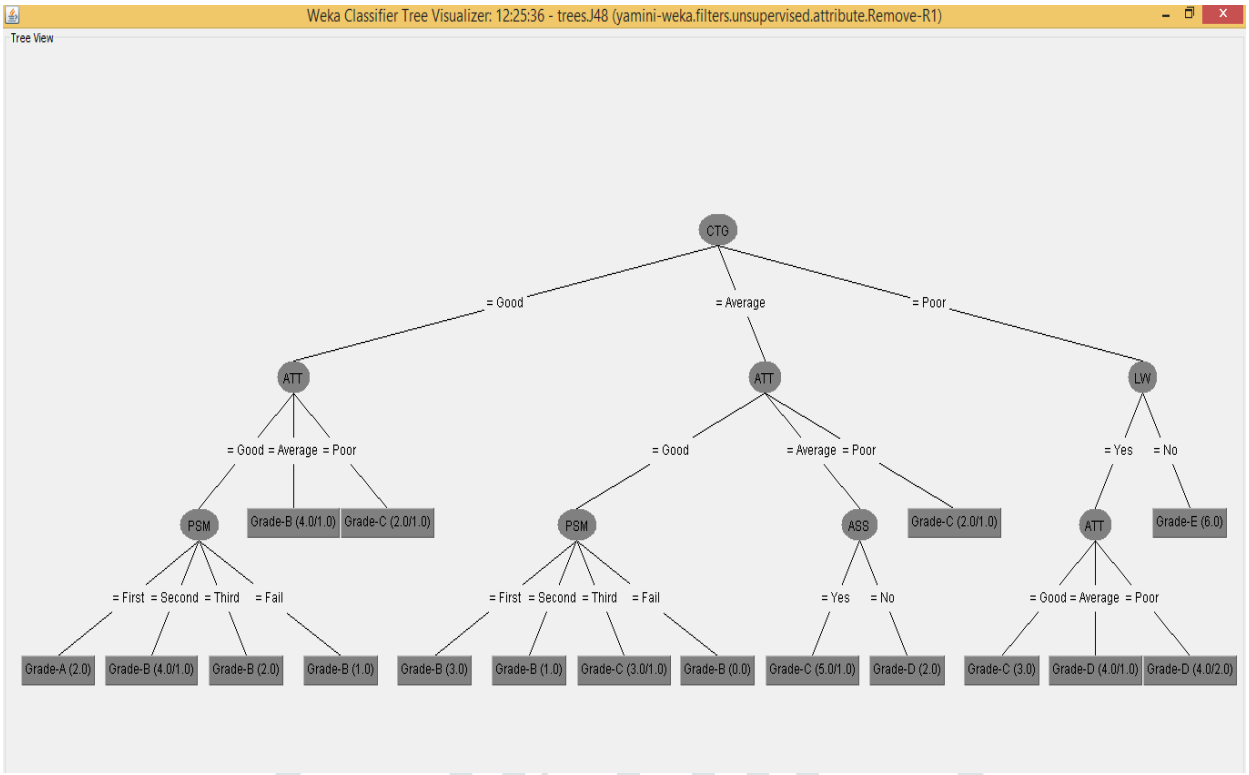


Fig3: Classification Decision tree of J48 Algorithm

V. ACKNOWLEDGMENT

The various data processing techniques are often effectively enforced on instructional information. From the higher than results it's clear that classification techniques are often applied on instructional information for predicting the student's outcome and improve their results. The effectiveness of different decision tree algorithms can be anatomized grounded on their delicacy. The splitting procedure stops if all instances in a subset belong to the same class. Then a leaf node is made within the call tree telling to settle on that category. But it may happen that none of the options offer any info gain.

In this case J48 creates a call node above within the tree victimization the expectation of the category. J48 will handle each continuous and separate attributes, coaching information with missing attribute values and attributes with differing prices.

REFERENCES

[1]. Gaganjot Kaur, Amit Chhabra Improved J48 Classification Algorithm for the Prediction of Diabetes *International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014.*

[2] M. Al-Razgan, A. S. Al-Khalifa, and H. S. Al-Khalifa, "Educational data mining: A systematic review of the published literature 2006-2013," in *Proc. the 1st International Conference on Advanced Data and Information Engineering*, 2013, pp. 711-719.

[3] F. Siraj and M. A. Abdoulha, "Mining enrolment data using predictive and descriptive approaches," *Knowledge-Oriented Applications in Data Mining*, pp. 53-72, 2007.

[4] Sumit Garg, Arvind K. Sharma, Comparative Analysis of Data Mining Techniques on Educational Dataset, International Journal of Computer Applications (0975 – 8887) Volume 74– No.5, July 2013.

[5] http://www.cs.waikato.ac.nz/~ml/weka.

[6] Manoj Bala et al., "Study of Application of Data Mining Technique in Education", International Journal of Research in Science and Technology, Vol. No. 1, Issue No. IV, Jan-March, 2012.

[7] Margaret H. Danham,S. Sridhar, " Data mining, Introductory and Advanced Topics", Person education, 1st ed., 2006

[8] Tina R. Patil, Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", International Journal of Computer Science and Applications Vol. 6, No.2, Apr 2013 ISSN: 0974-1011.

[9] Frank, E., and Whitten,I.H., (2005) Data Mining: Practical Machine learning and technique, San Francisco, Elsevier & Morgan Kaufmann

[10] Surjeet Kumar Yadav, Brijesh Bharadwaj, Saurabh Pal, "Data Mining Applications: A Comparitive Study for Predicting Students Performance ", International Journal Of Innovative Technology & Creative Enginnering, Vol.1 N0.12, December, ISSN: 2045-711.