International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-6,June 2016
ISSN: 2395-3470
www.ijseas.com

Prediction of Students Outcome Using Data Mining Techniques

R. Sumitha[1] , E.S. Vinothkumar [2]
PG Scholar[1], Associate Professor[2]

[1,2]Department of Computer Science and Engineering
[1,2]KLN College of Information Technology, Pottapalayam , TamilNadu, India
sumithar.r3@gmail.com[1], vinothkumar2k@gmail.com[2]

Abstract

Data mining focuses on collecting knowledge from databases or data warehouses and the information collected that had never been known before, it is valid and operational. Nowadays Educational Data Mining is an emerging discipline, concerned with various Approaches such as Predicting student performance, Analysis and visualization of data, Providing feedback for supporting instructors, Recommendations for students, Social network analysis and so on that automatically extracts meaning from large repositories of data generated by or related to people's learning activities in educational setting .One of the biggest challenges is to improve the quality of the educational processes so as to enhance student's performance. Thus, it is crucial to set new strategies and plans for a better management of the current processes. This model helps to predict student's future learning outcomes using data sets of senior students. Thereby to design student's data model using J48 algorithm which proved to be an efficient algorithm in terms of accuracy identified by a comparative study of data mining classification algorithms.

Keyword*:* Data Mining, Educational Data Mining

1. Introduction

Educational data has become a vital resource in this modern era, contributing much to the welfare of the society. [12]Educational institutions are becoming more competitive because of the number of institutions growing rapidly. To stay afloat, these institutions are focusing more on improving various aspects and one important factor among them is quality learning. For providing quality education and to face new challenges, the institutions need to know about their potentials which are explicitly seen and which are hidden. The truths behind today's educational institutions are a substantial amount of knowledge is hidden. To be competitive, the institutions should identify their own potentials hidden and implement a technique to bring it out. In recent years, Educational Data Mining has put on a mammoth recognition within the research realm as it has become a vital need for the academic institutions to improve the quality of education.

The higher education institutions has potential knowledge such as academic performance of students, administrative accounts, potential knowledge of the faculty, demographic details of the students and many other information in a hidden form. The technique behind the extraction of the hidden knowledge is Knowledge Discovery process. Recently Data mining is widely used on educational dataset. Educational Data mining (EDM) has become a very useful research area [1].

Data mining helps to extract the knowledge from available dataset and should be created as knowledge intelligence for the benefit of the institution. Higher education does categorize the students by their academic performance. Many factors influence the academic performance of the student. The model is mainly focused on exploring various indicators that have an effect on the academic performance of the students. The extracted information that describes student performance can be stored as intelligent knowledge for decision making to improve the quality of education in institutions. The knowledge stored is used for predicting the student's performance in advance.

2. 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 [11].

[1] The paper depicts the users, components as well as the various approaches in EDM. [2] In this paper a strategy to improve the student's performance is mentioned by mapping the student's record using K-mean clustering algorithm and grouping datasets into cluster but there is no future performance prediction.

[5] 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 .

[7] The reference introduces a CHAID prediction model to identify the factors influencing the performance of students in final examinations and predicting the grade of students using .NET framework. Decision Tree, and Multilayer Perception but attributes other than grades of student are not considered.

[9] The reference paper depicts an Efficient grouping of online Students data with similar Characteristics using CRISP – DM Methodology data in Moodle Database moreover No real time data collection is done. [6] The reference provides a methodology for Improving Quality of Educational Process by using clustering , correlation analysis and association rules on the other part decision making process for enhancing the quality of educational activities is not mentioned.

[8]The paper have a direct effect on quality assurance in e-learning and on the improvement of the teaching process through the adaptation of content by predicting behavior patterns using OLAP Analysis, Moodle, LMS but this study gives improvement in terms of the report system in the field of e-learning. But not concentrate on development and implementation of new modules, as well as user authentication.

[4] The paper analyzes the distressed students with the aim of detecting their identifying patterns and then possibly implementing actions to help them using K- Means Clustering Algorithm but there is no consideration about the learning systems. [3][12]The study Focus on predicting students performance and identifying the slow learners among students using a comparative study of Classification algorithms such as Multilayer Perception, J48 and REP Tree but not focused on Integration of data mining techniques with DBMS and E- learning [14].

[10] The paper depicts a comparative analysis for management of student retention using CRISP – DM, 10 fold Cross validation Approach using prediction model that can be used as decision aid in predicting students retention but Students outcome prediction in real time is loophole.

3. Proposed Framework

Educational data mining has vast amount of data that has to be organised in a consistent manner .To organise, analyse and classify students details K-mean Clustering algorithm is been used based on academic records. Thereby forming three clusters based on students record

- Low performance Student
- Average Student
- Smart Student

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

- best
- good
- average

- poor

Student model is designed for the prediction of the outcome of the student based on the framework given below in Figure 1.

This system provides an efficient analysis on student performance by data collection and result prediction.

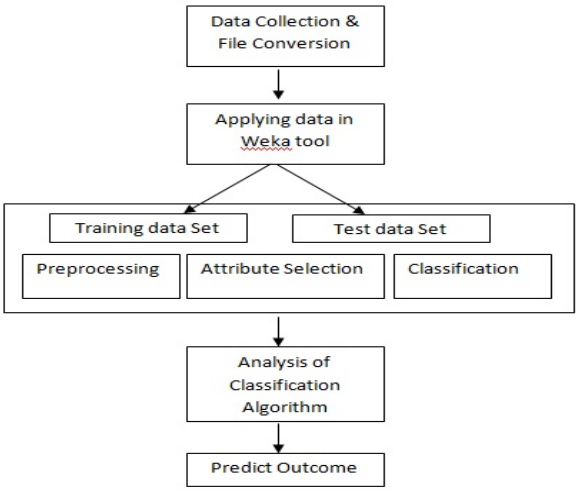

Figure 1: System Framework

4. Methodology

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. Attribute selection, Nominal conversion, file conversion and WEKA tool implementation. Comparative analysis of efficient classification algorithm is done to predict student's performance by creation of student model.

4.1 Data Collection and Preparation

The datasets of about 350 students are collected from I , II, III, IV year B.E CSE of KLN College of Information Technology (Affiliated To Anna University). In this process, a questionnaire form is used to collect the real data from the students that describe the relationship between learning behaviour and their academic performance. The variables for judging the learning and academic behaviour of students used in the questionnaire are student demographic details, School details, Attendance, CGPA and Final grade in last semester. These data's are thereby recorded in excel sheets for analysis.. Data sets about 300 students were collected by 3weeks.Among the dataset around 250 are been used as training dataset and 50 datasets as test data to design student model.

4.2 Data Selection and Transformation

In this stage only the data required for data mining are selected. A few derived variables were selected. From the available database, some of the information for the variables is collected. The data collected from Feedback forms and database .initially attribute selection is done. 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. The process of attribute selection deals with selecting the most appropriate attributes for classifying the data sets. By the analysis among the 24 attributes, attributes of higher ranking are used for classifying the training dataset.

The attributes are

- CGPA
- Arrears
- Attendance
- 12marks
- Engineering Cut-off
- Medium of Education
- Type of Board

All the predictor and response variables which were derived from the database are given in Table 1.On attribute selection the analysis is done for school and college dataset separately based on certain conditions as given below in figure and finally analysing the performance using final condition on both the records of college and school based on conditions.

For college details the conditions are applied and grouped as Best, Good, Average and Poor based on CGPA, ATD and ARR .but for school details the grouping is based on MOE, TOB, TWM and ECUT.Data collected from students as feedback and from database. The profile of students is defined based on the academic and demographic details of students.

The students' academic background is measured using the entry requirements to be fulfilled to get entry into the university/college. In this stage the only the data required for data mining are selected.

A few derived variables were selected. From the available database, some of the information for the variables is collected. The data collected from Feedback forms and database are entered in excel sheets and converted to ARFF format for further processing in WEKA tool.

Variable	Description	Possible values
TWM	Twelfth total marks	{best,good,average,poor}
MOE	Medium of Education	{English, Tamil}
TOB	Type of Board	{Stateboard,Matriculation,CBSE, Diploma}
ECUT	Engineering Cut-off	{best,good,average,poor}
CGPA	Current CGPA	{best,good, average, poor}
ARR	No. of arrears	{no,average,poor,very poor}
ATD	Attendance percentage	{best,good,average,poor}

Table 1: Dataset Description

4.3 Implementation of Model

The main objective is to explore if it is possible to predict the performance of the student (output) based on the various explanatory (input) variables which are retained in the model. The classification model was built using several different algorithms and each of them using different classification techniques. The WEKA Explorer application is used at this stage.

The implementation of the dataset is done using a data mining tool WEKA. WEKA is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. WEKA stands for Waikato Environment for Knowledge Analysis. From the above data, the student.xl file is converted to ARFF (Attribute Relation File Format)[11]for data analysis WEKA explorer. WEKA tool console is depicted in figure 6.The ARFF file i.e. student.arff is opened in WEKA console. The classify panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions, or the model itself. [14][13]The algorithm used for classification is Naive Bayes, Multilayer Perception (MLP), REP tree and J48. Each classifier is applied for two testing options - use Training set and supplied test set. After pre-processing attribute selection is done using select attribute option in WEKA tool to identify the attributes which has higher rank of contribution to the analysis. Since there is no separate evaluation data set, this is necessary to get a reasonable idea of accuracy of the generated model. This predictive model provides way to predict the student's future learning outcome.

A. Applying Training Data in WEKA Tool

WEKA is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. WEKA stands for Waikato Environment for Knowledge Analysis. From the above data, student.arff file is created, and then this file is loaded into WEKA explorer for processing.

- Choose "WEKA 3.7.x" from Programs. The first interface that appears looks like the one given below.
- Explorer: An environment for exploring data. It supports data preprocessing, attribute selection, learning and visualization
- Get to the WEKA Explorer environment and load the training file using the Preprocess mode.
- Get to the Classify mode (by clicking on the Classify tab) as shown below:
- Now, you can specify the test options (by checking the corresponding button):
- Use training set means that you use the training set (the file you loaded in Preprocess) for testing.

B. Comparative identification of efficient Classification algorithm

- Initially the datasets are filtered with attributes of higher rank for classification based on select attribute option.
- the datasets are thereby tested with various classification algorithms
- Classifiers simulated are :
  - ❑ Naïve bayes
  - ❑ Multilayer Perception

- SMO
- J48
- REP Tree

- Running a Test
  Click on the Choose button and choose a classifier (the default is ZeroR – the majority predictor).
- After selecting a classifier and setting its parameters (you can always start with the defaults), click on OK and then on Start. You get the output from the classifier in the Classifier output window.

C. Result of J48 Classifier

On evaluating the data set under J48 classifier the result generated is as shown in Figure 2.which shows the correctly classified instance by 97% .

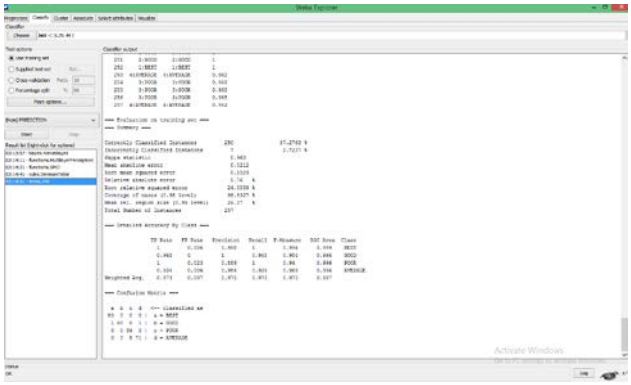

Figure 2: Applying J48 Algorithm

D. Applying test data in WEKA tool

To analyze the performance of students using training data result test data's are given to predict performance based on training data.

- Get to the WEKA Explorer environment and load the training file using the Preprocess mode.
- Get to the Classify mode (by clicking on the Classify tab) .
- Now, you can specify the test options (by checking the corresponding button):
- Use test set means that you use the test data set (the file you loaded by select test file) for testing.

The below Figures 3 and Figure 4 represent the j48 algorithm implementation for test data and tree visualisation of J48 algorithm.

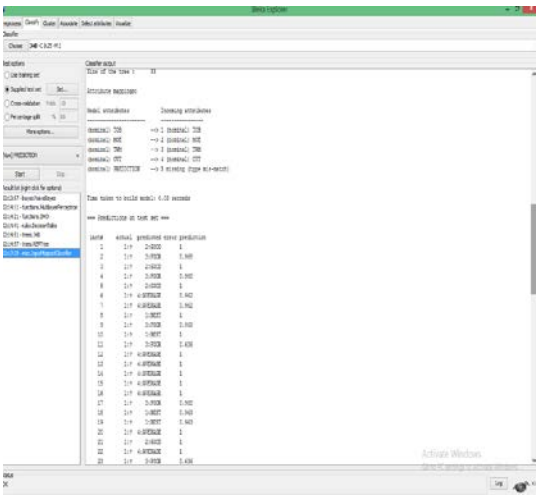

Figure 3: Applying J48 Algorithm for Test data set

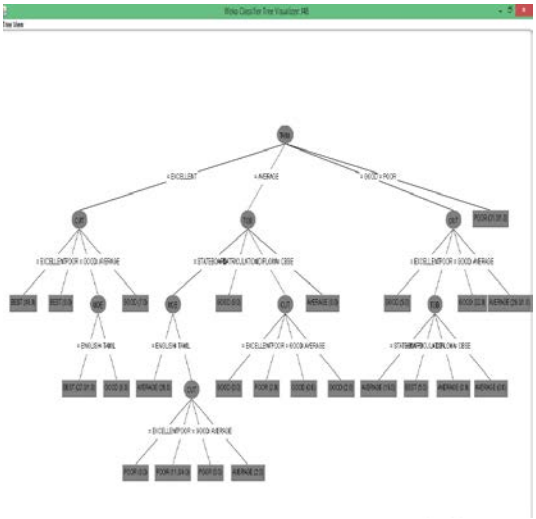

Figure 4: J48 Tree Visualisation

5. Design of Model

In the designing of student model J48 algorithm provide a maximum accuracy in classifying the instances in an efficient way. The student model is created in Net Beans using java coding. The J48 algorithm identified by comparative analysis of classification algorithm is taken for designing.

The student model is designed to view models such as

1. Login credential
2. Student detail record
3. Student outcome analysis

J48 is a tree based learning approach, based on iterative dichtomiser (ID3) algorithm. It uses divide-and-conquer algorithm to split a root node into a subset of two partitions till leaf node (target node) occur in tree.

Given a set T of total instances the following steps are used to construct the tree structure.

Step 1: If all the instances in T belong to the same group class or T is having fewer instances, than the tree is leaf labelled with the most frequent class in T.

Step 2: If step 1 does not occur then select a test based on a single attribute with at least two or greater possible outcomes. Then consider this test as a root node of the tree with one branch of each outcome of the test, partition T into corresponding T1, T2, T3 ....., according to the result for each respective cases, and the same may be applied in recursive way to each sub node.

Step 3: Information gain and default gain ratio are ranked using two heuristic criteria by algorithm J48.

The design of form includes a login form as shown in Figure 5 depicting username and password for access. Student details are gathered using the form as in Figure 6.students performance are analysed as per the prediction with suggestions as in Figure 7. The outcome of student's are analysed as if Best, Good, Average or Poor from the student model generated and thereby providing suggestions for upliftment of students performance.

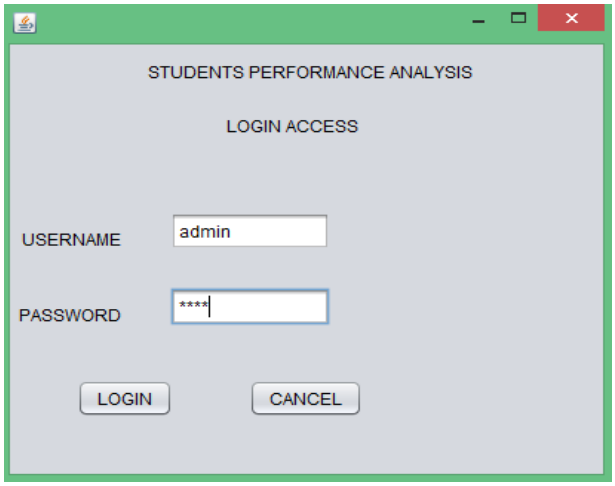

Figure 5: Login Form

Figure 6: Student Details Record

Figure 7: Student Outcome Anlaysis

6. Performance Analysis

By the simulation of dataset with the various classifiers the accuracy of correctly classified instances is as below in Table 2.

Classification Algorithm	Accuracy (In %)
Naive Bayes	85.92
MultilayerPerception	94.94
SMO	94.34
Decision table	96.10
J48	97.27
REP Tree	95.33

Table 2 : Classifier Accuracy

As J48 algorithm classifies the instance with maximum accuracy ,it is used in designing the student model to predict students performance by analyzing training data and test data .thereby predicting students performance as Best,Good ,Average or Poor and providing valuable suggestions.

7. Conclusion and Future Work

Education data mining main focus is to analyse the education system. The model focuses on analyzing the prediction accuracy of the student's performance .The dataset that comprises of all academic and personal factors of the students. This model can be useful in the educational system like Universities and Colleges. By this model we can know the academic status of the students in advance and can concentrate on students to improve their academic results and placements. Thereby improve their standards and reputations. As a result the quality of education can be improved. The results of the data mining algorithms for the classification of the students based on the attributes selected reveals that the prediction rates are not uniform among the algorithms. The range of prediction varies from (80-98%).Thereby by comparative analysis of classification algorithms (such as Naïve bayes, MLP ,SMO ,Decision Table, REP tree, J48) using WEKA tool, it is proven that the attributes chosen from the original dataset have high influence using J48 with an accuracy of 97% under analysis and used for predicting test data set for future outcome as best,good,average or poor.

The work can be further extended out by designing the student model analysing records of students extra-curricular skills and provide a suggestions on communication and technical skill development by which students can be build in professional aspect of talents. .

References

[1] Crist´obal Romero, *Member, IEEE, and Sebasti´an Ventura, Senior Member, IEEE,* "Educational Data Mining: A Review of the State of *the* Art " VOL. 40, NO. 6, NOVEMBER 2010.

[2] Parneet Kaura, Manpreet Singhb ,Gurpreet Singh Josanc "Classification and Prediction based DataMining Algorithms to Predict Slow Learners in Education Sector" Science Direct Procedia Computer Science 57 (2015) 500 – 508 2015 (ICRTC- 2015).

[3] Renza Campagni, Donatella Merlini, Renzo Sprugnoli, Maria Cecilia Verri, "DataMining Models for Student Careers", Science Direct - Expert Systems with Applications 42 (2015) 5508–5521.

[4] Nurbiha A Shukora , Zaidatun Tasira, Henny Vander Meijden, "An Examination of Online Learning Effectiveness using DataMining" ,Science Direct - Procedia - Social and Behavioral Sciences 172(2015) 555 – 562.

[5] Harwatia, Ardita Permata Alfiania, Febriana Ayu Wulandari," Mapping Student's Performance Based on Data Mining Approach", Science Direct Agriculture and Agricultural Science Procedia3 (2015) 173 – 177.

[6] Manolis Chalaris, Stefanos Gritzalis, Manolis Maragoudakis, Cleo Sgouropoulou and Anastasios Tsolakidis," Improving Quality of Educational Processes Providing New Knowledge using Data Mining Techniques", Science Direct - Procedia - Social and Behavioral Sciences 147 (2014) 390 – 397.

[7] V.Ramesh Assistant Professor Department of CSA, SCSVMV University Kanchipuram India "Predicting Student Performance: A Statistical and Data Mining Approach", International Journal of Computer Applications (0975 – 8887) Volume 63– No.8, February 2013 35.

[8] Krpan, Slavomir Stankov, "Educational Data Mining for Grouping Students in E-learning System", Proceedings of the ITI 2012 34th Int. Conf. on Information Technology Interfaces, June 25 – 28, 2012, Cavtat, Croatia.

[9] Dursun Delen, "A Comparative Analysis of Machine Learning Techniques for Student Retention Management", Science Direct - Decision Support Systems 49 (2010) 498–506.

[10] Ali Buldua, Kerem Üçgüna, "DataMining Application on Students' Data", Science Direct - Procedia Social and Behavioral Sciences 2 (2010) 5251–5259.

[11] Jiawei Han , Micheline Kamber ,Jian Pei ,"Data Mining: Concepts and Techniques", Third Edition (The Morgan Kaufmann Series in Data Management Systems) 3rd Edition.

[12]http://www.intechopen.com/books/theory-and-applications-for-advanced-text-mining.

[13] http://www.wekaexplorer/wekatutorial.pdf.

[14] https://weka.wikispaces.com/Netbeans/(weka-src.jar).