International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 02, February 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

PREDICTING STUDENT PERFORMANCE USING DATA MINING CLASSIFICATION TECHNIQUES

Lakshmipriya. K [1], Dr. Arunesh P.K [2]

[1]*M.Phil Scholar, Department of Computer Science, Sri S.R.N.M College(India)*

[2]*Associate Professor, Department of Computer Science, Sri S.R.N.M College, Sattur (India)*

ABSTRACT

Data mining applications are becoming a common tool in perceptive to solve educational and administrative problems in education system. The Educational Data Mining (EDM) focuses on modeling and evaluate student's performance based on examination scores and college atmosphere evaluation questionnaire. In the academic information system the data mining is mainly used for predicting the students' performance. This article provides a comprehensive literature review and classification methods for data mining techniques applied to academic information. The Data mining techniques, namely Decision Trees, Naive Bayes, Associative classification, and SVM are analyzed on student information and the data are converted in to decision tree model using R Studio programming Open Source Tool. The real data set for the college students was collected and filtrations of desired potential variables are also performed.

Keywords: Classification Algorithm, Decision Tree, Support Vector Machines, Apriori Algorithm, Naïve Beyes Algorithm, Student Performance Evaluation.

I. INTRODUCTION

Data mining, also known as knowledge discovery in databases, can be defined as the process of analyzing large information repositories and of discovering implicit, but potentially useful information (Han, Kamber, & Pei, 2011). Data mining techniques originate in this context, with the aim of discovering hidden and non-trivial relationships among information of various nature. This collection of techniques, used in different sectors, including the educational environment,comes from the traditional methods of data analysis and have the characteristic of being able to treat large amounts of data (Renza Campagni, 2015). Data mining is used to extract the useful important information from large repositories using knowledgeable patterns. It has been used in many applications such as educational data mining, opinion mining, web mining, text mining, health care, Agriculture, etc. Currently huge amount of data storages in educational databases are used to predict of students academic performances.

In the field of academic, educational data mining is a recent research area that explores and analyzes the information stored in student databases in order to analyze, understand and improve the learning process and academic performance of students. Data are analyzed by using statistical, machine learning and data mining algorithms, with the aim of resolving problems of educational research and improve the entire educational process. Recently there has been an increase in the use of educational software instruments and of databases

ISSN (O) 2454-9665
ISSN (P) 2454-0663

containing students information, so we have large repositories of data reflecting how students learn. In addition, the use of Internet in education has created the context of e-learning or web-based education which continuously generates large amounts of data concerning the interactions between teaching and learning. Educational data mining tries to use all this information to better understand learners and learning, and to develop methodologies which, integrating the data with the theory, allow to improve the educational process. Educational data mining is a growing research area that involves researchers all over the world from different and related research areas and since 2008 an annual International Conference on Educational Data Mining has been established (http://www.educationaldatamining.org) (Renza Campagni et.al., 2015). In the recent years researchers made great efforts in the direction of relating the state of the art of EDM research and, several model proposals and survey papers have been published on this research area (Lorena, 2015; Xing Wanli, 2015; Harwati, 2015; Renza Campagni,2015).

EDM classification is one among the data mining technique and it is the process of supervised learning to separate data into different class data set. Example, the students are categorizes into three categories, which are good, average and bad performance. Students in bad performance category may have a higher probability of failing. Educational data mining uses many techniques such as Decision trees, Neural networks, Naïve bayes, K-nearest neighbor, association mining etc. In this paper, the factors which are associated with student whose academic performance is not good are identified using four different classification techniques- Decision tree algorithms, Support Vector Machine (SVM), Apriori algorithm and naïve beyes algorithm. These techniques are used to build classifier models. Their performance are compared over a real world data set, college student info are collected and filtration of desired potential variables is done using R programming language, open source tool. The data set of student academic records is tested and applied on above four algorithms. The result, statistics are generated based on the classification algorithm and comparison is performed on the four classifiers algorithm in order to predict the accuracy and to find the suitability of the classification algorithm.

II. LITERATURE REVIEW

Data mining in higher education is a recent research field and this area of research is gaining recognition because of its potentials to educational institutions. It can be used in educational field to enhance our understanding of learning process to focus on to identifying and extracting variables related to the learning process of students are described in [1]. John Jacob et, al., [2] describes that higher education institutions goal is to improve the quality of education.Mustafa Agaoglu [3] applied the educational data mining concerns with developing methods for discovering knowledge from data that come from educational data mining. This research focuses on modeling student's performance is utilized to consider course evaluation questionnaires. Here, the most vital variables that split ‘satisfactory" and ``not satisfactory" instructor performances based on students' perception are establish.

Kamal Bunkar [4] describes the data mining classification enhancing the worth of the higher educational system by evaluating student data to study the major attributes that may affect the student performance in courses. The analytical models obtained from the student data set by three machine learning algorithms: the C4.5 decision tree algorithm, the CART algorithm, and ID3 decision tree algorithm. Suhem Parack [5] applied the data mining techniques to improve the efficiency of higher education institutions. Authors concluded that if data mining techniques such as Apriori and k-means are applied to higher education processes, it would help to improve

ISSN (O) 2454-9665
ISSN (P) 2454-0663

academic performance. Nguyen Thai Nghe [6] explained the Decision Tree and Bayesian Network algorithms are common for predicting the academic performance of undergraduate and postgraduate students. This algorithm accurately predicting student performance, accuracy of data mining algorithms are compared and authors demonstrated the maturity of open source tools.

David L. la et. al.,[16] suggested that education-oriented data mining allows to predict determined type of factor or characteristic of a case, phenomenon or situation. In this article the mining models used are described and the main results are discussed. Mining models of clustering, classification and association are considered especially. In all cases seeks to determine patterns of academic success and failure for students, thus predicting the likelihood of dropping them or having poor academic performance, with the advantage of being able to do it early, allowing addressing action to reverse this situation. This work was done in 2013 with information on the years 2009 to 2013, students of the subject Operating Systems tertiary career Superior Technical Analyst (TSAP) Higher Institute of Curuzú Cuatiá (ISCC), Corrientes, Argentina [16].

Renza Campagni et.al., [17] proposed a data mining methodology to study the behavior of university graduated students in terms of their careers, by analyzing the corresponding database, and, as a matter of fact, many interesting relationships among data are found in this work. By introducing the concept of ideal career, that is, the career of the ideal student who takes each exam at the end of the corresponding course without delay, various types of distance

between the career of any student and the ideal one; in particular, authors used the Bubble sort distance are computed. Authors defined the career of a student in different ways, by considering the temporal information or not, according to the method of analysis used; for traditional clustering analysis, also introduced an ideal career that was a sequence of exams, without the temporal information. For the frequent patterns analysis we represented a career as a sequence with temporal information, as this technique requires [17].

Xing Wanli et.al., [18] Building a student performance prediction model that is both practical and understandable for users is a challenging task fraught with confounding factors to collect and measure. Most current prediction models are difficult for teachers to interpret. This poses significant problems for model use e.g. personalizing education and intervention as well as model evaluation. In this paper, learning analytics approaches, educational data mining (EDM) and HCI theory to explore the development of more usable prediction models and prediction model representations using data from a collaborative geometry problem solving environment: Virtual Math Teams with Geogebra are synthesized. Harwati et.al., [19] focuses on mapping students using K-mean Cluster algorithm to reveal the hidden pattern and classifying students based on their demographic (gender, origin, GPA, grade of certain courses), and average of course attending. About 300 student's data is covered. From the calculation using SPSS 16, it is found that there are four clusters formed based on six variables, namely: smart students (45.74), standart student (33.33%), and 20.92% belongs to low performance students.

Lorena et.al., [20] provides a comprehensive literature review and classification method for data mining techniques applied to academic libraries. To achieve this, forty-one practical contributions over the period 1998–2014 were identified and reviewed for their direct relevance. Each article was categorized according to the main data mining functions: clustering, association, classification, and regression; and their application in the four main library aspects: services, quality, collection, and usage behavior.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

III. CLASSIFICATION MODEL

Classification is one of the most general application domains of data mining. The aim of classification is to exactly predict the target class for each case in the data [7]. Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations whose category membership is known. After knowledge phase, the classification, performance the classifier model built is evaluated on an independent test set before used.

In classification, there are special types of techniques and algorithms likely to use for building a classifier representation. Classification model contains various algorithms as Decision tree, Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression, Discriminant Analysis (DA), Rule Based System and Bayesian Belief Networks. In this paper the decision tree algorithm are used [8]. The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree and applying the tree to the dataset. Decision tree algorithm is one of the predictive modeling approaches used in statistical or impurity, data mining and machine learning. Its goal to generate a model that predicts the value of a target variable based on several input variables. The impurity measures that are used to get the homogeneity of instances in a node of the tree, Information Gain, Gain Ratio and Gini Index are the mainly identified ones. Mostly Information Gain is used in Iterative Dichotomiser (ID3), but Gini Index value is used in Classification and Regression Trees. A sample representation of a decision tree was described in [9]:

3.1. C5.0 Algorithm

C5.0 algorithm is used to create a decision tree which can be used for classification so it referred to as a statistical classifier. This model works through splitting the model based on the field that provides the maximum information gain. Every subsample defined in the first split is then split again, generally based on a different field, and the process repeats until the subsamples cannot be split any further[10]. Finally, the lowest-level splits are reexamined, and those that do not contribute significantly to the value of the model are removed or pruned. In C4.5 and C5.0 Algorithms, there are some improvements in terms of handling the bias toward tests with many outcomes performance and pruning. Gain Ratio is used for these algorithms and its extension to ID3 Calculations which has a kind of normalization to Information Gain using SpiltInfo values [9].

3.2. Naive beyes

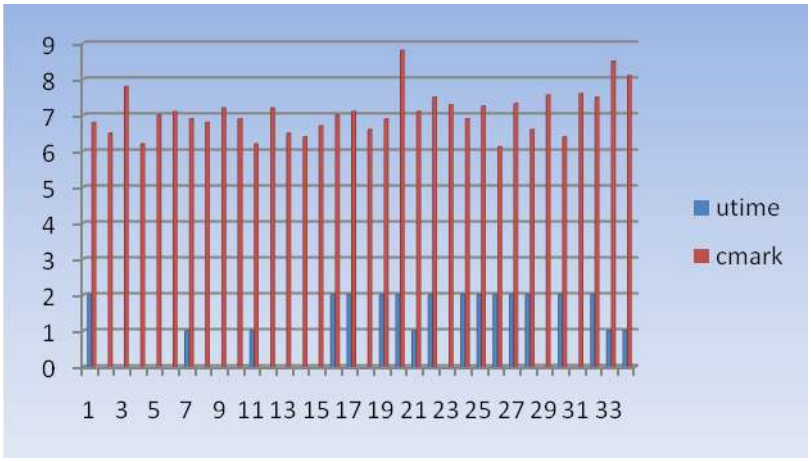
Naïve beyes classification technique is based on Bayes' Theorem with an assumption of independence among predictors. It is very simple, which assumes that the classification attributes are independent and they do not have any connection between them. A lot of researchers have found that this assumption of liberty do not work in the entire cases for which other different methods are proposed to raise the performance. The creative Naïve Bayesian technique is based on the conditional probability and the maximum likelihood incidence. The Naive Bayesian algorithm based on the description provided in [13]:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 02, February 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

IV. RESULT AND ANALYSIS

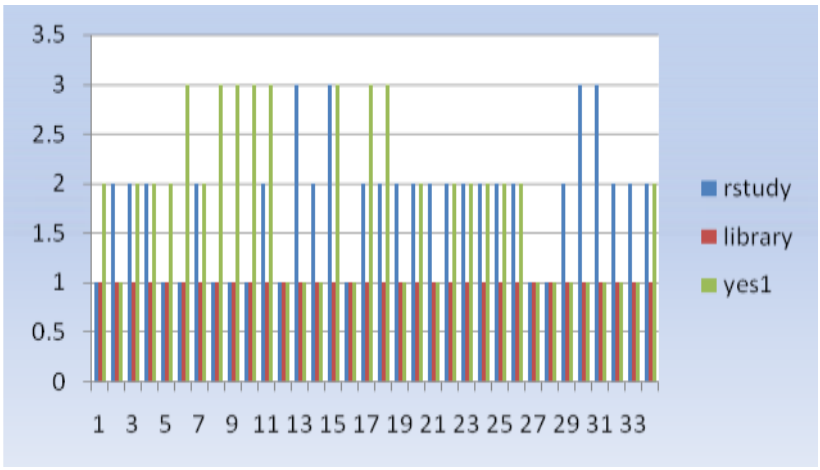
In this study, four classification models decision tree algorithm (C5.0), SVM, Apriori algorithm and naïve beyes algorithm are examined. The performances of these models are evaluated on the test data in terms of accuracy and specificity.



Decision Tree

utime->Mobile usage time

cmark->Cumulative mark

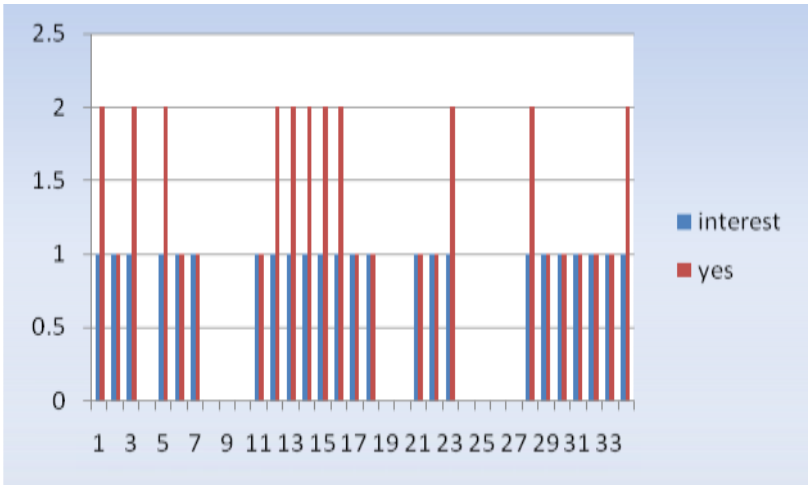


SVM

rstudy->Regular study

library->Library usage

yes-> library useage



Naïve Beyes

interest-> Particular , interesting subject

yes-> subject is interested or not

ISSN (O) 2454-9665
ISSN (P) 2454-0663

This process goes on until all the data classification is perfectly classified or run out of attributes. The knowledge represented in the form of IF-THEN rules are

If mobile='yes' or 'no' THEN cmark

If rstudy='yes' AND library='yes' THEN use

If interest='yes' THEN why

Pruning technique was executed by removing nodes with less than desired number of objects.

V. CONCLUSION

In this paper, we analyzed the data mining methodologies to study the behavior of university graduated students in terms of their career interests, by analyzing the corresponding database, and, as a matter of fact information like mobile phone usage, internet connectivity , subject assignments, study timings, interestingness and grade. We found the interesting relationships among data. The classification task is used in this work is to predict student performance. This study will also work to identify those students which needed special attention to reduce fail ratio and taking necessary action for the future career. This analysis shows that different technologies can be used to predict student performance in education system with different attributes.

REFERENCE

[1] Norlida Buniyamin, "Educational Data Mining for Prediction and Classification of Engineering Students Achievement", IEEE, 2015

[2] John Jacob "Educational Data Mining Techniques and their Applications", IEEE, 2015, pp: 1344-1348

[3]MUSTAFA AGAOGLU "Predicting Instructor Performance Using Data Mining Techniques in Higher Education", IEEE, 2016, pp: 2379-2387

[4]Kamal Bunkar "Data Mining: Prediction for Performance Improvement of Graduate Students using Classification", IEEE, 2012

[5]Suhem Parack "Application of Data Mining in Educational Databases for Predicting Academic Trends and Patterns", IEEE,

[6]Nguyen Thai Nghe "A Comparative Analysis of Techniques for Predicting Academic Performance", IEEE, 2007

[7] www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

[8] https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

[9] http://www.saedsayad.com/decision_tree.htm

[10]http://www.patricklamle.com/Tutorials/DecisiontreeR/Decisiontreesin0R%20using%20C50.html

[11] http://www.code2learn.com/2015/02/frequent-itemsets-apriori-algorithm-and.html

ISSN (O) 2454-9665
ISSN (P) 2454-0663

[12Jayshree Jha, "Educational Data Mining using Improved Apriori Algorithm", International Journal of Information and Computation Technology,2013,pp:411-418

[13]S. Karthika "A Naïve Bayesian Classifier for Educational Qualification", Indian Journal of Science and Technology, July 2015

[14] R. Quinlan. (2004). C5.0: An Informal Tutorial. [Online]. Available: http://www.rulequest.com/see5-unix.html

[15] O. K. Oyedotun, S. N. Tackie, E. O. Olaniyi, and A. Khashman, ``Data mining of students' performance: Turkish students as a case study," Intell. Syst. Appl., vol. 7, no. 9, pp. 20_27, 2015.

[16] David L. la Red Martínez1,2,*, Carlos E. Podestá Gómez, Contributions from Data Mining to Study Academic Performance of Students of a Tertiary Institute, American Journal of Educational Research, 2014, Vol. 2, No. 9, 713-726

[17] Renza Campagni, Donatella Merlini ⇑, Renzo Sprugnoli, Maria Cecilia Verri, Data mining models for student careers, Expert Systems with Applications 42, 2015, 5508–5521

[18] Xing Wanli 1, Guo Rui, Petakovic Eva 2, Goggins Sean, Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory, Computers in Human Behavior 47 (2015) 168–181

[19] Harwati, Ardita Permata Alfiani, Febriana Ayu Wulandari, Mapping Student's Performance Based on Data Mining Approach (A Case Study), 2210-7843, 2015 Published by Elsevier, ScienceDirect, The 2014 International Conference on Agro-industry (ICoA): Competitive and sustainable Agro industry for Human Welfare, Agriculture and Agricultural Science Procedia 3 (2015) 173 – 177

[20] Lorena Siguenza-Guzman , Victor Saquicela, Elina Avila-Ordóñez, Joos Vandewalle, Dirk Cattrysse, Literature Review of Data Mining Applications in Academic Libraries, Else vier , The Journal of Academic Librarianship 41 (2015) 499–510