Data Mining Techniques for Predicting Student Performance on Educational Data

Sindhurani G. S.[1] Shashirekha H.[2]
[1]M.Tech Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Department of PG Studies, Visvesvaraya Technological University, Mysuru, India

Abstract— The methods of Data mining are used in evaluating the given data and to mine the unknown facts and knowledge which greatly supports the researchers to take effective decisions. Due to the tremendous growth in recent technology like social media, it may divert the students from their actual track, and this is one of the reasons for the students to perform poor in academic activities and it even leads to course drop outs. This paper reviews the previous research works done on students' performance prediction, analysis, early alert and evaluation by using different methods of data mining.

Key words: Data Mining, Methodology, Data Mining Techniques, Machine Learning Process

I. INTRODUCTION

In the recent decays, the exploration in the educational field is increasing rapidly due to the colossal growth of data that related to the performance of students' academics. It increases the accuracy and quality of students' performance by predicting it in-advance. Early Prediction of students' performance is to enhance the quality of education in various traits. It helps in predicting at risk students during the course time itself and not at the end of the course. It acts as an effective tool that provides information to change educators' practices and make an alert to help students get back on track. It will be a successful method for improving academic success and retention. This early prediction benefits the students to take necessary steps in advance to avoid poor performance and to improve their academic scores. It benefits both the course instructors as well as the students whose performance is lagging in class. The feature of this prediction system is that it can be used early and as needed in prior of semester for faculty to communicate their concerns to students when signs of risk occur. So that students graduate on time without reappearing in the semester and are prepared to flourish in actual course and get back to careers on time. Student's academic performance in educational environment is based upon the mental and environmental factor can predicted by using various data mining techniques. The data collected for this educational practice i.e. data from traditional learning system will greatly avoid proxy method where as in E-Learning system it is not possible. This approach will support to predict the students' performance by analyzing the student's academic record such as internal assessment marks, assignment submission, and attendance percentage.

II. LITERATURE SURVEY

(Xinga, Chen, Steinc, & Marcinkowsk, 2016) [1] , present a paper to predict a dropouts in online courses: for this prediction they applied, two algorithms called General Bayesian Network (GBN) and decision tree (C4.5) are implemented to reduce the higher dropout rates in online education .

(Han Hu a, Lo a & Shih, 2014) [2], Develops the paper to alert students by some warning to improve their performance earlier. In their study they used three well-known single classification techniques, C4.5, CART, and LGR and made a comparative study to develop a system to predict student's E- learning performance by giving early alert.

(Wanli, Eva & Sean, 2015) [3], presents a paper to predict final year student performance as a model through Genetic Programming, and to Integrate the learning analytics, educational data mining and theory they demonstrate the structure for connecting trace data to a hypothetical framework, the processing data using the algorithm of Genetic Programming approach which outperforms the traditional models in prediction rate and also in interpretability.

(Campagni, Merlini, Sprugnoli, & Verri 2015) [4], Develops a model for student careers, in this paper they present different approaches based on clustering and sequential patterns techniques in order to identify strategies for improving the performance of students and the scheduling of exams. (Sembiring, et. al., 2011) [5], Made the presentation on Predicting students' performance academically by applying some of the techniques of data mining. The results of this study stated a model of student academic performance predictors by employing (psychometric factors) as variables predictors.

(Kalpana, et. al., 2014) [6], in this paper they made analysis on students Intellectual Performance by Using Data Mining Techniques. This presentation intends to analysis the student's performance in different categories of measurements. (Kaura, Singh & Josanc, 2015)

[7] ,In this paper they made a comparison study to predict slow learners in educational sector using Classification and prediction built data mining algorithms, in this paper they targeted the slow learners and the output dataset is tested and analyzed with five classification algorithms which are Multilayer Perception, Naïve Bayes, SMO, J48 and REP Tree. (Sullare, Thakur, Mishra, 2016)

[8], presents a paper on students' Performance, based on Grouping up of Neighbors Students in Progressive Education Datasets. In this paper they used Naive Bayes clustering method to assess student's performance in the end semester examination from education databases. (Nagar et. al., 2015)

[9], presents a paper on (Data Mining Clustering Methods), in this paper they made a detailed comparison study on different clustering techniques, an unsupervised learning method which makes the cluster of bits and pieces or forms according to their similarity and dissimilarity bases.

This paper gives review about various clustering methods. (Baradwaj, Pal, 2011)

[10], presents a paper to mine the educational data by analyzing the students' academic performance, for this analysis they used decision tree method for identifying the dropouts and predict students who need special attention and makes the work of educators easier in providing some appropriate warning or advising.

III. STEPS OF DATA MINING

Data mining is the method of discovering numerous models, derived values and summaries from a given collection of information. It's necessary that the problem of discovering or estimating dependencies from information or discovering new information is simply one part of the overall experimental procedure utilized by engineers, scientists and others who apply standard steps to draw conclusions from information. The overall method of finding and decoding patterns and models from information involves the recurrent application of the subsequent steps:

A. Understand the application domain

Understand the application domain, the relevant previous knowledge and the goals of the end-user (formulate the hypothesis).

B. Data Collection

Determining how to find and extract the right data for modeling. First, we need to identify the different data sources are available. Data may be scattered in different spreadsheets, files, and hard-copy (paper) lists.

C. Data Integration

Integration of multiple data cubes, databases or files. A big part of the integration activity is to build a data map, which expresses how each data element in each data set must be prepared to express it in a common format and record structure.

D. Data Selection

First of all the data are collected and integrated from all the various sources, and we select only the data which useful for data mining. Only relevant information is selected.

E. Pre-Processing

The Major Tasks in Data Preprocessing are: Cleaning, Transformation and Reduction.

1) Data cleaning

Additionally known as data cleansing. It deals with errors detection and removing from information so as to improve the quality of information. Information cleaning sometimes includes fill in missing values and identify or remove outliers.

2) Data Transformation

Data transformation operations are additional procedures of data pre-processing that would contribute toward the success of the mining process and improve data-mining results.

Some of Data transformation techniques are Normalization, Differences and ratios and Smoothing.

3) Data Reduction

For large datasets there's an increased probability that an intermediate, data reduction step should be performed before applying data mining techniques. While massive datasets have potential for higher mining results, there's no guarantee that they'll produce better knowledge than small datasets. Data Reduction obtains a reduced dataset representation that's much smaller in volume, however produces constant analytical results.

F. Building the model

In this step we elect and implement the appropriate data mining task (ex. association rules, serial pattern discovery, classification, regression, clustering, etc.), the data mining technique and also the data processing algorithm(s) to create the model.

G. Interpretation of the discovered knowledge (model /patterns)

The interpretation n of the detected pattern or model reveals whether or not the patterns are interesting. This step is additionally known as Model Validation/ Verification and uses it to represent the result in an appropriate approach so it may be examined completely.

H. Decisions / Use of Discovered Knowledge

It helps to make use of the knowledge gained to take better decisions.

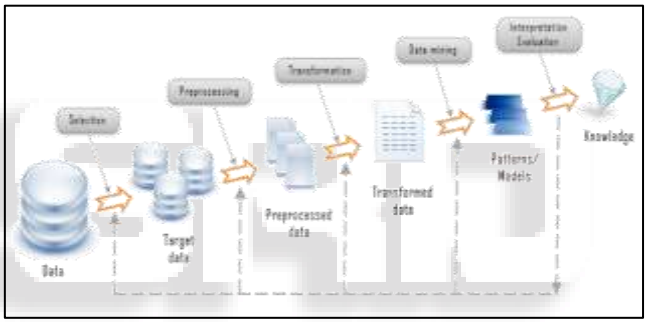

Fig. 1: Datamining Process

IV. PROPOSED SYSTEM

A. Data Collection

For this study real world data"s are collected from first year engineering students. A sample of 464 students was taken from a group of colleges. Students were grouped in a classroom they were briefed clearly about the questionnaire and it took on average half an hour to fill this questionnaire. Selection of students was at random. The primary data was collected using a questionnaire. Which include questions (i.e. with predefined options) related to several personal, socio-economic, psychological and school and college related variables that were expected to affect student performance. The questionnaire was reviewed by professionals and tested on a small set of 50 students in order to get a feedback. The final version contained 23 questions in a single A4 sheet and it was answered by more than 700 students. Latter we selected a sample of 464 from the whole. All questionnaires were filled with the response rate of 100% out of which 316 were females and 184 were males. The secondary data such as semester mark details, attendance percentage, and class test performance were collected from the college and from the directed website.

B. Feature Selection

Feature selection is a pre-processing stage used to reduce dimensionality and delete irrelevant data to increase learning accuracy and improve result comprehensibility. Irrelevant attributes may add noise to the data and also will affect the accuracy of the model. Furthermore, noise will increase the time in model building.

C. Algorithms

Although many classification models exist, only some have been selected within the scope of this study. The selected algorithms are Naive Bayesian algorithm; MLP, SMO, J48, REP tree, RANDOM tree and Decision table are used. The Naïve Bayesian model defines the classification problem with respect to probabilistic idioms, and supplies statistical methods to classify the instances based on probabilities. Multilayer preceptor is a type of artificial neural network algorithm which regards the human brain as the modeling tool. It provides a generic model for learning real, discrete and vector target values. The ability to understand the hidden model is hard and training times may be long. In decision tree algorithms, the classification process is summarized by a tree. After the model is built, it is applied to the database

V. IMPLEMENTATION

First, data cleaning was applied on the datasets. According to the missing data analysis, missing data have been removed from the datasets. Other than missing data analysis, datasets were also cleaned to remove noisy data. Unnecessary space characters or other spelling mistakes were also cleaned in the datasets. Another usual step in data pre-processing is data discretisation. Although some algorithms are said to perform better when the numerical input variables are discredited, in this study numerical variables have not been put into binned intervals in order to maintain the same conditions for all algorithms. Once the data pre-processing steps have been completed, the dataset have been used to run the classification algorithms Naïve Bayesian algorithm, MLP, SMO, J48, REP tree, RANDOM tree and Decision table. For all algorithms, splitting the data into train and test splits has been selected as the validation method. 66% of the data has been set as the training part and the rest has been set as the testing part.

Bayes' Theorem provides a way of calculating the posterior probability, P (x|c), from P(c), P (x|c), and P(x). Bayes' Theorem is:

Where

- P(c) is the prior probability of class that reflects background knowledge due to the chance of c to be correct. P(x) is the probability of x to be observed.
- P(x|c) is the probability of observing x given a world in c holds.
- P(c|x) is the posterior probability of class (target) given predictor (attribute).

$$P(c|x) = \frac{P(x|c)\,P(c)}{P(x)}$$

Naïve Bayes classifier work as follow: Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x1,x2,...,xn), depicting n measurements made on the tuple from n attributes, respectively, A1,A2,...,An. Suppose that there are m classes, C1,C2,...,Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if P(Ci|X)> P(Cj|X) for 1≤ j≤m, j6=i. Thus the P(Ci|X) need to be maximize.The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis.

VI. CONCLUSION

In this paper, we reviewed different classification method used on student database to predict the student's performance in the upcoming semester on the basis of previous student's database and the work done on this till now. As we have seen, predicting students' performance earlier is a difficult task because it is a multifaceted problem and because the available data are normally imbalanced. To resolve this problem and to improve the accuracy and quality, the Support Vector Machine algorithm can be used which is showing the greatest accuracy among other techniques. Clustering technique and ensemble cluster can also be used to fine tune the quality of resulting dataset. Information's like Attendance, Seminar and Assignment marks were collected from the student's database, to predict the performance at the mean time of the course. The other attributes are collected by students and their respective faculties who know the behavior of students. This study will help to the students and the teachers to improve the performance of the students who are at the risk of failure. This study will also work to identify those students who needed special attention to reduce fail ration and taking appropriate action for the current semester examination.

VII. FUTURE WORK

In place of future work, supposed to do the research by using various classifications and clustering applications to enhance the prediction speed and accuracy in the field of education.

REFERENCES

[1] Nkitaben Shelke, Shriniwas Gadage,"A Survey of Data Mining Approaches in Performance Analysis and Evaluation", (2015), International Journal of Advanced Research in Computer Science and Software Engineering

[2] Harwatia, Ardita Permata Alfiania, Febriana AyuWulandaria, "Mapping Student's Performance Based on Data Mining Approach (A Case Study)" ScienceDirect,Agriculture and Agricultural Science Procedia 3 (2015) 173 – 177.

[3] Renza Campagni, Donatella Merlini, Renzo Sprugnoli, Maria Cecilia Verri "Data mining models for student careers", at Science Direct Expert Systems withApplications,pp55085521,2015,www.elsevier.com.

[4] S. Archna, Dr. K. Elangovan ―Survey of Classification Techniques in Data Mining‖, International Journal of Computer Science and Mobile Applications vol 2, Issue 2,, February 2014, p.g. 65-71

[5] Rajni Jindal and Malaya Dutta Borah, "A SURVEY ON EDUCATIONAL DATA MINING AND RESEARCH

TRENDS, International Journal of Database Management Systems (IJDMS) Vol.5, No.3, June 2013.
[6] Random Forest Algorithm : Data Science Control
[7] An Overview of Data Mining Techniques Excerpted from the book Building Data Mining Applications for CRM by Alex Berson, Stephen Smith, and Kurt Thearling
[8] Advantages of Bayesian Networks in Data Mining and Knowledge Discovery By Petri Myllymäki , Ph.D., Academy Research Fellow, Complex Systems Computation Group, Helsinki Institute for Information Technology.
[9] Sajadin Sembiring, M. Zarlis, Dedy Hartama, Ramliana S, Elvi Wani "Prediction of Student Academic Performance by an Application of Data Mining Techniques" 2011 International Conference on Management and Artificial Intelligence IPEDR vol.6,pp 110-114,2011.
[10] J. K. Jothi Kalpana, K. Venkatalakshmi " Intellectual Performance Analysis of Students by Using Data Mining Techniques" ,International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014, & 2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14) On 21st&22ndMarch, Organized by K.L.N. College of Engineering, Madurai, Tamil Nadu, India ISSN (Online) : 2319 - 8753 ISSN (Print) : 2347 – 6710.pp 1922-1929, 2014.
[11] 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.

