IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A SURVEY ON DATA CLASSIFICATION AND MACHINE LEARNING FOR FORECASTING OF STUDENT PERFORMANCE

Neelam Peters*, Aakanksha S. Choubey
[*] MTech Student (CTA) Shri Shankaracharya Technical Campus, Bhilai, India
Asst. Prof. (CSE) Shri Shankaracharya Technical Campus, Bhilai, India

DOI: 10.5281/zenodo.222225

ABSTRACT

Students' academic performance is perilous for educational institutions because tactical programs can be prearranged in cultivating or maintaining enactment of the students for the duration of their period of studies in the institutions. The upsurge of student's dropout rate in higher education is one of the significant problems in most organizations. The unearthing of hidden information from the educational data system by the operative process of data mining technique to investigate factors affecting student waster can lead to a healthier academic planning and administration to moderate students drop out frequency, as well as can apprise cherished information for outcome making of policy makers to mend the quality of higher educational system. In this paper, we consider issues of factors affecting students' dropout rate, discussed about different techniques of data mining, machine learning which will predict the student performance index and what the parameters are which affects the accuracy of prediction model.

KEYWORDS: Regression;J-48;REPTree;RBTree.

INTRODUCTION

Roused by the continuous yearning to enhance the nature of training and address the constantly expanding expense of higher education by guaranteeing that students graduate within four years, data mining procedures have been progressively sent to dissect the boundless measures of chronicled information being gathered at different Colleges and Universities that relate to student academic result. One of the issues that these strategies are attempting to settle is to recognize the understudies that are at danger of coming up short a course and accordingly permit the establishment to take restorative activities by giving extra administrations and assets to the students as well as educators.

Nowadays, higher educational administrations are facing a very high viable environment and are targeting to get more modest rewards over the other business oppositions. These organizations should develop the excellence of their services and gratify their customers. They use to contemplate teachers and students as their main resources and they want to.

Henceforth there should be an a constructive analysis which helps policy makers of education system for taking quality parameters or quality decisions so as to improve the student academic performance.
In this paper we will discuss about data mining and machine learning techniques which would provide statistical analysis over the student dataset which leads to forecast the academic performance of the student. Significance of forecasting comprises following points:-

- Forecasting delivers relevant and consistent information about the past and present events and the likely future events. This is indispensable for sound planning and desired outcome.
- It gives self-reliance to the managers for making important decisions.
- It is the basis for construction planning premises.
- It keeps managers active and alert to face the challenges of future events and the changes in the environment.

There are some limitations of prediction model which encompasses following points:-

- The gathering and analysis of historical data, present and future implicates a lot of time and cost. Consequently, managers have to balance the cost of prediction with its benefits. Most small organizations don't do involve in forecasting because of the high cost.
- Forecasting can only guesstimate the future events. It cannot make assurance that these events will take place in the future. Long-term predictions will be less truthful as compared to short-term forecast.
- Forecasting is based on certain conventions. If these conventions are wrong, the forecasting will be wrong. Forecasting is completely depend on historical events. However, history may not repeat itself at all times.
- Forecasting requires proper judgment and skills on the part of managers. Forecasts may go erroneous due to wicked judgment and skills on the part of some of the managers. Consequently, forecasts are subject to human error.

Figure-1 demonstrate how prediction over data set works in different phases.

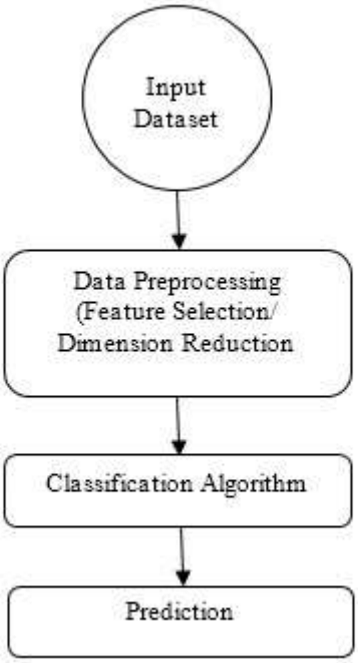

Fig.-1 Prediction Model

MOTIVATION

These days, the Institutions of Higher Learning database contains such a great amount of data about their understudies. The data is continued expanding by times, however there is no move made to pick up information from it. Data mining and machine learning are the reasonable systems in dealing with the student database, information to find new data and learning about understudies.

Correct prediction of student performance is significant in academic management due to its noteworthy impacts in an assortment of aspects. This attempt to examine and suggest a computer platform which can be used in enduring for higher studies comes very useful. It's recognized that the system is with no trouble scalable to include supplementary knowledge for different education institutes. Forecasting of student performance can be useful for:-

- Assessing the student result.
- In which area student is lagging behind.
- Predicting student result.
- Limiting the impact of number of students pass and fail ratio.
- Student performance deviation.
- Identification of factors which would not affect student performance.

LITERATURE SURVEY

While studying about prediction we have gone through different literatures, some authors concluded as:
P. Sunil Kumar et. al. [IJACECT 2013] concluded that Association rules are useful t o find the association between two elements and shows relationship between them. In this paper seven different parameters are used to find the relationship between two different factors affecting the school dropout. From the above analysis it can be concluded that the students who are disinterested are more prone to dropouts than due to Poverty and Teaching Environment. Another conclusion is extracted from confidence, cosine, AV analysis, lift, correlation and conviction analysis is that most of the Poor students are disinterested as well as students not satisfied with the teaching environment are also not interested to continue their education. So as desired by the study made by ASSOCHAM, government of Odisha has to take necessary steps to raise the expenditure level per family per month and also organize awareness camps to increase the interest level in the young minds.

Hina Gulati [IEEE 2015] concluded that Predicting student's dropout reasons can be difficult task due to multiple factors that can affect the decision. Moreover the complete procedure is long and time consuming. Data gathered is from different sources and need data preprocessing to be done first to clean and balance data. In preprocessing step feature selection algorithms are used to identify features that will affect the prediction process the most. It can be seen many factors like demographic factors, socio-economic factors, family factors, etc. can be responsible for students from dropping out from there courses. After preparing data for mining classification algorithms are applied and by analysis of decision tree and induction rules we get the prediction model that is tested on test data can help to find useful knowledge. Result obtained from such models can help teachers and management to identify the problem areas and reasons that affect dropout the most. We have considered three cases and accuracy when compared for classification in all three cases will lead to understanding that which is most effective way to analyze student performance and help in identifying reasons for drop-out. Author presents analysis of data set using data mining algorithms. After analysis the outcome will be the major factors that affect student dropping out of the open courses the most (dropout rate). Before applying classification algorithms some feature selection algorithms are also used so as to get refined prediction results. Such analysis and prediction information will help college management and teachers to make necessary changes for imparting better education. Mining of useful knowledge can be done by using many other mining techniques like association, clustering. Tool used for feature selection and mining is weka.

Ashish Dutt et. al. [IJIEE 2015] concluded that the application of data mining methods in the educational sector is an interesting phenomenon. It sets to uncover the previously hidden data to meaningful information that could be used for both strategic as well as learning gains. In this author have detailed the various disparate entities that are widely spread across in the educational foray. However, collectively they have not been addressed and this paper serves to bridge this gap. We would continue to pursue our research in clustering algorithms as applied to educational context and will also be working towards generating a unified clustering approach such that it could easily be applied to any educational institutional dataset without any much overhead.

Wilairat Yathongchai et. al. [LAET 2015] discussed that factors Analysis in Higher Educational Student's Drop Out is an important. In this paper we presented the effectiveness of classification techniques (J48 and Naïve Bayes algorithms) on the data set used from the database of Academic MIS at BRU. Sample data were faculty of science. The three issues of factors analysis affecting to student drop out are: factors related to the student before admission, factors related to the students during the study periods in the university, and all factors. Our experimental results are shown as the rules that transformed from decision tree by accuracy value between 75% and 88%. Based on the three issues analysis, we found the fundamental factors about student before admission to planning to qualify for admission. The knowledge about students during the study periods in the university factors can use for academic planning to improve the quality of students.

S. No.	Author Name/Tile/Publication/Year	Algorithm Used	Description
1.	Hina Gulati/ Predictive Analytics Using Data Mining Technique/IEEE/2015	Decision Tree Algorithm (J48, REPTree, NB Tree)	Author presents analysis of data set using data mining algorithms. After analysis the outcome will be the major factors that affect student dropping out of the open courses the most (dropout rate). Before applying classification algorithms some feature selection algorithms are also used so as to get refined prediction results. In future to increase the accuracy of prediction data must be tuned and need to improve efficiency of classification algorithms.
2.	Mr. M. N. Quadri et. al./ Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques/ GJCST/ 2010	Weka J48 Classifier	Author introduced the data mining approach to modeling drop out feature and some implementation of this approach The key to gaining a competitive advantage in the educational industry is found in recognizing that student databases, if properly managed, analyzed and exploited, are unique, valuable assets. Data mining uses predictive modeling, database segmentation, market basket analysis and combinations to more quickly answer questions with greater accuracy.
3.	Wilairat Yathongchai et. al. / Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out/ LAET / 2015	Naïve Bayes classifier, 10-fold cross validation	Author presented the effectiveness of classification techniques (J48 and Naïve Bayes algorithms) on the data set used from the database of Academic MIS at BRU. Sample data were faculty of science. The three issues of factors analysis affecting to student drop out are: factors related to the student before admission, factors related to the students during the study periods in the university, and all factors. Our experimental results are shown as the rules that transformed from decision tree by accuracy value between 75% and 88%.
4.	Asmaa Elbadrawy et. al./Personalized Multi-Regression Models for Predicting Students' Performance in Course Activities/ACM/2015	Multi-regression model	Author used a multi-regression model to predict student performance in course activities and analyze the resulting student populations. Author have shown that a multiregression model performs better than single linear regression as it captures personal student differences through the student-specific membership weights. We have also shown that the RMSE tends to decrease with increasing the number of linear regression models and thus allowing room for more personalized predictions. We have also shown that using the Moodle interaction features lead to an improved prediction accuracy.
5.	Shreyansh Kakadiya et. al./ Analyzing Start-up Success Possibility Using Data Mining Technique/ IJIACS/ 2015	Decision-tree algorithm	Author proposes a method that analysis attributes of successful startup and finds out the most important attributes using data mining technique. The attributes obtained as a result can then be compared with the new ventures to find out its success possibility.

6.	Dr. Mamta Madan et. al./ A Review on: Data Mining for Telecom Customer Churn Management/ IJARCSSE /2015	Discussed Support Vector Machine, Logistic Regression, Decision Tree	Author aims at reviewing the recent literature in the area of telecom customer churn mainly with two perspectives, i.e. technique being applied to telecom churn prediction and the publication year. Also the aim is to help the researchers to gain insight into the recent trends in this area, which will guide them in finding the possible reasons for the churn and consequently reducing them and helping the telecom sector to reduce their financial loss.

REGRESSION TECHNIQUES

Regression analysis is a procedure of predictive modelling technique which inspects the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for predicting, time series modelling and finding the casual effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression. Regression analysis is an important tool for modelling and analyzing data. There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the significant relationships between dependent variable and independent variable.
2. It indicates the strength of impact of multiple independent variables on a dependent variable. Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

There are different types of regression techniques:

Linear Regression: It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete and nature of regression line is linear. Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line). It is represented by an equation

$Y=a+b*X + e$

Where "a" is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s). In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables.

Logistic Regression: Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represented by following equation.

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrence

$\ln(odds) = \ln(p/(1-p))logit(p) = \ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk$

Above, p is the probability of presence of the characteristic of interest.

Polynomial Regression: A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

$y=a+b*x^2$

CONCLUSION

Nowadays, higher educational administrations are facing a very high viable environment and are targeting to get more modest rewards over the other business oppositions. These organizations should develop the excellence of their services and gratify their customers. Henceforth there should be an a constructive analysis which helps policy makers of education system for taking quality parameters or quality decisions so as to improve the student academic performance. In this paper we have discussed about data mining and machine learning techniques which would provide statistical analysis over the student dataset which leads to forecast the academic performance of the student. In section IV we have discussed some regression techniques which is technique of machine learning.

After studying different literature we found that applying clustering data mining technique and by use of machine learning predictive model we can increase the accuracy of prediction. Hina Gulati IEEE 2015 concluded that there is loss of efficiency during classification of input data henceforth efficient classification method will increase the accuracy of prediction. From different literatures we can draw comparison based on accuracy of numerous predictive modelling system which uses different types of classifier as J48, Naïve Bayes, REPTree, NBTree, etc.

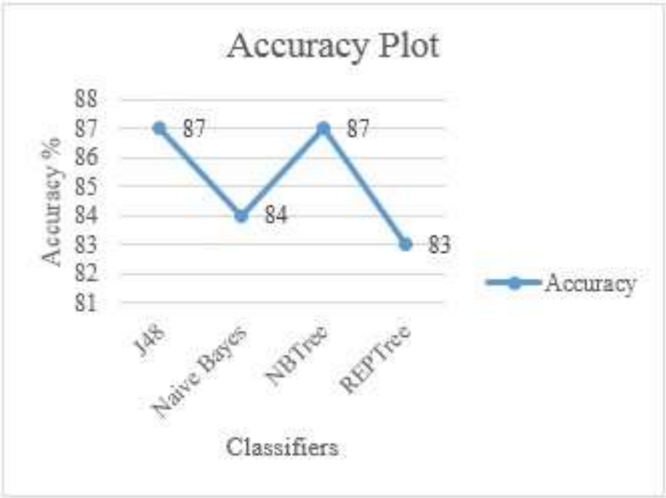

Fig.-2 Accuracy Plot among different classifier technique used by literatures

RREFERENCES

[1] C. Márquez-Vera, C.R.Morales, and S.V.Soto,"Predicting School Failure and Dropout by Using Data Mining Techniques", IEEE journal of Latin-American learning technologies, vol. 8, no. 1, February 2013, pp.7-14.

[2] M. N. Quadri1, N.V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques", Global Journal of Computer Science and Technology, Vol. 10 Issue 2 (Ver 1.0), April 2010, pp. 2-5.

[3] M. Nasiri, B. Minaei, F. Vafaei, "Predicting GPA and Academic Dismissal in LMS Using Educational Data Mining: A Case Mining" ,IEEE, 6th National and 3rd International conference of e-Learning and eTeaching(ICELET),2012,pp.53-58.

[4] W. Yathongchai, C. Yathongchai, K. Kerdprasop, N. Kerdprasop, "Factor Analysis with Data Mining Technique in Higher Educational Student Drop Out", Latest Advances in Educational Technologies, 2003.

[5] P. S. Kumar, A. K. Panda, D. Jena, "Mining the factors affecting the high school dropouts in rural areas", International Journal of Advanced Computer Engineering and Communication Technology (IJACECT), Volume-2, Issue – 3, 2013, pp.1-6.

[6] E. Yom-Tov, G.F. Inbar,"Feature Selection for the Classification of Movements From Single MovementRelated Potentials", IEEE Transactions on neural systems and rehabilitation engineering, vol. 10, no. 3, September 2002,pp. 170-177.

[7] M. S. Chen, J. Han, P. S, Yu, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, December 1996.

[8] Y. Kurniawan, E. Halim, "Use Data Warehouse and Data Mining to Predict Student Academic Performance in Schools: A Case Study (Perspective Application and Benefits)", IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), August 2013,pp.98-103.

[9] M. Wook, Y. Hani Yahaya, N. Wahab, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques", IEEE, Second International Conference on Computer and Electrical Engineering,2009,pp.357-361.

[10] E. Gharavi, M. J. Tarokh, “Predicting customers' future demand using data mining analysis: A case study of wireless communication customer”, IEEE,5th Conference on Information and Knowledge Technology,2013,pp.338- 343.

[11] Hina Gulati “Predictive Analytics Using Data Mining Technique” 978-9-3805-4416-8/15/$31.00 c 2015 IEEE.

[12] P. Sunil Kumar, Ashok Kumar Panda, D. Jena "Mining The Factors Affecting The High School Dropouts In Rural Areas" ISSN (Print): 2278-5140, Volume-2, Issue – 3, 2013.

[13] Azwa Abdul Aziz, Nur Hafieza Ismail, Fadhilah Ahmad "Mining Students’ Academic Performance" ournal of Theoretical and Applied Information Technology 31st July 2013. Vol. 53 No.3.