IJRSET AUGUST 2017 Volume 4, Issue 7 Pages: 46-54

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

A REAL TIME DATA MINING MODEL TO PREDICT ACADEMIC ATTRITION

[1] Vishal Mittal, [2] Anuradha
[1, 2] Department of Computer Science & Engineering,
[1, 2] Shree Siddivinayak Group of Institutions, Bilaspur,Yamuna Nagar,Haryana

ABSTRACT: Quality of education system is very important for a country growth. Today education sector is facing challenges, the major challenges of higher education being decrease in students success rate and their leaving a course without completion. An early prediction of student's failure can avoid poor performance, which will help to enhance their performance. It can help not only the current students but also the future students to predict thier performance. Data mining provides powerful techniques to analysis student performance. For this purpose, In this dissertation various educational data mining techniques have been used such as Naive Bayes, Decision Tree, K-Nearest Neighbour, Random Forest, Rpart , C5.0 to build a model for academic attrition based on students social integration, academic integration and various emotional skills considered. In order to future through data mining techniques data was collected from mullana university. Data from the admission process are complemented with the academic information that is gathered for each academic period; however, the causes of low academic performance occur on day-to-day basis and waiting until the academic period ends could be crucial. This leads to think that new, and possibly, non –traditional ways, for collecting information close to real time are needed. In this dissertation new attributes are identified which represent real time student academic attrition. The implementation of different data mining techniques is done on R language develop at the university of Auckland,New Zealand. It is an open source language. It is an interactive language used for easy input,output and large data manipulation and used for various statistical analysis and modelling. Many classification and regression algorithms are used, which are attribute dependent.some are used on categorical, nominal data others on numerical data. The experimental results are validated against test data and interesting co-relations are observed. The comparison of their accuracy is done to find the most accurate predictions. Graphs are also used for illustrative comparison, along with numerical values.

Keywords: [R language; data mining; attributes]

1. INTRODUCTION

Education means obtaining knowledge. A person who has knowledge of his surrounding can survive happily in the society. To get acquired with it, people join educational institutes, were time and money both being extremely precious thing themselves, must be spent efficiently. Higher education has gained important

manifolds in the past few decades. The higher education institute is forced to revise its scope and objects because of the private participation. Universities today, similar to business organizations, are operating in a very dynamic and strongly competitive environment. The education globalization leads to more and better opportunities for students to receive high quality education at institutes all over the world. [1].For higher education institution whose goal is to contribute to the improvement of quality of higher education, the success of creation of human capital is the subject of a continuous analysis. Therefore, the prediction of student's success is crucial for higher education institutions, because the quality of teaching process is the ability to meet student's needs. In the sense important data and information are gathered on a regular basis, and they are considered at the appropriate authorities, and standards in order to maintain the quality are set. The quality of higher education institutions implies providing the services, which most likely meet the needs of students, academic staff, and other participants in the education system, the participants in the educational process, by fulfilling their obligations through appropriate activities, create an enormous amount of data which needs to be collected and then integrated and utilized. By converting this data into knowledge, the gratification of all participates is attained: students, professors, administration, supporting administration, and social community. [2]

All participants in the educational process could benefit by applying data mining on the data from the higher education system (figure 1.1). Since data mining represents the computational data process from different perspectives, with the goal of extracting implicit and interesting samples (written and frank, 2000), trends and information from the data, it can greatly help every participant in the educational process in order to improve the understanding of the teaching process, and it centres on discovering, detecting and explaining educational phenomenon's (EI-Halees, 2008) [3]

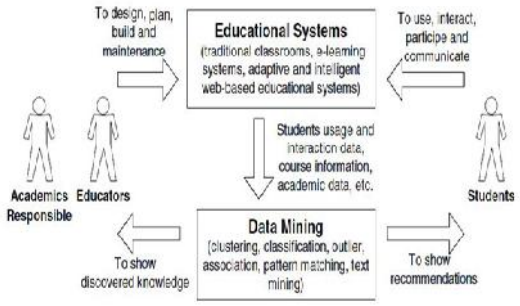

Figure- 1.1 The cycle of applying data mining in education system [7]

So with data mining techniques, the cycle is built in educational system which consists of forming hypotheses, testing and training.Thus, application of data mining in educational systems can be directed to support the specific needs of each of the participants in the educational process. The student is required to recommended additional activities, teaching materials and tasks that would favour and improve his/her learning.

2. PROBLEM FORMULATION

Performance of the students should be calculated on regular basics so that the results can be improved and the corrective action can be taken at right time. Therefore, it is very necessary that the day to day activities should be taken into consideration along with marks or grades in each semester and other factors.In this research work proposes an effective methodology for measuring the pivotal causes which can be related to student's performance at regular basis and prediction of student academic attrition. Different results of the classification algorithms of data mining are analyzed and compared based on the accuracy, precision, recall and specificity of the model. The outcome results can be stored as intelligent knowledge for decision making to improve the quality of education in ins1titutions. This stored knowledge is used for predicting the student's academic attrition due to low academic performance in advance.

3. METHODOLOGY

This section introduces the Data mining model to predict the student academic attrition due to low academic performance,

which uses the student data; socioeconomic, demographic, initial academic information and the academic records of previous academic periods as well as recurrent information of day to day activities of students. The model based on the framework given below in fig. 3.1

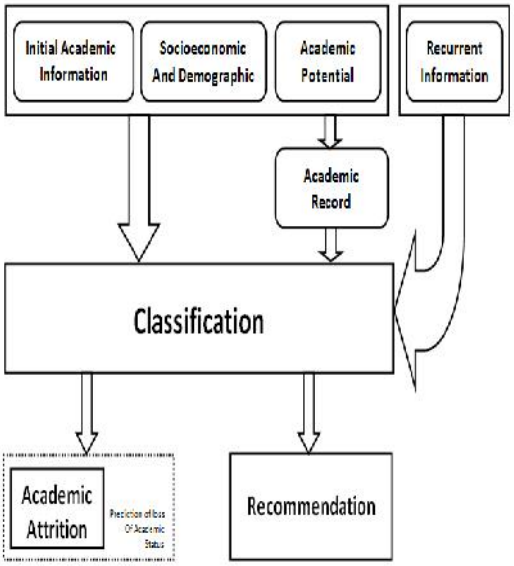

Figure- 3.1 Real Time Data Mining Model

This model uses the factors, which affects on the student's academic attrition and helps students to understand their learning status through the predicted academic attrition. The teachers also can view the predicted student's academic attrition and help earlier in identifying weak students who need special attention and give the required recommendations according to student's current performance. Through extensive search of literature and discussion with experts on student performance, a number of factors that have influence on the Performance of a student are identifying.

4. WORKFLOW STEPS

The main steps of the workflow are as explained below:

(A) Data fetching

Data fetching combines all the available data that can be used to resolve the data mining problem, into a set of instances. The data of our work is fetched from M.M University students. Each and every detail of student such as socioeconomic, demographic, initial academic information and the academic records of previous academic periods as well as recurrent information of day to day activities of students are taken for the efficient study of the approach. Intermittent Information is also collected on regular basis so that it can be applied to make an efficient data mining model.

(B) Data Preprocessing

Data pre - processing is an important step in the data mining process. Its involves following steps.

(i) Data Cleaning

The data cleaning process detects erroneous or irrelevant data and discards it. The data collected by us also had most common mistakes like inaccuracies, missing data and inconsistent data. Some of the attributes like tagging for, program code, section branch, semester type, wake up time of the student, mode of transportation and dressing sense of a student etc were irrelevant for our study. Therefore, they were cleaned from the data so that the main attribute could be used.

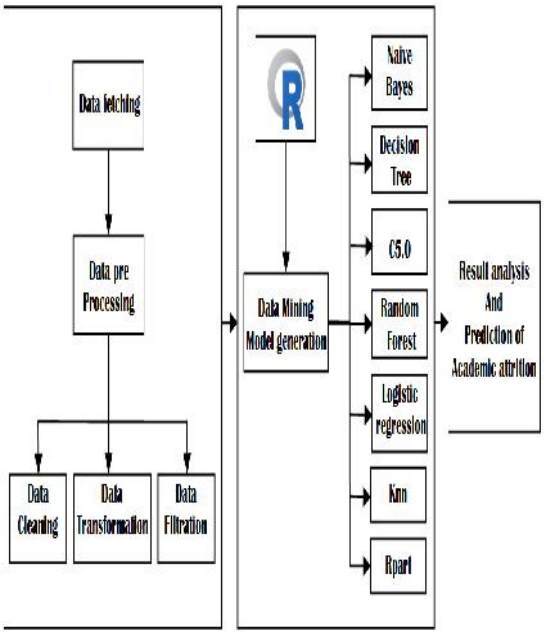

Figure- 3.2 Steps of the Workflow

(ii) Data Transformation

Data transformation is the process of deriving new attributes from beforehand available attributes to assist a better interpretation of information. The data is

separated according to the subjects and the grades obtained in those particular subjects. In addition, various formats were formed where the information is specific in accordance with the subjects, students, grade and semester

(iii) Data Filtering

Data filtering helps to reduce the large amount of information available to us. The most common types of filtering techniques are usually the selection of data subsets for educational data relevant to the intended reason. Huge amount of academic information is available at educational institute as all the subjects and the grades attained accomplished by the students, name and the enrolment number admitted in the course. However, we were only interested on a few subsets of courses or students depending on the proposed approach. For this reason, filtering is be used to select only a particular subset of desired data.

5. IMPLEMENTATION AND RESULTS

Data mining is the extraction of information from large data sets and transformation of that information into some understandable structure for further use. It is the process of selecting, exploring and modeling large amount of data by using different techniques to find useful patterns or models. Proper data mining technique should be available so that it can be used in many applications such as, social science, bank transactions, businesses, and psychology. In order to make this possible, proper data mining techniques should be available. As the datasets grow in terms of complexity and size, some automated techniques comes into picture. Data mining exploits actual learning, discovers algorithms that are more efficient, and allows such methods to be applied to larger data sets. Data mining process is applied with the intention of uncovering hidden patterns in large data warehouses [50]. The purpose of applying any data mining effort can be divided in two types: to generate descriptive models to solve problems and can be used to predict and solve problems.

As the datasets grow in terms of complexity and size, some automated techniques comes into picture. Data mining exploits actual learning, discovers algorithms that are more efficient, and allows such methods to be applied to larger data sets. Data mining process is applied with the intention of uncovering hidden patterns in large data warehouses. The purpose of applying any data mining effort can be divided in two types: to generate descriptive models to solve problems and can be used to predict and solve problems.

6. IMPLEMENTATION METHODOLOGY

In this work, multiple classification algorithms were used on regular interval of time to predict the student academic attrition due to low academic performance. Different algorithms work differently, and hence they have varied range of accuracy. Their accuracy depends on many factors like training dataset, type of algorithm, independent attribute has and values. This approach was used because it can provide a broader look and understanding of the results and output as well as it will lead to a comparative conclusion over the outcomes of the work.

For this study, the dataset was tested with five different classification algorithms: Naive Bayes, Decision Tree, K-Nearest Neighbour, Random Forest, and C5.0. All classification algorithms were executed on several input attributes. Sixty percent of dataset is used for training the model and remaining forty percent was used for model testing.

During the implementation phase, the algorithms for building models that would classify the student in to two classes – Fail, Pass, depending on their current Academic performance and based of the student information: socioeconomic, demographic, previous academic data coming from admissions as well as recurrent information of day to day activities of the student. Each student record contain several input attributes as shown in Table 4.2

Type	Attributes
Initial Academic information	ID No, School Type, Program type, Type of Access, PPC, Failure
Demographic And Socioeconomic	Age ,Sex, Address, Pstatus, Medu, Fedu, Mjob, Fjob, Famsize, Guardian
Academic Potential	Ad test score, FEGrade, Drop
Recurrent Information	Attendance, G1, G2, G3, CCR, Teachers feedback, Health, Travelling time, Free time, Internet, Activities, Study time, Gout, Dalc, Walc, Paid

Table 6.1

In this dissertation different experiments are carried out to get the objectives. The one objective was evaluating student's informational data to determine the student performance and predicting the academic attrition of student due to low performance on regular basis.

The objective was to implement and compare the various algorithms which one uses the attributes value at its best and thus perform better in term of accuracy. The performance of classification model is measured by evaluating the correctness of the classification decision of the algorithms. For the purpose of evolution of the accuracy of the algorithms the following terms has been used in matrix.

1)TP : Number of True Positives (Algorithm correctly labelled record as Positive).

2)TN : Number of True Negatives (Algorithm correctly labelled record as Negative).

3)FP : Number of False Positives (Algorithm incorrectly labelled record as positive).

4)TP : Number of False Negative (Algorithm incorrectly labelled record as negative).

To evaluate and Compare Algorithms performance we used accuracy, precision, recall, and specificity.

Accuracy: - Proportion of total number of correct prediction.

$$\frac{TP+TN}{P+N}$$

Precision: - Proportion of correct positive observations.

$$\frac{TP}{TP+FP}$$

Recall: - Proportion of positives correctly predicted as positive.

$$\frac{TP}{P}$$

Specificity: - Proportion of negatives correctly predicted as negative.

$$\frac{TN}{N}$$

Implementation and Results of Random forest algorithm during regular interval of time

Random Forest:- Random forest are an ensemble learning method that operate by constructing a multitude of decision tree at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forest correct for decision tree habits of over fitting to their training set.

Fig. 4.1 shows Result of random forest algorithm in R language after first midterm, second midterm, third midterm and after final exam.

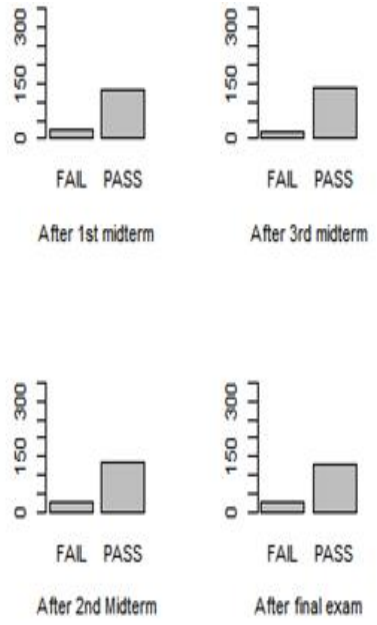

Figure - 6.2 Results of forest algorithm

Table 4.2 shows total result of Random forest algorithm in respect of accuracy, precision, recall and specificity that were taken on regular interval of time.

Random forest	Accuracy	Precision	Recall	Specificity
AfterFirst midterm	84.18	90.15	90.84	51.85
After Second midterm	87.34	91.73	93.13	59.26
After Third midterm	88.61	91.24	95.42	55.56
After final exams	89.24	94.53	92.37	74.07

Table 6.2: Results of Random Forest Algorithm using Attributes

CONCLUSION AND FUTURE SCOPE

Predicting Student Academic Attrition is most useful to help the teachers and student for improving their learning and teaching process. This work has reviewed previous studies on predicting student performance with various analytical methods. five algorithms Random forest, C5.0, Naive bayes, Decision tree, KNN were applied on student dataset. Result shows that after the comparison of this five algorithms on output attributes like Accuracy, Precision, Recall, Specificity the Random forest algorithm perform best for designed model by obtaining the overall prediction accuracy of 89.24, precision 94.53, recall 92.37 and Specificity 74.07 This work assist teacher to early detect student who is expected to fail the course. Teacher can provide special recommendation to those students and help them to improve their academic performance. In the current work, it was found that the performance of student's is not very dependent on their academic effort. In spite, there are many other non academic factors as well as day to day activities that have equal to greater influence on student performance and it vary from one institution to other institution, one country to other country and one culture to other culture. Teacher role is important in this regards. They have to be more interactive with student to, provide proper recommendation and motivate the student. In present work the role of recommendation is limited for alerting the student for improving their current poor performance

In future this work can be carried out with more experiment with bigger dataset by finding various factors and attributes. Addition of new attributes would increase the capability of model to predict the performance of students at interuniversity, country and worldwide level and role of recommendation would be increased by finding suitable and best recommendation according to student current situation.

REFERENCES

[1]. Ognjanovic, D. Gasevic and S. Dawson, "Using institutional data to predict student course selections in higher education," The Internet and Higher Education, vol. 29, pp. 49–62, Apr. 2016.

[2]. D. P. Nithya, B. Umamaheswari, A. Umadevi, "A survey on educational data mining in field of education," International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 5, no. 1, pp. 69–78, Jan. 2016.

[3]. Amjad Abu Saa, "Educational Data Mining & Students' Performance Prediction" International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016

[4]. Kamaljit Kaur and Kuljit Kaur, "Analyzing the Effect of Difficulty Level of a Course on Students Performance Prediction using Data Mining" 2015 Ist International Conference on Next Generation Computing Technologies(NGCT-2015) Dehradun, India, 4-5 september 2015.

[5]. A. M. Shahiri, W. Husain, N. A. Rashid, "A review on predicting student's performance using data mining techniques," Procedia Computer Science, vol. 72, pp. 414–422, 2015.

[6]. Sagar S. Nikam, "A comparative study of classification techniques in data mining" Oriental journal of computer science & technology-vol 8 no(1) pgs 13-19 issue april 2015

[7]. Rutvija Pandya and Jayati Pandya, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning" International Journal of Computer Applications (0975 – 8887) Volume 117 – No. 16, May 2015.

[8]. S. Venkata Krishna Kumar and P. Kiruthika, "An Overview of Classification Algorithm in Data mining" International Journal of Advanced Research in Computer

and Communication Engineering Vol. 4, Issue 12, December 2015
[9]. M. Goga, S. Kuyoro, N. Goga, "A Recommender for improving the student academic performance", Procedia - Social and Behavioral Sciences, vol. 180, pp. 1481–1488, May 2015.
[10]. W. Xing, R. Guo, E. Petakovic, S. Goggins, "Participation-based student final performance prediction model through interpretable genetic programming: Integrating learning analytics, educational data mining and theory", Computers in Human Behaviour, vol. 47, pp. 168–181, Jun. 2015.
[11]. K. Kaur and K. Kaur, "Analyzing the effect of difficulty level of a course on students performance prediction using data mining", Next Generation Computing Technologies (NGCT), 2015 1st International Conference, pp. 756-761, 2015.
[12]. Harwati, A. P. Alfiani, and F. A. Wulandari, "Mapping student's performance based on data mining approach (A case study)," Agriculture and Agricultural Science Procedia, vol. 3, pp. 173–177, 2015.
[13]. K. Parmar, D. Vaghela & P. Sharma, "Performance prediction of students using distributed Data mining." In Innovations in Information, Embedded andCommunication Systems (ICIIECS), 2015 International Conference on, pp. 1-5, March, 2015.
[14]. R. Campagni, D. Merlini, R. Sprugnoli, and M. C. Verri, "Data mining models for student careers," Expert Systems with Applications, vol. 42, no. 13, pp. 5508–5521, Aug. 2015.
[15]. P. S. Pradnya, "Overview of predictive and descriptive data mining techniques",International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5,no. 4, April-2015.
[16]. Mariammal.D,Jayanthi.S and Dr.P.S.K.Patra, "Classification Methods in Data Mining:A Detailed Survey" International Journal of Research in Computer and Communication Technology, Vol 3, Issue 4, April- 2014.
[17]. S. Natek, M. Zwilling, "Student data mining solution–knowledge management system related to higher education institutions", Expert Systems with Applications, vol. 41, no. 14, pp. 6400–6407, Oct. 2014.
[18]. R. Asif, A. Merceron, and M. K. Pathan, "Predicting student academic performance at degree level: A case study", International Journal of Intelligent Systems and Applications, vol. 7, no. 1, pp. 49–61, Dec. 2014.
[19]. J.M. Mativo and S. Huang, "Prediction of students' academic performance Adapt a methodology of predictive modelling For as mall sample size."In Frontiers in Education Conference (FIE), 2014 IEEE, pp. 1-3, Oct. 2014.
[20]. T. Mishra, D. Kumar & D.S.Gupta, "Mining Students' Data for Performance Prediction." In Proceedings of International Conference on Advanced Computing & Communication Technologies, pp. 255-263, 2014.
[21]. K. Parmar, D. Vaghela & P. Sharma, "Prediction and Analysis of Student Performance using Distributed Data mining." International journal of emerging technologies and applications in engineering, technology and sciences (ij-eta-ets), Dec. 2014.
[22]. P. Guleria, N. Thakur, M. Sood, "Predicting student performance using decision tree classifiers and information gain," Parallel, Distributed and Grid Computing (PDGC), International Conference on, Solan, pp. 126-129, 2014.
[23]. T. Mishra, D. Kumar & D.S.Gupta, "Mining Students' Data for Performance Prediction." In Proceedings of International Conference on Advanced Computing & Communication Technologies, pp. 255-263, 2014.
[24]. P.A. Patil & R.V. Mane, "Prediction of Students Performance Using Frequent Pattern Tree." In Computational Intelligence and Communication Networks (CICN), 2014 International Conference on , pp. 1078-1082, Nov. 2014.
[25]. Priyanka Saini and Ajit Kumar Jain, "Prediction using Classification Technique for the Students' Enrolment Process in Higher Educational Institutions" International Journal of Computer

Applications (0975 – 8887)Volume 84 – No 14, December 2013.
[26]. D. Solomatine, L.M. See and R.J. Abrahart "Data-Driven Modelling: Concepts, Approaches and Experiences" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 2, august 2013.
[27]. D. Suthers, K. Verbert, E. Duval, X. Ochoa, "Clow, MOOCs and the funnel of participation", In: International Conference on Learning Analytics and Knowledge, pp. 185–189. ACM New York, 2013.
[28]. V. Ramesh, P. Parkav, K. Rama, "Predicting student performance: A statistical and data mining", International Journal of Computer Applications, vol. 63, no. 8, pp. 35-39, 2013.
[29]. P.M. Arsad, N. Buniyamin & J.L.A Manan, "A neural network students' performance prediction model (NNSPPM)." In Smart Instrumentation, Measurement and Applications (ICSIMA), 2013 IEEE International Conference on, pp. 1-5, Nov. 2013.
[30]. Prof. R. A. Gangurde and Prof. M. R. Sonar, "Knowledge Extraction using Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 2, February 2013.
[31]. S.Neelamegam and Dr.E.Ramaraj, "Classification algorithm in Data mining: An Overview" International Journal of P2P Network Trends and Technology (IJPTT) – Volume 4 Issue 8- Sep 2013.
[32]. E. Osmanbegović, M. Suljić, "Data mining approach for predicting student performance", Economic Review – Journal of Economics and Business, vol., no. 1, pp. 3-12, 2012.
[33]. K. Bunkar, U.K. Singh, B. Pandya & R. Bunkar, "Data mining: Prediction for performance improvement of graduate students using classification." In Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on, pp. 1-5, Sept. 2012.
[34]. S. Huang & N. Fang, "Work in progress: Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques." In Frontiers in Education Conference (FIE), 2012 , pp. 1-2, Oct. 2012.
[35]. Edin Osmanbegović and Mirza Suljic, "Data mining approach for predicting student performance" Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.