Knowledge Discovery with Data Mining for Predicting Students' Success Factors in Tertiary Education System in Sri Lanka

K. T. S. Kasthuriarachchi[1], S. R. Liyanage[2]

[1]Faculty of Graduate Studies, University of Kelaniya, Dalugama, Sri Lanka

[2] Faculty of Computing and Technology, University of Kelaniya, Dalugama, Sri Lanka.

Abstract

Knowledge discovery in educational data would be so basic to determine better expectations on the undergraduates. Distinguishing proof of the components influence to the execution of undergraduates in light of various attributes will be supportive for instructors, educators and managers viewpoints. This paper endeavors to utilize different data mining ways to deal with find forecast manages in undergraduates' data to distinguish the components influence to the scholarly accomplishment in their tertiary education. The approach of this exploration observed the aftereffects of three mining algorithms with about 3800 undergraduates' records and the calculation which demonstrated the most elevated exactness has chosen as the best model and the connections acquired through that were gotten to foresee various elements against the objective of whether they will get the degree or not following three years of the university life. Naïve Bayes, Decision Tree and Support Vector Machine were used in predicting the most affecting factors to the performance of students. According to the prediction accuracy levels, the results of Decision Tree were selected since it outperforms the rest for the selected data set. Finally, the results were evaluated using a correlation analysis to select the most prominent factor. According to the test, the age, past failure modules, performance of past semesters were selected as the most influencing factors to the success or failure of the students in tertiary education system in Sri Lanka.

Keywords: *Educational Data Mining, Algorithms, Knowledge Discovery, Feature Extraction, Validation*

1. Introduction

Education is a key zone for accomplishing a long haul financial advance. Amid the most recent decades, innovative education in Sri Lankan tertiary education framework has enhanced with the blossom of mechanical establishments, which drives the undergraduates towards the degree level. In Sri Lanka, the tertiary education is proving by both state universities and degree awarding institutes. The University Grant Commission (UGC) was established to allocate funds to the universities and university institutes, serves as the central admission agency for undergraduate studies in universities, planning and monitoring of academic activities of the university system in order to maintain academic standards and implement national policies with respect to university education in Sri Lanka. The UGC selects students for admission to undergraduate courses based on the z-score [1]. The degree awarding institutes are recruiting students based on the performance of the selection test that they conduct and never worried about z-score. However, the majority of the students who enroll to institutes for their degree are having the pass mark to the Advanced Level examination, which is the minimum criteria for the university education. However, the statistics grabbed from institutes have proven that, only few students are completing the degree due to various reasons.

Data Mining (DM) / Business Intelligence (BI) which are very interesting methods used to predict what are the main reasons for student' failures?, what are the factors which affect mostly to the success of academic activities? and many more. Educational Data Mining (EDM) is a creating region of data mining, which the use of data mining strategies to a particular kind of data that chose from educational airs to address vital educational inquiries. There is a boundless uses of EDM to examine the conduct of data gathered from infant education, primary education, secondary education, higher education, auxiliary training, advanced education and option instruction have a place with conventional training conditions. Computer based educational condition incorporates Learning Management Systems (LMS), Intelligent Tutoring Systems (ITS), test and quiz frameworks, recreations and wikis discussions which use for e-learning and ementoring, online guideline frameworks and so forth. [2].

Predicting the factors, which affect to student performance, is very important for educators, educational decision makers and even for students. It would be a great benefit to educators and course designers to identify the characteristics of the students who have to be enrolled to the degree program if, a quality educational outcome is expecting. Students also be able to get away from the negative criteria to reach to the maximum level of their education by these predicted outputs.

This research study is focusing to identify the factors, which affect to student performance in their tertiary education. The paper is organized as follows. First, discusses the background of EDM. Then, the materials and methods used to mine educational data to prediction are described. Then, the results of the analysis is discussed.

2. Background

Knowledge discovery and data mining can be thought of as apparatuses for basic leadership and authoritative adequacy. It has its root in machine learning, computerized reasoning, software engineering, and measurements. There are different data mining strategies and methodologies, for example, classification, cluster and association [3]. Each of these methodologies can be utilized to quantitatively examine huge data sets to discover shrouded significance and examples. While data mining has been connected in an assortment of enterprises, government, military, retail, banking, telecommunication and healthcare services, data mining has not gotten much consideration in educational setting. Educational data mining is a field of concentrate that examinations and applies data mining to take care of educational related issues.

Educational data mining characterized scholastic examination as the utilization of factual methods and data mining in ways that assistance workforce and guides turn out to be more proactive in recognizing at-risk students and comparing in like manner. It would fairly supportive in enhancing understudy maintenance, advancement of recommender framework, overseeing new courses and so forth. Data mining focuses on identifying the useful insights of improving student success and processes directly related to student learning, build the patterns of student behaviour, assist faculty members of institutes to enhance learning and supporting educational processors [4].

In effect, several studies were conducted in similar area of predicting students' performance. Students' grades were predicted by feed forward neural networks and back propagation algorithms [5]. Feature examination from logged data in a web-based system has applied by another study for the prediction, monitoring and evaluation of students' performance [6]. Prediction of university students' satisfaction was another research that has been done using regression and decision tree algorithms [7]. Another set of researches were used Naïve Bayes algorithm to predict the performance of the students [8] and different rule based systems were used to predict the performance in an e-learning environment [9]. Different regression techniques were used to predict students' marks in an open university using locally weighed linear regression, linear regression, model trees, neural networks, and support vector machines and for predicting high school students' likelihoods of success in university was another study conducted in educational data mining domain [10]. There was another study about a student performance recommender system, which was developed to provide recommendations to the students during the enrolment period using Decision tree, Naïve Bayes, Neural Networks and K-nearest neighbor algorithms [11].

In this research, Naïve Bayes (NB), Decision Tree (DT) algorithm and Support Vector Machine (SVM) were used as mining approached to identify the most correlated factors with the final year performance.

3. Materials and Methods

A. Sample

This study will consider the data collected by querying the students' database of an institute, which offers Information Technology (IT) degree in three years. The dataset consists of 3794 instances with 10 attributes as shown in table 1.

Table 1. Description of Data Set

Attribute	Details
Sex	Gender (binary: Male “M”, Female “F”
Age	Age (numeric: from 18 – 45)
Failure	Has past failure modules (binary: yes, no)
S1	GPA of semester 1 (nuemeric: from 0-4)
S2	GPA of semester 2 (nuemeric: from 0-4)
S3	GPA of semester 3 (nuemeric: from 0-4)
S4	GPA of semester 4 (nuemeric: from 0-4)
S5	GPA of semester 5 (nuemeric: from 0-4)
S6	GPA of semester 6 (nuemeric: from 0-4)
IsPassDegree	Is the student pass the degree (binary: yes, no)

Gender and age are the demographic facts and all the others are academic facts related to the selected sample. During the pre-processing stage, all discriminative values were identified. All missing values, incomplete values were handled using, median imputation algorithm. Case deletion has been done to simply, removing some tuples from the data set.

Following histogram in Figure 1 describes the behaviour of our target variable.

According to the above figure1, 65% of students have successfully completed their degree while the remaining 35% have failed to complete the degree.

In prediction, the popular methods that are used by researches are classification, regression, and density estimation.

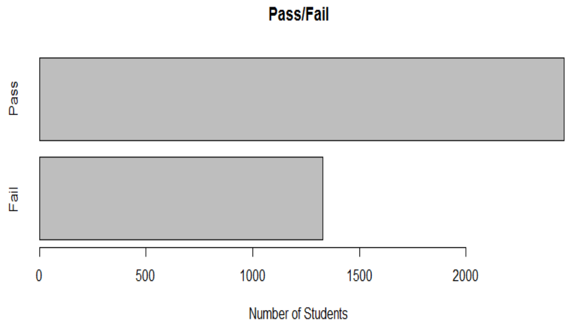


Figure 1. Histogram for the students’ Pass/ Fail state

B. *Data Mining Algorithms*

Three classification type data mining methods were applied in this study. They are;

i) Naïve Bayes algorithm
ii) C5.0 Decision Tree algorithm
iii) Support Vector Machine algorithm

Naïve Bayes algorithm is a classification method based on supervised learning approach. This Classification is named after Thomas Bayes based on his Bayes Theorem. Bayesian classification gives practical learning algorithms which prior knowledge and observed data can be consolidated. It gives a valuable point of view to comprehension and assessing many learning algorithms. Naive Bayes Algorithm is a quick, exceptionally versatile algorithm. It can be utilized for Binary and Multiclass classification. It gives distinctive sorts of Naive Bayes Algorithms like Gaussian Naïve Bayes, Multinomial Naïve Bayes and Bernoulli Naïve Bayes. It is a basic algorithm that relies upon doing a group of checks, which can be effectively, prepare on little data set.

The enhanced C5.0 decision tree algorithm was selected to obtain a better analysis result to produce a smaller decision tree with less memory in less time with the support of boosting, weighting and winnowing.

Support vector machine (SVM) is a supervised learning methods used for classification, regression and outlier detection tasks in data mining. The analysis could be able to perform using Linear or Gaussian methods.

C. *Mining Environment*

The results of this experiment were gathered using R studio statistical software package, which runs on windows platform. R is a free and high-level matrix programming language with an intense tools for statistical and data analysis [12]. It consists of several helpful packages/ libraries to perform almost all mining tasks including Naïve Bayes C5.0 Decision Tree analysis and Support Vector Machine.

The sex/ gender of the student has converted to a numerical value at the time of executing the algorithm and ‘1’ implies ‘male’ and ‘0’ implies ‘female’.

The data set has inputted to classification algorithms to categorize them. The input data set is comprising of vector values of attributes with equivalent classes. In the analysis, the data set has segregated into training sets and testing sets. Training dataset is to train the model and model learn from this. Testing data set is to measure how much the model has learnt using the train data.

3. Experiment Results

In the start of this investigation the impact of input attributes have been broken down in order to think of a better prediction result. For the identification of importance of every attribute, Info Gain and Gain Ratio were tried utilizing R package. According to the results of impact test, the greater part of the attributes in the dataset as in table 1 were included to the further analysis.

The results of the mining algorithms were assessed to think of a better classification model. For this assessment of classification accuracy, 5 fold cross validation technique has been utilized. The cross validation has repeated several times and in every time utilize one sub set as a test set. Under the cross validation; prediction accuracy, Kappa statistics, Precision, Recall and F-measure could be recorded. The performance of the selected algorithms were measured as shown in table 2.

Table 2. Performance of classification methods

Criteria	**Algorithm**		
	Naïve Bayes	Decision Tree (C5.0)	Support Vector Machine
Correctly classified instances	1153	1220	1149
Incorrectly classified instances	111	44	115
Prediction Accuracy	92.17%	97.1%	92.05%
Kappa Statistic	81.91%	93.57%	80.58%
Precision	89.403%	94.85%	89.24%
Recall	99.75%	98.71%	98.86%
F-Measure	87.5%	95.79%	86.95%

As per the above outcomes shown in table 2, it was watched that all the classification algorithms delivered moderately great outcomes, which were more like each other. However, the highest result is obtained by Decision Tree classification.

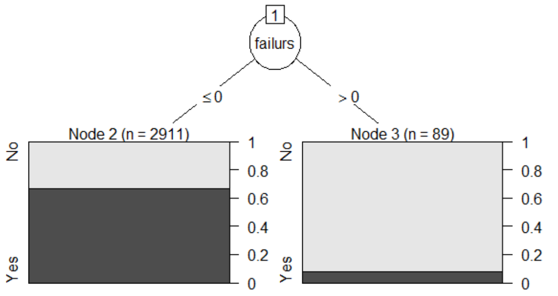


Figure 2. Decision Tree for fail modules and pass/fail status

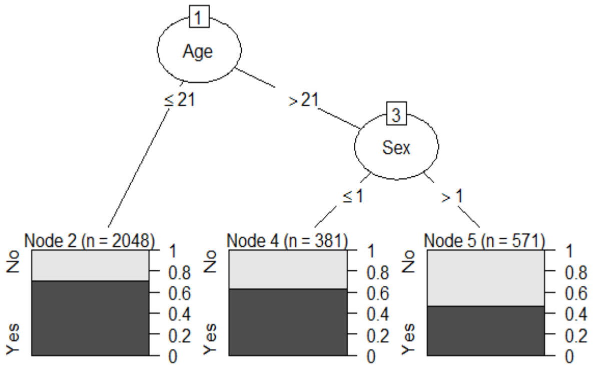


Figure 3. Decision Tree for age, sex and pass/fail status

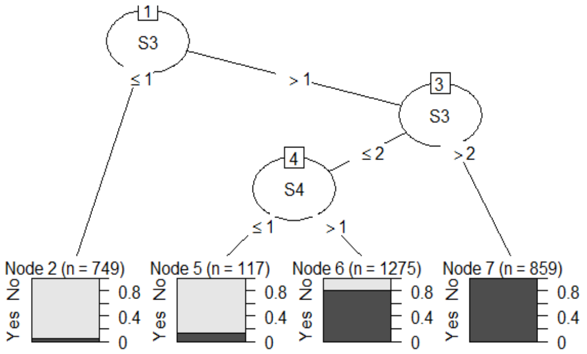


Figure 4. Decision Tree for semester 1, semester 2 performance and pass/fail status

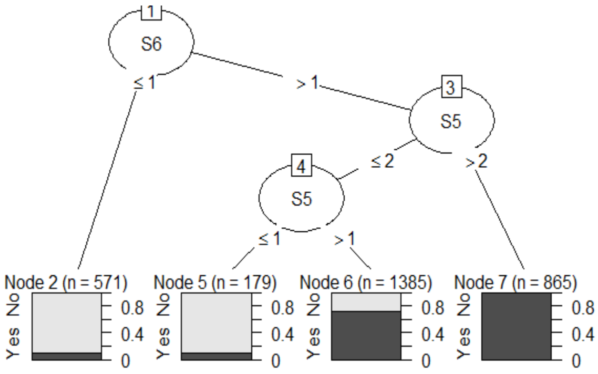


Figure 5. Decision Tree for semester 3, semester 4 performance and pass/fail status

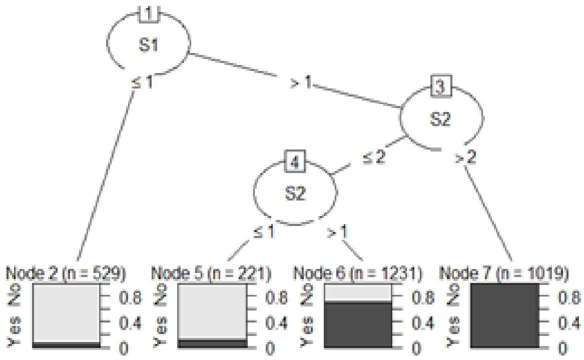
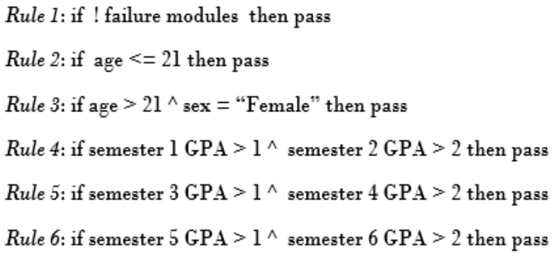


Figure 6. Decision Tree for semester 5, semester 6 performance and pass/fail status

According to the Decision Tree generated by R package, following interesting rules could be extracted.



Rule 1: if ! failure modules then pass

Rule 2: if age <= 21 then pass

Rule 3: if age > 21 ^ sex = "Female" then pass

Rule 4: if semester 1 GPA > 1 ^ semester 2 GPA > 2 then pass

Rule 5: if semester 3 GPA > 1 ^ semester 4 GPA > 2 then pass

Rule 6: if semester 5 GPA > 1 ^ semester 6 GPA > 2 then pass

Figure 7. Decision Tree Rules

The relationships discovered by the decision tree algorithms were evaluated using correlation analysis. The results of the correlation analysis was shown in table 3.

According to the results shown in above table, failure modules and pass/ fail status are significantly correlated with a correlation coefficient of -0.15 and p-value of 2.2e-16 which consists of a negative relationship among the variables.

Table 3. Results of Correlation Analysis

Variable	**P-value**	**Correlation**
Failure modules	2.2e-16	-0.1500
Age	1.82e-13	-0.119235
Semester1 GPA	2.2e-16	0.3429429
Semester2 GPA	2.2e-16	0.3705321
Semester3 GPA	2.2e-16	0.3901625
Semester4 GPA	2.2e-16	0.3191271
Semester5 GPA	2.2e-16	0.3393523
Semester6 GPA	2.2e-16	0.3160398

Age of the students and pass/ fail status of them are significantly correlated with a correlation coefficient of -0.12 and p-value of 1.82e-13 16 which consists of a negative relationship among the variables.

GPA of semester 1 and pass/ fail status of them are significantly correlated with a correlation coefficient of 0.34 and p-value of 2.2e-16.

GPA of semester 2 and pass/ fail status of them are significantly correlated with a correlation coefficient of 0.37 and p-value of 2.2e-16.

GPA of semester 3 and pass/ fail status of them are significantly correlated with a correlation coefficient of 0.39 and p-value of 2.2e-16.

GPA of semester 4 and pass/ fail status of them are significantly correlated with a correlation coefficient of 0.31 and p-value of 2.2e-16.

GPA of semester 5 and pass/ fail status of them are significantly correlated with a correlation coefficient of 0.33 and p-value of 2.2e-16.

GPA of semester 6 and pass/ fail status of them are significantly correlated with a correlation coefficient of 0.31 and p-value of 2.2e-16.

Based on the correlation results it was confirmed that the above decisions derived using Decision Tree algorithm were valid.

Therefore, the results were depicting that on the off chance that the undergraduates are not having past failure modules, clearly they will pass the degree. As the rule 1 and 2 says, dominant part of the undergraduates who completed the degree were enrolled to the degree at the age of 21 and below, which imagines that if the age is less, there exists a high likelihood of passing the degree.

4. Conclusion and Future Work

In this research study, the most prominent factors, which predict the performance of tertiary level students, are discussed. Three different data mining algorithms, i.e. Naïve Bayes, Decision Tree, and Support Vector Machine were used in the prediction stage. The results of mining algorithms were tested and the output of the algorithm, which had the highest prediction accuracy, was selected as the best prediction baseline. Accordingly, the extracted rules depicts that the age, past failure modules, performance of past semesters are having a significant correlation with the success or failure of the students of three-year degree program.

As the subsequent stage of the research study, the looks into hope to dissect the kind of the relationship measurably and will utilize other advanced mining algorithms to gauge the prediction accuracy to affirm whether similar outcomes will create by them.

Acknowledgement

The authors gratefully acknowledge Sri Lanka Institute of Information Technology for supplying students' data for the analysis.

References

[1] University of Colombo, Sri Lanka, "Student Handbook 2012/2013", http://www.bit.lk/downloads/handbook/2012-handbook.pdf. [online] accessed on 08/09/2017.

[2] Romero C, Ventura S, Data Mining in Education, WIREs Data mining knowledge discovery, Vol 3, pp. 12- 27, 2013.

[3] Kasthuriarachchi K.T.S., Chintan M. Bhatt, S.R.Liyanage, "A Review of Data Mining Methods for Educational Decision Support", International Postgraduate Research Conference, University of Kelaniya, Sri Lanka, pp- 30, December 2016.

[4] Richard A. Huebner, "A Survey of educational data-mining research," *Research in higher educational journal.*

[5] T. D. Gedeon and H. S. Turner, "Explaining student grades predicted by a neural network," in Int. Conf. Neural Netw., Nagoya, Japan, pp. 609–612, 1993.

[6] D. Shangping, Z. Ping, "A data mining algorithm in distance learning", Proc. Int. Conf. Comput. Supported Cooperative Work in Design, pp. 1014-1017, 2008.C.

[7] N. Myller, J. Suhonen, E. Sutinen, "Using data mining for improving web-based course design", Proc. Int. Conf. Comput. Educ., pp. 959-964, 2002.

[8] P. Haddawy, N. Thi, T. N. Hien, "A decision support system for evaluating international student applications", Proc. Frontiers Educ. Conf., pp. 1-4, 2007

[9] A. Nebot, F. Castro, A. Vellido, F. Mugica, "Identification of fuzzy models to predict students perfornance in an e-learning environment", Proc. Int. Conf. Web-Based Educ., pp. 74-79, 2006.

[10] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning systems using machine learning techniques," in Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst., Oxford, U.K., pp. 3–5, 2003

[11] Jorge Chue, Juan Pablo Peche, Gustavo Alvarado, Bruno Vinatea, Jhonny Estrella, Álvaro Ortigosa, "A data mining approach to guide students through the enrolment process based on academic performance", User Modeling and User-Adapted Interaction, Springer, Volume 21, Issue 1–2, pp 217–248

[12] "What is R?, R Development Core Team 2006", https://www.r-project.org/about.html [Online] accessed on 08/09/2017