2016-09-01

Data Analytics to Improve Students' Academic Performance

Ahmed, MD Sajib

Center for Pedagogy (CP) Established under the Sub-project Titled "Pedagogical Development at Undergraduate and Master's Level" (CP3357), Independent Univeristy, Bangladesh (IUB)

Data Analytics to Improve Students' Academic Performance

MD Sajib Ahmed,[66] Khawza Iftekhar Uddin Ahmed, Mohammad Tohidul Islam Miya and Hasan Sarwar
United International University

United International University (UIU) has been offering graduate degrees for more than 10 years. During this time period, we are regularly coming up with students who are suffering miserably in their academic career. In academic context, these students fall into the state of probation. A probationary state is when a student's CGPA falls below 2.00. Every new recruit is generating a larger batch of probationary students. Observation indicates that students suffer not only due to factors relating to teachers' quality and teaching practices; a lot of other issues like social surroundings, previous academic efforts, his/her real intention and other factors affect his/her performance. In this study, we would analyze some parameters of a group of students based on their historical data. Data were collected from their current and previous academic performances and survey questionnaires related to their habits and social involvements. Some data on teaching practices and delivery quality would be considered here. The analysis on data would help us to develop a model for prediction of students' future academic result. This prediction will help students to design their academic career more carefully and help develop a better sustainable nation.

Keywords: *Data analytics, data mining, decision tree, regression method*

Introduction

Data analytics is a process of collecting, cleaning, analyzing and modeling data for the purpose of sifting important insights, making prediction and reaching suggestive conclusions [1]. This helps the institution make an informed intelligent decision to maintain a sustainable growth. Data analytics is used in diverse fields such as business, industry, academia and societal relations.

Educational data analytics or educational data mining (EDM) is an emerging field that addresses the development of different methods for the exploration of volumes of data that are unique to educational institutes [2]. It models the student's performance and provides useful insight on the aspects vulnerability of teaching learning processes. Using these tools the institutional authority can predict the student's performance and take extra care and other remedial measures so that the deterioration of student's performance can be stopped and the drop-out can be prevented. Prediction of student's performance using data analytics tools have been discussed in [3], [4].

It has been observed that on an average 25% to 30% students go to the state of probation in United International University (UIU). At UIU a state of probation is defined as a status when a student achieves a CGPA below 2.0. Repeated state of probation in four consecutive years results in a student's cancellation of admission and eventual dropout from the institution. Therefore, to prevent such dropout, specific and concerted strategies need to be chalked out. Employing data analytics, the student's performance in

[66] Correspondence should be addressed to MD Sajib Ahmed, E-mail: sajib@iqac.uiu.ac.bd.

terms of CGPA and state of probation can be predicted. Based on the predicted results intrusive counseling can be arranged with the help of course instructors and administrative counselor. Such measures can improve the overall eco system of the academic institutions. In the subsequent sections we provide the description of data sets, the prediction algorithms and the performance of the predictors. Also, we explain the more insights that can be obtained from the results of the data analytics.

Data Sets

Primarily two sources of student's data are considered. One source is based on the data that are available during the process of admission test, such as, admission test mark, SSC and HSC results. Another source is the student's trimester-wise CGPA. First five consecutive trimesters CGPA of 759 students of Summer 2013, Fall 2013 and Spring 2014 trimesters are used. The students are from the different programs of the Department of Computer Science and Engineering, the Department of Electrical and Electronic Engineering and School of Business and Economics.

Methods

WEKA (Version 3.6.13) has been used for prediction of the status between probation and non-probation and prediction of the trimester CGPA. Decision tree is used for the prediction of probation, whereas regression is used for the prediction of CGPA. For decision tree, J48graft and Random Forest algorithm have been used. For regression method, Linear regression and M5P Regression Tree algorithm have been used. In our test model, 10-fold cross validation model has been used in each experiment where whole 759 instances of data have been divided into equal 10 partitions. 9 partitions are used for training and the remaining partition is used for the testing. This has been repeated for 10 times until all the partitions are employed for training one-by-one. The attributes of input data and output data for the prediction of probation are shown in Table 1. The attributes of input data and output data for the prediction of trimester CGPA are shown in Table 2. The attributes of output data are shown in bold font.

Table-1

The Attributes for the Prediction of Status of Probation for Different Trimesters

Trimester	Attributes with type
Trimester 1	Admission Test Mark (Numeric), SSC Result (numeric), HSC Result, *Probation (Yes, No)
Trimester 2	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric), 1st Trimester Result (Numeric), *Probation (Yes, No)
Trimester 3	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric), 1st Trimester Result (Numeric), 2nd Trimester Result (Numeric), *Probation (Yes, No)
Trimester 4	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric),

	1st Trimester Result (Numeric), 2nd Trimester Result (Numeric), 3rd Trimester Result (Numeric), *Probation (Yes, No)
Trimester 5	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric), 1st Trimester Result (Numeric), 2nd Trimester Result (Numeric), 3rd Trimester Result (Numeric), 4th Trimester Result (Numeric), *Probation (Yes, No)

*Probation – If a Trimester CGPA is below 2 then Probation **Yes.** The attributes of output data are shown in bold font.

Table-2

The Attributes for the Prediction of Trimester CGPA for Different Trimesters

Trimester	Attributes with type
Trimester 1	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric), **1st Trimester Result** (Numeric)
Trimester 2	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric), 1st Trimester Result (Numeric), **2nd Trimester Result** (Numeric)
Trimester 3	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric), 1st Trimester Result (Numeric), 2nd Trimester Result (Numeric), **3rd Trimester Result** (Numeric)
Data Set 4	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric), 1st Trimester Result (Numeric), 2nd Trimester Result (Numeric), 3rd Trimester Result (Numeric), **4th Trimester Result** (Numeric)
Data Set 5	Admission Test Mark (Numeric), SSC Result (Numeric), HSC Result (Numeric), 1st Trimester Result (Numeric), 2nd Trimester Result (Numeric), 3rd Trimester Result (Numeric), 4th Trimester Result (Numeric),

	5th Trimester Result (Numeric)

The attributes of output data are shown in bold font.

Performance Evaluation and Results

The performance of the binary predictor in predicting the state of probation can be measured in terms of observed accuracy and Kappa statistics. The performance of the regression algorithm in predicting the trimester CGPA is given by correction coefficients, mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE). The definitions of these performances are given below.

- **Kappa Statistic, k**

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e},$$

Here p_o is the observed accuracy and p_e is the estimated accuracy of the binary predictor.

- **Mean absolute error, MAE**

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i|$$

where y_i is the actual value and $\hat{y}_i$ is predicted value.

- **Root mean squared error, RMSE**

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **Relative absolute error, RAE**

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|}$$

where, $\bar{y}$ is the mean value of the actual values.

- **Root relative squared error**

$$\mathrm{RRSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Performance of predication of the state of probation is summarized in Table-3 and Table-4 when J48graft classification algorithm and Random Forest classification algorithm are used respectively. We observe that there is improvement of prediction accuracy in Trimester 5 compared to Trimester 1. In the prediction of the first Trimester, SSC and HSC Marks and Admission Test results are used. No prior trimester's CGPA is available before the completion of Trimester 1. However, in the prediction of the state of probation in the subsequent Trimesters, the past trimesters' CGPA are used as inputs to the predictors. Due to this, the prediction accuracies have improved with the availability of more trimesters'

CGPA. The effect of conditioning in the new environment of the university is reflected in CGPA. Therefore, prior trimester CGPA is an important parameter for the improvement of the prediction accuracy.

Table-3

Performance of Prediction Accuracy of State Probation Using J48graft Classification Algorithm

Trimester	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statistic	Total Number of Instances
Trimester-1	465 (61.2648%)	294 (38.7352%)	0.0261	759
Trimester 2	606 (79.8419%)	153 (20.1581%)	0.5628	759
Trimester 3	592 (77.9974%)	167 (22.0026%)	0.5124	759
Trimester 4	632 (83.2675%)	127 (16.7325%)	0.4821	759
Trimester 5	668 (88.0105%)	91 (11.9895%)	0.6314	759

Table-4

Performance of Prediction Accuracy of State Probation Using Random Forest Classification Algorithm

Trimester	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statistic	Total Number of Instances
Trimester 1	445 (58.6298%)	314 (41.3702%)	0.0944	759
Trimester 2	585 (77.0751%)	174 (22.9249%)	0.4982	759
Trimester 3	602 (79.3149%)	157 (20.6851%)	0.4969	759
Trimester 4	639 (84.1897%)	120 (15.8103%)	0.5468	759
Trimester 5	665 (87.6153%)	94 (12.3847%)	0.6413	759

Table-5

Performance of Prediction of CGPA Using Linear Regression Algorithm

Trimester	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Total Number of Instances
Trimester 1	0.3006	0.7625	0.9327	94.94%	95.2867%	759
Trimester 2	0.7992	0.3688	0.486	57.1588%	60.0723%	759
Trimester 3	0.8746	0.272	0.3594	47.0718%	48.4687%	759

Trimester 4	0.9158	0.2107	0.283	38.6542%	40.1658%	759
Trimester 5	0.9535	0.1467	0.2097	27.1086%	30.1365%	759

Table-6

Performance of Prediction of CGPA Using M5P RegressionTree Algorithm

Trimester	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Total Number of Instances
Trimester 1	0.3181	0.7637	0.9274	95.0935	94.7474	759
Trimester 2	0.7986	0.3674	0.4867	56.9403	60.1592	759
Trimester 3	0.8765	0.2679	0.3569	46.366	48.1381	759
Trimester 4	0.9158	0.2107	0.283	38.6542	40.1658	759
Trimester 5	0.9532	0.1474	0.2103	27.2493	30.2187	759

Table 5 and Table 6 tabulate the performance of the prediction of Trimester CGPA using Linear regression and M5P regression tree algorithms, respectively. Again we observe that predicted CGPA are better correlated with the actual CGPA with the passage of the trimesters when prior CGPAs are available.

In addition to the prediction of the student's performance data analytics can also be employed to derive valuable insight and intelligent on the causes of poor performance of students. To clarify this, let's consider the scenario of the 759 students who are considered in our study. Table-7 shows that the students who are from outside of Dhaka have a higher likelihood of going into probation. This may be due to the challenges of facing a new metropolitan city like Dhaka.

Table-7

Status of 1st year probation based on if the student is from outside Dhaka

	1st Trimester Probation					
From outside Dhaka (Yes/No)	No		Yes		Total	
Yes	226	56.4%	175	43.6%	401	52.8%
No	246	68.7%	112	31.3%	358	47.2%
Total	472	62.2%	287	37.8%	**759**	

Table 8: Relation Between Pre English and 1st Trimester Probation

	1st Trimester Probation					
Pre English	No		Yes		Total	
No	212	64.8%	115	35.2%	327	43.1%
Yes	260	60.2%	172	39.8%	432	56.9%
Total	472	62.2%	287	37.8%	**759**	

Table 8 shows the students who are comparatively poorer, i.e., who need to take a Pre-English course in English, are more prone to falling in to probation.

Conclusions

This paper presents some preliminary efforts on how data analytics can help in identifying the poor performing students in a university. The performance of prediction accuracy is convincing and it can be used to administer follow-up actions in containing the poor performer early so that their grades can be improved. In addition, data analytics can be employed to reveal the causes of poor performance so that early remedial and counseling activities can be designed. To increase the prediction performance of 1st year some more relevant data on the student's regular practice such as his/her regular study hours, and time he/she uses for socialization can be collected and afterwards can be fed into learning algorithms for the prediction. It is also evident that the SSC, HSC and Admission Test Results are not good features in predicting the state probation in the first trimester.

References

Bihani, Prateek, and S. T. Patil. "A comparative study of data analysis techniques." *International Journal of Emerging Trends & Technology in Computer Science* 3.2 (2014): 95-101.

Romero, Cristobal, and Sebastian Ventura. "Data mining in education." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 3.1 (2013): 12-27.

Dietz-Uhler, Beth, and Janet E. Hurn. "Using learning analytics to predict (and improve) student success: A faculty perspective." *Journal of Interactive Online Learning* 12.1 (2013): 17-26.

Kotsiantis, Sotiris B. "Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades." *Artificial Intelligence Review* 37.4 (2012): 331-344.