Predicting Student Academic Performance using Data Mining Methods

Raheela Asif[1], Saman Hina[1], Saba Izhar Haque[1]

[1]N.E.D University of Engineering & Technology /Department of Computer Science & Software Engineering, Karachi, 75270, Pakistan

Summary
The aim of this study is to use data mining techniques for predicting the students' graduation performance in final year at university using only pre-university marks and examination marks of early years at university, no socio-economic or demographic features are use.

Key words:
Educational data mining, predicting performance, decision trees

1. Introduction

In the past three decades the computer hardware technology has become very powerful. This has boosted up the database and information industry. As a result a large number of databases and information repositories are available and the organizations stored plenty of data. This has increased the need for powerful data analysis which is not possible without powerful tools. Data mining tools analyze data from different perspectives and summarize the results as useful information. They are employed to operate on large amounts of data to find out hidden patterns and associations that can be helpful in decision making [1]. The application of data mining methods to educational data is called Educational Data Mining (EDM) which is novel and promising field [2]. Researchers and experts in education are using EDM techniques in higher education institutions to enhance learning.

This paper focused on the capabilities of data mining in higher learning institutions for the study of educational data. It reflects on how data mining may help to improve decision-making processes in universities. This work aims on predicting students' academic performance at the end of four year bachelor's degree program and identifying effective indicators of at risk students in early years of their study. It provides the institution with the needed information using which it can outline measures to improve quality.

The paper is arranged as follows: The next section is devoted to literature review. Section 3 describes the data collection and methodology used for this study. Results and discussions are presented in Section 4. Finally, Section 5 concludes the paper.

Manuscript received May 5, 2017
Manuscript revised May 20, 2017

2. Literature Review

The literature review discloses that predicting performance at higher education level has involved substantial attention in the recent past and persists to remain focus of research and discussion. A number of studies investigated the performance of the students at higher level [3,4,5,6,7,].

The study conducted by [3] employs the Adaptive Neuro-Fuzzy Inference system (ANFIS) to predict student academic performance which will help the students to improve their academic success.

Acharya and Sinha [4] apply Machine Learning Algorithms for the prediction of students' results. They found that best results were obtained with the decision tree class of algorithms.

Kaur et al. [5] identify slow learners among students and displaying it by a predictive data mining model using classification based algorithms.

Gurlur et al. [6] attempt to find out student demographics that are associated with their success by using decision trees.

Vandamme et al. [7] use decision trees, neural networks and linear discriminate analysis to make early predictions of students' academic success in first academic year at university.

The literature review about predicting performance mentioned above show that it is possible to predict performance of students with a reasonable accuracy. All the mentioned works use cross validation to assess their results. However, we take one batch to train the classifier and the other batch to test the prediction results. This aspect differ our works from other works.

3. Data and Methodology

3.1 Data

In this study, we used the data of two academic cohorts or batches of Civil Engineering Department at NEDUET, Pakistan, which entailed altogether 214 undergraduate students enrolled in the academic batches of 2005–06 and

2006–07. We use pre-university marks i.e. HSC (High School Certificate) marks and the examination marks of students' in first and second year courses that are taught in first and second academic years, shown in Table 1. The prediction variable is the class interval which is calculated on the basis of the final marks of the degree. The final marks of the degree is divided into 5 class intervals: Class_A (90%–100%), Class_B (80%–89%), Class_C (70%–79%), Class_D (60%–69%), and Class_E (50–59%)

Table 1: List of variables used in study

Variable	Description
Class Interval	5 promising values(Class_A, Class_B, Class_C, Class_D and Class_E)
Adj_Marks	HSC Examination total marks
Maths_Marks	HSC Examination Mathematics marks
MPC	Maths+ Physics+ Chemistry marks
CE-101	Engineering Drawing-I
CE-102	Engineering Mechanics
CE-103	Surveying-I
CE-104	Engineering Materials
EE-102	Electrical Engineering
HS-101	English
HS-105/127	Pakistan Studies
ME-105	Applied Thermodynamics
MS-105	Applied Chemistry
MS-111	Calculus
MS-121	Applied Physics
CE-201	Surveying-II
CE-202	Introduction to Computing
CE-203	Engineering Drawing-II
CE-204	Fluid Mechanics-I
CE-205	Mechanics of Solids-I
CE-206	Engineering Geology
CE-209	Structural Analysis-I
MS-331	Applied Probability & Statistics
HS-205/206	Islamic Studies
MS-221	Linear Algebra & Ordinary Differential Equations
HS-303	Engineering Economics

3.2 Methodology

To predict the performance of the students as early as possible, we use HSC marks and the marks in first and second year courses to predict the performance of the students'. We used the data of batch 2005–06 to train the prediction models which were then used to predict the class intervals of students in the next batch i.e. 2006-07. Batch and Interval statistics are presented in Table 2.

Table 2: Batches and Class Interval Statistics

Academic Cohort	Total number of students' in Class	students in Class	students in Class	students in Class	students in Class
2005–06	99	-	3	46	44	6
2006–07	115	-	3	51	44	17

Table 2 shows that the distribution of students amongst the class intervals is unbalanced. 'Class_C' interval contains the most students. Predicting a class interval 'Class_C' would have an accuracy of 44.34%. This is the baseline of accuracy that we want to improve.

We ran a number of classifiers like Decision Tree produced with Gini Index (DT-GI), Decision Tree produced with Information Gain (DT-IG), Decision Tree produced with Accuracy (DT-Acc), Naive Bayes, Neural Networks (NN), Random Forest produced with Gini Index (RF-GI), Random Forest produced with Information Gain (RF-IG) and Random Forest produced with Accuracy (RF-Acc).

4. Analysis and Results

Table 3 shows the results of accuracy and kappa for the classifiers. We have applied other classifiers like Decision Tree with Gain Ratio, Rule Induction with Information gain, Rule Induction with Accuracy, I-NN, Linear Regression and Support Vector Machines. Their results are not mentioned here as the classification accuracies are not above the baseline.

Table 3: Prediction accuracy and Kappa results

Criterion	Accuracy	Kappa
Decision tree produced with the criterion Gini index (DT-GI)	67.71%	0.514
Decision tree produced with the criterion information gain (DT-IG)	66.67%	0.475
Decision tree produced with the criterion Accuracy (DT-Acc)	58.33%	0.356
Naïve Bayes	65.62%	0.466
Random Forest Trees produced with the criterion Gini index (RF-GI)	58.33%	0.378
Random Forest Trees produced with the criterion Information Gain (RF-IG)	66.67%	0.499
Random Forest Trees produced with the criterion Accuracy (RF-Acc)	50.00%	0.230

To improve the accuracy of the classifiers, we apply different feature selection techniques available in Rapid Miner. The Recursive Feature Elimination (RFE) operator available in RapidMiner has four criterions to weight attributes: Weight by Gini index (GI), weight by information gain ratio (IG), weight by chi-squared (Chi-SS) and weight by rule induction to choose subsets of variables. We have four different subsets of variables from the four criterions of the RFE operator. Each subset contains seven variables. It is interesting to observe that two subsets contain HSC marks. This means that HSC marks play an important role in student's university performance at Civil Engineering Department. The prediction models of Table 2, i.e. decision tree produced with the criterion Gini index (DT-GI), decision tree produced with the criterion information gain (DT-IG), decision tree produced with the criterion accuracy (DT-Acc), naive Bayes (NB), neural networks (NN) and random forest trees produced with the criterion Gini index (RF-GI), random forest trees produced with the criterion information gain (RF-IG) and random forest trees produced with the criterion accuracy (RF-Acc) were built again using these four subsets of variables. Figure 1 gives the results of feature selection algorithms.

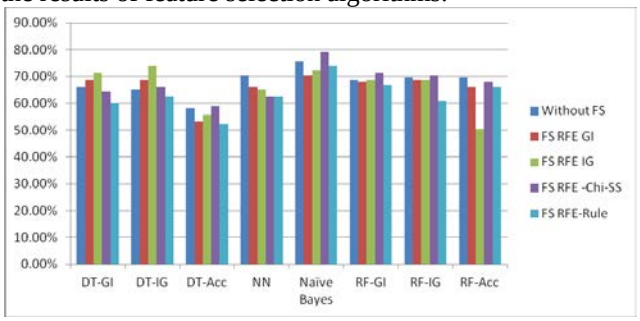

Fig. 1 Comparison of classifiers accuracy for Applying Feature Selection

We can see from the Figure 1, that there is no feature selection technique that improves the accuracy for all classifiers or a big majority of them. However, the accuracy for RFE-Chi-SS improves for two classifiers and stays the same for three classifiers. RFE- IG gives the best accuracies for two of the decision trees as compare to other feature selection techniques. We are more interested in decision trees result as they are understandable and can be used in implementing some policy. The set of attributes selected by RFE-Chi-SS is: CE-102, CE-103, CE-202, CE-203, CE-204, CE-206, MS-331. The set of attributes selected by RFE-IG is: Adj_Marks, CE-101, CE-102, CE-103, CE-202, CE-204, MS-331. If we take the intersection of these two sets we have 5 courses in common i.e. CE-102, CE-103, CE-202, CE-204 and MS-331. The meaning of these courses is given in Table 1.

We also investigated the Pearson's correlation of first and second year courses with the final marks obtained in the examination. The results of correlations are presented in Table 3.

Table 2: Correlation results between first and second year courses and final marks

Criterion	**Without Feature Selection (Accuracy / Kappa)**	**Features Selected by taking Intersection of the attributes selected by RFE-Chi SS and RFE-IG (Accuracy / Kappa)**
Decision tree produced with the criterion gini index (DT-GI)	66.09% /0.451 (with minimal leaf size 4)	74.78% / 0.593 (with minimal leaf size 6)
Decision tree produced with the criterion information gain (DT-IG)	65.22% / 0.399 (with minimal leaf size10)	73.91% / 0.582 (with minimal leaf size 6)
Decision tree produced with the criterion accuracy (DT-Acc)	58.26% / 0.344 (with minimal leaf size 4)	55.65% / 0.311 (with minimal leaf size 2)
Neural Networks	70.43% /0.497	68.70% / 0.508
Naïve Bayes	75.65% /0.616	74.78% / 0.591
Random Forest Trees produced with the criterion Gini Index (RF-GI)	68.70% / 0.464 (with minimal leaf size 2)	67.83% / 0.448 (with minimal leaf size 8)
Random Forest Trees produced with the criterion Information Gain (RF-IG)	69.57% / 0.481 (with minimal leaf size 6)	67.83% / 0.454 (with minimal leaf size 4)
Random Forest Trees produced with the criterion Accuracy (RF-Acc)	69.57% / 0.486 (with minimal leaf size 8)	64.35% / 0.385 (with minimal leaf size 8)

The five courses that we selected through the intersection of the subsets of RFE-IG and RFE-Chi-SS include one non-course of second year (i.e. MS-331), two core courses from first year and two core courses from second year. They are highlighted in Table 3. We can see from above table that all these five courses have high correlation with the final marks.

This subset of 5 courses was used with the same eight classifiers. The results are presented in Table 4. The three decision trees that are obtained by using these 5 courses are shown in Fig.1, Fig.2 and Fig. 3.

Table 4: Comparison of Prediction Accuracies after applying feature selection based on intersection of RFE-Chi SS and RFE IG

Variable	Correlation with final marks Batch 2005-06	Correlation with final marks Batch 2006-07
CE-101	0.446	0.647
CE-102	0.610	0.758
CE-103	0.671	0.760
CE-104	0.543	0.661
EE-102	0.483	0.626
HS-101	0.397	0.403
HS-105/127	0.288	0.654
ME-105	0.385	0.662
MS-105	0.549	0.653
MS-111	0.505	0.602
MS-121	0.550	0.613
CE-201	0.567	0.617
CE-202	0.715	0.693
CE-203	0.551	0.663
CE-204	0.765	0.842
CE-205	0.586	0.453
CE-206	0.694	0.603
CE-209	0.638	0.570
HS-205/206	0.463	0.665
HS-303	0.647	0.665
MS-221	0.546	0.674
MS-331	0.733	0.696

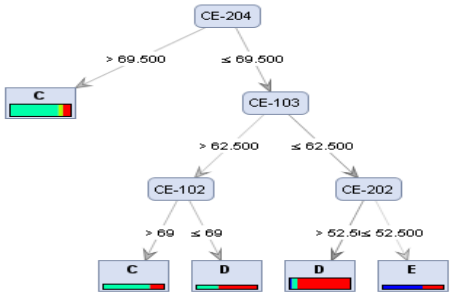

Fig. 2 Decision tree produced with the Gini index with K=5

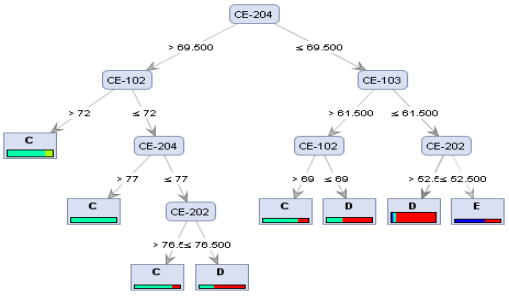

Fig. 3: Decision tree produced with the information gain with K=5

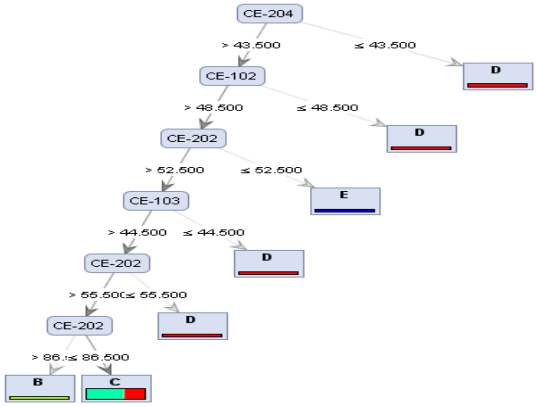

Fig.4: Decision tree produced with the accuracy with K=5

By examining the above trees, one can observed that there are two indicators of low performance: CE-102 and CE-202. A low performance in CE-102 leads to a leaf C or D and a low performance in CE-202 lead to a leaf with D or E interval in all the three trees. This suggests that a student having a mark lower than or equal to 48 in CE-102 are likely to achieve their degree with a poor mark. This suggests also that students having 52 or less in CE-202 are likely to obtain 52 or less in other subjects as well again because of the way the final mark is calculated.

The 2 indicators of low performance contain one course from first year and one from second year. CE-102, the first year course should be taken as indicator to warn students in first year. This can be abridged as follows:

- In first year, those students whose marks are around or less than 48 in CE-102, are likely to have a mark in the 'D' interval at the end of the degree.
- In second year, students whose marks are around or below 52 in CE-202 are likely to have a mark in the 'D' or 'E' interval at the end of the degree.

The above findings can be used to implement some policy. For example, the instructors of the course CE-102 could report about students with marks equal or less than 48. There is a possibility that these students are at risk and they need more academic support. A similar possibility of identifying at risk students could take place in second year, where the instructors could report about students whose marks are less than 52 in CE-102. These suggestions may help the University to pay extra attention to those students who are at risk by arranging more academic facilities e.g. extra classes or extra consultation hours with the instructors.

5. Conclusion

The result of the study shows that we can predict the graduation performance in a four-years university program using only pre-university marks and marks of first and second year courses, no socio-economic or demographic features, with a reasonable accuracy, and that the model established for one cohort generalizes to the following cohort. It makes the implementation of a performance support system in a university simpler because from an administrative point of view, it is easier to gather marks of students than their socio-economic data. The result also shows that decision trees can be used to identify the courses that act as indicator of low performance. By identifying these courses, we can give warning to students earlier in the degree program.

References

[1] J. Han, and M. Kamber, Data Mining Concepts and Techniques, 2nd ed. San Francisco: Morgan Kaufmann, pp.5-7, 2006.

[2] R.S.J.D Baker, and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions", 2nd International Conference on Educational Data Mining, Proceedings. Cordoba, Spain, pp. 1, 3-17, July 1-3, 2009.

[3] [3] A. Altaher , O. BaRukab,"Prediction of Student's Academic Performance Based on Adaptive Neuro-Fuzzy Inference", International Journal of Computer Science and Network Security, Vol.17 No.1, January 2017.

[4] [4] A. Acharya, D. Sinha, "Early prediction of student performance using machine learning techniques", International Journal of Computer Applications, Volume 107–No. 1, December 2014.

[5] [5] P. Kaur, M. Singh, G. S. Josan, "Classification and prediction based data mining algorithms to predict slow learners in education sector", 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015).

[6] [6] H. Guruler , A. Istanbullu , M. Karahasan. "A new student performance analysing system using knowledge discovery in higher educational databases". Computer and Education. 2010. 247-254.

[7] [7] J. P. Vandamme, N. Meskens , J. F. Superby, " Predicting Academic Performance by Data Mining Methods", Education Economics, Volume 15, No. 4, 2007.