I.J. Intelligent Systems and Applications, **2015, 01, 49-61**
Published Online December 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2015.01.05

Predicting Student Academic Performance at Degree Level: A Case Study

Raheela Asif
N.E.D University of Engineering & Technology /Department of Computer Science & I.T., Karachi, 75270, Pakistan
Email: engr_raheela@yahoo.com

Agathe Merceron, Mahmood K. Pathan
Beuth University of Applied Sciences /Department of Computer Science and Media, Berlin, 13353, Germany;
Federal Urdu University of Arts, Science & Technology, Karachi, 75300, Pakistan
Email: merceron@beuth-hochschule.de, mkpathan@hotmail.com

Abstract— Universities gather large volumes of data with reference to their students in electronic form. The advances in the data mining field make it possible to mine these educational data and find information that allow for innovative ways of supporting both teachers and students. This paper presents a case study on predicting performance of students at the end of a university degree at an early stage of the degree program, in order to help universities not only to focus more on bright students but also to initially identify students with low academic achievement and find ways to support them. The data of four academic cohorts comprising 347 undergraduate students have been mined with different classifiers. The results show that it is possible to predict the graduation performance in 4th year at university using only pre-university marks and marks of 1st and 2nd year courses, no socio-economic or demographic features, with a reasonable accuracy. Furthermore courses that are indicators of particularly good or poor performance have been identified.

Index Terms— Educational Data Mining, Knowledge Discovery, Predicting Performance, Electronic Performance Support System, Pedagogical Policy, Classification, Decision Trees

I. Introduction

Universities are working in a very dynamic and powerfully viable environment today. They gather large volumes of data with reference to their students in electronic form. However, they are data rich but information poor which results in unreliable decision making. The biggest challenge is the effective transformation of large volumes of data into knowledge to improve the quality of managerial decisions. Knowledge discovery in databases (KDD) refers to the discovery of interesting knowledge from the large volumes of data [1]. The KDD involves data selection, preprocessing of data, data transformation, data mining, understanding the results and reporting. However, since Data Mining is a crucial and significant part of the KDD process, many people uses data mining as a synonym for KDD [2]. The advances in the data mining field make it possible to mine the educational data and find information that allow for innovative ways of supporting both teachers and students.

There has been a big variety of research works using data mining techniques in higher education institutions to enhance learning, going from analyzing students' enrolment data to prevent drop-off and improve retention [3, 4, 5, 6], to predict student retention at an early stage from ePortfolios features [7], to analyzing the usage of learning materials uploaded in a E-Learning platform [8] or analyzing mistakes that students make together in a tutoring system [9]. The handbook of educational data mining [10] gives a good overview of representative works in the educational data mining area.

The review paper [11] has proposed 10 common tasks in education that have been tackled using data mining techniques and predicting students' performance is one of them. Predicting students' performance using data mining methods has been performed at various levels: at a tutoring system level to predict whether some specific knowledge or skills are mastered, at a course level or degree level to predict whether a student will pass a course or a degree, or to predict her/his mark. At a tutoring system level [12] predicts whether a student is likely to get the next training exercise right, and if yes, the tutoring system should skip it. For the course level, [13] has found that perceived ease of use of e-learning tools, perceived usefulness of e-learning tools and the ability to work independently were statistically significant contributors to the final course grade; [14] predicts whether a student will pass or fail a course based on his/her forum activity and [15] predicts course performance using students' performance in prerequisite courses and midterm examinations. The present contribution focuses on the latter level: predicting the mark of students at the end of a university degree. To predict students' performance at an early stage of the degree program helps universities not only to focus more on bright students but also to initially identify students with low academic achievement and find ways to support them.

The paper is organized as follows: Section II is devoted to the related works. Section III briefly describes the aim of the present study. Section IV gives a short overview of the data mining techniques used in this investigation. Section V describes the data and tools used for this case

study. The following section i.e. section VI describes the analysis and presents the results and is followed by a section called "discussion and implication". The conclusion elaborates on some findings, discusses them and presents future works.

II. Related Works

A number of works have investigated predicting performance at a university degree level.

The study in [16] determines the relationship between students' demographic attributes, qualification on entry, aptitude test scores, performance in first year courses and their overall performance in the program. Their sample data consisted of 96 students, 68 male and 28 females that were accepted to in the Bachelor of Science in Computing and Information Technology (BSCIT) at University of Technology, Jamaica (UTECH) in 1999-2000 academic years. The data was analyzed using stepwise multiple regression analysis. This study suggests that students who have done well in the foundation programming courses should be encouraged to continue in BSCIT program, while students who have not grasped the concepts should be channeled in the Bachelors of Science in Computing and Management Studies (BCMS) program. This study identifies an optimal set of admission indicators, which have the potential of predicting students' performance.

The investigation in [17] finds that performance in the first year of computer science courses is a determining factor in predicting students' academic performance at the conclusion of the degree. They consider the data of 85 students in the School of Computing and Information Technology at the UTECH and analyze this single cohort of students through the entire degree. They find that the first year gateway courses like C Programming, Introduction to Computer Networks and Computer Logic & Digital Design are strong predictors for overall academic performance (Grade Point Average GPA) in BSCIT program at UTECH. They use statistical methods like regression, no other data mining classifier, and find a strong correlation between the performance in first year computer science courses and students overall performance in BSCIT program with a correlation of 0.499 that explains 70.6% of students' performance. The authors also concluded that students' demographics do not have any significant relation to academic performance.

The work in [18] employs the data mining technique random forests, essentially a set of decision trees, to predict students' graduate level performance (Master of Science, M.Sc.) by using undergraduate achievements (Bachelor of Science, B.Sc.). In their study, they acquire the data of 176 undergraduate students of Computer Science at ETH Zurich. They use 125 predictor variables which include gender, age, single course achievement (first and final examination attempts), several GPA's (e.g. GPA 3rd year, GPA 3rd year core courses, GPA 3rd year elective courses, GPA 2nd year, GPA 2nd year compulsory courses, GPA 1st year etc.) and study duration; the target variable that is predicted is the GPA of M.Sc. program. They find that a small set of variables, namely 14, explains 55% of the variation in graduate performance and this set contains essentially grades. Further, they find that 3rd year B.Sc. achievements are more predictive than 1st year grades to predict the GPA of M.Sc. program. The evaluation of the prediction is done using an out-of-bag scheme: the model is created using all data except one and is tested on the left out data and the process is repeated n times with n being the number of data. This means that only one cohort of students is used to build the prediction model and to evaluate it.

The investigation in [19] predicts academic performance considering the data of two different universities. In the first case study, they use the data of undergraduate students of Can Tho University (CTU) in Vietnam to predict GPA at the end of 3rd year of their studies by using the students' records (e.g. English skill, field of study, faculty, gender, age, family, job, religion, etc.) and 2nd year GPA. In the second case study, they consider the data of masters' students of Asian Institute of Technology (AIT). By using students' admission information (like academic institute, entry GPA, English proficiency, marital status, Gross National Income, age, gender, TOEFL score etc.) they predict the GPA of students at the end of 1st year of the master degree. In the above studies, two data mining algorithms are employed namely decision trees and Bayesian networks and the accuracies of these algorithms are also compared. For these two case studies, the authors have done predictions for 4 classes (Fail, Fair, Good, and Very Good), 3 classes (Fail, Good, and Very Good) and 2 classes (Fail and Pass). They obtain higher accuracies using the decision tree classifier. The accuracies are as follows: For 2 classes the accuracies are: CTU, 92.86% and AIT, 91.98%; for 3 classes the accuracies are: CTU, 84.18% and AIT, 67.74%; and for 4 classes the accuracies are CTU, 66.69% and AIT, 63.25%. It is well known that good results for classification are less difficult to obtain when the classes are coarser; therefore, the prediction accuracy of 2 classes is much higher than that of 3 classes or 4 classes. Their results also show that highest accuracies are achieved for the largest classes, which are Good students in CTU dataset and Very Good students in AIT data. They measure the accuracy of predictions using cross-validation with 10 folds: 9/10 of the data is used to build the model that is tested on 1/10 of the data, and this process is repeated 10 times. This again means that a single cohort is used to build the prediction model and to evaluate it.

The study in [20, 21] predict students' university performance by using students' personal and pre-university characteristics. They take the data of 10330 students of a Bulgarian educational sector, each student being described by 20 attributes (e.g., gender, birth year and place, place of living, and country, place and total score from previous education, current semester, total university score, etc.). They have applied different data mining algorithms such as the decision tree C4.5, Naive Bayes, Bayesian networks, K-nearest neighbors (KNN)

and rule learner's algorithms to classify the students into 5 classes i.e. Excellent, Very Good, Good, Average or Bad. The best accuracy obtained by all these classifiers is 66.3%. The predictive accuracy for the Good and Very Good classes (which contain most students) for all classifiers is around 60%–75%.

The work presented in [15] does not predict performance at degree level but at a course level. However it is interesting as it suggests a kind of upper bound for the accuracy that can be achieved when predicting performance at the end of a degree. They employed four mathematical models namely multiple linear regression, multilayer perception networks, radial basis functions and support vector machines to predict students' academic performance in an engineering dynamics course. They worked on the data of 323 undergraduate students who took dynamics course at Utah State University in four semesters. Their predictor variables were the students' cumulative GPA; grades earned in four pre-requisite courses i.e. statistics, calculus I, calculus II and physics; and scores on three dynamics midterm examinations. They used six combinations of predictor variables to develop a total of 24 predictive mathematical models. For all the four models, they achieved an average prediction accuracy of 81%–91%. This work shows that previous marks can predict the grade in a course with high accuracy. It also gives some limit to what can be achieved when predicting graduation performance. Indeed the predictors include midterm examinations that can be expected to correlate well with the final exam of the course, more than marks of single courses with the graduation mark.

These works show that it is possible to predict performance at a degree level with an accuracy of more than 60% when several classes for the marks are considered. They also show that there is not necessarily some data mining classifier that is better than all the others to obtain a good prediction, though decision trees and Bayesian methods are quite commonly used. They suggest that some academic performance is needed for good results and that socio-demographic factors might be less relevant. All these works validate their approach on the same cohort and consequently leave unanswered the following question: do the models built for one cohort generalize to the next one? This question is important to implement some support policy in which information gained from one cohort can be used to enhance learning of the next cohort.

III. Aim Of The Present Study

The present paper seeks to answer the two following questions: "Can we predict the performance of students at the end of their degree in 4th year with a reasonable accuracy using only their marks in High School Certificate (HSC), and in first and second year courses at university, no socio-economic data, and utilizing one cohort to build the model and the next cohort to test it?" and "can we identify those courses in first and second years which are effective predictors of students' performance at the end of the degree?". From an administrative point of view, it is easier to gather marks of students than their socio-economic data. Therefore if a reasonable prediction can be reached without socio-economic data, it makes the implementation of a performance support system in a university easier. If courses can be identified with a major impact on graduation performance, then measures can be taken at the level of those courses, making also the implementation of a performance support system easier. In this study the performance of a student at the end of the degree will be a class A, B, C, D or E, which represents the interval in which her/his final mark lies. Intervals allow for differentiating between strong and weak students.

The present study differs from other works in three aspects. First, using the conclusions of others, it limits the variables to predict performance to marks only, no socio-economic data. Second, it takes one cohort to build a model and the next cohort to evaluate it, thus allowing for some measurement of how well findings generalize from one cohort to the next one. Third it is a longitudinal study as four cohorts of students have been considered.

IV. Data Mining Techniques For Classification

Data mining techniques for classification (or classifiers) predicts the class or label of a data object. A data object is described by a set of attributes; in our context an attribute is a mark. A training dataset contains data objects with a known label or class, in our case the interval of the graduation mark. A classifier makes use of a learning algorithm to find a model that best defines the relationship between the attributes and the class label of the training dataset. The generated model by the learning algorithm should both fit the training data well and correctly predict the class label of the testing data, the data which is independent of training data and therefore not used to build the classifier. Usually the performance of classification models is evaluated on the basis of the counting of test records that are correctly and incorrectly predicted by the model. These counts are put into a table called confusion matrix, see Fig. 2 below for an example. Summing up the number of correctly predicted objects in the confusion matrix gives a single number used to calculate accuracy. Accuracy is defined as the ratio of the number of correct predictions and total number of predictions.

They are many classifiers and none is known to perform better than the others in all situations. This also applies to educational data. Therefore, one has to investigate whether some classifier outperforms the others in a particular field of study. We briefly present the five well-known classification techniques, i.e. decision trees, rule induction, artificial neural networks, k-nearest neighbor and naive Bayes, that have given the best results in our study.

A decision tree is a kind of non-cyclic flowchart; see the decision trees in the appendix for an example. The tree consists of internal nodes (non-leaf nodes) that

correspond to a logical test on an attribute, and connecting branches that represent an outcome of the test. The nodes and branches form a sequential path through a decision tree that reaches a leaf node, which represents a class label. Any node in the tree corresponds also to a subset of the dataset. Ideally a leaf is pure, which means that all elements in a leaf have the same value for the target variable or class. In our study, this means that, ideally, all students of a leaf node have their graduation mark in the same interval, like A or C. If a leaf is not pure, its class label is determined by the most frequent value of the target variable or class. The uppermost node in a tree is the root node and contains the complete dataset. A tree is built by calculating which attribute can best separate an impure node into children nodes that are purer than the parent node. Several criteria can be used for this calculation. In this study, four criteria, namely information gain, Gini index, accuracy and gain ratio have been used. Information gain is based on information theory. If a node is pure, its entropy is 0. The bigger the entropy, the less pure is the node. The entropy of a node is calculated for all attributes, in our study for all marks. The variable or attribute that has the minimum entropy, or equivalently the biggest information gain, is chosen to split the node. Gini index is another measure of impurity of a node based on observed probabilities instead of entropy. As for entropy, the value is 0 if the node is pure and increases with the impurity of a node. Here too, the Gini index of a node is calculated for all variables. The variable that maximizes the decrease in impurity (means it has the smallest Gini index) is selected as the splitting variable. The accuracy is defined as above. The variable that maximizes the accuracy of the whole tree constructed so far is selected for split. Gain ratio, another criterion, is a variation of information gain as it has been observed that information gain tends to favor variables with a large number of distinct values. The results of decision trees can be written as IF-THEN rules, are simple to understand and interpretable by humans; hence they can be used in building policies, which is important in the present work.

In a rule induction algorithm, IF-THEN rules are extracted sequentially, i.e. one after the other, from the training data, as opposed to a decision tree that generate IF-THEN rules in parallel. Each rule for a given class should have a high coverage and a high accuracy, where coverage is measured by the proportion of the data to which the rule applies. Once a rule is learned, the corresponding subset is excluded from the data and a new rule is learned on the remaining dataset. In this study, we used Rule Induction with information gain as a criterion to learn rules. As for decision trees, the results of a rule induction algorithm are easily interpretable for humans.

An artificial neural network (ANN) comprises a set of interconnected units, supposed to represent neurons, in which each connection has a weight associated with it. The first layer of units receives the input, for us all the marks of a student, and the last layer produces the output, for us the interval. An activation function is associated to each unit. ANNs learn by adjusting the weights using a learning algorithm so that they are capable of predicting the correct class label of the input record. In this study, we use the well-known neural network architecture called multi-layer perceptron (MLP) with a back propagation as supervised learning algorithm. The functionality of MLPs is influenced by the number of hidden layers, units in hidden layers, the activation functions, weights, the number of training iterations etc. Some other parameters that play a role in training MLPs are the learning rate and the momentum. A learning rate manages the size of weight and bias changes during learning. Momentum is used to prevent the system from converging to a local minimum or saddle point [22]. ANNs have a high accuracy in many applications. However, their results are not understandable by humans, which is a drawback in our case, because we want to identify the courses that have a strong impact on the performance of students at a degree level.

The k-nearest neighbor (k-NN) algorithm is a method of classifying records based on learning by similarity. A distance has to be chosen to measure the likeness of two records. The unknown record is classified by a majority vote of its neighbors; it is assigned to the class most common amongst its k nearest neighbors or, in other words, the k records with the smallest distance to the unknown record. K is a positive integer, typically small. In the present study, we chose k=1 which meant that the class of a student would be predicted by taking the class of the student in the dataset with the most similar marks in all subjects. The similarity of two records is measured by using some distance metric, e.g. Euclidean distance, cosine similarity, correlation similarity, Jaccard similarity, etc. In this study, all our variables are marks or numbers, i.e. quantitative variables, as we will see in the next section. Therefore, we used the Euclidean distance to measure the closeness of records. Contrarily to the algorithms seen so far, a k-NN algorithm does not build a model, and therefore is not trained. As for ANNs the results of a k-NN algorithm are not easily interpretable by humans.

Bayesian classifiers use the observed probabilities of the data and are based on Bayes theorem. They calculate the probability that that a given record belongs to a particular class and use the training set to estimate a normal distribution for each class to be predicted, in our case each interval. A record is then assigned to the class with the highest probability. A Naive Bayes classifier makes the strong assumption that all attributes are independent in the probability sense, which allows for a considerable simplification of the calculations. Despite this “naive” approach, Naive Bayes classifiers are fast to train and are reported to give a high accuracy in many applications. However, their output is not easy to interpret, which is a disadvantage in our case.

The reader can consult [2, 23, 24] for a comprehensive introduction to classification and classifiers.

V. Data Description And Tools Used

In this study, we use the data of four academic cohorts or batches of Computer Science & Information

Technology (CSIT) department at NED University, Pakistan, which entailed altogether 347 undergraduate students enrolled in the academic batches of 2005–06, 2006–07, 2007–08 and 2008–09. The data contains variables related to students' pre-university marks used to select the students' prior entrance to university and examination marks of the courses that are taught in the first and second academic years of their study. Adj_Marks, Maths_Marks and MPC are variables associated with the admission data of students defined as follows: Adj_marks are the total marks in HSC Examination, Maths_Marks are the marks in mathematics, and MPC is the sum of the marks in mathematics, physics and chemistry in HSC examination. The rest of the variables are the examination marks of students in individual subjects from the first and second academic years. Admission data and the most important courses for this study are explained in Table 1. The data was gathered and consolidated from two university student databases. An integrated database was formed using Oracle 9i.

The mark at the end of the degree is calculated as follows. It is the sum of 10% of the first year average examination mark, 20% of second year, 30% of third year and 40% of fourth year average examination mark. At the time of graduation, the University awards class to the students as follows: First Division with Distinction (80% marks or above), First Division (marks between 60% and 80%) or Second Division (marks between 50% and 60%). An earlier work [25] has shown that the division can be predicted with an accuracy of more than 90% for the CSIT Department using only the first and second year average examination marks although they have little weight in computing the division as compared to the third and fourth year marks. In the present research, we want to investigate the first and second academic years in more details by considering the individual courses that are taught in these years and not only the average examination marks. In this way, we seek to identify those courses where more attention has to be focused so as to improve the students' overall performance at the end of the degree. Furthermore, instead of predicting division, in this study the output variable or target to be predicted is the interval of the graduation mark that has five possible values: A (90%–100%), B (80%–89%), C (70%–79%), D (60%–69%), and E (50–59%). Divisions classify students into 3 classes and intervals classify them into 5 classes and thus give a more precise measurement for success. One might wonder that the class F for fail is missing. Because of a strict selection process, the dropout rate of the students from the University is hardly 5% and very few fail in 4th year, and therefore not considered in this study. Batches and interval statistics of different batches are presented in Table 2.

Table 1. Variables in dataset

Role	Name	Description	Range of Dataset I	Range of Dataset II
target	Interval	5 possible values(A,B,C,D,E)	A(2),B(22),C(38),D(8),E(2)	A(1),B(41),C(46),D(14),E(4)
predictor	Adj_Marks	HSC Examination total marks	[791.00; 836.00]	[737.00; 949.00]
predictor	Maths_Marks	HSC Examination Mathematics marks	[115.00; 191.00]	[95.00; 193.00]
predictor	MPC	Maths+ Physics+ Chemistry marks	[397.00; 506.00]	[389.00; 561.00]
predictor	CT-153	Programming Languages	[41.00; 95.00]	[40.00; 99.00]
predictor	CT-157	Data Structures Algorithms and Applications	[38.00; 99.00]	[40.00; 96.00]
predictor	CT-158	Fundamentals of Information Technology	[40.00; 95.00]	[52.00; 91.00]
predictor	HS-205/206	Islamic Studies or Ethical Behaviour	[52.00; 85.00]	[44.00; 82.00]
predictor	MS-121	Applied Physics	[40.00; 90.00]	[40.00; 98.00]
predictor	CS-251	Logic Design and Switching Theory	[40.00 ; 94.00]	[34.00; 88.00]
predictor	CS-252	Computer Architecture and Organization	[36.00; 92.00]	[40.00; 95.00]
predictor	CT-251	Object Oriented Programming	[37.00; 95.00]	[40.00; 87.00]
predictor	CT-254	System Analysis and Design	[43.00 100.00]	[51.00; 90.00]
predictor	CT-255	Assembly Language Programming	[41.00; 94.00]	[36.00; 96.00]
predictor	CT-257	Data Base Management System	[43.00; 97.00]	[42.00; 92.00]
predictor	EL-238	Digital Electronics	[49.00; 93.00]	[40.00; 90.00]
predictor	HS-207	Financial Accounting and Management	[43.00; 95.00]	[40.00; 95.00]

Table 2. Statistics of batches and intervals

Academic Batch	Total No. of students	Total No. of instances in 'A' Interval	Total No. of instances in 'B' Interval	Total No. of instances in 'C' Interval	Total No. of instances in 'D' Interval	Total No. of instances in 'E' Interval
2005–06	72	2	22	38	8	2
2006–07	65	-	23	31	10	1
2007–08	106	1	41	46	14	4
2008–09	104	-	31	54	18	1

We made two datasets of the gathered data namely Dataset I and Dataset II. In Dataset I we used the data of the academic year 2005–06 as the training data and the data of academic year 2006–07 as the testing data, and in Dataset II, training data is of the academic year 2007–08, and testing data is of the academic year 2008–09. Because of a change in the curriculum, data of the academic year 2006–07 has only been used partially as training data to predict the performance of the 2007–08 batch (called Dataset III) in a later stage of the present study.

The tool RapidMiner 5.3 [26] was used for exploration, statistical analysis and mining of the data. To predict the graduation performance, several data mining classification algorithms have been used like decision trees with information gain, Gini index and accuracy, rule induction with information gain, 1-nearest neighbor with Euclidean distance, naive Bayes and neural networks. The default values proposed by RapidMiner were adopted. We have also applied other classifiers like decision tree with gain ratio, rule induction with accuracy, linear regression and support vector machines on both datasets. Their results are not presented in the next section because rule induction with accuracy and linear regression performed poorly on both the datasets while support vector machines and decision tree with gain ratio performed well on Dataset I but poorly on Dataset II.

The results of the decision trees and rule induction algorithms are important for our study, although other classifiers also give good or even better results. The first reason is that the classification model given by these two methods is user friendly as it represents rules which are easily interpretable by humans and therefore can be used in making policies. A second reason is that we can use them to discover courses in first and second years that are good predictors of the students' performance at the end of the degree.

VI. Analysis and Results

Datasets used in this study contained the students' pre-admission data and the examination scores of the courses of first and second academic years as described in Section IV. Admission data and the examination marks of students in individual subjects from the first and second academic years have been used to predict the students' overall performance at the end of the degree.

A. Trying out classifiers to predict graduation performance

The literature review in a previous section shows that in general there is no classifier that outperforms all the others in all situations. Therefore trials have to be performed to discover which classifiers work better with the data at hand.

As Table 2 shows, the repartition of the students among the intervals is unbalanced. Class 'C' interval contains the most students. Predicting a student class 'C' would have an accuracy of 47.69% on Dataset I and of 51.92% on Dataset II. These two accuracies formed our baseline that we sought to improve. Table 3 shows the accuracy results for the classifiers that do better than the baseline on both datasets. Figure 1 summarizes the results of the classifiers graphically.

Table 3. Comparison of Prediction Accuracies for Dataset I and Dataset II

Criterion	Dataset I	Dataset II
Decision Tree with Gini Index(DT-GI)	60.00%(with minimal leaf size 8)	68.27%(with minimal leaf size 2)
Decision Tree with Information Gain(DT-IG)	61.54%(with minimal leaf size 2)	69.23%(with minimal leaf size 6)
Decision Tree with Accuracy(DT-Acc)	60.00%(with minimal leaf size 4)	60.58%(with minimal leaf size 2)
Rule Induction with Information Gain(RI-IG)	55.38%	55.77%
1-NN	66.15%	74.04%
Naive Bayes	64.62%	83.65%
Neural Networks(NN)	61.54%	62.50%

Generally the classifiers gave better results on Dataset II, may be because the sets were bigger: There were more instances to train a better model. In comparison of all classification methods, decision tree with accuracy, rule induction with information gain and neural networks performed in a similar manner for Dataset I and Dataset II. Among all 3 criterions information gain gave the best results for decision trees. 1-NN and Naive Bayes outperformed all the classifiers for both datasets. Particularly on Dataset II, the accuracy of Naive Bayes reached 83.65%, which is a very good result. However the results of these two classifiers are not easy to interpret and therefore not actionable: One does not know which courses could be an indicator of poor performance for students, and hence could help to take action.

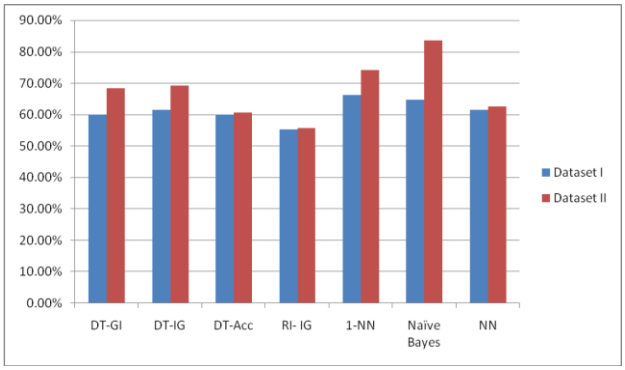


Fig. 1. Classification algorithms performance comparison

The resultant confusion matrices of this experiment are shown in Fig. 2. To understand these confusion matrices, let's take an example of the first confusion matrix of the classifier "Decision Tree with Gini Index'. In this confusion matrix, of the 23 (13+10) actual class 'B' students, the classifier predicted correctly 13 as 'B' and wrongly 10 as 'C'; from the 31 actual class 'C' students, 4 were predicted class 'B' and 3 class 'D'; similarly from the actual 10 class 'D' students, 8 were predicted class 'C', and the only actual 1 student who belongs to class 'E' was predicted 'D'. All correct predictions are located in the diagonals of the table. It's easy to visually check Fig. 2 for misclassifications, as all the incorrect predictions are present outside the diagonals.

Decision Tree with Gini Index		Dataset I Actual					Dataset II Actual				
		B	C	D	E	Class precision	B	C	D	E	Class precision
Predicted	B	13	4	0	0	76.47%	18	6	0	0	75.00%
	C	10	24	8	0	57.14%	13	38	2	0	71.70%
	D	0	3	2	1	33.33%	0	10	14	0	58.33%
	E	0	0	0	0	0.00%	0	0	2	1	33.33%
Class Recall		56.25 %	77.42 %	20.00 %	0.00 %		58.06 %	70.37 %	77.78 %	100.0 0%	

Decision Tree with Information Gain		Dataset I Actual					Dataset II Actual				
		B	C	D	E	Class precision	B	C	D	E	Class precision
Predicted	B	13	2	1	0	81.25%	16	4	0	0	80.00%
	C	9	25	6	0	62.50%	15	40	2	0	70.18%
	D	1	2	1	0	25.00%	0	10	16	1	59.26%
	E	0	2	2	1	20.00%	0	0	0	0	0.00%
Class Recall		56.25 %	80.65 %	10.00 %	100. 00 %		51.61 %	74.07 %	88.89 %	0.00 %	

Decision Tree with Accuracy		Dataset I Actual					Dataset II Actual				
		B	C	D	E	Class precision	B	C	D	E	Class precision
Predicted	B	14	6	0	0	70.00%	4	0	0	0	100.00%
	C	9	25	10	1	55.56%	27	44	2	0	60.27%
	D	0	0	0	0	0.00%	0	10	14	0	58.33%
	E	0	0	0	0	0.00%	0	0	2	1	33.33%
Class Recall		60.87 %	80.65 %	0.00 %	0.00 %		12.90 %	81.48 %	77.78 %	100.0 0%	

Rule Induction with Information Gain		Dataset I Actual					Dataset II Actual				
		B	C	D	E	Class precision	B	C	D	E	Class precision
Predicted	B	7	2	1	0	70.00%	22	8	3	1	57.89%
	C	16	28	8	1	52.83%	6	23	2	0	74.19%
	D	0	1	1	0	50.00%	3	23	13	0	33.33%
	E	0	0	0	0	0.00%	0	0	0	0	0.00%
Class Recall		30.43 %	90.32 %	10.00 %	0.00 %		70.97 %	42.59 %	72.22 %	0.00 %	

1-NN		Dataset I Actual					Dataset II Actual				
		B	C	D	E	Class precision	B	C	D	E	Class precision
Predicted	B	12	2	0	0	85.71%	26	11	0	0	70.27%
	C	10	28	6	0	63.64%	5	38	5	0	79.16%
	D	1	1	3	1	50.00%	0	5	13	1	68.42%
	E	0	0	1	0	0.00%	0	0	0	0	0.00%
Class Recall		52.17 %	90.32 %	30.00 %	0.00 %		83.87 %	70.37 %	72.22 %	0.00 %	

Naive Bayes		Dataset I Actual					Dataset II Actual				
		B	C	D	E	Class precision	B	C	D	E	Class precision
Predicted	B	11	0	0	0	100.00%	28	6	0	0	82.35%
	C	12	24	3	0	61.54%	3	47	6	0	83.92%
	D	0	7	7	1	46.67%	0	1	12	1	85.71%
	E	0	0	0	0	0.00%	0	0	0	0	0.00%
Class Recall		47.83 %	77.42 %	70.00 %	0.00 %		90.32 %	87.04 %	66.67 %	0.00 %	

Neural Networks		Dataset I Actual					Dataset II Actual				
		B	C	D	E	Class precision	B	C	D	E	Class precision
Predicted	B	11	2	0	0	84.62%	30	24	0	0	55.56%
	C	12	29	10	1	55.77%	1	23	3	0	85.18%
	D	0	0	0	0	0.00%	0	6	12	1	63.15%
	E	0	0	0	0	0.00%	0	1	3	0	0.00%
Class Recall		47.83 %	93.55 %	0.00 %	0.00 %		96.77 %	42.59 %	66.67 %	0.00 %	

Fig 2. Confusion Matrices of Dataset I and Dataset II

"Fig. 2" revealed that the class 'C' interval, our majority class, was better predicted by most classifiers, as the line recall shows. Recall is the ratio of the number of predicted elements and the number of actual elements. Previous studies also commented on predicting well the largest classes that contain the majority elements [19, 20]. Many classifiers are optimistic for class D in Dataset I: they predict most actual D students as C.

From Table 3 and Fig. 1, it is clear that the first research question is answered positively i.e. the performance of students at the end of their degree can be predicted with a reasonable accuracy using their marks in HSC and in first and second year courses. We wished to identify the courses at an early stage that could be effective predictors of students' performance at the end of the degree to answer the subsequent part of our research question. From the classifiers with interpretable models that could help identify those courses, decision trees gave the best results. In the sequel we present our endeavours to improve the accuracy of all classifiers and particularly the one of decision trees.

B. Trying to Improve Accuracy

Table 2 shows that the intervals/classes are not balanced. It is known that unbalanced datasets can lead to a poor accuracy. To balance the classes all the samples from the minority classes (i.e. the 'A' interval, 'D' interval and 'E' interval) were taken and copied multiple times in the dataset to nearly balance the classes. All prediction models using the balanced datasets were redeveloped and their accuracy was compared to the accuracy of the original models, but there was no improvement: All of these models had lower prediction accuracy with rebalanced datasets. The attempt of including minority classes data from earlier cohorts led also to poorer results, which suggests that timeliness of the data matters.

Second, feature selection techniques have been used to choose a subset of variables and eliminate others that could be irrelevant or of no predictive information and therefore could prevent the classifiers from reaching a good accuracy. The Recursive Feature Elimination (RFE) operator available in RapidMiner and employed in this study 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 select subsets of variables. For all four criterions, the number of features to select has been fixed to 8. The reason for fixing the number 8 was that when the decision trees were built with the full set of attributes, the decision tree with Gini index used 9 attributes, decision tree with information gain 8 and decision tree with accuracy 7, which gives an average of 8. In other words, the decision tree algorithm did perform a selection of the variables and this fact has been used later in this study. The four criterions of the RFE operator did not select the same 8 features so four different subsets of variables were returned. It is interesting to observe that all four subsets did not contain the admission marks. This means that admission marks do not seem to play an essential role in student's university performance.

However, admission marks are important in selecting the students for admission at NED University. Because of a strict selection process there is not much dropout of students from the University. The prediction models of Table 3, i.e. decision tree with Gini index (DT-GI), decision tree with information gain (DT-IG), decision tree with accuracy (DT-Acc), rule induction with information gain (RI-IG), 1-NN, naive Bayes (NB) and neural networks (NN) were built again using these four subsets of variables. Fig. 3 and Fig. 4 give the results of feature selection algorithms for Dataset I and Dataset II. We can see from the figures that the classification accuracies obtained using only the attributes given by the feature selection technique are not higher than the results obtained with the full set of attributes except for RFE-GI. Altogether, there are 14 classifiers for Dataset I and Dataset II, out of which RFE- GI stays same or improves for 8 cases. However, RFE-GI performs less well for Dataset II in general contradicting the findings of Table 3. Rule Induction with RFE-GI performs better on Dataset II, but still less well than other classifiers without selection of features.

In order to identify a subset of variables that could improve the accuracy of all classifiers, we selected those features that were common in four, three or two subsets given by the feature selection techniques mentioned above. This gave a total of 9 features. A dataset restricting the variables to these 9 features only was formed and the classifiers were applied again. However, the classification accuracies were not higher.

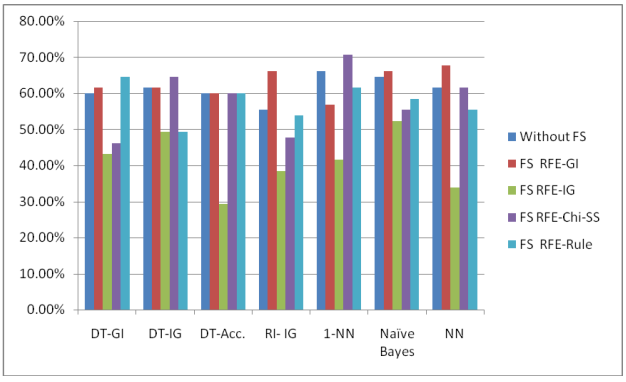


Fig. 3. Comparison of classifiers accuracy for Applying Feature Selection on Dataset I

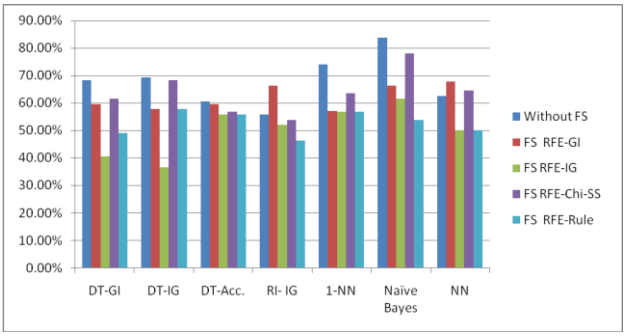


Fig. 4. Comparison of classifiers accuracy for applying feature selection on Dataset II

C. Improving Accuracy of Decision Trees with 5 Courses

As noticed above, decision trees not only classified the data, but did also some selection of the attributes. We built decision trees employing four criterions Gini index, information gain, gain ratio and accuracy, using the dataset with the full set of attributes. Decision tree with gain ratio was included to have a kind of majority. Those features that were present in all the four, three or two trees were selected. They were 5 features for Dataset I and 8 features for Dataset II. The 5 features included 2 courses from first year and 3 courses from second year, see Table 4. The 8 features include two courses from first year and 6 courses from second year, and are listed also in Table 4. These two subsets had 3 courses in common: MS-121, a first year course, and CS-251 and CT-255, two second year courses. The meaning of these courses is given in Table 1. First the 5 features for Dataset I and the 8 features for Dataset II were used with the same seven classifiers. The results are presented in Table 4. As far as Dataset I is concerned, we can see that accuracy stays the same or improves for all techniques, except for rule induction with information gain and 1-NN. For Dataset II the picture is quite different: Accuracy diminishes for all methods.

Next we switched the selected features, which meant that the 8 features that were selected for Dataset II were applied with Dataset I, and the 5 features that were selected for Dataset I were applied with Dataset II. The results are presented in Table 5. As far as Dataset I with 8 features was concerned, accuracy became worse for five methods, did not change for Naives Bayes and improved only for Neural Networks. For Dataset II with 5 features, accuracy improved for Decision Tree with Gini Index, stays the same for Decision Tree with Information Gain, and decreased for the other five methods.

Summing up, from the 14 models obtained on both Dataset I and Dataset II with k=5, accuracy stayed the same or improved in 7 cases. These 5 features tended to improve the accuracy of the decision trees. The 6 decision trees obtained with 5 features are shown in the appendix.

VII. Discussion and Practical Implications

Table 3 shows that it is possible to improve the baseline a lot and to answer positively the first research question. Table 4 and Table 5 show that it is possible to improve the accuracy of the decision trees and reach 73.08%, a nice result, but not to the extent of doing better than in Table 3.

A. Indicators of very good and low performance, and a pragmatic policy

By looking at the trees in the Appendix, one notices two indicators of very good performance: HS-207 and CT-255. A high performance in HS-207 leads to a leaf with graduation performance B or B mixed with A in the 3 trees of Dataset I and one tree of Dataset II and a high performance in CT-255 leads to a leaf with graduation performance B or B mixed with A in one tree of Dataset I

and 2 trees of Dataset II. This suggests that students having a mark bigger or equal to 80 in HS-207 and bigger or equal to 86 in CT-255 are likely to achieve their degree with a mark in the A or B interval. This suggests also that students having 80 or more in HS-207 are likely to obtain 80 or more in other subjects as well because of the way the final mark is calculated.

Table 4. Comparison of Prediction Accuracies after Applying Feature Selection for Dataset I and Dataset II

Criterion	**Without feature selection** **Dataset I**	**Features selected by Decision Trees** **K=5** **Selected Features (HS-205/206, MS-121,CS-251, HS-207, C T-255)** **Dataset I**	**Without feature selection** **Dataset II**	**Features selected by Decision Trees** **K=8** **Selected Features (CT-153, MS-121 CS-251, CS-252, CT-254, CT-255, CT-257, EL-238)** **Dataset II**
Decision Tree with Gini Index	60.00% (with minimal leaf size 8)	60.00% (with minimal leaf size 8)	68.27% (with minimal leaf size 2)	63.46% (with minimal leaf size 12)
Decision Tree with Information Gain	61.54% (with minimal leaf size 2)	64.62% (with minimal leaf size 10)	69.23% (with minimal leaf size 6)	67.31% (with minimal leaf size 8)
Decision Tree with Accuracy	60.00% (with minimal leaf size 4)	60.00% (with minimal leaf size 2)	60.58% (with minimal leaf size 2)	59.62% (with minimal leaf size 2)
Rule Induction with Information Gain	55.38%	50.77%	55.77%	44.23%
1-NN	66.15%	60.00%	74.04%	73.08%
Naive Bayes	64.62%	66.15%	83.65%	72.12%
Neural Networks	61.54%	67.69%	62.50%	56.73%

Table 5. Comparison of Prediction Accuracies after applying feature selection for Dataset I and Dataset II

Criterion	**Without feature selection** **Dataset I**	**Features selected by Decision Trees** **K=8** **Selected Features (CT-153, MS-121 CS-251, CS-252, CT-254, CT-255, CT-257, EL-238)** **Dataset I**	**Without feature selection** **Dataset II**	**Features selected by Decision Trees** **K=5** **Selected Features (HS-205/206, MS-121,CS-251, HS-207, CT-255)** **Dataset II**
Decision Tree with Gini Index	60.00% (with minimal leaf size 8)	53.85% (with minimal leaf size 6)	68.27% (with minimal leaf size 2)	73.08% (with minimal leaf size 4)
Decision Tree with Information Gain	61.54% (with minimal leaf size 2)	58.46% (with minimal leaf size 2)	69.23% (with minimal leaf size 6)	69.23% (with minimal leaf size 6)
Decision Tree with Accuracy	60.00% (with minimal leaf size 4)	30.77% (with minimal leaf size 2)	60.58% (with minimal leaf size 2)	56.73% (with minimal leaf size 2)
Rule Induction with Information Gain	55.38%	56.92%	55.77%	52.88%
1-NN	66.15%	63.08%	74.04%	60.58%
Naive Bayes	64.62%	64.62%	83.65%	59.62%
Neural Networks	61.54%	67.69%	62.50%	56.73%

One notices also two indicators of low performance: CS-251 and HS-207. A low performance in CS-251 leads to a leaf with label D or E in one tree of Dataset I and 3 trees of Dataset II and a low performance in HS-207 leads to a leaf with label D or E in one tree of Dataset I and 2 trees of Dataset II. This suggests that students having a mark in lower than 43 in CS-251 or lower than 60 in HS-207 are likely to achieve their degree with a poor mark. This suggests also that students having 60 or less in HS-207 are likely to obtain 60 or less in other subjects as well again because of the way the final mark is calculated.

The 2 indicators of low performance are courses of second year and therefore cannot help to warn students in first year. MS-121 and HS-205/206 are courses in first year. MS-121 should be taken as additional indicator as mark lower than 63 leads to a leaf with label D or E in 2 trees of Dataset II. This can be summarized as follows:

- In first year, those students whose marks are around or less than 63 in MS-121, are likely to have a mark in the 'D' or 'E' interval at the end of the degree.
- In second year, students whose marks are below 60 in HS-207 or whose marks are below 43 in CS-251, are likely to have a mark in the 'D' or 'E' interval at the end of the degree.
- In second year, students whose marks are equal or higher than 80 in HS-207 or students whose marks are bigger than 86 in CT-255, are likely to have a mark in the 'A' or 'B' interval at the end of the degree.

These findings are sensible and can be used to implement some policy. For instance, the instructors of the course MS-121 in first year could report about students with marks equal or less than 63. These students are at risk and they need more academic assistance. A similar reporting could take place in second year with reference to the courses HS-207 and CS-251. These suggestions may help the University to pay extra attention to those students who require more academic facilitation e.g., extra classes or extra consultation hours with the instructors. On the contrary students with high marks in HS-207 or CT-255 could be selected for special advanced program in third year.

B. Reflecting on the Indicators

Surprisingly, the five courses selected through the decision tree feature selection technique include three non-core courses (MS-121 and HS-205/206 in first year and HS-207 in second year), which are in general not seen as decisive courses for the degree. This sounds different from the findings reported in [17]. Therefore, we investigated the correlation of these non-core courses with the core courses of first and second years. The results of correlations are presented in Table 6.

Table 6. Correlation Results between non-core courses and core courses

	Core Courses	**Batch 2005–06**	**Batch 2006–07**	**Batch 2007–08**	**Batch 2008–09**	**Avg. Correlation**
HS-205/206	CT-153	0.369	0.320	0.310	0.395	0.3485
HS-205/206	CT-157	0.468	0.066	0.263	0.442	0.30975
HS-205/206	CT-158	0.456	0.507	0.341	0.633	0.48425
HS-205/206	CT-251	0.289	0.359	0.192	0.465	0.32625
HS-205/206	CT-254	0.391	0.347	0.448	0.468	0.4135
HS-205/206	CT-257	0.578	0.409	0.493	0.418	0.4745
HS-205/206	CS-252	0.431	0.300	0.390	0.402	0.38075
MS-121	CT-153	0.633	0.455	0.554	0.607	0.56225
MS-121	CT-157	0.627	0.358	0.612	0.495	0.523
MS-121	CT-158	0.473	0.245	0.423	0.581	0.4305
MS-121	CT-251	0.385	0.460	0.536	0.484	0.46625
MS-121	CT-254	0.568	0.304	0.510	0.424	0.4515
MS-121	CT-257	0.625	0.359	0.605	0.336	0.48125
MS-121	CS-252	0.543	0.451	0.582	0.567	0.53575
HS-207	CT-153	0.527	0.400	0.516	0.430	0.46825
HS-207	CT-157	0.628	0.299	0.505	0.489	0.48025
HS-207	CT-158	0.616	0.350	0.465	0.523	0.4885
HS-207	CT-251	0.447	0.629	0.534	0.395	0.50125
HS-207	CT-254	0.530	0.454	0.482	0.523	0.49725
HS-207	CT-257	0.640	0.598	0.644	0.437	0.57975
HS-207	CS-252	0.737	0.545	0.693	0.588	0.64075

We can see from Table 6 that HS-205/206 correlates positively but relatively weakly with the core courses of first and second years. MS-121 and HS-207 correlate better with the core courses of first year and second year, supporting the proposition of selecting these two courses as indicators of particularly weak or strong results in the degree and supporting the findings of the decision trees.

VIII. Conclusion

The present study shows that it is possible to predict the graduation performance in 4th year at university using only pre-university marks and marks of 1st and 2nd 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. Thus the first research question is answered positively. Naïve Bayes has given an accuracy of 83.65% on Dataset II. The accuracy obtained in this study is better than the one obtained in related works that have used socio-economic or demographic features and pre-university marks, but no marks at university level like [20]. This suggests that including marks obtained in the first semesters or year of university is important to obtain a reasonable accuracy. Other related works that have obtained a good accuracy did include marks at university level.

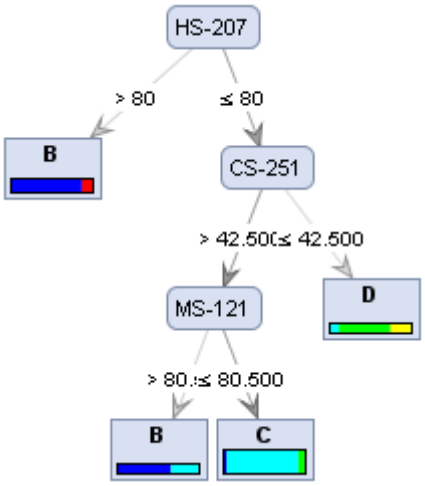
The investigation of the second research question led to identify 5 courses to predict the graduation performance. Considering this set of 5 courses only instead of pre-university subjects, all 1st year and 2nd year courses tends to increase the accuracy of the decision trees. However, the five selected courses do not lead to a better accuracy with the naive Bayes and 1-NN classifiers, which gave the best accuracy in the first place. Even if these five courses allow finding sensible course that are detectors of poor or strong performance in first and second year, some more research is needed to understand these limited findings.

This set of five courses includes 2 courses of 1st year and 3 courses of 2nd year, i.e. a majority of courses nearer to the graduation. These findings are consistent with the findings in [18] that conclude the following: marks in 3rd year Bachelor are better predictors of performance at the Master level than marks in 1st year Bachelor. These 5 courses are made up of 2 core courses and 3 courses seen as non-core courses of the degree, which came as a surprise for the faculty members. Though the non-core courses correlate positively with the core courses, some further work is needed to investigate in more depth this matter. However other works have shown that data mining results do not always match the beliefs of faculty members, as reported for example in [3].

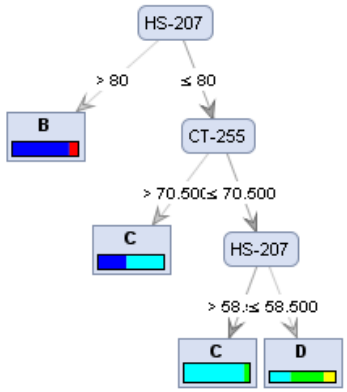
As already mentioned in the related works, the comparative work of [15] gives accuracies of 81% to 91% while predicting performance at a course level, which can be seen as an easier task than predicting performance at graduation level. Therefore it might be difficult to predict performance at a degree level with some accuracy well above 80%-85%. This might not be a limit of the data or the classifiers used but reflect the fact that students develop during their studies. Therefore another future work is to study progression of students during their 4 years of bachelor and investigate whether typical developments can be identified. Work along this line has already started.

Appendix A Dataset I Decision Trees with K=5

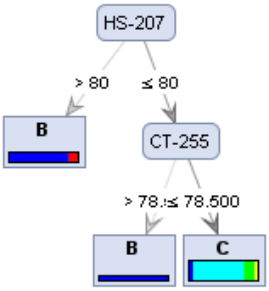
Decision Tree with Gini Index



Decision Tree with Information Gain

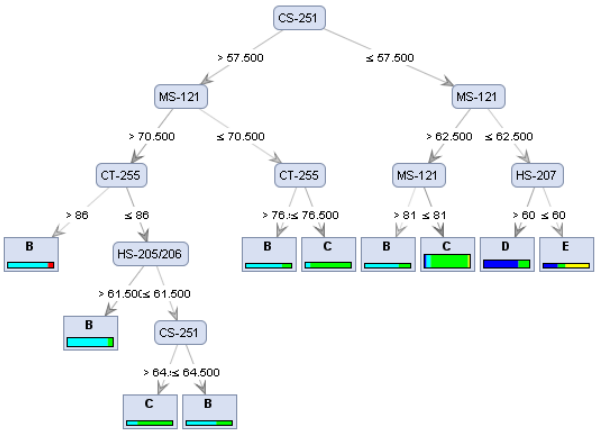


Decision Tree with Accuracy

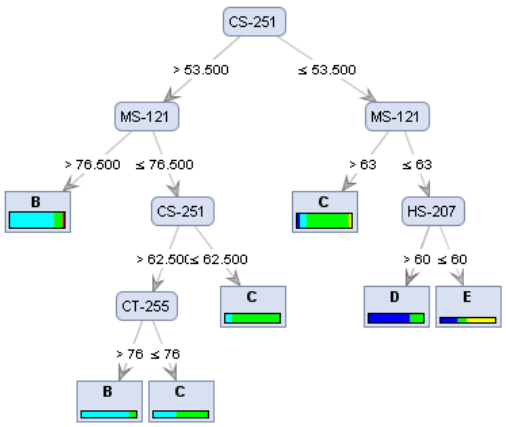


Appendix B Dataset II Decision Trees with K=5

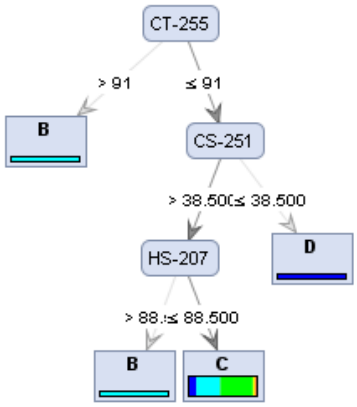
Decision Tree with Gini Index



Decision Tree with Information Gain



DECISION TREE WITH ACCURACY



ACKNOWLEDGMENT

We thank Dr. Sajida Zaki for proof reading this paper. This work is supported in part by a grant from NED University of Engineering & Technology, Karachi, Pakistan.

REFERENCES

[1] D. P. Acharjya ,D. Roy, and M.A. Rahaman, "Prediction of Missing Associations Using Rough Computing and Bayesian Classification," *International Journal of Intelligent Systems and Applications*, vol. 11, pp. 1-13, 2012. DOI: 10.5815/ijisa.2012.11.01

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

[3] G. Dekker, M. Pechenizkiy, and J. Vleeshouwers, "Predicting Students Drop Out: a Case Study," 2nd International Conference on Educational Data Mining, Proceedings. Cordoba, Spain, pp. 41-50, 2009.

[4] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, pp. 498–506, 2010.

[5] Z. J. Kovačić, "Predicting student success by mining enrolment data," *Research in Higher Education Journal*, vol. 15, pp. 1–20, 2012.

[6] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: predicting at-risk students by analysing clicking behaviour in a virtual learning environment," Proceedings of the Third International Conference on Learning Analytics and Knowledge, pp. 145-149, 2013.

[7] E. Aguiar, N.V. Chawla, J. Brockman, G. A. Ambrose, V. Goodrich, "Engagement vs Performance: Using Electronic Portfolios to predict first semester engineering student retention," International Conference on Learning Analytics and Knowledge, ACM, 2014.

[8] S. Valsamidis, and S. Kontogiannis, "E-Learning Platform Usage Analysis," *Interdisciplinary Journal of E-Learning and Learning Objects*, vol. 7, pp. 185-204, 2011.

[9] A. Merceron, and K. Yacef, "Measuring Correlation of Strong Symmetric Association Rules in Educational Data," In: Handbook of Educational Data Mining , edited by C. Romero, S. Ventura, M. Pechenizkiy & R.S.J.d. Baker, CRC Press, ISBN: 978-1-4398-0457-5, pp. 245 -256. 2010

[10] C. Romero, S. Ventura, M. Pechenizkiy, and R.S.J.d. Baker, Handbook of Educational Data Mining. CRC Press, 2010, ISBN: 978-1-4398-0457-5.

[11] C. Romero, and S. Ventura, "Educational Data Mining: A Review of the State of the Art," *IEEE transactions on Systems, Man and Cybernetics*, vol. 40(6), pp.601-618, 2010.

[12] Z. Pardos, N. Hefferman, B. Anderson, and C. Hefferman, "The effect of Model Granularity on Student Performance Prediction Using Bayesian Networks," Proceedings of the international Conference on User Modelling, Springer, Berlin, pp. 435-439, 2007

[13] E. Galy, C. Downey, and J. Johnson, " The Effect of Using E-Learning Tools in Online and Campus-based Classrooms on Student Performance," *Journal of Information Technology Education*, vol. 10, pp. 209-230, 2011.

[14] M. I. Lopez, R. Romero, V. Ventura, and J.M. Luna," Classification via clustering for predicting final marks starting from the student participation in Forums," In (Yacef, K., Za ïane, O., Hershkovitz, H., Yudelson, M., and Stamper, J. Hrsg.): Proceedings of the 5th International Conference on Educational Data Mining, Chania, Greece, June15-21, pp. 148-151 ,2012.

[15] S. Huang, N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Computer and Education*, pp. 133-145, 2013.

[16] P. Golding, S. McNamarah, "Predicting Academic Performance in the School of Computing & Information Technology (SCIT)," Proceedings of 35th ASEE /IEEE Frontiers in Education Conference, 2005.

[17] P. Golding, O. Donaldson, "Predicting Academic Performance", Proceedings of 36th ASEE /IEEE Frontiers in Education Conference, 2006.

[18] J. Zimmermann, K. H. Brodersen, J. P. Pellet, E. August, J. M. Buhmann, "Predicting graduate-level performance from undergraduate achievements," Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands. July 6-8, 2011.

[19] T. N. Nghe, P. Janecek, P. Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance," Proceedings of 37th ASEE /IEEE Frontiers in Education Conference, 2007.

[20] D. Kabakchieva , K. Stefanova, V. Kisimov, Analyzing University Data for Determining Student Profiles and Predicting Performance, Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands. July 6-8, 2011.

[21] D. Kabakchieva, Predicting Student Performance by Using Data Mining Methods for Classification, *Cybernetics and Information Technologies*, vol. 13, No. 1, pp. 61-72, 2013.

[22] S. Haykin, Neural Networks: A comprehensive Foundation. 2nd ed. Prentice Hall, Upper Saddle River, New Jersey, 1999, p.157, 171, 184.

[23] P. N. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining. 1st ed. Pearson Addison Wesley, US ed edition, 2005, p. 148-149.

[24] B. Liu, Web Data Mining – Exploring Hyperlinks, Contents and Usage Data. Springer. 2011.

[25] R. Asif, A. Merceron, M. K. Pathan, "Mining Student's Admission Data and Predicting Student's Performance using Decision Trees," Proceedings of the 5th International Conference of Education, Research and Innovation, Madrid: Spain, pp. 5121-5129, 2012.

[26] RapidMiner retrieved from www.rapid-i.com

Authors' Profiles

Raheela Asif received her master's degree in Computer Science & I.T. in 2008 from NED University of Engineering & Technology, Karachi, Pakistan. She is a Ph. D student in the same university and also serving the university as Assistant Professor since 2006. Her research interests include Educational Data Mining, Data Bases and Artificial Intelligence.

Agathe Merceron received her master's degree in Applied Mathematics in 1979 and her Ph. D. in Computer Science in 1981 from the University Paris VII, France, and her habilitation in Computer Science in 1986 from the University Paris XI, France. After occupying different positions in several countries (France, Germany and Australia) she has been a professor of Computer Science at Beuth University of Applied Sciences, Berlin, Germany since 2006. She is responsible for the online-degrees Bachelor and Master Computer Science and Media. Her current research interests include Information Systems and Knowledge Management, application to E-learning, Technology Enhanced Learning, Educational Data Mining and Learning Analytics. She is a member of the board of the international educational data mining society.

Mahmood Khan Pathan received his M.Sc. degree in Pure Mathematics in 1974 from the University of Karachi and Ph. D. degree in Applied Algebra from Brunel, The University of West London, United Kingdom in 1992. He served NED University of Engineering & Technology, Karachi-Pakistan as Dean of Faculty of Information, Sciences & Technology from 2005 to 2013. His research interests include finite fields, cryptography and Educational Data Mining.

