Computers & Education 113 (2017) 177–194

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Analyzing undergraduate students' performance using educational data mining

Raheela Asif [a, *], Agathe Merceron [b], Syed Abbas Ali [c], Najmi Ghani Haider [a]

[a] N.E.D University of Engineering & Technology, Department of Computer Science & Software Engineering, Karachi, 75270, Pakistan
[b] Beuth University of Applied Sciences, Department of Computer Science and Media, Berlin, 13353, Germany
[c] N.E.D University of Engineering & Technology, Department of Computer and Information Systems Engineering, Karachi, 75270, Pakistan

ARTICLE INFO

Article history:
Received 23 December 2015
Received in revised form 21 April 2017
Accepted 17 May 2017
Available online 22 May 2017

Keywords:
Data mining
Decision trees
Clustering
Performance prediction
Performance progression
Quality of educational processes

ABSTRACT

The tremendous growth in electronic data of universities creates the need to have some meaningful information extracted from these large volumes of data. The advancement in the data mining field makes it possible to mine educational data in order to improve the quality of the educational processes. This study, thus, uses data mining methods to study the performance of undergraduate students. Two aspects of students' performance have been focused upon. First, predicting students' academic achievement at the end of a four-year study programme. Second, studying typical progressions and combining them with prediction results. Two important groups of students have been identified: the low and high achieving students. The results indicate that by focusing on a small number of courses that are indicators of particularly good or poor performance, it is possible to provide timely warning and support to low achieving students, and advice and opportunities to high performing students.

1. Introduction

1.1. Background and theoretical framework

Owing to digitization of academic processes, universities are generating a huge amount of data pertaining to students in electronic form. It is crucial for them to effectively transform this massive collection of data into knowledge which will help teachers, administrators and policy makers to analyze it to enhance decision making. Furthermore, it may also advance the quality of the educational processes by providing timely information to different stakeholders. The purpose of data mining methods is to extract meaningful knowledge from data (Han & Kamber, 2006). The application of data mining methods to educational data is referred to as Educational Data Mining (EDM) (Baker & Yacef, 2009).

Baker (2010) proposes five primary categories or approaches in EDM: prediction, clustering, relationship mining, discovery within models, and distillation of data for human judgment. The present work combines three approaches: prediction, clustering and, to some extent, distillation of data for human judgment.

* Corresponding author.
E-mail addresses: engr_raheela@yahoo.com (R. Asif), merceron@beuth-hochschule.de (A. Merceron), saaj.research@gmail.com (S.A. Ali), najmi@neduet.edu.pk (N.G. Haider).

http://dx.doi.org/10.1016/j.compedu.2017.05.007
0360-1315/

In prediction, the goal is to predict the class or label of a data object. A major key application area of prediction in EDM is predicting student educational outcomes. Research within this area has been carried out at different levels of granularity: at a tutoring system level, at course level, or at degree level etc. At the level of intelligent tutoring system (Feng, Heffernan, & Koedinger, 2006) for instance, EDM predicts students' test scores by integrating timing information and the amount of assistance a student needs to solve problems; Pardos, Hefferman, Anderson, and Hefferman (2007) have devised a system to predict whether a student is likely to get the next training exercise right, and if so, whether the tutoring system should skip it. At the course level, Strecht, Cruz, Soares, Merdes-Moreria, and Abren (2015) predict students' success/failure and grade in a course by using social variables like age, sex, marital status, nationality, displaced (whether the student lived outside the district), scholarship, special needs, type of admission, type of student (regular, mobility, extraordinary), status of student (ordinary, employed, athlete, etc.), years of enrolment, delayed courses, type of dedication (full-time, part-time), and debt situation; ElGamal (2013) predicts students' grades in a programming course by considering different factors like the students' mathematical background, programming aptitude, problem solving skills, gender, prior experience, high school mathematics grade, locality, previous computer programming experience, and e-learning usage; Huang and Fang (2013) predict course performance on the basis of students' performance in prerequisite courses and midterm examinations; Romero, Lopez, Luna, and Ventura (2013) investigated the appropriateness of quantitative, qualitative and social network information about forum usage as well as the appropriateness of classical classification algorithms and clustering algorithms to predict students' success or failure in a course; Arnold and Pistilli (2012) provide an early intervention solution for difficult courses based on students' activity in a Learning Management System. A number of studies predict students' passing/failing or overall academic achievement (total marks/CGPA) at the end of a degree programme; these studies are described in greater detail in the 'Related work' section.

In clustering, the goal is to group objects into classes of similar objects. Though clustering has been used in educational data mining for a wide variety of tasks, an interesting sub-area is grouping students to study patterns of typical behaviours. The work by Cobo et al. (2012) finds typical behaviours in forums such as high-level workers, i.e. students that read all messages and post many messages in the forum, or lurkers, i.e. students who read all messages without posting any; Bower (2010) identifies groups of students with similar performance from Kindergarten till the end of high school; while Talavera and Gaudioso (2004) cluster students' interaction data to build profiles of students.

Distillation of data for human judgment accords with what others call overview statistics and visualizations (Baker, 2010). Its aim is to help in understanding the results of analyses. For example, Elkina, Fortenbacher, and Merceron (2013) use an intuitive visualization of analytic results that provides insight about learning processes to teachers, E-learning providers and researchers. Bower's (2010) work combines dendrograms with heat map to provide an intuitive visualization of distinctive groups of students.

1.2. Research goal and questions

This study aims to analyze the performance of students pursuing a 4-year Bachelor degree programme in the discipline of Information Technology. The rationale is to provide information regarding these students' performance to the concerned teachers and study programme directors which could help them in improving the programme. The approach delineated to achieve this goal is threefold.

- Firstly, several classifiers are generated to predict the performance of students at the end of the university degree as early as possible. To build these classifiers, only admission marks from high school certificate and final marks of first and second-year courses at university are used; no socio-economic or demographic features are considered. This approach should enable the university administration to develop an educational policy that is simpler to implement. This is the reason to investigate whether acceptable results can be obtained with marks only.
- Secondly, using these classifiers we aim to derive courses that can serve as effective indicators of students' performance in a degree programme. In doing so, we will be able to support at-risk students or to stimulate further the students showing promise. Not every classifier is suitable to derive such indicator courses. A trade-off might have to be met between the predictive power of a classifier and the interpretability of its model; this is indeed the case in the present investigation. In this study, decision trees, a classifier type explained in section 3, are used to derive those indicators. With our datasets, decision trees rank first for the interpretability of the model but third in terms of accuracy. Therefore, the goodness of the indicators needs to be further investigated in order to devise a pragmatic policy for intervention.
- Thirdly, we investigate how students' academic performance progresses over the 4-year degree programme as a kind of triangulation. Using clustering techniques, we divide students into groups such that students of the same group share the same typical progression. This puts in evidence interesting typical progressions, in particular, students who have low marks all the way through their studies and students with high marks throughout their studies. The key contribution of our work is to understand the benefits of the indicators proposed in the second step. Investigating the groups of students returned by the indicators show that they include the interesting groups uncovered with clustering: students with low marks and students with high marks. Therefore, the indicator courses can be recommended to implement a pragmatic policy for intervention.

In light of the above, this study examines three questions:

Question 1: Can we predict students' performance with a reasonable accuracy at an early stage of the degree programme using marks only?

Question 2: Can we identify courses that can serve as indicators of a good or low performance at the end of the degree?

Question 3: Can we identify typical progressions of students' performance during their studies and relate them with the indicator courses?

The rest of the paper is arranged in the following order. The next section is devoted to literature review and is followed by an overview on data mining methods. Then, we describe the data and methodology for this study in section four, and the results are presented in the succeeding section. The final section presents conclusions and discusses emerging directions for future research.

2. Literature review

Our study comprises three areas of educational data mining i.e. prediction, clustering and distillation of data for human judgment. This review identifies strengths and shortcomings in the existing literature and highlights the unique contribution that the study makes to the field.

2.1. Related works on predicting students' academic performance at degree level

Golding and Donaldson (2006) investigated the relationship between students' demographic attributes, qualification on entry, aptitude test scores, performance in first year courses and their overall performance in their programme using regression technique. In their study based on the data of a single cohort comprising 85 students of the School of Computing and Information Technology at the University of Technology, Jamaica (UTECH), they found a strong correlation between performance in a first year computer science courses and the students overall performance in the programme, with a correlation of 0.499 that explains 70.6% of the students' overall performance.

Nghe, Janecek, and Haddawy (2007) applied data mining techniques in predicting students' academic performance by considering the data of two different academic institutes; Asian Institute of Technology (AIT), Thailand, and Can Tho University (CTU), Vietnam. The AIT datasets included the Master programmes. The students' GPA at the end of first year of their Master programme is predicted from their admission information, including academic institute, entry GPA, English proficiency, marital status, Gross National Income, age, gender, and TOEFL score. In the case of the CTU dataset, the students' GPA at the end of the third year is predicted using attributes such as English skill, entry marks range, field of study, faculty, gender, age, family, job, religion, and also second-year GPA. For both 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). Two data mining algorithms were applied, namely decision trees and Bayesian network. Decision trees produced better accuracies. For 2 classes the accuracy was: CTU 92.86% and AIT 91.98%; for 3 classes: CTU 84.18% and AIT 67.74%; and for 4 classes CTU 66.69% and AIT 63.25%. The accuracy of predictions was measured using a 10-fold cross-validation: 9/10 of the data was used to build the model that was tested on 1/10 of the data, and this process was repeated 10 times. Thus a single cohort was used to build the prediction model and to evaluate it.

Kabakchieva developed models to predict students' university performance based on students' personal, pre-university and university performance characteristics (Kabakchieva, 2013; Kabakchieva, Stefanova, & Kisimov, 2011). The studies encompass the data of 10,330 students in the Bulgarian educational sector. Each student was described by 20 attributes which included gender, birth year and place, place of living, and country, place and total score from previous education, current semester, and total university score. Data mining algorithms such as the decision tree C4.5, Naive Bayes, Bayesian networks, K-nearest neighbours (KNN) and rule learner's algorithms were applied to classify the students into 5 classes: Excellent, Very Good, Good, Average or Bad. The decision tree classifier performed best having the highest overall accuracy, followed by the rule learner (JRip) and the k-NN classifier. However, all classifiers performed with an overall accuracy below 70%. The predictive accuracy for the Good and Very Good classes (which contained the most students) were between 60% and 75%. As above, the accuracy of predictions was evaluated using 10-fold cross validation.

Oskouei and Askari (2014) studied the academic performance of students in high school and bachelor degree studies in Iran, and compared their results with the results of a similar study done in India. They considered the data of 500 students having a high school level and 600 students having a Bachelor degree level. They applied various classifiers such as Naive Bayes, C4.5 decision tree, Random Forest and Neural Networks, and meta-classifiers such as Bagging, Boosting or Adaboost, to classify students into 2 classes: Pass, Fail. The results revealed that features such as parent educational level, past examination results and gender impact the prediction. Best accuracy of 96% was obtained with C4.5 decision tree. The results were comparable with similar studies conducted in India.

Yehuala (2015) applied the decision trees and Naïve Bayes algorithms to predict the likelihood of success/failure at university. The dataset consisted of 11,873 undergraduate students from the Debre Markos University, Ethiopia. His findings

indicated that EHEECE (Ethiopian Higher Education Entrance Certificate Examination) result, gender, number of students in a class, number of courses given in a semester, and field of study were the major factors affecting the student performance. The highest prediction accuracy was 92.34% obtained with the decision tree algorithm using 10-fold cross validation.

Campagni, Merlini, Sprugnoli, and Verri (2015) proposed a different methodology to study the behaviour of graduated students in terms of their academic career. They used the data of 141 graduated students enrolled in the degree of Computer Science at the University of Florence (Italy) from 2001 to 2002 up to 2007-08 academic years. The variables used in this study described general information about students and their exams at university, such as the year of enrolment at university, the chosen curriculum, the type of high school and the corresponding grade, the date of the final examination and grade. Information about exams comprised the exam identifier, the exam date and the grade of the students in the exam, the semester in which the course was given by the teacher and the semester in which the exam was taken by the student. Authors apply the concept of an ideal career which means the career of an ideal student who had taken each examination just after the end of the corresponding course, without delay. The career of each student was then compared with the ideal career by using K-means clustering and sequential pattern analysis. The results demonstrated that the more the students stick to the order given by the ideal career, the better they perform in terms of graduation time and final grade.

Zimmermann, Brodersen, Heinimann, and Buhmann (2015) analyzed how well undergraduate achievements can predict graduate-level performance. They used the data of 171 student records in the Bachelor and Master programmes in Computer Science at ETH Zurich, Switzerland. Employing linear regression models in combination with different variable-selection techniques, their findings showed that undergraduate level performance can explain as much as 54% of the variance in graduate-level performance. They identified the third-year grade point average as the most significant explanatory variable, whose influence exceeds the one of grades earned in challenging first-year courses.

The impression from review of the aforementioned works is that it is possible to predict performance of students with reasonable accuracy; the more aggregated the performance, e.g. pass/fail, the higher the accuracy. The studies mentioned differ in the features they select from students' personal information like age, gender, religion, place of living, family, job, total score from previous education etc., to predict students' performance, but recognize that earlier marks are essential for good prediction (Golding & Donaldson, 2006; Oskouei & Askari, 2014; Zimmermann et al., 2015). This appears to confirm the different accuracies obtained in the two studies presented by Nghe et al. (2007): accuracies are higher for the study at CTU. The CTU dataset includes former marks along with socio-economic or demographic variables; the AIT dataset does not use former marks. Building on this recognition and taking into account the strict selection process of students' admission in our context, the present work uses only admission marks and marks of first and second year courses for the prediction of students' graduation performance. It does not consider other socio-economic factors such as age, gender, as these kinds of data are more difficult to obtain and, therefore, make the adoption of a workable policy more difficult. The literature review also reveals that there is no classifier that does better than the others in all contexts, though decision trees is often quoted to give good results (Kabakchieva, 2013; Kabakchieva et al., 2011; Oskouei & Askari, 2014; Yehuala, 2015). These works used cross-validation to evaluate their results. This means that the same cohort is used to build and test a classifier. In the present work, we take one cohort to build a model and evaluate the classifier by using the next cohort in order to obtain a more realistic generalization of results, thus, reflecting the intended application of this study. This feature distinguishes our work from other works.

2.2. Related works on clustering students and distillation of data for human judgment

In the handbook of educational data mining, Vellido, Castro, and Nebot (2010, pp. 75–92) mentions that clustering students is a proper technique to find similar learning behaviours. An investigation that has strong connection with our work is given in Bower (2010). Bower uses all K-12 marks in all topics to cluster school students and applies a hierarchical clustering once in order to uncover typical learning progressions. We adopt a similar approach in the third step of the present work. However, we do not perform one single clustering but cluster students year by year. This strategy allows us to discover small but interesting clusters of atypical students that are more difficult to exhibit with the approach used in (Bower, 2010).

Visualizing clusters calculated over several features is not an easy task and often results are simply given in the form of tables. By contrast, dendrograms also called cluster-trees, show graphically how clusters aggregate when a single agglomerative hierarchical clustering is performed. The study presented in Bower (2010) provides an intuitive visualization by combining dendrograms with heat map. As we perform several clustering, we use hierarchical histograms to show the resulting clusters. We, then, use heat map on students grouped by clusters as a visual help to check the indicators of high and low performance.

3. Data mining methods

Various techniques of data mining like classification and clustering can be applied to uncover hidden knowledge from educational data. This section gives an overview of the two methods used in this work: classification and clustering. Classification is a particular case of prediction when a class also called label or a discrete value is predicted by using a classifier (Han & Kamber, 2006). A classifier produces a classification model based on training data, which contains objects described

by the values they have on a set of attributes; one attribute is distinguished as the class. The generated model should fit well with the training data and suitably predict the class or label of unknown data, i.e. the test data, which is a separate set of data not used to generate the classifier (Han & Kamber, 2006). Usually, the performance of the classifier is evaluated by counting the test objects that are correctly predicted by the model. Accuracy is the overall correctness of the model and is calculated as the number of correct predictions divided by the total number of test objects. Calculating the kappa coefficient, which takes into account that a correct prediction could occur by chance, often completes the calculation of accuracy. As for accuracy the closer to 1, the better is the performance of the classifier. A kappa value above 0.3 usually indicates that the classifier is better than chance. In this study, we elaborate only on the classifiers that achieved a reasonable accuracy, as will be explained in the following section. These are: decision trees, rule induction, artificial neural networks, k-nearest neighbours, Naive Bayes and random forest trees (Han & Kamber, 2006). Classifiers such as artificial neural networks, k-nearest neighbours, Naive Bayes or random forest trees are like black boxes. They deliver an outcome, i.e. a prediction, but the results are not interpretable for humans. For our purposes, this means that one cannot understand which courses impact mainly their prediction. On the contrary, human can make sense of decision trees and rule induction. Decision trees play a substantial role in the sequel; therefore, they are discussed in more detail below.

3.1. Decision trees

Decision trees are a kind of non-cyclic flowchart. 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 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 context, this means that, ideally, all students of a leaf node have their graduation mark in the same interval, like 'A' or 'C', as we will see later. 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 represents 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. This attribute is then used to split the node. This process is repeated until a node is pure or too small to be split further. Several criteria can be used for this calculation. In this study, four criterions, 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 greater the entropy value, the less pure is the node. Gini index is another measure of impurity of a node based on observed probabilities. The accuracy is defined as above. The variable that maximizes the accuracy of the whole tree constructed so far is selected for split. Another criterion is Gain ratio which is a variation on the information gain method which is biased towards variables with a large number of distinct values. Fig. 1 shows an example of a decision tree. The attraction of Decision trees is that they are simple to understand and interpret, which is important in our context as we need to be able to reason how and why a particular result occurred.

3.2. Clustering

Clustering is the process of grouping a set of objects into classes of similar objects. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters (Han & Kamber, 2006). Measuring the similarity of two objects is done by calculating a distance measure such as the Euclidean Distance attributes having numerical values. There are a number of clustering algorithms. The K-means algorithm divides the objects

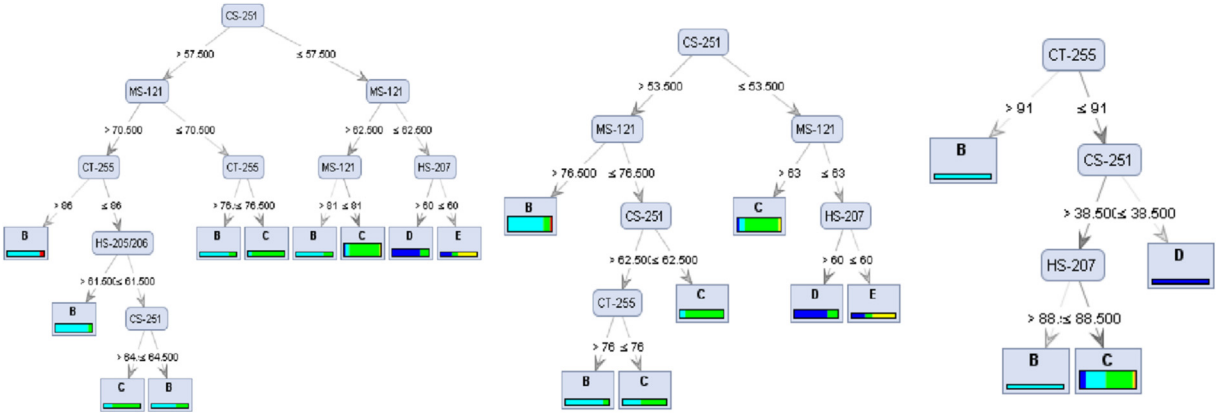

Fig. 1. Decision tree with Gini Index, Information Gain and Accuracy built with feature selection.

into *k* clusters, and iterates through the division-process as long as the distance between all objects and the centre or mean of the clusters can be reduced. A characteristic of this algorithm is that the number k of clusters has to be fixed. Quite often this number k is not known in advance, therefore several values of k need to be tried until a good balance is found between a small value, which might produce too coarse clusters, and a bigger value, which might produce too detailed clusters. The clustering method employed in this study is the X-means algorithm (Pelleg & Morre, 2000), which is a modification of K-means algorithm to include an automatic estimation of the optimal number of clusters based on the Bayesian Information Criterion (BIC).

4. Data and methodology

4.1. Data

The data used in this study comprises students' marks in a 4-year Information Technology bachelor degree of a public sector engineering university in Pakistan. This study employs data of two academic batches/cohorts using a sample of 210 undergraduate students who had enrolled in the academic batches of 2007–08 and 2008–09. The data set comprises variables related to students' pre-admission marks (used in selecting the students' for admission at university) and the marks for all the courses that are taught in the four years of the degree programme, shown in Table 1. Adj_Marks, Maths_Marks and MPC are variables related with the admission data of students defined as follows: Adj_Marks are the total marks in High School Certificate (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 remaining variables are the examination marks of students for all the courses that are taught in different academic years. Admission data and the courses that are mentioned in the sequel of this study are explained in Table 1. The full list of the courses used in the study is given in the appendix.

The marks at the end of the degree programme are calculated as 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. The interval of the graduation mark is divided into five possible values/categories: A (90%–100%), B (80%–89%), C (70%–79%), D (60%–69%), and E (50–59%) as these intervals are well understood by teachers and students alike. Batches and interval statistics at the end of the degree for the two cohorts are given in Table 2. It may be noted that there is no column F for fail. Indeed, owing to a strict selection process, drop-outs and failures are marginal and therefore not investigated in this work. Table 3 is an overview of the distribution of marks of the students over the four years. The average marks for each student in each year have been calculated as a number from 0 (worst mark) to 100 (best mark).

All data mining techniques in this study have been performed with the RapidMiner software (www.rapid-i.com).

4.2. Methodology

For predicting the students' graduation performance at the end of the degree and answer Research Question 1, several classification algorithms were used. The literature review suggests that in general there is no single classifier that works best in all contexts to provide good prediction. Therefore there is a need to investigate which classifiers are more suited to the data being analysed. As Table 2 shows, the repartition of the students among the intervals is unbalanced. Class 'C' interval contains the most students for both cohorts. Predicting a student class 'C' would have an accuracy of 51.92%. This forms the baseline that we want to improve. As discussed earlier, the classification models to predict students' performance are generated on

Table 1
Variables in dataset.

Variable	Description
Adj_Marks	HSC Examination total marks
Maths_Marks	HSC Examination Mathematics marks
MPC	HSC Maths + Physics + Chemistry marks
HS-205/206	Islamic Studies or Ethical Behaviour
MS-121	Applied Physics
CS-251	Logic Design and Switching Theory
CT-255	Assembly Language Programming
HS-207	Financial Accounting and Management

Table 2
Statistics of batches and intervals for degree mark.

Academic Cohort	Total No. of students	Total No. of students in 'A' Interval	Total No. of students in 'B' Interval	Total No. of students in 'C' Interval	Total No. of students in 'D' Interval	Total No. of students in 'E' Interval
Cohort 1	106	1	41	46	14	4
Cohort 2	104	–	31	54	18	1

Table 3
Distribution of yearly marks; left cohort 1, right cohort 2.

	A (90–100)	B(80–89)	C(70–79)	D(60–69)	E(50–59)	A (90–100)	B(80–89)	C(70–79)	D(60–69)	E(50–59)
First year	0	9	56	31	9	0	14	55	29	6
Second year	0	13	55	25	12	0	13	46	34	11
Third year	1	47	37	16	4	0	30	48	22	4
Fourth year	6	62	26	11	0	0	31	54	18	1

cohort 1 data and evaluated against cohort 2. Preliminary exploration of the data included building decision trees with different criteria (refer to section 3.1) to predict the division of the final marks of the degree. Divisions are coarser than intervals. The university categorizes a student's academic achievement in one of three divisions when awarding a degree. *First division with distinction* matches roughly intervals 'A' and 'B'; *first division* matches roughly intervals 'C' and 'D' while *second division* matches roughly the 'E' interval. Pre-admission marks and yearly marks from all four years were used to generate these trees. However, the obtained decision trees did not contain any branch labelled with 3rd year or 4th year marks, or, in other words, did not select 3rd year or 4th year marks to classify students according to division. Building on this exploration, only pre-admission marks and marks from 1st and 2nd years have been considered as variables or attributes for prediction in this study.

Research Question 2 deals with exploring the courses that may be indicators of a good or bad performance at the end of the degree. It is known that selecting attributes might improve the performance of classifiers, since they leave out attributes that do not have much impact on the prediction, and thus are not likely to be indicators of good or low performance. In the first step, we have explored different attributes selection techniques to narrow down the variables or attributes used to learn the classifiers, in particular all well-known selection operators as available in RapidMiner. However, each selection technique improved only the performance of a minority of the classifiers, and, most important for this study, did not improve the performance of the decision trees. Decision trees show which attributes lead to the prediction of a particular label and are therefore helpful to make sense of the findings. An original attribute selection technique has been used since it improves the performance of almost half of the classifiers, the decision trees in particular. This selection technique is based on the choices made by the decision trees in selecting the attributes while learning the trees. Extending the data set by considering two more consecutive cohorts, we have built decision trees with the four criteria given in section 3, taking one cohort to build the model and the follower cohort to test it. We have selected the variables or courses appearing in at least half of all decisions trees in the spirit of ensemble methods, as explained in (Asif, Merceron, & Pathan, 2015b). This gives us a set of courses that appear to be impacting performance most. In the second step, we considered parts of paths occurring in at least half of the decision trees and leading to nodes labelled 'B' -pure nodes or, if impure, containing elements with class 'A'- and paths leading to nodes labelled by 'D' or 'E'; in the latter case, impure nodes might include elements with class 'C'. Courses occurring on those paths leading to nodes labelled 'B' constitute indicators of good performance and courses on paths leading to nodes labelled 'D' or 'E' constitute indicators of low performance.

For Research Question 3, the focus is on investigating how performance of students progresses during their studies, and on relating this progress with the indicator courses. For this purpose, we look for typical progression patterns of students during their studies. Progressions of the marks are considered not on an absolute scale, as has been done in the study by Asif, Merceron, and Pathan (2014), but in comparison with other students. Each student is represented by a vector comprising of the marks obtained in each individual course of each of the four years of study. To discover typical progression patterns over the years, students are clustered each year taking their final marks of each course in each year. X-means clustering with Euclidean distance is applied. Thus, each student's progression is represented by a 4-tuple indicating the cluster the student belongs to in each year. We expect to find progressions representing high-performing students and progressions representing low-achieving students. We relate these results to the prediction results by investigating whether the courses that are indicators of good performance reveal high-performing students and, similarly, whether the courses that are indicators of low performance disclose low-achieving students.

5. Results and discussion

This section presents the results obtained by following the methodology mentioned above: (1) Predicting graduation performance using classifiers (2) Deriving actionable predictors for students' performance at the end of the degree programme (3) Investigating how academic performance of students progresses over the years and (4) Linking the results of prediction and progression.

5.1. Predicting graduation performance using classifiers

Table 4 presents the accuracy and kappa results of 10 classifiers that have achieved accuracy above the baseline. These results show that cohort 1 data can predict students' graduation performance of cohort 2 with a reasonable accuracy at the

end of the degree by using pre-university marks and marks of 1st and 2nd year courses; no socio-economic or demographic features were taken into consideration. Therefore, Research Question 1, i.e. *Can we predict students' performance with a reasonable accuracy at an early stage of the degree programme using marks only?*, is answered positively.

The resultant confusion matrices are shown in Table 5. To understand these confusion matrices, the first confusion matrix of the classifier "Decision Tree with Gini Index' is discussed. In the first column, out of the 31 (i.e. 18 + 13) actual class 'B' students, the classifier predicted correctly 18 as 'B'; the recall for class 'B' is 58.06 (i.e. 18/31). In the first line, 24 (18 + 6) were predicted as 'B'; the precision for class 'B' is 18/24 = 75%. Remaining columns and lines are similar for the other classes. All correct predictions are located in the diagonals of each matrix. One notices that 'A' is absent. This is due to the reason that for cohort 1 there is only one student that belongs to interval 'A' and no student for cohort 2. So, in the confusion matrices, there are only zeros for class 'A'; therefore the column for class 'A' is dropped. Table 5 shows that classifiers have difficulties with classes that are underrepresented such as 'A' and 'E'. Note that rebalancing techniques have been applied, but they have not lead to any improvement of the results. Best accuracies are obtained for well-represented classes such as 'B' and 'C'.

5.2. Exploring actionable predictors for students' performance

Table 4 shows that the classifiers with the best performance are Naive Bayes followed by 1-nearest neighbour and random forests with Gini Index. A drawback of these three classifiers is that they are not interpretable for humans: it is not possible to understand which variables or attributes, in our context which courses impact the prediction. In contrast, decision trees are understandable by humans. As established in the methodology section, we have devised our own selection technique that improves the performance of almost half of the classifiers and, in particular, of the decision trees. This technique performs better with our data than other well-known selection operators available in RapidMiner. The five selected attributes are HS-205/206, MS-121, CS-251, HS-207 and CT-255, two courses from the first year and three courses from the second year. The explanation of these courses is given in Table 1. The decision trees with accuracy above the baseline were built using these 5 courses with cohort 1 and tested with cohort 2 and are shown in Fig. 1. As already explained in the methodology section, a bigger data set considering two more cohorts has been used to select these five courses. The decision trees for the two additional cohorts are given in the appendix.

As explained in the methodology section and detailed in (Asif et al., 2015b), a pragmatic policy is derived by considering parts of paths occurring the most in the decision trees to characterize two groups: group of students who are likely to achieve their degree with a mark in the 'A' or 'B' interval, and group of students who are likely to achieve their degree with a poor mark i.e. 'D' or 'E' interval. It is summarized below:

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

Table 4
Classifier Prediction Accuracy and kappa.

Classifier	Accuracy/Kappa
Decision Tree with Gini Index (DT-GI)	68.27%/0.493 (with minimal leaf size 2)
Decision Tree with Information Gain (DT-IG)	69.23%/0.498 (with minimal leaf size 6)
Decision Tree with Accuracy (DT-Acc)	60.58%/0.325 (with minimal leaf size 2)
Rule Induction with Information Gain (RI-IG)	55.77%/0.352
1-Nearest Neighbour (1-NN)	74.04%/0.583
Naive Bayes	83.65%/0.727
Neural Networks (NN)	62.50%/0.447
Random Forest Trees with Gini Index (RF-GI)	71.15%/0.543 (with minimal leaf size 8)
Random Forest Trees with Information Gain (RF-IG)	69.23%/0.426 (with minimal leaf size 10)
Random Forest Trees with Accuracy (RF-Acc)	62.50%/0.269 (with minimal leaf size 2)

Table 5
Confusion matrices.

Decision Tree with Gini Index		Actual				Class precision
		B	C	D	E	
Predicted	B	18	6	0	0	75.00%
	C	13	38	2	0	71.70%
	D	0	10	14	0	58.33%
	E	0	0	2	1	33.33%
Class Recall		58.06%	70.37%	77.78%	100.00%	

Naïve Bayes		Actual				Class precision
		B	C	D	E	
Predicted	B	28	6	0	0	82.35%
	C	3	47	6	0	83.92%
	D	0	1	12	1	85.71%
	E	0	0	0	0	0.00%
Class Recall		90.32%	87.04%	66.67%	0.00%	

Decision Tree with Information Gain		Actual				Class precision
		B	C	D	E	
Predicted	B	16	4	0	0	80.00%
	C	15	40	2	0	70.18%
	D	0	10	16	1	59.26%
	E	0	0	0	0	0.00%
Class Recall		51.61%	74.07%	88.89%	0.00%	

Neural Networks		Actual				Class precision
		B	C	D	E	
Predicted	B	30	24	0	0	55.56%
	C	1	23	3	0	85.18%
	D	0	6	12	1	63.15%
	E	0	1	3	0	0.00%
Class Recall		96.77%	42.59%	66.67%	0.00%	

Decision Tree with Accuracy		Actual				Class precision
		B	C	D	E	
Predicted	B	4	0	0	0	100.00%
	C	27	44	2	0	60.27%
	D	0	10	14	0	58.33%
	E	0	0	2	1	33.33%
Class Recall		12.90%	81.48%	77.78%	100.00%	

Random Forest Trees with Gini Index		Actual				Class precision
		B	C	D	E	
Predicted	B	27	13	1	0	65.85%
	C	4	35	5	0	79.55%
	D	0	6	12	1	63.16%
	E	0	0	0	0	0.00%
Class Recall		87.10%	64.81%	66.67%	0.00%	

Rule Induction with Information Gain		Actual				Class precision
		B	C	D	E	
Predicted	B	22	8	3	1	57.89%
	C	6	23	2	0	74.19%
	D	3	23	13	0	33.33%
	E	0	0	0	0	0.00%
Class Recall		70.97%	42.59%	72.22%	0.00%	

Random Forest Trees with Information Gain		Actual				Class precision
		B	C	D	E	
Predicted	B	22	4	0	0	84.62%
	C	9	50	18	1	64.10%
	D	0	5	0	0	0.00%
	E	0	0	0	0	0.00%
Class Recall		70.97%	95.29%	0.00%	0.00%	

1-NN		Actual				Class precision
		B	C	D	E	
Predicted	B	26	11	0	0	70.27%
	C	5	38	5	0	79.16%
	D	0	5	13	1	68.42%
	E	0	0	0	0	0.00%
Class Recall		83.87%	70.37%	72.22%	0.00%	

Random Forest Trees with Accuracy		Actual				Class precision
		B	C	D	E	
Predicted	B	11	1	0	0	91.67%
	C	20	53	17	0	58.89%
	D	0	0	1	1	50.00%
	E	0	0	0	0	0.00%
Class Recall		35.48%	98.15%	5.56%	0.00%	

5.3. Investigating progression of students

As written in the methodology section, students have been clustered each year by taking into account their final marks for each course in each year to determine typical progression patterns over the years. The results presented in this section summarize the results of the work done by Asif, Merceron, and Pathan (2015a). Table 6 represents the clustering of cohort 1 for the first year. One notices that three clusters have been found. The centroids give the average mark of the cluster in each course. All these courses are listed in the appendix. In each course, the average mark of the cluster *Low* is less than the average mark of cluster *Intermediate* and this one is less than the average mark of cluster *High*.

Table 7 represents the total number of students in each cluster for all four years for cohort 1 and cohort 2 respectively. The sets of clusters obtained show the same trends in the complete data: clusters of students with low marks in all courses, intermediate marks in all courses and high marks in all courses. As a contrasting example, no cluster has been found containing students with low marks in courses $\times 1$ and $\times 2$, intermediate marks in courses $\times 3$ and $\times 4$, and high marks in course $\times 5$.

Table 7 also indicates that for cohort 1, most of the students belongs to the intermediate cluster in all the years except the second year where the "High" cluster is particularly large. Notice that in the third year the intermediate cluster splits into two clusters. Similar to cohort 1, most of the students of cohort 2 belong to the intermediate cluster in all years except in the first year, where the intermediate cluster does not exist. These findings match well with Table 3, which has the highest number of students in the interval 70–80 in all four years for both cohorts, except for the fourth year of cohort 1.

Having clustered each cohort four times, the typical progression of a student is determined as follows. First, the mean of all centres of each cluster in each year is calculated and rounded. For example, for cohort 1 year 1 (see Table 6), the mean of cluster 'Low' is 60, the mean of cluster 'Intermediate' is 71 and the mean of cluster 'High' is 78; for cohort 2 year 1, there are only 2 clusters, the mean of cluster 'Low' is 60 and the mean of cluster 'High' is 73. Then, to obtain an intuitive overview of how the performance of students globally evolves over the four years, each student is described by a 4-tuple whose elements are the mean of the centres of the clusters the student is in. For example, for cohort 1, a student belonging to the cluster with the lowest mark in first year and second year, with intermediate-low marks in third year and intermediate marks in fourth year will be represented by the tuple (60, 52, 66, 77) while a student belonging to the cluster with high marks in all four years will be described by the tuple (78, 73, 83, 85).

Fig. 2 and Fig. 3 aggregate all the tuples of all students for cohort 1 and cohort 2 respectively in a kind of hierarchical histogram. The height of the bar represents the number of students characterized by the same tuple. The diagram is ordered from right to left: low values on the right, and high values on the left. The colour indicates the first year clusters. The 2nd year is depicted at the bottom of the diagram and divided into different parts corresponding to the mean of the clusters. Each of these parts is divided into the clusters of 3rd year. Finally, the highest level of the hierarchy divides further each part with the clusters of the 4th year, which is drawn right below the bars.

We can see from the above figures that there are more high bars towards the left of Fig. 2 as compared to Fig. 3. This shows that there are more students with higher marks in the second year and high or intermediate marks in the third and fourth year in cohort 1 than in cohort 2, which visualizes the trend given by Table 3.

Table 6
Cluster centroids of first year for Cohort 1.

Attribute	Low	Intermediate	High
CT-153	61.714	75.636	87.259
CT-157	66.257	80.636	86.259
CT-158	76.543	82.114	82.963
CT-162	58.371	71.614	82.741
EE-115	49.571	62.273	76.259
EL-134	51.686	68.977	76.778
HS-102	53.714	56.409	64.185
HS-105/127	60.114	61.955	68.444
HS-205/206	58.743	64.500	69.296
MS-121	60.400	73.136	82.778
MS-171	60.686	78.341	82.926

Table 7
No. of Students Cluster Wise in Four Years; left cohort 1, right cohort 2.

Cluster	First Year	Second Year	Third Year	Fourth Year	First Year	Second Year	Third Year	Fourth Year
Low	32	15	14	23	49	20	18	34
Intermediate	47	25	27 (Intermediate - Low) 32 (Intermediate - High)	49	–	41	43	44
High	27	66	33	34	55	43	43	26

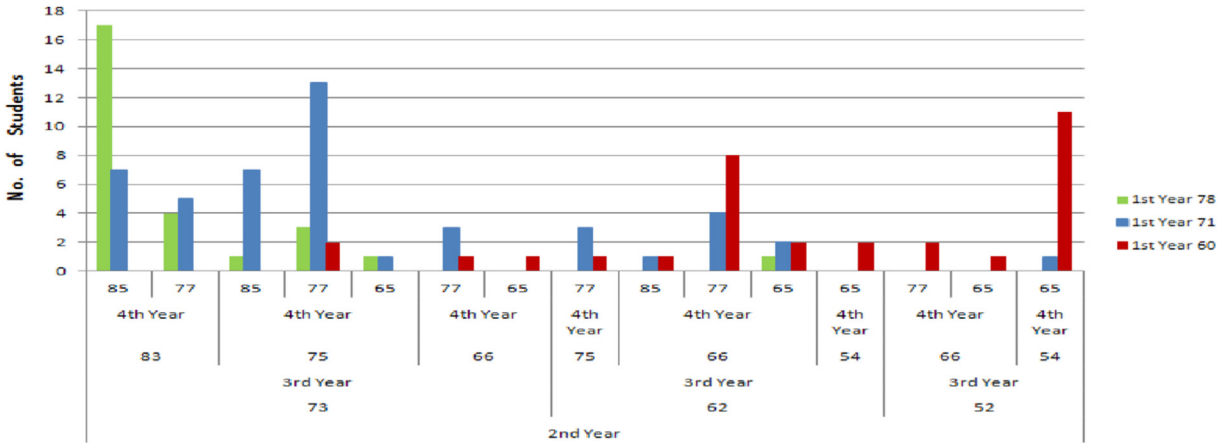

Fig. 2. Tuples summary of Cohort1.

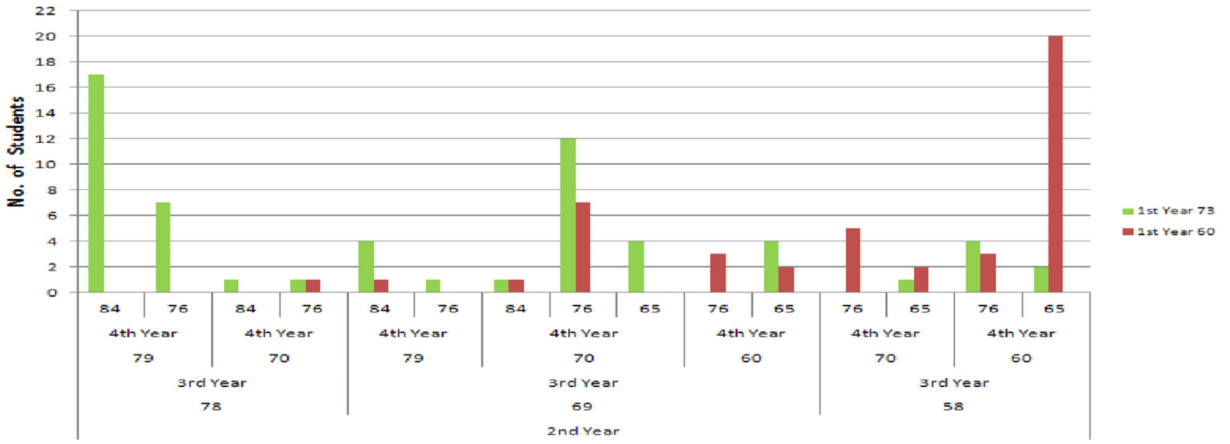

Fig. 3. Tuples summary of Cohort2.

The above two figures demonstrate two critical groups of students; one group of students is identified by the high green bar on the far left; this group represent the high-performing students. The second group is identified by the red bar on the far right and contains the low-achieving students. High-performing students have high marks in each of the four years and this group is essential in both cohorts. Low-achieving students have low marks in all four years and this group is somewhat bigger in cohort 2 than in cohort 1. Other important groups for both cohorts are those groups which contain students who have intermediate marks in all years or who have intermediate marks in all years with the exception of one year. The second and fourth highest bars of Fig. 2 show such groups for cohort 1: students with intermediate marks in all years but in the second year, where they have high marks, make up the second highest bar, (71, 73, 75, 77), and students with intermediate marks but in 1st year, where they have low marks, make up the fourth highest bar, (60, 62, 66, 77). The third and fourth highest bars of Fig. 3, (60, 69, 70, 76) and (73, 69, 70, 76), are further illustrations of students having intermediate marks but in one year, here the first year, for cohort 2.

Figs. 2 and 3 also demonstrate that there are very few atypical students like those who have low marks in 1st year but then progress and finish with high marks in the 4th year. They are discovered searching for small red bars towards the left of the diagram. There is one such student in cohort 2 given by the tuple (60, 69, 70, 84). Another interesting, and also small cluster, is constituted by students who have high marks in 1st year but low marks in all following years. Small green bars towards the right of the diagram depict them. There are two of them in cohort 2 (73, 58, 60, 65), and one in cohort 1 (78, 62, 66, 65).

The results obtained in this section show that there are two important groups of students: the high-performing students, and the low-achieving students. They also show that many students tend to stay in the same kind of groups all four years: many clusters are constituted of students with intermediate marks all the way but in one year. Finally, our approach of clustering students year by year allows for discovering small but interesting clusters of atypical students who begin with low marks and finish with high marks or vice versa. These students are not found when clustering is conducted by considering the marks of all years together. This kind of atypical group is also difficult to discover in the work done by Bower (2010).

5.4. Reflecting on the pragmatic policy

In this section, cohort 1 and cohort 2 are tackled in turn. For cohort 1, 39 students are predicted by the pragmatic policy as likely to have a degree mark in the 'A' or 'B' interval and 29 students in the 'D' or 'E' interval. Precisely, one student with an 'A' degree mark, 28 students with a 'B', and 10 with 'C'. Therefore, the accuracy of the pragmatic policy for the degree mark in the 'A' or 'B' interval is 29/39 or 74.35%. A very legitimate question arises here whether the pragmatic policy detects all students who obtained a degree mark in the 'A' or 'B' interval or not. This is technically known as the recall measure. This is a tricky question, as the pragmatic policy has not been designed to detect all students with graduation in the 'A' or 'B' interval, but only those where the graduation in a 'B' interval would not be too far from the 'A' interval as explained in the methodology section, see also (Asif et al., 2015b). Therefore, a high recall is not expected. Indeed, recall is 29/42 = 69.05%, as the data set of cohort 1 contains one student with a degree mark in the 'A' interval and 41 students with marks in the 'B' interval (refer back to Table 2).

Most important is to know whether the pragmatic policy detects all high-performing students as identified in the previous section about progression, and indeed it does. Fig. 2 shows that there are 17 high-performing students, which can be visualized by looking at the green bar at the extreme left of the figure, and defined by the cluster (78, 73, 83, 85). By inspecting these 17 students in the actual dataset, it may be noticed that these students are all predicted by the pragmatic policy as likely to have a degree mark in the 'A' or 'B' interval. The remaining 22 students that are predicted with an 'A' or 'B' degree mark by the pragmatic policy are grouped in different clusters as can be seen at the left side of Table 8. In Table 8, the

Table 8
Clusters and identification by the pragmatic policy for cohort 1, left: high performing students, right: low achieving students.

Cluster	Total No. of Students in each cluster	No. of students identified by pragmatic policy along with the interval count	Cluster	Total No. of Students in each cluster	No. of students identified by pragmatic policy along with the interval count
(60, 73, 66, 65)	01	01 ('C')	(60, 52, 54, 65)	11	11(8 'D' and 3 'E')
(60, 73, 75, 77)	02	02 (All 'C')	(60, 52, 66, 77)	02	02(1'C' and 1′D′)
(71, 73, 75, 77)	13	06 (1'B' and 5 'C')	(60, 62, 54, 65)	02	02(All 'D')
(71, 73, 75, 85)	07	01 ('B')	(60, 62, 66, 65)	02	02(1'C' and 1′D′)
(71, 73, 83, 77)	05	03 (All 'B')	(60, 62, 66, 77)	08	04(All 'C')
(71, 73, 83, 85)	07	05 (All 'B')	(71, 52, 54, 65)	01	01('D')
(78, 73, 75, 77)	03	02 (All 'C')	(71, 62, 66, 65)	02	01('C')
(78, 73, 75, 85)	01	01 ('B')	(71, 73, 66, 77)	03	01('C')
(78, 73, 83, 77)	04	01 ('B')	(71, 73, 75, 77)	13	02(1 'B' and 1 'C')
(78, 73, 83, 85)	17	17 (1 'A' and 16 'B')	(71, 73, 83, 85)	07	01('B')
			(78, 62, 66, 65)	01	01('C')
			(78, 73, 83, 85)	17	01('B')

second column shows the number of students in each cluster and the third column indicates the number of students identified by the pragmatic policy as well as the real interval of their degree mark. Interestingly, the clusters containing students identified as likely to obtain a high graduating mark by the pragmatic policy are classified in the category "high" for the second year.

Looking at the students of cohort 1 predicted as likely to have a degree mark in the 'D' or 'E' interval, 3 students have 'B' interval, 10 have 'C' interval, 13 have 'D' interval and 3 have 'E' interval. So, the accuracy of the pragmatic policy when predicting 'D' or 'E' is 16/29 = 55% and recall is 16/18 = 88.88%. In the actual dataset of cohort 1, there are 14 students with 'D' interval and 4 students with 'E' interval (refer Table 2). These figures reflect that the pragmatic policy has been designed to be cautious: in doubt, better give a warning. Fig. 2 indicates that there are 11 low-achieving students with lowest marks in all years and described by the tuple (60, 52, 54, 65). All of them are predicted by the pragmatic policy as likely to have a degree mark in the 'D' or 'E' interval, which is indeed the case. The remaining 18 students that are predicted by the pragmatic policy as likely to have a degree mark in the 'D' or 'E' interval are in different clusters, see right side of Table 7. We can also observe that the students who have low marks in all years but one are also flagged by the pragmatic policy, see clusters (60, 52, 54, 71), (60, 62, 54, 65) and (60, 52, 66, 65) at the right side of Table 7.

Table 9
Clusters and identification by the pragmatic policy for cohort 2, left: students likely to obtain high marks, right: students likely to obtain low marks.

Cluster	Total No. of Students in each cluster	No. of students identified by pragmatic policy along with the interval count	Cluster	Total No. of Students in each cluster	No. of students identified by pragmatic policy along with the interval count
(60, 58, 60, 65)	20	01('C')	(60, 58, 60, 65)	20	19(2'C', 16'D', 1'E')
(60, 69, 60, 76)	03	01('C')	(60, 58, 60, 76)	04	03(2'C, 1'D)
(60, 69, 70, 76)	07	02(All 'C')	(60, 58, 70, 65)	02	02(All 'C')
(60, 78, 70, 76)	01	01('C')	(60, 58, 70, 76)	05	05(All 'C')
(73, 69, 60, 65)	04	01('C')	(60, 69, 60, 65)	02	01('D')
(73, 69, 70, 76)	12	04(All 'C')	(60, 69, 60, 76)	03	02(All 'C')
(73, 69, 79, 76)	01	01('B')	(60, 69, 70, 76)	08	07(All 'C')
(73, 69, 79, 84)	04	01('B')	(60, 69, 70, 84)	01	01 ('C')
(73, 78, 70, 76)	01	01('B')	(60, 78, 70, 76)	01	01 ('C')
(73, 78, 79, 76)	07	05('B')	(73, 58, 60, 65)	02	02(1 'C', 1'D')
(73, 78, 79, 84)	17	17(1'A', 16'B')	(73, 58, 60, 76)	04	02(All 'C')
			(73, 58, 70, 65)	01	01('C')
			(73, 69, 60, 65)	04	03(All 'C')
			(73, 69, 70, 65)	04	02(All 'C')
			(73, 69, 70, 76)	15	03(All 'C')
			(73, 69, 79, 84)	04	02(1'B', 1'C')
			(73, 78, 79, 76)	07	02(All 'B')
			(73, 78, 79, 84)	18	01('B')

A further examination of Table 8 shows that the pragmatic policy foresees some students with a degree mark 'D' or 'E' who have intermediate or high marks in the first and second year but low or intermediate marks in the third and fourth year. Example clusters are (71, 62, 66, 65), (71, 73, 66, 77), (71, 73, 75, 77) and (78, 62, 66, 65). This confirms that the pragmatic policy warns too much. Interestingly, the atypical student beginning with high marks and finishing with low marks is predicted with degree mark 'D' or 'E' by the pragmatic policy: (78, 62, 66, 65) though actually s/he manages a 'C'. The pragmatic policy is also contradicting; it detects one student that belongs to the cluster (78, 73, 83, 85) as high-performing as well as low-achieving student.

Analysis of cohort 2 reveals that the accuracy of the pragmatic policy while predicting degree mark in the interval 'A' or 'B' is 68.85% and recall 70.96%. These figures are similar to those obtained for cohort 1. For predicting degree mark 'D' or 'E', accuracy is 30% and recall 94.74%, which shows that almost no targeted student will be missed, but with the price of having many false warnings issued. Table 9 is similar to Table 8, but for cohort 2. As for cohort 1, the pragmatic policy detects all high-performing students and all low-achieving students (except 1) in the sense of progression of the previous section. With the exception of the cluster (60, 58, 60, 65), groups containing students predicted with a degree mark in the 'A' or 'B' interval by the pragmatic policy are clusters with high or intermediate marks in year 2. Most of the groups containing students identified as likely to have a degree mark in the 'D' or 'E' interval by the pragmatic policy are clusters with low marks in year 1 or 2. Like for cohort 1, the pragmatic policy is also conflicting for cohort 2; it identifies one student that belongs to the high cluster (73, 78, 79, 84) as a high-performing as well as a low-achieving student.

To have a better visualization of how students are evolving over the years, we make a heat map by sorting the students of a cohort clusters-wise according to the results of progression. The number 1, 2, 3 and so on at the left of Fig. 4 shows the

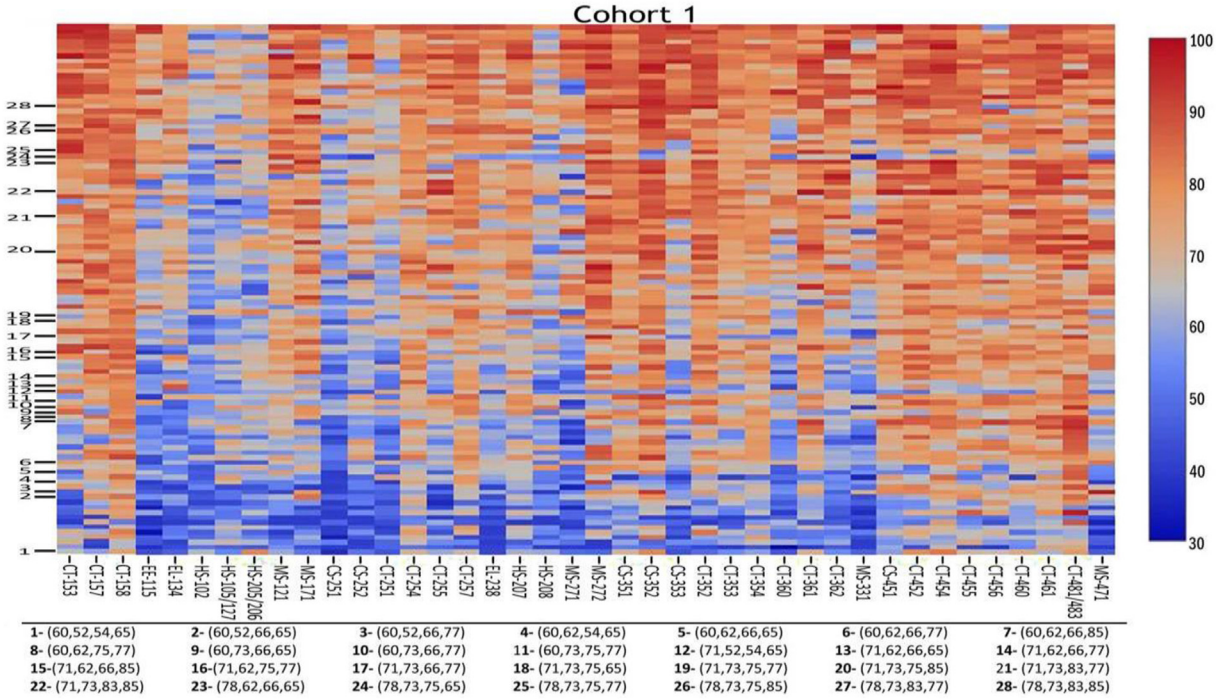

Fig. 4. Cohort 1 Heat map with sorted students.

Table 10
HSC marks of high performing and low achieving students.

Cohort No.	Total No. of high-performing students	Total no. of students with marks lower than average HSC marks	Total No. of low achieving students	Total no. of students with marks higher than average HSC marks
Cohort 1	17	05	11	04
Cohort 2	17	08	20	04

clusters sorted by the columns of their tuples. The bottom line shows all the courses of the four-year degree with the courses of the 1st year on the left, followed by the courses of the 2nd year, 3rd year and finally the courses of the 4th year on the right. Fig. 4 shows the heatmap of cohort 1 while the heatmap of cohort 2 is given in the appendix. It turns out that this heatmap visualizes several results of this contribution. The shift in the colour of the columns towards more red on the left matches the shift towards better marks shown in Table 3. In the section on actionable predictors, five courses have been selected to improve the performance of classifiers. These courses are HS-205/206, MS-121, CS-251, HS-207 and CT-255. These courses go from low to high marks (from blue to red) and the change in the colour follows roughly the clusters; this is particularly true for the courses identified as predictors of low performance (MS-121, CS-251, HS-207) or high performance (CT-255, HS-207) of students. Accidentally, the heatmap also points out courses where all students tend to have marks in a small range like CS-158, third from left, which is mostly red, or like HS-102, sixth from left, which is mostly blue.

Our aim is to uncover students likely to obtain a low graduation mark or a high graduation mark as early as possible. However, it turns out that our pragmatic policy does not use admission marks, which would allow putting in place some measures already at the beginning of the first year. This is an interesting and important result of the prediction part of this work. We now take the complementary look by considering the HSC marks of the high-performing and the low-achieving students, as the university selects the students for the entrance exam on their HSC marks. In cohort 1, out of 17 high-performing students, 5 have HSC marks lower than the average HSC marks. Similarly, for cohort 2, out of 17 high-performing students, 8 have HSC marks lower than average. Conversely, we find that 4 out of 11 low-achieving students have HSC marks above average in cohort 1 and 4 out of 20 low-achieving students in cohort 2. Table 10 gives an overview. These findings suggest that students are evolving at university and this evolution does not necessarily follow their performance behaviour prior to university.

6. Conclusion and future works

The present study has investigated three research questions with the final aim of providing teachers and study programme directors with information that might help them to improve the educational programmes at their institution. The first question concerns predicting students' performance using marks only, no socio-economic data. The results show that it is possible to predict the graduation performance in a four-year university program using pre-university marks and marks of first and second year courses only with a reasonable accuracy. Further, the model established for one cohort generalizes to the following cohort.

The second question strives for deriving courses that can serve as effective indicators of good or poor performance in the degree programme. With the help of decision trees, four courses have been put in evidence that can serve as such indicators.

The third question involves investigating how students' academic performance progresses over the four-year degree. Surprisingly, in each year students tend to have the same kind of marks: low marks, intermediate marks or high marks in all courses. This pattern repeats over the years: students tend to remain in the same kind of groups. Thus, two major groups are put in evidence: the group of high-performing students who acquired high marks during the four years, and the group of low-achieving students who got low marks during the four years. We observe that the proposed pragmatic policy is reliable in the sense that it detects these two groups (but for one student) in the two cohorts that we have studied. It should be noticed that the pragmatic policy is very cautious and may issue a warning to students who are not necessarily struggling students at the end of the degree program. In summary, the pragmatic policy appears to be workable and allows “automatically flag students who show early signs of struggle or opportunity”, a scenario that is ranked as very important by teachers in the report compiled by Pea (2014).

An important future work is to deepen the generalizability of the results. Studies along these lines have already begun and indicate that using the same approach, graduation performance can be predicted in two other degree programmes of the same university. Prediction of performance at the end of the course could be investigated for the courses identified as indicators of low and high performance, thus giving the university another leverage to improve educational outcomes. Further, the university has recently changed from an annual to a semester system. The framework of the present study should be adapted to this new context. Another future direction of research is to study whether the design of the heatmaps can be refined to extract the indicators of low and high performance without running the algorithms for prediction before. If conclusive, the whole approach would be easier to implement in practice.

Funding

This work was supported in part by a grant from NED University of Engineering & Technology, Karachi, Pakistan.

Acknowledgements

We thank Dr. Sajida Zaki and Ms. Rahila Huma Anwar for providing language help and proof reading this article. This work is supported in part by a grant from NED University of Engineering & Technology, Karachi, Pakistan.

Appendix

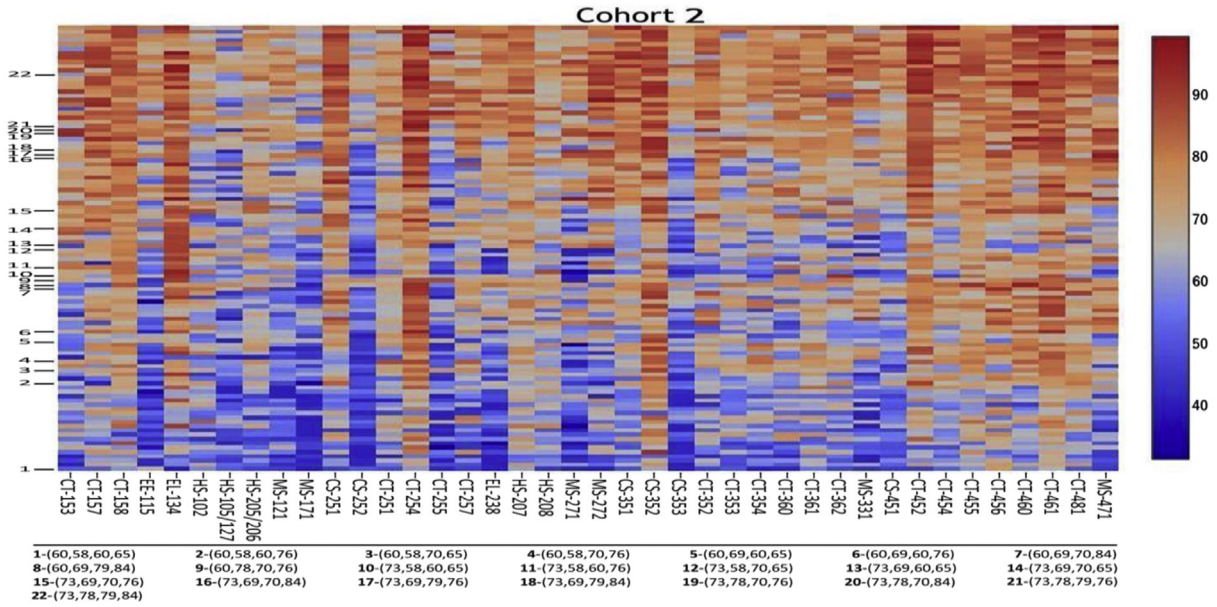

Fig. A.1. Cohort 2 Heat map with sorted students.

Admission data and the list of courses over the four years are explained in Table A.I. Besides for just a few exceptions, the first digit in the three digit-numeric code assigned to each course indicates the year in which the course is offered. For instance CT-153 is taught in first year, CS-251 is taught in second year, while CT-352 is taught in third year and CT-452 in fourth year.

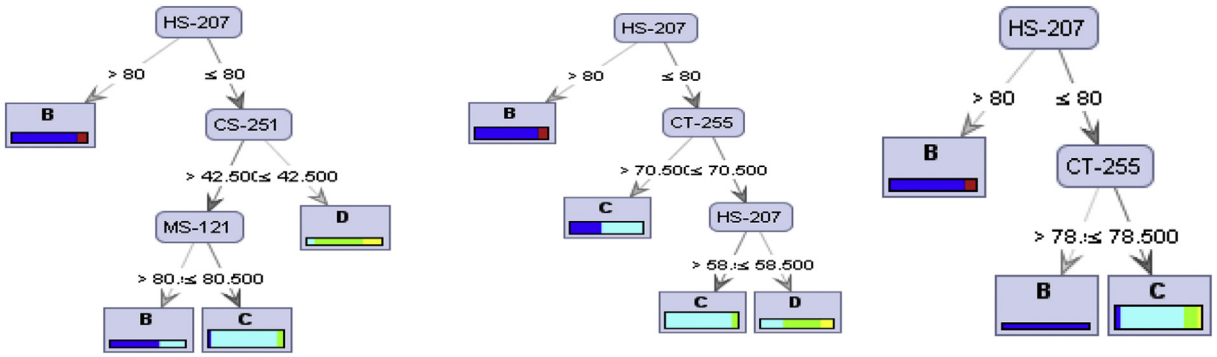

Fig. A.2. Decision tree with Gini Index, Information Gain and Accuracy built with feature selection for Dataset1.

Table A.1
Full list of variables of the dataset.

Name	Description
Adj_Marks	HSC Examination total marks
Maths_Marks	HSC Examination Mathematics marks
MPC	HSC Maths + Physics + Chemistry marks
CT-153	Programming Languages
CT-157	Data Structures Algorithms and Applications
CT-158	Fundamentals of Information Technology
CT-161	Computing Lab
EE-115	Electrical Technology Fundamentals
EL-134	Basic Electronics
HS-102	English
HS-105/127	Pakistan Studies
MS-171	Differential & Integral Calculus
HS-205/206	Islamic Studies or Ethical Behaviour
MS-121	Applied Physics
MS-172	Discrete Structures
CS-251	Logic Design and Switching Theory
CS-252	Computer Architecture and Organization
CT-251	Object Oriented Programming
CT-254	System Analysis and Design
CT-255	Assembly Language Programming
CT-257	Data Base Management System
EL-238	Digital Electronics
HS-208	Business Communication & Ethics
MS-271	Ordinary Differential Equation & Complex Variable
MS-272	Linear Algebra & Geometry
HS-207	Financial Accounting and Management
CT-352	Computer Graphics
CT-353	Operating Systems
CT-354	Software Engineering
CT-360	Visual Programming
CT-361	Artificial Intelligence & Expert System
CT-362	Web Engineering
CS-351	Computer Communication Networking
CS-352	Digital Communication Systems
CS-353	Microprocessor & their Applications
MS-331	Probability & Statistics
CT-452	Modelling & Simulation
CT-455	Distributed Database Client Server Programming
CT-456	Data Warehouse Methods
CT-460	Network & Information Security
MS-471	Applied Numerical Methods
CS-451	Parallel Processing
CT-454	Compiler Design
CT-461	E-Commerce
CT-481	Wireless Network & Mobile Computing
CT-483	System Administration

References

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In *Proceedings of the 2nd international conference on learning analytics and knowledge*. Vancouver, BC, Canada, 29 April – 2 May.

Asif, R., Merceron, A., & Pathan, M. (2014). Investigating performances' progress of students. In *Workshop Learning Analytics, 12th e_Learning Conference of the German Computer Society (DeLFI 2014)* (pp. 116–123). Freiburg, Germany, September 15.

Asif, R., Merceron, A., & Pathan, M. (2015a). Investigating performance of students: A longitudinal study. In *5th international conference on learning analytics and knowledge* (pp. 108–112). Poughkeepsie, NY, USA, March 16-20 http://dx.doi.org/10.1145/2723576.2723579.

Asif, R., Merceron, A., & Pathan, M. (2015b). Predicting student academic performance at degree level: A case study. *International Journal of Intelligent Systems and Applications (IJISA), 7*(1), 49–61. http://dx.doi.org/10.5815/ijisa.2015.01.05.

Baker, R. S. J. . d. (2010). Data mining for education. In B. McGaw, P. Peterson, & E. Baker (Eds.), *International encyclopedia of education* (pp. 112–118). Oxford, UK: Elsevier, 7(3).

Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining, 1*(1).

Bower, A. J. (2010). Analyzing the longitudinal K-12 grading histories of entire cohorts of Students: Grades, data driven decision making, dropping out and hierarchical cluster analysis. *Practical Assessment, Research & Evaluation, 15*(7), 1–18.

Campagni, R., Merlini, D., Sprugnoli, R., & Verri, M. C. (2015). Data Mining models for student careers. *Expert Systems with Applications, 42*(13), 5508–5521.

Cobo, G., Garcia, D., Santamaria, E., Moran, J. A., Melenchon, J., & Monzo, C. (2012). Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums. In *2nd international conference on learning analytics and knowledge* (pp. 248–251). Vancouver, Canada, April 29 – May 2.

ElGamal, A. F. (2013). An educational data mining model for predicting student performance in programming course. *International Journal of Computer Applications, 70*(17).

Elkina, M., Fortenbacher, A., & Merceron, A. (2013). .The learning analytics application lemo - rationals and first results. *International Journal of Computing, 12*(3), 226–234.
Feng, M., Heffernan, N. T., & Koedinger, K. R. (2006). Predicting state test scores better with intelligent tutoring systems: Developing metrics to measure assistance required. In *8th international conference on intelligent tutoring systems* (pp. 31–40). Berlin: Springer-Verlag.
Golding, P., & Donaldson, O. (2006). Predicting academic performance. In *36th ASEE/IEEE frontiers in education conference*.
Han, J., & Kamber, M. (2006). *Data mining concepts and techniques* (2nd ed.). San Francisco: Morgan Kaufmann (Chapter 1, Chapter 6 & Chapter 7).
Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*, 133–145.
Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies, 13*(1), 61–72.
Kabakchieva, D., Stefanova, K., & Kisimov, V. (2011). Analyzing university data for determining student profiles and predicting performance. In *4th international conference on educational data mining* (Eindhoven, the Netherlands).
Nghe, T. N., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *37th ASEE/IEEE frontiers in education conference*.
Oskouei, R. J., & Askari, M. (2014). Predicting academic performance with applying data mining techniques (Generalizing the results of two different case studies). *Computer Engineering and Applications Journal*, 79–88.
Pardos, Z., Hefferman, N., Anderson, B., & Hefferman, C. (2007). The effect of model granularity on student performance prediction using Bayesian networks. In *International conference on user modelling* (pp. 435–439). Berlin: Springer.
Pea, R., & the Learning Analytics Working Group. (2014). *A report on building the field of learning analytics for personalized learning at scale*. Stanford University, 2014, retrieved from https://ed.stanford.edu/news/stanford-professor-spurs-movement-build-new-field-learning-analytics.
Pelleg, D., & Morre, A. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. In *17th international conference on machine learning* (pp. 727–734). San Francisco, CA, USA.
RapidMiner retrieved from www.rapid-i.com.
Romero, C., Lopez, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education, 68*, 458–472.
Strecht, P., Cruz, L., Soares, C., Merdes-Moreria, J., & Abren, R. (2015). A comparative study of classification and regression algorithms for modelling students' academic performance. In *8th international conference on educational data mining* (pp. 392–395). Madrid: Spain.
Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *16th European Conference Artificial Intelligence (ECAI)*.
Vellido, A., Castro, F., & Nebot, A. (2010). *Clustering educational data. Handbook of educational data mining*. CRC Press.
Yehuala, M. A. (2015). Application of data mining techniques for student success and failure prediction (The case of Debre Markos university). *International Journal of Scientific & Technology Research, 4*(4), 91–94.
Zimmermann, J., Brodersen, K. H., Heinimann, H. R., & Buhmann, J. M. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining, 7*(3), 151–176.